PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

ESCUELA DE INGENIERIA

COMBINING DATA MINING AND ECONOMETRIC TECHNIQUES: A CASE STUDY ON ACADEMIC ACHIEVEMENT

VERÓNICA ANDREA PUGA DURÁN

Thesis submitted to the Office of Research and Graduate Studies in partial fulfillment of the requirements for the Degree of Master of Science in Engineering

Advisor:

MIGUEL NUSSBAUM VOEHL

Santiago de Chile, January 2018

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

COMBINING DATA MINING AND ECONOMETRIC TECHNIQUES: A CASE STUDY ON ACADEMIC ACHIEVEMENT

VERÓNICA ANDREA PUGA DURÁN

Members of the Committee:

MIGUEL NUSSBAUM VOEHL

PATRICIA GALILEA ARANDA

ERNESTO TREVIÑO VILLAREAL

ALDO CIPRIANO

Thesis submitted to the Office of Research and Graduate Studies in partial fulfillment of the requirements for the Degree of Master of Science in Engineering

Santiago de Chile, January 2018

To educational communities,
the protagonists of school improvement,
whose voice has been dimmed by our
educational system

ACKNOWLEDGEMENTS

On a personal level, I would like to thank my friends and family, who supported and accompanied me throughout this process. I cannot name them all, but they know how important their encouragement has been to me. I would like to specially acknowledge Bernard, Mónica and Julio. Bernard, father, thank you for your unconditional support during all my life, thank you for never judging me or pushing me to be someone else. Mónica, aunt, thank you for your unconditional love and for backing me up every time you could. Julio, partner, thank you for always being there for me, in the best and the worst.

I profoundly thank Miguel Nussbaum, my advisor. Thank you for believing in me and in this unconventional project, also for connecting me with people who nurtured my thesis. Thank you for the conversations and advice, both research and non-research related. I am grateful for working with an intelligent, humane and passionate man. I also thank Ernesto Treviño and Karim Pichara for helping me and guiding me in their areas of expertise, despite my limited knowledge in econometrics and data mining. This work was possible because of you.

I also thank four organizations. *Agencia de Calidad de la Educación*, for facilitating the datasets employed in this study and for responding my inquiries, specially Diego Nuñez, who responded the plentiful requests I submitted. *Mesa de Ayuda de Ingeniería UC*, for allowing me to use the cluster to run the algorithms, specially Mario Aguilera, who installed the packages I requested and helped me with the use of the cluster. I thank the Computer Science community, who welcomed me and helped me with the many programming doubts I encountered. Finally, *Comisión Nacional de Investigación Científica y Tecnológica*, CONICYT, for the scholarship *Beca de Magíster Nacional, Año Académico 2017*.

Finally, I thank the people and projects who inspired me through my undergraduate and graduate life. After all, the main motivation of starting this thesis was to learn a little bit more about how to build a fairer society, in this case, learning more about education. I thank Acerca Chile, Preuniversitario FEUC, Convive, Centro de Alumnos de Ingeniería UC and Ingenieros Sin Fronteras Chile, for keeping my spirit and convictions alive these 7 years. These projects, and more importantly, the people I met, will accompany me through the rest of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Data mining is changing econometric research. Although collaboration is expected between these two disciplines and that they can cope each other's limitations, to our best knowledge, there are not published attempts that show how data mining tools can complement econometric ones. This research proposes the Econometrics and Data Mining Dialogue approach, where an econometric model is built just from the data through data mining, without selecting variables based on bibliographic research or expert opinion. The approach was applied to a case study, predicting academic achievement in a longitudinal database. In total, we analyzed 142,457 students with 1,287 independent variables. We employed Random Forest, a data mining algorithm, to select a subset of variables to, posteriorly build an econometric model, an ordinal logistic multilevel model. Finally, we used Decision Trees, a data mining algorithm, to define student's achievement profiles. Most findings of our case study are consistent with academic achievement literature, like the relevance of prior academic achievement to present academic achievement. Other results offer fresher insights, like the impact of student's evaluation of their school on their academic achievement. This paper aims to contribute to the hands-on dialogue of how computer scientists and econometricians can collaborate to deepen the knowledge databases can offer and to improve econometric models.

RESUMEN

La minería de datos está cambiando la investigación en econometría. Aunque se espera colaboración entre estas dos disciplinas, hasta donde sabemos, no hay publicaciones que intenten mostrar cómo herramientas de la minería de datos pueden complementar herramientas de la econometría. Esta investigación propone el enfoque Econometría Dialoga con la Minería de Datos, donde un modelo econométrico se construye solo desde los datos a través de la minería de datos, sin basarse en investigación bibliográfica o la opinión de expertos para seleccionar variables. El enfoque se aplicó a un caso de estudio, predecir el rendimiento académico en una base de datos longitudinal. En total, analizamos 142.457 estudiantes asociados a 1.287 variables. Empleamos un Bosque de Árboles, algoritmo de la minería de datos, para seleccionar un subconjunto de las variables para luego construir un modelo econométrico, un modelo multinivel logístico ordenado. Finalmente, utilizamos un Árbol de Decisión, algoritmo de la minería de datos, para definir perfiles de desempeño de los estudiantes. La mayoría de los hallazgos de nuestro caso de estudio son consistentes con la literatura de desempeño escolar, como la relevancia del desempeño académico pasado para el desempeño académico actual. Otros resultados ofrecen ideas más frescas, como el impacto de la evaluación de los estudiantes de su escuela en el desempeño académico de los estudiantes. Esta investigación busca contribuir al dialogo aplicado de cómo científicos de la computación y econometristas pueden colaborar para profundizar los conocimientos que las bases de datos pueden ofrecer y para mejorar los modelos econométricos.

1. INTRODUCTION

Prediction is an old problem with a long history in econometric research (Mullainathan & Spiess, 2017). Econometric models tend to be theory-driven; traditionally they start from the thesis that specific variables have effect on a dependent variable (Fayyad et al., 1996). The selection of variables is based in theory built on previous research, expert opinion or the proposition of a new hypothesis; in any of these cases, the model is built in a highly subjective, slow and expensive way (Fayyad et al., 1996).

New disciplines such as big data and data mining are gaining attention (Sagiroglu & Sinanc, 2013; Yu et al., 2016). Einav and Levin (2014) state that these will probably change economic research, new methods will not replace common sense or economic theory, but will complement them and enable novel research designs. For example, when there is a prediction problem, an economist will probably think in a linear or logistic regression; however, data mining techniques provide other nonlinear and automated methods (Varian, 2014), e.g. Random Forest (Breiman, 2001), Support Vector Machines (Cortes & Vapnik, 1995), Neural Networks (Werbos, 1974), Nearest Neighbors (Cover & Hart, 1967), among others. These methods can improve the efficiency of treatment effects studies when there are many variables (Einav & Levin, 2014). These data-driven methods look for new relationships and findings instead of testing prior hypotheses (Slater, et al., 2017).

These theory- and data-driven modes of analysis have always coexisted and do not need to be in conflict (Mullainathan & Spiess, 2017). These two approaches complement each other; economic applications provide robust estimations of parameters that model relationships between variables and data mining provides useful tools to hear what data has to say (Mullainathan & Spiess, 2017). In fact, it is expected that collaboration between computer scientists and econometricians will be productive in the future (Varian, 2014).

In practice, data mining has been applied in diverse domains, for example, healthcare (Wang et al., 2017; Raghupathi, 2016), astronomy (Massaro et al., 2017; Ivezić et al.,

2014), business (Shmueli et al., 2017; Provost & Fawecett, 2013) and education (Angeli et al., 2017; Romero & Ventura, 2013). Applications can be as diverse as image processing (Gamal et al., 2017; Panda et al., 2017), speech recognition (Mustafa et al., 2017; Amodei et al., 2017), text analysis (Niekler et al., 2017; Cambria et al., 2013), among others. All these research evidences the productivity of data mining applications across domains.

The application of data mining tools to educational data is referred to as Educational Data Mining. An evidence of this growing research community is the International Data Mining Society, the Journal of Educational Data Mining and the Annual Conference of Educational Data Mining (Baker & Inventado, 2014). Education is an interesting domain for the application of data mining because it counts with diverse needs and many information from different actors, these data can be mined to obtain invaluable information (Ma, et al., 2000; Dutt et al., 2017). Tomar & Agarwal (2013) argue that data mining plays an essential role in uncovering new trends and hidden information in the healthcare area. There is no reason to believe that in the educational field data mining cannot play the same role.

However, to our best knowledge, there are not published attempts that show how data mining tools can complement econometric ones. In this work, we show a collaboration between computer scientists and econometricians proposing an approach that employs a data mining algorithm and an econometric model, and its application in a given domain (predicting academic achievement in a longitudinal database using student, classroom and school characteristics).

Data mining allows us to analyze enormous amounts of data and find novel and useful information (Chen et al., 2015). Econometric models produce good estimates of parameters that quantify relations between dependent and independent variables (Mullainathan & Spiess, 2017). A core idea of this study is that these different approaches combined can achieve better results and cope each other's limitations.

The research questions that drive this work are the following:

i) How can we combine data mining and econometric techniques?
ii) Can we build an econometric model just from data?
iii) In our specific case study (predicting academic achievement in a longitudinal database using student, classroom and school characteristics), what new findings do we discover?

2. THE APPROACH PROPOSAL: EDMD

In this section, we describe our proposed approach, EDMD, Econometrics (E) and Data Mining (DM) Dialogue (D). It is important not to look for a dividing line between these disciplines (Witten & Frank, 2005). To say that econometrics is more concerned with testing hypothesis and data mining with formulating a process to search possible hypotheses is an oversimplification (Witten & Frank, 2005). With this in mind, we propose a four steps approach that consists of: 1) define the objective, 2) prepare the data, 3) implement a data mining algorithm and 4) construct an econometric model. Below, we detail each step.

2.1. Objective: Collaboration within econometrics and data mining techniques

Our approach involves a dialogue between econometric and data mining techniques, nurturing one from each other. It combines the ability of data mining to analyze big data sets with the consistency of econometric models to estimate relations between variables. Therefore, a study employing our approach should satisfy the following:

i) Have access to an ample data set.
ii) Aim to find robust relations between variables.

Many domains have access to ample data sets and aim to find robust relations between variables, for example, healthcare, education, transportation, among others.

2.2. Data preparation: Statistical analysis and the pre-processing step

The statistical analysis allows to gain insights of the data, which will help with further analyses (Han et al., 2011). Before pre-processing the data, it is important to gain familiarity with the databases, e.g., examine variables and their distribution (Han et al., 2011).

Another relevant step is pre-processing. Building an adequate database means to organize the data together in the desired format. Problems may arise because data has to be collected from different areas, different areas may have different formats of registering data and data may need clean up (Witten & Frank, 2005). Some data preprocessing techniques are data integration, data cleaning, data reduction and data transformation (Han et al., 2011).

Our approach favors the analysis of the original data, hence we employed a small pre-processing stage to avoid modifying it. Specifically in this step, econometricians will provide theories related to the domain and some processing methods. Computer scientists provide knowledge in processing techniques that are outside the econometric domain. It is important to consider the data mining algorithm that will be used in the next step. A database adapted to the data mining algorithm will facilitate that the algorithm achieves its objective.

A small pre-processing stage would probably involve only data integration. Problems that arise when merging data from multiple sources are redundancy and entity identification. A variable is redundant if it can be deduced from other variables. The identification problem has to do with matching variables from different databases; it is important to match name, meaning, data type and allowed range values (Han et al., 2011). Data cleaning, reduction and transformation will modify in a deeper sense the original database. Nevertheless, the researcher may consider necessary to impute data, transform the data, among others.

To better organize the data preparation step we consider three sub steps: i) statistical analysis, ii) definition of the sample and data pre-processing and iii) adjust the database to the data mining algorithm. Table 2-1 presents the sub steps.

Table 2-1: Sub steps for the data preparation step.

Sub step	Objective
1. Statistical analysis.	Gain insights of the data, it will help with further analyses.
2. Definition of the sample and data pre-processing.	State the sample employed in the study and pre-process the data.
3. Adjust the database to the data mining algorithm.	Create a database that facilitates that the data mining algorithm achieves its objective.

2.3. Data mining: Selecting the algorithm

Choosing the proper data mining technique is a critical and difficult task. The main parameters to consider for selecting an algorithm are the goal of the problem to be solved and the data employed (Gibert et al., 2010).

An interesting proposal to start with is Witten et al. (2016), the authors provide an introduction to data mining, describe several algorithms and explain how to implement the algorithms through WEKA, an open-source software. It is central to review the applications of data mining in the domain being analyzed because each domain may favor different data mining algorithms. Since our case study is to predict academic achievement, below we refer to data mining work in the educational domain.

Silva & Fonseca (2017) analyze data mining techniques that have been applied to educational data. The main algorithms identified by the authors are neural networks, support vector machines, decision tree based methods and Bayesian classifiers. Han et al. (2011) and Witten et al. (2016) provide descriptions and examples of these algorithms. None of these methods is superior to all others for all types of data (Fernández-Delgado et al., 2014; Wolpert, 1996).

For example, Strecht et al. (2015) employed K-nearest neighbors (Fix & Hodges, 1951), Random Forest (Breiman, 2001), AdaBoost (Dietterich, 1997), Classification and Regression Trees (Breiman et al., 1984), Support Vector Machines (Vapnik, 2000) and Naive Bayes (Lewis & Ringuette, 1994) to predict if university students pass/fail a class and Ordinary Least Squares (Stigler, 1981), Classification and Regression Trees, Support Vector Machines, K-nearest neighbors, Random Forest, and AdaBoost.R2 (Drucker, 1997) to predict grades. In the pass/fail case, Support Vector Machines and Decision trees have better results. In the grades case, Support Vector Machines, Random Forest and AdaBosst.R2 obtained the best results. However, there was no statistical difference in terms of performance between algorithms.

Asif et al. (2017) predict the undergraduate grade at the end of a 4-year program with Decision trees, Naïve Bayes, Random Forests, Neural Networks, 1-nearest neighbor and rule induction. Naïve Bayes obtains the best results, followed by 1-nearest neighbor and Random Forest. The classifiers achieved better results than the baseline the authors built, the classifiers used pre-university grades and grades of the first and second year courses. The authors conclude that performance at the end of the 4-year program can be predicted at an early stage using grades only.

Finally, Cortez & Silva (2008) predict Portuguese and Mathematics achievement of students in secondary school with Decision Trees, Random Forest, Neural Networks and Support Vector Machines. During the school year, students are evaluated in three periods of time, the last period corresponds to the final grade. The authors build three datasets with different variables. One dataset contains grades of the first period, second period and information from questionnaires and school assistance. A second dataset excludes second period grades. The last dataset excludes first and second period grades. Predictions with information of prior grades were superior than predictions without these information, but some variables from the questionnaires were relevant like parent's job, parent's education and student's alcohol consumption. The algorithm with best results varied within dataset.

Decision Trees and Random Forest were superior than the other methods for most datasets.

These three studies compare algorithms and the best varies between each case. This is not only true for these cases; other studies compare algorithms and the best method varies between papers. More studies can be find in educational data mining reviews (Baker & Yacef, 2009; Romero & Ventura, 2007).

Some measures to compare algorithms are:

i) Accuracy or the quality of the results, it is the ability of the algorithm to predict correct results (Han et al., 2011).
ii) Speed and scalability, different algorithms involve different computational costs. This is of special interest when working with big data sets (Han et al., 2011).
iii) Robustness, i.e., ability of the algorithm to work well with noise or missing values.
iv) Interpretability, i.e., the level of insight provided by the algorithm (Han et al., 2011). For example, decision trees can be easy to interpret. More "black box" algorithms, for example, trained neural networks, require the implementation of algorithms to extract the knowledge embedded in them (Han et al., 2011).

To select the algorithm it is important to acknowledge the goal. The goal can be variable-related, for example to select, group or transform variables to employ on an econometric model (Variable, in Table 2-2). Or, to group or label samples to asses group characteristics in an econometric model, for example in education, to assess different student or teacher's profiles (Group, in Table 2-2). Another goal can be to analyze proximity between samples. For example, a data mining algorithm can specify which classes are similar according to some criteria (Proximity, in Table 2-2). These examples, summarized in Table 2-2, illustrate different possible objectives; it is not an exhaustive list of them.

We suggest contemplating the following questions to guide the process of analyzing the goal:

i) Once the goal is defined, what data mining algorithm is adequate?

ii) How will the data mining algorithm nurture the econometric model?

To help with these questions we provide Tables 2-2 and 2-3. Table 2-2 helps with question (i) by summarizing the examples described earlier per type of goal. Table 2-3 supports question (ii), offering examples of data mining work in the educational domain and associates them with a type of goal. In addition, we propose how each work could have nurtured an econometric model, if collaboration had existed.

Table 2-2: Examples per type of goal.

Goal	Examples
Variable	Select, group or transform variables to employ on an econometric model.
Group	Group or label samples to asses group characteristics in an econometric model.
Proximity	Specify which classes are similar according to some criteria to warn econometric models of possible difficulties.

Table 2-3: Type of goal matched with an example of data mining work in education and a mode in which the data mining algorithm could have nurtured an econometric model.

Goal	Authors	Data mining approach	Alternative econometric collaboration
Variable	Martínez & Chaparro, 2017	Use decision trees to detect student and school factors related to academic achievement.	The factors identified could have been used to build a multilevel model.
Grouping	Bresfelean et al., 2008	Build up a student's exams failure profile.	A variable could be created to acknowledge different profiles, and then it could be employed in an econometric model.
Proximity	Asif et al., 2017	Found that it is difficult for classification methods to predict minority classes.	Minority classes could be excluded from the sample. New thresholds that separate classes in a less unbalanced way can be proposed.

As mentioned before, the main parameters to select the algorithm are the goal and the data. We suggest contemplating the following questions to consider the data in this process:

i) How much computer power do I have available?
ii) What characterizes the structure of the sample?
iii) What types of variables does my sample have?

Question (i) refers to the computational resources available. Some data mining algorithms demand more computer power than others. Questions (ii) and (iii) refer to the structure and nature of data. These concepts are important for the data mining step and to the econometric step, which we refer to in the following section.

2.4. Econometric model: Selecting the model

Econometrics is about developing statistical methods for estimating relationships, testing economic theories and evaluating policies (Wooldridge, 2013). Just as in data mining, an important aspect to consider in econometric analysis is data (Wooldridge, 2013).

Models should be compatible with data structures, the most important structures are cross-sectional data, time series data, pooled cross-sectional data and panel data (Wooldridge, 2013). Cross-sectional data include a variety of units sampled at a given point in time. Time series data are observations on one or more variables over time. Pooled cross-sectional data have cross-sectional and time series features, for example by combining data of a questionnaire employed in two different years. Panel data is like pooled cross-sectional data, but units are followed over time (Wooldridge, 2013).

Models should also be compatible with types of variables. What defines the type of variable is the possible values that it can assume. The typical ones are nominal, binary, ordinal and numeric (Han et al., 2011). A description of these types of variables can be found in data mining and econometric books (Han et al., 2011; Witten & Frank, 2005; Wooldridge, 2013).

The amount of econometric methods available may seem confusing (Angrist & Pischke, 2008). Each domain counts with books and reviews of how to model problems, e.g., Scott & Usher, (2010) for education, Bowling (2014) for healthcare and Brooks (2014) for finance. It is important to know which econometric models are appropriate for the problem analyzed.

2.5. The iterative process

Our approach consists of four steps: a) objective, b) data preparation, c) data mining algorithm, and, d) econometric model. It is an iterative and sequential process with more than one possible successive step. Figure 2-1 summarizes our approach and describes each connection with letters. Connection a), b) c) and d) shows the sequential implementation of the four steps. Connection e) represents the path when data needs a second processing stage before building the econometric model. Connection f), g) and h) close the cycle and allow the process to be iterative. The researcher may want to finish the investigation once they finish the econometric model or may want to iterate. In path f), to nurture a second data mining algorithm with the results of the econometric model. In path g), to process the data and go back to the data preparation step. Finally, path h) states a new objective and starts with the process again.

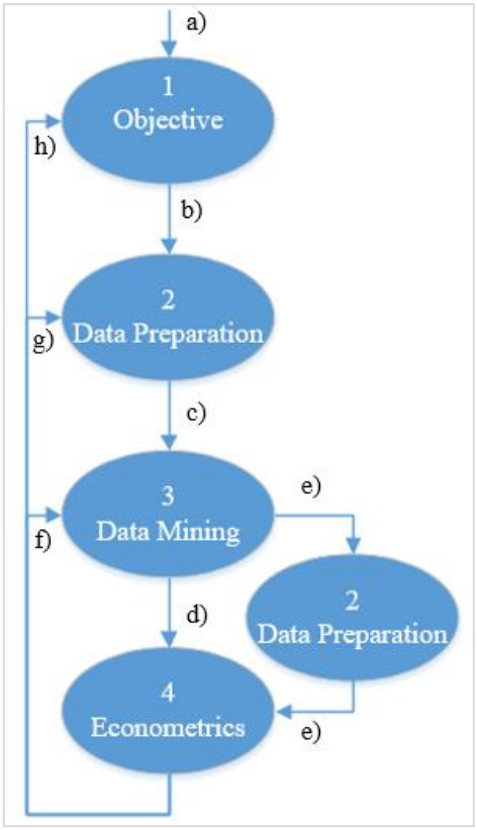


Figure 2-1: Diagram of our model.

3. LITERATURE REVIEW OF OUR CASE STUDY

In this section, we explore literature about academic achievement and academic achievement as a cumulative process. Then, we provide examples of data mining work in predicting academic achievement.

3.1. Academic achievement

Academic achievement started to grow as a research line with Coleman's (1966) and Plowden's (1967) reports. These studies wanted to determine school's effect in academic achievement; both concluded that school had little impact on it. The socioeconomic status of the student's family was the most important predictor of academic achievement, contradicting the notions deeply ingrained in the academic community, which criticized the methodology and the interpretations of these results.

Academic achievement is still a widely researched topic. An adequate statistic technique to analyze academic achievement are Hierarchical Lineal Models, also known as Multilevel Models. These models seek to explain a variable recognizing that the unit belongs to a group. Therefore, the variance of the student's academic achievement can be decomposed in these groups. As a result, the model takes into account the nature of the data, for example, that students are clustered in schools and share common influences (Raudenbush & Bryk, 2002).

Today, academic achievement has been studied from several perspectives. Some novel examples are Hattie (2009) and Bryk et al (2010). Hattie compares the effectiveness of interventions that seek to increase academic achievement. Bryk et al. study school improvement from an organizational perspective. Both studies identify factors related to academic achievement, e.g., learning climate, professional expertise of educators, motivation to learn, self-perception and ties with the community and parents. Yet, the both conclude that there is no silver bullet and improvement requires orchestrated initiatives in multiple areas.

3.2. Academic achievement as a cumulative process

Several studies have shown that academic achievement is highly correlated to prior academic achievement (Arnold & Doctoroff, 2003; Duncan et al., 2007; Adelson et al., 2016). Success in early school experience enhances motivation, while students who struggle become discouraged and disengaged (Herbers et al., 2012).

Learning is a cumulative process in which students improve existing skills while developing new ones (Duncan et al., 2007). Complex forms of learning build on simpler forms of learning, therefore, inequalities in a stage create greater differences in future stages (Caro, 2009). Studies have found that the gap between high and low socioeconomic status students increases with time, students with high socioeconomic status improve over time while students with low socioeconomic status fall further and further behind (Caro, 2009; Jimerson et al., 1999), even when controlling for early achievement (Herbers et al., 2012). Other variables related to changes in gap achievements are gender, ethnicity, parent involvement and home environment (Herbers et al., 2012; Jimerson et al., 1999).

3.3. Data mining work in academic achievement

Data mining work in academic achievement prediction focuses mainly in predicting evaluation performance or student dropout. There are several works involving university contexts, below we describe three of them. Oladokun et al. (2008) use Neural Networks to predict performance on graduation of Nigeria's undergraduate students; the model correctly predicts the performance of more than 70% of the students. Asif et al. (2017) predict grades at the end of a 4-year bachelor degree at a University in Pakistan. The authors employ Decision trees, Naïve Bayes, Random Forests, Neural Networks, 1-nearest neighbor and rule induction. Naïve Bayes obtains the best results, the model correctly predicts more than 80% of the sample. Nghe et al. (2007) predict final year grades in a Thai and a Vietnamese University, they compare Bayesian Network and Decision Tree, this latter algorithm produces more accurate results. The authors worked with several datasets, for example, predicting performance according to grades (fail, fair, good, very

good) versus predicting performance according to failing status (fail, pass). The percentage of correctly predicted performance varies between 64% and 94%.

Schools also seek to predict academic achievement. Cortez & Silva (2008) use Decision Trees, Random Forest, Neural Networks and Support Vector Machines to predict grades in Mathematics and Portuguese classes of secondary students. The authors run several models, but, in general, the nonlinear methods (Neural Networks and Support Vector Machines) are outperformed by the tree based ones (Decision Trees and Random Forest). The authors argue that this may be because of the high number of irrelevant inputs.

However, students are not only evaluated by their schools; many countries have national standardized tests but there is less research employing this data. Ma et al. (2000) use scoring techniques to predict students with low results in A-Level, a Singapore's National test, to assist them before they give the test. Martínez & Chaparro (2017) use Decision Trees to predict students result in ENLACE, a national standardized test in Mexico.

A common tendency of the research mentioned above is that their main objective is to make an accurate prediction. Understanding relationships between variables or quantifying their impact on the prediction becomes secondary or generally, not even assessed. Therefore, one of the limitations is failing to understand in a deeper sense the academic achievement phenomena.

4. METHODOLOGY

In this section, we describe the application of the proposed model to our specific case study, predicting academic achievement in a longitudinal database using student, classroom and school characteristics.

4.1. Objective

The Chilean educational system employs a nationwide standardized test, SIMCE, to measure school's educational quality. The evaluation includes not only tests in different subjects, but questionnaires to students, teachers and parents or guardians. In fact, between 2016 and 2017 a questionnaire to school principals was introduced (Agencia de Calidad de la Educación, 2016a; Agencia de Calidad de la Educación, 2016b). These questionnaires seek to identify and validate school internal and external factors that influence academic and non-academic results (Agencia de Calidad de la Educación, 2014).

We studied the cohort of students that had been assessed the most with standardized tests in recent years, i.e., fourth, sixth and eighth grade students during 2011, 2013 and 2015 respectively. We used as the dependent variable the achievement in the language test. The independent variables included school context and the questionnaires, in total, each student was associated to 1,287 independent variables.

This is a perfect context to apply our model. We employed a data mining algorithm, Random Forest, to reduce the number of variables. With the reduced set, we built an econometric model, the ordinal logistic multilevel model. Finally, we used a data mining algorithm, Decision Trees, to identify and define a student's academic achievement profile.

4.2. Data preparation for data mining

Table 4-1 presents the three sub steps of data preparation stated before and describe what we did in each sub step. These sub steps organize this section.

Table 4-1: Sub steps of the data preparation for the data mining algorithm.

Sub step	Objective	What we did
1. Statistical analysis.	Gain insights of the data, it will help with further analyses.	i) Described the database employed. ii) Described the dependent and independent variables. iii) Analyzed relations between the dependent and independent variables.
2. Definition of our sample and data pre-processing.	State the sample employed in the study and pre-process the data.	i) Described how we defined our sample. ii) Created a socioeconomic indicator. iii) Compared the sample with the complete database and stated any bias in our sample.
3. Adequate the database to the data mining algorithm.	Create a database that facilitates that the data mining algorithm achieves its objective.	i) Defined different models we were interested in analyzing. ii) Defined student's profiles and separate the data according to it. iii) Merge similar outputs.

4.2.1 Statistical analysis

In this sub step, we describe the database employed, describe the dependent and independent variables and analyze relations between the dependent and independent variables.

Regarding the database employed, we needed students that gave the same test in all three grades, i.e., that gave the fourth grade test in 2011, the sixth grade test in 2013 and the eighth grade test during 2015. 151,332 students satisfy this condition. Table 4-2 presents the number of students enrolled nationally in each grade and the number of students that gave the standardized test per grade.

Table 4-2: Number of students per level vs number of students that give the test.

	Fourth grade (2011)	Sixth grade (2013)	Eighth grade (2015)
Total students enrolled nationally	247,666	262,421	255,607
Students that give the test	216,133	219,856	214,510

Regarding the dependent variable, language academic achievement, we used Chile's Quality of Education Agency standards (Agencia de Calidad de la Educación, 2014). The Agency sets thresholds according to the student's fulfillment of the national curriculum and defines three achievement levels: adequate, elemental and insufficient (Agencia de Calidad de la Educación, 2014). These academic achievement levels are particular to the Chilean educational system; hence, we created a more generalizable threshold based on percentiles: an upper, medium and lower third. We used these two thresholds throughout our study. Appendix A gives a detailed explanation of the Agency's thresholds.

Figure 4-1 shows the percentage of students per language achievement level for fourth, sixth and eighth grade (we used the Agency threshold, the division by percentiles divides the sample in equal thirds). The figure shows the distribution for the total universe of students that give the test and for the group of students that satisfied our requirement.

Our sample has better academic achievement results than the whole universe of students. A reason may be that students that failed a school year between 2011 and 2015 are out of our study, this bias is acknowledged as a limitation of our study. Nevertheless, the tendency persists: In fourth grade, most students have an adequate level of fulfillment of the national curriculum (41% in the total universe and 46% in our sample), whereas in eighth grade, most students have an insufficient level of fulfillment of the national curriculum (49% in the total universe and 43% in our sample).

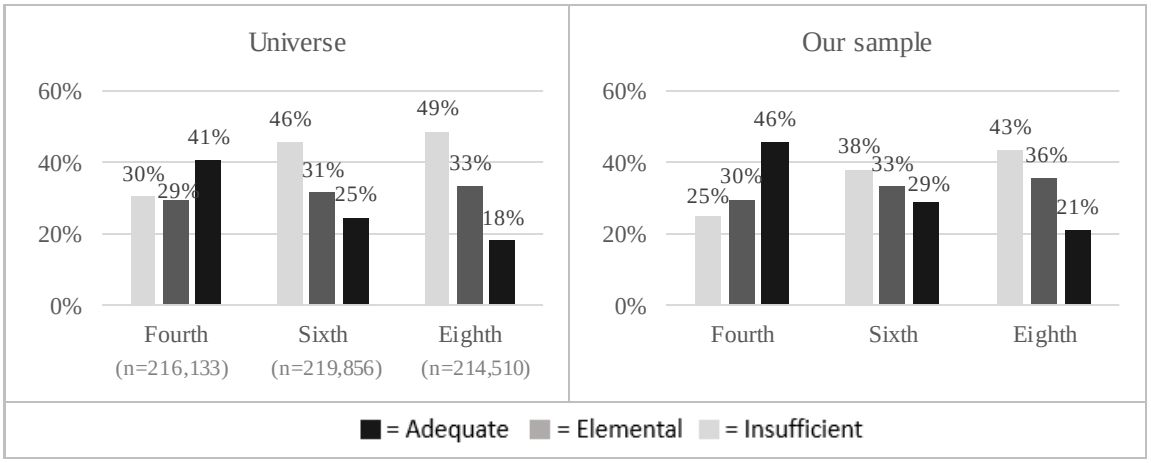


Figure 4-1: Students per achievement level for the total students that sit for the test and for the sample employed in our research.

Appendix B analyzes the distribution of academic achievement according to the achievement level in a prior grade. For example, from all fourth grade students that start on the upper level, 20% decrease to the lower one in eighth grade. From all students that start in the lower level in fourth grade, only 2% end in the upper level in eighth grade.

Regarding the 1,287 independent variables, 1,264 come from questionnaires to students, teachers and parents or guardians. 397 variables come from the fourth grade's questionnaires, 361 from the sixth grade's questionnaires and 506 from the eighth grade's questionnaires. Questionnaires gather information about school internal and external factors related to academic and non-academic results like school climate, self-perception, healthy habits, among others (Agencia de Calidad de la Educación, 2014).

From the 23 variables left, 15 characterize schools. Per grade (fourth, sixth and eighth grade), each school has five variables. Two are socioeconomic status variables (one is nominal provided by the Quality of Education Agency and the other one is continuous built by us). The other three are school size (total number of students), rurality (located in a rural or urban area) and financial dependency (privately paid, public subsidized or public).

From the eight variables left, each student counts with three binary variables that acknowledge change of school status; changed of school between fourth and sixth grade, fourth and eighth grade and sixth and eighth grade. Finally, each student counts with five non-temporal variables. One variable is gender, the remaining four are socioeconomic status indexes built from the questionnaires: number of books at the house, educational level of the mother, household's income and an indicator that combines all three prior indexes.

Regarding relations between the dependent and independent variables, Appendix C shows how scores varied for different variables. The achievement gap did not change for gender, school's socioeconomic status and school financial dependency. These results contradicts the literature presented in section 3.2. Academic Achievement as a cumulative process.

4.2.2 Definition of our sample and data pre-processing

In this section, we describe how we defined our sample, describe how we created a socioeconomic indicator and compare the sample with the complete database.

We acknowledge the importance of socioeconomic status to academic achievement research (Sirin, 2005). Hence, we demanded that all samples had information about number of books in the house, mother's educational level and household's income. With this information, we created a socioeconomic status indicator. The fourth, sixth and eighth grade questionnaires have information about these three variables, i.e., each student has different socioeconomic status indexes for each level. However, only 62% of the sample has information about these three topics for all three grades. We compared correlation between grades and found that number of books at the house, educational level of the mother and the household's income correlated highly for the same student. Table 4-3 shows the correlation coefficients.

Table 4-3: Correlation coefficients for number of books, mother's educational level and household's income.

	Books			Educational level			Income		
	Fourth	Sixth	Eighth	Fourth	Sixth	Eighth	Fourth	Sixth	Eighth
Fourth	1			1			1		
Sixth	0.63	1		0.85	1		0.87	1	
Eighth	0.59	0.63	1	0.83	0.83	1	0.83	0.85	1

Hence, we used these variables as constants through time, i.e., we used one indicator for number of books, one for the mother's educational level and one for the household's income. We used the eighth grade information if it was available, if it was not, we used the sixth grade information, if it was not available either, we used the fourth grade information. We grouped these variables into one socioeconomic status indicator with factor analysis using the *polychoric* command to acknowledge the ordinal nature of the variables (Kolenikov, 2016).

With this we lose less than 1% of the sample, leaving us with 149,825 samples instead of 151,132. Before, we mentioned two school socioeconomic status variables. The Quality of Education Agency provides a nominal one and we added one of our own. The one we added comes from the socioeconomic status index we created, we averaged the values of students per school and created a school mean socioeconomic status, this socioeconomic status variable is continuous. We built this socioeconomic status indicator because the Quality of Education Agency does not provide a socioeconomic status index for each student. Furthermore, the construction of attributes based on others may improve the accuracy of data mining algorithms in high dimensional data (Hat et al., 2011), which is our case.

Finally, we compared our sample with the complete database to identify any bias in our sample. In Appendix D, we analyze the distribution of students per gender, school's socioeconomic status and school dependency for all students coursing nationally each level, students that gave the test and the students in our sample. We found that female

students, students that attend privately paid and public subsidized schools and students that attend schools with higher socioeconomic status are slightly over represented in our sample.

4.2.3 Adjusting the database to the data mining algorithm

In this section, we define models we were interested in analyzing, separate the data according to student's profiles and merge similar outputs.

First, we decided to analyze if there were any differences depending on the grades analyzed (fourth, sixth and eighth), for example, if the relevant variables to predict achievement level in sixth grade were different from the relevant variables to predict achievement in eighth grade. We defined three different models to predict trajectories: from fourth to sixth grade (1, in Figure 4-2), from sixth to eighth grade (2, in Figure 4-2) and from fourth to eighth grade (3, in Figure 4-2). Figure 4-2 illustrates the three different trajectories analyzed.

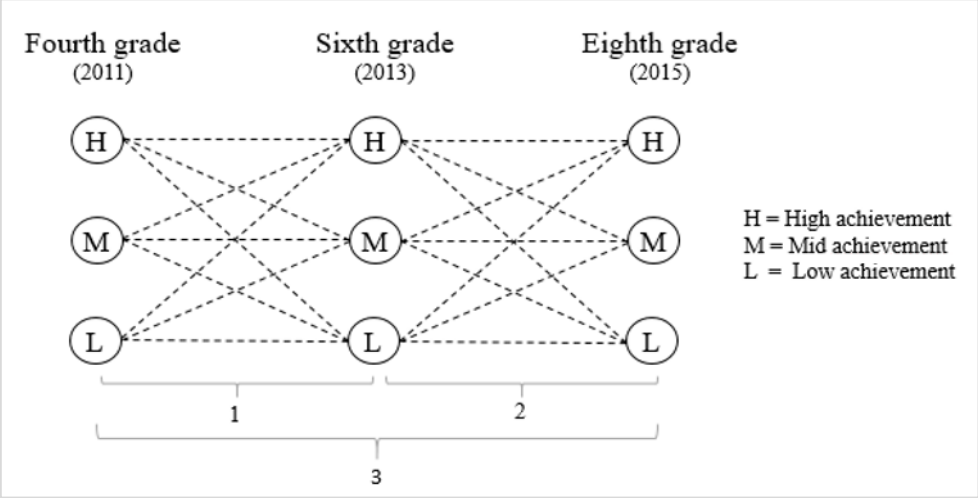


Figure 4-2: Illustration of the three different time lapses analyzed and the trajectories a student can have.

The model that predicts the trajectory between fourth and sixth grade uses variables from these two grades (774 independent variables), the same happens for the trajectory between sixth and eighth grade (883 independent variables). The model that predicts the trajectory between fourth and eighth grade employs all the 1,287 independent variables. Table 4-4 illustrates how many variables each trajectory has.

Table 4-4: Number of variables per trajectory.

	Trajectory Fourth - Sixth			Trajectory Sixth - Eighth			Trajectory Fourth - Eighth		
Variables	Fourth	Sixth	Eighth	Fourth	Sixth	Eighth	Fourth	Sixth	Eighth
Questionnaire	397	361	-	-	361	506	397	361	506
School	5	5	-	-	5	5	5	5	5
Change of school	1 (between fourth and sixth)			1 (between sixth and eighth)			1 (between fourth and sixth) 1 (between sixth and eighth) 1 (between fourth and eighth)		
Non-temporal variables	5			5			5		
TOTAL	774			883			1,287		

We applied algorithms in each database (fourth-sixth, sixth-eighth and fourth-eighth), but, the algorithms did not achieve good predictions. Hence, we separated the datasets according to initial achievement levels (high, mid and low) and build different datasets and models.

Finally, we merged similar outputs together. A standard way of modelling a problem with more than two classes is modelling it as a two-class situation (Witten & Frank, 2005). We compared the results of predicting between three academic achievement levels and two academic achievement levels, the case with two possible outcomes obtained better results. Since there are three achievement levels, we grouped the two most similar ones together. We tried the two possible groupings: i) high achievement with mid achievement and ii) mid achievement with low achievement. We grouped the levels that obtained better results.

If a student started in the upper or medium level (for fourth and sixth grade), the prediction was if the student decreased the achievement level or not (a. and b. in Figure 4-3). If a student started in the lower level (for fourth and sixth grade), the prediction was if the student increased the achievement level or not (c. in Figure 4-3). Figure 4-3 illustrates the grouping between trajectories; one group is marked with solid lines and the other with dashed lines. As in Figure 4-2, "H" stands for high, "M" for mid and "L" for low achievement.

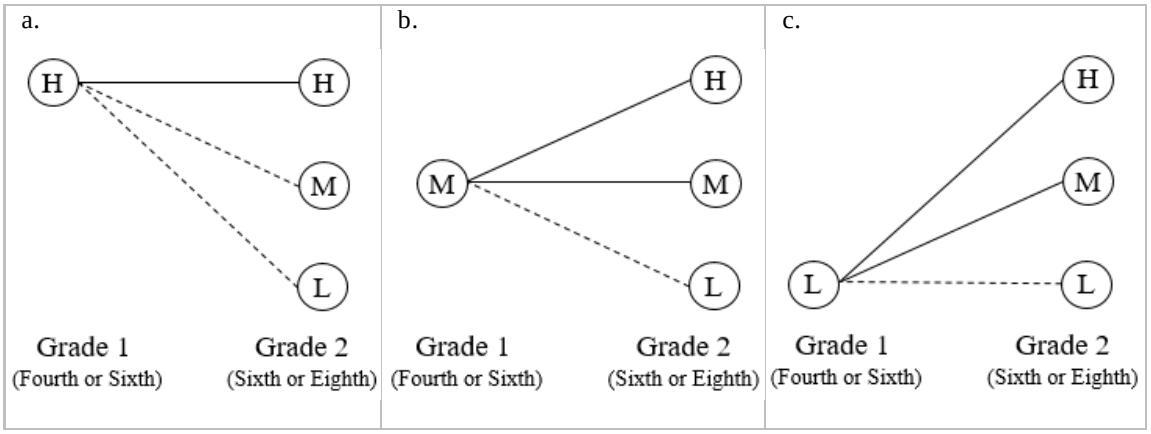


Figure 4-3: Illustration of the grouping between trajectories.

Because of the three time lapses (fourth-sixth, sixth-eighth and fourth-eighth) and the three achievement levels (high, mid and low achievement), we had to prepare nine datasets for nine different models. Since we are assessing two thresholds (The Agency's and the percentile's threshold), in total we had to elaborate 18 datasets and models. Table 4-5 illustrates the nine different models for a given threshold.

Table 4-5: Detail of the nine different data sets and models for a given threshold.

Achievement level	Fourth – Sixth	Fourth – Eighth	Sixth – Eighth
High initial achievement	1	4	7
Mid initial achievement	2	5	8
Low initial achievement	3	6	9

4.3. Data mining for variable reduction

Just as stated in the description of our proposed approach, it is important to acknowledge the goal of the data mining algorithm and the data structure to select the algorithm.

Our goal is to select relevant variables from the 1,287 independent variables; as a quality measure we checked the ability of the algorithm to predict the academic achievement level. The variables selected are used in the econometric model.

Regarding the structure of our data, it has a panel structure because we follow students through time. In addition, the database is high dimensional because it counts with 1,287 independent variables and it counts with different types of variables: binary, nominal, ordinal and numeric. It is probable that many variables will be irrelevant for predicting academic achievement. Moreover, it is highly unbalanced because academic achievement is highly correlated with prior academic achievement. For example, the database with low achievement students in fourth grade has 80% students that remain in the low level in eighth grade, and only 20% students that increase the academic achievement level.

We selected the Random Forest algorithm for the following reasons:

i) High dimensional data, probably with irrelevant features.
   Random Forests tend to perform well with high dimensional data and irrelevant features, this is not the case of other classifiers like Support Vector Machines and Neural Networks (Cortez & Silva, 2008).
ii) Simple and easily parallelized.

Random Forests are easy to implement and parallelize, hence, they can run fast and require few computer memory. In addition, it is easy to extract knowledge from the algorithm. Our goal is to identify relevant variables and Random Forest has implemented methods to extract a ranking of variable importance (Breiman, 2001).

iii) Categorical and quantitative variables.

Random Forest is an algorithm that does not require much pre-processing and is able to work with different types of variables (Luan, 2002; Martínez & Chaparro, 2017).

iv) Unbalanced datasets, probably with outliers and noise.

Random Forest works well with unbalanced datasets and are robust to outliers and noise (Liu, et al., 2013; Breiman, 2001). Our datasets are unbalanced and may have outliers and noise because we did not employ a cleaning process. We intentionally did not employ a cleaning process because we wanted to analyze the original data with the data mining algorithm.

The basic unit of Random Forest are Decision Trees, hence, we provide a brief explanation of them in the following section.

4.3.1 Decision Trees

Decision Trees are tree-like graphs in which each node uses a variable to separate the sample according to criterions. Normally, variables are compared with constants, but sometimes two variables can be compared with each other (Witten & Frank, 2005). For example, all samples that have variable "v" larger than "x" go to one branch and the rest of the sample goes to another branch. Variables can be assessed as categorical values too, for example, variables that have a specific value go to one branch, in fact, missing values can be treated as a special category (Witten & Frank, 2005). To better illustrate Decision Trees, Figure 4-4 is an examples provided by Witten & Frank (2005).

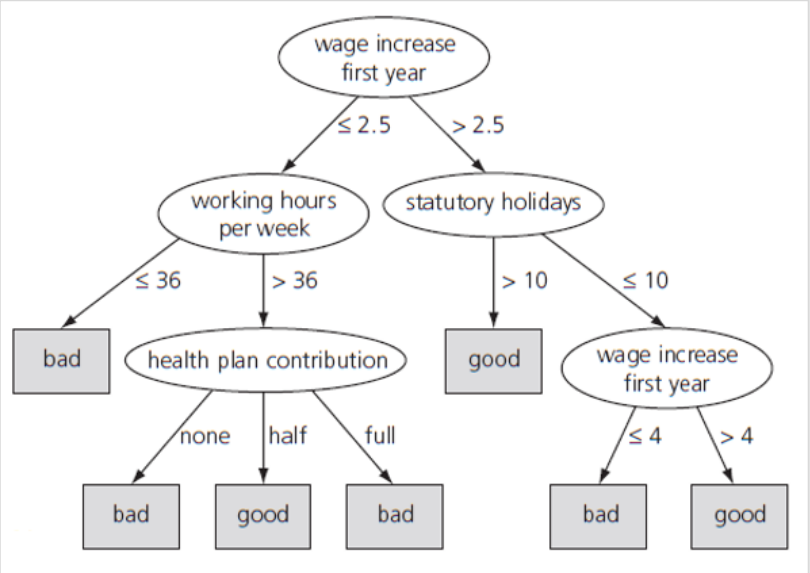


Figure 4-4: Decision Tree example (Witten & Frank, 2005).

Decision Trees are built in a recursive process. First, they select the variable to place at the root node to separate the sample, this splits the sample into leaves (Witten & Frank, 2005). If the leave has samples of different classes it becomes a node that will employ a variable to continue separating the sample (Witten & Frank, 2005). When a leave has samples of only one class the process is finished for that branch (Witten & Frank, 2005). After the tree is built, a pruning process can me employed to simplify the Decision Tree. To clarify the process of building a tree, we show an example explained in Witten & Frank (2005).

The weather dataset

According to the weather, a decision has to be made whether to go out and play or not. To model this decision a dataset has information about four variables and the decision made. Table 4-6 show the dataset available to model the decision. Table 4-7 summarize the number of "yes" and "no" for the outcome "Play" for each variable.

Table 4-6: Weather dataset (Witten & Frank, 2005).

Instance	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Table 4-7: Weather data with counts and probabilities of "yes" and "no" for the outcome "Play" (Witten & Frank, 2005).

Outlook			Temperature			Humidity			Windy			Play	
	Yes	No		Yes	No		Yes	No		Yes	No	Yes	No
Sunny	2	3	Hot	2	2	High	3	4	High	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	Normal	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	High	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	Normal	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

The variable selected to place at the root node is the one that separates the "yes" and "no" the most, i.e., creates the purest daughter nodes. The measure of purity of nodes is normally referred as information gain and is modeled with the entropy function. When the function has only one type of class, the value is zero. When the number of "yes" and "no" is equal it reaches a maximum value. Equation 4-1 shows the entropy function:

$$entropy(p_1, p_2, ..., p_n) = \sum_i -p_i \log_2 p_i$$
$$p_1, p_2, ..., p_n = probability\ of\ choosing\ class\ p_n \quad (4\text{-}1)$$

Figure 4-5 shows the separation of "yes" and "no" each variable achieves. With this, we calculate the information gain for each variable and select the one to place at the root node.

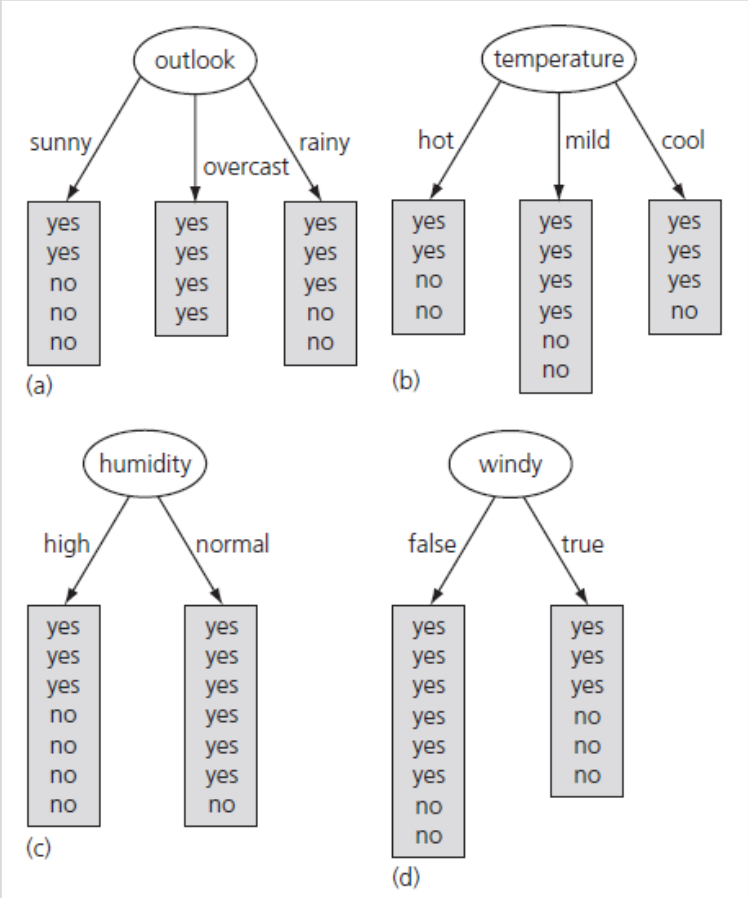


Figure 4-5: Separation of "yes" and "no" for each variable (Witten & Frank, 2005).

The variable "outlook" separates the sample into three leaves (image (a) in Figure 4-5) . The entropy of each leave is the following:

$$sunny = entropy(\frac{2}{5}, \frac{3}{5}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971 \quad (4\text{-}2)$$

$$overcast = entropy(\frac{0}{5}, \frac{5}{5}) = 0 \quad (4\text{-}3)$$

$$rainy = entropy(\frac{3}{5}, \frac{2}{5}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971 \quad (4\text{-}4)$$

Then, we average proportionally these values to obtain the entropy of the variable "outlook":

$$overlook = \frac{5}{14}\,0.971 + \frac{4}{14}\,0 + \frac{5}{14}\,0.971 = 0.693 \quad (4\text{-}5)$$

The total sample contains 9 "yes" and 5 "no", hence, the initial entropy is the following:

$$entropy(\frac{9}{14}, \frac{5}{14}) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.940 \quad (4\text{-}6)$$

The information gain obtained by the variable "overlook" is 0.940 – 0.693 = 0.247. The information gain of the other variables are 0.029 for "temperature", 0.152 for "humidity" and 0.048 for "windy". The variable to place at the root node is "overlook" because it achieves the maximum information gain. Figure 4-6 shows the complete Decision Tree.

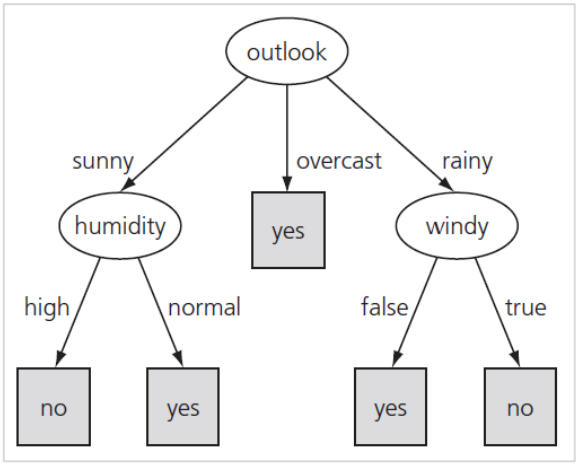


Figure 4-6: Complete Decision Tree for the weather dataset (Witten & Frank, 2005).

Decision Trees can grow until classes are completely separated. Pruning is a technique to reduce the size of trees, reduce their complexity and to avoid overfitting. Overfitting occurs when the classifier is overly fit to the sample, this traduces in poor classification results when new samples are available, i.e., the model cannot be generalized for out of the sample instances (Witten & Frank, 2005).

4.3.2 Random Forest

Random Forest (Breiman, 2001) is an algorithm that uses Decision Trees and incorporates two main ideas: "bagging" (Breiman, 1996) and "the random subspace" (Ho, 1998).

"Bagging" refers to generating new training sets by randomly sampling the original dataset, this is also referred as bootstrap aggregating. Therefore, each Decision Tree is trained with a different training set built using the original sample. This technique reduces variance and avoids overfitting (Breiman, 2001). The number of training sets generated, hence the number of Decision Trees trained, is a parameter that has to be set.

"The random subspace" is about employing a random subset of the features to select the one to separate the sample, i.e., the best variable is selected among a subset of them, not

from all available variables; this is also referred as feature bagging. This practice reduces correlation between classifiers and can increase accuracy of the algorithm (Breiman, 2001). The number of variables analyzed for each split is a parameter that has to be set too. We tuned this parameter by trying different values and selecting the one with smallest error.

Hence, each tree has two different bagging processes. First, bagging the training set, second, bagging the features for each node. The samples from the original dataset that are left out of the training set are used for the testing set.

For each sample of the original dataset, the prediction of all Decision Trees that do not contain that sample in the training set are aggregated. This is called out-of-bag classifier, with this, an out-of-bag error is estimated. Out-of-bag error represents the prediction error of the algorithm, it converges as the number of Decision Trees in the Random Forest increases. After a threshold, the error converges to a limit.

4.3.3 Variable selection with Random Forest

In R, we ran Liaw & Wiener's (2015) application of Random Forest, based in Breiman and Cutler's original Fortran code (Breiman & Cutler, 2004). It assesses variable importance with two indicators: Mean gini decrease and mean accuracy decrease.

Mean gini decrease is the total decrease in node impurity from separating samples with the variable, averaged over all trees, the indicator is the Gini Index (Liaw & Wiener, 2015). The Gini Index has the same characteristics as the entropy formula explained before: when a leaf has only one type of class, the value is zero, when the number of different classes is equal it reaches a maximum value. The Gini Index is calculated with Equation 4-7.

$$gini(p_1, p_2, ..., p_n) = \sum_i p_i(1 - p_i) \tag{4-7}$$

Mean accuracy decrease is based on out-of-bag error. For each Decision Tree, the out-of-bag error is recorded, then the same is done after dropping a predictor variable. The

difference between this new out-of-bag error and the original out-of-bag error is employed as the indicator of how much the accuracy of the model decreases due to the elimination of that variable (Liaw & Wiener, 2015). This indicator is calculated for each variable too. We selected the variables with biggest decrease in these indicators. Hence, each of the 18 models (nine for each threshold) has two rankings, one based on mean decrease in gini and one based on mean decrease in accuracy. In total, we get 36 variable's importance rankings.

4.3.4 Re-balancing and partitioning data

During the implementation process, we took special care of two issues: re-balancing data and partitioning of it. Since the datasets are highly unbalanced, we used two methods to re-balance it. One method was to over-sample the minority class with the SMOTE algorithm (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002; Han et al., 2011). The other method was to under-sample the majoritarian class and have equal proportion of classes train the algorithm (Han et al., 2011). Results reported in this work correspond to this latter method because predictions were more accurate. To partition data and to obtain statistical robust results we employed stratified 10-fold cross validation (Witten & Frank, 2005). Stratified 10-fold cross validation means to split randomly the data into 10 parts holding the proportion of classes as in the complete dataset. Instead of training the algorithm one time, the algorithm is trained 10 times with nine-tenths of the sample and tested with the one-tenth held out. Hence, the training procedure is executed 10 times on different training sets. Finally, the 10 errors are averaged to obtain an overall error estimate (Witten & Frank, 2005).

We employed 1,000 trees for each fold, since we employed 10-fold cross validation, each model was built with 10,000 Decision Trees. We tuned this value by building larger forests. We observed that after 50 trees the out-of-bag error decreased slowly, after 300 trees the out-of-bag error almost did not decreased. To have a big margin we used 1,000 Decision Trees.

4.3.5 Evaluating the classification algorithm through confusion matrices

Confusion matrices are useful for visualizing the performance of a classification algorithm. The predicted values go on columns and the actual values go on rows, or inversely. This helps to see if the algorithm confuses certain classes, or how is the prediction for each class.

Recall and precision are two indicators to evaluate the quality of the prediction per class. Recall is the number of items selected correctly divided by the total number of items of that class. Precision is the number of items selected correctly divided be the total prediction of that class. Table 4-8 is an example provided by Witten & Frank (2005), it illustrates the use of confusion matrices and the indicators mentioned before.

Table 4-8: Example of a confusion matrix (Witten & Frank, 2005).

		Predicted class				
		a	b	c	TOTAL	recall
Actual class	a	88	10	2	100	88%
	b	14	40	6	60	67%
	c	18	10	12	40	30%
TOTAL		120	60	20		
precision		73%	67%	60%		

The dataset contains 100 samples of class “a”, 60 of class “b” and 40 of class “c”, as shown in column “TOTAL” in Table 4-8. The algorithm predicts 120 samples of class “a”, 60 of class “b” and 20 of class “c”, as shown in row “TOTAL” in Table 4-8. The diagonal of the matrix shows the samples that were correctly predicted, 88 of class “a”, 40 of class “b” and 12 of class “c”. The samples that were miss classified are outside of the diagonal. For example, 14 samples were classified as class “a” but their actual class is “b”, as shown in column “a” and row “b” in Table 4-8.

Table 4-8 also shows precision and recall. For example, class “c” has 30% as recall because from the 40 actual samples of class “c”,12 were correctly predicted, 12 divided

by 40 gives the 30% of recall. The precision of class "c" is 60% because from the 20 samples predicted as class "c" only 12 were correctly predicted, 12 divided by 20 gives the 60% of precision.

4.4. Data preparation for the econometric model

We processed the data again to prepare it for the econometric model. Table 4-9 details what we did in each sub steps. First we did a statistical analysis, then we state the sample employed and preprocessed the data and finally, we adjusted the database to the econometric model.

Table 4-9: Sub steps of the data preparation for the econometric model.

Sub step	Objective	What we did
1. Statistical analysis	Gaining insights into the data will help with posterior analysis.	We did this sub step in section 4.2 Data preparation for data mining. The original dataset remains the same, so we did not conduct a second statistical analysis.
2. Definition of our sample and data pre-processing.	State the sample employed in the study and pre-process the data.	i) Merged the datasets. ii) Deleted variables with too many missing or double marked values. iii) Imputed the remaining missing and double marked responses. iv) Added control variables. v) Grouped similar variables
3. Adjust the database to the econometric model.	Create a database that facilitates that the data econometric model achieves good results.	i) Standardized indexes and questions with mean zero and standard deviation one to facilitate the interpretation of variables, ii) Defined three versions of variables: the school averaged, the group-mean-centered (centered to each school) and the un-centered original value. These variables allow us to analyze separately school level effects from student level effects. iii) Averaged teacher variables to use the variable as a school indicator. iv) Deleted students that in eighth grade belonged to a school that had information of less than 10 students. The econometric model has a multilevel structure and requires a minimum number of students per school to obtain statistical robust results.

4.4.1. Definition of our sample and data pre-processing

In this section, we describe how we process the data after the input from the data mining algorithm. We merge the datasets, delete variables with too many missing or double

marked values, impute the remaining missing and double marked responses, add control variables and group similar variables.

First, we merged all datasets together. For the data mining algorithm we had three different databases for each initial achievement level (high, mid and low achievement) because the algorithm achieved better results this way. For the econometric model we consolidated all databases in one that contained all students (high, mid and low achievement together), we used prior achievement levels as variables. With Random Forest we analyzed academic achievement between fourth and sixth grade, fourth and eighth grade and sixth and eighth grade. With the econometric model, we only analyzed academic achievement in eighth grade and used as independent variables the achievement levels in fourth and sixth grade. We decided to build only one econometric model because the focus of this document is our proposed approach, rather than the case study on academic achievement. Future work may analyze academic achievement in sixth grade with an econometric model.

We deleted variables with too many missing or double-marked values. Random Forests, works with missing and double marked responses, but the econometric model needs to impute these values. Therefore, we deleted all variables that had more than 15% of the responses missing or double marked. Also, we deleted questions where the respondent was not the main source of information of the question, for example, questions to parents asking how the student feels about certain topic. Afterwards, we imputed the remaining missing and double marked responses with the *missForest* package from R (Stekhoven, 2013).

In this step, we added control variables that were not identified as important by the data mining algorithm but have been widely used in the academic achievement domain (Battistich et al.,1995; Marks & Printy, 2003; Goddard et al., 2007). The school level variables added were type of school (public, private subsidized and private), rurality, school size and region where the school is located. The student level variables added were two binary variables that acknowledge the mother and father belonging to an ethnic group.

Finally, we grouped related questions to create indexes. First we employed R's *hclust()* algorithm (Müllner, 2017) to suggest the grouping of variables, with this information we started the process to build Structural Equation Models in STATA to create the indexes. In the following section, we describe the algorithm employed and detail the complete process.

4.4.2. Variable grouping with hierarchical clustering and Structural Equation Models

Hierarchical clustering can either be agglomerative or divisive. Agglomerative clustering is formed in a bottom-up approach, each object starts in its own cluster and the algorithm iteratively merges clusters into bigger ones, eventually, all objects are grouped in the same cluster (Han et al., 2011). Divisive clustering is formed in a top-down approach, all objects start in one big cluster and the algorithm iteratively starts separating the big cluster into smaller ones, the algorithm ends when there is one object in each cluster or objects in a cluster are sufficiently similar (Han et al., 2011).

We ran *hclust()*(Müllner, 2017), an agglomerative algorithm, in R. In each merging step, the algorithm finds the two closest clusters, according to similarity measures, and combines them. The algorithm offers several similarity or distance measures, the available methods are: i) ward.D, ii) ward.D2, iii) single, iv) complete, v) average, vi) mcquitty, vii) median and viii) centroid. What varies between methods is how to measure the distance between clusters. For example, the single method uses the minimum distance between two objects from two different clusters, whereas the complete method uses the maximum distance between two objects from two different cluster. We employed the complete method. As an output, the algorithm provides a dendogram that shows which clusters were grouped together and the distance between them. We employed the dendogram to identify which variables were closer, with this information we started the process to build Structural Equation Models in STATA.

Structural Equation Modeling is a causal inference method that requires three inputs and generates three outputs (Pearl, 2012). The inputs are: i) a set of qualitative causal

hypothesis, ii) a set of questions about causal relations between variables, and iii) data (Pearl, 2012). The outputs are: i) estimates of parameters for hypothesized effects, ii) logical implications of the model outside parameter estimations, e.g., that two variables are unrelated, and iii) the degree to which the implications of the model are supported by the data (Pearl, 2012).

Covariance is the basic statistic of Structural Equation Modelling, for example, one goal of the model is to explain as much variance as possible (Kline, 2015). Another important aspect is the importance of the theory behind the models and the hypothesis tested. This is because everything, from the initial specification of the model to posterior modifications, must be guided by theoretical and empirical research of the domain being analyzed (Kline, 2015).

Structural Equation Modeling is not a single statistical technique, but a family of related procedures (Kline, 2015). These procedures have its origins in regression analysis of observed variables and in factor analysis of latent variables (Kline, 2015). Figure 4-7, shows an example of a Structural Equation Model provided by Keith (2014).

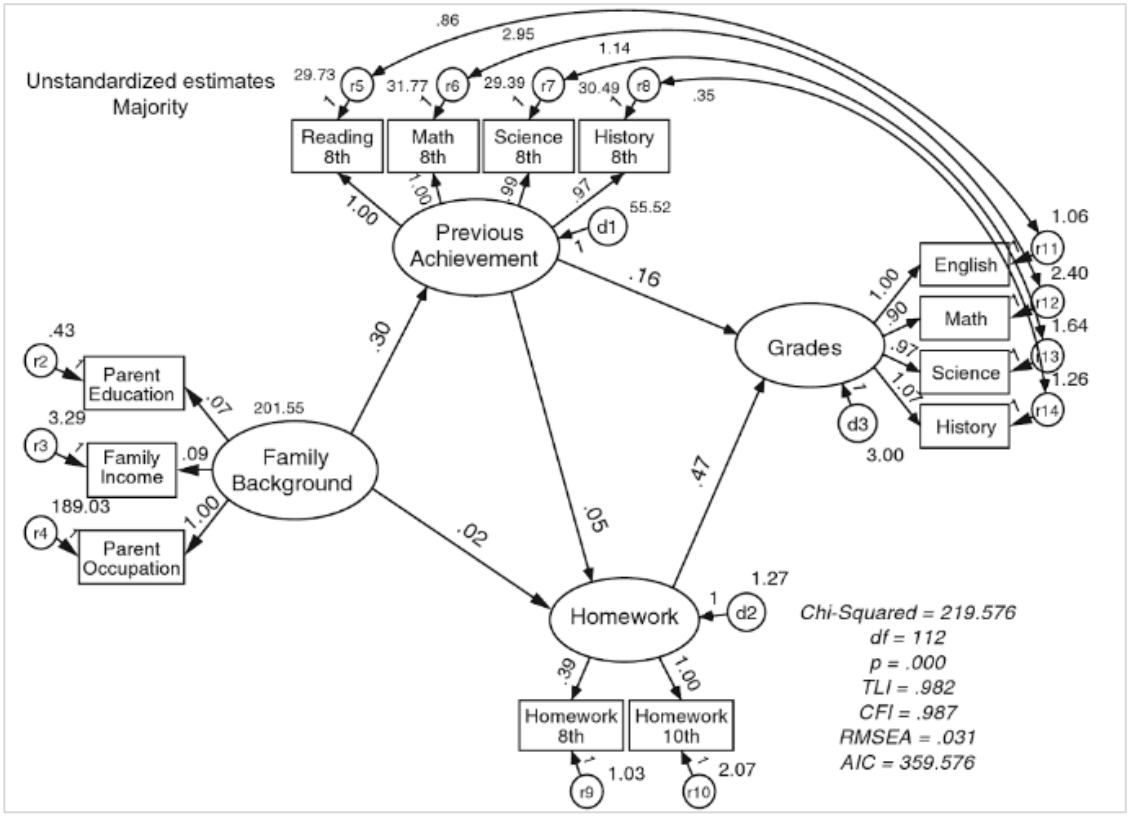


Figure 4-7: Example of a Structural Equation Model (Keith, 2014).

The variables that appear in circles in Figure 4-7, are latent variables. Variables that appear in rectangles are observed variables. The model hypothesizes that the latent variable "Family Background" can be measured with three observed variables: "Parent Education", "Family Income" and "Parent Occupation". The numbers between variables are the parameters that quantify their relation. The statistic indicators in the bottom right side of Figure 4-7 measure the level of fitness of the model to the data.

We used the following thresholds to approve the models: i) Root mean square error of approximation (RMSEA) equal or lower than 0.05, ii) Comparative fit index (CFI) higher than 0.9, iii) Tucker-Lexis index (TLI) higher than 0.9, and iv) the parameters that quantify the relationship between observed and latent variables are equal or higher than 0.40, with the variance of each latent variable standardized to 1.

4.4.3. Adjust the database to the econometric model.

In this section, we adjust the database to the econometric model. All these procedures were executed in STATA. We standardize indexes and questions, define three versions for variables (the school average, the group-mean-centered, i.e. centered to each school and the un-centered original value), average teacher's variables to use the variable as a school indicator and delete samples of students that in eighth grade belonged to a school that had information of less than 10 students.

First, we standardized all indexes and ungrouped questions (numeric and ordinal, binary and nominal variables were not standardized) with mean zero and standard deviation one. This facilitates the interpretation of the outputs of the model.

We calculated the mean per school to create school level variables. In addition, we created group-centered variables (centered to each school mean) to separate the school level effect from the student level effect. Hence, all variables had three versions; the school averaged, the group-mean-centered (centered to each school) and the un-centered original value. These variables allow us to separate school level effects from student level effects.

Then, we averaged teacher's variables and used this information as a school indicator. Most schools have only one teacher's questionnaire, hence, we did not create class level variables because most schools have information of only one class. Instead, we averaged the teacher's values for schools with more than one teacher's questionnaire. For schools with one teacher's questionnaire, we used the information of the teacher as a representation of the school.

Finally, we deleted students that in eighth grade belonged to a school that had information of less than 10 students. The econometric model employed has a multilevel structure, this means that students are nested in schools and the variance of the dependent variable is divided into school level variance and student level variance. For this analysis, each school must count with a minimum number of students to obtain statistical robust results, we set this threshold to 10 students. With this, we end up with a database of 142,457 students.

4.5. Econometric model

An important aspects of our data, regarding econometric and academic achievement research, is the hierarchical structure: Students attend school and share common experiences. Hence, it is probable that many aspects of their academic experience, such as academic achievement, have an important degree of correlation. Because of this structure, the assumptions of independence between observations and uncorrelated errors are not fulfilled; multilevel models acknowledge these issues (Raudenbush & Bryk, 2002). Our dependent variable has a small number of ordered categories as outcomes, consequently, we used the ordinal logistic multilevel model (Snijders & Bosker, 2012). The models were built with STATA.

The process to build the model was the following. First, we built an empty model to check if the proposed structure, ordinal logistic multilevel, is more adequate than the simple correspondent structure, ordinal logistic. The empty model is also useful to check the Intraclass Correlation Coefficient, an indicator of correlation between samples that belong to the same group, in our case schools. Then, we built a model with prior academic achievement variables because of the probable relevance of these variables, just as presented in section 3.2. Academic achievement as a cumulative process.

Posteriorly we built five models, each model corresponds to a group of variables: i) school variables, ii) variables from the teachers questionnaire, iii) variables from the parents questionnaire, iv) ungrouped variables from the students questionnaire, and v) indexes from the students questionnaire. In these five models, the prior academic achievement variables are present because of their probable relevance. In all these models, we left variables that had p-value less than 0.01 and an odds ratio larger than 1.1 or smaller than 0.9. A variable with an odds ratio of 1.1 means that one increase in this variable increases in 10% (1.10 – 1.00 = 10%) the probability of having a higher academic achievement level. Whereas, a variable with an odds ratio of 0.9 implies that one increase in this value decreases in 10% the probability of having a higher academic achievement (0.90 – 1.00 =

-10%). Hence, we left variables that impacted the probability of having a higher academic achievement level in 10% or more, positively or negatively.

After we obtained the five separate models that satisfied our requirements, we built one model that included all the selected variables in the previous stage. Again, we deleted variables that did not satisfied the previous requirements. In this step we eased the requirements in case vast literature argued the importance of a variable, this only occurred with the socioeconomic status variable. We ended up with a model with 35 variables. Finally, we tried to obtain a more parsimonious model and deleted variables that had small impact in the percentage of variance explained.

4.6. Data mining for variable selection

We made modifications to the database employed in the econometric model in comparison to the one employed in the data mining algorithm, we wondered if the new database would nurture a data mining algorithm. Our goal for this data mining algorithm is to identify the most important variables to separate students with high achievement, from students with low achievement in eighth grade. We used the Decision Tree algorithm because it is easy to extract insights from it, we employed R's implementation of Decision Trees (Therneau, et al., 2017). With this process, we propose student's achievement profiles.

To sum up all steps described before, Figure 4-8 summarizes the application of our proposed approach to our case study. The objective was to model the achievement level in the language test ('a' in Figure 4-8). Then, we statistically analyzed variables and preprocessed the data, for example, we created a socioeconomic status variable ('b' in Figure 4-8). We used R's implementation of Random Forest (Liaw & Wiener, 2015) to compare the 1,287 independent variables and selected a subset of them ('c' in Figure 4-8). Then, we imputed missing and double marked responses with R's *missForest* package (Stekhoven, 2013) and grouped variables that were about similar topics into indexes. To create indexes, first we employed *hclust()* from R (Müllner, 2017). We used these results as a starting point to build Structural Equation Models in STATA ('e.1' in Figure 4-8).

Then, we built an ordinal logistic multilevel model in STATA ('e.2' in Figure 4-8). Finally, we used R's implementation of Decision Trees (Therneau, et al., 2017) to identify variables to define a student's achievement profile ('f' in Figure 4-8).

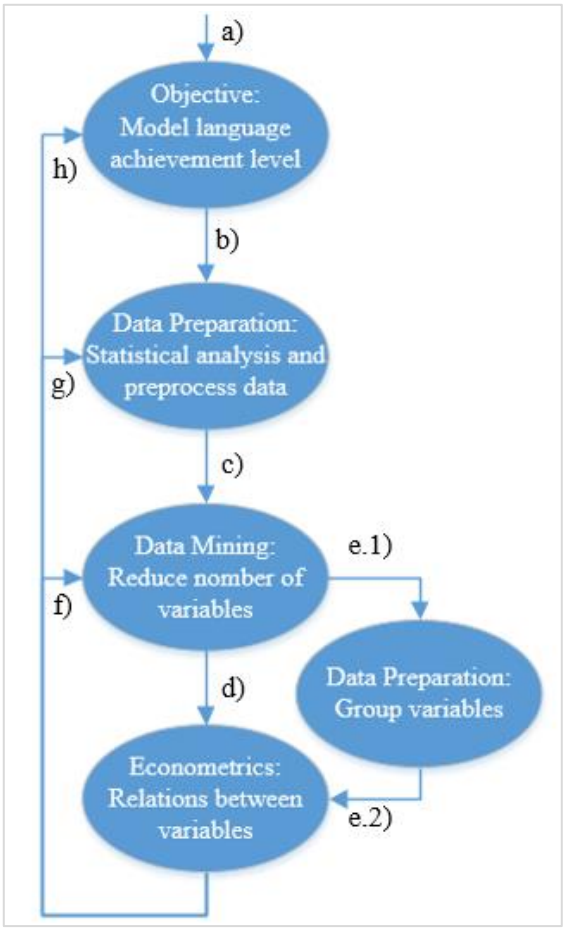


Figure 4-8: Application of our proposed approach.

5. RESULTS

Regarding the first research question, "How can we combine Data Mining and Econometric techniques?", we presented our proposed approach, Econometrics and Data Mining Dialogue, that combines techniques from both disciplines and applied it to a case study. We employed data mining techniques to reduce the number of independent variables from 1,287 to 275, a selection process that was done without assumptions from academic achievement literature. Then, we used a combination of data mining and econometric techniques to group variables and built an ordinal logistic multilevel model. We kept the assumption free approach from data mining and we tested all these variables in the model and deleted those that did not have high significance or had a small impact in the dependent variable (p-value less than 0.01 and an odds ratio larger than 1.1 or smaller than 0.9).

Regarding the second research question, "Can we build an econometric model just from data?" section 5.3. Econometric model, shows the econometric model built just from the data.

We refer to our third research question, "In our specific case study (predicting academic achievement in a longitudinal database using student, classroom and school characteristics), what new findings do we discover?" in section 6.Discussion.

5.1. Random Forest

The goal of our data mining algorithm is to select relevant variables from the 1,287 independent variables. However, as a quality measure, we checked the ability of the algorithm to predict the academic achievement level. We supposed that if the algorithm did not predict adequately the achievement levels, then the variables identified as relevant would not be so useful for the econometric model. Figure 5-1 shows the confusion matrices obtained for the 9 models for each threshold. As in section "4.3.5 Evaluating the classification algorithm through confusion matrices", columns represent actual value and rows represent predicted value.

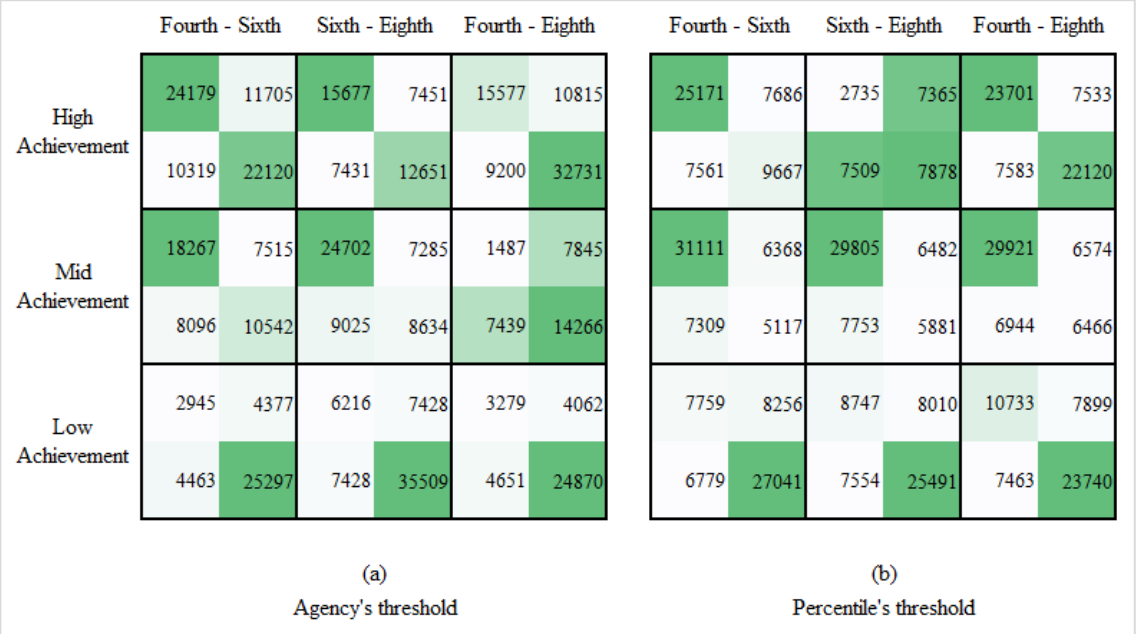


Figure 5-1: Confusion matrices obtained.

More detail on the confusion matrices obtained and a deeper clarification about them can be found in Appendix E. We built a baseline to compare our results based in a method presented in Witten & Frank (2005), details about this baseline are also present in Appendix E. All nine models, for both thresholds, have better results than their respective baseline.

Regarding the selection of variables, few were responsible for most of the decrease in accuracy and gini, the indicators that measure the ability of the variable to separate between classes, (to make class "a" go to one branch and class "b" go to the other branch). Figure 5-1 shows the mean decrease accuracy for one model. The vertical axis is the percentage of decrease that each variable provokes in comparison with the total decrease that all variables achieve together (we divided the decrease each variable provoked by the sum of the decrease of all variables). The horizontal axis has the 1,287 variables in decreasing order of mean decrease accuracy. The figure shows that the most important variables are more or less the 30 to 50 first ones, then the indicator declines dramatically.

This tendency is present in all variable's importance rankings; the threshold varies between the first 30 and 50 variables. Hence, we looked at the top 50 variables of each ranking. Since we are analyzing 18 models (nine for each threshold), and each model counts with two rankings (the gini and accuracy indicator), we analyzed 36 rankings of variable's importance. The graphs of the 36 rankings are in Appendix F. We combine the top 50 variables of all these rankings and identify 275 relevant variables.

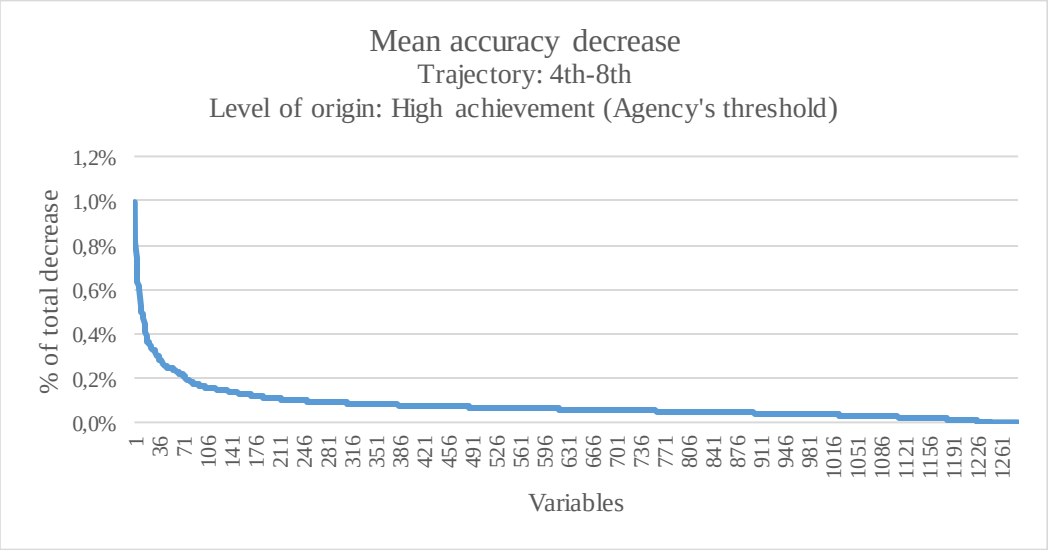


Figure 5-2: Example of mean decrease accuracy per variable.

To show the most important variables identified by the algorithm we associated them to concepts. Table 10 shows the five most important concepts (not in order of importance) for each model and variables that were important only to the specified model. Socioeconomic status, self-perception and enjoyment of reading appear in all nine models.

Each of the nine models counts with four variable's importance rankings (two thresholds and two indicators: i) percentile-gini, ii) percentile-accuracy, iii) Agency's standards-gini and iv) Agency's standards-accuracy). The variables presented in Table 5-1 as "the most important concepts" were selected by looking at the top 10 variables of the four correspondent rankings and requiring them to appear in the top 10 list of at least three of

the four rankings. For a given model, the four rankings were very similar. A tendency that the table shows is that school climate concepts (bullying, teacher's ability to resolute conflicts and weapons in school) are more important to predict academic achievement of students that start with a low or mid achievement level.

Another interesting pattern is that three concepts are important only for low initial achievement students. The concepts are: i) Students report or feel discriminated. ii) Teachers believe that all students can learn despite their differences. iii) School enrollment requirements, these requirements are marriage status, baptism or catholic marriage, income certificate, to attend a school playdate and to give an enrollment test.

Table 5-1: Important variables for each model.

Achievement	Fourth - Sixth	Sixth - Eighth	Fourth - Eighth
High initial achievement	1) Socioeconomic status 2) Self-perception 3) Enjoyment of reading 4) Parents Expectations 5) Sports Self-perception - Grade repetition	1) Socioeconomic status 2) Self-perception 3) Enjoyment of reading 4) Parents Expectations 5) Organize sport activities	1) Socioeconomic status 2) Self-perception 3) Enjoyment of reading 4) Parents Expectations 5) Organize sport activities
Mid initial achievement	1) Socioeconomic status 2) Self-perception 3) Enjoyment of reading 4) Parents Expectations 5) Bullying - Grade repetition	1) Socioeconomic status 2) Self-perception 3) Enjoyment of reading 4) Weapons in school 5) Teachers resolute conflict	1) Socioeconomic status 2) Self-perception 3) Enjoyment of reading 4) Weapons in school 5) Teachers resolute conflict
Low initial achievement	1) Socioeconomic status 2) Self-perception 3) Enjoyment of reading 4) Extracurricular participation 5) Bullying - Students are/feel discriminated - All students can learn focus - School enrollment requirements - Grade repetition	1)Socioeconomic status 2) Self-perception 3) Enjoyment of reading 4) Weapons in school 5) Teachers resolute conflict - Students are/feel discriminated - All students can learn focus - School enrollment requirements	1) Socioeconomic status 2) Self-perception 3) Enjoyment of reading 4) Weapons in school 5) Teachers resolute conflict - Students are/feel discriminated - All students can learn focus - School enrollment requirements

5.2. Data preparation for the econometric model

The data mining algorithm identified 275 independent variables. Just as described in section 4.5. Data Preparation, we deleted all variables with more than 15% of the responses missing or double marked, this deletes 20 variables. We also deleted variables were the respondent was not the main source of information of the question, this deleted 24 variables. Then, we imputed the remaining missing and double marked values with *missForest* package from R (Stekhoven, 2013). In addition, we added control variables that were not identified by the data mining algorithm as important, but have been extensively used in the academic achievement domain, this adds 11 variables. Also, added variables employed as keys, the key to identify each student through time and the keys to identify school belonging in fourth, sixth and eighth grade; this adds 4 variables.

With the process described above, we end up with 246 independent variables. Some of the 246 variables were about similar topics so we employed *hclust()* from R (Müllner, 2017) to start grouping them. Figure 5-2 shows the dendogram obtained, it basically shows which variables are closer to each other. The highest division separates the information obtained by the student's questionnaire with the rest of the information, mainly questions from teacher and parent's questionnaires. This shows that although these three actors are asked about common features, their answers do not relate to each other's. For example, teachers, parents and students are asked about school climate but their answers about these topic are not close from each other's in the dendogram.

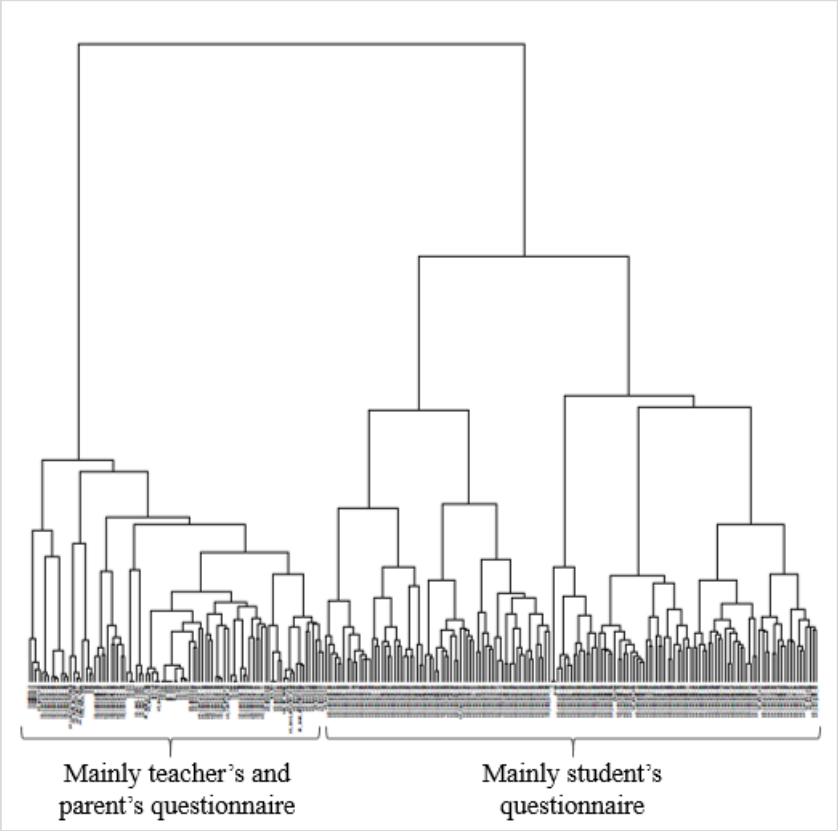


Figure 5-3: Dendogram obtained with *hclust()*.

Appendix G details the results obtained and offer close ups of the dendogram, these allows to read the keys of each variable. We used these latter results as a starting point to create indexes through Structural Equation Models with STATA. After some iterations, we obtained the indexes. Appendix H shows the results of the Structural Equations Models.

We built a structural equation model for each actor and grade. Table 5-2 summarizes the indexes created for each actor and grade. It is important to highlight that the indexes represent perceptions of the actors towards a topic. For example, the index "basic teaching practices" for fourth grade students, represent the perception of students about their teacher's practices. The rest of actors and grades did not have variables that could be grouped, they were employed ungrouped in the econometric model.

Table 5-2: Indexes created per actor and grade.

Actor	Grade	Indexes created
Student	Fourth	i) Extracurricular participation, ii) Effort and hard work, iii) Basic teaching practices, iv) Advance teaching practices, v) General self-perception, vi) Math self-perception, vii) school climate.
Student	Sixth	i) Extracurricular participation ii) General self-perception, iii) Math self-perception, iv) Language self-perception, v) Sport self-perception, vi) Fast food consumption, vii) school climate, viii) Tolerance to misconduct, ix) Bullying, x) Bullying victim.
Student	Eighth	i) Extracurricular participation, ii) Extracurricular organization, iii) Math self-perception, iv) Language self-perception, v) Sport self-perception, vi) Enjoyment of reading, vii) Fast food consumption, viii) School climate – between students, ix) School climate – between teachers and students, x) School climate – weapons, xi) School climate – drugs, xii) Discrimination, xiii) Evaluation of the school, xiv) teachers resolute conflicts.
Teacher	Sixth	i) School climate.
Teacher	Eighth	i) School climate, ii) basic teaching practices, iii) advance teaching practices

5.3. Ordinal logistic multilevel model

Table 5-3 shows the ordinal logistic multilevel models. Model 0 is the model without variables, the Intraclass Correlation Coefficient is 21.7%, which falls between often ranges: 10 and 25% (Hedges & Hedberg, 2007). This means that 21.7% of variation in academic achievement is explained by variation between schools. The test that compared the ordinal logistic multilevel models against an ordinal logistic model showed that grouping by schools is significant and the multilevel structure is correct. Ordinal logistic multilevel models set the variance of the student level error to $\pi^2/3$. The variance of the school level error is 0.91.

Model 1 has prior academic achievement variables. The Intraclass Correlation Coefficient decreased to 14,5%. The percentage of variance explained (PVE) by the model is 41.3%.

Model 2 is the complete model. It contains all the variables that satisfied the requirements detailed in the methodology; odds ratio are equal or larger than 1.1 or equal or smaller than 0.9 and have a p-value smaller than 0.01. The only variable that does not satisfy these requirements is student level socioeconomic status centered to the school mean (it does

not satisfy the odds ratio requirement because it has an odds ratio of 1.09); we kept it because of the vast research about socioeconomic status influence on academic achievement (Sirin, 2005). Variables that have "(g)" are group-mean-centered to schools, variables that have "(a)" are the average value for each school. Variables that have "(m)" are mean-centered variables with mean zero and a standard deviation of one.

Model 3 is a simpler model. We sought for parsimony and deleted variables that had the smallest impact on the percent of variance explained. After we deleted the ten variables with least effect, the eleventh variable reduced the variance explained dramatically so we stopped when the model had 25 variables. Model 4, our selected model, turns two fixed effects coefficients to random effects coefficients, this allows coefficients to vary across schools. The percentage of variance explained is 53.6%.

For Model 4 we changed the dependent and independent academic achievement variables to the percentiles division version, all variables remained significant with a p-value equal or smaller than 0.001. The percentage of variance explained in this case is 53.1%

Table 5-3: Ordinal logistic multilevel models.

	Model 0	Model 1	Model 2	Model 3	Model 4
Number of variables	0	4	35	25	25
Prior academic achievement					
Fourth grade					
High achievement		5.30 (85.5)	4.12 (69.9)	4.10 (69.8)	4.10 (69.8)
Mid achievement		2.39 (46.9)	2.12 (39.2)	2.11 (39.1)	2.11 (39.1)
Sixth grade					
High achievement		17.23 (148.6)	10.91 (120.6)	10.93 (120.7)	10.93 (120.7)
Mid achievement		3.59 (82.6)	3.01 (69.2)	3.01 (69.3)	3.01 (69.3)
Parent's questionnaire					
Fourth grade					
Grade repetition (g)			0.79 (-9.4)		
Grade repetition (a)			0.75 (-3.5)		
Grade repetition (m)				0.79 (-9.8)	0.79 (-9.7)
Eighth grade					
SES (c)			1.09 (8.8)	1.09 (8.8)	1.09 (8.8)
SES (a)			1.17 (7.2)	1.11 (5.4)	1.10 (5.2)
Students' grades are reported to parents (a)			1.12 (3.5)		
Student's questionnaire: Questions					
Sixth grade					
"We obey our teachers and work in order" (g)			0.89 (-13.0)	0.89 (-13.0)	0.89 (-13.0)
"We obey our teachers and work in order" (a)			0.79 (-6.6)	0.81 (-6.4)	0.80 (-6.5)
Eighth grade					
"I understand what they teach me in class" (g)			1.18 (21.4)		
"I understand what they teach me in class" (a)			1.31 (4.2)		
"I understand what they teach me in class" (m)				1.18 (21.8)	1.18 (21.8)
"In language tests I do better than most of my classmates" (g)			1.27 (27.3)	1.27 (27.2)	1.27 (27.2)
"In language tests I do better than most of my classmates" (a)			1.48 (6.5)	1.57 (7.7)	1.55 (7.5)
"Being a good reader is important to me" (a)			1.32 (3.5)		
"Teachers have had to shout to keep order" (a)			0.80 (-4.4)		
"Teachers promote participation in the classroom." (a)			1.65 (8.2)	1.69 (8.7)	1.71 (8.8)
"Places for physical activity outside school." (a)			0.74 (5.1)	0.74 (-5.5)	0.73 (-5.6)
Student's questionnaire: Indexes					
Fourth grade					

General self-perception	1.13 (18.0)	1.13 (17.8)	1.13 (17.8)
School climate (a)	0.86 (-5.0)		
Eighth grade			
Math self-perception	1.18 (25.2)	1.18 (25.2)	1.18 (25.2)
Sport self-perception	0.88 (-18.2)	0.88 (-18.0)	0.88 (-18.0)
Enojoyment of reading (g)	1.39 (47.0)	1.39 (47.0)	1.39 (47.0)
Enojoyment of reading (a)	1.36 (5.7)	1.55 (9.7)	1.56 (9.5)
Fast food consumption (g)	0.90 (-16.3)		
Fast food consumption (a)	0.84 (-5.3)		
Fast food consumption (m)		0.89 (-17.2)	0.89 (-17.2)
Frequency of threats or assaults with weapons (g)	0.83 (-21.3)	0.83 (-21.2)	0.83 (-21.2)
Frequency of threats or assaults with weapons (a)	0.70 (-8.1)	0.62 (-12.8)	0.61 (-12.9)
Organization of extracurricular activities (g)	0.84 (-23.8)		
Organization of extracurricular activities (a)	0.80 (-5.3)		
Organization of extracurricular activities (m)		0.84 (-24.2)	0.84 (-24.2)
Frequency with which you have felt discriminated (a)	0.80 (-4.0)		
Students evaluation of the school (a)	1.72 (11.8)	1.94 (15.7)	1.95 (15.3)
School variables			
Eighth grade			
Rurality	1.16 (3.4)		
Random effects			
Variance of enjoyment of reading (a)			0.16
Variance of students evaluation of the school (a)			0.23

Intercept Variance Component	0.91	0.56	0.37	0.38	0.33
ICC	21.7%	14.5%	10.1%	10.4%	9.5%
PVE	0%	41.3%	53.6%	53.4%	53.6%

Note: Intercept cut points are excluded from the output. The values between parentheses are the z-value of each variable. All variables have a p-value equal or smaller than 0.001.

5.4. Decision Tree

We used the processed database employed in the econometric model in a data mining algorithm. Our goal for the data mining algorithm is to propose different student's achievement profiles. The idea is to identify the most important variables to separate students with high and low achievement level in eighth grade.

The decision tree uses prior academic achievement as the first variables to separate between students. Table 5-4 summarizes the three student's achievement profiles.

Table 5-4: Student's achievement profile based on prior academic achievement.

	Achievement level in 8th Number of students				Achievement level in 8th Percentage of students			
Student's achievement profiles	Low	Mid	High	Total	Low	Mid	High	Total
Probably high achievement (High achievement level in 4th and 6th)	2,646	11,832	20,358	34,836	8%	34%	58%	100%
Probably mid achievement (The rest)	34,742	35,292	10,008	80,042	43%	44%	13%	100%
Probably low achievement (Low achievement level in 4th and 6th)	23,832	3,550	197	27,579	86%	13%	1%	100%

We expected that prior academic achievement would probably define student's academic achievement profiles because of the literature presented in section 3.2. Academic achievement as a cumulative process and the high z-values of prior academic achievement variables in the ordinal logistic multilevel model. We built another decision tree excluding prior academic achievement variables. In this case, the variables selected come from the student's questionnaires: i) general self-perception in fourth grade, ii) enjoyment of reading in eighth grade and iii) school's evaluation in eighth grade. Table 5-5 presents these new profiles.

Table 5-5: Student's achievement profile excluding prior academic achievement variables.

Student's achievement profiles	Achievement level in 8th Number of students				Achievement level in 8th Percentage of students			
	Low	Mid	High	Total	Low	Mid	High	Total
Probably high achievement (High: Enjoyment of reading, school evaluation and self-perception)	1,047	3,290	5,502	9,839	11%	33%	56%	100%
Probably mid achievement (The rest)	34,636	36,675	22,542	93,853	37%	39%	24%	100%
Probably low achievement (Low: Enjoyment of reading, school evaluation and self-perception)	25,537	10,709	2,519	38,765	66%	28%	6%	100%

The Decision Trees employed to define both profiles and more detail about this process can be found in Appendix I.

6. DISCUSSION

In this section, we discuss the results obtained. We refer to the algorithms and the model's performance. Also, we state our major findings and answer our third research question, "In our specific case study (predicting academic achievement in a longitudinal database using student, classroom and school characteristics), what new findings do we discover?".

6.1. Algorithm and model performance

The Random Forest model outperformed the baseline we built, but did not predict correctly the achievement of all students. For example, the model that predicted the trajectory between fourth and sixth grade with low achievement students in fourth grade had a precision and recall of 40% for the students that were able to increase and a precision and recall of 85% for students that remained in the lower achievement level in sixth grade Thus, the model was not good enough to predict students that increased their achievement level between fourth and sixth grade.

The multilevel model is able to explain 53.6% of the total variance, there is still a percentage that the model cannot explain. We did not find an ordinal logistic multilevel model that predicts academic achievement to compare the 53.6% against. Just to compare the 53.6% against something, we refer to some linear multilevel results. Linear multilevel models work with continuous outcomes. This is different to ordinal logistic multilevel models, because when the outcome is transformed from a score to an ordinal value, variance is lost. Pitsia et al. (2017) employ linear multilevel models to predict PISA 2012 Math scores in Greece, the model is able to explain 39.4% of the total variance. Mancebón et al. (2012) use linear multilevel models to predict PISA 2006 Science scores in Spain, the model is able to explain 43.2% of the total variance. Finally, Zambrano (2016) use linear multilevel models to predict SERCE 2006 Math scores in 15 Latin American countries, the model is able to explain 18.5% of the total variance.

Although our models have space to improve, we are satisfied with their results. Random Forest and the ordinal logistic multilevel model identified common variables as important, for example, enjoyment of reading, self-perception, socioeconomic status, among others.

In addition, the data mining algorithm (Random Forest) and the econometric model were not modified dramatically by the change in thresholds (the thresholds defined by the Agency and the division based on percentiles). The variables defined as important by the data mining algorithm and the quality of the academic achievement prediction did not vary much with the change of thresholds. In the econometric model, specifically Model 4, all 25 independent variables remained significant when changed to the percentile threshold from the Agency's threshold (p-value equal or smaller than 0.001). The percentage of variance explained only declined from 53.6% to 53.1%.

These two aspects: i) That variables identified as relevant are common to the data mining algorithms and the econometric model and ii) that the models are not affected by the change in the thresholds, make us conclude that our findings are robust.

6.2. Major findings

Some of our major finding are new findings, which helps us answer our third research question, "In our specific study (predicting academic achievement in a longitudinal database using student, classroom and school characteristics), what new findings do we discover?

6.2.1 Random Forest findings

Regarding the first data mining algorithm, Random Forest, we highlight the following results:

i) School climate concepts (bullying, teacher's ability to resolute conflicts and weapons in school) are more important to predict achievement of students that start with a low or mid achievement level. It seems to be a differential effect of school

climate on academic achievement according to the initial level of achievement of students.

There is vast literature that shows the impact of school climate in academic achievement, but we did not found research that proposes that school climate variables are more important for low or mid achieving students that for high achieving students. The only research found in this line was from Brookover et al. (1978); it showed that school climate explained more variance in academic achievement in majority black schools than majority white schools. But the study does not refer to relationship between achievement and school climate.

ii) Three concepts important only for low achievement students. The concepts are: i) Students report or feel discriminated. ii) Teachers believe that all students can learn despite their differences. iii) School enrollment requirements, which are marriage status, baptism or catholic marriage, income certificate, to attend a school playdate and an enrollment test.

We did not find research that shows that the previous concepts are relevant for low achieving students. Regarding school enrollment requirements Contreras et al. (2010) analyze the effects of schools selecting students. The authors find that a student that attends a school that uses selection criteria obtains between 6% and 14% higher results (in comparison to students that attend schools that do not select) in the SIMCE mathematics test.

6.2.2 Ordinal logistic multilevel findings

Regarding the econometric model, the ordinal logistic multilevel model, we highlight the following results:

i) After previous academic achievement variables, student's evaluation of the school has the biggest effect on academic achievement. In model 4, students with a higher level in this variable have 95% more probabilities of having a higher academic achievement level in eighth grade.

This is consistent with literature (Wang & Holcombe, 2010), but we did not find this variable present in many studies. Student's perceptions about their school are not included in econometric studies so often.

ii) The negative effect of variables related to a learning environment and school climate.

Normally, these variables have a positive effect on academic achievement (Hattie, 2009; Bryk, et al., 2010), opposite to what we found. For example, sixth grade students that belong to a school with a higher level of "We obey our teachers and work in order" have 21% less probabilities to have a higher achievement level in eighth grade than students who belong to a school with a lower level in this variable. We highlight that this is the perception of students.

iii) Organization of extracurricular activities and sports self-perception have a negative effect on academic achievement.

According to literature, organization of extracurricular activities has null or positive effect on academic achievement (Hattie, 2009). Regarding sport self-perception, it normally has null or positive effect on academic achievement (Hattie, 2009; Trudeau & Shephard, 2008). Few studies found a slight negative relation between physical activity and academic achievement (Tremblay et al., 2000). For example, when school banding exists and schools group high and low achieving students separately. The higher band experiences a positive relation between physical activity and grades, whereas the low band experiences a negative relation between physical activity and grades (Lindner, 2002). However, this refers to physical activity, not sports self-perception.

iv) Most of the variables belong to the student's questionnaire.

This may suggest that the perceptions that have more impact on academic achievement are the student's perceptions, rather than their parents or teachers.

6.2.1 Decision Tree findings

Regarding the second data mining algorithm, Decision Tree, we highlight the following:

i) The decision tree selected prior academic achievement variables to build student's achievement profiles.

It separates the sample into probably high achievement (34,836 students), probably mid achievement (54,213 students) and probably low achievement (27,579). We expected the algorithm to select these variables because of the literature review and the high z-value of these variables in the econometric model.

ii) When we excluded prior academic achievement variables, the decision tree selected variables we constructed from the student's questionnaire. The variables are i) enjoyment of reading, ii) evaluation of school and iii) self-perception.

The profiles separates the sample into probably high achievement (10,033 students), probably mid achievement (95,475 students) and probably low achievement (27,949). The profiles built with these variables are more unbalanced than the ones created with prior academic achievement variables. We consider that these profiles are weaker, but we highlight these results because is a less intuitive way of profiling students.

7. CONCLUSIONS

This study's contribution is our proposed approach, Econometrics and Data Mining Dialogue, which details how data mining and econometric techniques can collaborate and support each other. Data mining applications in academic achievement has focused on accurately predicting variables, while econometric work has focused on understanding education by building over past knowledge and theory. On one hand, data mining can nurture econometrics by providing a wider and less biased way of analyzing the education phenomena; it allows to look over present knowledge. On the other hand, econometrics nurtures data mining by providing a deeper understanding of the education phenomena

and by asking vital questions that the data mining community has not questioned yet; for example, providing new tools to better understand interactions between variables.

Here, we made the econometric and data mining domains dialogue by complementing each other's methods. Our first research question was "How can we combine data mining and econometric techniques?, we detailed our proposed approach, and applied it to a specific case study: academic achievement. The methods and results sections provide a detailed explanation of how we combined Random Forests, ordinal logistic multilevel models and Decision Trees, to study academic achievement and the results obtained by this methodology. Our second research question was "Can we build an econometric model just from data?". Throughout this paper, we showed it was possible. In fact, our econometric model was able to explain 53.6% of the total variance; and it is robust to the threshold change from the Agency's division to the percentile's division. Finally, our third research question was " In our specific case study (predicting academic achievement in a longitudinal database using student, classroom and school characteristics), what new findings do we discover?". We identified several findings in section 6. Discussion. Some of them are i) the special importance of school climate concepts (bullying, teacher's ability to resolute conflicts and weapons in school) for low and mid achieving students, ii) the negative relation between sport self-perception, extracurricular organization and school climate with academic achievement; contradicting most literature, and the importance of students perceptions, specially the evaluation they make to their schools.

We identify two kinds of limitations to our research. Data limitations and modeling limitations.

Regarding our data; first, we acknowledge the limitations of employing a standardized test as an indicator of academic achievement and the simplification of continuous scores to three levels of achievement. Academic achievement is a complex concept and it is hard to reduce its meaning to a single indicator. Second, the use of self-reported questionnaires. Although these questionnaires have been studied and improved throughout the years by national agencies and the academia, all questionnaires have biased responses (Paulhus,

1991; Van de Mortel, 2008). Third, the sample. From our data, on average 255,000 students are enrolled in each grade nationally. However, on average only 217,000 give the test. Hence, the Quality of Education Agency systematically does not have information of 15% of Chilean students; neither their scores on tests nor their perceptions measured by the questionnaires. In fact our sample was even smaller, 142,457 students.

Regarding our models; the Random Forest, the ordinal logistic multilevel model and Decision Trees, cannot predict or explain perfectly our dependent variable. In fact, some authors propose improvements to the variable selection process of Random Forest (Strobl et al., 2007; Strobl et al., 2008; Touw et al., 2012). Another limitation is that we did not use all the information from the Random Forest model to nurture the multilevel model. Our Random Forest model is an ensemble of 10,000 Decision Trees. These Decision Trees use variable interactions for the prediction and estimation of variable importance, the importance value provides the combined importance of variables but does not specify which variables interact with each other (Touw et al., 2012). That we know of, it does not exist a tool to extract easily information about variable interactions from Random Forest (Liaw & Wiener, 2015).

Future studies can focus in two lines: i) Provide more examples of the usefulness of combining data mining and econometric techniques and ii) Develop new tools that facilitate cooperation between data mining and econometric techniques.

First of all, in the introduction of this work we highlighted how several authors argue that collaboration between computer scientists and econometricians is expected. But, to our best knowledge, there is no work that attempts to show how these two disciplines can work together. We proposed an approach and applied it to a case study; we hope this example encourages future work to advance in this line. Future studies can provide new examples of collaboration between computer scientists and econometricians.

Specifically to the educational domain, both disciplines count with strong research communities. The Educational Data Mining community and the Learning Analytics community are related to the data mining discipline. Whereas the communities related to

econometrics in education are plentiful, e.g., academic achievement, school improvement, economics of education, among others. The communities related to data mining have dialogued with each other, and communities related to econometrics have dialogued between each other too. However, we did not find evidence that communities from different disciplines (data mining and econometrics) have dialogued. Future work should bring actors from different areas together.

Secondly, one limitation we encountered throughout this study was to extract all the information embedded in the data mining algorithms. Some software like WEKA and Orange offer easy ways to implement some data mining techniques, but, these software count with limited tools in comparison to what is implemented in R or python. However, other tasks like extracting information about variable interactions from Random Forest or Neural Networks are not implemented in R or python. Future work should focus in developing new tools that enhance and facilitate cooperation between data mining and econometric techniques.

REFERENCES

Adelson, J. L., Dickinson, E. R., & Cunningham, B. C. (2016). A Multigrade, Multiyear Statewide Examination of Reading Achievement: Examining Variability Between Districts, Schools, and Students. *Educational Researcher, 45*(4), 258-262.

Agencia de Calidad de la Educación (2014). Informe Técnico SIMCE 2014.

Agencia de Calidad de la Educación (2016a). Manual de aplicación de pruebas censales y experimentales SIMCE 2016.

Agencia de Calidad de la Educación (2016b). Resultados educativos SIMCE 2016.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Chen, J. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning* (pp. 173-182).

Angeli, C., Howard, S., Ma, J., Yang, J., & Kirschner, P. A. (2017). Data mining in educational technology classroom research: Can it make a contribution?. *Computers & Education*.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arnold D. H., Doctoroff G. L. (2003). The early education of socioeconomically disadvantaged children. *Annual Review of Psychology,* 54, 517–545.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics* (pp. 61-75). Springer New York.

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining, 1*(1), 3-17.

Battistich, V., Solomon, D., Kim, D. I., Watson, M., & Schaps, E. (1995). Schools as communities, poverty levels of student populations, and students' attitudes, motives, and performance: A multilevel analysis. *American educational research journal*, 32(3), 627-658.

Bekele, R., & Menzel, W. (2005). A bayesian approach to predict performance of a student (bapps): A case with ethiopian students. *Algorithms*, *22*(23), 24.

Bowling, A. (2014). *Research methods in health: investigating health and health services*. McGraw-Hill Education (UK).

Breiman, L. (1996) Bagging Predictors, *Machine Learning*, 26, No. 2, 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Breiman, L. & Cutler. A. (2004). Random Forests source code (Version 5.1) [Source code]. https://www.stat.berkeley.edu/users/breiman/RandomForests.

Bresfelean, V. P., Bresfelean, M., Ghisoiu, N., & Comes, C. A. (2008, June). Determining students' academic failure profile founded on data mining methods. In *Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on* (pp. 317-322). IEEE.

Brookover, W. B., Schweitzer, J. H., Schneider, J. M., Beady, C. H., Flood, P. K., & Wisenbaker, J. M. (1978). Elementary school social climate and school achievement. *American educational research journal*, *15*(2), 301-318.

Brooks, C. (2014). *Introductory econometrics for finance*. Cambridge university press.

Bryk, A. S., Sebring, P. B., Allensworth, E., Easton, J. Q., & Luppescu, S. (2010). *Organizing schools for improvement: Lessons from Chicago*. University of Chicago Press.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, *28*(2), 15-21.

Caro, D. H. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education*, *32*(3), 558.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

Chen, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A. V., & Rong, X. (2015). Data mining for the internet of things: literature review and challenges. *International Journal of Distributed Sensor Networks*.

Coleman, J. S. (1966). Equality of educational opportunity.

Contreras, D., Sepúlveda, P., & Bustos, S. (2010). When schools are the ones that choose: The effects of screening in Chile. *Social Science Quarterly*, 91(5), 1349-1368.

Cortes, C.; Vapnik, V. (1995). "Support-vector networks". *Machine Learning*. 20 (3): 273–297.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21-27.

Dietterich, T. G. (1997). Machine-learning research. *AI magazine*, 18(4), 97.

Drucker, H. (1997, July). Improving regressors using boosting techniques. In *ICML* (Vol. 97, pp. 107-115).

Duncan G. J., Dowsett C. J., Claessens A., Magnuson K., Huston A. C., Klebanov P., . . . Japel C. (2007). School readiness and later achievement. *Developmental Psychology*, 43(6), 1428–1446

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A Systematic Review on Educational Data Mining. *IEEE Access*.

Education and Science, Department of, & Plowden, B. B. H. P. (1967). *Children and Their Primary Schools: A Report. Research and Surveys*. HM Stationery Office.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14*(1), 1-24.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15(1), 3133-3181.

Fix, E., & Hodges Jr, J. L. (1951). *Discriminatory analysis-nonparametric discrimination: consistency properties*. California Univ Berkeley.

Gamal, A., Sayed, G. I., Darwish, A., & Hassanien, A. E. (2017). A New Proposed Model for Plant Diseases Monitoring Based on Data Mining Techniques. In *Plant Bioinformatics* (pp. 179-195). Springer, Cham.

Gibert, K., Sànchez-Marrè, M., & Codina, V. (2010). Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation. *International Congress on Environmental Modelling and Software*, 367.

Goddard, Y. L., Goddard, R. D., & Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teachers college record*, 109(4), 877-896.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hattie, J. A. (2009). Visible learning: A synthesis of 800+ meta-analyses on achievement. Abingdon: Routledge.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), 832-844.

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. (2014). *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton University Press.

Jimerson, S., Egeland, B., & Teo, A. (1999). A longitudinal study of achievement trajectories: Factors associated with change. *Journal of educational psychology*, *91*(1), 116.

Kane, M. J., Price, N., Scotch, M., & Rabinowitz, P. (2014). Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC bioinformatics*, *15*(1), 276.

Keith, T. Z. (2014). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*. Routledge.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Kolenikov, S. (2016, August). polychoric, by any other'namelist'. In 2016 Stata Conference (No. 15). Stata Users Group.

Lewis, D. D., & Ringuette, M. (1994, April). A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval*, (Vol. 33, pp. 81-93).

Liaw, A., & Wiener, M. (2015). Breiman and Cutler's random forests for classification and regression, R package version 4.6-12. *R Foundation for Statistical Computing*, Vienna.

Lindner, K. J. (2002). The physical activity participation–academic performance relationship revisited: Perceived and actual performance and the effect of banding (academic tracking). *Pediatric Exercise Science*, 14(2), 155-169.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Luan, J. (2002). Data mining and its applications in higher education. *New directions for institutional research*, *2002*(113), 17-36.

Ma, Y., Liu, B., Wong, C. K., Yu, P. S., & Lee, S. M. (2000, August). Targeting the right students using data mining. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 457-464). ACM.

Mancebón, M. J., Calero, J., Choi, Á., & Ximénez-de-Embún, D. P. (2012). The efficiency of public and publicly subsidized high schools in Spain: Evidence from PISA-2006. *Journal of the Operational Research Society*, 63(11), 1516-1533.

Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. *Educational administration quarterly*, 39(3), 370-397.

Martínez Abad, F., & Chaparro Caso López, A. A. (2017). Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement*, *28*(1), 39-55.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-124.

Massaro, F., Allegretti, M., & Ferrari, A. (2017, September). Data mining in modern radio astronomy. In *Antennas and Propagation in Wireless Communications (APWC), 2017 IEEE-APS Topical Conference on* (pp. 15-16). IEEE.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Müllner, D. (2017). Fast Hierarchical Clustering Routines for R and Python. R package version. 1.1.24

Mustafa, M. K., Allen, T., & Appiah, K. (2017). A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition. *Neural Computing and Applications*, 1-9.

Nghe, N. T., Janecek, P., & Haddawy, P. (2007, October). A comparative analysis of techniques for predicting academic performance. In *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual* (pp. T2G-7). IEEE.

Niekler, A., Wiedemann, G., & Heyer, G. (2017). Leipzig corpus miner-a text mining infrastructure for qualitative data analysis. *arXiv preprint arXiv:1707.03253*.

Oladokun, V. O., Adebanjo, A. T., & Charles-Owaba, O. E. (2008). Predicting students' academic performance using artificial neural network: A case study of an engineering course. *The Pacific Journal of Science and Technology*, *9*(1), 72-79.

Panda, M., Hassanien, A. E., & Abraham, A. (2017). Hybrid Data mining approach for image segmentation based Classification. In *Biometrics: Concepts, Methodologies, Tools, and Applications*(pp. 1543-1561). IGI Global.

Paulhus, D. L. (1991). Measurement and control of response bias.

Pearl, J. (2012). *The causal foundations of structural equation modeling* (Vol. 370). CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE.

Pitsia, V., Biggart, A., & Karakolidis, A. (2017). The role of students' self-beliefs, motivation and attitudes in predicting mathematics achievement: A multilevel analysis of the Programme for International Student Assessment data. *Learning and Individual Differences*, 55, 163-173.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.".

Raghupathi, W. (2016). Data mining in healthcare. *Healthcare Informatics: Improving Efficiency through Technology, Analytics, and Management*, 353-372.

Ramaswami, M., & Bhaskaran, R. (2010). A CHAID based performance prediction model in educational data mining. *IJCSI International Journal of Computer Science Issues*, Vol. 7,

January 2010, Issue 1, No 1, 10-18.

*R*audenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.

Scott, D., & Usher, R. (2010). *Researching education: Data, methods and theory in educational enquiry*. Bloomsbury Publishing.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons.

Silva, C., & Fonseca, J. (2017). Educational Data Mining: A Literature Review. In *Europe and MENA Cooperation Advances in Information and Communication Technologies* (pp. 87-94). Springer International Publishing.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, 75(3), 417-453.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85-106.

Stekhoven, D. (2013) MissForest: nonparametric missing value imputation using random forest. R package version. 1.4

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 465-474.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society*.

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2017). Recursive Partitioning and Regression Trees, R package version 4.1-11.

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241-266.

Tremblay, M. S., Inman, J. W., & Willms, J. D. (2000). The relationship between physical activity, self-esteem, and academic achievement in 12-year-old children. *Pediatric exercise science*, 12(3), 312-323.

Trudeau, F., & Shephard, R. J. (2008). Physical education, school physical activity, school sports and academic performance. *International Journal of Behavioral Nutrition and Physical Activity*, *5*(1), 10.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3-27.

Van de Mortel, T. F. (2008). Faking it: social desirability response bias in self-report research. *The Australian Journal of Advanced Nursing*, 25(4), 40.

Wang, F., Li, X. L., Wang, J. T., & Ng, S. K. (2017). Guest Editorial: Special Section on Biological Data Mining and Its Applications in Healthcare. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *14*(3), 501-502.

Wang, M. T., & Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. *American Educational Research Journal*, *47*(3), 633-662.

Werbos, P. J. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. *Doctoral Dissertation, Applied Mathematics, Harvard University, MA*.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

David H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 9:1341–1390, 1996.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach*. Nelson Education.

Yu, Z. J., Haghighat, F., & Fung, B. C. (2016). Advances and challenges in building engineering and data mining applications for energy-efficient communities. *Sustainable Cities and Society*, 25, 33-38.

Zambrano Jurado, J. C. (2016). Un estudio multinivel del rendimiento escolar en matemáticas para tercer grado de educación básica primaria en América Latina. *Revista Sociedad y Economía*, (30).

APPENDIX

Appendix A: Chile's Quality of Education thresholds

All the information presented in this appendix comes from *Informe Técnico SIMCE 2014* (Agencia de Calidad de la Educación, 2014).

The Agency defines "The Learning Standards", which describe what students should know and be able to do in order to demonstrate, through the standardized test SIMCE, their achievement level of the learning objectives defined in the national curriculum. "The Learning Standards" consist of three levels: insufficient, elemental and adequate.

The Agency defined threshold scores for fourth and eighth grade. Sixth grade thresholds were not defined because the sixth grade test had not been analyzed by the time the thresholds were defined. Hence, we established the sixth grade thresholds based on a linear interpolation of the fourth and eighth grade thresholds. Table A-1 shows the thresholds employed in this research.

Table A-1: Agency's thresholds define three academic achievement levels.

	Insufficient	Elemental	Adequate
Fourth grade	$X < 241$	$241 \leq X < 284$	$284 \leq X$
Sixth grade*	$X < 243$	$243 \leq X < 288$	$288 \leq X$
Eighth grade	$X < 244$	$244 \leq X < 292$	$292 \leq X$

*The sixth grade thresholds were defined by us based on the thresholds of fourth and eighth grade.

Appendix B: Academic achievement distribution based on prior academic achievement

The following table shows the distribution of academic achievement per achievement level in a prior grade. For example, the upper section in the left ("4º-6º Agency's threshold") shows that from the 37,082 students that have a low achievement in fourth grade, 80% have a low achievement in sixth grade. The ones named with "Agency's threshold" use the division of achievement levels provided by Chile's Quality of Education Agency, the ones named with "Percentile's threshold" use the division we used to separate the sample in three thirds.

Table B-1: Academic achievement distribution based on prior academic achievement.

4º-6º Agency's threshold					4º-6º Percentile's threshold				
		Fourth grade achievement					Fourth grade achievement		
		Low	Mid	High			Low	Mid	High
Sixth grade achievement	Low	80%	42%	13%	Sixth grade achievement	Low	68%	25%	7%
	Mid	18%	43%	35%		Mid	27%	45%	27%
	High	2%	15%	53%		High	5%	30%	66%
		100%	100%	100%			100%	100%	100%
	Total students	37,082	44,420	68,323		Total students	49,835	49,905	50,085
4º-8º Agency's threshold					4º-8º Percentile's threshold				
		Fourth grade achievement					Fourth grade achievement		
		Low	Mid	High			Low	Mid	High
Eighth grade achievement	Low	80%	49%	20%	Eighth grade achievement	Low	63%	27%	10%
	Mid	18%	41%	41%		Mid	30%	43%	27%
	High	2%	10%	39%		High	7%	31%	62%
		100%	100%	100%			100%	100%	100%
	Total students	37,082	44,420	68,323		Total students	49,835	49,905	50,085
6º-8º Agency's threshold					6º-8º Percentile's threshold				
		Sixth grade achievement					Sixth grade achievement		
		Low	Mid	High			Low	Mid	High
Eighth grade achievement	Low	76%	36%	10%	Eighth grade achievement	Low	66%	27%	6%
	Mid	22%	50%	37%		Mid	28%	47%	25%
	High	2%	14%	54%		High	5%	26%	69%
		100%	100%	100%			100%	100%	100%
	Total students	56,969	49,646	43,210		Total students	49,802	49,921	50,102

Appendix C: Academic achievement gap between different group of students

This appendix shows the academic achievement gap between different students. The characteristics analyzed are gender, school's financial dependency and school's socioeconomic status.

Figure C-1 shows the distribution of gender in our sample for all three levels. We analyzed mean score per gender and found that the decreasing tendency was similar for both genders. In addition, we analyzed the 25th and 75th percentile; the decreasing tendency of the percentiles is also similar.

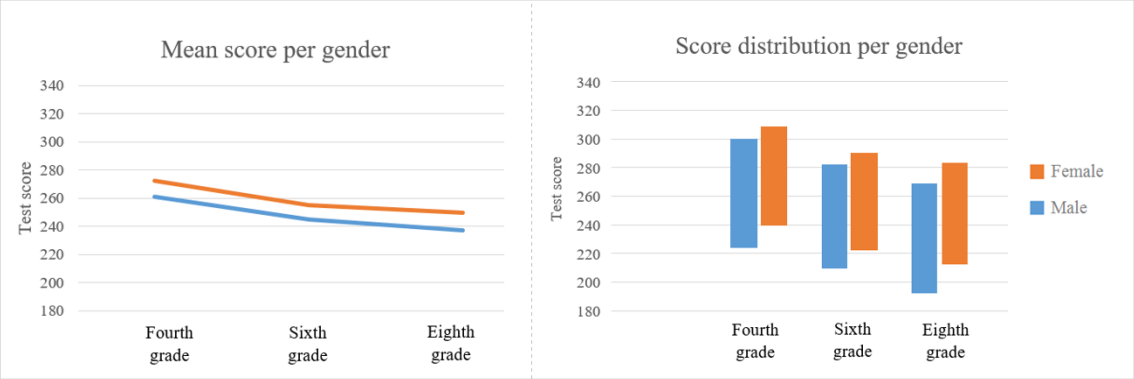


Figure C-1: Mean score and distribution per gender.

We analyzed mean score per school dependency and found that the decreasing tendency was similar for all three dependencies; private, private subsidized and public. The 25th and 75th percentile of the three school dependency options decreased similarly.

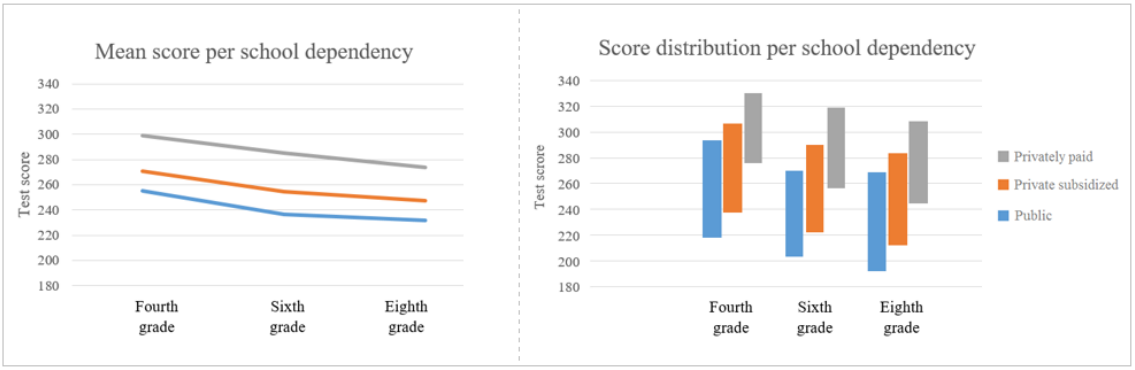


Figure C-2: Mean score and distribution per school financial dependency.

Figure C-3 shows the distribution of school's socioeconomic status (SES) in our sample. First, we analyzed mean score per school's socioeconomic status and found that the decreasing tendency was similar though all socioeconomic status levels. Since means hide a lot of information, we analyzed the 25th and 75th percentile. The decreasing tendency is similar for all socioeconomic status levels.

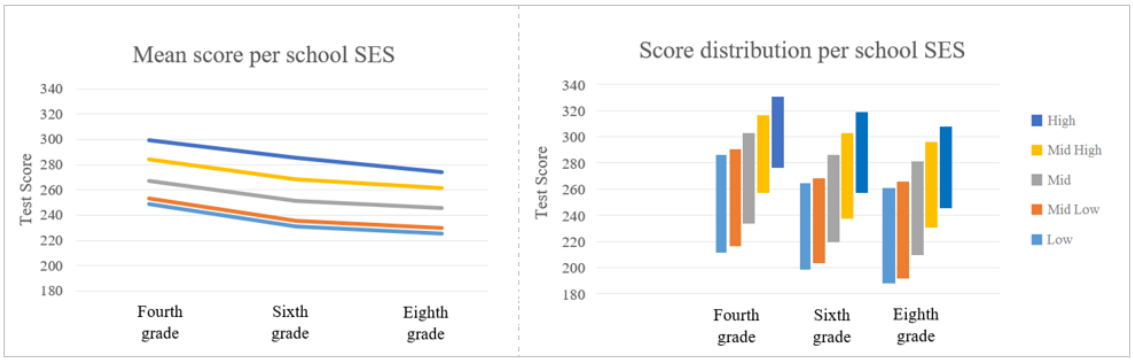


Figure C-3: Mean score and distribution per school's socioeconomic status.

Appendix D: Characterization of students

This appendix shows the distribution of students per gender, school financial dependency and school socioeconomic status (based on the nominal indicator provided by the Quality of Education Agency) for all the students enrolled nationally, students that give our test and the 149,825 students of our initial sample. The following table shows the number of students per grade nationally and the number of students that give the test per grade.

Table D-1: Number of students per level vs number of students that sit for the test.

	Fourth grade (2011)	Sixth grade (2013)	Eighth grade (2015)
Total students enrolled nationally	247,666	262,421	255,607
Students that give the test	216,133	219,856	214,510

The following table shows the distribution of students per gender per grade.

Table D-2: Gender distribution for all students, students that sit for the test and for the sample employed in our study.

Gender							
		Fourth grade		Sixth grade		Eighth grade	
		Female	Male	Female	Male	Female	Male
Percentage of students	All students	49%	51%	48%	52%	49%	51%
	Give test	49%	51%	49%	51%	49%	51%
	Our sample	51%	49%	51%	49%	51%	49%

The following table shows the distribution of students per school's financial dependency per grade.

Table D-3: School financial dependency distribution for all students, students that sit for the test and for the sample employed in our study.

School dependency		Fourth grade			Sixth grade			Eighth grade		
		1	2	3	1	2	3	1	2	3
Percentage of students	All students	41%	51%	7%	42%	51%	7%	43%	50%	8%
	Give test	40%	53%	8%	39%	53%	8%	41%	52%	8%
	Our sample	37%	55%	8%	37%	55%	8%	37%	55%	8%

1 = public, 2 = public subsidized, 3 = privately financed

The following table shows the distribution of students per school's socioeconomic status per grade.

Table D-4: School's socioeconomic status distribution for all students, students that sit for the test and for the sample employed in our study.

SES		Fourth grade					Sixth grade					Eighth grade				
		1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Percentage of students	All students	10%	32%	34%	16%	8%	12%	33%	33%	15%	8%	13%	33%	31%	15%	9%
	Give test	9%	31%	35%	17%	8%	9%	32%	35%	15%	8%	11%	31%	32%	16%	9%
	Our sample	7%	29%	37%	19%	9%	8%	29%	36%	18%	9%	8%	29%	36%	18%	9%

1 = low socioeconomic status, 2 = mid-low socioeconomic status, 3 = mid socioeconomic status, 4 = mid-high socioeconomic status, 5 = high socioeconomic status

Appendix E: Confusion matrix, a brief clarification about them and the baseline to compare them

Table E-1 shows the confusion matrices obtained. The first confusion matrix, "4º-6º Agency's threshold: High achieving students", is discussed to clarify them.

In total, 68,323 students start with a high achievement level in fourth grade, we obtain this number by adding the four values of the matrix. By the time students reach sixth grade, 35,884 did not decrease their achievement level and 32,439 did, we obtain these values by adding horizontally on the "Actual" direction. The algorithm predicts that 34,498 students will not decrease their achievement level by the time they reach sixth grade and 33,825 students will decrease. We obtain these values by adding vertically, on the "Predicted" direction. From the 34,498 students predicted as "Non decrease", the algorithm predicted correctly 24,179 and incorrectly 10,319. From the 33,825 students predicted as "Decrease", the algorithm predicted correctly 22,120 and incorrectly 11,705.

Recall and precision are two indicators to evaluate the quality of the prediction per class. Recall is the number of items selected correctly divided by the total number of items of that class. Precision is the number of items selected correctly divided be the total prediction of that class. So, recall for "Non decrease" students is 24,179 divided by all the 35,884 students that actually did not decrease, 67%. Recall for "Decrease" students is 22,120 divided by all the 32,439 students that actually decrease, 68%. Precision for "Non decrease" students is 24,179 divided by all the 34,498 students predicted as "Non decrease", 70%. Precision for "Decrease" students is 22,120 divided by all the 33,825 students predicted as "Decrease", 65%.

We built a baseline to compare the results, this method was taken from Witten & Frank (2005). We kept the total number of predicted students per class; 34,498 students will not decrease and 33,825 students will decrease. Then, we divided them according to overall proportions, the probability of choosing a student that did not decrease is 53% and the probability of selection a decreasing student is 47%. With this random allocation, recall and precision have the same value as the probability of choosing each class. Hence,

precision and recall for "Non decrease" students is 53% and for "Decrease" students is 47%.

Our predictions obtained higher values of recall: 65% for "Non decrease" students and 68% for "Decrease", higher than 53% and 47% respectively. And higher values of precision: 70% for "Non decrease" students and 65% for "Decrease", higher than 53% and 47% respectively. All nine models, for both thresholds, have better results than their respective baseline. The baseline values can be found in table E-2.

Table E-1: Confusion matrixes for the nine models.

4º-6º Agency's threshold High achieving students		Predicted			4º-6º Percentile's threshold High achieving students		Predicted		
		Non Decrease	Decrease	recall			Non Decrease	Decrease	recall
Actual	Non Decrease	24,179	11,705	67%	Actual	Non Decrease	25,171	7,686	77%
	Decrease	10,319	22,120	68%		Decrease	7,561	9,667	56%
precision		70%	65%		precision		77%	56%	

4º-6º Percentile's threshold Mid achieving students		Predicted			4º-6º Percentile's threshold Mid achieving students		Predicted		
		Non Decrease	Decrease	recall			Non Decrease	Decrease	recall
Actual	Non Decrease	18,267	7,515	71%	Actual	Non Decrease	31,111	6,368	83%
	Decrease	8,096	10,542	57%		Decrease	7,309	5,117	41%
precision		69%	58%		precision		81%	45%	

4º-6º Percentile's threshold Low achieving students		Predicted			4º-6º Percentile's threshold Low achieving students		Predicted		
		Increase	Non Increase	recall			Increase	Non Increase	recall
Actual	Increase	2,945	4,377	40%	Actual	Increase	7,759	8,256	48%
	Non Increase	4,463	25,297	85%		Non Increase	6,779	27,041	80%
precision		40%	85%		precision		53%	77%	

4º-8º Agency's threshold High achieving students	Predicted	4º-8º Percentile's threshold High achieving students	Predicted

		Non Decrease	Decrease	recall
Actual	Non Decrease	15,577	10,815	59%
	Decrease	9,200	32,731	78%
precision		63%	75%	

		Non Decrease	Decrease	recall
Actual	Non Decrease	23,701	7,533	76%
	Decrease	7,583	22,120	60%
precision		76%	60%	

4º-8º Agency's threshold
Mid achieving students

		Predicted Non Decrease	Predicted Decrease	recall
Actual	Non Decrease	14,870	7,845	65%
	Decrease	7,439	14,266	66%
precision		67%	65%	

4º-8º Percentile's threshold
Mid achieving students

		Predicted Non Decrease	Predicted Decrease	recall
Actual	Non Decrease	29,921	6,574	82%
	Decrease	6,944	6,466	48%
precision		81%	50%	

4º-8º Agency's threshold
Low achieving students

		Predicted Increase	Predicted Non Increase	recall
Actual	Increase	3,279	4,062	45%
	Non Increase	4,651	24,870	84%
precision		41%	86%	

4º-8º Percentile's threshold
Low achieving students

		Predicted Increase	Predicted Non Increase	recall
Actual	Increase	10,733	7,899	58%
	Non Increase	7,463	23,740	76%
precision		59%	75%	

6º-8º Agency's threshold
High achieving students

		Predicted Non Decrease	Predicted Decrease	recall
Actual	Non Decrease	15,677	7,451	68%
	Decrease	7,431	12,651	63%
precision		68%	63%	

6º-8º Percentile's threshold
High achieving students

		Predicted Non Decrease	Predicted Decrease	recall
Actual	Non Decrease	27,350	7,365	79%
	Decrease	7,509	7,878	51%
precision		78%	52%	

6º-8º Agency's threshold
Mid achieving students

		Predicted Non Decrease	Predicted Decrease	recall
Actual	Non Decrease	24,702	7,285	77%
	Decrease	9,025	8,634	49%
precision		73%	54%	

6º-8º Percentile's threshold
Mid achieving students

		Predicted Non Decrease	Predicted Decrease	recall
Actual	Non Decrease	29,805	6,482	82%
	Decrease	7,753	5,881	43%
precision		79%	48%	

6º-8º Agency's threshold
Low achieving students

		Predicted Increase	Predicted Non Increase	recall

6º-8º Percentile's threshold
Low achieving students

		Predicted Increase	Predicted Non Increase	recall

Actual	Increase	6,216	7,428	46%	Actual	Increase	8,747	8,010	52%
	Non Increase	7,428	35,509	82%		Non Increase	7,554	25,491	77%
precision		44%	83%		precision		54%	76%	

Table E-2: Baseline for the nine models.

4º-6º Agency's threshold High achieving students		4º-6º Agency's threshold High achieving students	
	Precision and recall baseline		Precision and recall baseline
Non Decrease	53%	Non Decrease	66%
Decrease	47%	Decrease	34%
4º-6º Agency's threshold Mid achieving students		4º-6º Agency's threshold Mid achieving students	
	Precision and recall baseline		Precision and recall baseline
Non Decrease	58%	Non Decrease	75%
Decrease	42%	Decrease	25%
4º-6º Agency's threshold Low achieving students		4º-6º Agency's threshold Low achieving students	
	Precision and recall baseline		Precision and recall baseline
Non Decrease	20%	Non Decrease	32%
Decrease	80%	Decrease	68%
4º-8º Agency's threshold High achieving students		4º-8º Agency's threshold High achieving students	
	Precision and recall baseline		Precision and recall baseline
Non Decrease	39%	Non Decrease	51%
Decrease	61%	Decrease	49%
4º-8º Agency's threshold Mid achieving students		4º-8º Agency's threshold Mid achieving students	
	Precision and recall baseline		Precision and recall baseline
Non Decrease	51%	Non Decrease	73%
Decrease	49%	Decrease	27%
4º-8º Agency's threshold Low achieving students		4º-8º Agency's threshold Low achieving students	
	Precision and recall baseline		Precision and recall baseline
Non Decrease	20%	Non Decrease	37%
Decrease	80%	Decrease	63%
6º-8º Agency's threshold High achieving students		6º-8º Agency's threshold High achieving students	
	Precision and recall baseline		Precision and recall baseline

Non Decrease	54%	Non Decrease	69%
Decrease	46%	Decrease	31%

6º-8º Agency's threshold Mid achieving students		6º-8º Agency's threshold Mid achieving students	
	Precision and recall baseline		Precision and recall baseline
Non Decrease	64%	Non Decrease	73%
Decrease	36%	Decrease	27%

6º-8º Agency's threshold Low achieving students		6º-8º Agency's threshold Low achieving students	
	Precision and recall baseline		Precision and recall baseline
Non Decrease	24%	Non Decrease	34%
Decrease	76%	Decrease	66%

Appendix F: Variable importance rankings

In this appendix, we show the rankings of decrease in accuracy and gini. In total, there are 36 rankings. The following equation helps clarify where the 36 rankings come from:

36 rankings = 2 thresholds (Agency's and percentiles) x 2 indicators (accuracy and gini) x 9 different models

Figures F-1 to F-3 show the rankings for models that predict achievement between fourth and sixth grade, these models count with 774 independent variables. Figures F-4 to F-6 show the rankings for the models that predict the achievement between fourth and eighth grade, these models count with 1,287 independent variables. Finally, Figures F-7 to F-9 show the rankings for the models that predict the achievement between sixth and eighth grade, these models count with 883 independent variables.

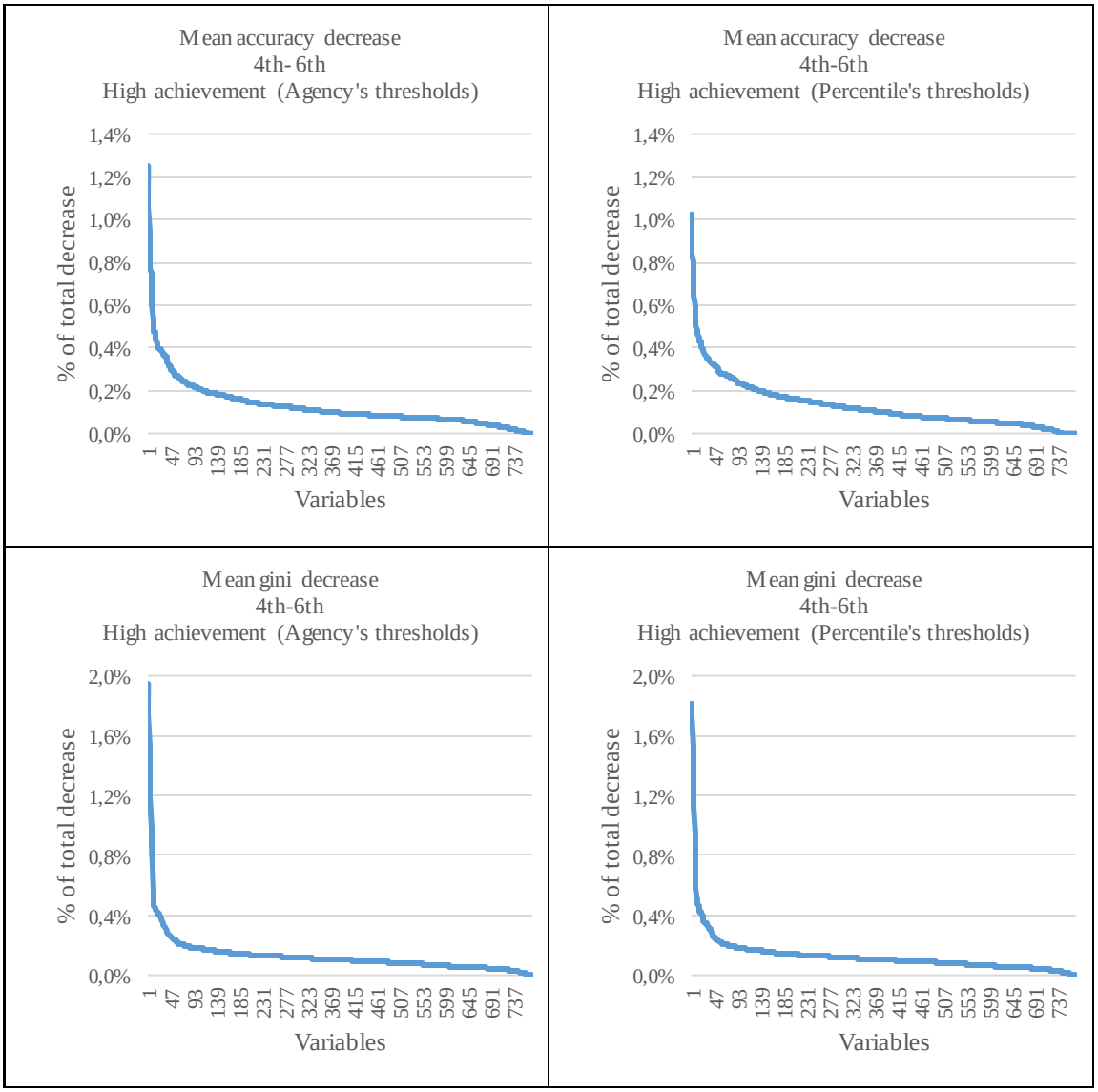


Figure F-1: Ranking of variable's importance for models that predict the trajectory between fourth and sixth grade and start with a high achievement level in fourth grade.

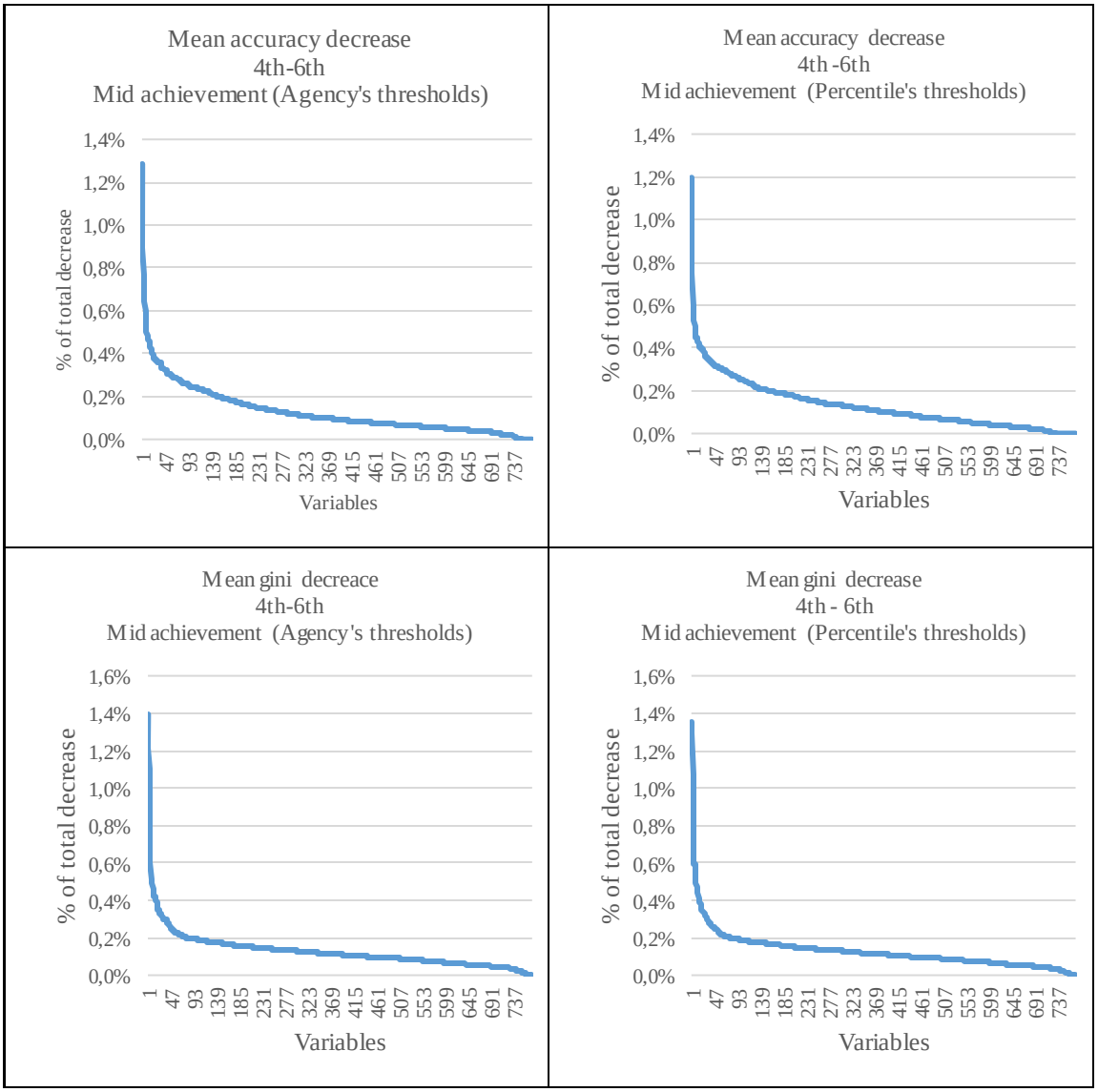


Figure F-2: Ranking of variable's importance for models that predict the trajectory between fourth and sixth grade and start with a mid achievement level in fourth grade.

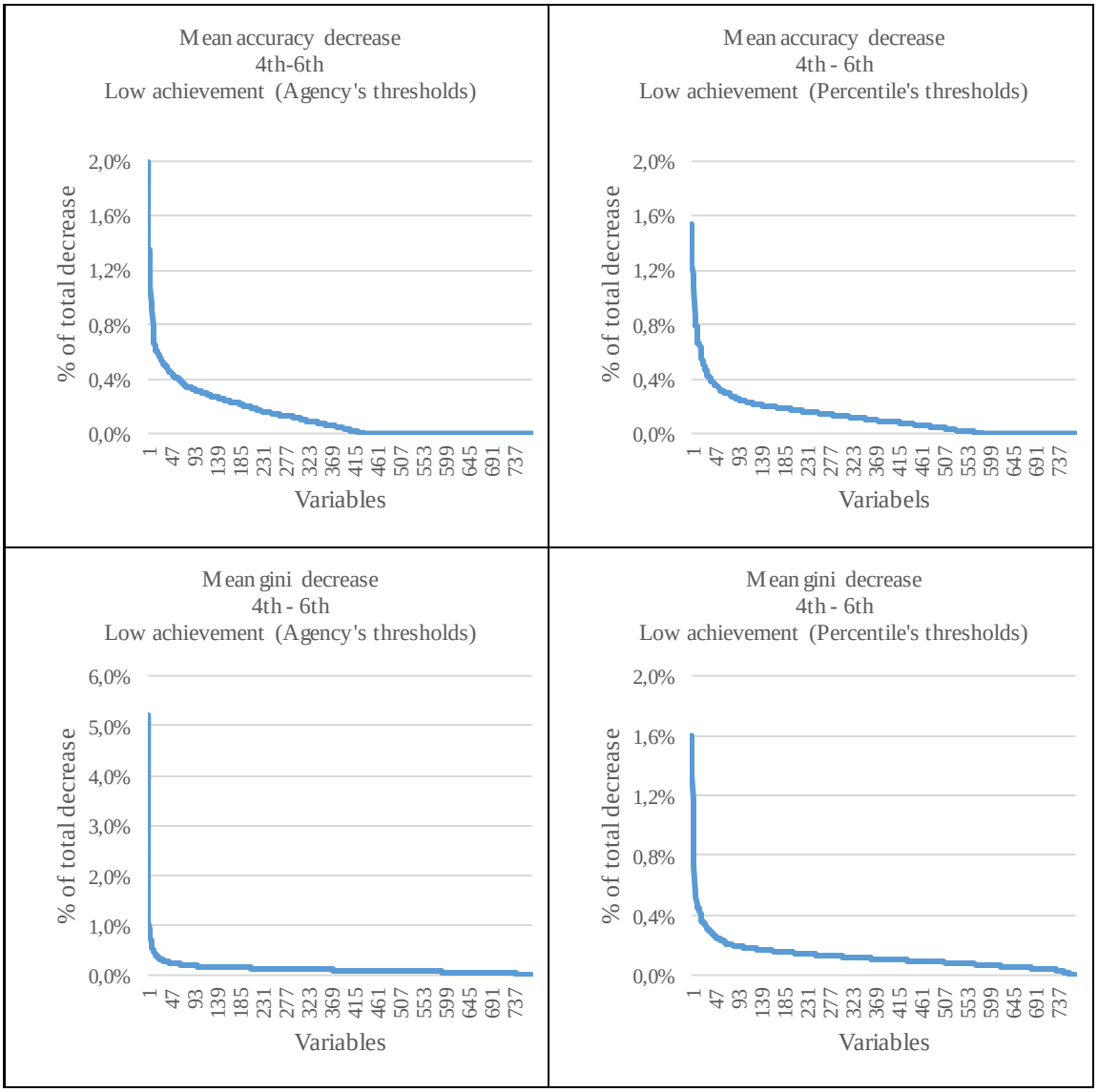


Figure F-3: Ranking of variable's importance for models that predict the trajectory between fourth and sixth grade and start with a low achievement level in fourth grade.

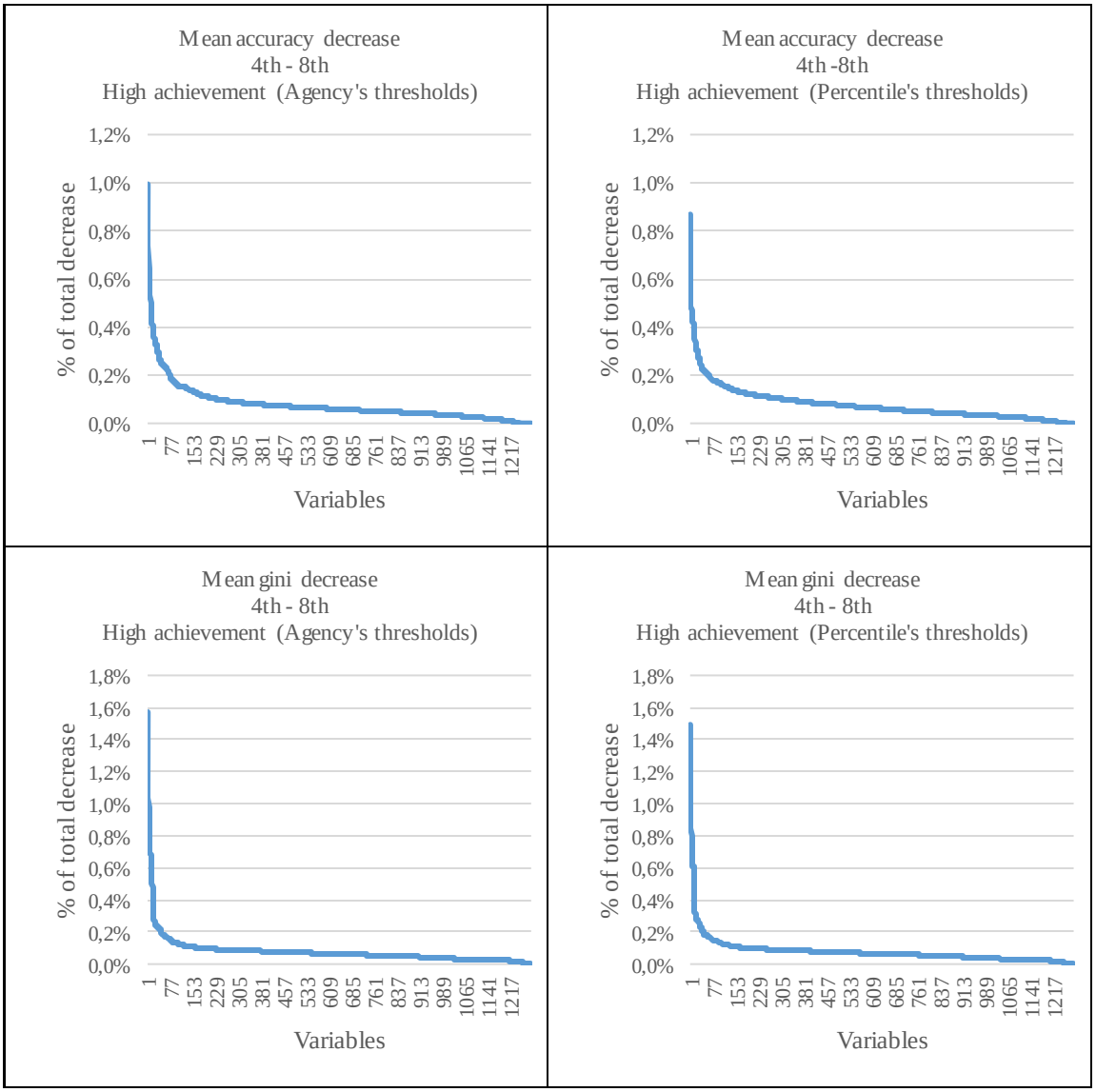


Figure F-4: Ranking of variable's importance for models that predict the trajectory between fourth and eighth grade and start with a high achievement level in fourth grade.

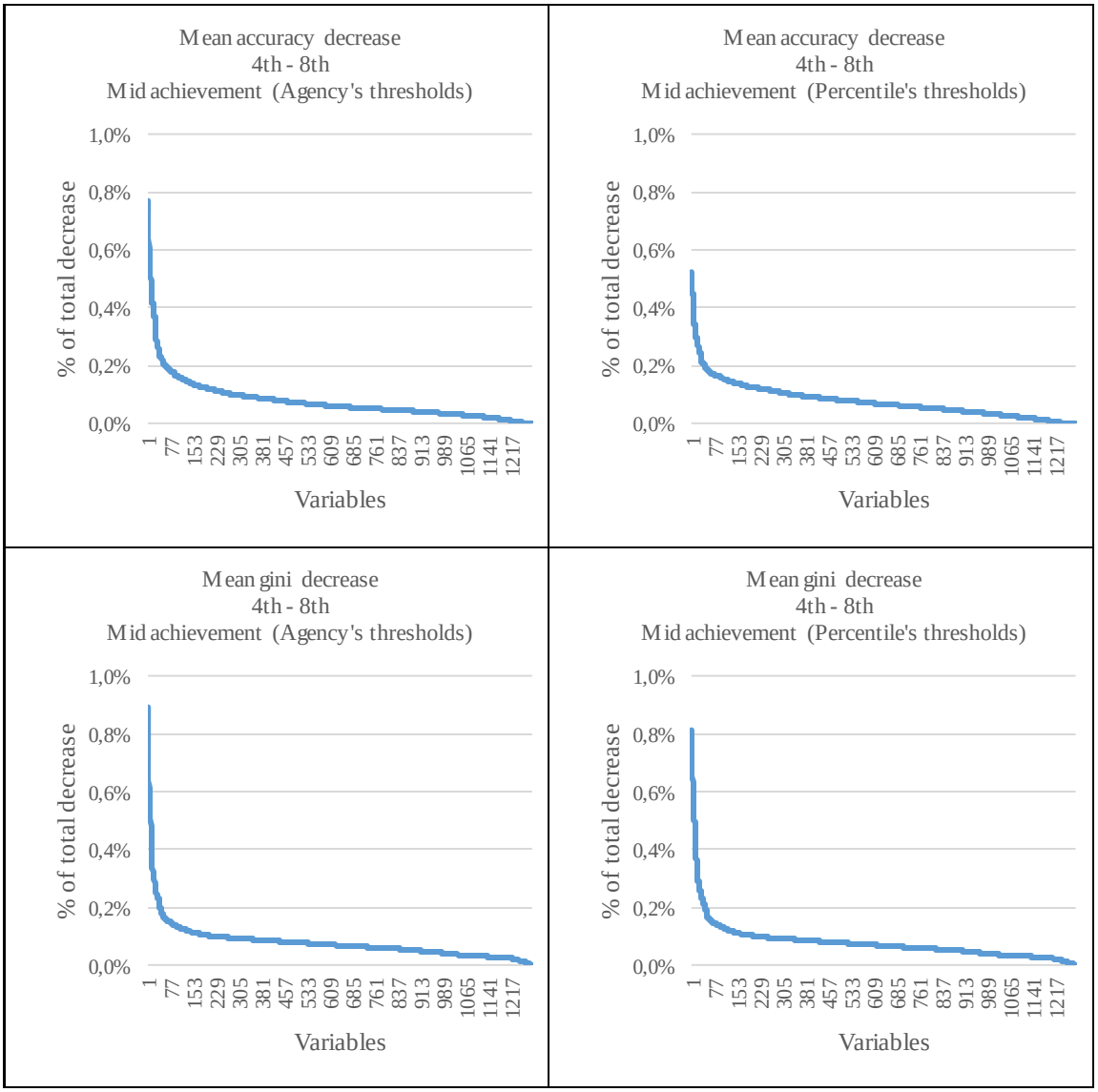


Figure F-5: Ranking of variable's importance for models that predict the trajectory between fourth and eighth grade and start with a mid achievement level in fourth grade.

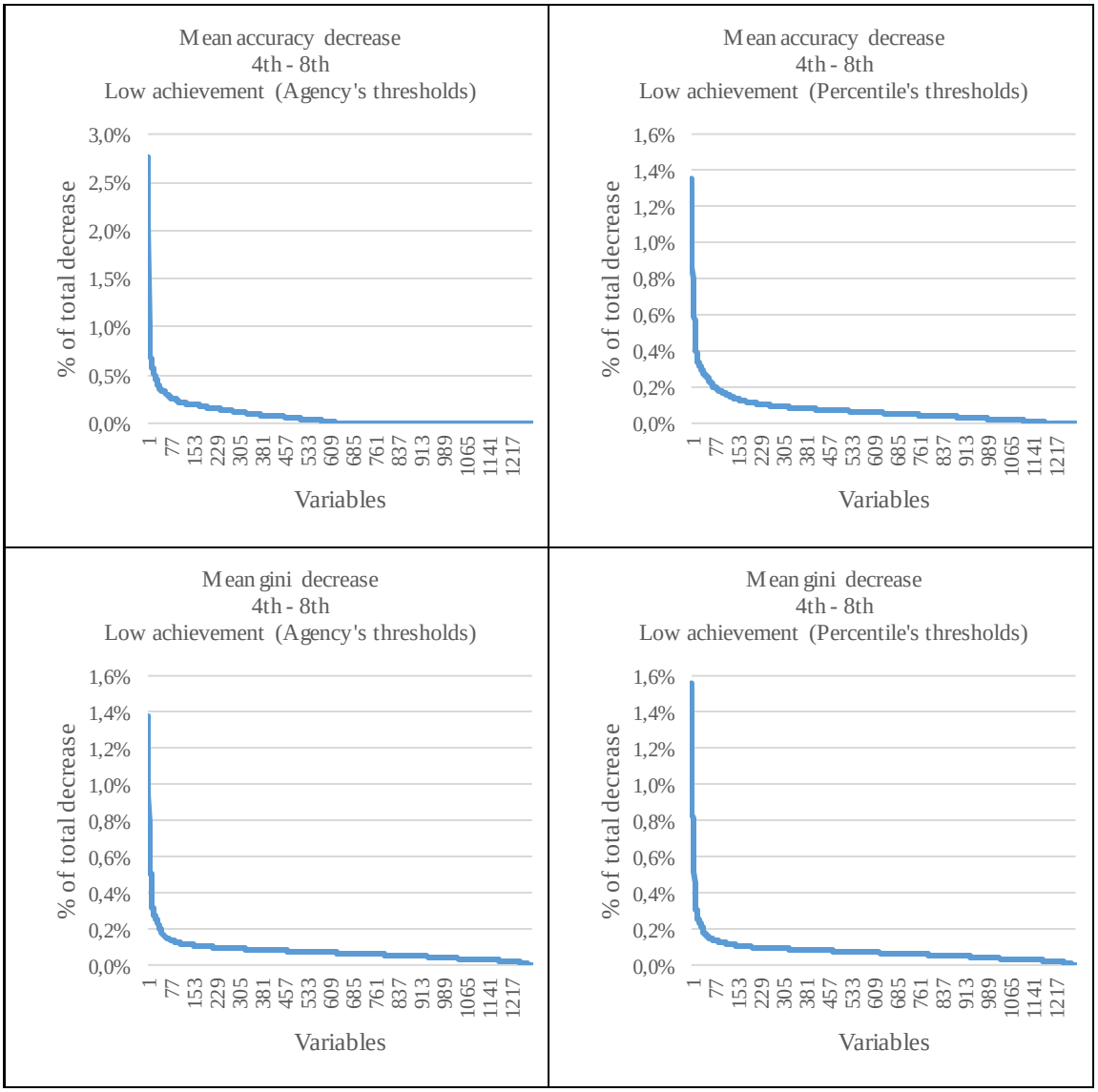


Figure F-6: Ranking of variable's importance for models that predict the trajectory between fourth and eighth grade and start with a low achievement level in fourth grade.

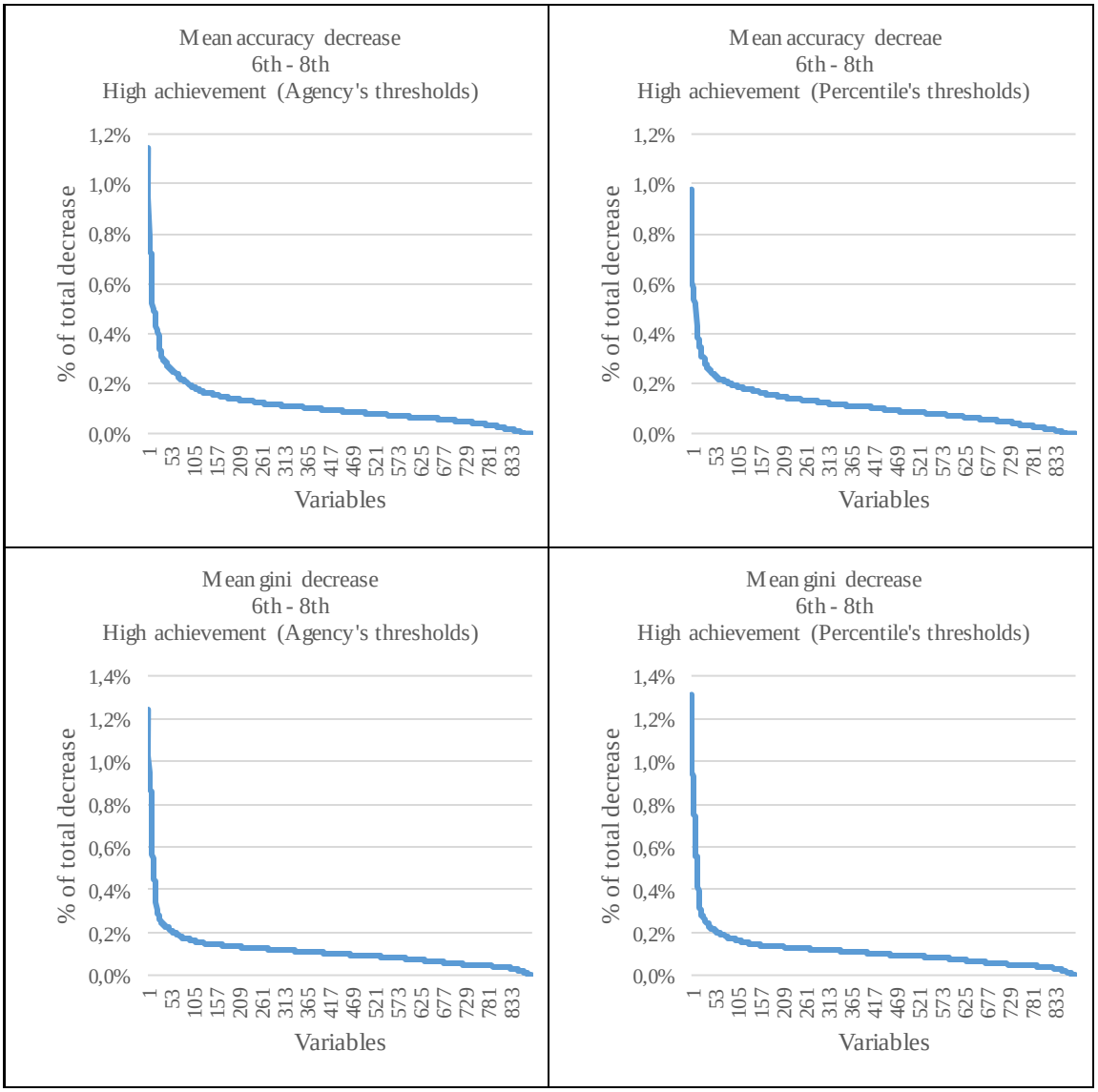


Figure F-7: Ranking of variable's importance for models that predict the trajectory between sixth and eighth grade and start with a high achievement level in fourth grade.

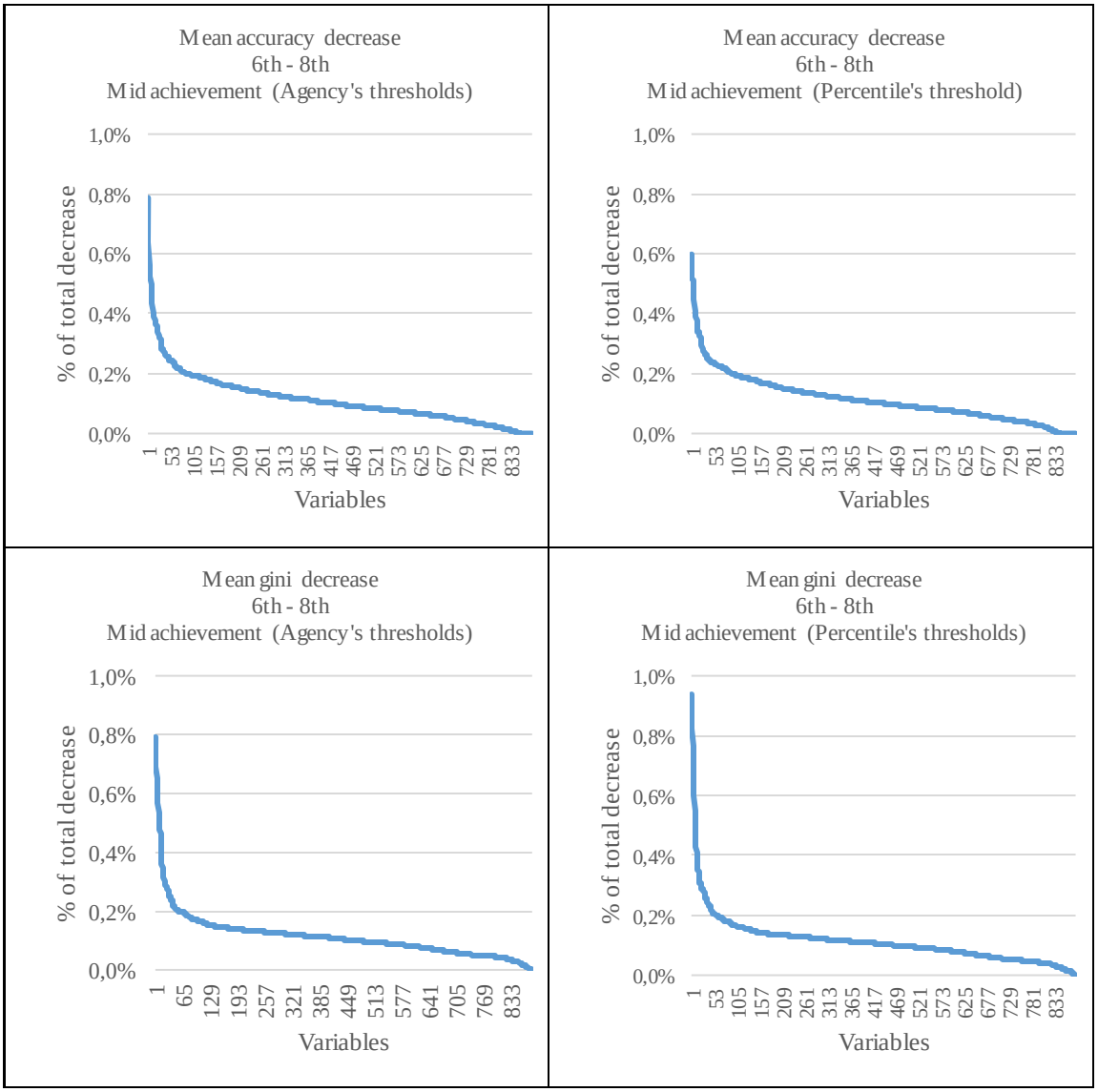


Figure F-8: Ranking of variable's importance for models that predict the trajectory between sixth and eighth grade and start with a mid achievement level in fourth grade.

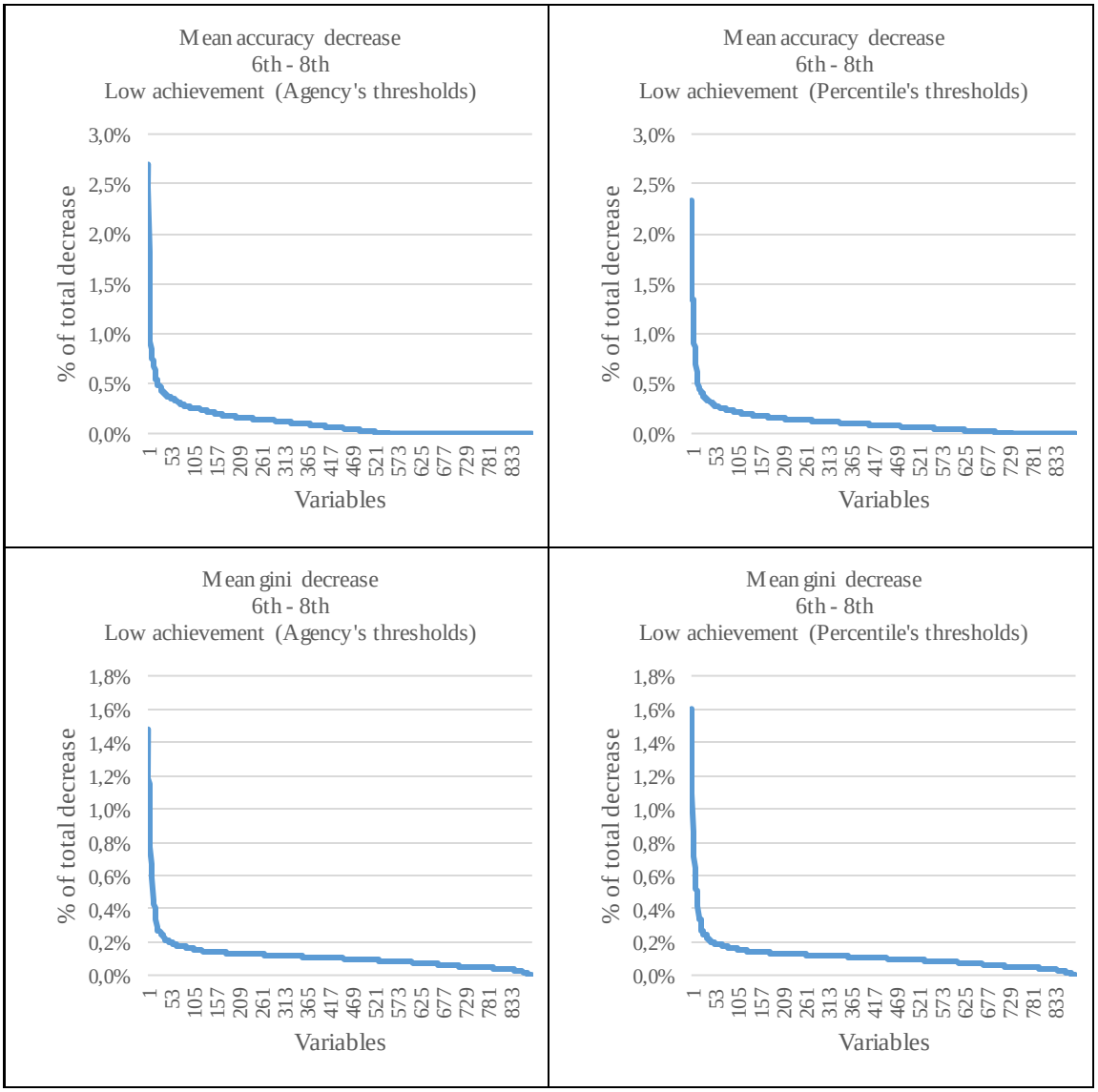


Figure F-9: Ranking of variable's importance for models that predict the trajectory between sixth and eighth grade and start with a low achievement level in fourth grade.

Appendix G: Hierarchical clustering using *hclust()* from R

In this appendix we present the results of the hierarchical clustering obtained by *hclust()* from R (Müllner, 2017). We applied this algorithm to the 246 variables selected with the first data mining algorithm, Random Forest, we obtained a dendogram that represents which variables are closer to each other. Figure G-1 shows the dendogram obtained. Figures G-2 to G-4 are close ups of the dendogram, these allows us to read the keys of the variables. Figure G-2 shows the left side of Figure G-1. Figure G-3 shows the middle of Figure G-1 and Figure G-4 shows the right side of Figure G-1.

Since the questionnaires are not public and must be solicited from Chile's Quality of Education Agency, we do not present the details or meanings of each key in this document.

Details we can precise are:

i) The number that follows the first letter represents the grade the student is coursing when that information was obtained. The variables that do not have this are timeless.
ii) Keys that have "cprof", "cpad" and "cest" are questions from the teacher, parent and student's questionnaires respectively.
iii) The key to identify students is "mrun", we also used variable "x" to enumerate the samples. The key to identify schools are "rbd4", "rbd6" and "rbd8" for fourth, sixth and eighth grade respectively.
iv) We constructed the variables "CAMBIO4_6", "CAMBIO4_8" and "CAMBIO6_8" to acknowledge change of school between fourth and sixth, fourth and eighth, and sixth and eighth respectively. We also constructed the socioeconomic variables "nse_libr", "nse_ing", "nse_ed_madre" that represent number of books, household's income and mother's educational level per student. We grouped these three variables and created the index "nse_est". Finally, we averaged the values of this last index per school and obtained "a4_nse", "a6_nse" and "a8_nse" to obtain a school level indicator for fourth, sixth and eighth grade respectively

v) The rest of the variables characterize the school.

The highest division in Figure G-1 separates the information obtained by the student's questionnaire with the rest of the information, mainly questions from teacher and parent's questionnaires. This shows that although these three actors are asked about common features, their answers do not relate to each other's. For example, teachers, parents and students are asked about school climate but their answers about these topic are not close from each other's.

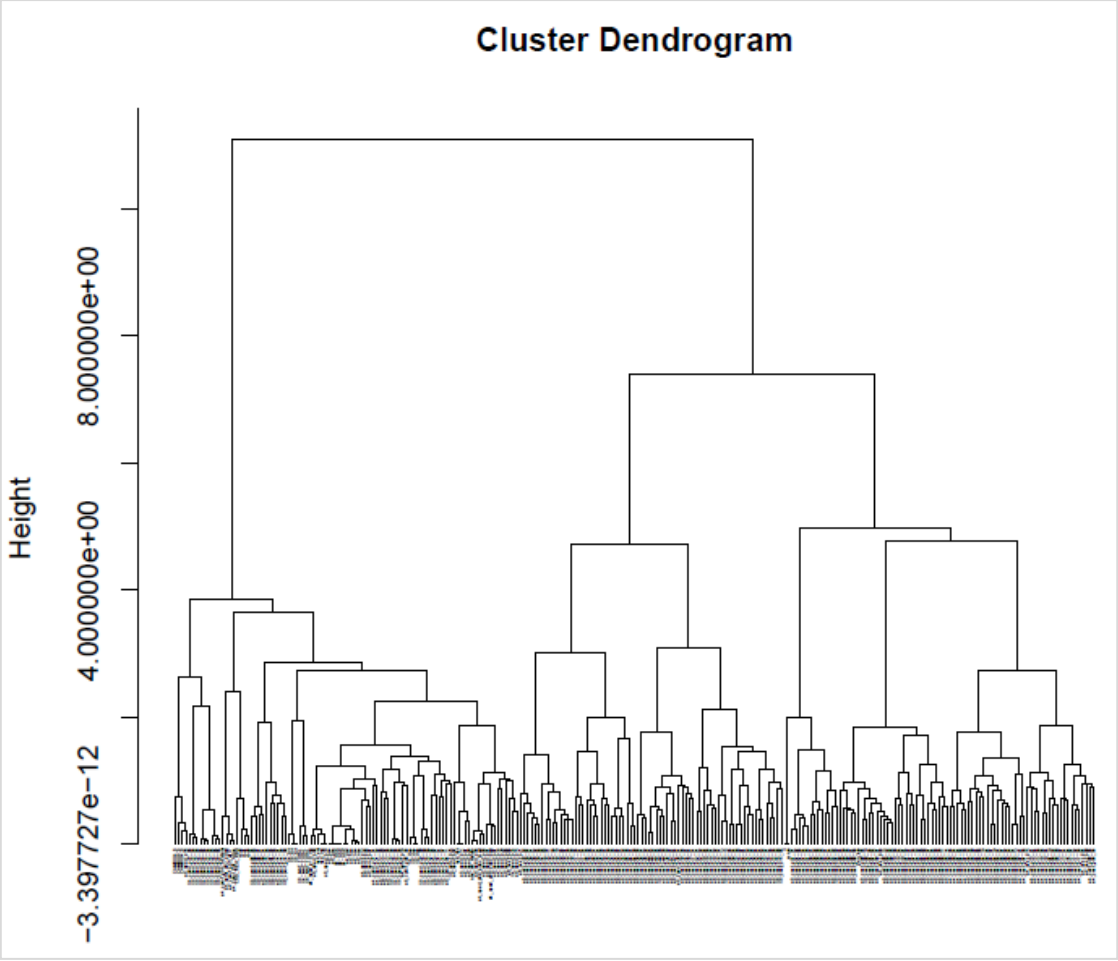


Figure G-1: Dendogram obtained with *hclust()*.

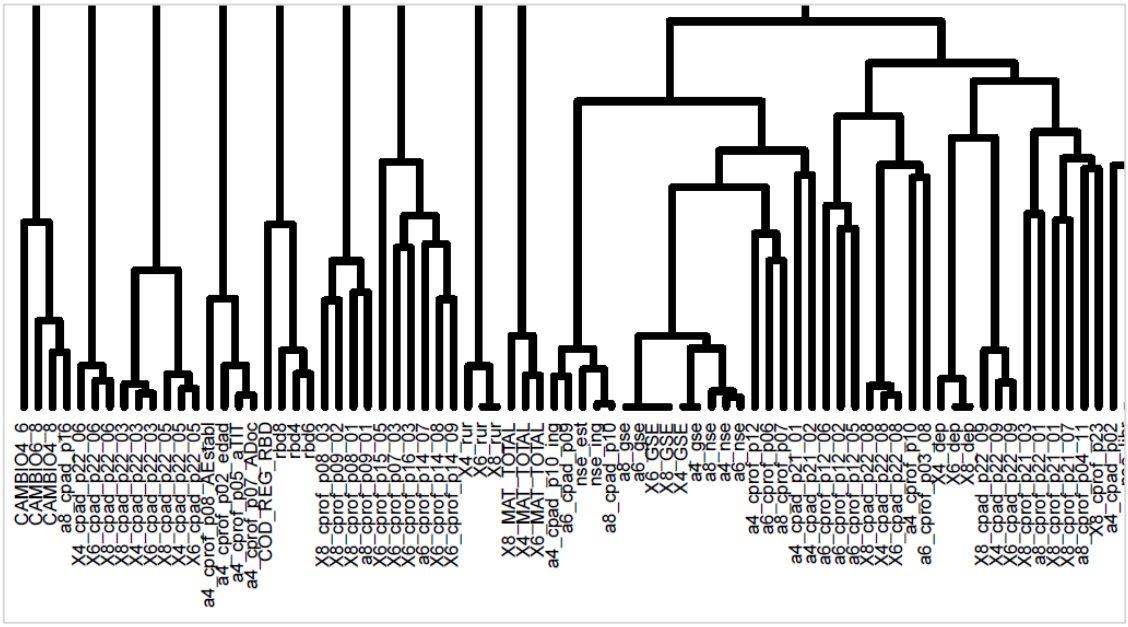


Figure G-2: Close up of the left side of Figure G-1.

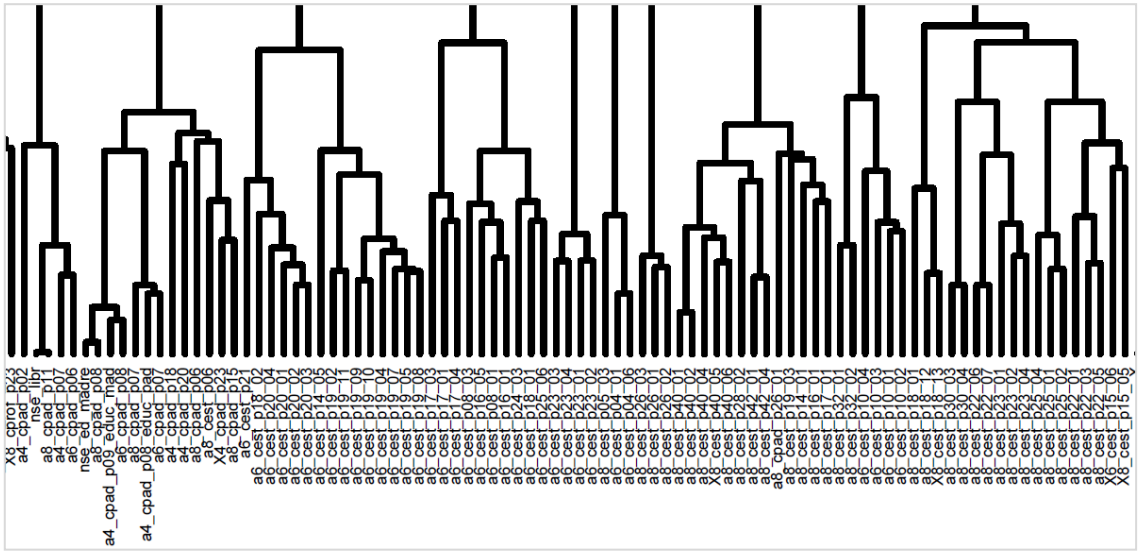


Figure G-3: Close up of the middle of Figure G-1.

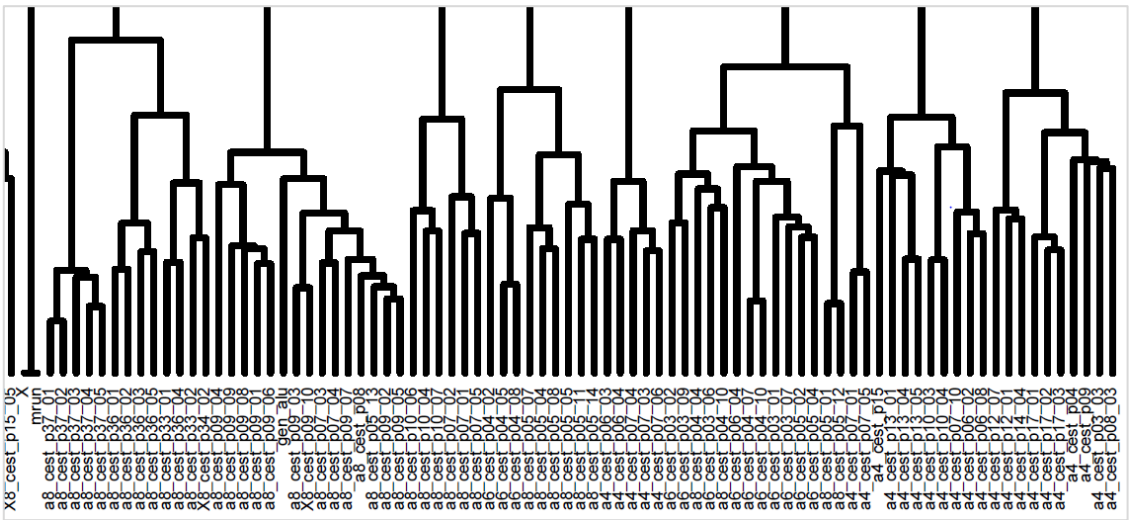


Figure G-4: Close up of the right side of Figure G-1.

We assigned a number to represent variables that where close to each other. Table G-1 shows the number assignment, it was done from right to left.

Table G-1: Assignment of numbers to variables to identify variables close to each other.

Key	Grade	Actor	Dendogram order
X4_cest_p03_03	4	Student	1
X4_cest_p04	4	Student	1
X4_cest_p08_03	4	Student	1
X4_cest_p09	4	Student	1
X4_cest_p17_01	4	Student	2
X4_cest_p17_02	4	Student	2
X4_cest_p17_03	4	Student	2
X4_cest_p12_01	4	Student	3
X4_cest_p12_07	4	Student	3
X4_cest_p14_04	4	Student	3
X4_cest_p06_02	4	Student	4
X4_cest_p06_08	4	Student	4
X4_cest_p07_10	4	Student	4
X4_cest_p10_03	4	Student	5
X4_cest_p10_04	4	Student	5
X4_cest_p13_04	4	Student	6
X4_cest_p13_05	4	Student	6
X4_cest_p13_01	4	Student	7
X4_cest_p15	4	Student	7
X4_cest_p07_01	4	Student	8
X4_cest_p07_05	4	Student	8
X8_cest_p05_01	8	Student	9
X8_cest_p05_12	8	Student	9
X6_cest_p03_01	6	Student	10
X6_cest_p03_07	6	Student	10
X6_cest_p05_02	6	Student	10
X6_cest_p05_04	6	Student	10
X6_cest_p04_07	6	Student	11
X6_cest_p04_10	6	Student	11
X6_cest_p06_04	6	Student	12
X6_cest_p03_02	6	Student	13
X6_cest_p03_06	6	Student	13
X6_cest_p03_09	6	Student	13
X8_cest_p04_04	8	Student	13
X8_cest_p04_10	8	Student	13
X4_cest_p06_03	4	Student	14
X4_cest_p06_04	4	Student	14
X4_cest_p07_03	4	Student	14
X4_cest_p07_04	4	Student	14
X4_cest_p07_06	4	Student	14
X8_cest_p05_05	8	Student	15
X8_cest_p05_11	8	Student	15
X8_cest_p05_14	8	Student	15
X8_cest_p05_04	8	Student	16
X8_cest_p05_07	8	Student	16
X8_cest_p05_08	8	Student	16
X6_cest_p04_05	6	Student	17
X6_cest_p04_08	6	Student	17
X6_cest_p04_02	6	Student	18
X8_cest_p07_01	8	Student	19
X8_cest_p07_02	8	Student	19
X8_cest_p07_05	8	Student	19
X8_cest_p10_04	8	Student	20
X8_cest_p10_06	8	Student	20
X8_cest_p10_07	8	Student	20
X8_cest_p05_13	8	Student	21
X8_cest_p07_03	8	Student	21
X8_cest_p07_04	8	Student	21
X8_cest_p08	8	Student	21
X8_cest_p09_02	8	Student	21
X8_cest_p09_05	8	Student	21
X8_cest_p09_07	8	Student	21
X8_cest_p09_03	8	Student	22
X8_cest_p09_10	8	Student	22
gen_alu	-	Student	23
X8_cest_p09_01	8	Student	24
X8_cest_p09_06	8	Student	24
X8_cest_p09_08	8	Student	24
X8_cest_p09_09	8	Student	24
X8_cest_p09_04	8	Student	25
X8_cest_p33_02	8	Student	26
X8_cest_p34_02	8	Student	26
X8_cest_p33_01	8	Student	27
X8_cest_p36_04	8	Student	27
X8_cest_p36_03	8	Student	28
X8_cest_p36_05	8	Student	28
X8_cest_p36_01	8	Student	29
X8_cest_p36_02	8	Student	29
X8_cest_p37_04	8	Student	29
X8_cest_p37_05	8	Student	29
X8_cest_p37_01	8	Student	30
X8_cest_p37_02	8	Student	30
X8_cest_p37_03	8	Student	31
X6_cest_p15_06	6	Student	32
X8_cest_p15_05	8	Student	32
X8_cest_p22_01	8	Student	33
X8_cest_p22_03	8	Student	33
X8_cest_p22_05	8	Student	33
X8_cest_p25_01	8	Student	34
X8_cest_p25_02	8	Student	34
X8_cest_p25_04	8	Student	34
X8_cest_p23_01	8	Student	35

X8_cest_p23_02	8	Student	35
X8_cest_p23_04	8	Student	35
X8_cest_p22_06	8	Student	36
X8_cest_p22_07	8	Student	36
X8_cest_p30_03	8	Student	37
X8_cest_p30_04	8	Student	37
X8_cest_p18_12	8	Student	38
X8_cest_p18_13	8	Student	38
X8_cest_p18_01	8	Student	39
X6_cest_p10_01	6	Student	40
X6_cest_p10_02	6	Student	40
X6_cest_p10_03	6	Student	40
X6_cest_p10_04	6	Student	41
X8_cest_p32_01	8	Student	42
X8_cest_p32_02	8	Student	42
X8_cest_p14_01	8	Student	43
X8_cest_p16_01	8	Student	43
X8_cest_p17_01	8	Student	43
X8_cest_p19_03	8	Student	43
X8_cpad_p26_01	8	Parents	43
X8_cest_p42_01	8	Student	44
X8_cest_p42_04	8	Student	44
X8_cest_p28_02	8	Student	45
X8_cest_p40_05	8	Student	46
X8_cest_p40_06	8	Student	46
X8_cest_p40_04	8	Student	47
X8_cest_p40_01	8	Student	48
X8_cest_p40_02	8	Student	48
X8_cest_p26_01	8	Student	49
X8_cest_p26_02	8	Student	49
X8_cest_p26_03	8	Student	49
X6_cest_p04_01	6	Student	50
X6_cest_p04_06	6	Student	50
X8_cest_p05_03	8	Student	50
X6_cest_p23_01	6	Student	51
X6_cest_p23_02	6	Student	51
X6_cest_p23_03	6	Student	52
X6_cest_p23_04	6	Student	52
X6_cest_p17_01	6	Student	53
X6_cest_p17_03	6	Student	53
X6_cest_p17_04	6	Student	53
mrun	-	Student	54
X6_cest_p18_01	6	Student	55
X6_cest_p24_03	6	Student	55
X6_cest_p25_06	6	Student	55
X6_cest_p08_01	6	Student	56
X6_cest_p08_03	6	Student	56
X6_cest_p16_01	6	Student	56
X6_cest_p16_05	6	Student	56
X6_cest_p19_04	6	Student	57
X6_cest_p19_05	6	Student	57
X6_cest_p19_07	6	Student	57
X6_cest_p19_08	6	Student	57
X6_cest_p19_09	6	Student	58
X6_cest_p19_10	6	Student	58
X6_cest_p19_02	6	Student	59
X6_cest_p19_11	6	Student	59
X6_cest_p14_05	6	Student	59
X6_cest_p20_02	6	Student	60
X6_cest_p20_03	6	Student	60
X6_cest_p20_01	6	Student	61
X6_cest_p20_04	6	Student	61
X6_cest_p18_02	6	Student	62
X6_cest_p21	6	Student	63
X4_cpad_p23	4	Parents	64
X8_cpad_p15	8	Parents	64
X8_cest_p06	8	Student	65
X8_cpad_p06	8	Parents	66
X4_cpad_p18	4	Parents	67
X4_cpad_p20	4	Parents	67
X4_cpad_p08	4	Parents	68
X6_cpad_p07	6	Parents	68
X8_cpad_p07	8	Parents	68
nse_ed_madre	-	Parents	69
X4_cpad_p09	4	Parents	69
X6_cpad_p08	6	Parents	69
X8_cpad_p08	8	Parents	69
nse_libr	-	Parents	70
X4_cpad_p07	4	Parents	70
X6_cpad_p06	6	Parents	70
X8_cpad_p11	8	Parents	70
X4_cpad_p02	4	Parents	71
X8_cprof_p04_11	8	Teacher	72
X8_cprof_p23	8	Teacher	72
X8_cprof_p21_01	8	Teacher	73
X8_cprof_p21_07	8	Teacher	73
X8_cprof_p21_03	8	Teacher	74
X8_cprof_p22_01	8	Teacher	74
X4_cpad_p22_09	4	Parents	75
X6_cpad_p22_09	6	Parents	75
X8_cpad_p22_09	8	Parents	75
X4_dep	4	School	76
X6_dep	6	School	76
X8_dep	8	School	76
X4_cprof_p10	4	Teacher	77
X6_cprof_p03_08	6	Teacher	77
X4_cpad_p22_08	4	Parents	78
X6_cpad_p22_08	6	Parents	78
X8_cpad_p22_08	8	Parents	78
X6_cprof_p12_02	6	Teacher	79
X6_cprof_p12_05	6	Teacher	79
X6_cprof_p12_06	6	Teacher	79

X4_cpad_p21_01	4	Parents	80
X4_cpad_p21_02	4	Parents	80
X4_cprof_p12	4	Teacher	81
X6_cprof_p06	6	Teacher	81
X8_cprof_p07	8	Teacher	81
X4_GSE	4	School	82
X4_nse	4	School	82
X6_GSE	6	School	82
X6_nse	6	School	82
X8_GSE	8	School	82
X8_nse	8	School	82
nse_est	-	Parents	83
nse_ing	-	Parents	83
X4_cpad_p10	4	Parents	83
X6_cpad_p09	6	Parents	83
X8_cpad_p10	8	Parents	83
X6_cprof_p14_07	6	Teacher	84
X6_cprof_p14_08	6	Teacher	84
X6_cprof_p07_03	6	Teacher	85
X8_cprof_p08_01	8	Teacher	86
X8_cprof_p09_01	8	Teacher	86
X8_cprof_p08_03	8	Teacher	87
X8_cprof_p09_02	8	Teacher	87
X4_MAT_TOTAL	4	School	88
X6_MAT_TOTAL	6	School	88
X8_MAT_TOTAL	8	School	88
X4_rur	4	School	89
X6_rur	6	School	89
X8_rur	8	School	89
X4_cpad_p22_05	4	Parents	90
X6_cpad_p22_05	6	Parents	90
X8_cpad_p22_05	8	Parents	90
X4_cpad_p22_03	4	Parents	91
X6_cpad_p22_03	6	Parents	91
X8_cpad_p22_03	8	Parents	91
X4_cpad_p22_06	4	Parents	92
X6_cpad_p22_06	6	Parents	92
X8_cpad_p22_06	8	Parents	92
CAMBIO4_6	6	Student	93
CAMBIO4_8	8	Student	93
CAMBIO6_8	8	Student	93
X8_cpad_p16	8	Parents	93
rbd4	4	School	94
rbd6	6	School	94
rbd8	8	School	94
COD_REG_8	8	School	95
COD_REG_6	6	School	95
COD_REG_4	4	School	95
X4_cprof_p02	4	Teacher	96
X4_cprof_p05	4	Teacher	96
X4_cprof_p07	4	Teacher	96
X4_cprof_p08	4	Teacher	96

Appendix H: Structural Equation Models

We built a structural equation model for each actor and grade. In this appendix we describe the structural equation models for i) students in fourth grade, ii) students in sixth grade, iii) students in eighth grade, iv) teachers in sixth grade and v) teachers in eighth grade. The combinations of actors and grades not mentioned did not have variables that could be grouped and are used individually in the econometric model. The variables analyzed here are the 275 variables identified as relevant by the data mining algorithm, minus variables that had more than 15% of missing or double marked values (-20 variables), minus questions were the respondent was not the main source of information of the question (-24 variables), plus some control variables described in section 4.5. Data Preparation and 5.2. Data Preparation (+11 variables, plus key variables to follow students and school belonging in fourth, sixth and eighth grade (+4 variables). In total, these add up 246 variables.

To start the modelling process we used as input the information provided by the dendogram detailed in appendix G. Table H-1 shows all the important variables from the fourth grade student's questionnaire. It also shows the indexes created, "No group" means that we did not group the question into an index and we used it individually in the model.

Table H-1: Variables from the fourth grade student's questionnaire.

Key	Dendogram group	Index
X4_cest_p03_03	1	No group
X4_cest_p04	1	No group
X4_cest_p08_03	1	No group
X4_cest_p09	1	No group
X4_cest_p17_01	2	Extracurricular participation
X4_cest_p17_02	2	Extracurricular participation
X4_cest_p17_03	2	Extracurricular participation
X4_cest_p12_01	3	School climate
X4_cest_p12_07	3	School climate
X4_cest_p14_04	3	School climate
X4_cest_p06_02	4	Effort and hard work
X4_cest_p06_08	4	Effort and hard work
X4_cest_p07_10	4	Effort and hard work
X4_cest_p10_03	5	Basic teaching practices
X4_cest_p10_04	5	Basic teaching practices
X4_cest_p13_04	6	Advance teaching practices
X4_cest_p13_05	6	Advance teaching practices
X4_cest_p13_01	7	No group
X4_cest_p15	7	No group
X4_cest_p07_01	8	No group
X4_cest_p07_05	8	No group
X4_cest_p06_03	14	General self-perception
X4_cest_p06_04	14	General self-perception
X4_cest_p07_03	14	Math self-perception
X4_cest_p07_04	14	Math self-perception
X4_cest_p07_06	14	Math self-perception

Figure H-1 shows the structural equation model obtained. Figure H-2 shows the level of fit of the structural equation model. We used the following thresholds to approve the models: i) Root mean square error of approximation (RMSEA) equal or lower than 0.05, ii) Comparative fit index (CFI) higher than 0.9, iii) Tucker-Lexis index (TLI) higher than 0.9, and iv) the parameters that quantify the relationship between observed and latent

variables are equal or higher than 0.40, with the variance of each latent variable standardized to 1.

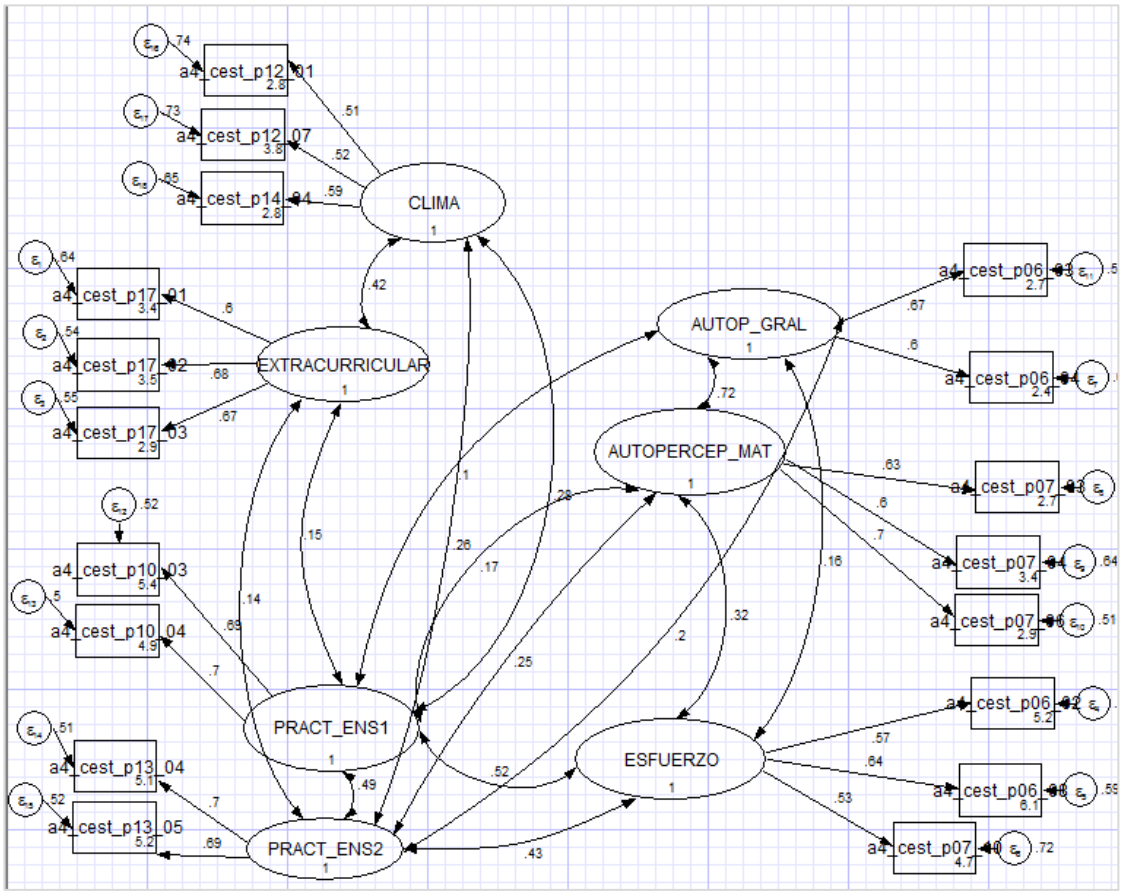


Figure H-1: Structural equation model for variables from fourth grade student's questionnaire.

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(120)	19717.650	model vs. saturated
p > chi2	0.000	
chi2_bs(153)	460471.467	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.033	Root mean squared error of approximation
90% CI, lower bound	0.000	
upper bound	.	
pclose	.	Probability RMSEA <= 0.05
Information criteria		
AIC	6.616e+06	Akaike's information criterion
BIC	6.617e+06	Bayesian information criterion
Baseline comparison		
CFI	0.957	Comparative fit index
TLI	0.946	Tucker-Lewis index
Size of residuals		
SRMR	0.037	Standardized root mean squared residual
CD	0.998	Coefficient of determination

Figure H-2: Level of fit of the structural equation model detailed in Figure H-1.

Table H-2, Figure H-3 and Figure H-4 correspond to variables from the sixth grade student's questionnaire. Table H-3, Figure H-5 and Figure H-6 stand for variables from the eighth grade student's questionnaire. Table H-4, Figure H-7 and Figure H-8 are variables from the sixth grade teacher's questionnaire. Finally, Table H-5, Figure H-9 and Figure H-10 correspond to variables from the eighth grade teacher's questionnaire.

Table H-2: Variables from the sixth grade student's questionnaire.

Key	Dendogram group	Index
X6_cest_p03_01	10	General self-perception
X6_cest_p03_07	10	General self-perception
X6_cest_p05_02	10	General self-perception
X6_cest_p05_04	10	No group
X6_cest_p04_07	11	Math self-perception

X6_cest_p04_10	11	Math self-perception
X6_cest_p06_04	12	No group
X6_cest_p03_02	13	No group
X6_cest_p03_06	13	General self-perception
X6_cest_p03_09	13	No group
X6_cest_p04_05	17	Language self-perception
X6_cest_p04_08	17	Language self-perception
X6_cest_p04_02	18	No group
X6_cest_p15_06	32	No group
X6_cest_p10_01	40	Fast food consuption
X6_cest_p10_02	40	Fast food consuption
X6_cest_p10_03	40	Fast food consuption
X6_cest_p10_04	41	No group
X6_cest_p04_01	50	Sport self-perception
X6_cest_p04_06	50	Sport self-perception
X6_cest_p23_01	51	Extracurricular participation
X6_cest_p23_02	51	Extracurricular participation
X6_cest_p23_03	52	Extracurricular participation
X6_cest_p23_04	52	Extracurricular participation
X6_cest_p17_01	53	Missconduct impressions
X6_cest_p17_03	53	Missconduct impressions
X6_cest_p17_04	53	Missconduct impressions
X6_cest_p18_01	55	No group
X6_cest_p24_03	55	No group
X6_cest_p25_06	55	No group
X6_cest_p08_01	56	School climate
X6_cest_p08_03	56	No group
X6_cest_p16_01	56	School climate
X6_cest_p16_05	56	School climate
X6_cest_p19_04	57	Bullying
X6_cest_p19_05	57	Bullying
X6_cest_p19_07	57	Bullying
X6_cest_p19_08	57	Bullying
X6_cest_p19_09	58	Bullying
X6_cest_p19_10	58	Bullying
X6_cest_p19_02	59	No group
X6_cest_p19_11	59	Bullying
X6_cest_p14_05	59	No group
X6_cest_p20_02	60	Bullying victim
X6_cest_p20_03	60	Bullying victim

X6_cest_p20_01	61	Bullying victim
X6_cest_p20_04	61	Bullying victim
X6_cest_p18_02	62	Bullying victim
X6_cest_p21	63	Bullying victim

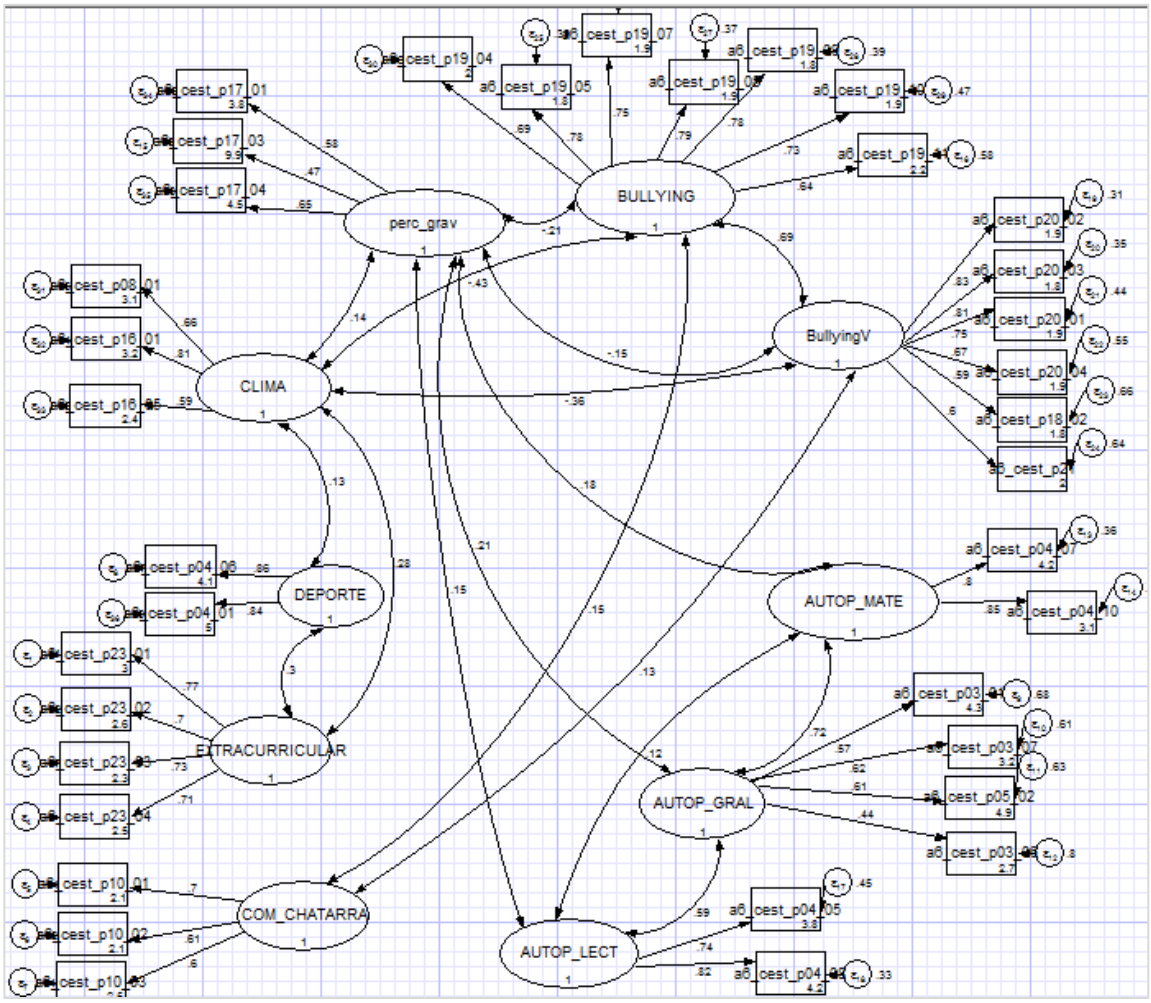


Figure H-3: Structural equation model for variables from sixth grade student's questionnaire.

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(577)	137552.151	model vs. saturated
p > chi2	0.000	
chi2_bs(630)	1.954e+06	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.040	Root mean squared error of approximation
90% CI, lower bound	0.000	
upper bound	.	
pclose	.	Probability RMSEA <= 0.05
Information criteria		
AIC	1.107e+07	Akaike's information criterion
BIC	1.107e+07	Bayesian information criterion
Baseline comparison		
CFI	0.930	Comparative fit index
TLI	0.923	Tucker-Lewis index
Size of residuals		
SRMR	0.048	Standardized root mean squared residual
CD	1.000	Coefficient of determination

Figure H-4: Level of fit of the structural equation model detailed in Figure H-3.

Table H-3: Variables from the eighth grade student's questionnaire.

Key	Dendogram group	Index
X8_cest_p05_01	9	Math self-perception
X8_cest_p05_12	9	Math self-perception
X8_cest_p04_04	13	No group
X8_cest_p04_10	13	No group
X8_cest_p05_05	15	No group
X8_cest_p05_11	15	No group
X8_cest_p05_14	15	No group
X8_cest_p05_04	16	No group
X8_cest_p05_07	16	Language self-perception
X8_cest_p05_08	16	Language self-perception
X8_cest_p07_01	19	No group

X8_cest_p07_02	19	No group
X8_cest_p07_05	19	No group
X8_cest_p10_04	20	No group
X8_cest_p10_06	20	No group
X8_cest_p10_07	20	No group
X8_cest_p05_13	21	Enjoyment of reading
X8_cest_p07_03	21	Enjoyment of reading
X8_cest_p07_04	21	Enjoyment of reading
X8_cest_p08	21	Enjoyment of reading
X8_cest_p09_02	21	Enjoyment of reading
X8_cest_p09_05	21	Enjoyment of reading
X8_cest_p09_07	21	Enjoyment of reading
X8_cest_p09_03	22	Enjoyment of reading
X8_cest_p09_10	22	Enjoyment of reading
X8_cest_p09_01	24	Enjoyment of reading
X8_cest_p09_06	24	Enjoyment of reading
X8_cest_p09_08	24	Enjoyment of reading
X8_cest_p09_09	24	Enjoyment of reading
X8_cest_p09_04	25	No group
X8_cest_p33_02	26	No group
X8_cest_p34_02	26	No group
X8_cest_p33_01	27	No group
X8_cest_p36_04	27	Extracurricular participation
X8_cest_p36_03	28	Extracurricular participation
X8_cest_p36_05	28	Extracurricular participation
X8_cest_p36_01	29	No group
X8_cest_p36_02	29	No group
X8_cest_p37_04	29	Extracurricular organization
X8_cest_p37_05	29	Extracurricular organization
X8_cest_p37_01	30	Extracurricular organization
X8_cest_p37_02	30	Extracurricular organization
X8_cest_p37_03	31	Extracurricular organization + Sports self-perception
X8_cest_p15_05	32	No group
X8_cest_p22_01	33	School climate - between students
X8_cest_p22_03	33	School climate - between students
X8_cest_p22_05	33	School climate - between students
X8_cest_p25_01	34	Bullying victim
X8_cest_p25_02	34	Bullying victim
X8_cest_p25_04	34	Bullying victim
X8_cest_p23_01	35	School climate -between students and teachers

X8_cest_p23_02	35	School climate -between students and teachers
X8_cest_p23_04	35	School climate -between students and teachers
X8_cest_p22_06	36	School climate - weapons
X8_cest_p22_07	36	School climate - weapons
X8_cest_p30_03	37	School climate - drugs
X8_cest_p30_04	37	School climate - drugs
X8_cest_p18_12	38	Discrimination
X8_cest_p18_13	38	Discrimination
X8_cest_p18_01	39	No group
X8_cest_p32_01	42	Fast food consumption
X8_cest_p32_02	42	Fast food consumption
X8_cest_p14_01	43	No group
X8_cest_p16_01	43	No group
X8_cest_p17_01	43	No group
X8_cest_p19_03	43	No group
X8_cest_p42_01	44	Evaluation of the school
X8_cest_p42_04	44	No group
X8_cest_p28_02	45	No group
X8_cest_p40_05	46	Evaluation of the school
X8_cest_p40_06	46	Evaluation of the school
X8_cest_p40_04	47	Evaluation of the school
X8_cest_p40_01	48	Evaluation of the school
X8_cest_p40_02	48	Evaluation of the school
X8_cest_p26_01	49	Teachers resolute conflicts
X8_cest_p26_02	49	Teachers resolute conflicts
X8_cest_p26_03	49	Teachers resolute conflicts
X8_cest_p05_03	50	Sport self-perception
X8_cest_p06	65	No group

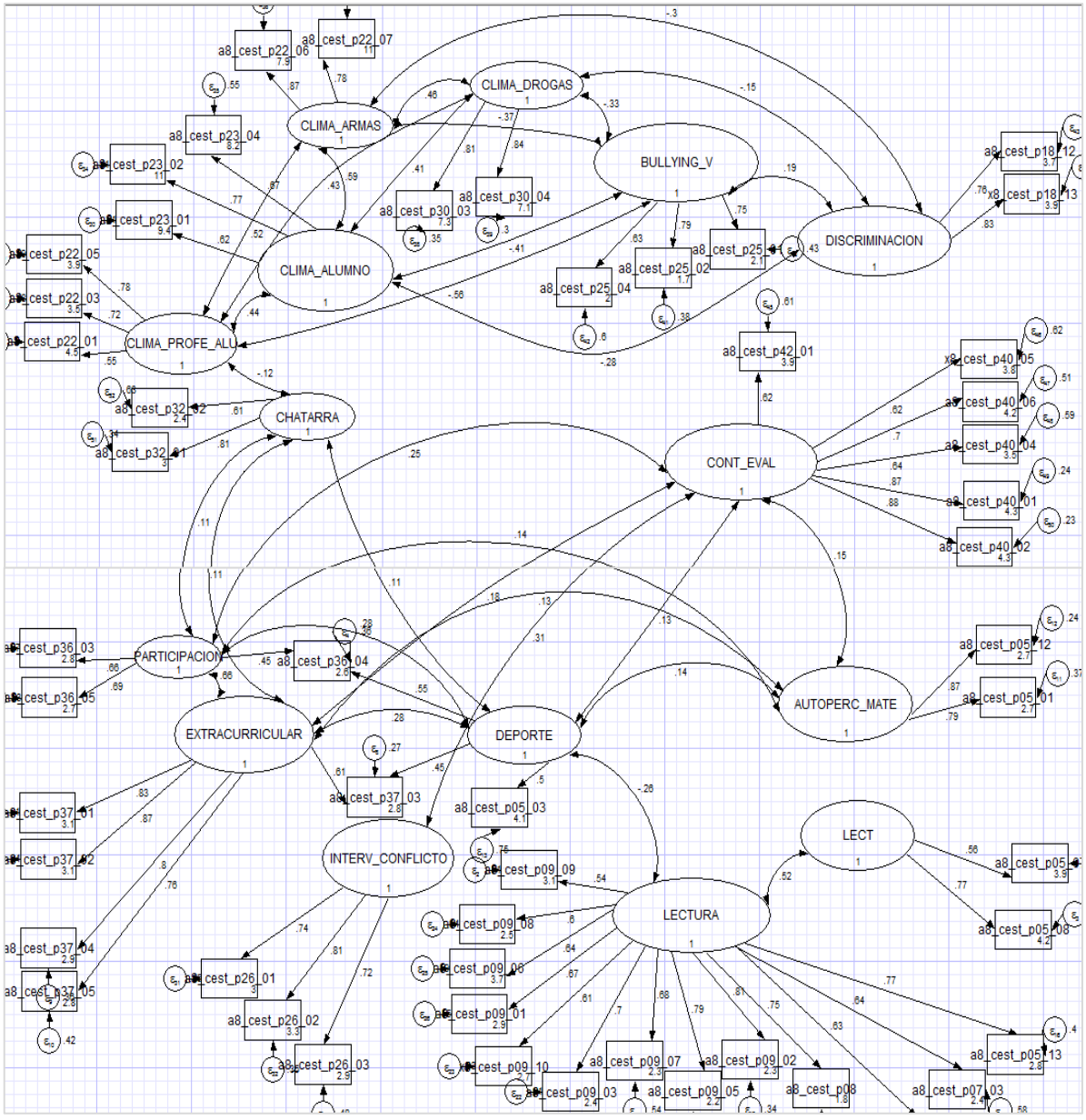


Figure H-5: Structural equation model for variables from eighth grade student's questionnaire.

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(1241)	247466.038	model vs. saturated
p > chi2	0.000	
chi2_bs(1326)	3.272e+06	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.036	Root mean squared error of approximation
90% CI, lower bound	0.000	
upper bound	.	
pclose	.	Probability RMSEA <= 0.05
Information criteria		
AIC	1.626e+07	Akaike's information criterion
BIC	1.626e+07	Bayesian information criterion
Baseline comparison		
CFI	0.925	Comparative fit index
TLI	0.920	Tucker-Lewis index
Size of residuals		
SRMR	0.062	Standardized root mean squared residual
CD	1.000	Coefficient of determination

Figure H-6: Level of fit of the structural equation model detailed in Figure H-5.

Table H-4: Variables from the sixth grade teacher's questionnaire.

Key	Dendogram group	Index
X6_cprof_p03_08	77	No group
X6_cprof_p12_02	79	No group
X6_cprof_p12_05	79	No group
X6_cprof_p12_06	79	No group
X6_cprof_p06	81	No group
X6_cprof_p14_07	84	School climate
X6_cprof_p14_08	84	School climate
X6_cprof_p07_03	85	School climate

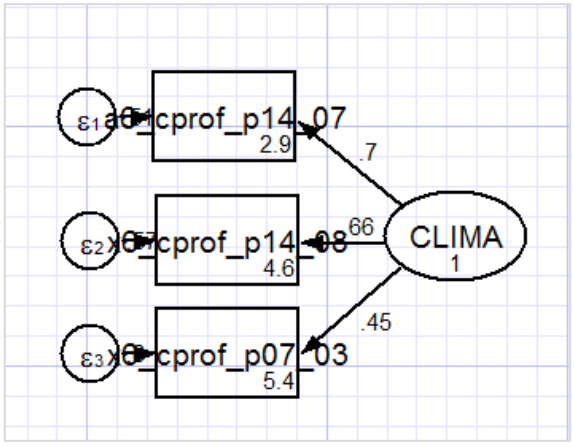


Figure H-7: Structural equation model for variables from sixth grade teacher's questionnaire.

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(0)	0.000	model vs. saturated
p > chi2	.	
chi2_bs(3)	53228.715	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.000	Root mean squared error of approximation
90% CI, lower bound	0.000	
upper bound	0.000	
pclose	1.000	Probability RMSEA <= 0.05
Information criteria		
AIC	904318.861	Akaike's information criterion
BIC	904407.662	Bayesian information criterion
Baseline comparison		
CFI	1.000	Comparative fit index
TLI	1.000	Tucker-Lewis index
Size of residuals		
SRMR	0.000	Standardized root mean squared residual
CD	0.663	Coefficient of determination

Figure H-8: Level of fit of the structural equation model detailed in Figure H-7.

Table H-5: Variables from the eighth grade teacher's questionnaire.

Key	Dendogram group	Index
X8_cprof_p04_11	72	No group
X8_cprof_p23	72	No group
X8_cprof_p21_01	73	Advanced teaching practices
X8_cprof_p21_07	73	Advanced teaching practices
X8_cprof_p21_03	74	Basic teaching practices
X8_cprof_p22_01	74	Basic teaching practices
X8_cprof_p07	81	No group
X8_cprof_p08_01	86	School climate
X8_cprof_p09_01	86	School climate
X8_cprof_p08_03	87	School climate
X8_cprof_p09_02	87	School climate

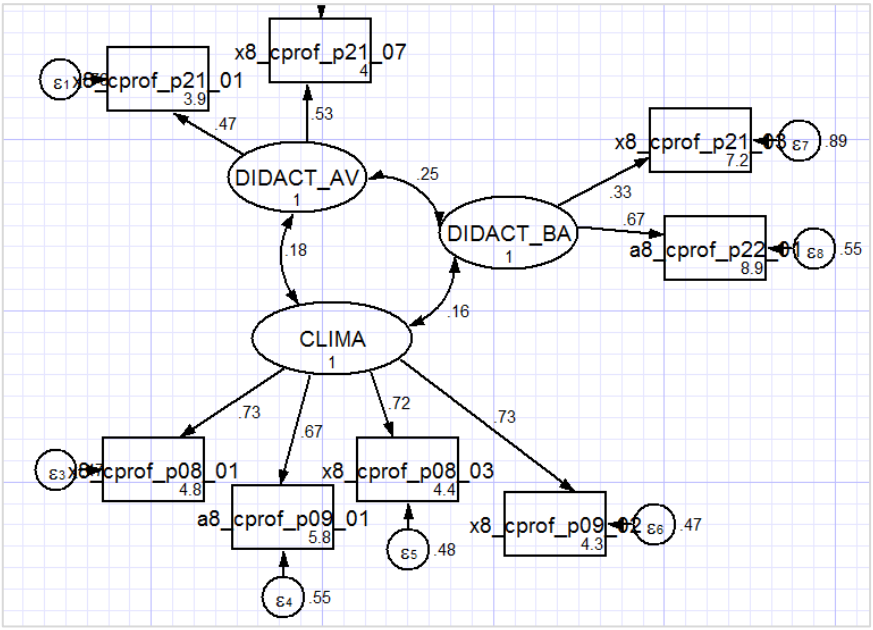


Figure H-9: Structural equation model for variables from eighth grade teacher's questionnaire.

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(17)	6283.039	model vs. saturated
p > chi2	0.000	
chi2_bs(28)	208543.616	baseline vs. saturated
p > chi2	0.000	
Population error		
RMSEA	0.050	Root mean squared error of approximation
90% CI, lower bound	0.000	
upper bound	.	
pclose	.	Probability RMSEA <= 0.05
Information criteria		
AIC	2.325e+06	Akaike's information criterion
BIC	2.325e+06	Bayesian information criterion
Baseline comparison		
CFI	0.970	Comparative fit index
TLI	0.951	Tucker-Lewis index
Size of residuals		
SRMR	0.019	Standardized root mean squared residual
CD	0.939	Coefficient of determination

Figure H-10: Level of fit of the structural equation model detailed in Figure H-9.

Appendix I: Decision Trees and student's profiles

In this appendix we describe how we used Decision Trees to select variables and built student's academic achievement profiles. Throughout our study, our indicator of academic achievement is the achievement level in the language standardized test.

Figure I-1 shows the Decision Tree we obtained when we ran the algorithm with 2 nodes of depth, this means that the algorithm builds a tree with 2 levels. The Figure shows that the two most important variables to separate achievement levels in eighth grade are "eda_lect6" and "eda_lect4", which are the keys of the language achievement level in fourth and sixth grade respectively. Number 3 corresponds to the higher achievement level and number 1 to the lower achievement level. Hence, if a student has the lowest level in fourth and sixth grade it will end up in "Node 3" (the bar graph in the left in Figure I-1). We used this information to build the academic achievement profiles.

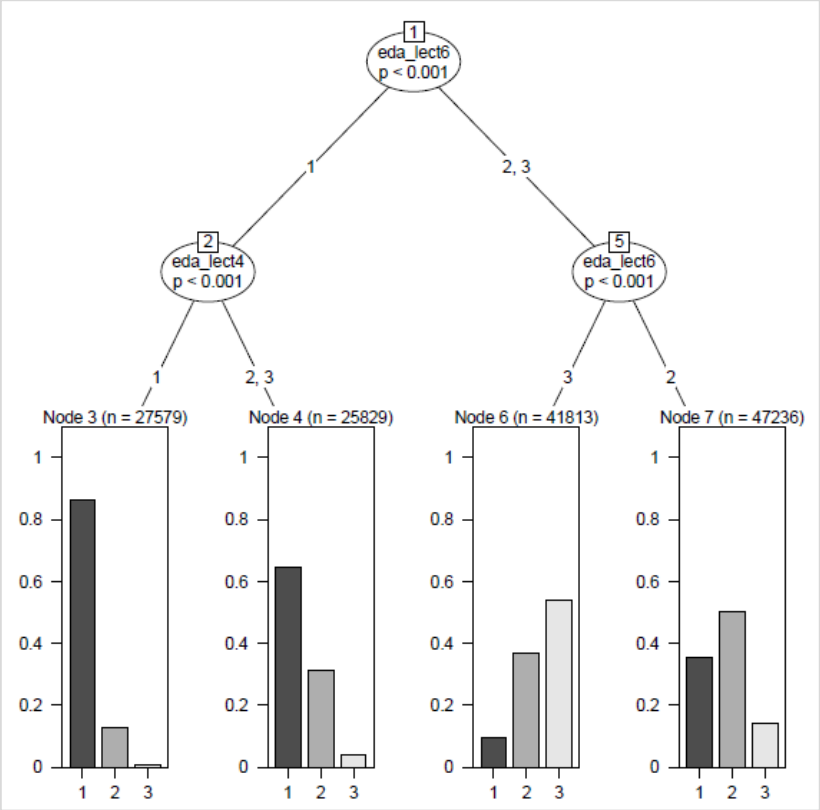


Figure I-1: Decision Tree with two nodes of depth.

Then, we excluded prior academic achievement variables and built a second Decision Tree. Figure I-2 is the Decision Tree obtained. In this case, if a student has a high level of enjoyment of reading in eighth grade ("z8_gusto_lect" higher than 0.672) and a high level of self-perception in fourth grade ("z4_autop_gral" higher than 0.202), it will belong to "Node 7" (the bar graph in the right in Figure I-2). Most students in "Node 7" have a high achievement level in eighth grade (the lightest bar in the bar graph in the right in Figure I-2).

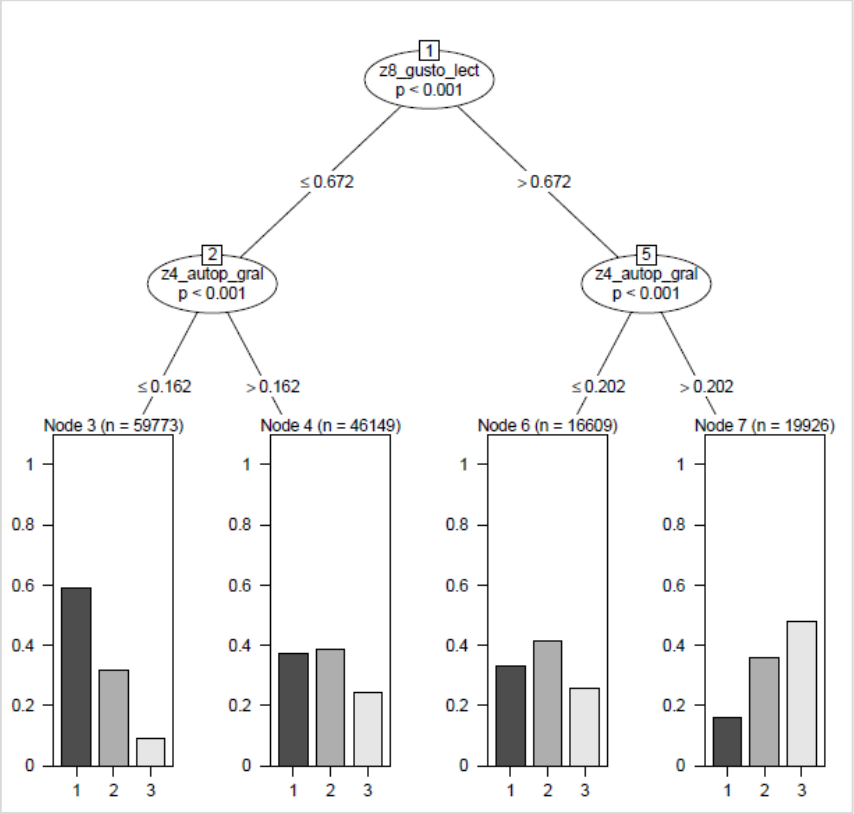


Figure I-2: Decision Tree excluding prior academic achievement variables with two nodes of depth.

Finally, we built a third Decision Tree, excluding prior academic achievement variables but with 3 nodes of depth. Figure I-3 shows the tree.

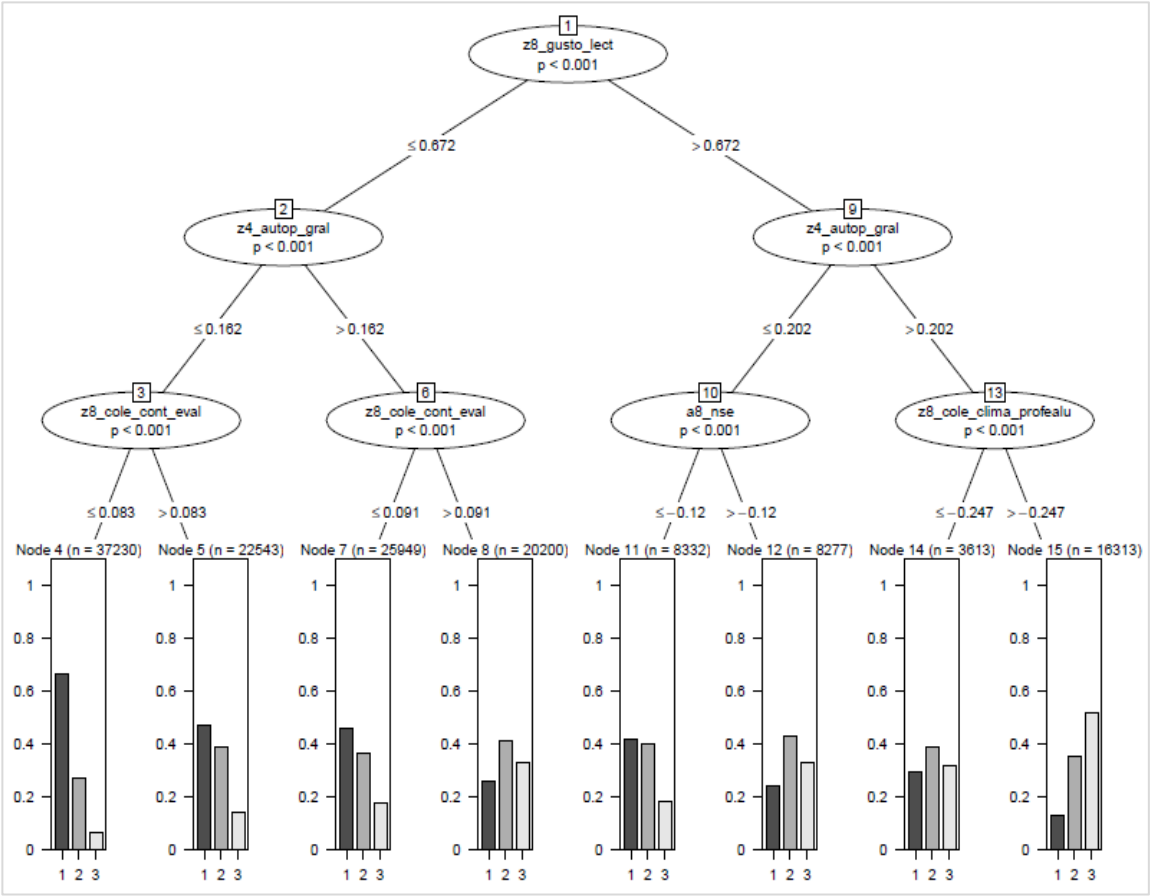


Figure I-3: Decision Tree excluding prior academic achievement variables with three nodes of depth.

We decided to build the second academic achievement profile with three variables: Enjoyment of reading in eighth grade ("z8_gusto_lect"), self-perception in fourth grade ("z4_autop_gral") and student's evaluation of school in eighth grade ("z8_cole_cont_eval").

Because enjoyment of reading in eighth grade and self-perception in fourth grade are in the higher levels, these variables are more relevant to separate academic achievement than

evaluation of school in eighth grade, socioeconomic status of the school in eighth grade ("a8_nse") and school climate in eighth grade, specifically how teachers and students treat each other ("z8_cole_clima_profealu"). From these three latter variables we chose student's evaluation of school because it appeared in two of four nodes of the third level and the bar graphs showed that it separates better achievement level in eighth grade. In fact, when we excluded the socioeconomic status and school climate variable, student's evaluation of school variable replaced the school climate variable, as shown in Figure I-4.

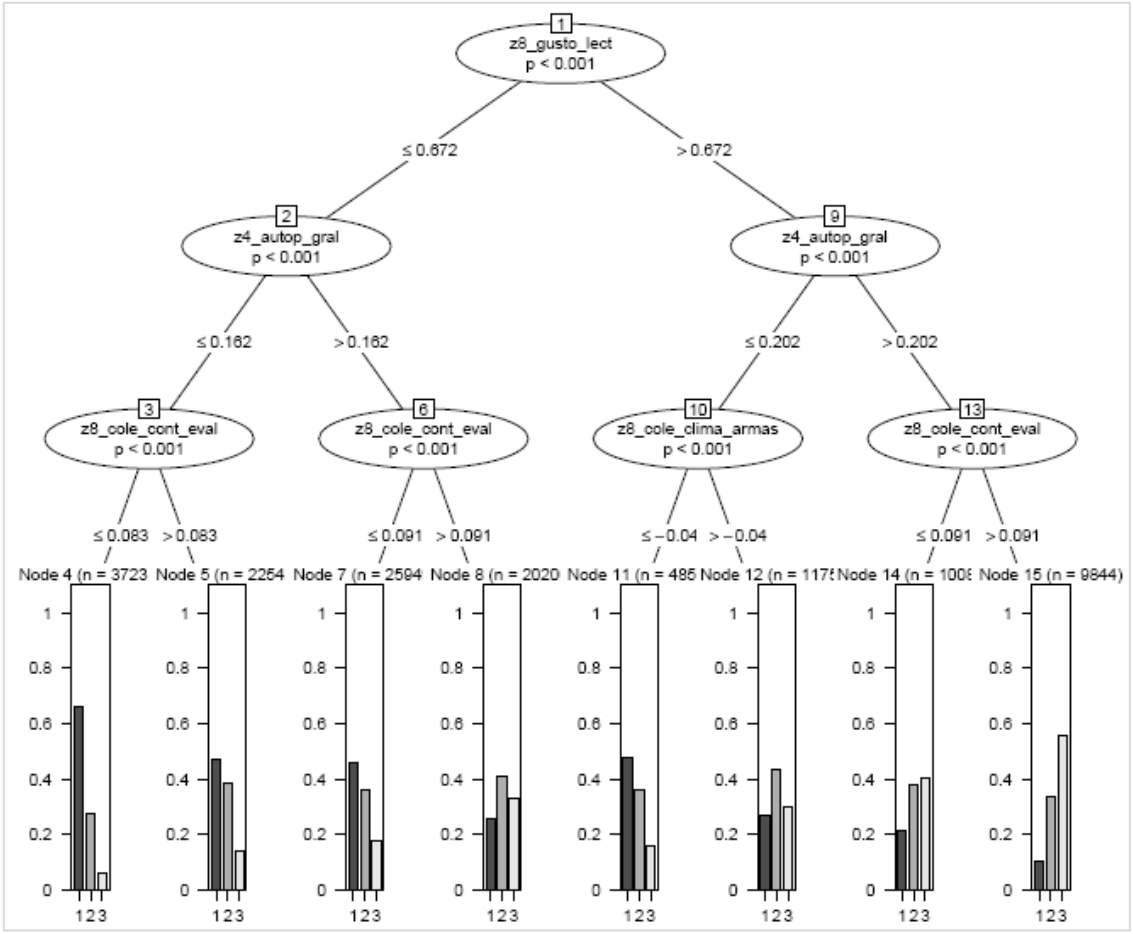


Figure I-4: Decision Tree excluding prior academic achievement, socioeconomic status in eighth grade and school climate in eighth grade.

In the Decision Tree, enjoyment of reading appears only once with 0,672 as the threshold to separate students. Self-perception appears two times with two threshold values: 0.162 and 0.202. Finally, student's evaluation of school appear three times with two threshold values: 0.083 and 0.091.

Tables I-1 to I-3 show the distribution of students per profile for three different combinations of thresholds. The distribution of students almost does not vary, hence, it is irrelevant what combination of thresholds we chose. The table presented in section 5. Results, is table I-2 because it matches the most thresholds of the Decision Tree.

Table I-1: Student's profile with 0.182 as the threshold for self-perception and 0.087 for student's evaluation of school.

	Achievement level in 8th Number of students				Achievement level in 8th Percentage of students			
Student's achievement profiles	Low	Mid	High	Total	Low	Mid	High	Total
Probably high achievement (High: Enjoyment of reading, school evaluation and self-perception)	1,088	3,362	5,583	10,033	11%	34%	56%	100%
Probably mid achievement (The rest)	35,067	36,860	22,548	94,475	37%	39%	24%	100%
Probably low achievement (Low: Enjoyment of reading, school evaluation and self-perception)	25,065	10,452	2,432	37,949	66%	28%	6%	100%

Table I-2: Student's profile with 0.202 as the threshold for self-perception and 0.091 for student's evaluation of school.

Student's achievement profiles	Achievement level in 8th Number of students				Achievement level in 8th Percentage of students			
	Low	Mid	High	Total	Low	Mid	High	Total
Probably high achievement (High: Enjoyment of reading, school evaluation and self-perception)	1,047	3,290	5,502	9,839	11%	33%	56%	100%
Probably mid achievement (The rest)	34,636	36,675	22,542	93,853	37%	39%	24%	100%
Probably low achievement (Low: Enjoyment of reading, school evaluation and self-perception)	25,537	10,709	2,519	38,765	66%	28%	6%	100%

Table I-3: Student's profile with 0.162 as the threshold for self-perception and 0.083 for student's evaluation of school.

Student's achievement profiles	Achievement level in 8th Number of students				Achievement level in 8th Percentage of students			
	Low	Mid	High	Total	Low	Mid	High	Total
Probably high achievement (High: Enjoyment of reading, school evaluation and self-perception)	1,127	3,445	5,680	10,252	11%	34%	55%	100%
Probably mid achievement (The rest)	35,416	37,069	22,528	95,013	37%	39%	24%	100%
Probably low achievement (Low: Enjoyment of reading, school evaluation and self-perception)	24,677	10,160	2,355	37,192	66%	27%	6%	100%*

* The percentages were approximated, that is why the direct sum does not add up 100%.

Appendix J: Journal of Economic Perspectives reception letter.

From: "Ann Norman" <anorman@aeapubs.org>
Date: Dec 5, 2017 12:07
Subject: Re: JEP - Research proposa
To: "Veronica Puga" <veropugaduran@gmail.com>
Cc:

Dear Professor,

I have put your paper on the agenda for discussion at the next editors' meeting. Unfortunately, the editors just met and I don't know if they will be meeting again before Christmas, after which we are all busy until the first week of January. So due to the Holidays, it may take longer than usual for a decision. If you hear nothing from the editors within 6 weeks, please get back to me for an update on the status of your proposal.

All the best to you and Happy Holidays!

Ann

Ann Norman
Assistant Editor, JEP
anorman@jepjournal.org

Appendix K: Paper proposal sent to Journal of Economic Perspectives

Combining data mining and econometric techniques: A case study on academic achievement

Abstract: Data mining is changing econometric research. Although collaboration is expected between these two disciplines, to our best knowledge, there are not published attempts that show how data mining tools can complement econometric ones. This research proposes the Econometrics and Data Mining Dialogue approach, where an econometric model is built just from the data through data mining, without selecting variables based on bibliographic research or expert opinion. The approach was applied to a case study, predicting academic achievement in a longitudinal database. In total, we analyzed 142,457 students with 1,287 independent variables. We employed Random Forest, a data mining algorithm, to select a subset of variables to, posteriorly build an econometric model, an ordinal logistic multilevel model. Finally, we used Decision Trees, a data mining algorithm, to define a student's achievement profile. Most findings of our case study are consistent with academic achievement literature, like the relevance of prior academic achievement to present academic achievement. Other results offer fresher insights, like the impact of student's evaluation of their school on their academic achievement. This paper aims to initiate a hands-on dialogue of how computer scientists and econometricians can collaborate to deepen the knowledge databases can offer and to improve econometric models.

I. Context and thesis statement

Prediction is an old problem with a long history in econometric research (Mullainathan & Spiess, 2017). Econometric models tend to be theory-driven; the selection of variables is based in theory built on previous research, expert opinion or the proposition of new hypothesis; in any of these cases, the model is built in a subjective, slow and expensive way (Fayyad et al., 1996). New disciplines such as big data and data mining are gaining attention (Sagiroglu & Sinanc, 2013; Hand et al., 2001) and will probably change economic research (Einav & Levin, 2014). New methods will not replace economic theory, but will complement them and enable new research designs (Einav & Levin, 2014). These data-driven methods look for new findings instead of testing hypotheses (Slater, et al., 2017).

These theory- and data-driven modes of analysis have always coexisted and do not need to be in conflict (Mullainathan & Spiess, 2017). In fact, collaboration between computer scientists and econometricians is expected to be productive in the future (Varian, 2014), but, to our best knowledge, no work attempts to show how these disciplines can

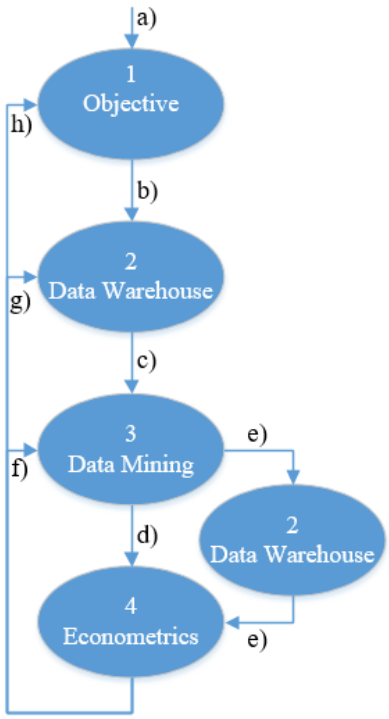
complement each other. The thesis of our paper is that an econometric model can be built just from data through data mining, without selecting variables based on bibliographic research or expert opinion.

Section II presents our proposed approach, Econometrics and Data Mining Dialogue. Section III describes the application of our approach to our case study, predicting academic achievement in a longitudinal database using student, classroom and school characteristics. Section IV describes the results obtained and section V presents the conclusions.

II. Econometrics and Data Mining Dialogue: EDMD.

Our approach combines the ability of data mining to analyze big data sets with the consistency of econometric models to estimate relations between variables. As shown in Figure 1, it consists of four steps: 1) objective, 2) data warehouse, 3) data mining algorithm and 4) econometric model.

Figure 1: Diagram of our proposed approach.



1) Objective: A study employing the proposed approach should have access to an ample data set and aim to find robust relations between variables.

2) Data warehouse: The objective is to adjust the database to the data mining algorithm. Statistical analysis and pre-processing allow to gain insights of the data to prepare it for further analyses (Han et al., 2011; Witten & Frank, 2005).

3) Data mining: We propose three main objectives. To select or group variables to employ on an econometric model (Variable, in Table 1), to label samples to assess group characteristics in an econometric model (Group, in Table 1), and to identify similar classes to warn econometric models (Proximity, in Table 1). Table 1 offers examples of data mining work in the educational domain and associates them with an of objective. We propose how each work could have nurtured an econometric model, if collaboration had existed.

Table 1: Type of goal matched with an example of data mining work in education.

Goal	Authors	Data mining approach	Alternative econometric collaboration
Variable	Martínez & Chaparro, 2017	Use decision trees to detect student and school factors related to academic achievement.	The factors identified could have been used to build an econometric model.
Group	Bresfelean et al., 2008	Build a student's exams failure profile.	A variable could be created to acknowledge different profiles. It could be employed in an econometric model.
Proximity	Asif et al., 2017	Found that it is difficult for classification methods to predict some classes.	Minority classes could be excluded from the sample or new thresholds can be proposed.

4) Econometric model: Considering that, economic applications produce good estimations of relations between variables (Mullainathan & Spiess, 2017), the objective of this step is to estimate relationships between independent and dependent variables.

The presented approach is an iterative and sequential process. In Figure 1, each connection is marked with letters. Connection a), b) c) and d) show the path for a sequential implementation of the previously defined four steps. Connection e) represents the path when data needs a second processing stage before building the econometric model. Connection f), g) and h) close the cycle to allow the process to be iterative.

III. Application to our case study

In this section, we describe how we applied our approach to our case study. We used data from a Chilean standardized test, SIMCE. The evaluation includes tests in different subjects and questionnaires to students, teachers and parents. These questionnaires seek to identify and validate factors that influence academic and non-academic results (Agencia de Calidad de la Educación, 2014). We selected a cohort of students that had been assessed three times with standardized tests, i.e., in fourth, sixth and eighth grade, during 2011, 2013 and 2015 respectively. The independent variables include school context and the questionnaires, in total; each student was associated to 1,287 independent variables

Figure 2: Application of our proposed approach.

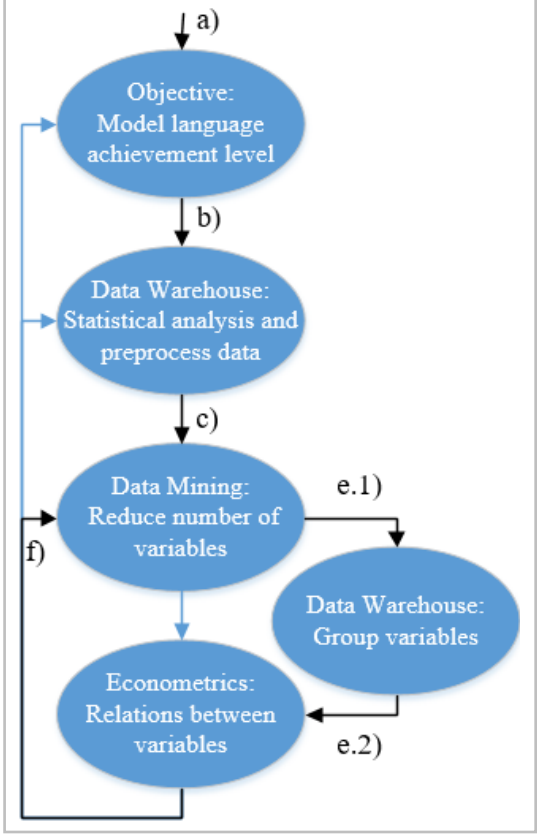


Figure 2 illustrates the application of our proposed approach to our case study. The objective was to model the achievement level in the national language test ('a' in Figure 2). Then, we analyzed statistically the variables and created a socioeconomic status variable ('b' in Figure 2). We used R's Random Forest (Liaw & Wiener, 2015) to compare the 1,287 independent variables and selected a subset of them ('c' in Figure 2). Then, we imputed missing responses with the R's "missForest" package (Stekhoven, 2013) and grouped variables that were about similar topics into indexes. To create indexes, first we employed *hclust()* from R (Müllner, 2017). We used these results as a starting point to build Structural Equation Models in STATA ('e.1' in Figure 2). Then, we build an ordinal logistic multilevel model ('e.2' in Figure 2). Finally, we used R's Decision Trees (Therneau, et al., 2017) and identified variables to define a student's achievement profile ('f' in Figure 2).

IV. Results

Regarding the data mining algorithm, Random Forest, we found that socioeconomic status, self-perception and enjoyment of reading are important for all students. However, school climate concepts are more important to predict achievement of students that start with a low or mid achievement level.

Regarding the econometric model, the ordinal logistic multilevel model, we built an empty model to check that the proposed structure, ordinal logistic multilevel, was more adequate than the simple correspondent structure, ordinal logistic. We also assessed the Intraclass Correlation Coefficient, an indicator of correlation between samples that belong to the same group, in our case schools. Our selected model explains 53.6% of the variance and has 25 independent variables. The variables with higher significance and impact on the dependent variable are prior academic achievement and student's evaluation of their school.

Regarding the data mining algorithm, Decision Trees, it selects prior academic achievement as the first variables to separate between students. We built another Decision Tree excluding prior academic achievement variables and the variables selected are general self-perception, enjoyment of reading and school's evaluation, all indexes built from the student's questionnaires. Table 2 presents these two student's achievement profile.

Table 2: Student's achievement profiles.

	Achievement level in 8th Number of students				Achievement level in 8th Percentage of students			
Student's achievement profile 1	Low	Mid	High	Total	Low	Mid	High	Total
1.- Probably high achievement (High achievement level in 4th and 6th)	2,646	11,832	20,358	34,836	8%	34%	58%	100%
2.- Probably mid achievement (The rest)	34,742	35,292	10,008	80,042	43%	44%	13%	100%
3- Probably low achievement (Low achievement level in 4th and 6th)	23,832	3,550	197	27,579	86%	13%	1%	100%
Student's achievement profile 2	Low	Mid	High	Total	Low	Mid	High	Total
1.- Probably high achievement (High enjoyment of reading, school evaluation and self-perception)	1,088	3,362	5,583	10,033	11%	34%	56%	100%
2.- Probably mid achievement (The rest)	35,067	36,860	22,548	94,475	37%	39%	24%	100%
3.- Probably low achievement (Low enjoyment of reading, school evaluation and self-perception)	25,065	10,452	2,432	37,949	66%	28%	6%	100%

V. Conclusion

This study's contribution is our proposed approach, which shows how an econometric model can be built just from data through data mining, evidencing how these two disciplines complement each other. Our approach can be applied to any domain that has big datasets and econometric models that reflect the data structure, e.g. healthcare, education, transportation.

Future studies can provide more examples that show how to combine data mining and econometrics and develop new tools that facilitate cooperation between these two disciplines. These can develop an applied dialogue of how computer scientists and econometricians can work together to extract novel knowledge from data and to improve econometric models.

References

Agencia de Calidad de la Educación (2014). Informe Técnico SIMCE 2014.

Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*.

Bresfelean, V. P., Bresfelean, M., Ghisoiu, N., & Comes, C. A. (2008, June). Determining students' academic failure profile founded on data mining methods. In *Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on* (pp. 317-322). IEEE.

Einav, L., & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, *14*(1), 1-24.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT press.

Liaw, A., & Wiener, M. (2015). Breiman and Cutler's random forests for classification and regression, R package version 4.6-12.

Martínez Abad, F., & Chaparro Caso López, A. A. (2017). Data-mining techniques in detecting factors linked to academic achievement. *School Effectiveness and School Improvement*, *28*(1), 39-55.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

Müllner, D. (2017). Fast Hierarchical Clustering Routines for R and Python. R package version. 1.1.24

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *Collaboration Technologies and Systems (CTS), 2013 International Conference on* (pp. 42-47). IEEE.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*, *42*(1), 85-106.

Stekhoven, D. (2013) MissForest: nonparametric missing value imputation using random forest. R package version. 1.4

Therneau, T., Atkinson, B., Ripley, B., & Ripley, M. B. (2017). Recursive Partitioning and Regression Trees, R package version 4.1-11.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3-27.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Authors

Verónica Puga-Durán, Computer Science Department, School of Engineering, Pontificia Universidad Catolica de Chile.
Miguel Nussbaum, Computer Science Department, School of Engineering, Pontificia Universidad Catolica de Chile.
Ernesto Treviño, Faculty of Education, Pontificia Universidad Catolica de Chile.
Karim Pichara, Computer Science Department, School of Engineering, Pontificia Universidad Catolica de Chile