Prediction of students' graduation time using a two-level classification algorithm

Vassilis Tampakas*, Ioannis E. Livieris*, Emmanuel Pintelas†, Nikos Karacapilidis‡ and Panagiotis Pintelas§
*Department of Computer & Informatics Engineering (DISK Lab),
Technological Educational Institute of Western Greece, Greece, GR 263-34.
Email: vtampakas@teimes.gr, livieris@teiwest.gr
†Department of Electrical and Computer Engineering,
University of Patras, Greece, GR 265-00. Email: ece6835@upnet.gr.
‡Department of Mechanical Engineering and Aeronautics,
University of Patras, Greece, GR 265-00. Email: karacap@upatras.gr
§Department of Mathematics, University of Patras,
Greece, GR 265-00. Email: ppintelas@gmail.com

***Abstract*—During the last decades, higher educational institutes have managed to accumulate a large volume of data about their students' characteristics and performance. Machine learning techniques offer a first step and a helping hand in extracting useful information from these data and gaining insights into the prediction of students' progress and performance. In this work, we present a two-level classification algorithm for predicting students' graduation time. The proposed algorithm has two major features. Firstly, it identifies with high accuracy the students at risk of not completing their studies; secondly, it classifies the students based on their expected graduation time. Our preliminary numerical experiments indicate that the proposed algorithm exhibits reliable predictions based on the students' performance in their courses during the first two years of their studies.**

I. INTRODUCTION

Higher education constitutes a significant and critical factor in human resources development, increasing people's knowledge and competencies and ensuring nations' economic prosperity. The main objective of a higher education institute and one of its biggest challenges is to provide quality education to its students. In 2001, the *National Research Council report* [7] illustrated the immediate need to develop innovative methodologies to assist higher institutes, which will further improve the quality of their studies, facilitate students' timely graduation and limit their dropout. To achieve a higher level of studies' quality, there are three main key aspects: the first two lay emphasis on refining teaching and knowledge acquisition methods, while the third one concerns the development of efficient systems for monitoring students' progress and identifies key aspects of their success.

An enduring issue in higher education is student retention of successful graduation [2]. As the cost of higher education (fees, living expenses, etc) has exploded over the past decade, prolonged graduation time consists a crucial factor in discouraging students ultimately leading them to dropout. In fact, recent studies show that a minority of students have succeeded to complete a four-year bachelor program on time [10, 25, 26]. Some of the causes of slow student progress are the inability to register for required courses, credits lost in transfer and remediation sequences that do not work.

Therefore, a crucial step towards effective intervention is to develop a system which can continuously keep track of students' academic progress and accurately predict their graduation time. The ability to accurately predict students' future performance is considered essential for effectively carrying out necessary pedagogical interventions to ensure students' on-time and satisfactory graduation. By analyzing students' progress, appropriate actions and strategic programs could be well planned in an institution in order to decrease the students' mean graduation time and limiting dropout.

Although the prediction of students performance in a course has been extensively studied in the literature [3, 16–18, 21, 22, 28, 29], the early prediction of their graduation time is a significantly different task which faces new challenges. This is due to many factors: firstly, students attend many courses during their studies, but not all courses are equally informative about their graduation; secondly, students differ in terms of knowledge backgrounds and specializations, as well as in the sequence they choose their courses; finally, the predictions need to be made based not only on the most recent students' accomplishments, but also on the evolution of their progress. Therefore, the application of machine learning techniques is considered essential, offering a first step in extracting useful and novel information from these students' records in order to gain a deeper insight in the prediction of students' progress and performance.

The main objective of this research is to predict students' graduation time, putting emphasis on the identification of students who are likely not to complete their studies within six years or dropout. Specifically, we are dealing with the following main tasks:

- Predict students at risk of not graduating within 6 years of studies.
- Classify the students' based on their graduation time.

Nevertheless, the development of an accurate prediction model is a very attractive and challenging task (see [3, 18, 21, 22] and references therein). The task of predicting students' graduation time becomes very complicated by the expanding volume of data, from the increasing student enrollments and by the continually shifting performance during their studies. Generally, educational datasets are imbalanced, hence standard learning algorithms face inability to detect a pattern based on the correct distribution governing the classes of dataset, frequently exhibiting an unacceptable error rate for the minority classes. Furthermore, the difficulty to distinguish between noise and rare cases is also responsible for poor performance on the minority classes [13–15].

In this paper, we present a model for predicting the years taken for a student to complete a bachelor's degree study. Our prediction model is based on a two-level classification algorithm, which accurately classifies the students based on courses' characteristics, students' demographic information and their performance in courses during the first two years of their studies. The early prediction of students' future progress enables the allocation of proper actions to support them and eliminate problems causing late graduation or dropout.

The remainder of this paper is organized as follows: Section II presents a survey of recent studies concerning the application of data mining in education. Section III presents a detailed description of the data collection and data preparation used in our study and the proposed two-level classification algorithm. Finally, Section IV presents experimental results while Section V sketches concluding remarks and future work directions.

II. Related studies

The development and adoption of machine learning systems for predicting students' performance has gained popularity during the last decades, addressing many issues and problems in the educational domain and providing useful outcomes about the learning process. Numerous research studies have been conducted to predict students' academic performance either to facilitate degree planning or to determine students at risk. Romero and Ventura [21, 22] and Baker and Yacef [3] have provided some extensive reviews of different types of educational systems and how data mining can be successfully applied to each of them. More specifically, they described in detail the most accurate models utilized for the prediction of students' performance available in the literature and summarized the diverse factors that influence the performance of students. Along this line, Peña-Ayala [18] presented a review aiming to preserve and enhance the chronicles of recent educational data mining advances and developments and analyze the outcomes produced by a machine learning approach.

Musso et al. [16] have proposed an artificial neural network approach for predicting general academic performance of university students identifying the differential contribution of participating variables using cognitive and non-cognitive measures of students, together with background information. The results showed that neural networks can achieve higher accuracy rate than traditional methods such as discriminant analysis.

Nagy et al. [17] developed an intelligent student advisory framework to improve the success rate of the first year university stage utilizing machine learning techniques. This framework can be used to provide pieces of consultations to a first year university student to pursue a certain education track where he/she will likely succeed in, aiming to decrease the high rate of academic failure among students. The framework acquires information from the datasets which store the academic achievements of students before enrolling to higher education together with their first year grade after enrolling in a certain department. After acquiring the relevant information, the intelligent system utilizes both classification and clustering techniques to provide recommendations for a certain department for a new student. Additionally, they presented a case study to prove the efficiency of the proposed framework. Students' data were collected from Cairo Higher Institute for Engineering, Computer Science and Management department, during the period 2000-2012.

Saa [24] explored multiple factors which theoretically affect students' performance in higher education and concluded to some interesting results. In particular, he showed that the students' performance is not totally dependent on their academic efforts; there are many other personal and social factors that have equal or greater influences as well. Moreover, he developed a qualitative model that classifies students and predicts their performance.

Yasmeen et al. [28] proposed a prediction model for students' academic performance and studied the identification of the courses which highly influence and affect low academic performance. Their study was based on records of 100 graduates from the Information Technology department of King Saud University. Their primary goal was to explore the high potential of data mining applications for university management, referring to the optimal usage of data mining methods and techniques to the collected historical data. Their experiments indicate that reliable predictions can be achieved based on the performance of students in second year courses.

Along this line, Yassein et al. [29] utilized machine learning and data mining techniques to deeply analyze students' data and identify features affecting student performance in selected courses in Najran University in Saudi Arabia. More specifically, they studied the relationship between both practical work and assignments in several courses and their success rate. Their results revealed the strong relationship between these factors; in addition, it was found that a large number of given assignments acts negatively on course academic performance.

Xu et al. [26] developed a model that predicts student performance in degree programs using a novel machine learning method based on students' progressive performance states. Their proposed method adopts a latent factor model-based course clustering method developed to discover relevant courses for constructing base predictors while an ensemble-based progressive prediction architecture was developed to incorporate students' evolving performance into the prediction.

The dataset contained 1169 undergraduate students over three years from Mechanical and Aerospace Engineering department at UCLA. Their results showed up the effectiveness of their proposed method achieving superior performance to benchmark approaches.

III. Research methodology

The primary goals of the present research are the accurate and early identification of the students who are at-risk of not completing their studies within six years and the accurate classification of students who have successfully graduated. We have adopted a two-stages methodology, where the first stage concerns data collection and data preparation, while the second one deploys the proposed two-level classification algorithm.

A. Dataset

We have collected 282 student records over four years (2010-2013) from the School of Health & Social Welfare of Technological Institute of Western Greece. The dataset consists of demographic information as well as information of the students' performance in courses of the first two years of their studies. Notice that the Bachelor's degree program consists of four (4) academic years (eight semesters). Each record comprised 127 variables divided in two groups: the "*Demographic-based group*" and the "*Performance-based group*".

Attribute	Values
Genger	male/female
Age	integer
Home location	nominal
High school type	technical/general/evening

TABLE I
DEMOGRAPHIC-BASED GROUP ATTRIBUTES

Attribute	Values
Type of course	core/laboratory/clinical
Number of times examined	integer
Final grade in the course	integer

TABLE II
PERFORMANCE-BASED GROUP ATTRIBUTES

The Demographic-based group represents attributes concerning students' gender, age, home location and type of high school, as presented in Table I. The reason why most researchers utilize students demographic information such as gender is because male and female students exhibit different styles of learning process [5].

The Performance-based group represents attributes concerning courses characteristics and students' progress in several courses. More specifically, the Bachelor's program includes 41 courses of which twenty five (25) are core, twelve (12) are laboratory and four (4) are clinical courses. For each course, we register the type of course (core/laboratory/clinical), the number of times the student was examined in the first three years of his/her studies and the final grade (Table II). It is noted that in case, the student has not successfully passed the course, the grade assigned is -1.

Finally, the students were classified utilizing a four-level classification scheme, based on the years needed to complete their studies, namely {*4 years*, *5 years*, *6 years*, *Fail*}.

B. Two-level classification algorithm

Subsequently, we present our proposed novel two-level classification scheme aiming to achieve the highest possible efficiency and efficacy. We recall that two-level classification schemes are heuristic pattern recognition tools that are supposed to yield better classification accuracy than single-level ones at the expense of a certain complication of the classification structure [4, 11, 12, 27].

An overview of our classification algorithm is depicted in Figure 1 while a high level description of the training process of our two-level classifier is presented in Table III. On the first level of our classification scheme, we utilize a classifier to distinguish the students who are likely to "*Graduate*" or "*Fail*". More analytically, this classifier predicts whether the student will manage to complete his/her studies within 6 years. In the rest of our work, we refer to this classifier as A-level classifier. Clearly, the primary goal of the classifier in this level is to identify the students' who are at-risk of not completing their studies. In case the verdict (or prediction) of the A-level classifier is "*Graduate*", we use a second-level classifier to conduct a more specialized decision and distinguish between "*4 years*", "*5 years*" and "*6 years*" to finish his/her studies. We refer this classifier as B-level classifier.

```
Input:   $D$ - Initial training dataset.
         $C_A$ - User selected A-level classifier.
         $C_B$ - User selected B-level classifier.
Output:  Use trained two-level classifier to predict class labels of
         the test cases.

/* Initialization phase */
Set $D_A = \emptyset$ and $D_B = \emptyset$ .
for each $(x, y) \in D$ do
    if ($y ==$ "Fail") then
        $D_A = D_A \cup (x, y)$.
    else
        $D_A = D_A \cup (x, \text{"Graduate"})$.
        $D_B = D_B \cup (x, y)$.
    end if

/* Training phase */
Train classifier $C_A$ on dataset $D_A$.
Train classifier $C_B$ on dataset $D_B$.
```

Remarks: For each instance (x, y) in the dataset D, x stands for the vector of attributes while y stands for the output variable, namely $y \in \{$"4 years", "5 years","6 years","Fail"$\}$

TABLE III
TWO-LEVEL CLASSIFIER

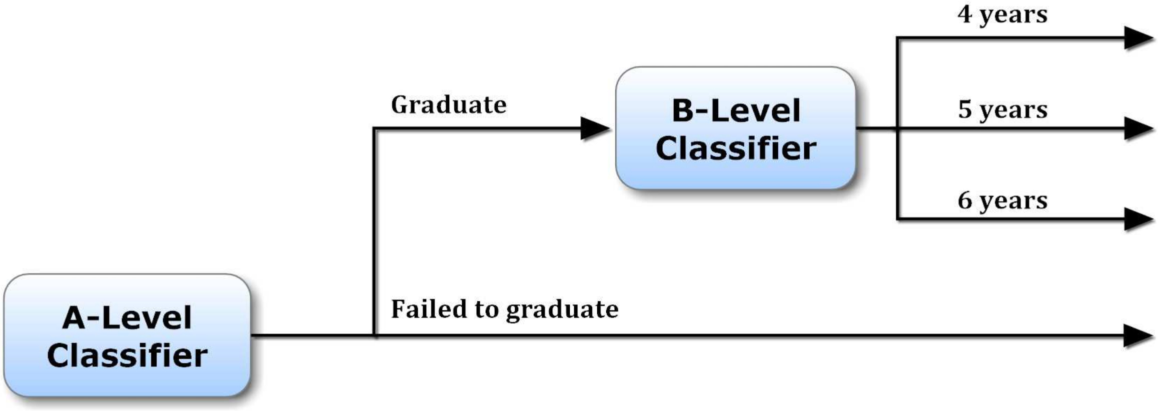

Fig. 1. An overview of the two-level classifier

IV. Experimental results

In this section, we present a series of tests conducted to evaluate the performance of the proposed two-level classification scheme utilizing the most popular and frequently used classification algorithms as A-level and B-level learners.

Naive Bayes (NB) algorithm was the representative of the Bayesian networks [8]. The Back-Propagation (BP) algorithm with momentum [23] was representative of the artificial neural networks while from the support vector machines, we have selected the Sequential Minimal Optimization (SMO) algorithm [19]. From the decision trees, C4.5 algorithm [20] was the representative in our study and RIPPER (JRip) algorithm [6] was selected as a typical rule-learning technique since it is one of the most commonly used methods for producing classification rules. Finally, 10NN algorithm was selected as instance-based learner [1] with Euclidean distance as distance metric.

The implementation code was written in JAVA using the WEKA Machine Learning Toolkit [9]. The classification accuracy was evaluated using the stratified 10-fold cross-validation i.e. the data was separated into folds so that each fold had the same distribution of grades as the entire data set.

To evaluate the performance of the proposed classification algorithm, we consider the following four performance metrics.

$$A_{NG} = \frac{\text{number of students correctly predicted Not to Graduate}}{\text{total number of not graduated students}}.$$

$$A_{G} = \frac{\text{number of students correctly predicted to Graduate}}{\text{total number of graduated students}}.$$

$$A_{CGT} = \frac{\text{number of graduated students Correctly predicted their Graduation Time}}{\text{total number of graduated students}}.$$

$$A_{C} = \frac{\text{number of Correctly classified students}}{\text{total number of students}}.$$

The first two metrics evaluate the performance of A-level classifier, the third metric evaluates the performance of B-level classifier while the last metric evaluates the overall performance of the proposed two-level classification scheme. Our aim is to find which learner is best suited for A-level and B-level for producing the highest performance.

Table IV presents the performance evaluation of A-level and B-level classifiers. Table V presents the performance of the proposed two-level scheme utilizing various learners as A-level and B-level classifiers. Notice that the highest classification accuracy is highlighted in bold. Additionally, a more representative visualization of the classification performance of the compared classifiers for each performance metric is presented in Figures 2-5.

Classifier	A_{NG}	A_G	A_{CGT}
NB	63.04%	75.38%	60.80%
BP	60.87%	98.99%	66.33%
SMO	63.04%	96.98%	68.84%
C4.5	52.17%	97.99%	**77.39**%
JRip	50.00%	95.48%	76.88%
10NN	**84.78%**	**98.99%**	75.38%

TABLE IV
Accuracy of A-level and B-level classifiers

		B-Level Classifier					
		NB	BP	SMO	C4.5	JRip	10NN
A-Level Classifier	NB	49.27%	49.70%	51.73%	59.10%	57.90%	55.87%
	BP	60.37%	64.88%	66.95%	73.88%	73.52%	72.28%
	SMO	60.75%	64.08%	65.72%	73.05%	73.50%	71.05%
	C4.5	57.90%	62.83%	64.90%	72.23%	71.87%	70.23%
	JRip	56.72%	61.65%	64.12%	70.25%	71.08%	69.87%
	10NN	64.82%	69.73%	71.80%	**78.73**%	76.75%	77.13%

TABLE V
Two-level classifier classification accuracy

Clearly, 10NN illustrates the best performance as A-level classifier, since it exhibits the highest accuracy of correctly identifying students who managed to graduate (or not), within 6 years. Moreover, C4.5 reports the best performance as B-level classifier, illustrating the highest percentage of correctly classified students who have successfully graduated, closely followed by JRip.

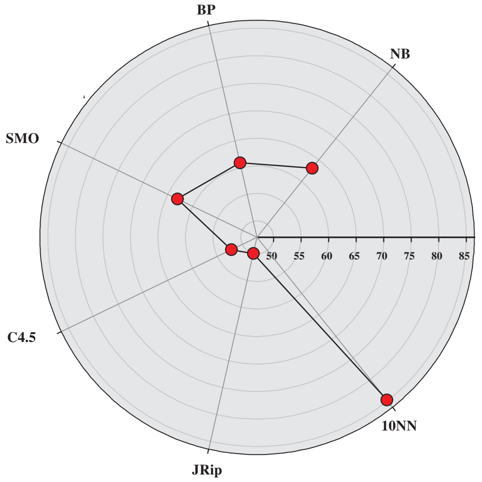

Fig. 2. Radar plot for performance comparison relative to the performance metric A_{NG}

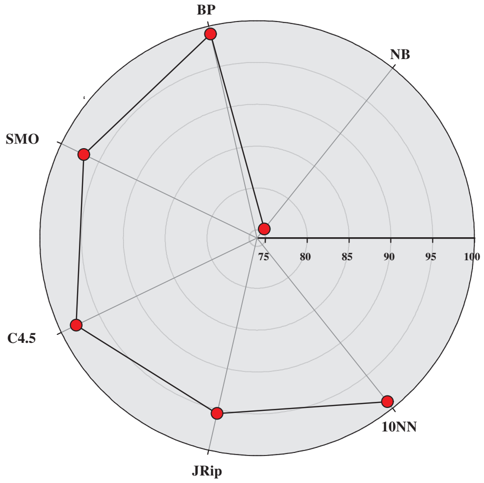

Fig. 3. Radar plot for performance comparison relative to the performance metric A_G

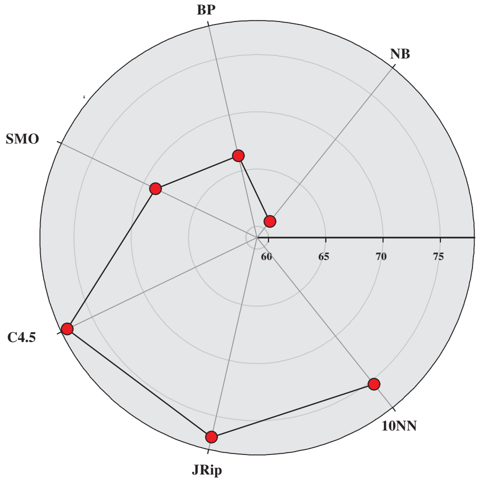

Fig. 4. Radar plot for performance comparison relative to the performance metric A_{CGT}

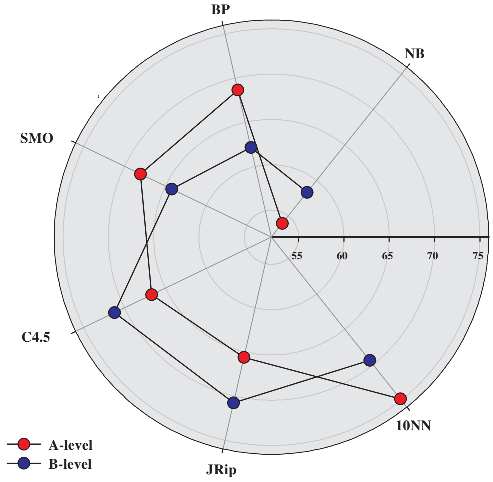

Fig. 5. Radar plot for performance comparison relative to average performance of A-level and B-level classifier

Figures 2, 3 and 4 present the classification performance of the compared learners, relative to each performance metric while Figure 5 presents the average performance of A-level and B-level classifiers. The interpretation of Figures 2 and 3 reveal that 10NN reported the highest average classification accuracy as A-Level classifier. As regards B-level classifier, C4.5 exhibited the best performance slightly outperforming JRip and 10NN. More specifically, C4.5 reported 71.21% average classification performance while JRip and 10NN exhibited 70.77% and 69.41%, respectively. Finally, it is worth noticing that based on the previous discussion, we conclude that the best classification performance of the two-level classifier was demonstrated in case 10NN was selected as A-level classifier and C4.5 as a B-level one.

Subsequently, in order to illustrate the efficacy of our two-level classification algorithm, we compared it with the performance of the supervised learning algorithms. Notice that our algorithm uses C4.5 and 10NN, as A-level and B-level classifiers, respectively. Table VI summarizes the accuracy of each individual classifier which reveals the efficacy of our two-level algorithm. Clearly, the proposed scheme significantly outperforms all individual classifiers, exhibiting higher classification performance.

Classifier	NB	MLP	SMO	JRip	C4.5	10NN	Two-level (10NN-C4.5)
Accuracy	47.80%	62.10%	64.97%	66.58%	69.73%	68.89%	**78.73%**

TABLE VI
PERFORMANCE OF EACH INDIVIDUAL CLASSIFIER

V. CONCLUSIONS

In this work, we present a two-level machine learning classifier for the accurate prediction of the students' graduation time. The reported experimental results reveal that the proposed algorithm is effective and practical for early student graduation prediction and early identification of students at-risk in order to take proper actions for improving their performance. Our work could provide valuable hints and insights for better educational support by offering customized assistance according to students' predicted performance. It can be used as a reference for decision making in the graduate program admission process.

Our future work directions include the application of the proposed scheme on data from several departments in order to extract useful information about key factors affecting students' performance. Another direction concerns the incorporation of the proposed algorithm in a semi-supervised framework.

Acknowledgments

The authors would like to thank the *Technological Institute of Western Greece* for granting the corresponding data collection utilized in this study.

References

[1] D. Aha. *Lazy Learning*. Dordrecht: Kluwer Academic Publishers, 1997.

[2] S. Aud, M. Planty, and W.J. Hussar. *Condition of education 2011*. Government Printing Office, 2011.

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

[4] Y.L. Barabash. Collective statistical decisions in recognition. *Radio i Sviaz*, 1983.

[5] U. bin Mat, N. Buniyamin, P.M. Arsad, and R. Kassim. An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. In *2013 IEEE 5th Conference on Engineering Education (ICEED)*, pages 126–130. IEEE, 2013.

[6] W. Cohen. Fast effective rule induction. In *International Conference on Machine Learning*, pages 115–123, 1995.

[7] National Research Council. *Building a workforce for the information economy*. National Academies Press, 2001.

[8] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: An update. *SIGKDD Explorations Newsletters*, 11:10–18, 2009.

[10] N. Johnson. Three policies to reduce time to degree. *Complete College America*, 2011.

[11] L.I. Kuncheva. "Change-glasses" approach in pattern recognition. *Pattern Recognition Letters*, 14:619–623, 1993.

[12] I.E. Livieris, K. Drakopoulou, Th. Kotsilieris, V. Tampakas, and P. Pintelas. DSS-PSP - A decision support software for evaluating students' performance. In *Engineering Applications of Neural Networks (EANN)*, volume 744, pages 63–74. Springer, 2017.

[13] I.E. Livieris, K. Drakopoulou, and P. Pintelas. Predicting students' performance using artificial neural networks. In *Information and Communication Technologies in Education*, September 2012.

[14] I.E. Livieris, K. Drakopoulou, V. Tampakas, T. Mikropoulos, and P. Pintelas. Predicting secondary school students' performance utilizing a semi-supervised learning approach. *Journal of Educational Computing Research*, 2018.

[15] I.E. Livieris, T. Mikropoulos, and P. Pintelas. A decision support system for predicting students' performance. *Themes in Science and Technology Education*, 9:43–57, 2016.

[16] M.F. Musso, E. Kyndt, E.C. Cascallar, and F. Dochy. Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks. *Frontline Learning Research*, 1(1):42–71, 2013.

[17] H.M. Nagy, W.M. Aly, and O.F. Hegazy. An educational data mining system for advising higher education students. *World Academy of Science, Engineering and Technology International Journal of Information Engineering*, 7(10):175–179, 2013.

[18] A. Peña-Ayala. Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462, 2014.

[19] J. Platt. Using sparseness and analytic QP to speed training of support vector machines. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in neural information processing systems*, pages 557–563, MA: MIT Press, 1999.

[20] J.R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Francisco, 1993.

[21] C. Romero and S. Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33:135–146, 2007.

[22] C. Romero and S. Ventura. Educational data mining: A review of the state of the art. *IEEE on Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, 40(6):601–618, 2010.

[23] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. In D. Rumelhart and J. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, pages 318–362, Cambridge, Massachusetts, 1986.

[24] A.A. Saa. Educational data mining & students performance prediction. *International Journal of Advanced Computer Science and Applications*, 7(5):212–220, 2016.

[25] M.C. Stetser and R. Stillwell. Public high school four-year on-time graduation rates and event dropout rates: School years 2010-11 and 2011-12. First Look. NCES 2014-391. *National Center for Education Statistics*, 2014.

[26] J. Xu, K.H. Moon, and M. van der Schaar. A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[27] L. Xu and A. Krzyzak C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on systems, man, and cybernetics*, 22(3):418–435, 1992.

[28] A. Yasmeen, A. Wejdan, A. Isra, and A. Muna. Predicting critical courses affecting students performance: A case study. *Procedia Computer Science*, 82:65–71, 2016.

[29] N.A. Yassein, R.G.M. Helali, and S.B. Mohomad. Predicting critical courses affecting students performance: A case study. *Journal of Information Technology & Software Engineering*, 7(5):1–5, 2017.