Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 7, No. 6, December, 2023

A Data Science Maturity Model Applied to Students' Modeling

L. Cavique [1, 2*], Paulo Pombinho [2], Luís Correia [2]

[1] *Universidade Aberta and Lasige-FCUL, Portugal.*

[2] *LASIGE, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal.*

Abstract

Maturity models define a series of levels, each representing an increased complexity in information systems. Data Science appears in the Business Intelligence (BI) and Business Analytics (BA) literature. This work applies the _IABE maturity model, which includes two additional levels: Data Engineering (DE) at the bottom and Business Experimentation (BE) at the top. This study uses the _IABE model for students' modeling in the ModEst project. For this purpose, the Public Administration organism is the Directorate-General for Statistics of Education and Science (DGEEC) of the Portuguese Education Ministry. DGEEC provided vast data on two million students per year in the Portuguese school system, from pre-scholar to doctoral programs. This work presents the comprehensible _IABE maturity model to extract new knowledge from the DGEEC dataset. The method applied is _IABE, where after the DE level, wh-questions are formulated and answered with the most appropriate techniques at each maturity level. This work's novelty is applying the maturity model _IABE to a unique dataset for the first time. Wh-questions are stated at the BI level using data summarization; at the BA level, predictive models are performed, and counterfactual approaches are presented at the BE level.

Keywords:

Maturity Model;
Wh-question;
Students' Modeling;
Business Intelligence;
Business Analytics;
Causality.

Article History:

Received:	20	August	2023
Revised:	13	October	2023
Accepted:	04	November	2023
Published:	01	December	2023

1- Introduction

In science and technology, asking the right questions is critical. The philosopher Claude Lévi-Strauss said, 'The scientist is not a person who gives the right answers; he asks the right questions'. In data science, given a dataset, asking the right questions, and having the necessary tools to obtain relevant information is even more challenging. With this target in mind, the utilization of wh-questions is preferred. Wh-questions start with wh-words, including what, when, where, which, who, whom, whose, why, and how. Wh-words or question words inquire about specific characteristics, reasons, qualities, times, places, facts, or people. A way to divide all the possible wh-questions is to create maturity levels of questions. Data maturity models are a valuable and current topic since they allow organizations to plan their medium- and long-term goals [1]. Maturity models are an essential business management tool, allowing organizations to improve the planning of actions that should lead to the desired results. This problem is even more relevant as new concepts, keywords, and products are launched annually in the information technology market, whose impact is rarely known.

The emergence of Data Science goes beyond Business Intelligence (BI) and Business Analytics (BA), abbreviated by BI&A. Data Science covers a broad spectrum of data and its derivatives, from initial data engineering (extraction, integration, transforming), exploration (aggregation, visualization), and modeling. Data Engineering (DE) arose in this decade in similar areas like Data Wrangling, Feature Engineering, and Data Pre-processing, partially replacing the well-known ETL (extraction, transformation, and loading). Data science overlaps multiple data-analytic disciplines, such as databases, statistics, operations research, and machine learning.

* **CONTACT**: luis.cavique@uab.pt

DOI: http://dx.doi.org/10.28991/ESJ-2023-07-06-08

The causal hierarchy [2] provides a framework for understanding causal inference. It is also called the causality ladder and consists of three rungs: association, intervention, and counterfactuals. Furthermore, most maturity models do not follow what is done in the technological sector, particularly in the GAFAM (Google, Amazon, Facebook, Apple, Microsoft) regarding Business Experimentation (BE) [3], closely related to the concept of intervention of Judea Pearl [2, 4].

Aiming to extend the hierarchy of causal inference to data science, the maturity model _IABE, an acronym of Intelligence, Analytics, and Business Experimentation [5], includes four levels. _IABE can be read as 'IAB' or 'in a bit', meaning 'in a short amount of time'. To the traditional BI&A, two more levels are added: Data Engineering (DE) at the bottom and Business Experimentation (BE) at the top. The proposed pipeline can be scratched as DE → BI → BA → BE. The _IABE model promotes the creation of the right wh-questions in the proper sequence.

The work described here is part of the project ModEst (student modeling, 'modelação de estudantes' in Portuguese) of the Portuguese education system developed from 2019 to 2022. For this purpose, data provided by the Directorate-General for Statistics on Education and Science (DGEEC) cover 12 academic years, from 2008 to 2019, from preschool to higher education, including data on enrollments of students, teaching modality, geographic location, as well as access to socio-economic data on students and parents.

Thus, to use these data, it was first necessary to understand the definitions used by the DGEEC and then proceed with Data Engineering to enable data extraction, cleaning, and loading into a dataset. It was necessary to consult the published legislation on each type of teaching modality existing in the database so that it was possible to understand how to use and frame this information.

1-1- Problem

Given a dataset, different kinds of questions can be asked. Several challenges arise in this context: determining the appropriate questions to pose, establishing a suitable sequence, identifying methods for acquiring answers, and recognizing which questions lead to new insights.

1-2- Objectives

The objective is to find a structured method to create a sequence of wh-questions and associate them with desirable techniques in order to find relevant knowledge. The sequence of questions is strategic to guide the exploration analysis by finding paths from data to insights. Moreover, to answer the questions, the appropriate tools are needed. Finally, only suitable data patterns with new knowledge are chosen.

1-3- Method

This work proposes a comprehensible maturity model with wh-questions for each maturity level. The maturity model _IABE is applied to the DGEEC dataset to achieve this end.

Figure 1 shows the workflow of the _IABE methodology. In this work, given the DGEEC, a sequence of wh-questions is employed to uncover insights from the data. The DIKW (data, information, knowledge, wisdom) Pyramid [6] outputs the process.

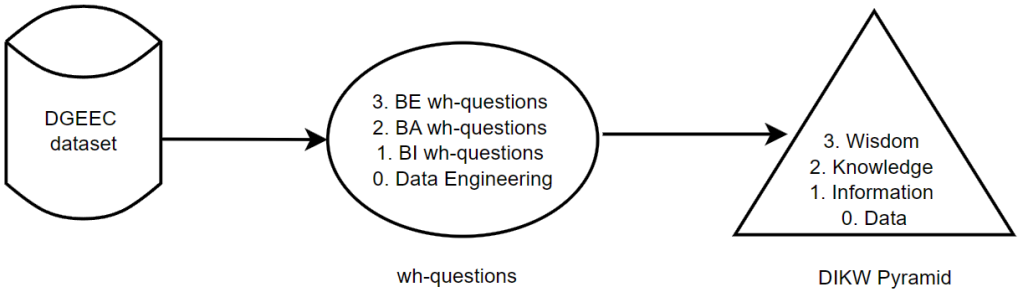

Figure 1. Workflow of the _IABE methodology

The DIKW Pyramid is often referred to in the information systems and knowledge management literature. Levels of data, information, and knowledge are consensual. However, the wisdom level needs further clarification. The term is replaced by awareness in agent-based programming since it implies a connection to the real world. Ackoff [7] associates wisdom with the term ‘why’ in the causal hierarchy [2]. In this work, wisdom is the result of the BE wh-questions.

1-4- Contributions

This work follows on from the _IABE maturity model [5] using the ModEst project data [8, 9] as a test case. This work's main contribution is applying the _IABE model to the DGEEC dataset by defining different wh-questions for each maturity model level and using the most appropriate techniques to obtain the answers. The novelty of this work is applying the maturity model _IABE to a unique dataset for the first time. Moreover, the goal is to create a sequence of wh-questions that unveil new knowledge from the dataset, unlike most published works that focus on a single research question. Multiple wh-questions can provide a comprehensive and holistic perspective of the student's modeling.

2- Students' Modeling

Education covers a range of sectors, from kindergarten, primary, and secondary schooling to higher education. The education sector can be seen as a series of components where each student follows a pathway that meets his aspirations [10]. Numerous methods have been proposed to enhance education due to abundant educational data. Many institutions analyze student conduct using Educational Assessment and Student Flow Analysis techniques. Many examples of mathematical models' application to education planning have existed since the late 1960s [11]. This state-of-the-art student flow problem identifies KPIs (key performance indicators), Visualization, Markov models, and What-if simulation approaches.

Viewing the education system in the framework of a Markov process is desirable. It is an ordered series of states linked by a transition matrix of probabilities of moving from one state to another. In the education system, Kwak et al. [12] considered students to be in the following states: studying full-time, studying part-time, on a temporary leave of absence, successfully graduating, or withdrawing. Markov chain models have proved particularly useful in providing not just predictions of students but also additional insights because they are based on student flows from state to state [13]. In Meece & Miller [14], the purpose is to examine the temporal stability of children's motivational goals for literacy activities. Achievement examined the stability of elementary school students' motivation goals (task mastery, performance, and work-avoidant). Task-specific assessments of students' goals were collected in the fall and spring of three grades, and changes were consistent across gender groups and ability levels.

In Science Education, two KPIs are usually established in student flow: dropout and failure (or retention) [15]. The knowledge and prediction of dropout and retention in the student flow are highly relevant since these KPIs directly influence the education system's performance. In Business Intelligence, What-if analysis fills this gap between Data Mining and Decision Making by enabling users to simulate and inspect the behavior of a complex system under some given hypotheses. The what-if dynamic simulation model allows analysis of potential changes in core curriculum policy, prerequisite structure, and staffing capacity to be tested before implementation [16]. Discrete Event Simulation is more adaptable to real-world applications than the pure Markov model. It accommodates more easily the complexities and interdependencies of the many components involved in the system [17].

Nese et al. [18] provide an overview of longitudinal educational data analysis. The primary purpose of longitudinal data analysis is to determine the variation pattern, growth or decrease, over time of the same variable. The authors refer to approaches such as Hierarchical Linear and Structural Equation Modeling. Kwok et al. [19] discuss the importance of the multilevel structure of educational data in longitudinal analysis using three different models. The authors use data from the English Language and Literacy Acquisition project to demonstrate the importance of capturing the complex data structure. Victorino et al. [20] investigate the impact of Design Thinking skills on Ph.D. students' academic and professional performance through a 7-year longitudinal mixed methods approach.

In recent years, innovative strategies have emerged to enhance teaching and learning, driven by expanding educational data and incorporating Artificial intelligence techniques, such as Educational Data Mining and Learning Analytics. Learning Analytics (LA) [21] refers to the importance of intensive use of data in education, with a view to planning and decision-making in education systems. Romero & Ventura [22] refer to Educational Data Mining as combining three main areas: Computer-based Education, Machine Learning, and Learning Analytics. LA takes a more holistic point of view, emphasizing understanding systems as a whole, whereas EDM focuses more on computational aspects.

Educational Data Mining identifies essential factors for academic success in a Portuguese business school's bachelor program, showing the influence of prior grades and school engagement on success at enrollment and the end of the first academic year [20, 23]. The study by Nunes et al. [24] combines machine learning efficiency with prototype analysis to quantify the impact of various factors on academic achievement in Mathematics and the mother tongue. Results highlight the significance of prior retention, legal guardian's education, and gender, offering insights for both research and practice.

The study by Costa-Mendes et al. [25] applies deep learning to predict high school grades in Portugal, comparing it with classical econometrics. Deep learning outperforms in student grade prediction and accurate prediction intervals. Feng et al. [26] address intelligent technologies in education by employing clustering, discriminant analysis, and convolutional neural networks to predict students' academic performance, enhancing prediction reliability through novel clustering methods and validation techniques.

Most of the previously reported works are focused only on one research question. In this work, the goal is to create a sequence of wh-questions revealing the most interesting patterns of the dataset.

3- Maturity model _IABE

The _IABE data science maturity model [5] comprises four levels: a previous level of Data Engineering (DE), followed by Business Intelligence (BI), Business Analytics (BA) at the second level, and Business Experimentation (BE) at the upper level. The model pipeline is DE → BI → BA → BE.

In education terminology, a rubric is a scoring guide used to evaluate students and is also known as a qualitative grading method. The rubric matrix includes the performance levels (in the columns), the performance criteria/dimensions (in the rows), and performance descriptors (in the cells). The most usual frameworks in organizational maturity models use wh-questions or rubrics (levels versus criteria). GAAM [27] is one of the most well-known wh-questions frameworks, and the Delta Plus model represents the rubric framework [28].

Table 1 shows the Data Science Maturity Model rubric framework with four levels and four criteria. The criteria/dimensions include the business information system, the approach to planning, the guidance of the level, and finally, a synthesis function where Tc means the algorithmic running time complexity, represented by the big O notation, and T denotes the treatment of the experiment.

Table 1. Data Science Hierarchy

		level 0	level 1	level 2	level 3
Criteria	Information System	Data Engineering	Business Intelligence	Business Analytics	Business Experimentation
	Approach to Planning	Inactive	Reactive	Proactive	Interactive
	Guidance	Data Pre-Processing	Data-Driven	Model-Driven	Experiment-Driven
	Function	N.A.	$y = f(X)$, $Tc(N) \leq O(N^2)$	$y = g(X)$, $Tc(N) > O(N^2)$	$y = h(X, T)$

The functions f(X), g(X), and h(X, T) are associated with BI, BA, and BE. Functions f(X) and g(X) use the same argument, X, where X denotes the set of attributes of the system. On the other hand, function h(X, T) has two arguments, where T represents the treatment. Since $O(N^2)$ is the time complexity between easy and hard problems, the threshold of $O(N^2)$ is defined for the time complexity to distinguish between BI and BA, where N is the number of lines of the dataset. The function f(X) can be exemplified by the sum of an attribute in an OLAP system, showing a running time complexity lower than $O(N^2)$, $Tc(N) \leq O(N^2)$. Moreover, function g(x) can be exemplified as a classification algorithm of a predictive model, with running time complexity usually greater than $O(N^2)$, $Tc(N) > O(N^2)$.

The second criterion in Table 1 is closely associated with how information system planning is developed. The BI behavior is reactive since it only cares about past events. Given the more complex models of BA, it is possible to elaborate on recommendations and take a proactive role in the company. Finally, in the BE stage, the ability to interact with customers by performing controlled trials moves the company to a new level in organizational learning with interactive planning. The previous level, named level 0, comprises Data Engineering tasks, where questions or answers are not presented, and there is no planning. This level includes pre-processing data features as an ETL process (Extraction, transformation, and loading).

Business Intelligence (BI) comprehends tools for data-driven decisions, emphasizing reporting and data visualization. Data warehouse design is essential to provide multidimensional data tables that can be analyzed using OLAP (online analytical processing) systems. Management by exception can be carried out based on the KPI, triggering alerts when threshold values are reached. The approach to planning is reactive based on the current information.

Business Analytics (BA) merges the areas of Data Mining with Decision-Making tools. In Data Mining, two sub-areas should be mentioned: descriptive and predictive. The descriptive approach looks for relevant patterns in the data, and the predictive models use supervised algorithms with labeled data to anticipate future events. Decision-making models include the techniques studied in Operations Research, like decision analysis, simulation, and optimization. These techniques aim to find the best solutions for each decision problem. BA comprehends tools to support model-driven decisions where the approach to planning is proactive.

In Business Experimentation (BE) [3], experimentation interacts with individuals (customers, patients, or users), generating more data and feeding back to the system. A low-cost business experiment can change the way organizations design decision-making. BE comprehends tools to support experiment-driven decisions in interactive planning. Pearl's causality hierarchy refers to association, intervention, and counterfactuals, where the association is closely related to the traditional data mining approach. In the BE level, a new sub-level, the explanatory one, is added, reflecting the title of the book "The book of why" [4]. Figure 2 shows the three stages associated with the standard techniques and the related comprehensive questions that can be asked. BI&A is highly effective at answering questions of 'what'. On the other hand, BE answers questions of 'what-if' and 'why', which implies causal relationships.

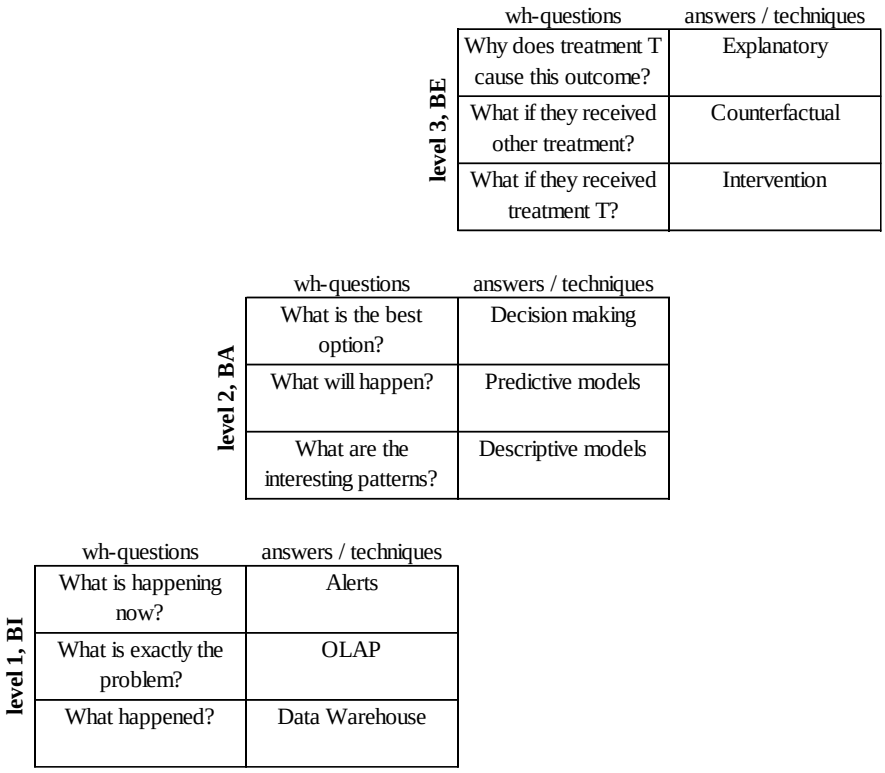

Figure 2. Sub-levels of the Data Science hierarchy with wh-questions

The sub-levels of BI and data warehouses, OLAP systems, and Alerts supported by KPI are included. The BA sub-levels comprehend the descriptive, predictive, and decision-making models. The sub-levels of BE include the last rungs of Pearl's causality ladder: the intervention, the counterfactual, and the explanatory reasoning. In this work, a sequence of wh-questions is formulated, one for each level, using the DGEEC dataset provided by the project ModEst. Table 2 shows the general wh-questions and the techniques used to answer them. Several wh-questions offer a thorough view of the student's modeling.

Table 2. Wh-questions and techniques of the ModEst project

_IABE	BI	BA	BE
wh-questions	Is there some significant difference?	What will happen?	What if they received other treatment?
answers/ set of techniques	OLAP	Predictive models	Counterfactual
technique	Pivot table	Time series forecasting	Regression discontinuity

In a previous level, level 0, the data engineering process was carried out to model the students' performance.

In level 1, using OLAP techniques, questions are formulated in order to find answers to previously encountered problems. Different pivot tables are built on understanding data variations. The concept of management by exception, i.e., focusing on significant deviations from previous events, is widely applied. In level 2, in the BA domain, predictive models foresee the number of students in the short term. Regarding the level 3, counterfactual events are studied. Regression discontinuity design is applied to measure the impact of national legislation on education.

In the following sections, the 0 to 3 levels are detailed. In levels 1 to 3, the most relevant wh-questions are proposed, and the insights are presented.

4- Data Engineering

The data provider of the ModEst study is the Directorate-General for Statistics of Education and Science (DGEEC) of the Portuguese Education Ministry. DGEEC has vast data regarding two million students per year in the Portuguese school system, from pre-scholar to doctoral programs. The data provided by DGEEC have been anonymized to meet the General Data Protection Regulation (GDPR) requirements.

Only the data from the 1st to the 12th grade are used in this work, leaving the study of higher education for future work. The available dataset contains the enrollment information of pre-primary, primary, secondary, and post-secondary education levels of ISCED [29] over twelve school years, from 2008-2009 to 2019-2020. The Portuguese educational system comprises four educational levels: basic-1 (1^{st}-4^{th} grade), basic-2 (5^{th}-6^{th} grade), basic-3 (7^{th}-9^{th} grade), and secondary (10^{th}-12^{th} grade).

The dataset used consists of information on student enrollments. Each entry concerns a student's enrollment in a specific school year, with information detailing the characteristics of the enrollment and its outcome. It has the following relevant attributes:

- School year (e.g., 2008–2009);
- Grade (e.g., 12th grade);
- Nature (e.g., public or private schools);
- Geographic information (NUTS continental Portugal);
- Modality of the course (e.g., regular or professional education);
- Age and gender;
- Ingress and Outcome event (dropout, retention, pass).

The geographical classification NUTS is derived from the French version of 'Nomenclature des Unités Territoriales Statistiques' to divide the economic territory of the European Union. Concerning NUTS I, the dataset only has information from mainland Portugal, with no information from the Autonomous Regions of Madeira and the Azores.

The DGEEC has annual data, and the ModEst project is interested in having a longitudinal view of the data [8]. So, it is necessary to correctly perceive how a student's path through the education system changes and what events are associated with them. The path of a student as a sequence of states with three different types of state inputs and three types of outputs (Figure 3):

- ***Dropout***: there is no information about the student in the next school year;
- ***Retention***: the student is in the same grade in the following school year;
- ***Pass***: the student is enrolled in a higher grade in the following school year;
- ***Ingress***: there is no information about the student in the previous school year.

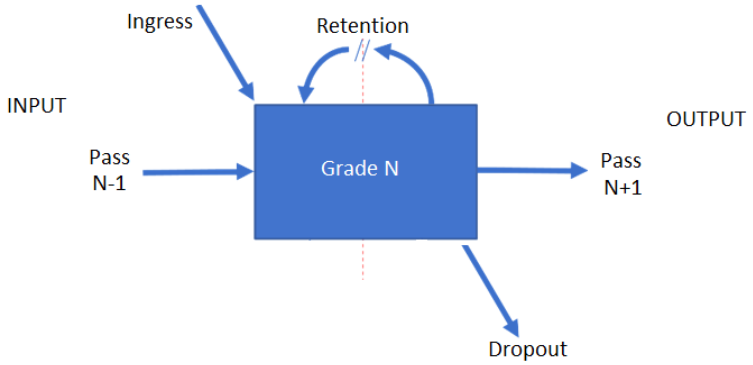

Figure 3. A state (concerning a grade N) with input and output state transitions

The outcome events for DGEEC are three equally but slightly different, as follows: (i) unknown or unspecified, (ii) retention and withdrawal, and (iii) transition or completion. The second event joins retention and dropout.

The dataset has around 19 million instances, corresponding to an average of 1.5 million students per school year. The DGEEC used an anonymization algorithm to comply with the privacy requirements. The algorithm led to some unintended inaccuracies in the data, which has made it more challenging to take a longitudinal view [9]. Data cleaning and normalization procedures were performed as usual in real data projects. In particular, the dimension grade was normalized from 0 (kindergarten) to 13 (post-secondary) for the different course modalities. During the transformation phase, the classifications were normalized, the duplicate student registrations were removed, and missing data were removed.

Given the anonymization process, student flow or cohort studies are not visible. Therefore, given the GDPR, only aggregated data could be used in this study, which is limited to annual enrollment counts. In the dataset, a counter defines the number of students for each combination of the dimensions. The aggregate information results in a dataset with 885 thousand instances.

The dataset with the events contains an aggregate view, with the students' ingresses, and allows us to obtain information about the students' outcome events. This information is combined with other attributes, such as nature, geographic information, course modality, age, and gender, for more detailed analysis.

5- Business Intelligence

OLAP (Online Analytical Processing) technology allows the organization to use large datasets to obtain new perspectives, being a relevant support for Business Intelligence. One of the components of OLAP systems is the use of cubes that aggregate different levels and hierarchies of the various attributes to be analyzed. Cuboids of N dimensions can be created. Using the OLAP model, it is possible to create pivot tables to explore and compare data.

In the case of the ModEst project, six dimensions were explored: school year, grade, NUTS, course modality, nature, and event, which can be grouped into cubes of different dimensions. There are several possible combinations of variables to create pivot tables and perform OLAP operations. The challenge is to find differentiating and meaningful information, that is, to create a clear and valuable picture that helps to understand reality. In this work, two examples are provided regarding the percentage of success by NUTS and how national exams can influence the attributes of Nature and Modality. The chosen data correspond to the last three years of available data to avoid contamination of data from being too old.

Figure 4 shows the percentage of success of each NUTS II compared with the mean success. The outcome of the transition /completion is used to define success. The success is more considerable in NUTS II of North and Center and decreases in A.M. Lisboa, Algarve, and Alentejo. There is a difference between the different NUTS II, although without clear statistical significance with a p-value=0.154. However, combining three dimensions, success, NUTS II, and nature (private and public), the ANOVA test reveals a statistically significant difference, with a p-value<0.001.

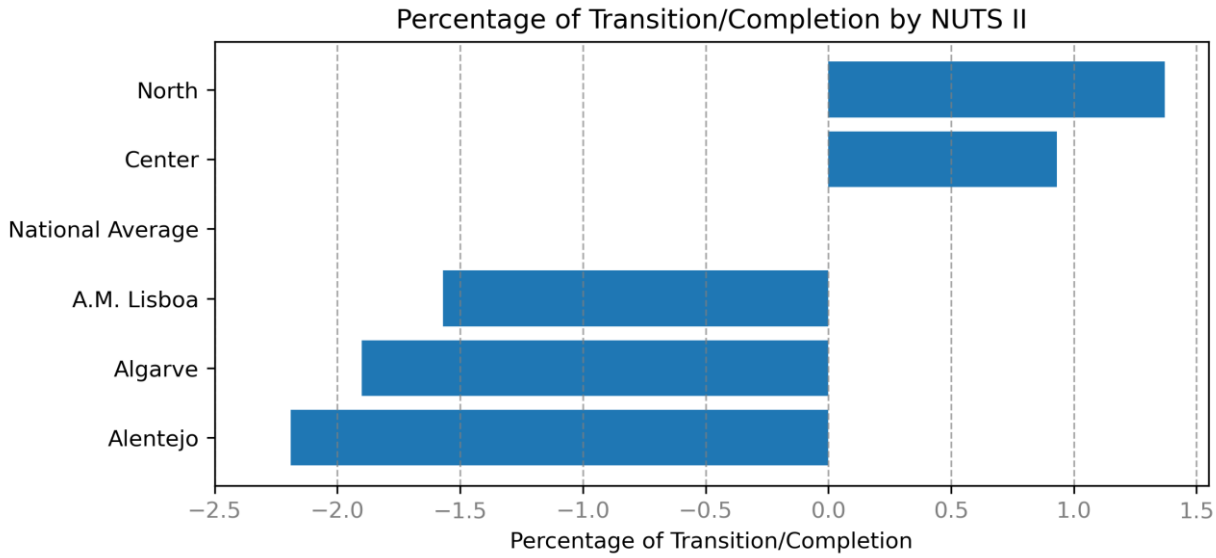

Figure 4. Percentage of success by NUTS II

There are two groups of national assessments: benchmarking tests and exams. The 2022-2023 school year calendar provides benchmarking tests for the 2nd, 5th, and 8th grades. The final tests are accomplished in the 9th year, and national final exams are in the 11th and 12th years. Figure 5 shows a pair of charts to reveal the eventual consequences of the 9th-grade final exams and the 12th-grade national exams.

Figure 5a shows the percentage of students in private schools from the 1st to the 12th grade. The percentage of students is almost constant until the 8th grade, starting to rise in the 9th grade and scaling up in the secondary level till the 12th grade. In private schools, the percentage of nearly 10% in the 8th grade grows to more than 20% of students in the 9th to 12th grades.

Figure 5b shows the number of students in traditional modality versus other modality courses. Traditional courses include general courses at the primary (basic) level and scientific and humanistic courses at the secondary levels. The other courses comprise a dozen of different course modalities. Two peaks in the number of students can be noticed in the grades with exams. The number of students in the 9th and 12th grades is more significant than in previous grades, given the number of retained students. Again, the traditional modality is constant until the 8th grade. The number of students in other modalities increases with the final exams of the 9th grade. On the other hand, the number of students in traditional modalities decreases with higher grades.

At the BI level, the wh-questions are related to past events. Statistical methods can be added to enhance the quality of the information. In this section, two wh-questions of the first maturity model level are formulated:

BI.1: Are there any differences in NUTS II regarding student success?

BI.2: Is there any difference between benchmarking tests and exams on the choice of schools (public, private) and the modality of courses?

Both wh-questions lead to affirmative answers. Regarding BI.1, success is more significant in the North and Center than in the other regions. Concerning BI.2, the 9th and 12th-grade evaluation drives the transition from public to private education and traditional courses to other courses. With OLAP, meaningful information that helps to understand reality can be extracted.

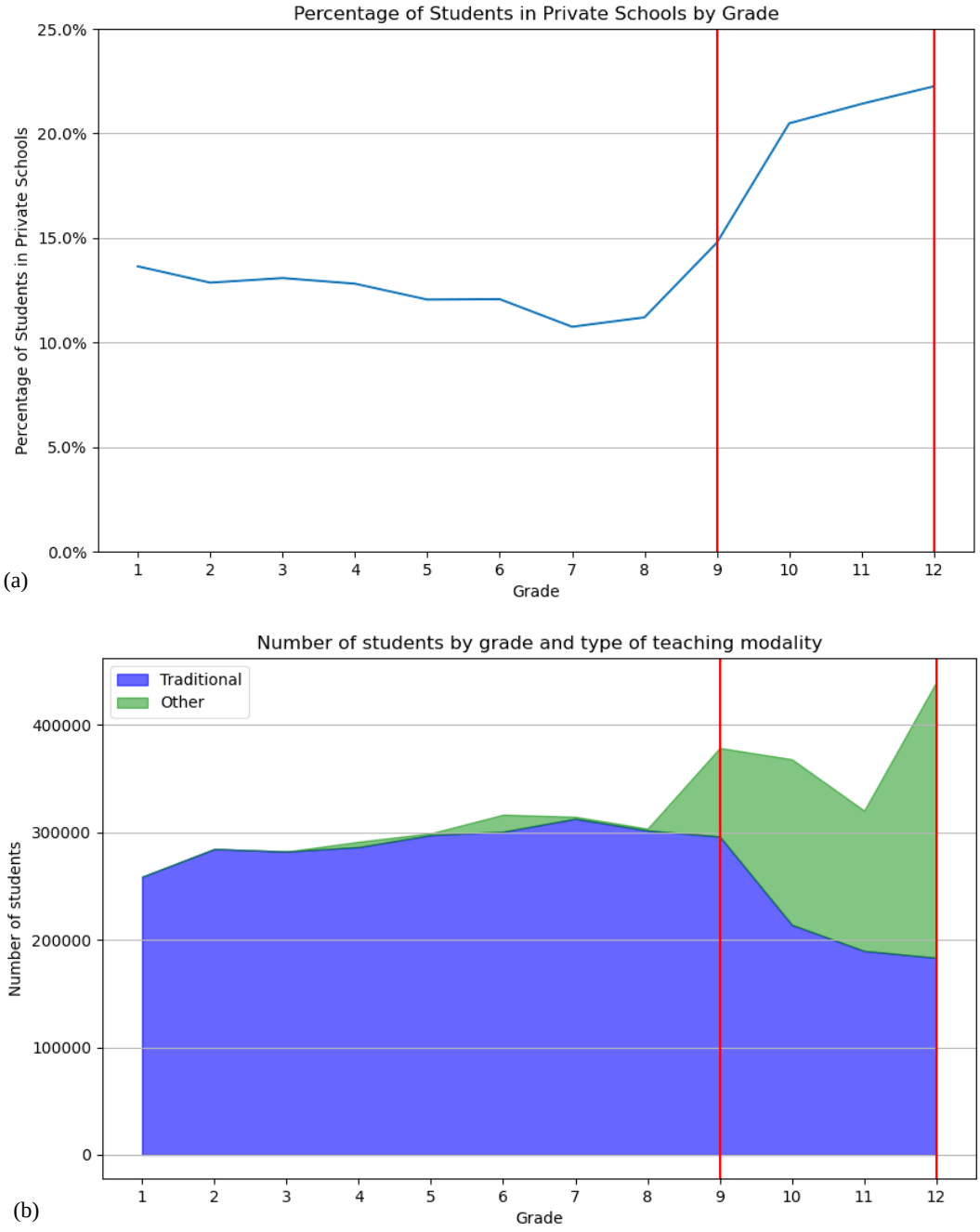

Figure 5. The variation in the attributes' nature (above) and modality (below) due to exams

6- Business Analytics

The reduction in the birth rate and consequent decrease in the number of students in the country is well-known. However, the exact reduction rate is not specified and needs to be quantified. At the BA level, the questions are related to predictions. In this section, the wh-question of the second maturity model level is formulated as follows:

BA.1: What is the prediction for the number of students in 2025?

A pivot table can provide the most complex information in a two-entry table. However, there are more techniques to extract information from tabular data. Given the information on the number of students in a pivot table with two entries

(grade versus school year), the goal is to transform it into tabular data, as shown in Figure 6. The tabular data can be seen as a time series where seasonality equals the number of grades. The time series, with 12 school years and 12 grades per year, results in 144 observations.

grade \ school year	2008	2009	2010	…
1	106,021	105,206	101,486	…
2	113,760	112,974	111,865	…
3	116,569	109,382	108,619	…
4	116,786	119,349	112,726	…
…	…	…	…	…

school year	grade	# students
2008	1	106,021
2008	2	113,760
2008	3	116,569
2008	4	116,786
2008	…	…
2009	1	105,206
2009	2	112,974
2009	3	109,382
2009	4	119,349
2009	…	…
2010	1	101,486
2010	2	111,865
2010	3	108,619
2010	4	112,726
2010	…	…

Figure 6. Transformation of a two-entry table (left) into tabular data (right)

In the time series visualization, the trend and the seasonality are pretty straightforward, which leads us to consider the model of classical decomposition. Economists first used the classical decomposition method in the 19th century. The basis of the method includes four components: trend T, seasonality S, cyclicality C, and randomness R. The observed variable X uses the previous components, X = f (T, S, C, R). C is neglected since cyclicality is only observed for long series, resulting in X = f (T, S, R). In the classical decomposition approach, two models can be referred to, the additive and the multiplicative one:

- Additive model: $X_t = T_t + S_t + R_t$
- Multiplicative model: $X_t = T_t \times S_t \times R_t$

The model results in a forecast F dependent on trend and seasonality F = f' (T, S). The undesirable error R arises by comparing the forecast F with the observed data X, where R = X - F or R = X ÷ F, depending on the model. The classical additive decomposition method was chosen since the additive and multiplicative models present the same error. The model considers that the predicted value for a time series equals the sum of the trend and seasonality, allowing a forecast for 2025. Figure 7 shows that the trend of the time series is decreasing, and the progression of grades has two peaks, in the 9 and the 12 grades, resembling seasonality in time series data

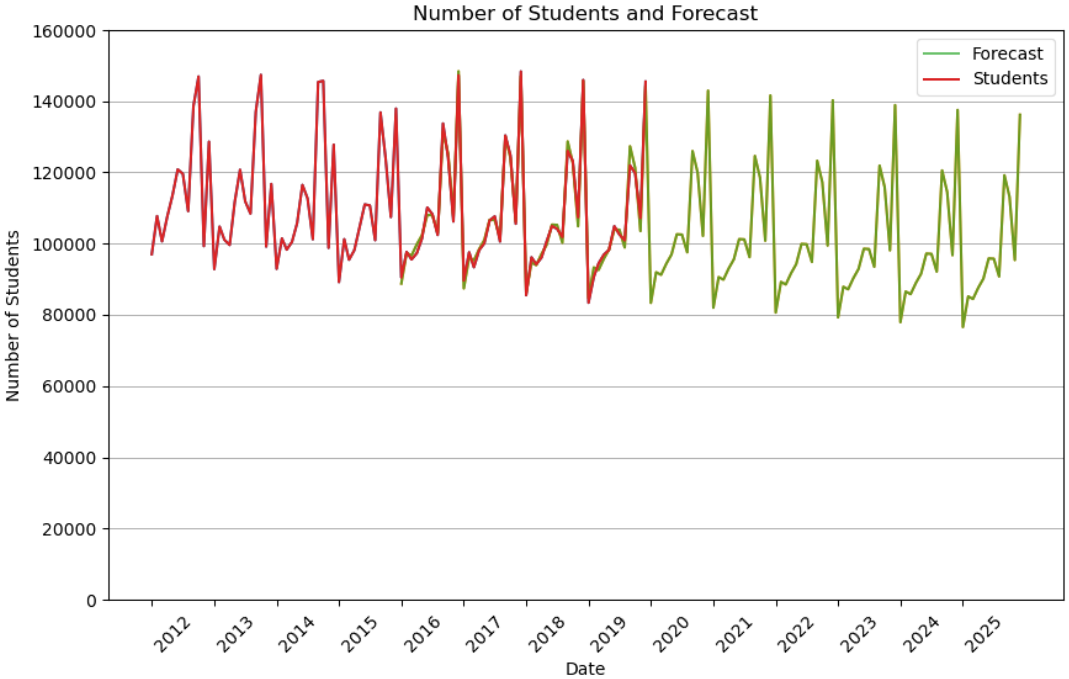

Figure 7. Additive model to predict the number of students in 2025

The observed data is from 2012 to 2019, the trained data is from 2016 to 2019, and the forecast is from 2016 to 2025. The observed and forecasted time series overlapped from 2016 to 2019 with a slight error of 1.24%. It was verified that the error was minor when using only the last four years of the series. Answering question BA.1, the prediction for the number of students in 2025 equals 1.17 M (million). Finally, suppose the decreasing trend is approximately constant, around 16 K (thousand) students per year. In that case, therefore, it is foreseeable that the threshold of one million students will be reached in the year 2036.

7- Business Experimentation

As already stated, analytical models discover or describe intriguing patterns and make predictions based on good fits of historical data. In evaluating data-driven models' recent movements, as in xAI (explainable artificial intelligence), are presented in favor of explanatory models [30]. The difference between correlational analysis and causality is at the heart of the controversy over prediction and explanation. In machine learning, two tasks must be distinguished: prediction and explanation. In prediction, two variables are used: the independent variable X and the dependent variable Y. The original data is divided into the training and testing data sets to find the Y=f(X) function, where X is a covariate, and Y is the outcome.

A new variable type should be included in the intervention/treatment T. In this task, outcome Y of treatment T is the subject of the study. For this purpose, test and control datasets are used for treatment accomplished T=1 and not accomplished T=0. In analogy with Y=f(X), the explanatory function uses three variables, Y=f (T, X). This dichotomy can be found in the ladder of causation [4]. Pearl [2] proposes three levels of causality — association, intervention, and counterfactual. The association has no causal consequences, contrary to the intervention and counterfactuals. Association corresponds to the predictive approach Y=f(X). The intervention is exemplified by the A/B testing, where treatment T appears in the equation Y=f (T, X). Finally, counterfactual (or unavailable data) involves imaginary worlds and specific approaches that compare treatments.

Counterfactual analyses have become popular since the philosophical developments in the 1970s. The best-known counterfactual analysis of causation is Lewis's theory [31]. Counterfactual truths are fictional since they occur in a different world. A counterfactual world is Spatio-temporal disconnected from our world, and there is no interaction between worlds, so its empirical existence cannot be checked.

The core idea behind the counterfactual theory of causation is causal dependence. A hypothetical scenario can define causal dependence: given that T and Y are different events, Y causally depends on T if and only if the counterfactual "if T were not to occur, Y would not occur" is a proper sentence. The same reasoning is addressed in law, considering that jurists have searched for a direct test of the defendant's guilt called 'but-for causation' or objective cause for centuries. This case is also known as a *sine qua non* or necessary condition.

However, the intense discussion over the last fifty years has not privileged the application of counterfactual concepts in causality. Currently, SEM (Structural Equation Modeling) and SCM (Structural Causal Models) [32] are the most well-known frameworks for studying counterfactuals in causation.

Like an experiment using interventions, the contractual approaches use a quasi-experimental design to establish a cause-and-effect relationship between an independent and dependent variable. A significant advance in the operability of counterfactual reasoning took place in econometrics [33]. Some counterfactual questions can be answered using appropriate statistical techniques, named counterfactual impact evaluation methods. Regression Discontinuity Design and Differences in Differences allow an intuitive graphical representation [34]. Regression Discontinuity Design is a quasi-experimental impact evaluation method that uses a cutoff point in addition to the input dataset. The cutoff point corresponds to an intervention, dividing the independent variable into control and treatment.

Next, the case study of introducing compulsory education for children and young people between 6 and 18 years old is presented, regulated by Decree-Law No. 176/2012.

The BE wh-question of the third maturity level:

BE.1: What is the effect of Decree-Law No. 176/2012 regarding the number of students in high school?

The time series shows the number of high-school students (grades 10, 11, and 12) for the school years from 2008 to 2019. Figure 8 shows a discontinuity in 2012, corresponding to the cutoff point. The cutoff point matches the year of introduction of the compulsory law. Two regressions are shown with different slopes, and the extension of the regressions represents the contrafactual data.

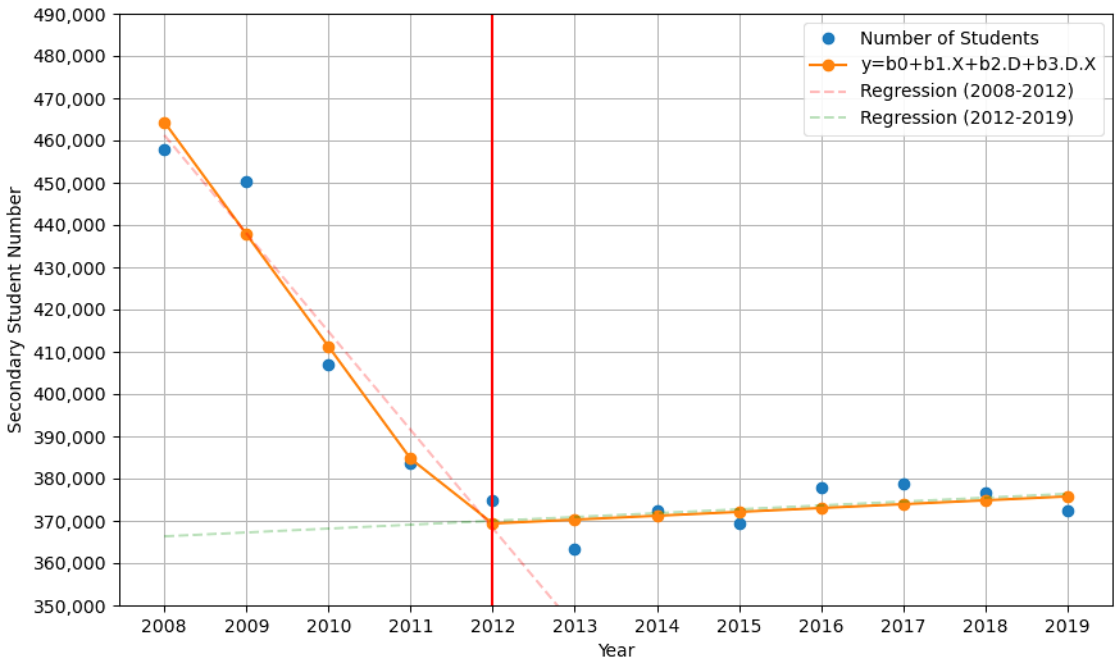

Figure 8. Observed data and regressions before and after the cutoff point

The number of high-school students had been decreasing by 27 K (thousand) per year before the implementation of the Decree Law, but after the cutoff point, there was an increase of nearly 1 K (thousand) students. The causal effect is the difference between the number of students who have not left the education system, and therefore, the system increased to more than 27 K (thousands).

Given the number of high-school students X and the dummy variable D corresponding to the cutoff point:

$$D = \begin{cases} 0 \; if \; year < 2012 \\ 1 \; if \; year \geq 2012 \end{cases} \quad (1)$$

The discontinuity regression equation using the interaction between variables D and X equals Y = β0 + β1. X + β2.D + β3. D. X, with the coefficients β0=53,784,799, β1=-26,554, β2=-55,254,393, and β3=27,468.

Regarding the quality of the model, comparing the observed data and regressions shows a slight error of 1.18%.

8- Discussion

The data of the ModEst project of Portuguese education support this work. The data provided by the Directorate-General for Statistics on Education and Science (DGEEC) cover 12 academic years, from 2008 to 2019, from preschool to higher education, including data on enrolments of students, teaching modality, geographic location, as well as access to socio-economic data on students and parents. The objective is to identify paths from data to insights within _IABE, illustrating them through wh-questions applied to aggregated data from the DGEEC dataset. Next, BI, BA, and BE wh-questions are summarized, along with their corresponding responses and potential insights.

At the BI level, two wh-questions of the first maturity model level are formulated:

BI.1: Are there any differences in NUTS II regarding student success?

BI.2: Is there any difference between benchmarking tests and exams on the choice of schools (public, private) and the modality of courses?

Regarding BI.1, success is more significant in the North and Center than in the other regions. Concerning BI.2, the 9th and 12th-grade evaluation drives the transition from public to private education and traditional courses to other courses. The BA wh-question of the second maturity model level was formulated as follows:

BA.1: What is the prediction for the number of students in 2025?

Answering question BA.1, the prediction for the number of students in 2025 equals 1.17 M (million). It is foreseeable that the threshold of fewer than one million students will be crossed in the year 2036.

The BE wh-question of the third maturity level:

BE.1: What is the effect of Decree-Law No. 176/2012 regarding the number of students in high school?

Regarding question BE.1, the causal effect shows the reversal of the downward trend and the increase of more than 27 K (thousands) high-school students.

In order to condense the key results of the students' modeling into a concise format, the following elements are recognized:

- The presence of exams triggers the mobility from public to private education and from traditional courses to other course types;
- A minor and non-alarming trend was identified in the predicted number of students in the coming years;
- The requirement for students to stay in the educational system until the age of 18 results in a notable rise in the enrollment of students in high school.

This study's novelty lies in applying the _IABE maturity model to an unprecedented and unique dataset. Additionally, the objective is to formulate a series of wh-questions that reveal novel insights from the dataset, diverging from the prevalent approach of single-focused research questions. Utilizing numerous wh-questions can offer a holistic view of the student's modeling.

9- Conclusion

Several maturity models study the Business Intelligence (BI) and Analytics (BA) domains [1]. Our first goal is to find a maturity model in the data science domain, including recent business experimentation (BE), approaches [3], and new causality visions [2, 33]. This study reuses the maturity model named _IABE, the Intelligence, Analytics, and Business Experimentation acronym [5]. The most common frameworks in organizational maturity models use wh-questions or rubrics. GAAM [27] is one of the most well-known wh-questions frameworks, and the Delta Plus model represents the rubric framework [28]. The _IABE model combines two approaches and is applied to a single dataset in this work.

In order to retrieve valuable answers, especially from complex systems, the right questions should be asked. The strategic sequence of questions directs the exploration and analysis process. Furthermore, addressing these questions requires the use of suitable tools. Ultimately, only relevant data patterns leading to new insights are selected and represented in the DIKW (data, information, knowledge, and wisdom) Pyramid. Within the ModEst project, the _IABE model defined the paths from data to insights through various stages of work.

For the first time, this work elucidated the _IABE maturity model by employing comprehensive wh-questions focused on the DGEEC (Directorate-General for Statistics of Education and Science) dataset. This article contributes to the knowledge of the students' modeling domain. The ModEst study results have contributed to a more holistic comprehension of the education system. Moreover, the wh-questions contribute to finding relevant insights illustrating the DIKW Pyramid's functioning in the ModEst project.

The ModEst project used a real dataset from DGEEC, and concrete wh-questions are asked guided by the _IABE model. The new knowledge extracted is crucial for making the right policy choices, leading to much-needed improvements, and ensuring education is fair, available to all, and of good quality. Therefore, this work motivates the academic community to engage in practical and actionable studies.

10- Declarations

10-1- Author Contributions

Conceptualization, L.Ca., P.P., and L.Co.; methodology, L.Ca.; software, P.P.; validation, L.Ca., P.P., and L.Co.; formal analysis, L.Ca.; data curation, P.P.; writing—original draft preparation, L.Ca.; writing—review and editing, P.P. and L.Co.; visualization, P.P.; supervision, L.Co.; project administration, L.Co.; funding acquisition L.Co. All authors have read and agreed to the published version of the manuscript.

10-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

10-3- Funding

The authors would like to acknowledge the LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020, and the support of ModEst project, DSAIPA/DS/0039/2018, FCT, Portugal.

10-4- Acknowledgement

The authors would like to thank the support of the Directorate-General for Statistics of Education and Science (DGEEC) of the Portuguese Education Ministry.

10-5- Institutional Review Board Statement

Not applicable.

10-6- Informed Consent Statement

Not applicable.

10-7- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

11- References

[1] Carvalho, J. V., Rocha, Á., Vasconcelos, J., & Abreu, A. (2019). A health data analytics maturity model for hospitals information systems. International Journal of Information Management, 46, 278–285. doi:10.1016/j.ijinfomgt.2018.07.001.

[2] Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. Communications of the ACM, 62(3), 54–60. doi:10.1145/3241036.

[3] Thomke, S. H. (2020). Experimentation works: The surprising power of business experiments. Harvard Business Press, Boston, United States.

[4] Pearl, J., & Mackenzie, D. (2018). The book of why: the new science of cause and effect. Basic Books, New York, United States.

[5] Cavique, L. (2023). Causality: The Next Step in Artificial Intelligence. In Philosophy of Artificial Intelligence and Its Place in Society, IGI Global, 1-17. doi:10.4018/978-1-6684-9591-9.ch001.

[6] Jackson R. (1999). Information Design. MIT Press, Cambridge, United States.

[7] Ackoff, R. L. (1989). From data to wisdom. Journal of Applied Systems Analysis, 16(1), 3-9.

[8] Cavique, L., Pombinho, P., Tallón-Ballesteros, A. J., & Correia, L. (2020). Data Pre-processing and Data Generation in the Student Flow Case Study. Intelligent Data Engineering and Automated Learning – IDEAL 2020, IDEAL 2020, Lecture Notes in Computer Science, 12490. Springer, Cham, Switzerland. doi:10.1007/978-3-030-62365-4_4.

[9] Pombinho, P., Cavique, L., & Correia, L. (2023). Errors of Identifiers in Anonymous Databases: Impact on Data Quality. 17th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2022), SOCO 2022, Lecture Notes in Networks and Systems, 531, Springer, Cham, Switzerland. doi:10.1007/978-3-031-18050-7_53.

[10] Tavares, L. V. (1995). On the development of educational policies. European Journal of Operational Research, 82(3), 409–421. doi:10.1016/0377-2217(95)98193-4.

[11] Lovell, C. C. (1971). Student Flow Models: A Review and Conceptualization. Technical Report 25, Preliminary Field Review Edition, National Center for Higher Education Management System at Western Interstate Commission for Higher education, Boulder, United States.

[12] Kwak, N. K., Brown, R., & Schiederjans, M. J. (1986). A Markov analysis of estimating student enrollment transition in a trimester institution. Socio-Economic Planning Sciences, 20(5), 311–318. doi:10.1016/0038-0121(86)90040-6.

[13] Bessent, E. W., & Bessent, A. M. (1980). Student Flow in a University Department: Results of a Markov Analysis. Interfaces, 10(2), 52–59. doi:10.1287/inte.10.2.52.

[14] Meece, J. L., & Miller, S. D. (2001). A longitudinal analysis of elementary school students' achievement goals in literacy activities. Contemporary Educational Psychology, 26(4), 454–480. doi:10.1006/ceps.2000.1071.

[15] Lima Junior, P., Silveira, F. L. da, & Ostermann, F. (2012). Survival analysis applied to the study of academic flow in undergraduate physics courses: an example from a Brazilian university. Brazilian Journal of Physics Teaching, 34(1), 1-10. doi:10.1590/s1806-11172012000100014. (In Portuguese).

[16] Saltzman, R. M., & Roeder, T. M. (2012). Simulating student flow through a college of business for policy and structural change analysis. Journal of the Operational Research Society, 63(4), 511–523. doi:10.1057/jors.2011.59.

[17] Fiallos, A., & Ochoa, X. (2017). Discrete event simulation for student flow in academic study periods. 2017 Twelfth Latin American Conference on Learning Technologies (LACLO), La Plata, Argentina. doi:10.1109/laclo.2017.8120908.

[18] Nese, J. F., Lai, C. F., & Anderson, D. (2013). A primer on longitudinal data analysis in education. Behavioral Research and Teaching. Technical Report#1320, University of Oregon, Eugene, United States.

[19] Kwok, O.-M., Lai, M. H.-C., Tong, F., Lara-Alecio, R., Irby, B., Yoon, M., & Yeh, Y.-C. (2018). Analyzing Complex Longitudinal Data in Educational Research: A Demonstration with Project English Language and Literacy Acquisition (ELLA) Data Using XXM. Frontiers in Psychology, 9. doi:10.3389/fpsyg.2018.00790.

[20] Victorino, G., Coelho, P. S., & Henriques, R. (2023). The Value of Design Thinking for PhD Students: A Retrospective Longitudinal Study. Emerging Science Journal, 7, 16–31. doi:10.28991/ESJ-2023-SIED2-02.

[21] Siemens, G., & Baker, R. S. J. d. (2012). Learning analytics and educational data mining. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 252-254. doi:10.1145/2330601.2330661.

[22] Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12–27. doi:10.1002/widm.1075.

[23] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1), 11. doi:10.1186/s40561-022-00192-z.

[24] Nunes, C., Beatriz-Afonso, A., Cruz-Jesus, F., Oliveira, T., & Castelli, M. (2022). Mathematics and Mother Tongue Academic Achievement: A Machine Learning Approach. Emerging Science Journal, 6, 137–149. doi:10.28991/esj-2022-sied-010.

[25] Costa-Mendes, R., Cruz-Jesus, F., Oliveira, T., & Castelli, M. (2022). Deep Learning in Predicting High School Grades: A Quantum Space of Representation. Emerging Science Journal, 6, 166–187. doi:10.28991/esj-2022-sied-012.

[26] Feng, G., Fan, M., & Chen, Y. (2022). Analysis and Prediction of Students' Academic Performance Based on Educational Data Mining. IEEE Access, 10, 19558–19571. doi:10.1109/ACCESS.2022.3151652.

[27] Gartner. (2012). Gartner Analytic Ascendancy Model. Gartner, Inc., Stamford, United States. Available online: https://www.gartner.com/en (accessed on May 2023).

[28] Davenport, T. (2018). DELTA Plus Model & five stages of analytics maturity: A primer. International Institute for Analytics, Portland, United Sates.

[29] ISCED. (2011). International Standard Classification of Education. UNESCO Institute for Statistics, Montreal, Quebec, Canada.

[30] Hanif, A., Zhang, X., & Wood, S. (2021). A Survey on Explainable Artificial Intelligence Techniques and Challenges. 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), Gold Coast, Australia. doi:10.1109/edocw52865.2021.00036.

[31] Lewis, D. (1973). Causation. The Journal of Philosophy, 70(17), 556. doi:10.2307/2025310.

[32] Pearl, J. (2000). Models, reasoning and inference. Cambridge University Press, Cambridge, United Kingdom.

[33] Angrist, J. D., & Pischke, J. S. (2014). Mastering 'metrics: The path from cause to effect. Princeton University Press, Princeton, United States. doi:10.1093/erae/jbv011.

[34] Crato, N., & Paruolo, P. (2018). Data-driven policy impact evaluation: How access to microdata is transforming policy design. Springer, Cham, Switzerland. doi:10.1007/978-3-319-78461-8.