





Article

Early Dropout Prediction in MOOCs through Supervised Learning and Hyperparameter Optimization

Theodor Panagiotakopoulos ^{1,2}, Sotiris Kotsiantis ^{3,*}, Georgios Kostopoulos ³, Omiros Iatrellis ⁴
and Achilles Kameas ¹

¹ School of Science and Technology, Hellenic Open University, 26335 Patras, Greece; panagiotakopoulos@eap.gr (T.P.); kameas@eap.gr (A.K.)

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

³ Department of Mathematics, University of Patras, 26500 Patras, Greece; kostg@sch.gr

⁴ Department of Digital Systems, University of Thessaly, 41500 Larissa, Greece; iatrellis@hotmail.com

* Correspondence: kotsiantis@upatras.gr

Abstract: Over recent years, massive open online courses (MOOCs) have gained increasing popularity in the field of online education. Students with different needs and learning specificities are able to attend a wide range of specialized online courses offered by universities and educational institutions. As a result, large amounts of data regarding students' demographic characteristics, activity patterns, and learning performances are generated and stored in institutional repositories on a daily basis. Unfortunately, a key issue in MOOCs is low completion rates, which directly affect student success. Therefore, it is of utmost importance for educational institutions and faculty members to find more effective practices and reduce non-completer ratios. In this context, the main purpose of the present study is to employ a plethora of state-of-the-art supervised machine learning algorithms for predicting student dropout in a MOOC for smart city professionals at an early stage. The experimental results show that accuracy exceeds 96% based on data collected during the first week of the course, thus enabling effective intervention strategies and support actions.

Keywords: MOOCs; smart cities; completion rates; dropout; early prediction; supervised learning; classification models



Citation: Panagiotakopoulos, T.; Kotsiantis, S.; Kostopoulos, G.; Iatrellis, O.; Kameas, A. Early Dropout Prediction in MOOCs through Supervised Learning and Hyperparameter Optimization. *Electronics* **2021**, *10*, 1701. <https://doi.org/10.3390/electronics1014>

Academic Editor:
Krzysztof Szczypiorski

Received: 22 June 2021
Accepted: 13 July 2021
Published: 16 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Over recent years, massive open online courses (MOOCs) have gained increasing popularity in the field of online education. Students with different needs and learning specificities are able to attend flexible and high-quality online courses offered by universities and educational institutions [1]. These courses vary significantly from the traditional online courses delivered by higher education institutions, particularly in terms of course length and content structure [2]. As a result, large amounts of data regarding students' demographic characteristics, activity patterns, and learning performance are generated and stored in institutional databases and repositories on a daily basis.

However, despite the potential benefits of MOOCs, these courses are characterized by low retention rates, which directly affect student success [3]. A number of studies have highlighted various factors influencing retention rates in MOOCs. Student motivation, challenge, future economic profit, growth of personal and professional identity, insufficient background knowledge, and lack of time have a significant impact on preventing students from completing a MOOC. Additionally, lack of mobile-friendly features (e.g., inability to watch videos via smart phones), lack of interaction with peers and instructors, and difficulty in following the language of the instructor have a major influence on student dropout [4]. Another key factor is course design, which comprises three components: course content, course structure, and information delivery technology. Among these components, course content is the most significant predictor of MOOC dropout [3].

Consequently, it is of utmost importance for educational institutions and faculty members to find more effective practices, provide efficient intervention strategies, support low performers, and decrease non-completers. Educational data mining (EDM) is the appropriate tool for efficiently analyzing students' learning behavior and predicting their performance. EDM is a fast-growing research field that is mainly focused on the implementation of data mining methods in educational settings for improving teaching and learning [5]. Predicting student dropout in distance education is considered to be one of the most important EDM tasks. In this context, the main purpose of the present study is to employ a plethora of state-of-the-art supervised ML algorithms for predicting student dropout in a three-month MOOC at an early stage. A plethora of experiments are conducted measuring the values of various metrics, such as accuracy, F1-score, and kappa. Further, it is examined whether an accurate prediction could be done in sufficient time to provide effective intervention strategies for at-risk students. The experimental results show a high degree of accuracy based on data collected during the first week of the course, thus enabling properly targeted support for potential non-completers.

The rest of the paper is organized as follows: Section 2 reviews recent studies concerning the implementation of ML techniques for detecting high-risk MOOC students. Section 3 provides a description of the dataset, while Section 4 presents the experimental procedure and a thorough analysis of the results obtained. The paper concludes by summarizing the most important elements of the study and considering some thoughts for future research directions.

2. Related Work

A number of studies have examined the impact of learners' characteristics on MOOC retention rates employing a variety of statistical methods. Guo and Reinecke (2014) analyzed student activity in four edX MOOCs [6]. To this end, data comprising 140,546 students were evaluated using multiple linear regression. The findings revealed that students usually ignore the linear structure of the learning content, while age and grade were found to correlate positively with the volume of the learning material studied by a learner. In a similar study, Cisel (2014) attempted to identify the indicators that significantly affect student completion rates in a French xMOOC [2]. Data regarding 3029 registered students were analyzed using the R statistical package. The experimental results indicated that completion rates are mainly dependent on employment status and time limitations. Additionally, student active participation in forums was found to enhance their overall performance. In the same context, Morris et al. (2015) explored the existence of relationships between completion and several demographic characteristics of students enrolled in five MOOCs using nonparametric methods [7]. The level of completion was found to be closely linked to prior online experience and educational attainment, employment status, and age of participants.

Several ML approaches have been adopted for detecting students who are likely to drop out from a MOOC. Kizilcec et al. (2013) applied a clustering method for categorizing students according to their engagement patterns [8]. The most notable cluster included learners who remained engaged through the course without taking assessments. Similarly, a systematic investigation of MOOC dropout was conducted using logistic regression (LR) on data from 100,000 students enrolled in 21 courses [9]. The study yielded that the probability of student dropout increases substantially if students disengage for 14 days or more. In addition, the probability of re-engaging increases with the number of released videos that students have viewed before the absence. LR was also used for predicting student dropout in MOOCs including feature generation and feature selection methods [10].

Feng et al. (2019) proposed a context-based feature interaction network (CFIN) for predicting the potential dropout students enrolled in two MOOCs [11]. Various experiments were conducted evaluating the effectiveness of CFIN against familiar classification methods, such as LR, support vector machines (SVM), and random forest (RF). In addition, an ensemble method was designed by combining CFIN with XGBoost. CFIN prevailed

over baseline methods in terms of area under the ROC curve (AUC) and F1-score. In this connection, the RF ensemble method was utilized with a view to finding the most important features that influence students' performance [12]. A set of familiar ML algorithms was applied for predicting student performance in MOOCs. The results revealed that RF prevailed in terms of accuracy, sensitivity, and Cohen's kappa coefficient (Kappa).

Deep neural networks (DNNs) were also used for predicting student success and dropout in a MOOC [13]. Several predictive models were built after applying multivariate analysis for selecting the most important features of students. In addition, association rules were extracted for discovering similarities in student behavior patterns. A feed-forward DNN was also employed for addressing the dropout problem in MOOCs [14]. The produced model achieved high accuracy and a low false-negative rate compared with familiar classification methods.

Liang et al. applied gradient boosting for predicting student dropout in MOOCs [15]. Hence, data regarding students' learning activities from 39 courses were collected and exploited for creating decision tree models with an accuracy of 89%. Very recently, a support vector regression model incorporating an improved quantum particle swarm optimization algorithm was applied for the same task [16]. The model was based on students' learning behavior data collected on a weekly basis and achieved better predictive performance compared with other classification methods.

3. Dataset Description

The study was performed in the context of the Erasmus+ Sector Skills Alliances project "DevOps Competences for Smart Cities" (<https://devops.uth.gr/dev/>, accessed on 20 June 2021). DevOps focuses on equipping current and prospective professionals in municipalities and regional authorities with appropriate competences to support the emerging smart city concepts, needs, and requirements [17]. Registrations in the DevOps MOOC (<https://devops.uth.gr/dev/about-the-mooc/>, <https://smartdevopsmooc.eu/moodle/pages/login.php>, accessed on 20 June 2021) lasted from 15 September to 15 October 2020, while the course started on 19 October 2020. It lasted approximately 3 months, and it was structured on a weekly format delivering one or two training modules (i.e., competences) per week. Each training module—available in English—comprised two to five learning units, each of which included an automatically graded assessment test. The course content was designed to address the European Qualifications Framework level 5, as this is the required level of autonomy and responsibility for smart city professionals. The registration form included a questionnaire asking applicants to provide personal and demographic data and notifying them that all data would be acquired and used according to the General Data Protection Regulation (EU 2016/679) for evaluating the quality of the DevOps MOOC. All applicants were requested to provide their consent to store and use these data; otherwise, they could skip the questionnaire and proceed with the registration providing only their full name and email.

The initial dataset used in the study included the following attributes regarding a wide range of personal and demographic information of students: gender, age, nationality, country of residence, mother tongue, education level, current employment status, current job role or occupation, years of experience in the role/occupation, average amount of daily working hours, level of technical English language skills, current digital proficiency, number of underage children, available amount of study hours on a weekly basis, and prior MOOC attendance. Figures 1–3 depict the interdependent relationship between two demographic attributes each time and their impact on MOOC dropout.

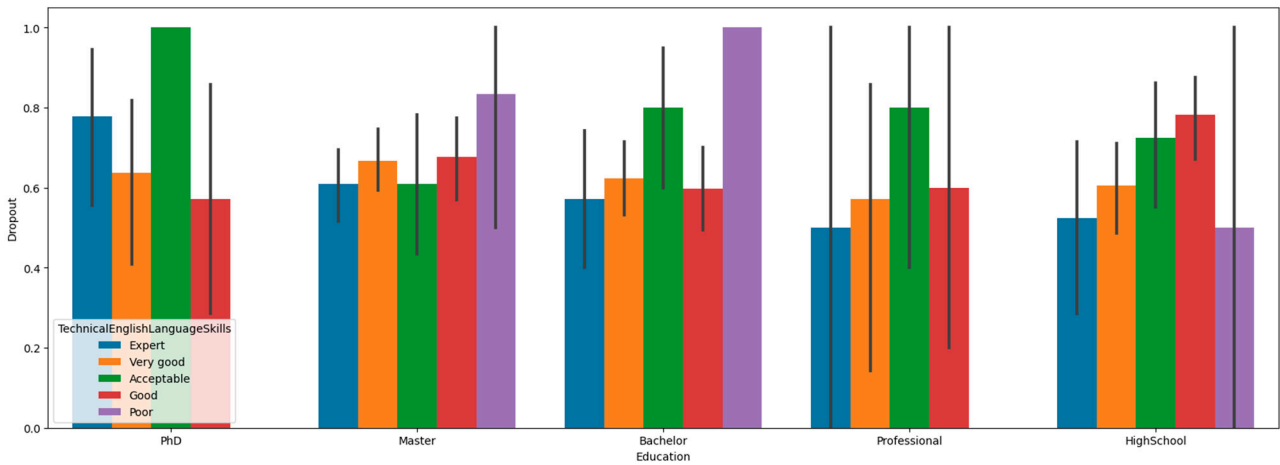


Figure 1. The interdependent relationship between current technical English language skills and education level and their impact on MOOC dropout.

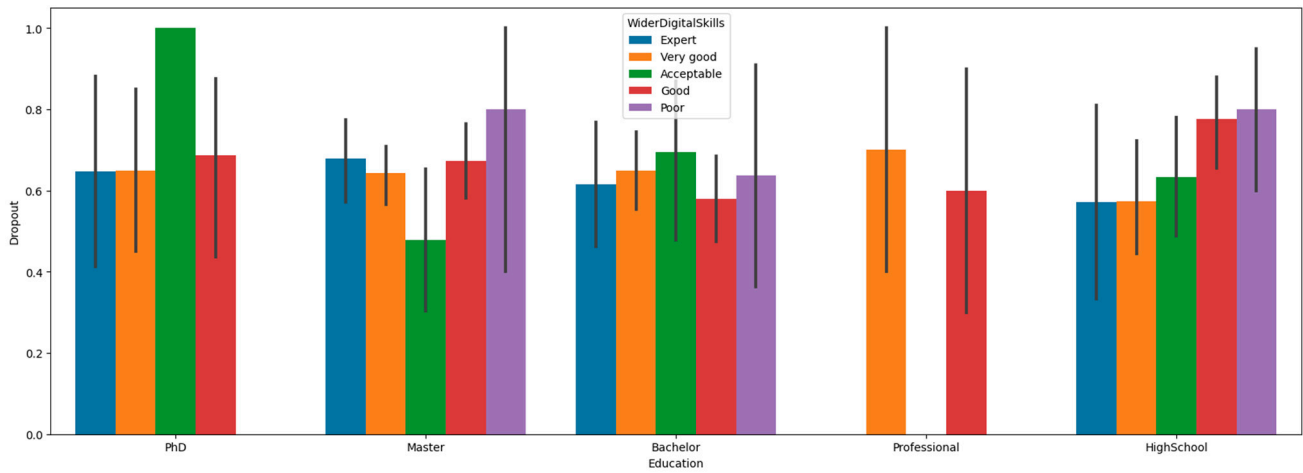


Figure 2. The interdependent relationship between current digital proficiency and education level and their impact on MOOC dropout.

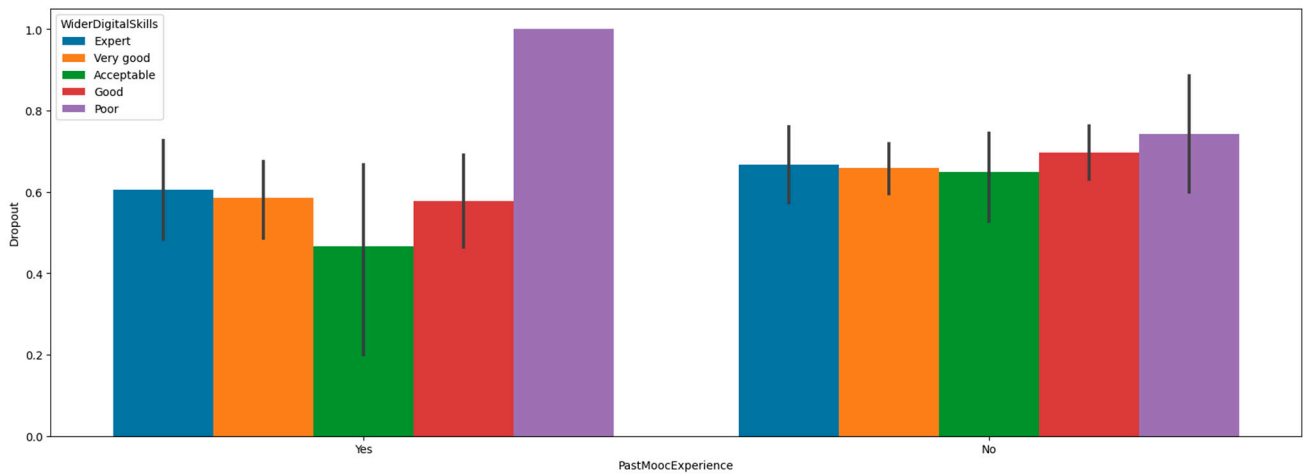


Figure 3. The interdependent relationship between current digital proficiency and previous MOOC experience and their impact on MOOC dropout.

Figure 4 provides additional information regarding the dropout rates in the dataset depending (Figure 4a) on the education level and (Figure 4b) on the current digital profi-

ciency skills. Furthermore, Figure 5 presents a correlation matrix heatmap for the dataset used in the study, where each correlation is shown by color. The red color indicates positive correlation, and the blue one negative. The deeper the color, the larger the correlation between two attributes. Overall, weak correlations appear between the attributes. The strongest positively correlated pairs are {Quiz Week 1 Unit 1 Assessment, Quiz Week 1 Unit 2 Assessment} and {Connections per Day Module 1, Course Dedication in Mins Module 1}.

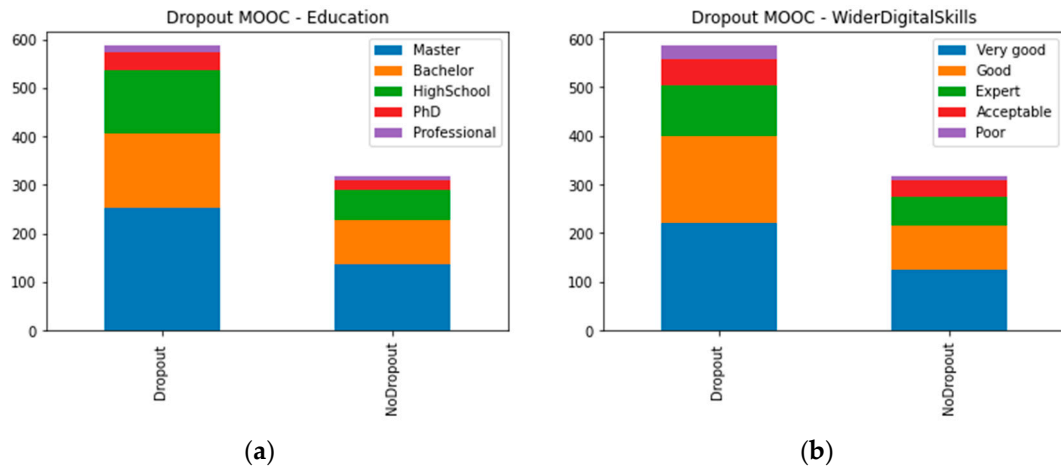


Figure 4. Dropout rates depending (a) on education level and (b) on current digital proficiency.

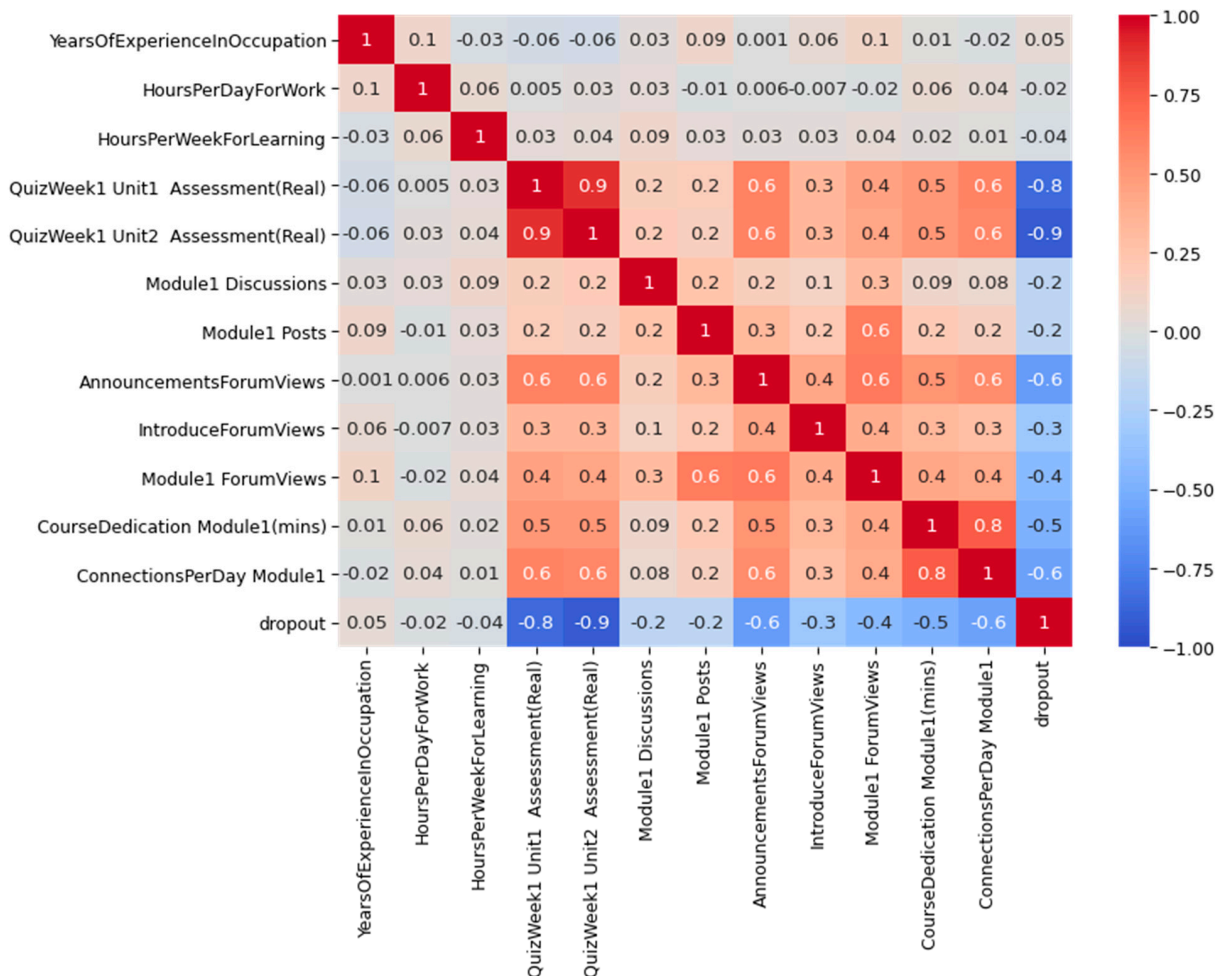


Figure 5. Correlation matrix heatmap.

4. Experimental Process and Results

A set of nine classification algorithms were used during the experimental process for building nine corresponding predictive models using PyCaret [18], an open-source ML library in Python. In addition, hyperparameter optimization was automatically performed on a set of optimal hyperparameters for all the examined learning algorithms using random search, which chooses random groupings of hyperparameters for training the learning model. These algorithms are:

- LightGBM, a gradient boosting decision tree implementation [19];
- Extremely randomized trees (Extra) algorithm [20];
- Ridge classification method (Ridge) [21];
- Gradient boosting classifier (GBC) [22];
- Random Forest (RF) ensemble method [23];
- Logistic regression (LR) [24];
- Classification and regression tree (CART) algorithm [25];
- AdaBoost boosting algorithm [26];
- Linear SVM with stochastic gradient descent (SVM-SGD) algorithm [27].

The initial dataset was aggregated with attributes regarding (a) student performance in the first week of the course (i.e., quiz week 1 unit 1, quiz week 1 unit 2) and (b) student activity in the online learning platform during the first week of the course (i.e., forum views, logins, and dedication time). All the attributes were acquired through custom software querying the Moodle database directly, except for the dedication time attribute, where the course dedication plugin (https://moodle.org/plugins/block_dedication, accessed on 15 June 2021) was implemented.

For evaluating the performance of the models, the 10-fold validation resampling technique was used [28] while calculating six metrics: accuracy, recall, precision, F1-score, kappa, and Matthews correlation coefficient (MCC). The results are shown in Table 1, whereas the best value for each metric is highlighted in bold font. Overall, it is observed that LightGBM is the top-performing model. Accuracy and F1-score range from 91% to 95.58% and 93.16% to 96.34%, respectively, showing that a very accurate prediction of potential dropout students could be performed after the first week of the course.

Table 1. Experimental results.

Classifier	Accuracy	Recall	Precision	F1-Score	Kappa	MCC
LightGBM	0.9558	0.9507	0.9777	0.9634	0.9076	0.9097
Extra	0.9497	0.9434	0.9755	0.9585	0.8946	0.8972
Ridge	0.9497	0.9483	0.9704	0.9587	0.8944	0.8964
GBC	0.9450	0.9482	0.9637	0.9551	0.8841	0.8866
RF	0.9435	0.9434	0.9656	0.9537	0.8812	0.8831
LR	0.9421	0.9459	0.9614	0.9530	0.8776	0.8793
CART	0.9405	0.9532	0.9531	0.9522	0.8732	0.8763
AdaBoost	0.9298	0.9384	0.9491	0.9429	0.8517	0.8543
SVM-SGD	0.9100	0.9315	0.9376	0.9316	0.7983	0.8097

Figure 6 illustrates the learning curve for each classification model along with the most important features for making a prediction. Additionally, the relative importance score for each attribute is recorded in descending order. The most important attributes are “Quiz Week1 Unit1 Assessment”, “Quiz Week1 Unit2 Assessment”, “Announcements Forum Views”, and “Introduce Forum Views”. Finally, it is seen that all models achieve high accuracy using only 100 instances during the training phase.

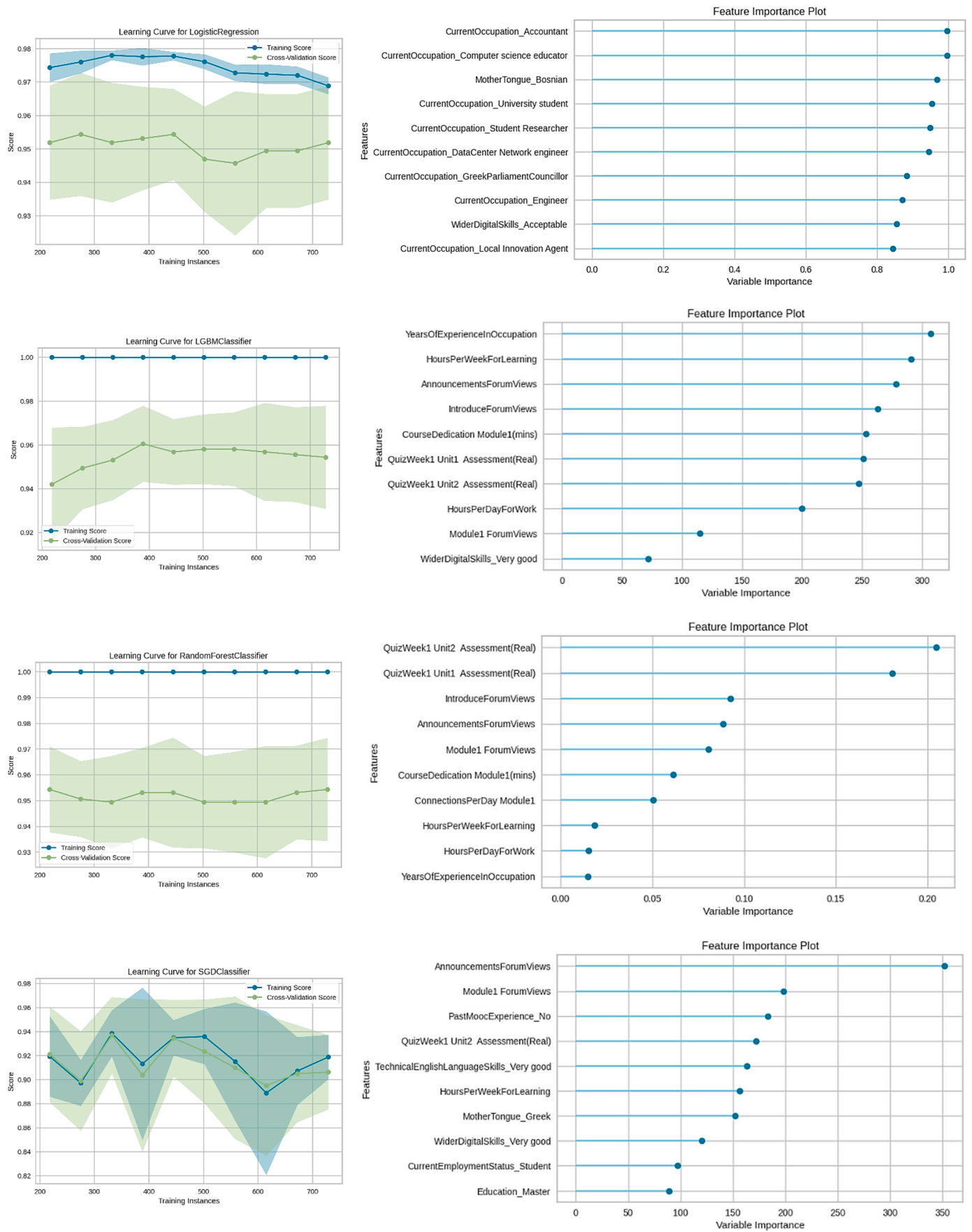


Figure 6. Cont.

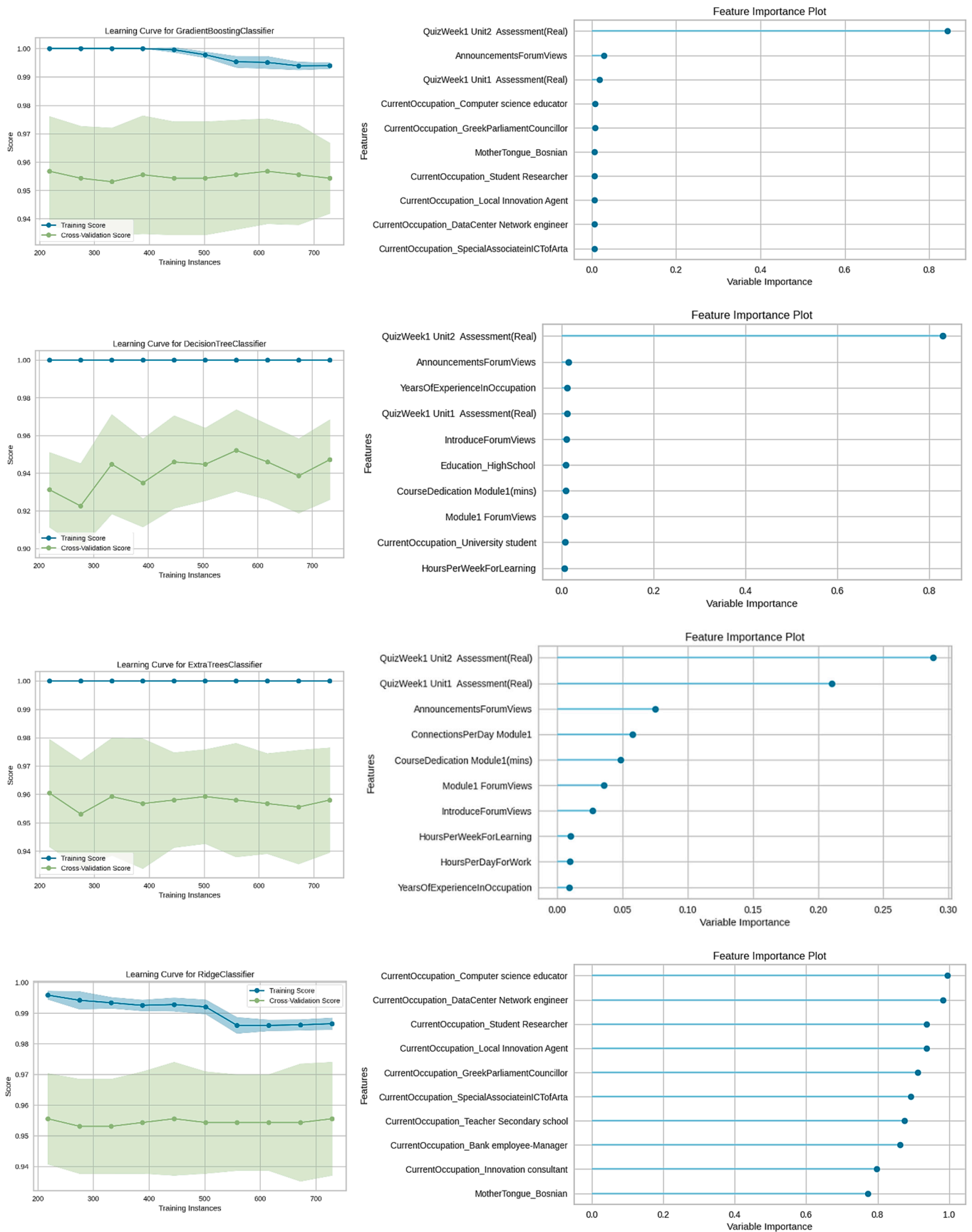


Figure 6. Cont.

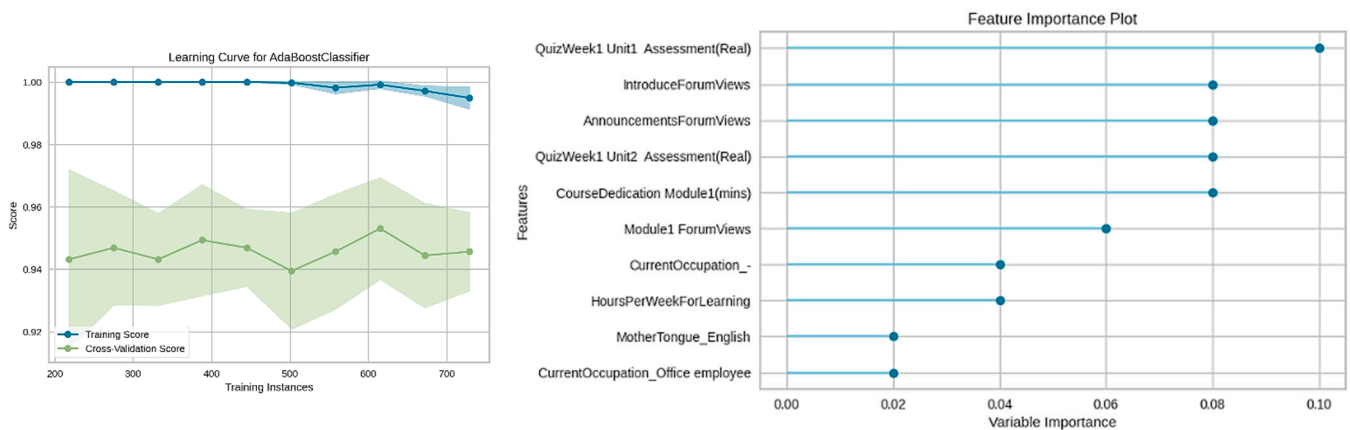


Figure 6. Learning curve for each algorithm and the relative importance score for each attribute.

To further improve the predictive accuracy, a stacked generalization approach, namely, stacking, was employed [29]. Stacking involves the creation of a strong and high-level classification model with superior generalized performance combining a set of different classifiers. To achieve an appropriate combination of the base predictions, a learning algorithm was employed, while the model utilized the following steps in the final forecast: (a) Level 0 data: all the base learners ran on the original dataset. (b) Level 1 data: after the 0 level, the predictions made by the classifiers were considered new data. (c) Final prediction: another learning process used the level 1 data as new inputs and as output, and the final prediction was gained. The default parameters of the learning algorithms were used during the stacked generalization for simplifying the process, and the k -nearest neighbor (k -NN) algorithm [30] as meta-learner. The results are shown in Table 2.

Table 2. Stacking results.

Classifier	Accuracy	Recall	Precision	F1-Score	Kappa	MCC
Stacking	0.9604	0.9632	0.9730	0.9677	0.9165	0.9175

5. Conclusions

The main purpose of the present study was to employ a plethora of state-of-the-art supervised ML algorithms for predicting student dropout in a MOOC. Several predictive models were produced and evaluated in terms of six well-known metrics. In addition, hyperparameter optimization was automatically performed to improve the performance of the learning algorithms through random search. The results indicated a high degree of accuracy based on data collected during the first week of the course. What is more, a stacked generalization approach was applied to further improve the classification performance. It was observed that using the default parameters of the learning algorithms on a stacked generalization procedure, results better than any single tuned learning algorithm were produced. Therefore, students who are prone to failure can be accurately predicted with an accuracy value exceeding 96%. Additionally, students' interaction data provide more information than their demographic data.

Being able to know and predict early on the training cycle of those more likely to drop out of the course is quite important for MOOC providers to create and implement timely learner engagement strategies. More personalized content and support could be offered, especially for people that relate their learning with career advancement and decreased available time for learning; microlearning [31] and micro-credentials could be employed for those people as well, opening up education to more people as they are supported by natural flexibility and inclusiveness. A complementary list of alternative candidates could be exploited so as to replace those who are likely to drop out. Moreover, alternative

registration policies could be coined, and personalized learning paths could be offered addressing different learning behaviors.

Author Contributions: Conceptualization, T.P.; methodology, S.K. and G.K.; validation, O.I. and A.K.; formal analysis, T.P. and O.I.; investigation, S.K. and T.P.; resources, A.K. and O.I.; writing—original draft preparation, T.P.; writing—review and editing, G.K.; visualization, S.K. and G.K.; supervision, S.K. and A.K.; project administration, A.K. and O.I.; funding acquisition, A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This paper received funding from the research project DevOps, “DevOps competences for Smart Cities” (Project No.: 601015-EPP-1-2018-1-EL-EPPKA2-SSA, Erasmus+ Program, KA2: Cooperation for innovation and the exchange of good practices-SSA).

Conflicts of Interest: The authors declare no conflict of interest.

References

- Dalipi, F.; Imran, A.S.; Kastrati, Z. MOOC dropout prediction using machine learning techniques: Review and research challenges. In Proceedings of the 2018 IEEE Global Engineering Education Conference (EDUCON), Canary Islands, Spain, 18–20 April 2018; pp. 1007–1014.
- Cisel, M. Analyzing completion rates in the first French xMOOC. *Proc. Eur. MOOC Stakehold. Summit* **2014**, *26*, 51.
- Hone, K.S.; El Said, G.R. Exploring the factors affecting MOOC retention: A survey study. *Comput. Educ.* **2016**, *98*, 157–168. [[CrossRef](#)]
- Bote-Lorenzo, M.L.; Gómez-Sánchez, E. Predicting the decrease of engagement indicators in a MOOC. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference, Vancouver, BC, Canada, 13–17 March 2017; pp. 143–147.
- Romero, C.; Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1355. [[CrossRef](#)]
- Guo, P.J.; Reinecke, K. Demographic differences in how students navigate through MOOCs. In Proceedings of the First ACM Conference on Learning@Scale Conference, Atlanta, GA, USA, 4–5 March 2014; pp. 21–30.
- Morris, N.P.; Swinnerton, B.J.; Hotchkiss, S. Can demographic information predict MOOC learner outcomes? In Proceedings of the Experience Track: Proceedings of the European MOOC Stakeholder, Mons, Belgium, 18–20 May 2015.
- Kizilcec, R.F.; Piech, C.; Schneider, E. Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. In Proceedings of the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium, 8–13 April 2013; pp. 170–179.
- Kizilcec, R.F.; Halawa, S. Attrition and achievement gaps in online learning. In Proceedings of the Second (2015) ACM Conference on Learning@Scale, Vancouver, BC, Canada, 14–18 March 2015; pp. 57–66.
- Qiu, L.; Liu, Y.; Liu, Y. An integrated framework with feature selection for dropout prediction in massive open online courses. *IEEE Access* **2018**, *6*, 71474–71484. [[CrossRef](#)]
- Feng, W.; Tang, J.; Liu, T.X. Understanding dropouts in MOOCs. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 517–524.
- Al-Shabandar, R.; Hussain, A.; Laws, A.; Keight, R.; Lunn, J.; Radi, N. Machine learning approaches to predict learning outcomes in Massive open online courses. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 713–720.
- Mourdi, Y.; Sadgal, M.; El Kabtane, H.; Fathi, W.B. A machine learning-based methodology to predict learners’ dropout, success or failure in MOOCs. *Int. J. Web Inf. Syst.* **2019**, *15*, 489–509. [[CrossRef](#)]
- Imran, A.S.; Dalipi, F.; Kastrati, Z. Predicting student dropout in a MOOC: An evaluation of a deep neural network model. In Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, Bali, Indonesia, 19–22 April 2019; pp. 190–195.
- Liang, J.; Li, C.; Zheng, L. Machine learning application in MOOCs: Dropout prediction. In Proceedings of the 2016 11th International Conference on Computer Science & Education (ICCSE), Nagoya, Japan, 23–25 August 2016; pp. 52–57.
- Jin, C. MOOC student dropout prediction model based on learning behavior features and parameter optimization. *Interact. Learn. Environ.* **2020**, *1*–19. [[CrossRef](#)]
- Iatrellis, O.; Panagiotakopoulos, T.; Gerogiannis, V.C.; Fitsilis, P.; Kameas, A. Cloud computing and semantic web technologies for ubiquitous management of smart cities-related competences. *Educ. Inf. Technol.* **2021**, *26*, 2143–2164. [[CrossRef](#)]
- Ali, M. PyCaret: An Open Source, Low-Code Machine Learning Library in Python, PyCaret Version 2.3. Available online: <https://www.pycaret.org> (accessed on 15 June 2021).
- Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
- Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [[CrossRef](#)]
- Grüning, M.; Kropf, S. A ridge classification method for high-dimensional observations. In *From Data and Information Analysis to Knowledge Engineering*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 684–691.

22. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
23. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
24. Ng, A.Y.; Jordan, M.I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Proceedings of the Advances in Neural Information Processing Systems*, Burlington, MA, USA; 2002; pp. 841–848.
25. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
26. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [[CrossRef](#)]
27. Platt, J. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. 1998. Available online: https://www.researchgate.net/publication/2624239_Sequential_Minimal_Optimization_A_Fast_Algorithm_for_Training_Support_Vector_Machines (accessed on 15 June 2021).
28. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009.
29. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [[CrossRef](#)]
30. Aha, D.W. *Lazy Learning*; Springer: Berlin/Heidelberg, Germany, 2013.
31. Emerson, L.C.; Berge, Z.L. Microlearning: Knowledge management applications and competency-based training in the workplace. *UMBC Fac. Collect.* **2018**, *10*, 2.