


# A Descriptive and Predictive Model for Data-Driven Decision Making in Higher Education: A Case Study


**Claudio Gutiérrez-Soto**

(Universidad del Bío-Bío, Concepción, Chile)

 <https://orcid.org/0000-0002-7704-6141>, [cogutier@ubiobio.cl](mailto:cogutier@ubiobio.cl))


**Marco A. Palomino\***

(University of Aberdeen, Aberdeen, UK)

 <https://orcid.org/0000-0001-7850-416X>, [marco.palomino@abdn.ac.uk](mailto:marco.palomino@abdn.ac.uk))


**Patricio Galdames**

(Universidad San Sebastián, Concepción, Chile)

 <https://orcid.org/0000-0003-3051-2413>, [patricio.galdames@uss.cl](mailto:patricio.galdames@uss.cl))

**Cristian Duran-Faundez**

(Universidad del Bío-Bío, Concepción, Chile)

 <https://orcid.org/0000-0002-2793-5880>, [crduran@ubiobio.cl](mailto:crduran@ubiobio.cl))

**Abstract:** The growth of online learning in higher education, particularly after the COVID-19 pandemic, has fostered the advancement of learning analytics, which nowadays relies greatly on capturing and mining data derived from systems such as Blackboard and Moodle. However, it remains difficult to identify all the variables having a direct bearing on academic success, and drawing advice from machine learning models trained to support data-driven decision making is challenging. Therefore, we have endeavoured to pair a descriptive model, which characterises the profiles of computer science students, with a predictive model, which relies on Bayesian networks to forecast academic success. To achieve this, we have looked for the factors directly influencing the academic performance of computing science students, and the common patterns of behaviour which characterise higher education students individually and as part of a cohort. Our approach has been tested with data provided by a Chilean institution—University of Bío-Bío. We have enhanced and supplemented the data employed in our investigation by means of two surveys distributed among all the different cohorts of the student population. Our predictive model can determine student outcomes with an accuracy rate above 97%.

**Keywords:** Learning Analytics; Bayes Nets; Educational Data Mining; Higher Education.

**Categories:** L.3.6, L.2.5, L.3.3, L.3.0

**DOI:** 10.3897/jucs.154610

## 1 Introduction

Most organisations are investing in data to support decision making. This applies not only to business organisations, but also to the education sector, which has grown steadily

---

\* Correspondence: [marco.palomino@abdn.ac.uk](mailto:marco.palomino@abdn.ac.uk)

since the end of the COVID-19 pandemic. Indeed, the total revenue in education worldwide is expected to have an annual growth rate of 9.11% from 2024, resulting in a projected market volume of USD \$11.83 billion by 2029 [Statista, 2024b].

The growing trend in education is perceived at a faster pace in some countries—for example, the total revenue in education in the UK is estimated to grow more than 12% annually for the rest of the decade [Statista, 2024a]—but it is observed globally. We are particularly familiar with the case of Chile, where the spending on education is forecast to reach USD \$11.1 billion by 2029, which will be a new historic peak [Degenhard, J, 2024]. Although the work that we introduce here has been done with data from a Chilean institution, it is applicable to any educational establishment.

Whilst the investment in education continues to grow, *learning analytics* continues to thrive, as more data related to learning and the environments in which learning occurs become available. There are many definitions of learning analytics, but they all agree that it comprises gathering, processing and interpreting data, and communicating results, including recommended decisions and actions related to the efficiency and effectiveness of the programmes designed to improve individual and institutional performance [Hall et al., 2020]. A substantial body of literature on this subject has concentrated on offering guidelines for the application of big data analytics in education [Viberg et al., 2018, Gladshiya and Sharmila, 2019, Franco Caballero et al., 2020], and several techniques involving machine learning, web analytics, and information visualisation have been applied to understand and optimise learning [Ranjeeth et al., 2020].

Although the primary focus of learning analytics is to determine if a learning experience is successful or not, there is still significant uncertainty about the depth of the students' knowledge and the factors that directly affect their academic performance [Guzmán-Valenzuela et al., 2021]. Moreover, pairing descriptive and predictive models that represent students' profiles for specific cohorts remains difficult. Drawing advice from such models to improve students' learning is either too challenging or unpractical, and this is where our main contribution lies.

We present here the results of an investigation carried out at the *University of Bío Bío* in 2023. The University of Bío Bío is a public institution located in the city of Concepción in Chile. As a country, Chile offers an abundant ground for research into higher education. It has been estimated that the dropout rate for students enrolled on Chilean higher education institutions is significant [Espinoza et al., 2024]—by the year 2022, female students had a dropout rate of 16.7%, while male students had a dropout rate of 20.5%, and only half of the student population attained their degrees within the expected time. This poses a rather challenging setting which demands improvement. In comparison, the dropout rate for UK universities is much lower: Northern Ireland's dropout rate is only 2.4%, while England's is 2.7% [Jack, 2023].

Chile's dropout rate is especially high for students whose socioeconomic status is low, and this is also present in technical training centres and professional institutes [Valenzuela and Kuzmanic, 2023]. Consequently, there is an appetite for our research, and we expect it to strengthen early warning systems which promptly identify students who require support, so that interventions can take place before it is too late.

To get to know the students involved in our study, and fill in the gaps identified by our analysis, we carried out two surveys. Such surveys were distributed among students from two different computing programmes—namely, *Ingeniería Civil en Informática* (Civil Engineering in Informatics) and *Ingeniería de Ejecución en Computación e Informática* (Computing and Informatics Engineering)—and all their levels of study—from the first to the fifth year. The first year of both programmes is largely dedicated to reinforcing the students' mathematical skills and providing introductory courses on engi-

neering and programming. Enabling the efficient organisation and manipulation of data begins in the second year, when the students attend the course on data structures. During the third year, databases and algorithms are covered.

Across the first three years of both programmes, students are expected to take English language courses on top of their other courses to improve their communication skills and expand their career opportunities. Software engineering and artificial intelligence (AI)—two common occurrences in modern curricula—are attended in the fourth year, in preparation for the research project, which is carried out in the final year and leads to the dissertation. The fifth year is also when courses that have a direct link with the industry, and address real engineering problems, are held. Both programmes included in our study have been certified by the Chilean *National Accreditation Commission* (<https://www.cnachile.cl/>), which monitors the quality of all Chilean higher education institutions and the programmes that they deliver.

The research questions that we shall address in this contribution are as follows:

- RQ1 (Academic Performance Factors):** What are the factors directly influencing the academic performance of computing science students? We aim to determine a set of normalised variables which lead to the success, or failure, of the students. Even though we study the specific case of computing science, the methodology and insights derived from our work should be applicable to different contexts and institutions, giving us the opportunity to generalise our research partially.
- RQ2 (Descriptive Model):** What are the common patterns of behaviour which characterise higher education students individually and as part of a cohort? We will be using techniques such as clustering to explore and describe these patterns.
- RQ3 (Predictive Model):** Can the likelihood of academic success, or failure, for a specific sample of higher education students be determined with a certain degree of confidence? Our objective is to create a predictive model based upon Bayesian networks. We appreciate that Bayesian networks are not the only alternative, but they are reasonably straightforward to apply as a starting point.

The remainder of this paper is organised as follows: Section 2 provides an overview of the state-of-the-art in learning analytics. Section 3 explains how we identified the variables which have a direct bearing on academic success. Section 4 presents the steps pursued to develop the descriptive and predictive models and what results we attained. We have also used this section to discuss our findings and how to exploit them. Finally, Section 5 states our conclusions and opportunities for future research.

## 2 Related Work

Although learning analytics is a relatively new area of research, it has already been the subject of several reviews. One of such reviews, which is fundamental to our work, is the one published by Peña-Ayala in 2018 [Peña-Ayala, 2018]. Such a review provided us with a structured perspective of the evolution of learning analytics, its status, and current trends. Peña-Ayala's taxonomy became indispensable to place our research within the context of existing work. However, we are not interested in delivering another review. Instead, we propose to develop a descriptive model to characterise higher education students and estimate their likelihood to succeed with the help of a predictive model.

We shall summarise in the next couple of pages the different developments undertaken within learning analytics. We have divided the content into subsections, which refer to the main areas of research in this field. It should be observed that more than a survey of current literature, we plan to describe how previous work relates to our own.

## 2.1 Dashboards

Jivet et al. [Jivet et al., 2018] surveyed the concept of *dashboards* in education. Dashboards are visual tools capable of aggregating different indicators about the learners, their processes, and perspectives into one or multiple displays [Jivet et al., 2018]. Dashboards help learners to reflect on their learning behaviour and progress towards the achievement of their learning outcomes.

All the students who participated in our study had access to a dashboard, which is maintained and hosted by the institution where we carried out our work. However, we did not limit our research to the specifics of dashboards. We are not concerned with graphical interfaces and visualisations, as our research is only starting. We opted to move a step further and suggest descriptive and predictive models derived from our analysis.

## 2.2 Learning Management Systems

After 2018, and largely due to the COVID-19 pandemic, most higher education institutions strengthen their online courses and introduced learning activities via institutional *learning management systems* [Turnbull et al., 2020], such as *Blackboard* and *Moodle* [Darko, 2021]. These systems constitute a fertile ground for learning analytics, because data can be readily mined from them. Hence, many institutions are adopting these systems to collect and analyse data, which, in turn, can equip students with advice to improve their learning [Leitner et al., 2017].

Our research is partly based on data derived from a corporate repository fed by a learning management system. However, it is important to listen to the student voice too to cover all perspectives. Therefore, we devised two surveys and distributed them among the student population. We collected the voluntary responses and use them to document the students' thoughts, especially with regards to academic success.

## 2.3 Theoretical Basis

Very few publications have approached learning analytics on a purely theoretical basis. However, this is of particular importance when recognising the challenges faced by higher education institutions. For instance, Daniel [Daniel, 2015] looked into the complexities of higher education and explored the potential of big data to tackle them. Although Daniel's conclusions can support the rationale behind learning analytics, we did not follow a similar approach, because we wanted to produce a practical implementation. As opposed to a theoretical investigation, we created a testable model which can be used to support the evaluation of individual students' performance, and such model can also be applied to evaluate a full cohort.

## 2.4 Higher Education

From an educational perspective, it is important to consider the large body of literature on factors that impact student failure and success. Both, failure and success at the higher

education level, have been widely scrutinised for a long time—well before learning analytics became known. For instance, Behr et al. [Behr et al., 2020] have pointed out that failure—and subsequent withdrawal—may be linked to the national education system, teaching quality, and the students themselves. Success, on the other hand, may have different meanings for different students, but generally varies depending on gender, age, and other factors [Alyahyan and Düşteğör, 2020].

We were keen on reading the literature on higher education as we had to ascertain how students interpret success. This is also the reason why previous work on learning analytics has delved into the literature. For example, García y García [García y García, 2021] found significant differences between men and women in their approach to academic studies. Men appear to give a much greater weight than women to “calm” to succeed, whereas women consider “effort” to be a more important cause for achievement than men. Women seem to attribute their success to a certain amount of “luck”. Indeed, according to García y García [García y García, 2021], the variables that predict good grades for male students are effort and good teachers, whereas for female students are a liking for teachers, luck, and attention.

Generally, there is no single reason that leads students to succeed academically, but rather a diverse and multicausal phenomenon. Each institution has specific conditions and peculiarities, and all these must be taken into account holistically [Gutiérrez et al., 2021]. Thus, we ensured that our study involved a survey to get to know the students and understand their views. In Section 3, we will detail what the students who participated in our study considered to be the reasons behind their academic success or failure, and how they feed into our models.

## 2.5 Machine Learning

A great deal of recent work on learning analytics focuses on the use of machine learning algorithms to train software to analyse data and derive insight into improvements to the learning experience [Hall et al., 2020]. As the training continues, the software moves forward and is likely to become more accurate and result in better insights. Leveraging on the state-of-the-art advances on machine learning, we have implemented this type of analytics. We have developed a descriptive model to characterise students’ academic performance, and a predictive model to forecast students’ results.

McKay et al. [McKay et al., 2012] published one of the earliest works on the use of machine learning in the field of learning analytics. After McKay et al.’s seminal contribution, many more researchers have attempted to discover hidden insights from data automatically. Previous research has employed both supervised and unsupervised machine learning algorithms, including clustering [Dawson et al., 2019], which we used too, neural networks [Rodríguez-Hernández et al., 2021], correlation [Romero and Ventura, 2010], regression [Yildirim and Gülbahar, 2022], and sequence and data mining [Matcha et al., 2019], which is our core contribution too.

According to Siemens and Baker [Baker and Siemens, 2014], understanding learners when interacting in a learning environment is fundamental to learning analytics, but our ultimate goal is to design interventions to adapt to new realities—such as, COVID-19. We can expect machine learning to be a part of every application in the future [Alzubi et al., 2018], and learning analytics will not be the exception, as it continues to assist decision makers to manage interventions in education.

## 2.6 Corporate Analytics

The latest literature recognises distinct types of learning analytics. One of those types focuses on evaluating whether the learning experiences are effective in the corporate sector [Hall et al., 2020]. There is evidence that the new corporate training is rapidly moving from strategies that offer broad and general content towards a curriculum that fulfils company-specific training needs [Ben Salamah et al., 2023].

The application of learning analytics to the corporate sector is beyond the scope of our work, though we recognise the importance of training employees to prepare them for their roles and responsibilities within organisations. Still, our methodology can certainly be reproduced at any higher education institution.

## 2.7 Criticism

As learning analytics have become common practice throughout all levels of education, concerns about their suitability and applicability have emerged [Selwyn, 2019]. Roberts et al. [Roberts et al., 2016] have pointed out that students are often unaware of how their behaviours and outcomes are used for research and evaluation, despite them being the main source of data for the purpose of learning analytics. To avoid this problem, we placed the students at the centre of our work, and we paid attention to their opinions. To this end, we conducted two surveys among the student population.

Other academics have claimed that learning analytics underplays the complexity of the teaching-learning processes [Guzmán-Valenzuela et al., 2021]. Institutions control the gathering and analysing of data, and students and teachers are relegated to a merely observational role [Archer and Prinsloo, 2020]. Hence, we have made all our development as transparent as possible to ensure that the entire academic community is aware of our approach and the way we apply it.

## 3 Materials and Methods

As mentioned above, we conducted two surveys. In addition to revealing information about the students' perspectives, the surveys helped us to complement the corporate repository. Subsequently, using all the data from the repository and the surveys, we identified twenty-three variables with a direct bearing on academic success. We then carried out a *principal component analysis* (PCA) to reduce the dimensionality of our dataset into a smaller and more manageable set of components [Greenacre et al., 2022].

Once eight principal components were determined, we proceeded to outline a descriptive model using *K-means* [Ikotun et al., 2023] and *expectation maximization* (EM) [Do and Batzoglou, 2008]. We detected three student clusters, which were classified according to their academic performance. Afterwards, we developed a predictive model using Bayesian networks—specifically, we applied the *K2* algorithm for Bayesian network structure learning [Behjati and Beigy, 2020]. Figure 1 depicts the step-by-step approach that we followed, starting with the gathering of data from the surveys and corporate repository, and concluding with the predictive model.

### 3.1 Surveys

To ensure that our survey had a solid foundation, we adopted and customised the questionnaire developed by Zollanvari et al. [Zollanvari et al., 2017] to enquire about the students' learning strategies. These questions covered self-regulatory learning behaviours—for example, time management, exam revision, study skills, and note-taking skills.

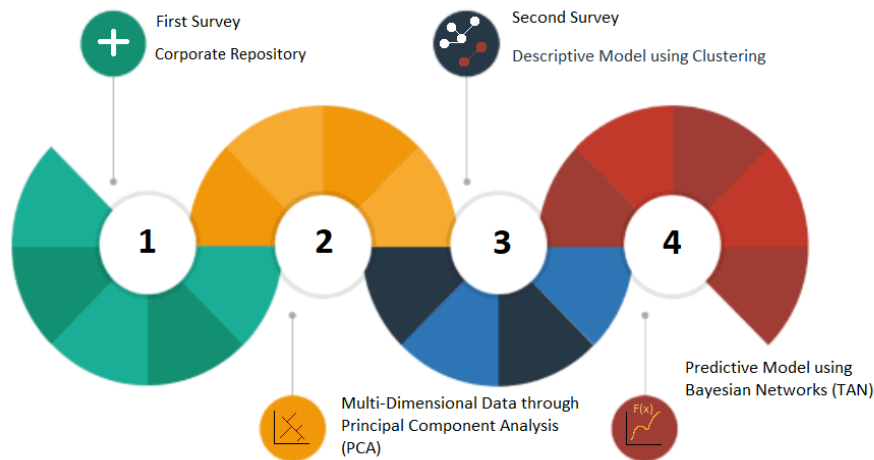


Figure 1: Step-by-step approach to build the descriptive and predictive models.

Our first batch of these questions is listed in Table 1. Note that, for each of the close-ended questions in Table 1, the students had four choices to express the extent of their agreement. To avoid repetition, these choices are displayed only after question  $Q_1$ . Readers can assume that all the other questions have the same response choices.

An expanding body of research has been developed around the place of study and how it shapes the students' experience [Butler and Sinclair, 2020]. The place of study—whether at home or in campus—provides the context where the students learn. Gutiérrez et al. [Gutiérrez et al., 2021] have found that the larger the study rooms, laboratories, and classrooms are, the lower the student withdrawal rate is. This issue is certainly interesting for planning institutional policies. Hence, We explored the students' opinions about their places of study in the second part of our survey. We included some binary questions for this purpose, which are listed in Table 2. Note that, for each question in Table 2, the students had two choices: “Yes” or “No”. To avoid repetition, these choices are displayed only after question  $Q_{1,14}$ . Readers can assume that all the other questions have the same response choices.

Finally, we included a couple of questions intended to capture the students' preferences for the modes of teaching offered in the two programmes under study, and whether the students had to work to support their studies. These questions are listed in Table 3.

After compiling the responses to our survey, we realise that further details were required. Thus, we distributed a second survey, largely dedicated to identifying what the students consider to be academic success and failure. This second survey employed the ranking questions showed in Table 4. For these questions, students were asked to sort a list of answers into a ranked order, providing us with quantitative research data.

### 3.2 Data Integration

All the participating students were asked to grant us permission to retrieve their records from the corporate repository to verify their *academic history* up to the moment when the surveys took place. A student's academic history consists of all the module marks

<b>Questions</b>	
<i>Q</i> <sub>1,1</sub>	Do you ask questions in class when you do not understand something? <input type="checkbox"/> More than 75% of the times <input type="checkbox"/> Between 50% and 75% of the times <input type="checkbox"/> Between 25% and 50% of the times <input type="checkbox"/> Less than 25% of the times
<i>Q</i> <sub>1,2</sub>	When you fail to understand something in class, do you ask your questions later and resolve your doubts?
<i>Q</i> <sub>1,3</sub>	Do you always finish the coursework?
<i>Q</i> <sub>1,4</sub>	Do you do your coursework without looking at the solutions?
<i>Q</i> <sub>1,5</sub>	Are you able to take notes in class, keep up with the explanations provided, and understand the concepts at the same time?
<i>Q</i> <sub>1,6</sub>	Do you review your notes after class?
<i>Q</i> <sub>1,7</sub>	Do you make notes and highlight them as you read the teaching materials before or after class?
<i>Q</i> <sub>1,8</sub>	Can you understand and concentrate on the materials you read without re-reading a second or third time?
<i>Q</i> <sub>1,9</sub>	Can you read and learn at the rate of 12-15 pages per hour?
<i>Q</i> <sub>1,10</sub>	Do you start your revision at least 3 days ahead of the exam?
<i>Q</i> <sub>1,11</sub>	Do you use the time between classes to study?
<i>Q</i> <sub>1,12</sub>	Do you study at least 2 hours per every hour of class?
<i>Q</i> <sub>1,13</sub>	How often do you attend class?

*Table 1: Survey 1: Closed-ended questions about self-regulatory learning behaviours.*

achieved by the student which contribute towards her progression and final award. Students' attendance records and information regarding their previously earned degrees were also retrieved from the corporate repository with the students' consent. Students who did not provide their consent were removed from our research.

<b>Question</b>	
Q <sub>1,14</sub>	Do you have a place where you always go to study? <input type="checkbox"/> Yes <input type="checkbox"/> No
Q <sub>1,15</sub>	Is your area of study free of noise and distractions and comfortable?
Q <sub>1,16</sub>	Do you study in the company of other classmates?
Q <sub>1,17</sub>	Can you study for at least half an hour without getting up, taking a snack, or having phone breaks?
Q <sub>1,18</sub>	Do you use the University library?
Q <sub>1,19</sub>	Do you use the self-study infrastructure provided by the University's learning management system in your spare time?

Table 2: Survey 1: Binary questions about the place of study and its environment.

<b>Question</b>	
Q <sub>1,20</sub>	What is your favourite type of teaching mode? Rank the answers below from 5 (your most preferred answer) to 1 (your least preferred answer): <input type="checkbox"/> Lectures <input type="checkbox"/> Workshops <input type="checkbox"/> Tutorials <input type="checkbox"/> Seminars <input type="checkbox"/> Online lectures and recordings
Q <sub>1,21</sub>	Do you have to work part-time to support your studies? <input type="checkbox"/> Yes <input type="checkbox"/> No

Table 3: Survey 1: Questions about teaching-mode preferences and part-time work.

<b>Question</b>	
<i>Q<sub>2,1</sub></i>	<p>What do you consider to be academic success? Rank the answers below from 4 (your most preferred answer) to 1 (your least preferred answer):</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Pass all the modules</li> <li><input type="checkbox"/> Pass the core modules which are prerequisites for the following year's modules</li> <li><input type="checkbox"/> Acquire knowledge, information, and skills</li> <li><input type="checkbox"/> Acquire knowledge which is applicable in the workplace</li> </ul>
<i>Q<sub>2,2</sub></i>	<p>Which of the following statements best identify the reason behind your academic success? Rank the answers from 4 (your most preferred answer) to 1 (your least preferred answer):</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> The amount of information delivered in class and comprised in the teaching materials</li> <li><input type="checkbox"/> The teaching modes applied by the academic staff: tutorials, work-shops, seminars, etc.</li> <li><input type="checkbox"/> An inspiring, motivational academic staff</li> <li><input type="checkbox"/> Your own skills and abilities</li> </ul>
<i>Q<sub>2,3</sub></i>	<p>Which of the following statements best identify the reason behind academic failure? Rank the answers from 4 (your most preferred answer) to 1 (your least preferred answer):</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> The academic staff did not deliver enough knowledge and information about the subject</li> <li><input type="checkbox"/> Failure to dedicate sufficient time to study</li> <li><input type="checkbox"/> Failure to understand the information delivered in class</li> <li><input type="checkbox"/> Lack of motivation or interest in the subject</li> </ul>
<i>Q<sub>2,4</sub></i>	<p>From the list below, prioritise the variables which have a direct bearing on academic success. Rank the variables from 6 (the most important one) to 1 (your least important one):</p> <ul style="list-style-type: none"> <li><input type="checkbox"/> Attendance at class</li> <li><input type="checkbox"/> Time dedicated to study</li> <li><input type="checkbox"/> Studying in a group</li> <li><input type="checkbox"/> Understanding the information delivered in class</li> <li><input type="checkbox"/> Interest in the subject</li> <li><input type="checkbox"/> Undertaking the coursework and studying the course materials</li> </ul>

*Table 4: Survey 2: Questions about academic success and failure.*

The IDs of the participating students were discarded after their academic histories and attendance records were retrieved to conceal their identities. This is crucial for ethical considerations and data handling, primarily to protect the privacy and confidentiality of the students, foster trust in research, and ensure compliance with legal guidelines.

The total number of students enrolled on the two programmes under study at the time of our investigation was 614, and we collected 87 voluntary responses for the first survey. The same students who responded to the first survey were later invited to participate in the second survey, and they agreed to do so. Thus, we received exactly 87 responses for each of the two surveys. Given the total size of the population for the two degrees under study—614 students—and the 87 responses received, we could guarantee that our confidence level was set to 5%, and we had a margin of error of 10%. The number of students at each level who participated in our surveys was proportional to the total number of students in each of the two programmes under study. A *level* is equivalent to an academic year—students must complete five levels or years to graduate.

The combination of the corporate repository and surveys yielded a total of twenty-three variables which have a direct bearing on academic success. These variables can be divided into the five categories listed in Table 5. Note that Table 5 also shows the associated variables per category, and these categories exhibit a clear overlap between our work and Tinto's theory of student withdrawal [Tinto, 2012]. Tinto's work on withdrawal is widely accepted to explain the students' decisions to dropout or persist [Samoila and Vrabie, 2023]. Hence, it was encouraging to discover the overlap between Tinto's theory and our own work. It demonstrates that our findings are not limited to the country or geographical region where the investigation was undertaken.

### 3.3 Principal Component Analysis (PCA)

Our next step was to undertake PCA to transform our large set of correlated variables—twenty-three variables to be precise—into a more manageable set, called *principal components*. We carried out this analysis using the *Weka Java API* [Hall et al., 2009], a Java machine-learning tool, which includes several built-in algorithms and a wide range of preprocessing methods, including normalisation, discretisation, and feature selection. PCA is often used in data analysis as the initial step for building predictive models, which is what we wanted to do to answer the third research question (RQ3).

Automatically, Weka normalises all the values by centring to the mean and dividing by the standard deviation. An extra advantage of employing PCA is that it helped us to prevent *overfitting* when developing the predictive algorithm—see Subsection 4.3 for further details. Schittenkopf et al. [Schittenkopf et al., 1997] and more recently Sarita et al. [Sarita et al., 2022] have demonstrated that PCA not only reduces the dimension of the input space—the dimensions of the training dataset—but also concentrates on the essential information in the internal representations. Hence, it reduces the number of free parameters, and the generalisation performance improves.

To validate the PCA, we applied two tests: *Kaiser–Meyer–Olkin* (KMO) [Shrestha, 2021] and *Bartlett's sphericity* [Mehmedinović, 2017]. KMO examines the strength of the partial correlation—how the original variables explain each other. Whilst KMO values below 0.5 are unacceptable, our test yielded 0.61, which is satisfactory. On the other hand, Bartlett's test validates whether the correlations in the data are strong enough to use a dimension-reduction technique, such as PCA. A significance level smaller than 0.05 suggests that the correlations are suitable. In our case, the chi-square test provided a value of 760.30, with 253 degrees of freedom, and a significance level of  $3.43e-52$ . These results confirm the feasibility of PCA.

Categories	Variables
Academic history	Number of modules failed; number of modules passed; number of modules dropped before completion; degrees previously earned; attendance at class.
Learning methods	Amount of time dedicated to self-study; doubt resolution approach; use of bibliographical materials; importance conferred to group study; reading and learning skills; revision habits; note-taking skills; coursework completion habits; library usage; interest in completing the programme.
Teaching methods	Preference for inspiring academic staff; preference for hands-on and interactive activities (tutorials); interest in specialised workshops and seminars.
Technological tools	Frequency of use of the University's learning management system.
Socioeconomic factors	Part-time work; year of admission at the University; year of birth (age); gender.

*Table 5: Variables which have a direct bearing on academic success.*

PCA produced eight components with correlated variables representing 70% of all variables, a statistically significant outcome. These components and their associated variables are displayed in Table 6, which not only includes the variables, but also the reasons for academic success or failure that were strongly correlated with the components. To strengthen our analysis, we made use of *varimax rotation* [Jackson, 2005], a statistical method used to simplify the results of PCA, making it easier to identify which variables are most strongly associated with each component. Figure 2 displays our results. Varimax gathers the greatest coefficients between  $-1$  and  $1$ , or the smallest ones—around zero—to indicate whether the variables have a strong relationship with the components or not.

## 4 Results and Discussion

*Clustering* is a technique to separate data of similar nature [Xu and Wunsch, 2008]. We used two clustering algorithms—K-means and EM—which are similar in the sense that they both allow us to refine an iterative process to find the best congestion [Ahmed et al., 2020]. To implement K-means and EM, we employed the *SimpleKMeans* and EM class libraries incorporated in Weka [Arthur and Vassilvitskii, 2007]. *SimpleKMeans* allowed us to identify three clusters using 40 seeds and resulting in a mean square error of 419.14. Unfortunately, these clusters were not representative enough and yielded a few outliers. Thus, we tested EM. One of the advantages of EM is that it can automatically determine the number of clusters. We kept 100 as the maximum number of iterations and  $1e-10$  as the minimum standard deviation. This resulted in three groups with 44 seeds and a likelihood of  $-30.41$ . Figure 3 depicts these clusters:  $C_0$ ,  $C_1$ , and  $C_2$ .

Principal component	Variables
PC1: Personal information	Year of admission; year of birth (age); pass all modules ( <i>Reason for academic success</i> ).
PC2: Dedication	Time dedicated to study ( <i>Reason for academic success</i> ); time dedicated to revision; use of bibliographical materials.
PC3: Need to work	Part-time work; failure to dedicate sufficient time to study ( <i>Reason for academic failure</i> ).
PC4: Group and online study	Importance conferred to group study; studying in a group ( <i>Reason for academic success</i> ); use of the University's online learning system; number of modules passed.
PC5: Comprehension and understanding	Doubt resolution approach; understanding the information delivered in class ( <i>Reason for academic success</i> ); failure to understand the information delivered in class ( <i>Reason for academic failure</i> ).
PC6: Use of teaching materials	Undertaking the coursework and revising course materials ( <i>Reason for academic success</i> ); use of library; note-taking skills.
PC7: Attendance	Attendance at class; the academic staff did not deliver enough information ( <i>Reason for academic failure</i> ).
PC8: Motivation	Interest in completing the programme ( <i>Reason for academic success</i> ); inspiring, motivational academic staff ( <i>Reason for academic success</i> ); lack of interest in the subject ( <i>Reason for academic failure</i> ).

Table 6: Principal components associated with academic success and failure.

#### 4.1 Descriptive Model

For practical reasons, we employed the students' cumulative average to label the clusters. This was helpful for identification purposes, as recognised by the academics interested in our results. Cluster  $C_2$  became known as "High Achievers", because it comprised the students with the highest cumulative average in our study, namely, 71.02. Cluster  $C_2$  consisted of 31 of the 87 participating students—that is, 35.63% of the total. Cluster  $C_1$ , whose students shared a cumulative average of 61.82, became known as the "Low Achievers". Cluster  $C_1$  consisted of 39 students—that is, 44.82% of the total. Finally, Cluster  $C_0$ , whose students' cumulative average was 62.71, became known as the "Middle Achievers". Cluster  $C_0$  comprises 17 students—that is, 19.54% of the total.

Then, we sketched profiles for the students within each cluster. Table 7 displays the details. A High Achiever is a student who did not fail in the previous semester, has at most one module failed overall, and has dropped less than two modules since the start.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Year of admission	-0.84	0.24	-0.11	0.02	-0.21	0.10	0.05	-0.07
Year of birth (age)	0.84	-0.06	0.17	-0.13	0.15	-0.02	-0.06	0.04
Pass all modules	0.78	-0.05	0.13	0.14	-0.13	-0.08	-0.02	0.12
Time dedicated to study	-0.03	0.75	-0.08	-0.15	-0.07	-0.06	0.04	-0.13
Time dedicated to revision	-0.14	0.76	-0.06	0.14	0.02	0.12	0.20	0.09
Use of bibliographical materials	-0.14	0.53	-0.19	0.30	-0.27	-0.24	0.04	0.04
Part-time work	0.20	-0.08	0.93	-0.11	0.03	-0.06	-0.06	0.03
Failure to dedicate sufficient time to study	0.18	-0.13	0.93	-0.01	0.08	-0.04	-0.01	0.07
Group study	-0.33	0.35	0.21	0.55	-0.03	0.19	-0.06	0.25
Study in a group	-0.04	0.24	-0.03	0.63	-0.13	0.05	0.24	-0.08
Use of the learning system	-0.03	0.11	0.26	-0.71	0.03	-0.02	0.16	0.16
Number of modules passed	0.28	-0.25	0.09	0.60	0.37	-0.11	0.02	0.13
Doubt resolution approach	-0.32	0.24	-0.03	-0.23	-0.58	-0.15	0.06	0.14
Note-taking skills	0.06	0.06	0.08	-0.19	0.82	-0.07	0.03	0.02
Understanding the information	0.10	0.13	0.01	-0.10	-0.60	0.41	0.33	-0.16
Undertaking the coursework...	0.08	0.05	0.13	-0.05	0.14	-0.85	0.03	0.00
Library usage	-0.04	-0.11	0.05	0.13	0.00	0.55	0.54	-0.04
Reading and learning skills	-0.27	0.49	-0.02	-0.05	0.07	0.52	-0.08	-0.11
Attendance at class	-0.11	-0.24	0.00	0.32	0.17	-0.03	-0.77	0.04
Academic staff did not deliver enough information	-0.29	0.07	-0.10	0.25	0.03	-0.11	0.69	0.07
Interest in completing the programme	0.38	0.15	0.21	-0.07	0.15	0.37	-0.17	0.45
Inspiring, motivational academic staff	0.19	0.08	0.18	0.08	-0.17	-0.13	0.01	0.79
Lack of interest in the subject	-0.03	-0.33	-0.17	-0.22	0.25	0.00	0.06	0.67

Figure 2: Varimax rotation.

According to Table 7, profiling differences among the three clusters  $C_0$ ,  $C_1$ , and  $C_2$  are evident. The number of failed modules in the previous semester and overall, and the number of dropped modules before completion, are remarkably different between High and Low Achievers. Also, there are more than 9 points of difference in the cumulative average between High and Low Achievers—see Table 7.

The students’ own definition of academic success offers an alternative viewpoint to understanding the differences in Table 7. Neither one of the top two reasons for academic success among the High Achievers appears to be considered significant by the Low Achievers. Nevertheless, High and Middle Achievers considered “Understanding the class contents” among the top two reasons for their academic success. High Achievers considered “Time dedicated to study” to be the main reason for their success.

Figure 4 displays on the horizontal axis the six reasons for academic success listed in Question  $Q_{2,4}$ , whereas the vertical axis displays the three clusters of students  $C_0$ ,  $C_1$ , and  $C_2$ . We can clearly visualise the students’ agreement per reason for each cluster.

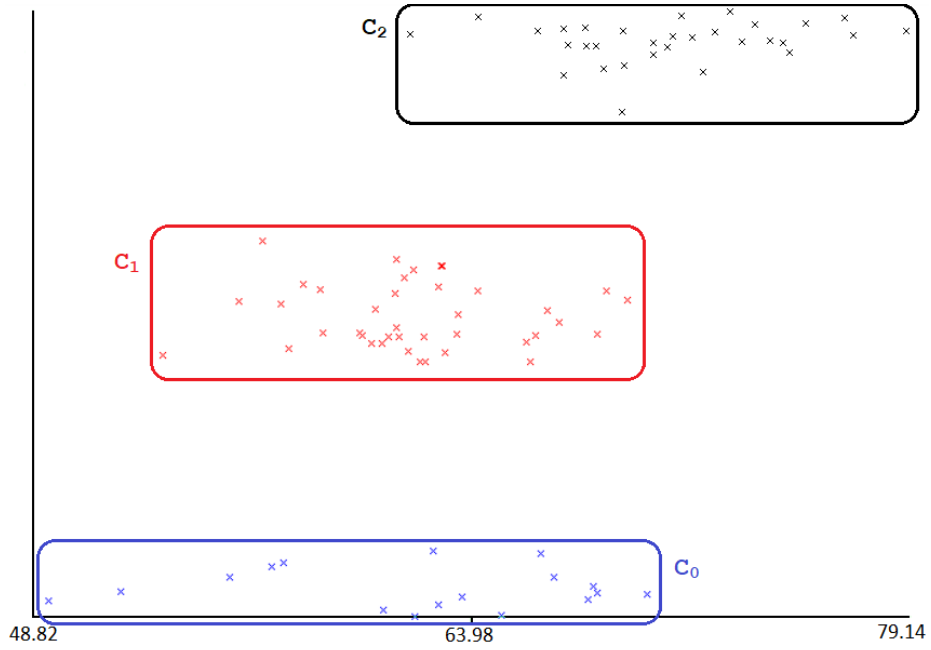


Figure 3: Student clusters: High Achievers (C<sub>2</sub>), Middle Achievers (C<sub>0</sub>), and Low Achievers (C<sub>1</sub>).

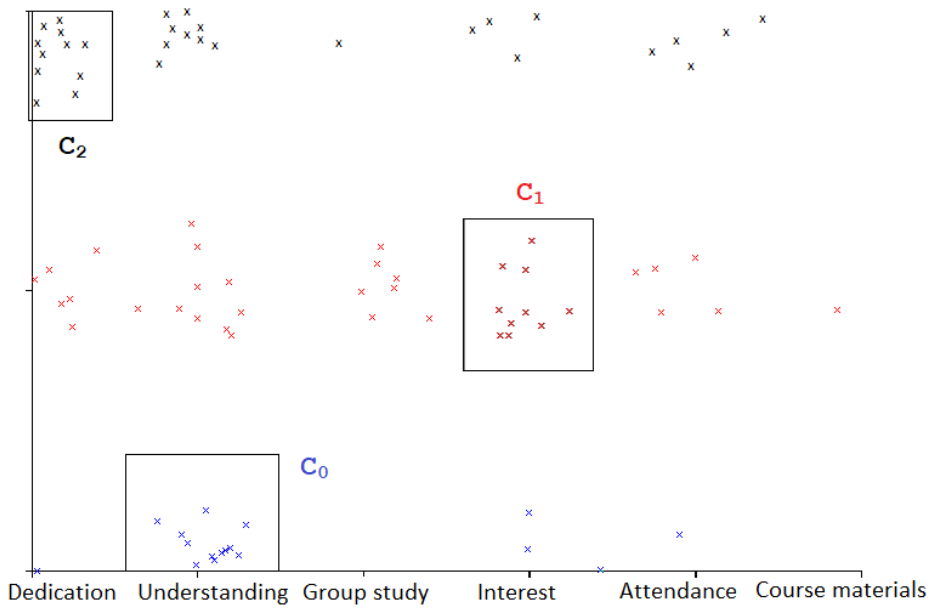


Figure 4: Reasons for academic success according to each cluster of students.

<b>High Achievers – Cluster <math>C_2</math></b>	
<b>Variables</b>	<b>Values</b>
Modules failed (previous semester)	0
Modules failed (overall)	1
Modules dropped	1.50
Cumulative average	71.02
Standard deviation of cumulative average	3.53
First reason for academic success	Time dedicated to study
Second reason for academic success	Understanding the class contents
Work part-time	Yes (55% of the students)
<b>Middle Achievers – Cluster <math>C_0</math></b>	
<b>Variables</b>	<b>Values</b>
Modules failed (previous semester)	0.31
Modules failed (overall)	5.64
Modules dropped	2.20
Cumulative average	62.71
Standard deviation of cumulative average	5.93
First reason for academic success	Understanding the class contents
Second reason for academic success	Attendance
Work part-time	No (77% of the students)
<b>Low Achievers – Cluster <math>C_1</math></b>	
<b>Variables</b>	<b>Values</b>
Modules failed (previous semester)	1.30
Modules failed (overall)	8
Modules dropped	3.50
Cumulative average	61.82
Standard deviation of cumulative average	3.30
First reason for academic success	Interest in the subject
Second reason for academic success	Undertaking the coursework and studying the course materials
Work part-time	No (60.40% of the students)

*Table 7: Representative profiles per cluster.*

## 4.2 Predictive Model

Along with the growth of learning analytics, performance predictors have been progressively incorporated into educational decision making. Although different approaches have been pursued—including neural networks [Rodríguez-Hernández et al., 2021], regression and correlation analysis [Romero and Ventura, 2010], and Bayesian networks [Amra and Maghari, 2017]—they all start by identifying the factors influencing academic success, which is what we did.

We favoured Bayesian networks, because they offer an intuitive modelling method, which does not require a great deal of statistical knowledge and therefore contributes to a straightforward interpretation and contextualisation [Fernández et al., 2011]. Bayesian networks have been successfully applied to many real-world domains [Pourret et al., 2008], and their classification accuracy in educational data mining has outperformed other options [Fida et al., 2022]. Moreover, Bayesian networks have been recommended to be used in the modelling complexities of Higher Education, given that they provide a “holistic” approach to answer common institutional questions [Di Pietro et al., 2015].

Given that “Pass all the modules of the semester” was considered by most of the participants in our surveys as the preferred way to describe academic success, we used the variable “Number of modules failed” as the predictor for evaluating students’ performance. This allowed us to anticipate a student’s failure in any course on a semester basis. To build our predictive model, we applied the K2 algorithm for Bayesian network structure learning [Behjati and Beigy, 2020], which is an extension of the *Tree Augmented Naive Bayes* (TAN) algorithm [Madden, 2008]. As a benchmark, we employed *hill climbing* (HC), which is a deterministic optimisation algorithm [Adhitama and Saputro, 2022], widely used to find the best possible solution to a given problem.

## 4.3 Overfitting and Generalisation

Overfitting is an undesirable behaviour present when the machine learning model gives accurate predictions for the training data, but not for new data [Montesinos López et al., 2022]. It occurs when a model learns too many details of the training data, including the noise and outliers, and is unable to generalise to new data.

We realise that our training data is small. Hence, to ensure that our model was not picking up noise, we applied *cross-validation* [Wong and Yeh, 2019]. To be precise, we applied a technique known as *Leave-One-Out Cross Validation* (LOOCV) [Silva and Zanella, 2024]. We separated our dataset into  $k$  subsets ( $k = 5$ ). Then, we tested each algorithm five times. However, for each time, we tested the algorithm using one of the subsets as the test set and we combined together the remaining  $k - 1$  subsets to form the training set. Rather than splitting the dataset randomly to generate the subsets, we made sure that each subset corresponded to a different year, or level, of study. Recall that the number students at each level of study who participated in our surveys was proportional to the total number of students in the two programmes involved in our study—remember that a level is equivalent to an academic year, and it is composed of two semesters.

## 4.4 Algorithm Comparison

Table 8 presents the comparison of the three algorithms involved in our study: HC, K2 and TAN. Once again, we employed Weka to implement all these algorithms and test them—the mean absolute error, root mean squared error, relative absolute error, and root relative squared error were all calculated using Weka.

Algorithm	Predc.	Correctly classified instances	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
TAN	N/A	93.10%	0.07	0.21	15.86%	42.84%
HC	1 pred.	95.70%	0.05	0.15	11.30%	30.25%
HC	2 pred.	93.10%	0.08	0.21	17.18%	42.43%
<b>K2</b>	<b>1 pred.</b>	<b>97.70%</b>	<b>0.05</b>	<b>0.14</b>	<b>11.36%</b>	<b>27.77%</b>
K2	2 pred.	96.55%	0.06	0.16	11.92%	33.51%
K2	4 pred.	95.40%	0.06	0.17	12.58%	35.77%

Table 8: Algorithm comparison: Average values after cross-validation.

The values displayed in Table 8 are the average values after applying cross-validation—that is, the average values computed after the execution of each algorithm on each of the five combinations derived from the ( $k = 5$ ) subsets. We have highlighted in bold font the row corresponding to K2 with one predecessor node, because this is the option which provided the highest accuracy and guaranteed the lowest root mean squared error. Note that accuracy was measured as the percentage of correctly classified instances—that is, all the correctly classified students. The accuracy of all the algorithms was reduced by a small percentage after cross-validation, but we still have K2 with one predecessor node as the best option. Of course, we are considering further training, testing, and validation with other datasets as part of our future work.

Some of the algorithms listed in Table 8, such as K2 with two predecessor nodes, are slightly less accurate than the best option, but may still be suitable for analysts' interpretations. In our experience, using the *Weka Explorer* [Attwal and Dhiman, 2020] as a visualisation tool, K2 with two predecessor nodes correctly maps all the nodes in the model into the variables which we identified as having a direct bearing on academic performance. This may be helpful for individual analysis and decision making.

Another interesting point to note is that K2 with four predecessors is equivalent to the *Bayesian Network Augmented Naive-Bayes* (BAN) algorithm [Sugahara and Ueno, 2021], which is similar to TAN, but generates an arbitrary graph, instead of a tree for the learning scheme. Cheng et al. [Cheng and Greiner, 2001] have highlighted that BAN generally outperforms TAN, which is exactly our case, as it can be seen in Table 8. Regrettably, BAN is expected to be slower than TAN.

Given that K2 with one predecessor node represents the best option at present, this is the algorithm that we have chosen to move forward with our predictive model. When assessing the predictive model's accuracy, we considered all the 87 participating students. Our model was able to classify correctly 84 of them, and this was indicated in the *confusion matrix* [Zeng, 2020] displayed in Table 9. The confusion matrix assesses where the errors in the model were made. The rows represent the actual classes that the outcomes should be—fail or pass—and the columns represent the predictions the model has made—"Yes" or "No". Our confusion matrix was derived from the *Weka's classification summary*. Note that 34 students were correctly classified as "Yes", meaning that they are likely to fail, whereas 50 students were correctly classified as "No", meaning that they are likely to pass all their modules. Only three students were wrongly classified.

A	B	Classified as
34	1	A = Yes
2	50	B = No

Table 9: Confusion matrix.

#### 4.5 Discussion

The literature is crammed with research offering guidelines and tips to apply tools and methods in educational environments. Although these contributions significantly advance the state-of-the-art, they are not necessarily useful at identifying an applicable model which is generic and sufficiently descriptive and predictive.

Unlike much of the existing research, our work is centred on the belief that gathering vast volumes of data is not enough. The emphasis must be placed on establishing a meaningful descriptive model. Later, we can progressively enrich such model with more data from different sources and tune up its accuracy. Naturally, our approach is not the only one available, but it has the advantage of economising time and resources.

An interesting point raised by our work is to do with the strong correlation between low attendance and students complaining that the academic staff did not deliver enough information to pass a module. This correlation is characterised by the seventh principal component, as stated in Table 6. In contrast, High Achievers—that is, students in Cluster  $C_2$ —have a higher attendance rate and do not think the academic staff fail to deliver enough information to pass the modules.

Part-time work is another interesting point. We expected to find that students who work dedicate less time to their studies and, consequently, accomplish lower marks. Contrary to our expectations, the High Achievers—Cluster  $C_2$ —comprise the largest number of students with a part-time job—see Table 7.

#### 4.6 Limitations and Extensibility

We recognise a key limitation of our study: all experiments were conducted using data from a single institution—Universidad del Bío-Bío. However, this constraint has also served as an advantage, as it enabled us to operate within a familiar context while designing solutions that have the potential to be adapted to other educational establishments.

Our work on data integration, which is described in Subsection 3.2, agrees with other more general studies. Our findings clearly overlap with Tinto’s theory, which is broadly documented as a foundational model for explaining dropout behaviour [Samoila and Vrabie, 2023]. Consequently, identifying a convergence between Tinto and our empirical results was both affirming and analytically significant. It means that our findings are not limited to the institution where the investigation took place but can be applied to others too. Preventing student withdrawal is critical for any institution [Behr et al., 2020], and our methodology can be reproduced anywhere.

Research conducted within American higher education institutions corroborates another one of our findings: the link between students’ success and part-time employment. It has long been established that students in American institutions, who engage in a moderate amount of paid work, tend to achieve higher academic performance, whereas excessive working hours are associated with adverse academic outcomes.

According to the *Bureau of Labor Statistics* (BLS) [Rones et al., 1997], students who work less than 20 hours per week have an average *grade point average* (GPA) of 3.13, while non-working students have an average GPA of 3.04. However, students who work more than 20 hours a week had a much lower GPA, namely, 2.95. This agrees with our discovery that the High Achievers—students in Cluster  $C_2$ —comprise the largest number of students with a part-time job—see Table 7. Again, despite limitations, our work confirms correlations found in other contexts and countries, and shows that the prospect of applying our methodology to other institutions is promising.

#### 4.7 Ethical Considerations and Data Availability

This study was approved by the Ethics Committee of Universidad del Bío-Bío. Informed consent was obtained from all the participating students. The data described and analysed here are not readily available, because the participating students did not consent to share their responses to the survey and other details outside Universidad del Bío-Bío or the scope of this study.

## 5 Conclusions

We have presented an application of data-driven decision making in the educational context at the University of the Bío-Bío. We have combined a corporate repository with two surveys distributed among the student population to identify a set of variables that influence academic success and dropout rates. We have applied PCA to determine eight principal components from these variables, and we have used such components to derive a descriptive model using clustering. Our clustering has allowed us to categorise students into three groups according to their academic performance: High, Middle and Low Achievers. Finally, we have produced a predictive model which offers more than 97% likelihood of forecasting module success or failure.

Our contributions are threefold,

**Reasons for success:** We have identified a set of normalised variables related to the academic success of computer science students. Even though we concentrated on computer science, the insights derived from our work are applicable to different contexts and institutions, giving us the opportunity to generalise our research.

**Descriptive model:** A model to characterise computer science students obtained using clustering and the normalised variables identified above.

**Predictive model:** A model based upon Bayesian networks used to determine the likelihood of academic success.

In the short term, we aim to analyse the variables with lower incidence within each cluster. This shall help us to suggest actionable policies to enhance academic performance. We also plan to examine the relationship between admission tests and future academic performance and investigate to other programmes and institutions.

#### Acknowledgements

This research was supported by Universidad del Bío-Bío, Chile, under Grants No. DI-UBB GI 195212/EF, and DIUBB 2130253 IF/R. Marco A. Palomino thanks the University of Aberdeen for supporting his participation in this work.

## References

- [Adhitama and Saputro, 2022] Adhitama, R. P. and Saputro, D. R. S. (2022). Hill Climbing Algorithm for Bayesian Network Structure. *AIP Conference Proceedings*, 2479(1).
- [Ahmed et al., 2020] Ahmed, M., Seraj, R., and Islam, S. M. S. (2020). The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. *Electronics*, 9(8):1295.
- [Alyahyan and Düşteğör, 2020] Alyahyan, E. and Düşteğör, D. (2020). Predicting Academic Success in Higher Education: Literature Review and Best Practices. *International Journal of Educational Technology in Higher Education*, 17(1):3.
- [Alzubi et al., 2018] Alzubi, J., Nayyar, A., and Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. In *Journal of Physics: Conference Series*, volume 1142, page 012012. IOP Publishing.
- [Amra and Maghari, 2017] Amra, I. A. A. and Maghari, A. Y. (2017). Students performance prediction using knn and naïve bayesian. In *8th International Conference On Information Technology (ICIT)*, pages 909–913. IEEE.
- [Archer and Prinsloo, 2020] Archer, E. and Prinsloo, P. (2020). Speaking the unspoken in learning analytics: Troubling the defaults. *Assessment & Evaluation in Higher Education*, 45(6):888–900.
- [Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). K-means++: The Advantages of Carefull Seeding. In *18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, New Orleans, Louisiana.
- [Attwal and Dhiman, 2020] Attwal, K. P. S. and Dhiman, A. S. (2020). Exploring Data Mining Tool-Weka and Using Weka to Build and Evaluate Predictive Models. *Advances and Applications in Mathematical Sciences*, 19(6):451–469.
- [Baker and Siemens, 2014] Baker, R. and Siemens, G. (2014). Learning Analytics and Educational Data Mining. *Cambridge Handbook of the Learning Sciences*, pages 253–272.
- [Behjati and Beigy, 2020] Behjati, S. and Beigy, H. (2020). Improved K2 Algorithm for Bayesian Network Structure Learning. *Engineering Applications of Artificial Intelligence*, 91:103617.
- [Behr et al., 2020] Behr, A., Giese, M., Teguiam Kamdjou, H. D., and Theune, K. (2020). Dropping out of University: A Literature Review. *Review of Education*, 8(2):614–652.
- [Ben Salamah et al., 2023] Ben Salamah, F., Palomino, M. A., Craven, M. J., Papadaki, M., and Furnell, S. (2023). An Adaptive Cybersecurity Training Framework for the Education of Social Media Users at Work. *Applied Sciences*, 13(17):9595.
- [Butler and Sinclair, 2020] Butler, A. and Sinclair, K. A. (2020). Place Matters: A Critical Review of Place Inquiry and Spatial Methods in Education Research. *Review of Research in Education*, 44(1):64–96.
- [Cheng and Greiner, 2001] Cheng, J. and Greiner, R. (2001). Learning Bayesian Belief Network Classifiers: Algorithms and System. In *14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence*, pages 141–151, Ottawa, Canada. Springer.
- [Daniel, 2015] Daniel, B. (2015). Big Data and Analytics in Higher Education: Opportunities and Challenges. *British Journal of Educational Technology*, 46(5):904–920.
- [Darko, 2021] Darko, C. (2021). An Evaluation of How Students Use Blackboard and the Possible Link to Their Grades. *SAGE Open*, 11(4):21582440211067245.
- [Dawson et al., 2019] Dawson, S., Joksimovic, S., Poquet, O., and Siemens, G. (2019). Increasing the Impact of Learning Analytics. In *9th International Conference on Learning Analytics & Knowledge*, pages 446–455.

- [Degenhard, J, 2024] Degenhard, J (2024). Total Consumer Spending on Education in Chile from 2014 to 2029. <https://www.statista.com/forecasts/1160936/education-consumer-spending-forecast-in-chile>. Accessed: 11 August 2024.
- [Di Pietro et al., 2015] Di Pietro, L., Mugion, R. G., Musella, F., Renzi, M. F., and Vicard, P. (2015). Reconciling Internal and External Performance in a Holistic Approach: A Bayesian Network Model in Higher Education. *Expert Systems with Applications*, 42(5):2691–2702.
- [Do and Batzoglou, 2008] Do, C. B. and Batzoglou, S. (2008). What is the Expectation Maximization Algorithm? *Nature Biotechnology*, 26(8):897–899.
- [Espinoza et al., 2024] Espinoza, O., Sandoval, L., González, L., Maldonado, K., Larrondo, Y., and Corradi, B. (2024). Reasons for University Dropout in Chile: Does Student Gender Play a Role? *Educational Review*, pages 1–16.
- [Fernández et al., 2011] Fernández, A., Morales, M., Rodríguez, C., and Salmerón, A. (2011). A System for Relevance Analysis of Performance Indicators in Higher Education Using Bayesian Networks. *Knowledge and Information Systems*, 27:327–344.
- [Fida et al., 2022] Fida, S., Masood, N., Tariq, N., and Qayyum, F. (2022). A Novel Hybrid Ensemble Clustering Technique for Student Performance Prediction. *Journal of Universal Computer Science*, 28(8):777.
- [Franco Caballero et al., 2020] Franco Caballero, P., Matas-Terron, A., and Leiva, J. (2020). Big Data Irruption in Education. *Pixel-Bit*, 57:59–90.
- [García y García, 2021] García y García, B. E. (2021). To What Factors Do University Students Attribute Their Academic Success? *Journal on Efficiency and Responsibility in Education and Science*, 14(1):1–8.
- [Gladshiya and Sharmila, 2019] Gladshiya, V. B. and Sharmila, D. (2019). A Review Study on Predictive Analytical Tools and Techniques in Education. *International Journal of Engineering Research and Technology (IJERT)*, 8(11):877–881.
- [Greenacre et al., 2022] Greenacre, M., Groenen, P. J., Hastie, T., d’Enza, A. I., Markos, A., and Tuzhilina, E. (2022). Principal Component Analysis. *Nature Reviews Methods Primers*, 2(1):100.
- [Gutiérrez et al., 2021] Gutiérrez, D. S., Domínguez, A. K., and Rivas, L. A. (2021). Incidencia de la Gestión Universitaria en la Deserción Estudiantil de las Universidades Públicas en Chile. *IE Revista de Investigación Educativa de la REDIECH*, 12:29.
- [Guzmán-Valenzuela et al., 2021] Guzmán-Valenzuela, C., Gómez-González, C., Rojas-Murphy Tagle, A., and Lorca-Vyhmeister, A. (2021). Learning analytics in higher education: A preponderance of analytics but very little learning? *International Journal of Educational Technology in Higher Education*, 18:1–19.
- [Hall et al., 2020] Hall, C., Mattox II, J. R., and Parskey, P. (2020). *Learning Analytics: Using Talent Data to Improve Business Outcomes*. Kogan Page Publishers, London, UK.
- [Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.
- [Ikotun et al., 2023] Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., and Heming, J. (2023). K-means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Information Sciences*, 622:178–210.
- [Jack, 2023] Jack, P. (2023). More than 40,000 Students Drop Out of UK University Courses. <https://www.timeshighereducation.com/news/more-40000-students-drop-out-uk-university-courses>. Accessed: 10 October 2024.
- [Jackson, 2005] Jackson, J. E. (2005). Varimax Rotation. *Encyclopedia of Biostatistics*, 8.

- [Jivet et al., 2018] Jivet, I., Scheffel, M., Specht, M., and Drachsler, H. (2018). License to evaluate: Preparing learning analytics dashboards for educational practice. In 8th International Conference on Learning Analytics and Knowledge, pages 31–40.
- [Leitner et al., 2017] Leitner, P., Khalil, M., and Ebner, M. (2017). Learning Analytics in Higher Education—A Literature Review. *Learning Analytics: Fundamentals, Applications, and Trends: A View of the Current State of the Art to Enhance E-learning*, pages 1–23.
- [Madden, 2008] Madden, M. G. (2008). On the Classification Performance of TAN and General Bayesian Networks. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 3–16. Springer.
- [Matcha et al., 2019] Matcha, W., Gašević, D., Uzir, N. A., Jovanović, J., and Pardo, A. (2019). Analytics of learning strategies: Associations with academic performance and feedback. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 461–470.
- [McKay et al., 2012] McKay, T., Miller, K., and Tritz, J. (2012). What to Do with Actionable Intelligence: E<sup>2</sup>Coach As an Intervention Engine. In *2nd International Conference on Learning Analytics and Knowledge*, pages 88–91.
- [Mehmedinović, 2017] Mehmedinović, S. (2017). Fundamentals of Application Factor Analysis in Education and Rehabilitation. *Journal for Interdisciplinary Studies HUMAN*, 7(1):61–65.
- [Montesinos López et al., 2022] Montesinos López, O. A., Montesinos López, A., and Crossa, J. (2022). Overfitting, Model Tuning, and Evaluation of Prediction Performance. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, pages 109–139. Springer.
- [Peña-Ayala, 2018] Peña-Ayala, A. (2018). Learning Analytics: A Glance of Evolution, Status, and Trends According to a Proposed Taxonomy. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(3):e1243.
- [Pourret et al., 2008] Pourret, O., Na, P., Marcot, B., et al. (2008). *Bayesian Networks: A Practical Guide to Applications*. John Wiley & Sons.
- [Ranjeeth et al., 2020] Ranjeeth, S., Latchoumi, T. P., and Paul, P. V. (2020). A Survey on Predictive Models of Learning Analytics. *Procedia Computer Science*, 167:37–46.
- [Roberts et al., 2016] Roberts, L. D., Howell, J. A., Seaman, K., and Gibson, D. C. (2016). Student Attitudes toward Learning Analytics in Higher Education: The Fitbit Version of the Learning World. *Frontiers in Psychology*, 7:1959.
- [Rodríguez-Hernández et al., 2021] Rodríguez-Hernández, C. F., Musso, M., Kyndt, E., and Cascallar, E. (2021). Artificial Neural Networks in Academic Performance Prediction: Systematic Implementation and Predictor Evaluation. *Computers and Education: Artificial Intelligence*, 2:100018.
- [Romero and Ventura, 2010] Romero, C. and Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.
- [Rones et al., 1997] Rones, P. L., Ilg, R. E., and Gardner, J. M. (1997). Trends in Hours of Work since the Mid-1970s. *Monthly Labor Review*, 120:3.
- [Samoila and Vrabie, 2023] Samoila, M. E. and Vrabie, T. (2023). First-Year Seminars through the Lens of Vincent Tinto’s Theories of Student Departure: A Systematic Review. In *Frontiers in Education*, volume 8, page 1205667. Frontiers Media SA.
- [Sarita et al., 2022] Sarita, K., Devarapalli, R., Kumar, S., Malik, H., Garcia Marquez, F. P., and Rai, P. (2022). Principal Component Analysis Technique for Early Fault Detection. *Journal of Intelligent & Fuzzy Systems*, 42(2):861–872.
- [Schittenkopf et al., 1997] Schittenkopf, C., Deco, G., and Brauer, W. (1997). Two Strategies to Avoid Overfitting in Feedforward Networks. *Neural Networks*, 10(3):505–516.

- [Selwyn, 2019] Selwyn, N. (2019). What's the Problem with Learning Analytics? *Journal of Learning Analytics*, 6(3):11–19.
- [Shrestha, 2021] Shrestha, N. (2021). Factor Analysis as a Tool for Survey Analysis. *American Journal of Applied Mathematics and Statistics*, 9(1):4–11.
- [Silva and Zanella, 2024] Silva, L. A. and Zanella, G. (2024). Robust Leave-One-Out Cross-Validation for High-Dimensional Bayesian Models. *Journal of the American Statistical Association*, 119(547):2369–2381.
- [Statista, 2024a] Statista (2024a). Education - United Kingdom. <https://www.statista.com/outlook/amo/app/education/united-kingdom>. Accessed: 11 August 2024.
- [Statista, 2024b] Statista (2024b). Education - Worldwide. <https://www.statista.com/outlook/amo/app/education/worldwide>. Accessed: 11 August 2024.
- [Sugahara and Ueno, 2021] Sugahara, S. and Ueno, M. (2021). Exact Learning Augmented Naive Bayes Classifier. *Entropy*, 23(12):1703.
- [Tinto, 2012] Tinto, V. (2012). *Completing College: Rethinking Institutional Action*. University of Chicago Press.
- [Turnbull et al., 2020] Turnbull, D., Chugh, R., and Luck, J. (2020). Learning management systems: An overview. *Encyclopedia of Education and Information Technologies*, pages 1052–1058.
- [Valenzuela and Kuzmanic, 2023] Valenzuela, J. P. and Kuzmanic, D. (2023). Dropouts and Transfers: Socioeconomic Segregation in Entrance to and Exit from the Chilean Higher Education. *Higher Education Research & Development*, 42(8):2048–2065.
- [Viberg et al., 2018] Viberg, O., Hatakka, M., Bälter, O., and Mavroudi, A. (2018). The Current Landscape of Learning Analytics in Higher Education. *Computers in Human Behavior*, 89:98–110.
- [Wong and Yeh, 2019] Wong, T.-T. and Yeh, P.-Y. (2019). Reliable Accuracy Estimates from k-Fold Cross Validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1586–1594.
- [Xu and Wunsch, 2008] Xu, R. and Wunsch, D. (2008). *Clustering*. John Wiley & Sons.
- [Yildirim and Gülbahar, 2022] Yildirim, D. and Gülbahar, Y. (2022). Implementation of Learning Analytics Indicators for Increasing Learners' Final Performance. *Technology, Knowledge and Learning*, 27(2):479–504.
- [Zeng, 2020] Zeng, G. (2020). On the Confusion Matrix in Credit Scoring and Its Analytical Properties. *Communications in Statistics-Theory and Methods*, 49(9):2080–2093.
- [Zollanvari et al., 2017] Zollanvari, A., Kizilirmak, R. C., Kho, Y. H., and Hernández-Torrano, D. (2017). Predicting Students' GPA and Developing Intervention Strategies Based on Self-Regulatory Learning Behaviors. *IEEE Access*, 5:23792–23802.