

Received November 13, 2021, accepted November 30, 2021, date of publication December 10, 2021, date of current version December 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3134787

Interpretable Models for Early Prediction of Certification in MOOCs: A Case Study on a MOOC for Smart City Professionals

GEORGIOS KOSTOPOULOS¹, THEODOR PANAGIOTAKOPOULOS^{1,2},
SOTIRIS KOTSANTIS³, CHRISTOS PIERRAKEAS⁴, AND ACHILLES KAMEAS¹

¹School of Science and Technology, Hellenic Open University, 26335 Patras, Greece

²School of Business, University of Nicosia, 2417 Nicosia, Cyprus

³Department of Mathematics, University of Patras, 26500 Patras, Greece

⁴Department of Management Science and Technology, University of Patras, 26500 Patras, Greece

Corresponding author: Sotiris Kotsiantis (kotsiantis@upatras.gr)

This work was supported by the Research Project DevOps, “DevOps Competences for Smart Cities” under Project 601015-EPP-1-2018-1-EL-EPPKA2-SSA, Erasmus⁺ Program, KA2: Cooperation for innovation and the exchange of good practices-SSA.

ABSTRACT Over the last few years, Massive Open Online Courses (MOOCs) have expanded rapidly and tend to become the most typical form of online and distance higher education. As a result, a tremendous amount of data is generated and stored on MOOCs online learning platforms. In any case, this data should be effectively transformed into knowledge, thus providing valuable feedback to learners, and enhancing decision making practices in the educational field. Despite the benefits and learning prospects that MOOCs offer to learners, there is a considerable divergence between enrollment and completion rates. In this context, the main scope of this study is to exploit predictive analytics and explainable artificial intelligence for the early prediction of student certification in a 11-week MOOC for smart cities, namely DevOps. A plethora of Machine Learning models were built employing familiar classification algorithms. The experimental results revealed that the models based on Gradient Boosting, Logistic Regression and Light Gradient Boosted Machine classifiers prevailed in terms of Accuracy, Area Under Curve, Recall, Precision, F1-score, Kappa, and Matthews Correlation Coefficient, getting a predictive accuracy of 94.41% at the end of the second week of the course. Therefore, students who are less likely to obtain a certificate could be envisaged at an early enough stage to provide sufficient support actions and targeted intervention strategies to them. Finally, the performance attributes (i.e., overall grades per week) proved to be the most important predictors for identifying students at risk of failure.

INDEX TERMS MOOCs, certification, early prediction, supervised learning, explainable artificial intelligence, interpretable predictive models, feature importance.

I. INTRODUCTION

Over the last few years, there is an increasing interest in Massive Open Online Courses (MOOCs) offered by top-quality universities worldwide. MOOCs have expanded rapidly nowadays and tend to become the most typical form of online and distance higher education [1]. In light of this new trend, diverse and large groups of learners, varying in several characteristics such as age, nationality, family obligations and educational level, may attend for free or at low-cost flexible

and short-term courses of their interest and study at their own pace without requiring physical attendance [2]. Additionally, even after registration has been completed, attendees differ significantly with regard to their objectives, motivation, interests, and interaction with the course content [3].

As a result, a tremendous amount of data is generated and stored on MOOC platforms regarding student learning behavior and performance, engagement, social interactions, assignments scores, learning outcomes, and demographic information such as gender, ethnicity, education level, employment status and employment type, to name a few [4]. However, this data should be mined and effectively trans-

The associate editor coordinating the review of this manuscript and approving it for publication was Juan A. Lara¹.

formed into knowledge, in order to provide valuable feedback to students and enhance decision making practices in the educational field [5]. Educational Data Mining (EDM) and Learning Analytics (LA) constitute powerful tools for uncovering valuable information from vast amount of raw data.

EDM and LA have significantly evolved over the past two decades and form an integral part of educational research [6]. These different but complementary scientific fields aim primarily at enhancing learning experience and optimizing the quality of teaching [7]. More precisely, EDM concerns the development and implementation of Data Mining (DM) methods in data derived from different educational settings for addressing a wide range of learning problems [8]. LA refers to the exploitation of educational data for supporting educational practices, understanding learning behavior and improving student performance [9]. Notwithstanding their common objective, the principal difference between them is that EDM tackles an educational problem from a technological viewpoint, while LA is mainly focused on its pedagogical aspects [10].

One of the most studied problems in the fields of EDM and LA is prediction, which involves building a Machine Learning (ML) model for inferring future learning characteristics of students based on historical data and ML methods [11]. This process is known as predictive analytics [12]. Regarding MOOCs, prediction is an umbrella term for a wide array of specific predictive problems which have attracted the attention of many researchers. These problems encompass primarily the prediction of student behaviors and outcomes, such as dropout, retention, completion, certification [13], final exam grade and course grade [14]. The first four problems refer to binary classification tasks (i.e., the output attribute has two class labels), whereas grade prediction is a typical multiclass or regression task (i.e., the output attribute is a numerical one). In particular, concerning the term “certification”, which is the main focus of our research, it refers to the successful completion of a MOOC course [15] (i.e., achieving an average score above a predefined threshold).

Despite the educational benefits and learning prospects that MOOCs offer to students, there is a considerable divergence between enrollment rates and completion rates [16]. Lack of time and interactivity, course time-table and length, motivation, subject interest, isolation feelings, and poor background knowledge are important factors of low completion rates [17]. These factors can be classified into two main groups: student-related and course-related factors [2]. Recent researches have shown that the success of a MOOC is directly linked to the provision of support services to potential low performer students, which could motivate them to successfully complete the course and receive a certification [18].

In this context, the main scope of this study is to exploit predictive analytics for predicting student certification in a MOOC. The contribution of this study is three-fold. First,

we examine whether predictive analytics could help us to gain insight into student online learning behavior and build highly effective learning models in predicting student certification at the end of a 11-week course. Second, we intend to understand and explain the predictions made by a predictive model and specify the features of students which are of great impact on earning a course certificate. Explainable Artificial Intelligence (XAI) is a new research field attempting to build ML models that are easier to interpret and understand than the so-called black-box models [19]. Finally, the results of the study provide evidence that students who are less likely to obtain a certificate could be envisaged at an early enough stage to provide sufficient support actions and properly targeted intervention strategies to them.

The rest of the paper is organized as follows. Section II presents a brief review of previous work in predicting student certification in MOOCs. The dataset used in the study is described in Section III, whereas Section IV formulates the experimental design and analyzes the produced results. Finally, the paper concludes considering some thoughts for future research directions.

II. RELATED WORK

A great deal of existing research has focused on building predictive models for predicting student behaviors and outcomes in MOOCs. Additionally, several ML models have been utilized in recent years confronting the problem of MOOC unsuccessful completion.

Coleman *et al.* applied Latent Dirichlet Allocation (LDA), an unsupervised probabilistic method, for uncovering behavioral patterns of students enrolled in a MITx course [20]. These patterns were then used for building a mixed-membership model to predict certification of students.

Al-Shabandar *et al.* examined the efficiency of various ML algorithms for predicting student certification in 15 MOOC courses offered by Harvard and MIT [21]. To this end, they considered two feature datasets: the clickstream dataset (i.e., video views, content interaction, access to assignments, and posts in discussion forums) and the demographic dataset (i.e., age, gender, and education level). Random Forests (RF) and Multilayer Perceptrons (MLPs) were found to prevail in terms of several metrics such as classification accuracy and kappa.

Four classification algorithms (i.e., Logistic Regression (LR), k Nearest Neighbors (k -NN), Gradient Boosting and RF) were applied in [22] for the early prediction of student certification. The produced models were evaluated in terms of F1-score and Area Under Curve (AUC) at the end of each one of the first four weeks of the course.

Very recently, Moore and Wang examined the influence of student motivational dispositions on completing a XarvardX MOOC [3]. For this purpose, they employed Latent Profile Analysis (LPA) for detecting specific groups among course completers. Two latent profiles were identified regarding intrinsic and extrinsic motivation of students. What is more,

it was found that educational background, gender, and latent profile were all substantially related to the course grade.

In a similar study, Gitinabard *et al.* combined forum posts and online activity log files for predicting student dropout and certification in a MOOC course delivered by Coursera [23]. To this end, LR and Support Vector Machines (SVMs) were applied, creating two classification models. The predictive accuracy was above 85% from the first week of the course, whereas AUC exceeded 90%. Additionally, submissions and video watching were found to be the most influential attributes. Video watching frequency per week was also confirmed to be a significant attribute for grade prediction in a weekly-organized MOOC [24], [25].

III. DATA DESCRIPTION

This study was conducted in the context of the Erasmus+ Sector Skills Alliance project called “DevOps Competences for Smart Cities”.¹ DevOps aims at equipping current and aspiring smart city professionals with appropriate skills to enable transformative urban innovation and support the technologically enhanced urban governance with a special emphasis on the DevOps methodology [26]. Registration in the DevOps MOOC² started at 15 September 2020 and lasted one month, whereas the start date of the course was 19 October 2020.

The course was organized in a weekly setup (11 weeks length in total), and it was structured at approximately 1-2 modules per week³ (15 modules in total). Each module was available in English and consisted of 2 to 5 learning units, each of which included an automatically graded multiple choice assessment quiz. Weeks 7 and 11 were used as reflection weeks to help learners catch up on their study or get some free time to reflect on what has been taught. The course content was designed to address the European Qualifications Framework (EQF) level 5, since this is the requisite level of autonomy and responsibility for smart city professionals.

The registration form comprised a questionnaire regarding personal, demographic and employment data of students, in which they were informed that all data would be acquired and used according to the GDPR regulation (EU 2016/679) for evaluating the quality of the course. Participants were asked to consent for the analysis of data; otherwise, they could skip the questionnaire and proceed with the registration providing only their full name and email. An overall number of 961 students enrolled in the course and 944 of them filled the questionnaire providing the full set of data.

Considering that the quality of data is essential for building effective and robust ML models [27], a preprocess analysis was performed for cleaning and preparing the data before applying a ML algorithm. For this purpose, the missing values of the numerical attributes were imputed employing the mean imputation method. Regarding the missing values of

TABLE 1. Demographics set of attributes.

Attributes
Gender
Mother tongue
Education level
Employment status
Current occupation
Occupation experience in years
Daily work hours
Technical English language skills
Digital proficiency skills
Previous experience in MOOC
Average available study hours per week

TABLE 2. Performance set of attributes.

Attributes
Week1 Unit_i Assessment Quiz, $i \in \{1,2,3,4,5\}$
Week2 Unit_i Assessment Quiz, $i \in \{1,2,3\}$
Overall grade in Module_i, $i \in \{1,2\}$

TABLE 3. Activity set of attributes.

Attributes
Number of views in introductory forum
Number of views in announcements forum
Number of forum views in Module_i, $i \in \{1,2\}$
Number of posts in Module_i, $i \in \{1,2\}$
Number of discussions in Module_i, $i \in \{1,2\}$
Average connections per day in Module_i, $i \in \{1,2\}$
Total time dedicated in Module_i (mins), $i \in \{1,2\}$

the qualitative attributes, they were replaced with the constant “unknown”. Besides that, all zero activity records, during the course, were excluded. Finally, 936 records were filtered and saved in a comma-separated values (.csv) format file, allowing to apply a plethora of DM methods and ML algorithms.

The initial dataset comprised eleven qualitative attributes regarding personal information of students such as gender, educational background, employment, skills, previous MOOC experience, and average available study hours per week (Table 1). Therefore, it is called Demographics set of attributes. This dataset was gradually enriched with two other sets of attributes, namely Performance set and Activity set.

The Performance set consisted of ten attributes about students’ learning achievements during the first two weeks of the course, mainly considering grading scores in quizzes (100-point scale) and overall grade in modules 1 and 2 (Table 2).

The Activity behavior set comprised twelve numerical attributes concerning students’ recording activity in the online learning platform such as number of views, posts, discussions, and connections, as well as the total time devoted to the first two modules of the course.

¹<https://devops.uth.gr/dev/>

²<https://smartdevopsmooc.eu/moodle/pages/login.php>

³<https://devops.uth.gr/dev/about-the-mooc/>

TABLE 4. Descriptive statistics of performance and activity attributes.

	count	mean	std	min	1 st quartile	median	3 rd quartile	max
Week1 Unit_1 Assessment Quiz	936.0	32.8846	39.6294	0.0	0.0	0.0	70.0	100.0
Week1 Unit_2 Assessment Quiz	936.0	35.1495	45.9990	0.0	0.0	0.0	100.0	100.0
Week1 Unit_3 Assessment Quiz	936.0	32.4572	44.7637	0.0	0.0	0.0	100.0	100.0
Week1 Unit_4 Assessment Quiz	936.0	32.9058	45.0500	0.0	0.0	0.0	100.0	100.0
Week1 Unit_5 Assessment Quiz	936.0	33.3152	45.1672	0.0	0.0	0.0	83.33	100.0
Week2 Unit_1 Assessment Quiz	936.0	27.2970	39.1256	0.0	0.0	0.0	62.5	100.0
Week2 Unit_2 Assessment Quiz	936.0	27.0619	39.4212	0.0	0.0	0.0	70.0	100.0
Week2 Unit_3 Assessment Quiz	936.0	28.5389	41.1279	0.0	0.0	0.0	75.0	100.0
Overall grade in Module_1	936.0	33.3428	42.7790	0.0	0.0	0.0	87.33	100.0
Overall grade in Module_2	936.0	27.6326	39.2956	0.0	0.0	0.0	72.5	100.0
Number of views in introductory forum	936.0	6.3194	19.9614	0.0	0.0	0.0	7.0	450.0
Number of views in announcements forum	936.0	2.3247	4.0814	0.0	0.0	1.0	3.0	33.0
Number of forum views in Module_1	936.0	3.7521	9.9886	0.0	0.0	0.0	3.0	142.0
Number of forum views in Module_2	936.0	2.0405	7.2677	0.0	0.0	0.0	1.0	98.0
Number of posts in Module_1	936.0	0.0416	0.3292	0.0	0.0	0.0	0.0	7.0
Number of posts in Module_2	936.0	0.0181	0.2282	0.0	0.0	0.0	0.0	6.0
Number of discussions in Module_1	936.0	0.0203	0.0214	0.0	0.0	0.0	0.0	1.0
Number of discussions in Module_2	936.0	0.1411	0.2917	0.0	0.0	0.0	0.0	8.0
Average connections per day in Module_1	936.0	0.1743	0.2217	0.0	0.0	0.16	0.31	0.94
Average connections per day in Module_2	936.0	0.1569	0.2357	0.0	0.0	0.0	0.31	1.1
Total time dedicated in Module_1	936.0	52.8664	107.1565	0.0	0.0	0.0	58.00	848.0
Total time dedicated in Module_2	936.0	60.8603	124.4969	0.0	0.0	0.0	64.5	859.0

The target attribute “Course Total Grade” is a binary one with two possible values {0,1}, where 0/1 means that a student has not/has obtained a certification of successful completion of the course (i.e., a student achieved an average quiz score of 80% or more). As regards the distribution of the two output classes in the dataset, 75.75% of the records represents students who obtained a certification, whereas 24.25% corresponds to students who didn’t get a course certificate.

Table 4 provides a descriptive statistic summary (i.e., count, mean, standard deviation, minimum, 1st quartile, median, 3rd quartile and maximum) for each one of the performance and activity attributes in order to get a better understanding of the distribution of the data.

In addition, a matrix heatmap is illustrated in Fig. 1 depicting the correlation dependence between the numerical attributes of the dataset. Therefore, each square of the matrix represents the correlation between the attributes paired on the two axes. The red color indicates positive correlation between two attributes, whereas the blue one negative. Moreover, the intensity of the color implies how strongly these attributes are correlated, meaning that the deeper color corresponds to stronger correlation. A cursory reading of the matrix heatmap reveals that there is a perfect positive correlation between the performance attributes.

IV. METHODOLOGY AND EXPERIMENTS

A plethora of supervised learning algorithms were employed for building corresponding predictive models using PyCaret [28]. PyCaret is an open-source software ML library package in Python, which enables the implementation of

several classification methods. Specifically, the algorithms used in the experiments are as follows:

- Adaptive Boosting (AdaBoost) classifier [29], which was originally used for binary classification problems. This is a meta-estimator that is seeking to build a strong classifier exploiting a set of weak classifiers.
- Gradient Boosting (GB) classifier [30], another common boosting method which fits a new estimator targeting the errors made by the previous one.
- Classification and Regression Tree (Cart) [31], a very powerful decision tree algorithm for both classification and regression problems, using Gini coefficient for splitting the dataset at the node with larger uncertainty.
- Extremely Randomized Trees (Extra) [32], a very fast ensemble-based algorithm which chooses randomly the split point at each node.
- Linear Discriminant Analysis (LDA) algorithm for two-class classification problems [33], which employs the Bayes theorem for calculating the probability of the output class given the input attributes under the assumption that the output class is linearly separable.
- Light Gradient Boosted Machine (LightGBM) ensemble method [34], a robust extended version of the GB classifier.
- Logistic Regression (LR) algorithm [35], a representative statistical method for binary classification problems, which models the probability of a student classified as certificated (with two possible outcomes: 0 and 1) given the values of the independent attributes.

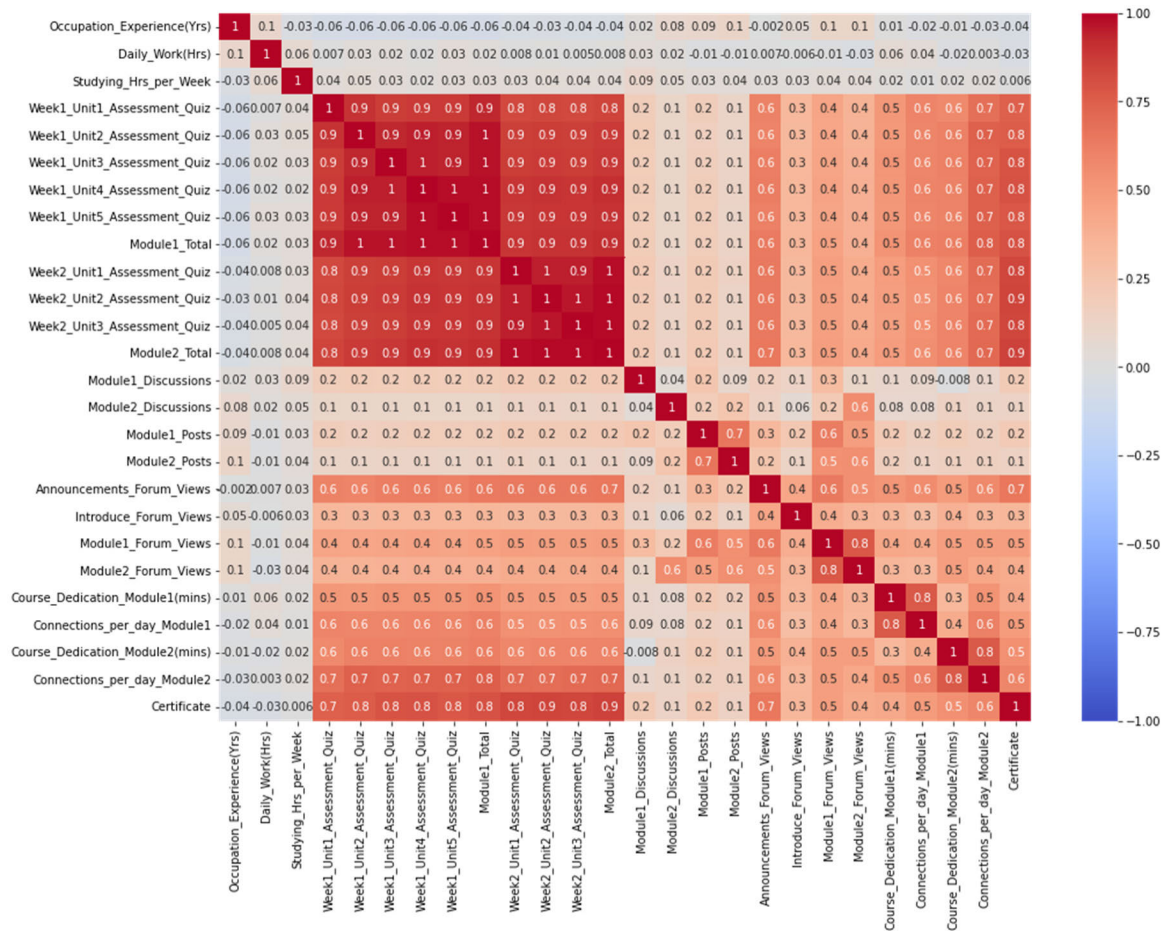


FIGURE 1. Correlation matrix heatmap.

- Random Forest (RF) [36], a popular bagging-based algorithm, which combines the output of several decision trees trained on sub-samples of the dataset for producing the final prediction via majority voting.

Several studies indicate the effectiveness of these methods for building highly accurate and robust predictive models in the fields of EDM and LA [11]. What is more, boosting and ensemble methods show top-performing results in both classification and regression problems without any special parameter configuration being required. The parameter settings of these methods are presented in Table 5. In most cases, the default parameter settings were adopted.

Since the distribution of the two output classes in the dataset was imbalanced, we applied the Synthetic Minority Oversampling Technique (SMOTE) [37] for augmenting the minority class. In addition, the k -fold ($k = 10$) cross validation resampling procedure was used for evaluating the performance of the predictive models [38], [39]. Accordingly, the dataset was randomly divided into k folds of equal size. Each fold was used for evaluating the performance of the model trained on the rest folds, whereas the final measure was

the average value of the computed performance measures on each test fold.

One of the main goals of our study was the early prediction of student certification. Therefore, the experiments were performed in three consecutive steps each of which was linked to a specific time-point. Step 1 (Week 0) corresponded to the data available before the start of the course, whereas steps 2 and 3 corresponded to the time-point at the end of the first (Week 1) and second week (Week 2) of the course respectively. The attributes used in each one of the experiment steps are shown in Table 6.

V. RESULTS

A broad variety of widely used evaluation metrics were calculated to quantify the performance of the classification models in each one of the experimental steps. More specifically, these metrics include accuracy (Acc), Area Under Curve (AUC), Recall, Precision, F1-score, Kappa, and Matthews Correlation Coefficient (MCC).

The results are shown in Tables 7-8, while the best value for each metric per step is bold highlighted. Overall, it is observed that GB, LightGBM and LR produce the best-performing models. All metrics increase over time, as could

TABLE 5. Parameters configuration settings.

Algorithm	Parameters Settings
AdaBoost	Solver = Singular value decomposition (Svd)
	Base estimator = SAMMER algorithm
	Maximum number of estimators = 50
	Learning Rate = 1.0
GB	Criterion = Friedman MSE
	Number of boosting stages = 100
	Learning Rate = 0.1
	Minimum number of samples to split node = 2
	Minimum number of samples to be at leaf = 1
	Maximum depth of regression estimators = 3
Cart	Criterion = Gini
	Splitter = Best
	Maximum depth of the tree = None
	Minimum number of samples to split node = 2
	Minimum number of samples to be at leaf = 1
Extra	Criterion = Gini
	Number of trees = 100
	Maximum depth of the tree = None
	Minimum number of samples to split node = 2
	Minimum number of samples to be at leaf = 1
LDA	Solver = Svd
	Threshold for rank estimation in SVD solver = 0.0001
LightGBM	Boosting type = Gradient Boosting Decision Tree
	Number of boosted trees to fit = 100
	Maximum tree depth for base learners = None
	Minimum number of data needed in leaf = 20
	Maximum tree leaves for base learners = 31
LR	Norm used in the penalization = L2
	Maximum number of iterations = 1000
	Algorithm used in the optimization problem = lbfgs
RF	Criterion = Gini
	Number of trees = 100
	Maximum depth of the tree = None
	Minimum number of samples to split node = 2
	Minimum number of samples to be at leaf = 1

TABLE 6. Attributes used in each step.

Step 1	Demographic attributes
	Number of views in introductory forum
	Number of views in announcements forum
Step 2	Attributes of step 1
	Number of forum views in Module_1
	Number of posts in Module_1
	Number of discussions in Module_1
	Average connections per day in Module_1
	Total time dedicated in Module_1 (mins)
	Week1 Unit_i Assessment Quiz, $i \in \{1,2,3,4,5\}$
	Overall grade in Module_1
Step 3	Demographic attributes
	Performance attributes
	Activity attributes

be expected, since new information about students is added from week to week. Accuracy is greater than 83% for all

TABLE 7. Performance results of predictive models (week 0).

Model	Acc	AUC	Recall	Precision	F1	Kappa	MCC
AdaBoost	0.8694	0.9168	0.7031	0.7650	0.7165	0.6341	0.6455
GB	0.8801	0.9373	0.7126	0.7825	0.7384	0.6617	0.6680
Cart	0.8397	0.7844	0.6781	0.6640	0.6655	0.5614	0.5652
Extra	0.8302	0.8981	0.3988	0.8025	0.5249	0.4366	0.4797
LDA	0.7528	0.7932	0.4233	0.7002	0.5135	0.4313	0.4591
LightGBM	0.8587	0.9218	0.6779	0.7302	0.6949	0.6043	0.6103
LR	0.8837	0.9376	0.6481	0.8373	0.7215	0.6509	0.6647
RF	0.8599	0.9265	0.5931	0.7918	0.6653	0.5804	0.5977

TABLE 8. Performance results of predictive models (week 1).

Model	Acc	AUC	Recall	Precision	F1	Kappa	MCC
AdaBoost	0.9216	0.9739	0.8495	0.8423	0.8422	0.7903	0.7933
GB	0.9346	0.9773	0.8926	0.8545	0.8705	0.8270	0.8296
Cart	0.9097	0.8781	0.8160	0.8266	0.8158	0.7563	0.7608
Extra	0.9204	0.9743	0.8931	0.8107	0.8461	0.7929	0.7979
LDA	0.9097	0.9663	0.9029	0.7745	0.8316	0.7707	0.7768
LightGBM	0.9204	0.9718	0.8736	0.8231	0.8452	0.7919	0.7946
LR	0.9275	0.9733	0.8883	0.8326	0.8577	0.8093	0.8117
RF	0.9275	0.9752	0.8979	0.8298	0.8596	0.8111	0.8147

TABLE 9. Performance results of predictive models (week 2).

Model	Acc	AUC	Recall	Precision	F1	Kappa	MCC
AdaBoost	0.9192	0.9777	0.8581	0.8214	0.8367	0.7832	0.7858
GB	0.9382	0.9852	0.8969	0.8575	0.8744	0.8335	0.8359
Cart	0.9216	0.8971	0.8493	0.8375	0.8404	0.7886	0.7910
Extra	0.9323	0.9833	0.9219	0.8238	0.8687	0.8233	0.8270
LDA	0.9312	0.9729	0.8976	0.8373	0.8644	0.8185	0.8211
LightGBM	0.9441	0.9857	0.9262	0.8568	0.8891	0.8518	0.8540
LR	0.9430	0.9863	0.9219	0.8590	0.8877	0.8496	0.8521
RF	0.9370	0.9841	0.9319	0.8325	0.8782	0.8360	0.8395

models (except LDA) in Week 0 reaching 93.46% at the end of the first week and 94.41% at the end of the second week. F1-score, which is the harmonic mean of precision and recall, achieves a value of 87.05% and 88.91% at Week 1 and Week 2 respectively. In addition, AUC ranges from 0.9376 to 0.9863, showing that very precise models are created with high-level measure of class separability.

Fig. 2 presents the learning curve for each one of the learning models together with the respective feature importance plot. Each learning curve depicts the performance of the corresponding ML model over the number of training instances. The blue line represents the learning performance of the model in the training set, while the green one provides the model's generalization ability. It is noticed that the difference between training and 10-cross validation accuracy is very small, especially for a training size more than 400, indicating the models' ability to minimize bias and variance.

The feature importance plot displays how useful is each one of the 10 most important input attributes for predicting the target attribute for each learning model. Besides that,



FIGURE 2. Learning curve and feature importance plot for each classification model.

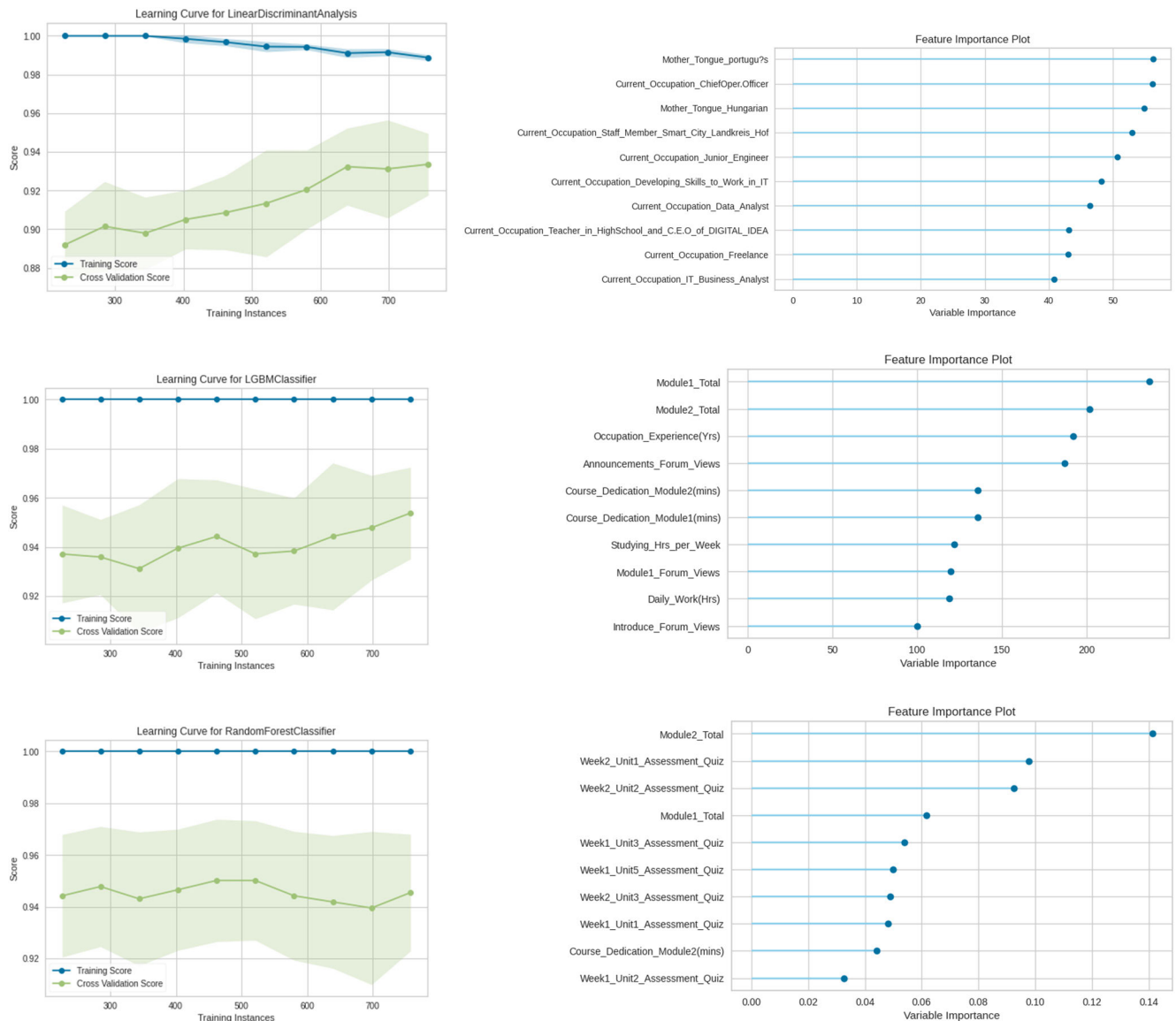


FIGURE 2. (Continued.) Learning curve and feature importance plot for each classification model.

it provides valuable information for better understanding both the data and the model. Mostly, it is observed that the performance attributes are the most important, and especially the total grade in modules 1 and 2, as well as the quiz grades. Regarding the activity attributes, the number of views in the announcement forum and the number of connections in modules 1 and 2 seems to be the most significant predictors.

These findings are in line with the results of recent research, showing that certificate achievers engage in more course-associated and graded assessment quizzes in the course [40].

Regarding the interpretability of the ML models, we also provide the SHAP (SHapley Additive exPlanations) summary plot [41] for the LightGBM (Fig. 3), GB (Fig.4) and RF models (Fig. 5). In these plots, the input attributes are ordered

in descending importance from top to bottom and present their impact on the model output (red color for positive impact and blue color for negative). It is shown that a large total grade in module 2 increases the chance of a learner to earn a certificate. Therefore, a result that was previously made by a black-box method is now reasonably explainable and easily understandable.

Finally, we made an attempt to build new classification models based only on the two most important attributes in each step (Table 10). It is observed (Tables 11-13) that the effectiveness of all methods was slightly improved for all metrics showing that the integration of XAI may also contribute towards building more effective and robust learning models regardless the ML method applied.

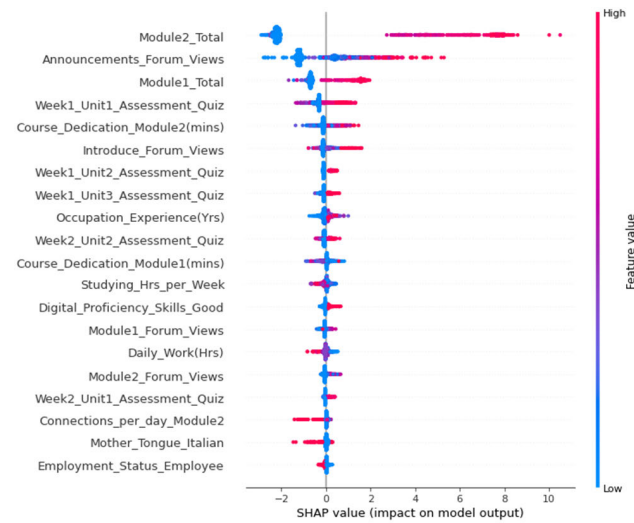


FIGURE 3. SHAP value (impact on LightGBM model output).

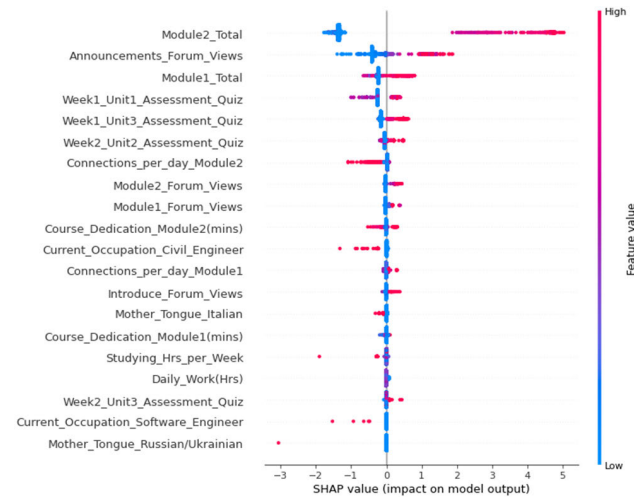


FIGURE 4. SHAP value (impact on GB model output).

TABLE 10. Attributes used in each step.

Step 1	Number of views in introductory forum
	Number of views in announcements forum
Step 2	Number of views in announcements forum
	Overall grade in Module_1
Step 3	Number of views in announcements forum
	Overall grade in Module_2

In conclusion, the results reveal that the produced predictive models could serve as an early alert system for identifying students at risk of failure from the very beginning of the course. Therefore, a series of targeted support actions and intervention strategies could be implemented to improve student learning outcomes [42]. It is important the prediction and therefore the intervention take place early enough, after the first two weeks of the course, thus providing the ability

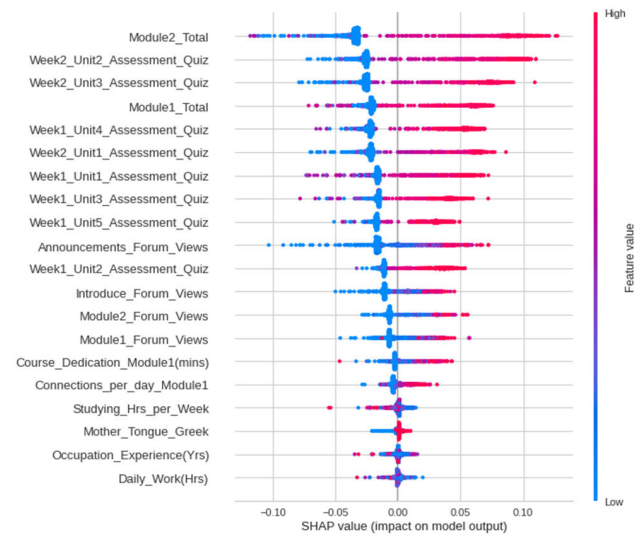


FIGURE 5. SHAP value (impact on RF model output).

TABLE 11. Performance results of predictive models (week 0).

Model	Acc	AUC	Recall	Precision	F1	Kappa	MCC
AdaBoost	0,8671	0,9306	0,7081	0,7372	0,7208	0,6339	0,6352
GB	0,8576	0,9298	0,6600	0,7245	0,6873	0,5964	0,5995
Cart	0,8362	0,7482	0,5531	0,7250	0,6182	0,5178	0,5309
Extra	0,8480	0,8289	0,5960	0,7393	0,6539	0,5589	0,5680
LDA	0,8659	0,9445	0,5129	0,8908	0,6452	0,5715	0,6072
LightGBM	0,8600	0,9226	0,6645	0,7361	0,6941	0,6046	0,6087
LR	0,8778	0,9445	0,6443	0,8204	0,7185	0,6423	0,6521
RF	0,8469	0,9003	0,6348	0,7173	0,6684	0,5701	0,5753

TABLE 12. Performance results of predictive models (week 1).

Model	Acc	AUC	Recall	Precision	F1	Kappa	MCC
AdaBoost	0,9371	0,9785	0,8957	0,8613	0,8739	0,8323	0,8362
GB	0,9336	0,9802	0,8912	0,8515	0,8675	0,8234	0,8267
Cart	0,9111	0,9041	0,8221	0,8185	0,8177	0,7591	0,7612
Extra	0,9206	0,9427	0,8669	0,8215	0,8415	0,7887	0,7909
LDA	0,9170	0,9834	0,9950	0,7489	0,8537	0,7976	0,8140
LightGBM	0,9360	0,9786	0,9060	0,8500	0,8733	0,8308	0,8347
LR	0,9395	0,9830	0,9210	0,8474	0,8810	0,8407	0,8434
RF	0,9230	0,9740	0,8814	0,8207	0,8475	0,7963	0,7993

TABLE 13. Performance results of predictive models (week 2).

Model	Acc	AUC	Recall	Precision	F1	Kappa	MCC
AdaBoost	0,9430	0,9826	0,9190	0,8682	0,8891	0,8510	0,8548
GB	0,9430	0,9768	0,9048	0,8776	0,8882	0,8501	0,8526
Cart	0,9323	0,9116	0,8524	0,8765	0,8618	0,8171	0,8192
Extra	0,9371	0,9555	0,8905	0,8664	0,8759	0,8338	0,8360
LDA	0,9324	0,9865	1,0000	0,7912	0,8822	0,8360	0,8483
LightGBM	0,9395	0,9809	0,8952	0,8711	0,8803	0,8399	0,8423
LR	0,9466	0,9866	0,9238	0,8722	0,8962	0,8603	0,8619
RF	0,9347	0,9790	0,8857	0,8640	0,8720	0,8283	0,8308

to educators to measure its effectiveness in addressing the learning needs and disabilities of learners during the next weeks of the course [43].

VI. CONCLUSION

In the present study, an attempt was made to exploit predictive analytics for predicting student certification in a MOOC for smart city professionals. Therefore, a plethora of high-performing ML models were produced employing a variety of classification algorithms. The experiments were performed in three consecutive time-steps in correspondence with the first two, out of eleven weeks of the course.

The results revealed that students at risk of failure can be identified with an accuracy measure greater than 94% at the end of the second week of the course. Hence, each model may serve as an early alert system for educators, since students less likely to obtain a certificate could be identified at an early enough stage to provide sufficient support actions and targeted intervention strategies to them. Moreover, it is of vital importance for universities to increase retention rates in MOOCs and offer high-quality education to learners. Furthermore, we identified the features which are of great impact on earning a course certificate. Our findings indicate that performance attributes regarding student grading in quizzes and activities during the first two weeks of the course are of great impact for accurately identifying low performers.

Although our research is a case study, the findings are in line with recent research. Explainable learning models could provide more meticulous and expressive information about student learning behavior and performance [19]. To this end, for future work we intend to build predictive models based on the most informative features of the dataset and apply them to other MOOCs. In addition, we intend to experiment with new ML methods, such as Semi-Supervised Learning, which has been proven to be very effective in the EDM field [44].

REFERENCES

- [1] E. Handoko, S. L. Gronseth, S. G. McNeil, C. J. Bonk, and B. R. Robin, "Goal setting and MOOC completion: A study on the role of self-regulated learning in student performance in massive open online courses," *Int. Rev. Res. Open Distrib. Learn.*, vol. 20, no. 3, pp. 1–21, Jul. 2019.
- [2] F. Dalipi, A. S. Imran, and Z. Kastrati, "MOOC dropout prediction using machine learning techniques: Review and research challenges," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Apr. 2018, pp. 1007–1014.
- [3] R. L. Moore and C. Wang, "Influence of learner motivational dispositions on MOOC completion," *J. Comput. Higher Educ.*, vol. 33, no. 1, pp. 121–134, Apr. 2021.
- [4] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. D. Kloos, "Prediction in MOOCs: A review and future research directions," *IEEE Trans. Learn. Technol.*, vol. 12, no. 3, pp. 384–401, Jul. 2019.
- [5] A. Pardo, O. Poquet, R. Martínez-Maldonado, and S. Dawson, "Provision of data-driven student feedback in LA & EDM," in *Handbook of Learning Analytics*. AL, Canada: Soc. Learn. Anal. Res., 2017, pp. 163–174.
- [6] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning Analytics*. New York, NY, USA: Springer, 2014, pp. 61–75.
- [7] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 10, no. 3, p. e1355, May 2020.
- [8] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 3, no. 1, pp. 12–27, Jan. 2013, doi: [10.1002/WIDM.1075](https://doi.org/10.1002/WIDM.1075).
- [9] D. Clow, "An overview of learning analytics," *Teach. Higher Educ.*, vol. 18, no. 6, pp. 683–695, 2013.
- [10] G. Siemens, "Learning analytics: Envisioning a research discipline and a domain of practice," in *Proc. 2nd Int. Conf. Learn. Anal. Knowl. (LAK)*, 2012, pp. 4–8.
- [11] C. Brooks and C. Thompson, "Predictive modelling in teaching and learning," in *Handbook of Learning Analytics*. AL, Canada: Soc. Learn. Anal. Res., 2017, pp. 61–68.
- [12] G. Shmueli and O. R. Koppius, "Predictive analytics in information systems research," *MIS Quart.*, vol. 35, no. 3, pp. 553–572, Sep. 2011.
- [13] J. Maldonado-Mahauad, M. Pérez-Sanagustín, P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. Delgado-Kloos, *Predicting Learners' Success in a Self-Paced MOOC Through Sequence Patterns of Self-Regulated Learning* (Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11082. Cham, Switzerland: Springer, 2018, pp. 355–369, doi: [10.1007/978-3-319-98572-5_27](https://doi.org/10.1007/978-3-319-98572-5_27).
- [14] J. Gardner and C. Brooks, "Student success prediction in MOOCs," *User Model. User-Adapted Interact.*, vol. 28, no. 2, pp. 127–203, Jun. 2018.
- [15] K. Jordan. (2015). *MOOC Completion Rates: The Data*. [Online]. Available: <http://www.katyjordan.com/MOOCproject.html>
- [16] K. Jordan, "Massive open online course completion rates revisited: Assessment, length and attrition," *Int. Rev. Res. Open Distrib. Learn.*, vol. 16, no. 3, pp. 341–358, Jun. 2015.
- [17] H. Khalil and M. Ebner, "MOOCs completion rates and possible methods to improve retention—A literature review," in *Proc. EdMedia+ Innovate Learn.*, 2014, pp. 1305–1313.
- [18] J. Reich and J. A. Ruipérez-Valiente, "The MOOC pivot," *Science*, vol. 363, no. 6423, pp. 130–131, Jan. 2019.
- [19] M. Saarela, V. Heilala, P. Jaaskela, A. Rantakaulio, and T. Karkkainen, "Explainable student agency analytics," *IEEE Access*, vol. 9, pp. 137444–137459, 2021.
- [20] C. A. Coleman, D. T. Seaton, and I. Chuang, "Probabilistic use cases: Discovering behavioral patterns for predicting certification," in *Proc. 2nd ACM Conf. Learn. @ Scale*, Mar. 2015, pp. 141–148.
- [21] R. Al-Shabandar, A. Hussain, A. Laws, R. Keight, J. Lunn, and N. Radi, "Machine learning approaches to predict learning outcomes in massive open online courses," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 713–720.
- [22] J. A. Ruipérez-Valiente, R. Cobos, P. J. Muñoz-Merino, Á. Andujar, and C. D. Kloos, "Early prediction and variable importance of certificate accomplishment in a MOOC," in *Proc. Eur. Conf. Massive Open Online Courses*, 2017, pp. 263–272.
- [23] N. Gitinabard, F. Khoshnevisan, C. F. Lynch, and E. Y. Wang, "Your actions or your associates? Predicting certification and dropout in MOOCs with behavioral and social features," 2018, *arXiv:1809.00052*.
- [24] D. J. Lemay and T. Doleck, "Grade prediction of weekly assignments in MOOCs: Mining video-viewing behavior," *Educ. Inf. Technol.*, vol. 25, no. 2, pp. 1333–1342, Mar. 2020.
- [25] C.-H. Yu, J. Wu, and A.-C. Liu, "Predicting learning outcomes with MOOC clickstreams," *Educ. Sci.*, vol. 9, no. 2, p. 104, May 2019.
- [26] O. Iatrellis, T. Panagiotakopoulos, V. C. Gerogiannis, P. Fitsilis, and A. Kameas, "Cloud computing and semantic web technologies for ubiquitous management of smart cities-related competences," *Educ. Inf. Technol.*, vol. 26, no. 2, pp. 2143–2164, Mar. 2021.
- [27] N. Fazakis, G. Kostopoulos, S. Kotsiantis, and I. Mporas, "Iterative robust semi-supervised missing data imputation," *IEEE Access*, vol. 8, pp. 90555–90569, 2020, doi: [10.1109/ACCESS.2020.2994033](https://doi.org/10.1109/ACCESS.2020.2994033).
- [28] M. Ali, "PyCaret: An open source, low-code machine learning library in Python," Apr. 2020. [Online]. Available: <https://www.pycaret.org>
- [29] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, pp. 119–139, Aug. 1995.
- [30] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
- [31] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.
- [32] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, pp. 3–42, Apr. 2006.
- [33] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, vol. 544. Hoboken, NJ, USA: Wiley, 2004.
- [34] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3146–3154.
- [35] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 841–848.
- [36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324).

- [37] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, Jan. 2002.
- [38] M. Stone, "Cross-validated choice and assessment of statistical predictions," *J. Roy. Stat. Soc. B, Methodol.*, vol. 36, no. 2, pp. 111–133, 1974.
- [39] P. Burman, "A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods," *Biometrika*, vol. 76, no. 3, pp. 503–514, Sep. 1989.
- [40] B. Liu, Y. Wu, W. Xing, G. Cheng, and S. Guo, "Exploring behavioural differences between certificate achievers and explorers in MOOCs," *Asia Pacific J. Educ.*, pp. 1–13, 2021, doi: [10.1080/02188791.2020.1868974](https://doi.org/10.1080/02188791.2020.1868974).
- [41] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
- [42] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauría, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: An open source analytics initiative," *J. Learn. Anal.*, vol. 1, no. 1, pp. 6–47, 2014.
- [43] D. Bevitt, C. Baldwin, and J. Calvert, "Intervening early: Attendance and performance monitoring as a trigger for first year support in the biosciences," *Biosci. Educ.*, vol. 15, no. 1, pp. 1–14, Jun. 2010.
- [44] G. Kostopoulos and S. Kotsiantis, "Exploiting semi-supervised learning in the education field: A critical survey," in *Advances in Machine Learning/Deep Learning-Based Technologies*. Cham, Switzerland: Springer, 2022, pp. 79–94.



is 12. His research interests include machine learning, learning analytics, and educational data mining.

GEORGIOS KOSTOPOULOS was born in Agrinio, Greece, in 1971. He received the master's degree in computer science and the Ph.D. degree in educational data mining from the Department of Mathematics, University of Patras, Greece, in 2015 and 2020, respectively. He is currently a High School Mathematics Teacher and teaching at the University of Peloponnese. He has several publications with enough citations. According to Google Scholar, his H-index is 11 and H10-index



SOTIRIS KOTSIANTIS received the B.Sc., master's, and Ph.D. degrees in computer science from the University of Patras, Patras, Greece, in 1999, 2001, and 2005, respectively. He is a Mathematician. He is currently an Assistant Professor with the Department of Mathematics, University of Patras. According to Google Scholar, his H-index is 34 and G-index is 118. He has a lot of publications to his credit in international journals and conferences. His research interests include data mining, machine learning, and learning analytics.



CHRISTOS PIERRAKEAS received the B.Sc. degree in mathematics and the Ph.D. degree in medical informatics from the University of Patras, Greece, in 1986 and 1994, respectively. He also holds two postgraduate certificates in "open and distance learning" and "adults education" from Hellenic Open University (HOU). He is currently an Associate Professor of design, analysis, and development of information technologies, with emphasis on educational technology, with the Department of Management Science and Technology, University of Patras, and has been a Tutor with the Department of Informatics, HOU, since 2000. His research interests include educational technology, applications of new ICT technologies in education, applications of innovative technologies in education (e-learning systems, tools, techniques, methodologies, applications in MOOC, and STEM/STEM education), digital competences development (tools, techniques, applications, and methodologies), development and evaluation of educational material and educational processes, user modeling and learning analytics, design and development of information (and educational information) systems, and distance education.



nals, and conferences, and has participated in 15 European and national research programs holding key positions at a research, technical, and managerial level. His research interests include pervasive computing, the Internet of Things, machine learning, smart city applications, mobile health, fuzzy systems, and digital literacy.

THEODOR PANAGIOTAKOPOULOS received the Diploma and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Patras, Greece, in 2006 and 2011, respectively. He is currently a Senior Research Fellow with the Mobile and Pervasive Computing, Quality, and Ambient Intelligence Laboratory, School of Science and Technology, Hellenic Open University. He has published more than 30 scientific articles in international book chapters, journals, and conferences, and has participated in 15 European and national research programs holding key positions at a research, technical, and managerial level. His research interests include pervasive computing, the Internet of Things, machine learning, smart city applications, mobile health, fuzzy systems, and digital literacy.



ACHILLES KAMEAS received the Engineering Diploma and Ph.D. degrees in human-computer interaction from the Department of Computer Engineering and Informatics, University of Patras, Greece, in 1989 and 1995, respectively. He also received formal education on adult education and open and distance education. He is a Professor at Hellenic Open University.

...