Student Profile Modeling Using Boosting Algorithms

Touria Hamim, Faculty of Sciences Ben M'sick, University of Hassan II, Casablanca, Morocco

Faouzia Benabbou, Faculty of Sciences Ben M'sick, University of Hassan II, Casablanca, Morocco

Nawal Sael, Faculty of Sciences Ben M'sick, University of Hassan II, Casablanca, Morocco

ABSTRACT

The student profile has become an important component of education systems. Many education systems objectives, such as e-recommendation, e-orientation, e-recruitment, and dropout prediction are essentially based on the profile for decision support. Machine learning plays an important role in this context, where several studies have been carried out either for classification, prediction, or clustering purpose. In this paper, the authors present a comparative study of different boosting algorithms, which have been used successfully in many fields and for many purposes. In addition, the authors applied feature selection methods Fisher score, information gain combined with recursive feature elimination to enhance the preprocessing task and models' performances. Using multi-label dataset to predict the class of the student performance in mathematics. This article shows that the light gradient boosting machine (LightGBM) algorithm achieved the best performance when using information gain with recursive feature elimination method compared to the other boosting algorithms.

KEYWORDS

AdaBoost, Boosting Algorithms, CatBoost, Feature Selection, Fisher Score, Information Gain, LightGBM, Prediction, Recursive Feature Elimination, Student Profile, XGBoost

INTRODUCTION

With the development of e-learning platforms and the availability of learner tracks, several classification and prediction systems have emerged. The information collected on the learner allows to establish different profiles and therefore to propose adapted material. The learner profile describes several sides of the student, such as personal information, social situation, academic background, skills, personal characters, preferred learning styles, online behavior, etc. An example of a potential use in the academic field is the student failure and dropout prediction which are two serious problems in every educational system nowadays. To overcome these problems, several research studies have been carried out using different Machine learning techniques such as K Nearest Neighbor (kNN), Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayesian (NB), etc. Recently, with the availability of huge volume of data, many studies have been conducted based on Deep Learning techniques to make efficient classification and/or prediction in the academic domain.

Machine learning is a field of artificial intelligence that relies on past experiences and observations to provide systems able to automatically learn without being explicitly programmed. They can be categorized as supervised, unsupervised, semi-supervised and reinforcement learning algorithms.

DOI: 10.4018/IJWLTT.20220901.oa4

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

Figure 1. The principle comes from the combination of classifiers, by successive iterations; the knowledge of a weak classifier is added to the final classifier (strong classifier)

In a previous research, Sael et al. (2019) show the benefit of using analytic hierarchy process as a multi-criteria decision Analysis technique to study the impact of success/failure in certain academic subjects on academic year result. For the purpose of improving the performance of prediction, this article conduct a study on ensemble methods as Boosting family which have proven to be efficient in several areas. The boosting is one of the supervised machine learning methods (Ferreira et al., 2012), which consists of aggregating classifiers developed sequentially on a learning sample, to predict new observation. The task of a boosting algorithm is to learn from the iterative application of weak classifiers and then correct the errors from the last classifier to obtain more accurate classifier (Mayr et al., 2014). There are several types of boosting algorithms, like adaptive boosting (AdaBoost), Gradient Boosting, eXtreme Gradient Boosting (XGBoost), LightGBM and CatBoost (Al Daoud, 2019). The Boosting algorithms are simple and to implement and aim to improve the prediction power by training a sequence of weak models. In this paper, the authors propose to explore the benefit of boosting algorithms potentials, to improve the student performance prediction rate. This article objective is to find out the most efficient method from AdaBoost, XGBoost, CatBoost and LightGBM for student performance prediction.

The rest of this paper is organized as follows: Section II presents a theoretical background of boosting algorithms and feature selection techniques. Section III explores the related works that used boosting algorithms in educational domain. Section IV describes the experiment process, and exhibits the performance of different techniques. The conclusion and future works are detailed in Section V.

THEORITICAL BACKGROUND

Boosting Algorithms

The boosting methods assume the use of multiple classifiers that are weighted, i.e. the classifier whose prediction is more correct will have the greater weight. The principle comes from the combination of classifiers, as a result of successive iterations. The knowledge of a weak classifier is added to the final classifier (strong classifier) as shown in Figure 1.

Adaptive Boosting (AdaBoost)

This was proposed by Yoav Freund and Robert Schapire, it is based on an adaptive parameter update and giving more importance to values that are difficult to predict, therefore by boosting the classifiers

who succeed when others have failed. AdaBoost relies on existing classifiers and seeks to assign them the right weights with regard to their performance (Schapire, 2013).

Gradient Boosting Algorithms

This is mainly used with decision trees, hence it is called Gradient Tree Boosting (GTB). It trains learners based upon minimizing the loss function of a learner, while adaptive learning focuses on training upon misclassified observations. The main idea is to aggregate several Decision Tree classifiers iteratively to build the super-classifier based on the weight of these weak classifiers. The three algorithms described below are gradient boosting methods.

XGBoost (eXtreme Gradient Boosting)

This comes from the application of boosting methods to regression trees due to the fact that it can handle missing values. In addition, the model is regularized to better controls overfitting which achieves good performance. The algorithm is also known for the use of other methods than CART trees such as linear regression (Linear Boosting), or DART trees (Dropouts Additive Regression Trees) (Chen and Guestrin, 2016).

LightGBM

This was introduced by Microsoft Research. The main difference between XGBoost and LightGBM is that, the later uses XGBoost as a baseline. It outperforms in training speed and the dataset sizes that can be handled. There are also major differences from the old implementations. First, the construction of vertical and non-horizontal trees, i.e., the algorithm chooses the leaf with the best loss to grow, second, the algorithm is very efficient and fast on sparse data and large volumes of data. This algorithm also consumes very few memory resources (KE et al., 2017).

CatBoost

A similar implementation to XGBoost and LightGBM, however, CatBoost considers categorical variables for learning. The algorithm built various statistics from the categorical variables and, takes the results into account for learning (Prokhorenkova et al., 2018; Dorogush et al., 2018).

Feature Selection

The feature selection consists of selecting from the dataset the most relevant features describing the phenomenon studied. The concept of relevance of a subset of features always depends on the objectives and criteria of the system. This subset must allow improving the model performance and the speed of learning. Feature selection is an active area of research and many methods have been proposed. Feature selection techniques can be categorized into two subcategories: feature-ranking and feature subset selection (Rathore and Gupta, 2014). In this paper, the authors are interested in feature ranking methods. The idea of this paper approach is to rank features according to their score using Fisher Score and Information Gain and remove uninformative and redundant features with the use of Recursive Feature Elimination.

Fisher Score

The principle of Fisher Score is to select each feature independently based on their scores under the Fisher criterion, which eventually leads to a subset of the most representative features where the distances between data points in different classes are as large as possible, and the distances between data points in the same class are as small as possible (Gu et al., 2012). Let μ^k and σ^k denote the mean and standard deviation of the whole dataset corresponding to the k-th feature, fisher score of this k-th feature is (1):

$$F(x^k) = \frac{\sum_{n=1}^c n_n (u_n^k - \mu^k)}{(\sigma^k)^2} \quad (1)$$

where n_n is the size of the n-th class.

Information Gain

Information gain is the amount of information gained by knowing the value of the attribute. Feature selection by Information Gain aims at ranking subsets of features based on high information gain entropy in decreasing order (Azhagusundari and Thanamani, 2013). The formula of IG is given by (2):

$$\text{Information Gain} = \text{Entropy}(p) - \left(\sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) \right) \quad (2)$$

where p is the proportion of samples of the parent node, n is the total number of samples, and “i” is the proportion of samples of a particular node.

Recursive Feature Elimination

This is a feature selection method that fits a model and removes the weakest features, and builds the model using the remaining attributes (Granitto et al., 2006).

K-Fold Cross Validation

The cross validation is a technique for evaluating machine-learning models by training several machine-learning models on subsets of the available input data and by evaluating them on the complementary data subset. In k-fold cross validation, the input data is divided into k subsets of data (or samples). You train a machine learning model on (k-1) subset, then the model is evaluated on the subset that was not used for training. This process is repeated k times, with a different subset reserved for evaluation (and excluded from training) each time (Rodriguez et al., 2009). The k-fold validation is very important to prevent overfitting problems.

Performance Evaluation

Evaluating the performance of a model is one of the key steps in the data science process; it indicates the efficiency of the notation (predictions) of a dataset by a trained model. The most common tool for measuring the performance of a machine learning model is the confusion matrix, which is a summary of the prediction results on a classification problem. Correct and incorrect predictions are highlighted and broken down by class. The results are thus compared with the actual values. From the confusion matrix, a whole performance criteria can be derived such as the Accuracy.

Table 1 shows the confusion matrix in case of 2-class classification problem, where:

- **True Positive (TP):** Observation (Actual Class) is positive and is predicted to be positive.
- **False Negative (FN):** Observation is positive but is predicted negative.
- **True Negative (TN):** Observation is negative and is predicted to be negative.
- **False Positive (FP):** Observation is negative but is predicted positive.

Table 1. Confusion matrix for binary classification

		Predicted Classes	
		Positives	Negatives
Actual Classes	Positives	True Positive (TP)	True Negative (TN)
	Negatives	False Positive (FP)	False Negative (FN)

$$Accuracy_{BinaryClass} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Accuracy_{Multi-Class} = \frac{\sum_{i=1}^k TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (4)$$

$$Accuracy_{mean} = \frac{Accuracy_{Train} + Accuracy_{Test}}{2} \quad (5)$$

In this work, the authors used the accuracy, weighted precision and weighted F-score to measure the performance of the model. Accuracy with a binary classifier is measured as number of items correctly identified as either truly positive or truly negative out of the total number of items (3), but accuracy for a multiclass classifier is calculated as the average accuracy per class (4). In weighted precision, the authors computed the weight of the precision of each class by the number of samples from that class and similarly for weighted F-score.

In this work, the authors make the choice to use the average accuracy, precision and F-score of evaluating the models for each fold, and finally calculate the average of accuracy of the whole 10 folds (5).

RELATED WORKS

Predict the student failure; dropout or academic performance is one of the most important researches topics in the academic field. Some researches focused on how to improve the student performance in both presential and e-learning context.

E-learning has become a major challenge in the educational world, which has given rise to new learning behaviors different from traditional ones. Wan et al. (2017), proposed extracting students 'behavior under an e-learning platform to help instructors to know how exactly students are learning. They used Gradient Boosting Decision Tree (GBDT) classifier and achieved 0.93 in AUC. The objective of Liang et al. (2016) is to predict dropout in Massive Open Online Course (MOOCs), using four models: SVM, LR, Random Forest (RF) and GBDT. The best result achieved was 88% for accuracy parameter with GBDT. A study on engagement prediction was proposed by Hew et al. (2018), to identify which aspects of MOOCs participants found attractive. The analysis was conducted on a large dataset of participant qualitative comments, using five machine learning algorithms kNN, GBT, SVM, LR and NB. The best model was GBT (Gradient Boosted tree) with an

accuracy varying between 80.6% and 96.3%. Hew et al. (2020), proposed to investigate the learner satisfaction in MOOCs, to uncover satisfaction factors related to their design. The learner satisfaction investigation was conducted based on different models as kNN, SVM, LR and GBT which gave the best performance. Fernandes et al. (2019) conducted a predictive analysis of the academic performance in public schools of the Federal District of Brazil. They have proposed a classification models based on the Gradient Boosting Machine (GBM) to predict academic outcomes of student performance in the end of the year, the model reached 99% in AUC. Akçapınar et al. (2019), proposed an early prediction system in the context of eBook-based for teaching and learning. The students' eBook reading data was exploited to develop an early warning system for students at-risk of academic failure. The purpose is to predict students whose academic performance is low by using 13 prediction algorithms (AdaBoost - Bayesian Additive Regression Trees (BART) - GBM - Generalized Linear Model (GLM) - C4.5-like Trees (J48) - Rule-Based Classifier (JRip) - kNN - NB - Neural Network (NN) - RF - Classification and Regression Tree (RPART) - SVM Linear and XGBoost. The results revealed that in a sixteen-week long course, all models reached their highest performance with the data from the 15th week. In terms of algorithms, RF outperformed other algorithms when raw data was used, however, with the transformed data, J48 algorithm performed better and when categorical data were used, NB outperformed other algorithms (AdaBoost, BART, GBM, etc.).

Sawant et al. (2019) proposed a student placement prediction model in higher education using the gradient boosted algorithm. Based on the results obtained the proposed algorithm achieved 100% accuracy in predicting the placement opportunity. Sekeroglu B. al. (2019) study's objective is the prediction and classification of student performance respectively using five machine learning algorithms, Backpropagation (BP), Support Vector Regression (SVR), Long-Short Term Memory (LSTM) and Gradient Boosting Classifier (GBC). The BP was the best and experiments gave 87.78% in accuracy. Miguéis et al. (2018) proposed a system to support an European Engineering School in promoting the academic potential of each student and experience through data mining models. The study's main goal is to early classify students into segments, based on the student's average grade and the number of years it takes to graduate. The results shown that random forests classifier presents the best results 96.1% in terms of accuracy. Han et al. (2017) proposed multiple classifiers as DT, NN, RF, SVM and AdaBoost to predict students' academic performance. The accuracy of each prediction model reveals that the prediction model constructed by AdaBoost algorithm was the most efficient, but relatively the cost of time was the longest. Stearns et al. (2017) used socioeconomic information to predict the student performance. Gradient Boosting and AdaBoost were used to construct decision tree ensembles and the mean absolute error was of 65.9 points using Gradient Boosting. Sagar et al. (2016), proposed a system to predict the programming performance of students (executable computer program), and analyze the impact factor of various attributes that influence this performance. XGBoost (Tuned) proved to be the most efficient, exhibiting an accuracy of 80% and 91% for two different datasets.

In Table 2, the authors present a comparison of various research studies aiming to produce a system for student profile modeling using boosting algorithms during the last four years beginning from 2016, in order to better analyze the contribution of these techniques in profile modeling.

The parameters used in comparison are:

Objective: The goal of the research study: predict failure, dropout, recommendation, academic performance, etc.

Student features: The authors carried out the categorization of students' features, all the student features are extracted from each paper, and gathered in categories as proposed in the author's last papers (Hamim et al., 2019). These features are: Personal Identity (defines the student in a unique way compared to other students, it encompasses: sex, Gender, Age, Nationality, Name/ Surname, Occupation, Student code Phone, Email.), Social Identity (It is an element derived from the belonging of an individual to a social group, it encompasses: Marital status, parents'

education, parents' job, religion, Address, Culture.etc.), Academic (it concerns every information related to the academic course of the student: diplomas, Bachelor, High school grade, Majors, marks, etc.), Online Behavior (Comments, Rank, Tests, Quizzes, Courses, etc.), Learning Behavior (Attention, Decision making, Responsibility, etc.) and Physical Conditions (refers to the state of the physical qualities).

Techniques: Boosting or machine learning techniques used: GBT, AdaBoost, XGBoost, etc.

Result: The performance metrics results of the most efficient technique (marked in bold in the Techniques column) used by the research study (A: Accuracy, AUC, P: Precision, and F: F-score).

From the state of art, improving the academic performance is the most common objective of the research studies. The most student characteristics used are the online behaviors, followed by Academic results, at last personal and social identity. Concerning classification algorithms, boosting algorithms show their importance by exceeding, in the majority cases, the other machine learning algorithms. This paper objective is to validate these results and to discover other functioning of the boosting algorithms by testing these algorithms on a multiclass dataset which belongs to the educational field, and by using these same boosting techniques to generate important features that can help us build sophisticated profile model.

COMPARATIVE STUDY OF BOOSTING ALGORITHMS

Methodology

To conduct our study, Figure 2 presents the proposed process. First, one label encoding preprocessing is applied to prepare the data before fitting a machine learning model, thus categorical variables are converted into numerical. Second, the authors applied Recursive Feature Elimination technique based on two feature ranking methods (Fisher Score and Information Gain) separately on the dataset, to select the most relevant features. A 10-fold cross validation is applied, and four boosting machines are tested to find the most efficient one, finally, the evaluation is done by three performance metrics: Accuracy, Precision and F-score.

First Step: The authors used the dataset called Kalboard 360 (Amrieh et al., 2015), (Amrieh et al., 2016) which is collected from learning management system (LMS). It consists of 480 student records with 3 classes identifying student success in 3 categories: low level (0-69), middle level (70-89) and high level (90-100). There is 16 features including personal identity, social identity, online behavior, learning behavior and especially the important characteristics extracted from the authors' categorization analysis (Hamim et al., 2019).

Second Step: data preprocessing: one label encoding preprocessing is applied.

Third Step: The 10-Fold cross validation is used in order to predict student performance using four boosting machine algorithms CatBoost, LightGBM, XGBoost and AdaBoost.

RESULTS AND ANALYSIS

Table 3 shows the effectiveness of the performance classification measured by the accuracy, precision and F-score with the number of features used. LightGBM algorithm with Fisher Score gave the most accuracy with an average of 89.05%, using 15 features, followed by XGBoost when using 14 features (88.96%). The AdaBoost algorithm is the weakest algorithm when using Fisher Score, with an accuracy, around 70%. In terms of precision, the two algorithms LightGBM and XGBoost reached 89% using 14 features in the modeling, and concerning the F-score, LightGBM performs 89% using only 12 features, whereas the XGBoost reaches the same performance value using 14.

Table 2. Comparative study of state of art

Ref.	Objective	Student Features'	Techniques	Result (%)
(Hew et al., 2020)	E-learning sentiment analysis	<ul style="list-style-type: none"> • Online behavior 	kNN, GBT , SVM, LR	F:76.25-88.32
(Fernandes et al., 2019)	Academic performance	<ul style="list-style-type: none"> • Personal Identity • Social Identity • Academic • Physical conditions 	GBT	AUC: 99
(Akçapınar et al., 2019)	Academic performance	<ul style="list-style-type: none"> • Academic • Online behavior 	AdaBoost, BART, GBM, GLM, J48 , JRip, kNN, NB , NN, RF , RPART, SVMLinear, xgbLinear	A: 83%
(Sawant et al., 2019)	Academic performance	<ul style="list-style-type: none"> • Personal Identity • Academic 	GBT	A:100
(Sekeroglu et al., 2019)	Academic performance	<ul style="list-style-type: none"> • Personal Identity • Social Identity • Academic • Physical Conditions • Learning Behavior • Online Behavior 	BP , SVR, LST, GBC	A:87.78
(Hew et al., 2018)	E-learning engagement	<ul style="list-style-type: none"> • Online Behavior 	kNN, GBT , SVM, LR, NB	A: 80.6-96.3
(Miguéis et al., 2018)	Academic performance	<ul style="list-style-type: none"> • Personal Identity • Social Identity • Academic 	NB, SVM, DT, RF , BT, ABT	A: 96.1
(Han et al., 2017)	Academic performance	<ul style="list-style-type: none"> • Academic 	DT, NN, RF, SVM, AdaBoost	A:91.67
(Wan et al., 2017)	E-learning performance	<ul style="list-style-type: none"> • Online Behavior 	GBDT	AUC: 93
(Stearns et al., 2017)	Academic performance	<ul style="list-style-type: none"> • Personal Identity • Social Identity • Academic 	AdaBoost, GB	
(Liang et al., 2016)	E-learning dropout	<ul style="list-style-type: none"> • Online behavior 	SVM, LR, RF, GBDT	A: 88
(Sagar et al., 2016)	Academic performance	<ul style="list-style-type: none"> • Online behavior 	RF, GBM, GLM, XGBoost	A: 80-91

The LightGBM algorithm is the most efficient in terms of accuracy with an average of 89.26% using 14 features with information gain, instead of 15 (found for Fesher Score), and for precision and F-score performance metrics, LightGBM reached 90% using 15. These results show that, information gain feature selection gave a better result with a minimal margin compared to Fisher Score method.

According to experiment analysis, gradient boosting algorithms give most important results compared to the adaptive boosting approach using AdaBoost. The use of boosting algorithms allowed to boost the performance of this paper model, but also to detect the most important features used when building the model. After the phase of boosting the performance results of the model, the second step of the process is to analyze the feature importance provided by the models. This can benefit in targeting feature that play an important role in the explanation of students' unsuccessful learning process, which is important in building a profile model by focusing only on the important variables and removing the one that are not significant. Feature selection can also help us subsequently to build

Figure 2. Proposed process

a simple and understandable model that consumes less time. The Table 4 shows the results of the feature importance score that was calculated for each fold among the 10 folds of cross validation and marked the importance of each selected feature for the two feature ranking methods for LightGBM prediction model. Student performance is strongly influenced by visited resources, followed Discussion, Announcements view and raised hands features using the two feature ranking methods, which show the effect of the e-learning aspect on student performance, and also reflects the learning behavior aspect and its influence on student achievement.

Table 3. Performance evaluation vs. Feature selection

Number of selected features	Adb			CatBoost			LightGBM			XGBoost		
	A (%)	P (%)	F (%)	A (%)	P (%)	F (%)	A (%)	P (%)	F (%)	A (%)	P (%)	F (%)
Fisher Score												
12	70	71	70	75.31	76	75	88.69	89	89	88.33	89	88
14	70	71	70	76.34	77	76	88.64	89	88	88.96	89	89
15	70	71	70	74.93	75	75	89.05	89	89	88.13	89	88
Information Gain												
12	73.97	75	74	75.6	76	75	87.58	88	87	88.02	89	88
14	74.12	75	74	75.05	75	75	89.26	90	89	88.34	89	88
15	74.12	75	74	75.39	75	75	88.88	90	90	88.54	89	88

Table 4. Features importance comparison

Attribute	LightGBM Weight by Fisher Score	LightGBM Weight by Information Gain
VisITedResources	0.7691	0.7478
Discussion	0.7316	0.7404
AnnouncementsView	0.676	0.6836
raisedhands	0.6635	0.6637
Topic	0.3656	0.3691
GradeID	0.2556	0.3077
StudentAbsenceDays	0.2045	0.1991
ParentAnsweringSurvey	0.1869	0.1934
gender	0.1729	0.1686
PlaceofBirth	0.1484	0.1445
Relation	0.1422	0.1437
NationalITy	0.1343	0.136
ParentschoolSatisfaction	0.1055	0.1024
Semester	0.086	0.0994
StageID	0.0692	0.011

CONCLUSION

In this study, the authors present an experiment on student profile modeling using boosting algorithms for the objective of predicting student performance in order to give the most efficient algorithm and the overall description of the student profile. The obtained results show that Light Gradient Boosting Machine is the most efficient than the algorithms used in this study (CatBoost, XGBoost and AdaBoost) followed by XGBoost and that AdaBoost algorithm was the least efficient from other algorithms. In addition, by using two feature ranking methods (Fisher score and Information Gain) combined with recursive feature elimination, it was found that the two methods gave good performance and

especially information gain. Thus, online behavior was the main feature that impacted the student performance, followed by learning behavior, social and personal identity. In future works, the authors aim to propose student profile model that can be exploited in many situations such as: prediction, classification, adaptive learning and e-recommendation.

REFERENCES

- Akçapınar, G., Hasnine, M. N., Majumdar, R., Flanagan, B., & Ogata, H. (2019). *Developing an early-warning system for spotting at-risk students by using eBook interaction logs*. Academic Press.
- Al Daoud, E. (2019). Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. *International Journal of Computer and Information Engineering*, 13(1), 6–10.
- Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015). Preprocessing and analyzing educational data set using X-API for improving student's performance. In *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)* (pp. 1-5). IEEE.
- Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, 9(8), 119–136. doi:10.14257/ijtda.2016.9.8.13
- Azhagusundari, B., & Thanamani, A. S. (2013). Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering*, 2(2), 18–21.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM. doi:10.1145/2939672.2939785
- Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: gradient boosting with categorical features support*. arXiv preprint arXiv:1810.11363.
- Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, 94, 335–343. doi:10.1016/j.jbusres.2018.02.012
- Ferreira, A. J., & Figueiredo, M. A. (2012). Boosting algorithms: A review of methods, theory, and applications. In *Ensemble machine learning* (pp. 35–85). Springer. doi:10.1007/978-1-4419-9326-7_2
- Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*, 83(2), 83–90. doi:10.1016/j.chemolab.2006.01.007
- Gu, Q., Li, Z., & Han, J. (2012). *Generalized fisher score for feature selection*. arXiv preprint arXiv:1202.3725.
- Hamim, T., Benabbou, F., & Sael, N. (2019, October). Toward a Generic Student Profile Model. In *Proceedings of the Third International Conference on Smart City Applications* (pp. 200-214). Springer.
- Hamim, T., Benabbou, F., & Sael, N. (2019, October). Student profile modeling: an overview model. In *Proceedings of the 4th International Conference on Smart City Applications* (pp. 1-9). Academic Press.
- Han, M., Tong, M., Chen, M., Liu, J., & Liu, C. (2017, July). Application of Ensemble Algorithm in Students' Performance Prediction. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)* (pp. 735-740). IEEE.
- Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*, 145, 103724. doi:10.1016/j.compedu.2019.103724
- Hew, K. F., Qiao, C., & Tang, Y. (2018). Understanding student engagement in large-scale open online courses: A machine learning facilitated analysis of student's reflections in 18 highly rated MOOCs. *International Review of Research in Open and Distributed Learning*, 19(3). Advance online publication. doi:10.19173/irrodl.v19i3.3596
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., . . . Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154). Academic Press.
- Liang, J., Li, C., & Zheng, L. (2016, August). Machine learning application in MOOCs: Dropout prediction. In *2016 11th International Conference on Computer Science & Education (ICCSE)* (pp. 52-57). IEEE.

Mayr, A., Binder, H., Gefeller, O., & Schmid, M. (2014). *The evolution of boosting algorithms-from machine learning to statistical modelling*. arXiv preprint arXiv:1403.1452.

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, 115, 36–51. doi:10.1016/j.dss.2018.09.001

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 6638–6648.

Rathore, S. S., & Gupta, A. (2014, February). A comparative study of feature-ranking and feature-subset selection techniques for improved fault prediction. In *Proceedings of the 7th India Software Engineering Conference* (pp. 1-10). Academic Press.

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3), 569–575. doi:10.1109/TPAMI.2009.187 PMID:20075479

Sael, N., Hamim, T., & Benabbou, F. (2019). Implementation of the Analytic Hierarchy Process for Student Profile Analysis. *International Journal of Emerging Technologies in Learning*, 14(15), 78–93. doi:10.3991/ijet.v14i15.10779

Sagar, M., Gupta, A., & Kaushal, R. (2016, September). Performance prediction and behavioral analysis of student programming ability. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1039-1045). IEEE.

Sawant, T. U., Pol, U. R., & Patankar, P. S. (2019). *Student Placement Prediction Model using Gradient Boosted Tree Algorithm*. *Journal of Emerging Technologies and Innovative Research*.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37-52). Springer.

Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student performance prediction and classification using machine learning algorithms. In *Proceedings of the 2019 8th International Conference on Educational and Information Technology* (pp. 7-11). Academic Press.

Stearns, B., Rangel, F. M., Rangel, F., de Faria, F. F., Oliveira, J., & Ramos, A. A. D. S. (2017). Scholar Performance Prediction using Boosted Regression Trees Techniques. *ESANN 2017 Proceedings. European Symposium on Artificial Neural Networks. Computational Intelligence*.

Touria Hamim received the engineer degree in Information and Communication Systems Engineering from ENSAJ, Morocco, in 2017. Currently she is preparing her Ph.D. in Computer Science in Faculty of Science Ben M'sik. Her research interests profile modeling using machine learning.

Faouzia Benabbou is a professor of Computer Science and member of Compute Science and Information Processing laboratory. She is Head of the team "Cloud Computing, Network and Systems Engineering (CCNSE)". She received his Ph.D. in Computer Science from the Faculty of Sciences, University Mohamed V, Morocco, 1997. His research areas include cloud Computing, data mining, machine learning, and Natural Language Processing.

Nawal Sael is a professor of Computer Science and member of Computer Science and Information Processing laboratory at faculty of science Ben M'sik (Casablanca, Morocco). She received her Ph.D. in Computer Science from the Faculty of Sciences, University Hassan II Casablanca, Morocco, 2013 and her engineer degree in software engineering from ENSIAS, Morocco, in 2002. Her research interest include Data Mining, Machine Learning, Natural Language Data Preprocessing and Internet of Things.