Exploiting LSTM Neural Network Algorithm Potentiality for Early Identification of Delayed Graduation in Higher Education

THEODOROS ANAGNOSTOPOULOS[1], DIMITRIS PAPAKYRIAKOPOULOS[1], YANNIS PSAROMILIGKOS[1], SYMEON RETALIS[2]
Department of Business Administration,
University of West Attica,
250 Thivon & P. Ralli Str, Egaleo, 12241 Athens,
GREECE

[2]Department of Digital Systems,
University of Piraeus,
150 Androutsou Str, 18532 Piraeus,
GREECE

Abstract: - Adoption of deep learning classification algorithms in the domain area of higher education provides exploratory predictive data analytics able to exploit students' academic behavior. Concretely, student retention and success are critical concerns in higher education globally. Timely identification of potential delays in graduation is essential for universities to provide effective interventions and support, ensuring students' progress efficiently and maintaining high graduation rates, thereby enhancing institutional reputation. This study examines data from a typical computer science department of a central Greek university, covering student performance for almost two decades (1999-2018). Through extended data preprocessing, we developed a robust dataset focusing on key courses indicative of students' likelihood to graduate on time or experience delays. We employed a deep learning Long Short-Term Memory (LSTM) Neural Network algorithm, leveraging this dataset to classify and predict students' final academic outcomes. Our findings reveal that early-semester performance data can successfully forecast graduation timelines, enabling proactive educational strategies to support student success during their studies at the university.

Key-Words: - Deep Learning LSTM Neural Network Algorithm, Data Preprocessing, Predictive Data Analytics, Binary Classification, Evaluation Method and Metrics, Early Identification, Delayed Graduation in Higher Education.

Received: March 7, 2024. Revised: September 3, 2024. Accepted: October 6, 2024. Published: November 8, 2024.

1 Introduction

The complex phenomenon of a timely or a delayed graduation as well as university dropout has garnered significant attention within higher education research communities due to its profound impact on students' lives and societal development, [1]. Timely graduation refers to the ability of students to complete their degrees within an expected duration of study, while delayed graduation refers to the prolonged time taken by students to complete their academic programs beyond the expected duration. Concretely, dropout denotes the premature withdrawal of students from their educational endeavors before achieving their intended qualifications, [2]. These phenomena not only impact individual students but also have broader implications for society, affecting workforce readiness, economic productivity, and social mobility.

According to the United States of America (USA) National Center for Education Statistics, [3], [4], only a small percentage of students 64% who began seeking a bachelor's degree at a 4-year institution in the fall 2014 completed that degree at the same institution within 6 years which is comparable to rates seen in other countries worldwide, [5]. Subsequently, spanning time of study in higher educational institutes reveals that in the case of a 6-year studying process, the graduation rate was higher for females than for males (67% vs. 60%, respectively).

Concretely, in the European Union (EU), [6], [7], a study composed of 14 countries out of the 35 bachelor completion rates reported a range from 53% to 83%. However, these numbers should be interpreted with caution due to the different evaluation methods, which were used to calculate bachelor completion rates in European higher education institutes. Intuitively, the unique national context of each European country needs to be taken into consideration when comparing educational indicators, which are applied in countries with different education cultures.

In Greece, according to data from the Hellenic Authority for Higher Education, [8], only one in five students (19.62%) of Greek universities graduate on time, i.e. within the minimum study time of their school. It takes an average of six years for a student to complete a four-year course of study, 7.5 years for five-year courses, and nine years for medical studies, which are six-year courses. In fact, the phenomenon of late or no graduation has worsened greatly in the last decade after the economic crisis. Observed data indicate that graduates have been declining continuously since then from 10.04% of all enrolled students in 2013 (9.91% in 2015, 9.41% in 2017, 8.91% in 2019, and 8.60% in 2020). Compared to the EU, the corresponding average for EU countries in 2020 was 23.15%, up from 22.13% in 2018 and 22.28% in 2019.

In this paper, we conducted a research study in the Department of Digital Systems at the University of Piraeus, Greece to predict a timely or delayed student's graduation. Data sources incorporated include detailed information on students' efforts during the fall and the spring semesters in an extended total period of time containing two decades of university studies. The examined data cover educational experiences over a wide range of periods from 1999 to 2018. Extensive data pre-processing was performed on the initial data sources to produce a detailed derived dataset containing meaningful information.

Intuitively, the resulting derived dataset comprises specific courses that describe students' educational behavior from their first semesters of study. This educational behavior provides insights into students' likelihood of achieving a timely or delayed graduation. To assess the potential of predictive analytics, a Long Short-Term Memory (LSTM) Neural Network deep learning algorithm is employed to perform classification and prediction based on the derived dataset. Concretely, the adopted LSTM Neural Network algorithm successfully classifies students' final learning performance and efficiently predicts whether a student will graduate on time or experience delays. These results are primarily based on data from the first semesters of study in the examined university.

The rest of the paper is organized as follows. In Section 2 it is presented the prior work in the research effort area. Section 3 defines the adopted data model. In Section 4 evaluation parameters are defined. In Section 5 experiments are performed and results are observed. Section 6 discusses the strengths and the weaknesses of the proposed research effort, while Section 7 concludes the paper and proposes future work.

2 Prior Work

Delayed graduation and overeducation are studied in research efforts, [9], where the focus is given to the social impact of delayed graduation and overeducation on salaries observed after graduation. Research on delayed time to get a degree with regards to post-graduation earnings, [10], correlates the graduation process along with capital spent by students' families to achieve graduation. Degree completion prediction is analyzed, [11], where special focus is given to a timely and a delayed graduation while examining dropout possibilities within a certain time horizon. An ensemble prediction model is elaborated, [12], to define students' timely and/or delayed graduation assessing the potentiality of students' learning behavior enhanced with university domain knowledge. Data analytics uplift modeling, [13], is used to model a timely or delayed graduation while simultaneously preventing a possible student dropout from the university.

Machine learning explainable modeling technology is incorporated, [14], to predict student-delayed graduation as well as prevent student dropout. Students' educational behavior is modeled as a timestamped academic trajectory, [15], which is able to predict proactively a timely and/or delayed graduation during university studies. Learning analytics can be provided, [16], to understand students' academic behavior thus explaining delayed graduation by assessing the potentiality of a randomized control experiment. A data mining research approach is incorporated, [17], which exploits the ensemble classification algorithm's capabilities to proactively predict students' dropout and delayed graduation behavior. An evaluation method is proposed, [18], which aims to explain the academic dropout rate with regard to labor market conditions and salaries gained from dropout university students.

The university graduation process is linked with the salary earned by graduated students, [19], to understand the conditions leading to a timely or delayed graduation behavior in academia. Social network parameters enhance the capabilities of a pattern-augmented algorithm, [20], which is able to accurately measure timely and/or delayed student graduation process as well as a possible dropout from universities. Delayed graduation and dropout behavior are examined, [21], with regard to success or failure rates in core and direction courses during contemporary studies in higher education. The impact of financial support on students from their families, [22], is analyzed to specify the reasons for timely and/or delayed graduation, which affect their observed wages in the labor marketplace. Predicting accurately the timely graduation of students, [23], is possible by exploiting the impact of crowdfunding and institutional accompaniment on the continuous educational process within the university.

Understanding the reasons for delayed graduation and university dropout resulted in the creation of degree roadmaps, [24], which used to provide supportive educational services for timely graduation. The discovery of potential reasons resulting in delayed graduation and dropout is examined, [25], to propose countermeasures, which aim to eliminate in large scale such inefficient academic behavior. Social networking analysis from data belonging to student university networks, [26], is a promising method to predict proactively observed delayed graduation and dropout in universities. Timely graduation is able to be predicted accurately based on input information provided by academia [27], thus enabling the manipulation and measurement of student activities within the universities.

Research efforts examined in the literature face timely and/or delayed graduation by using data sources provided by the universities. Such data are generated by core or discipline courses offered by higher education institutes. Several machine learning algorithms are incorporated to perform predictive analytics to identify students' further activities during their studies. Subsequently, there is also examined the relation between a timely or delayed graduation as well as students' dropout with regards to the family financial support provided as well as salaries observed by students after their graduation in the open marketplace?

Concretely, data preprocessing in the majority of the research efforts is rather straightforward thus not exploiting the rich information inherent within the provided data sources. Intuitively, in most of the cases, data stem from a limited time horizon of studies in higher education, which deteriorates incorporated algorithms' prediction accuracy since the more the stochastic data provided to inference models the more the prediction accuracy observed. Subsequently, in the observed research efforts focus is given mainly on predicting a timely or a delayed graduation for the total number of semesters of a university, thus underestimating the potentiality of a specific department's domain knowledge, which can provide accurate prediction results from the first semesters of study.

Subsequently, such inefficiencies are faced by current research studies, which are able to understand students' behavior and predict proactively a timely or delayed graduation based on specific domain knowledge observed by the first semesters of study in the university. Provided data sources cover a stochastic time horizon of almost twenty years of students' activity. This piece of information is exploited by assessing the potentiality of an LSTM Neural Network deep learning algorithm, which is used to perform classification and prediction based on a meaningful derived data source of students' behavior in the Department of Digital Systems at the University of Piraeus, Greece.

3 Data Model

Data provided for analysis by the university department needs an extensive data preprocessing phase. Such data cleaning and transformation is necessary to derive a data source meaningful for performing valuable data analytics by the adopted LSTM Neural Network deep learning algorithm.

3.1 Provided Data Source

The provided data source has a time span of almost twenty years of undergraduate student activity in the Department of Digital Systems at the University of Piraeus from 1999 to 2018. The initial number of students examined was 1278. The personal data of students are preserved by the department, which has anonymized their sensitive data (i.e., name and age) with certain identifiers provided by incorporating a hash function. The provided data source has information for both core and discipline courses, while the number of courses is 42. For each type of course, the dataset includes three discrete data quantities: (1) the number of times a student has attempted the course [1-8], (2) the highest grade achieved [0-10], and (3) the number of years the course has been taken [1-12]. Total number of predictive attributes for the provided data source is calculated by multiplying the number of core and discipline courses with observed attribute quantities stored for each course, which results in 126 predictive attributes. The data source has a class attribute, which denotes if a student has timely or delayed graduation. Specifically, Class 1 denotes students with delayed graduation, while Class 2 denotes students with timely graduation. It is observed that for the provided data source delayed graduation students (i.e., Class 1) are 386, while the number of timely graduation students (i.e., Class 2) is 892.

3.2 Derived Data Source

Since the time span of studies is wide it is observed that undergraduate courses have changed through time either their names or their context. To avoid overlaps focus is given to core courses, thus omitting discipline courses. A number of core courses is 15 in the undergraduate program of studies spanning from the first to the last semester of the undergraduate program. Feature selection is applied to the observed data courses and it is found that only a subset of them was significant for further experimentation, that is only 5 core courses, [28]. Concretely, such courses span only in the first four semesters of study, that is a timely or a delayed graduation could be predicted from the first two years of study thus enabling proactive recovery of students that will face some problems in finishing their studies. Specifically, these courses are: (1) Mathematical Analysis & elements of Linear Algebra taught in the first semester, (2) advanced mathematical Analysis taught in the second semester, (3) discrete mathematics taught in the second semester, (4) introduction to telecommunications taught in the third semester, and (5) algorithms & complexity taught in the fourth semester of studies. Intuitively, it is derived that only the maximum number of grades information observed in the interval $[0, 10]$ is significant for performing data analytics with the adopted LSTM Neural Network. Subsequently, grades aggregation leads to better predictive results, thus initial grades were transformed according tothe following rules: (1) grade equals to 0 for grade interval between $[0, 4.9]$, (2) grade equals to 1 for grade interval between $[5, 6.9]$, (3) grade equals to 2 for grade interval between $[7, 8.9]$, and (4) grade equals to 3 for grade interval between $[9, 10]$. After data preprocessing it is observed that the final number of students prepared for experimentation is 674. Intuitively, derived delayed graduation students (i.e., Class 1) are 257, while number of timely graduation students (i.e., Class 2) is 417.

Concretely, derived data source class instances are visualized as presented in Figure 1. Intuitively, Class 1 instances data distribution (i.e., delayed graduation students) is depicted in the bottom left corner with blue 'x' marks. Subsequently, Class 2 instances data distribution (i.e., timely graduation students) is depicted in the upper right corner of the plot with red 'x' marks.

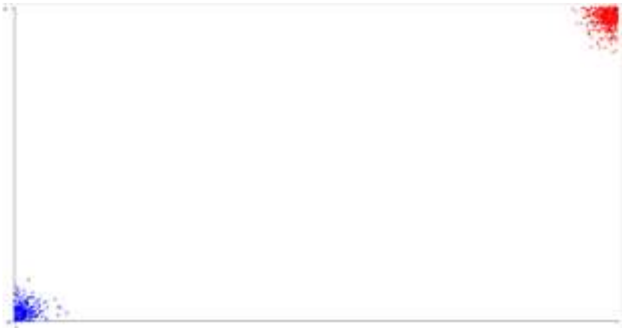

Fig. 1: Class instances of derived data source

4 Evaluation Parameters

Assessing the performance of the adopted LSTM Neural Network deep learning algorithm, certain valuation methods and evaluation metrics should be incorporated to perform specific experiments and observe derived results.

4.1 Evaluation Method

To evaluate the adopted LSTM Neural Network deep learning algorithm there are used certain evaluation methods. Authors adopt one of the widely used evaluation methods, due to its simplicity and optimum results, which is 10-fold cross-validation. Specifically, such an evaluation method divides the input dataset into 10 equal-sized parts and then in a certain loop incorporates the first 9 parts to train the LSTM Neural Network classification algorithm and the remaining 1 to test the classifier. This process is repeated until all the parts are used for training and testing.

4.2 Evaluation Metrics

Given the evaluation method, which is proposed to support the experimental setup there is a need to adopt specific evaluation metrics. Such metrics are: (1) prediction accuracy, (2) correctly classified instances, and (3) confusion matrix that are able to assess the efficiency of a deep learning classification algorithm, such as the adopted LSTM Neural Network algorithm.

4.2.1 Prediction Accuracy

The effectiveness of the adopted LSTM Neural Network algorithm is assessed by incorporating prediction accuracy evaluation metric, $a \in [0, 1]$, which is defined in the following mathematical equation, (1):

$$a = \frac{tr_{pos}+tr_{neg}}{tr_{pos}+fl_{pos}+tr_{neg}+fl_{neg}} \quad (1)$$

Where, tr_{pos}, are the instances, which are classified correct as positives, and, tr_{neg}, are the instances, which are classified correct as negatives. In addition, fl_{pos}, are the instances, which are classified false are positives, and, fl_{neg}, are the instances, which are classified false as negatives. A low value of a means a weak classifier while a high value of a indicates an efficient deep learning classifier.

4.2.2 Correctly Classified Instances

In deep learning literacy, such as in the adopted LSTM Neural Network algorithm, it is common to express prediction accuracy as a percentage thus observed results being more easily interpreted and presented. Concretely, it is used the term correctly classified instances, $c \in [0\%, 100\%]$, which is defined according to the following mathematical equation, (2):

$$c = \alpha\% \quad (2)$$

Where, a value close to 0% means that the classification algorithm is not efficient, while a value close to 100% indicates that the deep learning algorithm is able to classify instances optimally.

4.2.3 Confusion Matrix

We also evaluated the adopted deep learning LSTM Neural Network classification algorithm with the confusion matrix evaluation metric. A confusion matrix is a special form of data matrix, which in the case of a binary classification of 2 classes, (i.e., Class 1: denoting the categorical class data value of delayed graduation students, and Class 2: denoting the categorical class data value of timely graduation students) has the following encoded form, as described in Table 1.

Table 1. Confusion matrix evaluation metric

Class 1	**Class 2**	← Classified as
A	B	**Class 1**
C	**D**	**Class 2**

Where, "A" quantity depicts the number of Class 1 instances, which are classified correctly as instances of Class 1. "B" quantity depicts the number of Class 1 instances, which are falsely classified as instances of Class 2. "C" quantity depicts the number of Class 2 instances, which are falsely classified as instances of Class 1. "D" quantity depicts the number of Class 2 instances, which are correctly classified as instances of Class 2. It holds that "A", "B", "C", and "D" are denoting certain numerical data values, which are observed during the evaluation process.

A given classification model, such as the adopted deep learning LSTM Neural Network algorithm, is considered efficient if it maximizes the numerical data elements of the main diagonal of the confusion matrix (i.e., "A", and "D" should have high numerical data values) and also minimizes the other numeric data elements of the matrix (i.e., "B", and "C" should have low numerical data values).

5 Experiments and Results

Experiments are performed by the adopted LSTM Neural Network deep learning algorithm, which is evaluated based on certain evaluation method and metrics. Results observed have an impact on the potentiality of the selected deep learning model for the stated problem of the current research effort.

Table 2. LSTM Neural Network tuning parameters

Algorithm tuning parameter	**Value**
Input layer	1 node
Number of hidden layers	1
Hidden layer	2 nodes
Output layers	2 nodes
Hidden layer activation function	ReLu
Output layer activation function	SoftMax

5.1 Experimental Setup

An LSTM Neural Network deep learning algorithm is incorporated to perform predictive analytics based on the derived data source. Since there exist two classes, namely, Class 1 of delayed graduation students and Class 2 of timely graduation students this is a binary class classification problem. The adopted inference model is provided by Weka machine learning software, [29]. However, to observe optimal results there is a need to fine-tune the provided algorithm according to certain parameters' values. Specifically, the adopted model requires one input layer of 1 node and one hidden layer. The hidden layer is composed of 2 nodes, while the output layer is also composed of 2 nodes (i.e., one node for each of the two classes). Intuitively, the hidden layer activation function is ReLu, while the output layer activation function is SoftMax. Tuning model parameters along with their values are described in Table 2.

5.2 Derived Results

A fine-tuned LSTM Neural Network learning algorithm is then used for experimentation, which provides certain results.

5.2.1 Observed Prediction Accuracy

The evaluation method incorporated to evaluate the adopted LSTM Neural Network binary class classification algorithm is 10-fold cross-validation. According to this evaluation method observed prediction accuracy is: $a = 0.781899$, which is a high value for prediction accuracy thus proving that the adopted machine learning algorithm is suitable for the examined binary class classification problem.

5.2.2 Observed Correctly Classified Instances

According to the evaluation method of 10-fold cross-validation correctly classified instances it occurred to be: $c = 78.18\%$, which indicated that the selected LSTM Neural Network algorithm is an optimal choice for the examined classification problem.

5.2.3 Observed Confusion Matrix

Confusion matrix results are derived based on the 10-fold validation evaluation method for the examined binary class classification problem. Derived results are presented in Table 3.

Table 3. Confusion matrix of observed results

Class 1	Class 2	← Classified as
110	147	**Class 1**
0	**417**	**Class 2**

It can be observed that most of the classified instances are located in the main diagonal of Table 3. Specifically, the quantity of elements in the main diagonal depicts the significant number of certain instances, which are correctly classified. Concretely, such an optimal prediction behavior indicates a robust deep learning LSTM Neural Network classification algorithm for the examined binary class classification problem.

Intuitively, the adopted LSTM Neural Network algorithm has certain classification errors, which affects its prediction accuracy. Such classification errors can be visualized as presented in Figure 2. Subsequently, it can be observed that Class 1 error instances (i.e., depicted with blue 'squares') have been plotted within the Class 2 correct instances area (i.e., depicted with red 'x' marks) in the upper right corner of the visualized classification results. In the bottom left corner, it can be also observed the Class 1 correct instances area (i.e., depicted with blue 'x' marks).

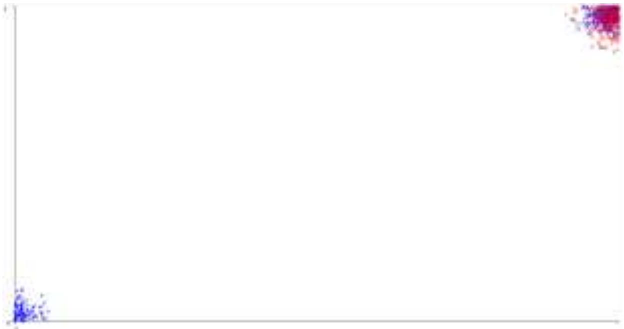

Fig. 2: Visualization of classification errors

6 Discussion

The current research effort has specific strengths and certain weaknesses, which are analyzed to assess qualitatively the potentiality of the study.

6.1 Strengths of the Study

The provided data source contains information on almost twenty years of students' activity at the Department of Digital Systems at the University of Piraeus, Greece. Concretely, a data preprocessing phase is required to mine rich context, which in turn can be used to further perform predictive analytics. During this phase, there were separated cores from discipline courses. Research focuses on core courses that are fundamental for determining a timely or a delayed graduation. Intuitively, courses of only the first four semesters of studies are capable of providing proactive predictions on students' graduation. Subsequently, an LSTM Neural Network deep learning algorithm is incorporated to predict sufficiently a timely or a delayed graduation of certain students contained in the derived data source

6.2 Weaknesses of the Study

Discontinuity of undergraduate programs' courses during the time period span of almost twenty years affects the quality of the derived data source. Specifically, rich information is lost during the data preprocessing phase, which leads to the deletion of more precise information. However, such inefficiencies are part of the provided data source and are inherent to the initial data source, thus it could not be feasible to exploit further their potentiality. In addition, although prediction accuracy is high and Class 2 has been successfully predicted there is not the same for Class 1 prediction. This is explained due to the fact that Class 1 students either have decided to follow a delayed graduation behavior or are in the process of actively performing their studies at a more relaxed pace, which affects the observed prediction accuracy.

7 Conclusions and Future Work

Defining a timely or a delayed graduation of the students population is of significant importance for a university since such knowledge has an impact on academic ratings worldwide. Concretely, proactive prediction of students' behavior can affect the countermeasures incorporated by the university to help students with their activities to have an interesting and valuable daily activity. Intuitively, timely graduation can positively affect the amount of salary gained on the open market, while delayed graduation might lead to low wages.

In this research effort, we exploit the potentiality of students' data from the Department of Digital Systems at the University of the Piraeus, Greece to proactively predict a timely or a delayed graduation based mainly on core courses of the first four semesters of study. Future work will be towards the direction of being able to proactively infer students' academic behavior based on more detailed information existing in the undergraduate programs such as the discipline courses thus adding more knowledge to the adopted LSTM Neural Network deep learning algorithm, which is incorporated in this research study.

Acknowledgement:
The authors would like to thank the Department of Digital Systems at the University of Piraeus, Greece for providing the initial data on students' performance during their course of study in the university, which are further exploited by the current research effort.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the authors used classification and visualization services of the Weka AI-enabled open source software in order to analyse provided data sources and exploit their visual potentiality. After using the services of the AI-enabled Weka tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

[1] What Matters to Students Success: A Review of the Literature, *National Postsecondary Education Cooperative (NPEC)*, July 2006, [Online]. https://nces.ed.gov/npec/pdf/kuh_team_report.pdf (Accessed Date: February 12, 2024).

[2] B. Castleman, and K. Meyer, "Financial Constraints & Collegiate Student Learning: A Behavioral Economics Perspective", *Daedalus*, vol. 148, no. 4, 2019, pp. 195–216, DOI: https://doi.org/10.1162/daed_a_01767.

[3] Report of the Condition of Education 2022, *National Center for Education Statistics (NCES)*, [Online]. https://nces.ed.gov/pubs2022/2022144.pdf (Accessed Date: February 12, 2024).

[4] Impact of the Coronavirus Pandemic on Fall Plans for Postsecondary Education, *National Center of Education Statistics (NCES)*, [Online]. https://nces.ed.gov/programs/coe/indicator/tpb/covid-impact-postsecondary-plans (Accessed Date: February 12, 2024).

[5] T. Pascarella, and P. Terenzin, "How College Affects Students, A Third decade of Research", *Journal of Student Affairs in Africa*, vol. 2, no. 2, 2014, pp. 47–50, DOI: https://doi.org/10.14426/jsaa.v2i2.70.

[6] Dropout and completion in higher education in Europe, *Publications Office of the European Union (POEU)*, [Online]. https://op.europa.eu/en/publication-detail/-/publication/4deeefb5-0dcd-11e6-ba9a-01aa75ed71a1/language-en (Accessed Date: February 12, 2024).

[7] Delayed Graduation and University Dropout: A review of Theoretical Approaches, *Institute of Labor Economics (IZA)*, [Online]. https://www.iza.org/publications/dp/12601/delayed-graduation-and-university-dropout-a-review-of-theoretical-approaches (Accessed Date: February 12, 2024).

[8] Quality Assurance, *Hellenic Authority for Higher Education (HAHE)*, [Online]. https://www.ethaae.gr/en/quality-assurance (Accessed Date: February 12, 2024).

[9] C. Aina, and F. Pastore, "Delayed Graduation and Overeducation in Italy: A Test of the Human Capital Model Versus the Screening Hypothesis", *Social Indicators Research*, vol. 152, 2020, pp. 533–553, DOI: https://doi.org/10.1007/s11205-020-02446-0.

[10] D. Witteveen, and P. Attewell, "Delayed Time-to-Degree and Post-college Earnings",

Research in Higher Education, vol. 62, 2021, pp. 230–257, DOI: https://doi.org/10.1007/s11162-019-09582-8.

[11] E. Demeter, M. Dorodchi, E. A. Hossami, A. Benedict, L. S. Walker, and J. Smail, "Predicting first-time-in-college students' degree completion outcomes", *Higher Education*, vol. 84, 2022, pp. 589–609, DOI: https://doi.org/10.1007/s10734-021-00790-9.

[12] A. A. Priyambada, T. Usagawa, and M. Er, "Two-layer ensemble prediction of students' performance using learning behavior and domain knowledge", *Computers and Education: Artificial Intelligence*, vol. 5, 2023, pp. 1–12, DOI: https://doi.org/10.1016/j.caeai.2023.100149.

[13] D. Olaya, J. Vasquez, S. Maldonado, J. Miranda, and W. Verbeke, "Uplift Modeling for preventing student dropout in higher Education", *Decision Support Systems*, vol. 134, 2020, pp. 1–11, DOI: https://doi.org/10.1016/j.dss.2020.113320.

[14] J. G. C. Kruger, A. S. Britto, and J. P. Barddal, "An explainable machine learning approach for student dropout prediction", Expert Systems with Applications, vol. 233, 2023, pp. 1–9, DOI: https://doi.org/10.1016/j.eswa.2023.120933.

[15] M. F. Musso, C. F. R. Hernandez, and E. C. Cascallar, "Predicting key educational outcomes in academic trajectories: a machine-learning approach", *Higher Education*, vol. 80, 2020, pp. 875–894, DOI: 10.1007/s10734-020-00520-7.

[16] J. Hellings, and C. Haelermans, "The effect of providing learning analytics on student behavior and performance in programming: a randomized controlled experiment", *Higher Education*, vol. 83, 2022, pp. 1–18, DOI: https://doi.org/10.1007/s10734-020-00560-z.

[17] K. Okoye, J. T. Nganji, J. Escamilla, and S. Hosseini, "Machine learning model (RG-DMML) and ensemble algorithm for prediction of students' retention and graduation in education", *Computers and Education: Artificial Intelligence*, vol. 6, 2024, pp. 1–13, DOI: https://doi.org/10.1016/j.caeai.2024.100205.

[18] D. Mikola, and M. D. Webb, "Finish it and it is free: An evaluation of college graduation subsidies", *Economics of Education Review*, vol. 93, 2023, pp. 1–18, DOI: https://doi.org/10.1016/j.econedurev.2023.102355.

[19] L. Finamor, "Labor market conditions and college graduation: Evidence from Brazil", *Economics of Education Review*, vol. 94, 2023, pp. 1–13, DOI: https://doi.org/10.1016/j.econedurev.2023.102403.

[20] B. S. Assis, E. Ogasawara, R. Barbastefano, and D, Carvalho, "Frequent pattern mining augmented by social network parameters for measuring graduation and dropout time factors: A case study on a production engineering course", *Socio-Economic Planning Sciences*, vol. 81, 2022, pp. 1–12, DOI: https://doi.org/10.1016/j.seps.2021.101200.

[21] C. S. Sargent, T. Sullivan, and H. M. Alum, "DFW in gateway courses not always a graduation problem: A study in Intermediate Accounting I from 2007 to 2018", *Journal of Accounting Education*, vol. 60, 2022, pp. 1–7, DOI: https://doi.org/10.1016/j.jaccedu.2022.100795

[22] V. Rattini, "The effects of financial aid on graduation and labor market outcomes: New evidence from matched education-labor data", *Economics of Education*, vol. 96, 2023, pp. 1–15, DOI: https://doi.org/10.1016/j.econedurev.2023.102444.

[23] C. A. Silvera, J. C. Nino, J. D. Manotas, N. S. Quintero, and Z. H. Fontalvo, "Model based in Crowdfunding and institutional accompaniment for the timely graduation in undergraduate students", *Procedia Computer Science*, vol. 220, 2023, pp. 991 – 997, DOI: https://doi.org/10.1016/j.procs.2023.03.137.

[24] X. Su, M. Chen, J. Y. Austin, and Y. Liu, "Restructuring degree roadmaps to improve timely graduation in higher education", *International Journal of Educational Management*, vol. 34, no. 2, 2020, pp. 432-449, DOI: 10.1108/IJEM-07-2019-0257.

[25] V. Bocsi, T. Cegledi, Z. Kocsis, K. E. Kovacs, K. Kovacs, A. Muller, K. Pallay, B. E. Szabo, F. Szigeti, and D. A. Toth, "The discovery of the possible reasons for delayed graduation and dropout in the light of a qualitative research study", *Journal of Adult Learning, Knowledge and Innovation*, vol. 3, no. 1, 2019, pp. 27–38, DOI: https://doi.org/10.1556/2059.02.2018.08.

[26] N. Nur, N. Park, M. Dorodchi, W. Dou, M. J. Mahzoon, X. Niu, and M. L. Maher, "Student Network Analysis: A Novel Way to Predict Delayed Graduation in Higher Education",

Lecture Notes in Computer Science (LNAI), vol. 11625, 2019, pp. 370–382, DOI: https://doi.org/10.1007/978-3-030-23204-7_31.

[27] N. L. Wade, "Measuring, Manipulating, and Predicting Student Success: A 10-Year Assessment of Carnegie RI Doctoral Universities Between 2004 and 2013", *Journal of College Student Retention: Research, Theory & Practice*, vol. 21, no. 1, 2019, pp. 119–141, DOI: https://doi.org/10.1177/1521025119831456.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update", *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, 2009, DOI: https://doi.org/10.1145/1656274.1656278.

[29] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition, 2017, DOI: https://doi.org/10.1016/C2015-0-02071-8.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research work was supported in part by the University of Piraeus Teaching and Learning Center (TLC), Greece under Grant C.947.

Conflict of Interest

The authors have no conflicts of interest to declare.