Comparative Analysis of Supervised Machine Learning Algorithms to Build a Predictive Model for Evaluating Students' Performance

https://doi.org/10.3991/ijoe.v17i02.20025

Inssaf El Guabassi (✉)
Abdelmalek Essaadi University, Tetouan, Morocco
elguabassi@gmail.com

Zakaria Bousalem
Hassan 1st University, Settat, Morocco

Rim Marah
Abdelmalek Essaadi University, Tetouan, Morocco

Aimad Qazdar
Cadi Ayyad University, Marrakech, Morocco

Abstract—In recent years, the world's population is increasingly demanding to predict the future with certainty, predicting the right information in any area is becoming a necessity. One of the ways to predict the future with certainty is to determine the possible future. In this sense, machine learning is a way to analyze huge datasets to make strong predictions or decisions. The main objective of this research work is to build a predictive model for evaluating students' performance. Hence, the contributions are threefold. The first is to apply several supervised machine learning algorithms (i.e. ANCOVA, Logistic Regression, Support Vector Regression, Log-linear Regression, Decision Tree Regression, Random Forest Regression, and Partial Least Squares Regression) on our education dataset. The second purpose is to compare and evaluate algorithms used to create a predictive model based on various evaluation metrics. The last purpose is to determine the most important factors that influence the success or failure of the students. The experimental results showed that the Log-linear Regression provides a better prediction as well as the behavioral factors that influence students' performance.

Keywords—Student Performance, Prediction, Machine Learning, Regression, Predictive Modeling, Educational Data Mining

1 Introduction

In the real world, with a remarkable growth within the universe of measured data warehouse sizes, analyzing the data and extracting the useful information is becoming

a necessity and a rich topic for several researchers [1]. Many application areas adopt machine learning techniques in their systems such as finance, shopping platforms, restaurants, economy, medicine, tourist targets, and marketing. Over the last two decades, machine learning has entered the e-learning space as well [2] [3] [4] [5]. Thus, several machine learning algorithms have been exploited by researchers to predict hidden patterns from educational settings [6] [7] [8].

The prediction of students at risk for academic failure is of the utmost importance and it must be identified as soon as possible during the academic year. The early prediction of student performance is necessary for higher education for providing high-quality education, reducing dropout rates, increasing school completion rates, and improving educational outcomes.

However, the real and major problems are:

- How to identify the "weak" students who will need additional help to improve their performance?
- Which the best machine learning algorithm (i.e., model) for predicting students' academic performance?
- What factors can affect students' academic performance?

This research work evaluates and compares the effectiveness of different machine learning algorithms. While there are many algorithms for creating predictive models, this work concentrates on seven of them, which are ANCOVA, Logistic Regression, Support Vector Regression, Log-linear Regression, Decision Tree Regression, Random Forest Regression, and Partial Least Squares Regression. The present paper also determines the factors affecting students' academic performance.

The outline of the present paper is as follows: Section 2 presents recent studies regarding the specified area. The background of machine learning is briefly described in Section 3. Section 4 concentrates on the proposed approach. A description of the materials, as well as the methods, is presented in Section 5. In Section 6 our implementation and results are presented. Section 7 concentrates on experimental evaluation. Section 8 contains the discussion. Finally, Section 9 presents the main conclusions considering some future research directions.

2 Related Work

In recent decades, many studies by several research teams have focused on predicting the performance of students based on divers' factors using various machine learning algorithms.

Bravo-Agapito et al [13] explained their study based on the prediction of 802 undergraduate student's academic performance in completely online learning. They used exploratory factor analysis, multiple linear regressions, and cluster analysis. They concluded the "age" is a factor that affects the academic achievement of the student. Gray and Perkins [14] conducted a study on predicting student outcomes as early as week 4 of the Fall semester using machine learning techniques. Hamsa et al [15] applied two classification methods which are decision tree and fuzzy genetic algorithm to predict

the student's performance for the Bachelor and Master degree students in Computer Science and Electronics and Communication. Hussain et al [16] described a performance study on predicting student difficulties from learning session data. They have used artificial neural networks, support vector machines, logistic regression, Naïve Bayes classifiers, and decision trees. Their results show that artificial neural networks and support vector machines are the best algorithms to predict the performance of the student. Karthikeyan et al [17] investigated the performance of the students by developing a hybrid educational data mining model called HEDM. Their model combines two techniques which are the J48 Classifier and Naive Baye's classification. Their results show that HEDM outperforms the results obtained in EDM.

In summary, many researchers in their recent papers have made significant results in educational data mining. However, most of them use classification methods for predicting Student' academic performance. Moreover, there was very little focus on interactional and parental involvement features.

3 Machine Learning

Machine learning reproduces behavior using learning algorithms that are themselves fueled by immense sources of information. The computer trains and improves, hence the word learning; it "learns" from data and extracts knowledge from it.

The algorithms are the engines of machine learning. In general, three main types of machine learning algorithms are used: supervised learning, unsupervised learning, and reinforcement learning.

- Supervised learning: The system learns a function from examples.
- Unsupervised learning: The system does not rely on predefined elements.
- Reinforcement learning: consists of letting the algorithm learn from its own mistakes. Faced with a random choice at the start, it uses rewards and punishment as signals for a bad and good decision.

After briefly describing the background of machine learning, in the next section, we will present our proposed approach.

4 Proposed Approach

Increasingly, E-learning has become an important tool of teaching and learning around the world. Further, Learners have the opportunity to switch to distance learning in various scientific fields anytime and anywhere [9]. It is therefore evident that many researchers work on the various aspects of e-learning [10] [11] [12]. The identification of the "weak" students and the factors affecting students' academic performance is a crucial step for successful learning. Hence, in the present paper, we aim to evaluate the student's academic performance and identifying the factors that influence academic performance using supervised machine learning algorithms.

This research work focuses on the following steps:

- Applying several machine learning algorithms which are ANCOVA, Logistic Regression, Support Vector Regression, Log-linear Regression, Decision Tree Regression, Random Forest Regression, and Partial Least Squares Regression.
- Comparing and evaluating machine learning algorithms for identifying which are most suitable by using several evaluation metrics which are Mean Square Error (MSE), Root Mean Square Error (RMSE), and R-squared (R^2).
- Identifying which factors influence the final prediction of students' results.

The next section describes the materials and methods used in our research work which are the dataset, the applied methods, and evaluation methods.

5 Materials and Methods

5.1 Dataset

The data used for this work's experimentation (available here) is collected from a dataset named "Students' Academic Performance Dataset (xAPI-Edu-Data)" [18] [19]. It is, therefore, an open-source dataset available publicly on the Kaggle dataset repository for academic and research purposes. The primary source of the dataset is from Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah, The University of Jordan, Amman, Jordan, http://www.Ibrahimaljarah.com, www.ju.edu.jo. This data is obtained from the Learning Management System known as Kalboard 360 [20]. Kalboard 360 has been created to support schools to improve their learning through the use of cutting-edge technology. Typically, any such system share and provides users synchronous access to educational resources from any device that already has internet access. Table 1 provides a summary of the dataset characteristics, including name, abbreviation, source, characteristics, number of samples, area, attribute characteristics, number of attributes, date, associated tasks, missing value and file formats.

Table 1. Summary of the Dataset

Name	Students' Academic Performance Dataset
Abbreviation	xAPI-Edu-Data
Source	Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah, The University of Jordan, Amman, Jordan.
Characteristics	Multivariate
Number of Samples	480
Area	E-learning, Education, Predictive models, Educational Data Mining
Attribute Characteristics	Integer/Categorical
Number of Attributes	16
Date	2016-11-8
Associated Tasks	Classification
Missing Values?	No
File formats	xAPI-Edu-Data.csv

As shown in Table 1, the dataset considered consists of 480 student records from various countries and 17 features. On the other hand, the features are classified into three main categories, named "Demographic features", "Academic background features", and "Behavioral features" category:

- **Demographic features**: Include qualities such as gender, nationality, and Place of birth.
- **The academic background features**: Represents the background characteristics of students such as educational stage, grade Level, section, and semester.
- **Behavioral features**: Illustrate the behavior such as a raised hand-on class, opening resources, answering surveys by parents, and school satisfaction.

Table 2 contains an overview of Dataset features used for training and testing. It contains three fields: feature, description, and type. It should be noted that there are two major feature types, named "Nominal" and "Numeric".

- **Nominal**: It labels variables by providing non-numeric value.
  - Examples: Sex {Male or Female}, level {low, middle, high}, eye color {blue, green, brown, hazel, amber, red, and gray}
- **Numeric**: It labels variables by providing quantitative value.
  - Examples: Rankings, Size, humidity, temperature, and time.

Table 2. Dataset Features

Feature	Description	Type
Gender	Gender of Student (i.e., Male or Female)	Nominal
Nationality	Nationality of Student (e.g., Morocco, Kuwait, Lebanon, Jordan, Egypt, SaudiArabia, USA, etc.)	Nominal
Place of birth	Country of birth for the student (e.g., Morocco, Kuwait, Lebanon, Jordan, Egypt, SaudiArabia, USA, etc.)	Nominal
Educational Stages	The educational level of the student (i.e., Lowerlevel, MiddleSchool or HighSchool)	Nominal
Grade Levels	Grade level of the student (i.e., G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11 or G-12)	Nominal
Section ID	Classroom of the student (i.e., A, B or C)	Nominal
Topic	Course topic (i.e., English, Spanish, French, Arabic, IT, Maths, Chemistry, Biology, Science, History, Quran or Geology)	Nominal
Semester	Semester of the year (i.e., First or Second)	Nominal
Parent responsible	The parent responsible for the student (i.e., mother or father)	Nominal
Raised hand	Number of times the student raised hand on the classroom (i.e., from 0 to 100)	Numeric
Visited resources	Number of times the student visited a course content (i.e., from 0 to 100)	Numeric
Viewing announcements	Number of times the student checked the new announcements (i.e., from 0 to 100)	Numeric
Discussion groups	Number of times the student participated in discussion groups (i.e., from 0 to 100)	Numeric

Parent Answering Survey	Parent answered the surveys which are provided from school or not (i.e., Yes or No)	Nominal
Parent School Satisfaction	The degree of parent satisfaction for the school (i.e., Yes or No)	Nominal
Student Absence Days	The number of absence days for each student (i.e., above-7or under-7)	Nominal
Class	Grade of student for the course (i.e., Low-Level, Middle-Level, or High-Level)	Nominal

After seeing the dataset used in our experimentation, in the next section we will present the selected methods for predicting students' academic performance.

5.2 Selected methods

It is impossible to predict the future with certainty, but it can determine a highly successful outcome by looking at existing data sources. Nowadays, there are many algorithms for predictive modeling machine learning. In this present work, we focus especially our concentration upon supervised machine learning algorithms because they are the most appropriate (see section III for more details).

In the next sections, we will present the algorithms used to build predictive models which are ANCOVA, Support Vector Regression, Decision Tree Regression, Random Forest Regression, Partial Least Squares Regression, Log-linear Regression, and Logistic Regression.

ANCOVA (ANalysis of VAriance) [21] is a statistical test that makes it possible to compare globally the mathematical expectation of several samples. The name of this test is explained by its way of proceeding: we decompose the total variance of the sample into two partial variances, the inter-class variance, and the residual variance, and we compare these two variances. The ANCOVA model is written as follows (1):

$$y_{ij} = \mu + \tau_j + \beta\left(x_{ij} - \bar{x}\right) + \varepsilon_{ij} \quad (1)$$

Where:

y_ij: is the jth observation in the ith group.

μ: is a constant common to all individuals.

τ_j: is the treatment effect of the jth group.

β: is the regression slope corresponding to the covariate xij.

x_ij: is the covariate for the ith subject in the jth group.

$\bar{x}$: is the overall mean of x.

ε_ij: is a Gaussian error term.

As shown in figure 2, ANCOVA help to compare two or more regression lines to each other.

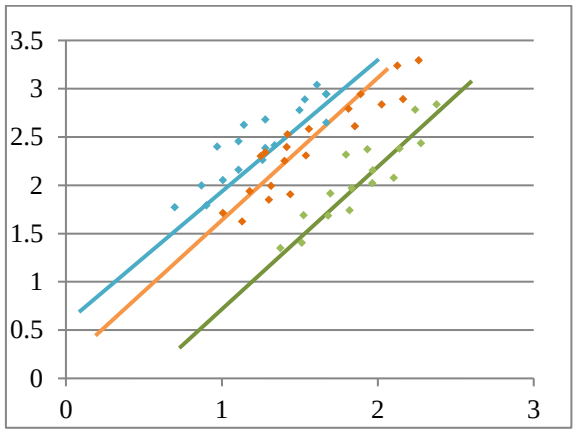

Fig. 1. ANCOVA

Logistic Regression or Logit Regression (Logit-R) [22] is a statistical method for performing binary classifications such as healthy/sick, win/lose, pass/fail, or alive/dead.

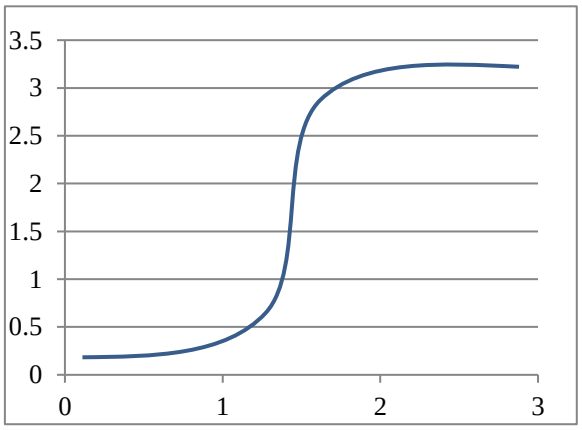

Fig. 2. Logistic Regression

It takes qualitative and/or ordinal predictor variables as input and measures the probability of the output value using the sigmoid function shown in figure 2 and defined by the formula (1):

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

Support Vector Regression (SVR) [23] is a binary classification algorithm. Just like the Logistic Regression. If we take the image above, we have two classes (e.g., suppose these are e-mails, and Spam mails are in red and non-spam emails are in blue). The Logistics regression can separate these two classes by defining the line in red. The SVR will opt to separate the two classes by the green line (see figure 3).

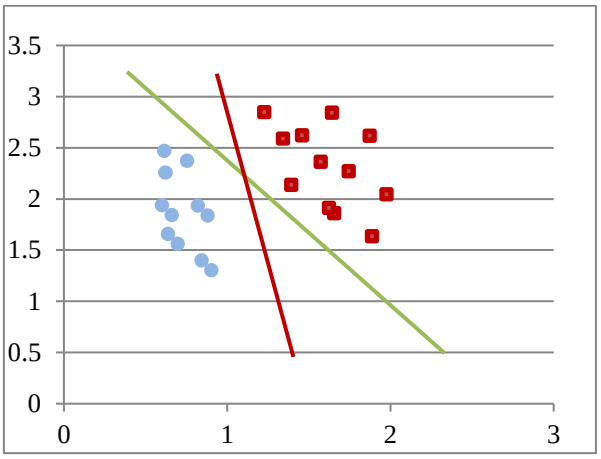

Fig. 3. Support Vector Regression

Decision Tree Regression (DTR) [25] is an algorithm that uses a graph model (trees) to define the final decision. Each node has a condition, and the branches are based on this condition (True or False). The further down the tree you go, the more conditions we accumulate. Figure 4 illustrates this operation.

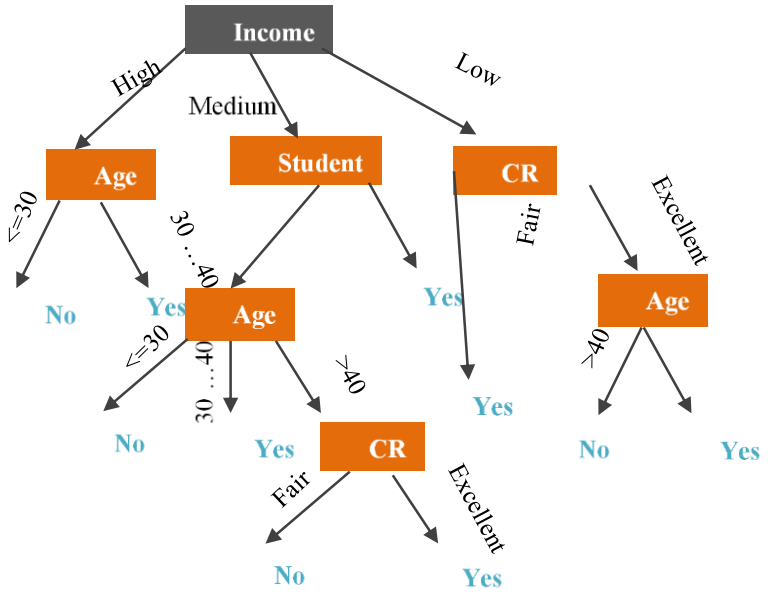

Fig. 4. Decision Tree

Log-Linear Regression (Log-LR) [24] is part of the family of generalized linear models for Exponential-distributed, Gamma, or Poisson data. This method is a linear

approach to modeling the relationship between a response variable and one or more explanatory variables. We assume that the response variable is written as the logarithm of an affine function of the explanatory variables

Random Forest Regression (RFR) [26] is a supervised learning algorithm that combines multiple predictions to make a more accurate prediction than a single model (see Figure 5)

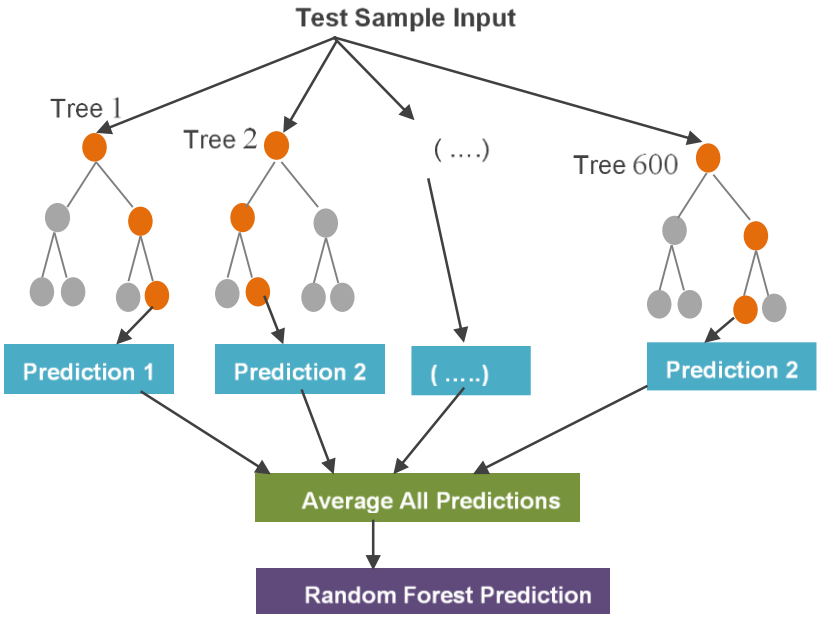

Fig. 5. Random Forest

Partial Least Squares Regression (PLS-R) [27] is a flexible statistical technique applicable to any form of data. It allows modeling the relationships between inputs and outputs, even when the inputs are correlated and noisy, the outputs multiple and the inputs more numerous than the observations. In the next section, we will concentrate on the evaluation metrics used in our experimental study for identifying the best machine learning algorithm.

5.3 Evaluation methods

Evaluating a model is a core part of building an effective machine learning model. There are many methods of evaluation that can be used. However, the question is: which metrics should we use to evaluate regression techniques in machine learning? Figure 6 is represented to answer this question.

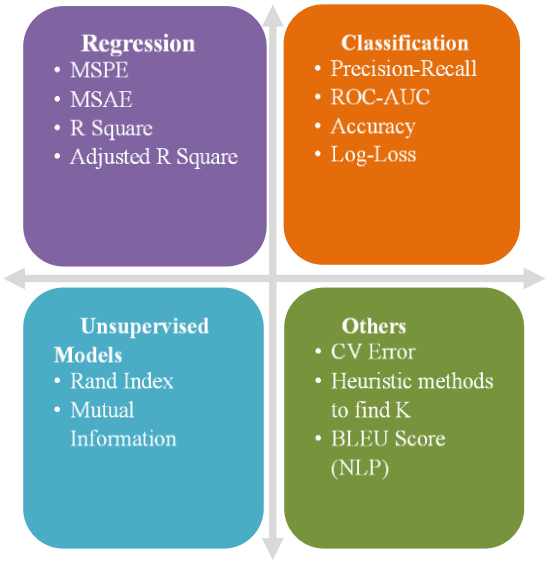

Fig. 6. Right metrics for evaluating machine learning models [28]

In the following, we will discuss the three main metrics which we will use in our evaluation.

R-Squared (R2 or the coefficient of determination) [29] is an indicator that allows judging the quality of simple linear regression. It measures the fit between the model and the observed data or how well the regression equation is to describe the distribution of points.

- If the R^2 is zero, it means that the equation of the regression line determines 0% of the distribution of points. This means that the mathematical model used does not explain the distribution of points.
- If the R^2 is 1, it means that the equation of the regression line can determine 100% of the distribution of points. This then means that the mathematical model used, as well as the parameters a and b calculated, are those which determine the distribution of the points.

In short, the closer the coefficient of determination is to 0, the more the scatter plot disperses around the regression line. On the contrary, the more the R^2 tends towards 1, the more the cloud of points narrows around the regression line. When the points are exactly aligned on the regression line, then $R^2 = 1$.

Mean Square Error (MSE) [30] is the arithmetic mean of the squares of the predictions between the model and the observations. This is the value to be minimized in the context of a single or multiple regressions. The method is based on the nullity of the mean of the residuals. But the average of their squares is generally not zero.

Root Mean Square Error (RMSE) is a standard way to measure the error in model evaluation studies. It is the square root of the mean of the square of all of the errors.

6 Implementation and Results

The present paper represents a comparison and evaluation of supervised machine learning algorithms for predicting students' academic performance. Many experiments were conducted in seven major steps depending on the regression methods namely ANCOVA, Logistic Regression (Logit-R), Support Vector Regression (SVR), Log-linear Regression (Log-LR), Decision Tree Regression (DTR), Random Forest Regression (RFR), and Partial Least Squares Regression (PLS-R). These regression methods were applied using the XLSTAT environment [31]. In the following, the experimental result of each algorithm is presented.

Table 3. Experimental results

	MSE	**RMSE**	**R^2**
ANCOVA	0.157256464	0.396555752	0.71890384
Logit-R	0.156250000	0.395284708	0.73799242
SVR	0.212447120	0.460919863	0.6271547
Log-LR	0.158611894	0.398261088	0.71667276
DTR	0.195293449	0.441920184	0.65025193
RFR	0.171994444	0.414722128	0.69480482
PLS-R	0.205659323	0.453496773	0.63238366

The table above therefore represents summary results for the seven algorithms used in this research work. The evaluation metrics used in this experiment are Mean Square Error (MSE), Root Mean Square Error (RMSE), and R-squared (R^2). It should be noted that RMSE is just the square root of the MSE.

7 Evaluation

After rigorously evaluating all the seven algorithms on the 480 students of our dataset, we compare the performances to determine which model predicts better. According to the experimental results, it is clear that Log-linear Regression (Log-LR) provides better performance because it has a low MSE, low RMSE, and high R^2 score, closely followed by ANCOVA. On the other hand, we observed that Support Vector Regression (SVR) isn't suitable for predicting students' academic performance because it has a high MSE, high RMSE, and low R^2 score.

8 Discussion

Given the R^2= 73% of the variability of the dependent variable, Class is explained by the 16 explanatory variables. The remainder of the variability is due to other explanatory variables that have not been considered during the present experiment research. Table 4 displays the Type III Sum of Squares analysis. This table is very important to determine whether or not the explanatory variables provide significant information.

Table 4. Type III Sum of Squares analysis

Feature	DF	Sum of Squares	Mean Squares	F	Pr > F
Raised Hand	1.000	2.470	2.470	14.380	0.000
Visited Resources	1.000	4.327	4.327	25.197	0.000
Viewing Announcements	1.000	0.625	0.625	3.638	0.057
Discussion Groups	1.000	0.353	0.353	2.056	0.152
Gender	1.000	1.432	1.432	8.336	0.004
Nationality	8.000	1.981	0.248	1.442	0.177
Place of Birth	8.000	2.018	0.252	1.469	0.166
Educational Stages	1.000	0.120	0.120	0.699	0.403
Grade Levels	9.000	1.723	0.191	1.115	0.350
Section ID	2.000	0.013	0.007	0.038	0.963
Topic	11.000	2.347	0.213	1.243	0.256
Semester	1.000	0.050	0.050	0.294	0.588
Parent Responsible	1.000	2.497	2.497	14.541	0.000
Parent Answering Survey	1.000	2.319	2.319	13.506	0.000
Parent School Satisfaction	1.000	0.276	0.276	1.606	0.206
Student Absence Days	1.000	23.444	23.444	136.517	0.000

According to Fisher's F-test. lower the F probability corresponding to a given variable. the stronger the impact of the variable on the model. In the table above. we can see that the p-value for the "Viewing Announcements". "Discussion Groups". "Gender". "Nationality". "Place of Birth". "Educational Stages". "Grade Levels". "Section ID". "Topic". "Semester" and "Parent School Satisfaction" are 0.057. 0.152. 0.004. 0.177. 0.166. 0.403. 0.350. 0.963. 0.256. 0.588. and 0.206 respectively. This confirms the weak impact of these parameters on the model. On the other hand. it is clear that the p-value for "Raised Hand". "Visited Resources". "Parent Responsible". "Parent Answering Survey" and "Student Absence Days" is 0. Therefore, these parameters bring significant information to our model. Furthermore. based on type III errors. it can be inferred that the most influential explanatory variable is "Student Absence Days". The following chart indicates the predicted values versus the observed values. Also. Confidence intervals for the mean allow for the detection of potential outliers.

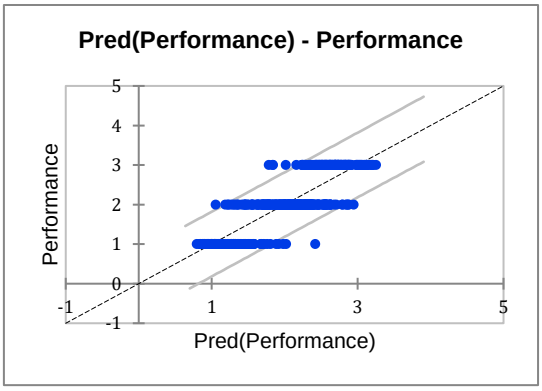

Fig. 7. Predicted values versus the observed values

The following histogram represents the standardized residuals versus the performance. It indicates that the residuals grow with the Performance. As we can see in Figure 8 the residuals bar chart allows to quickly showing the residuals that are out of the range [-2. 2].

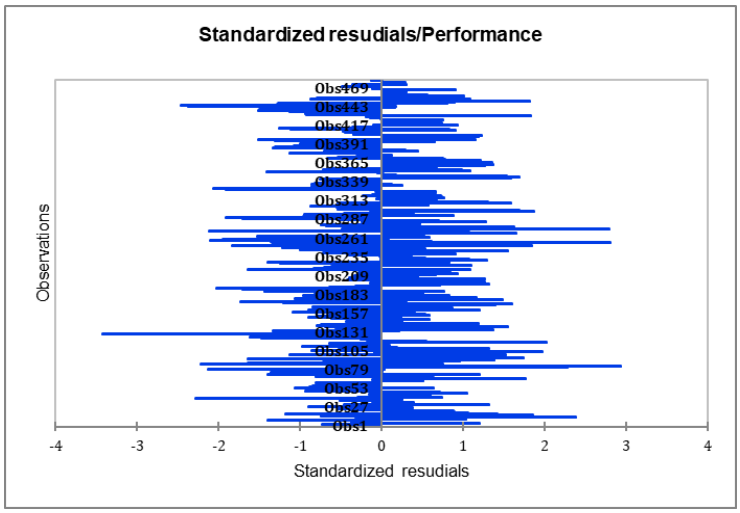

Fig. 8. Standardized residuals versus the performance

As conclusion. “Raised Hand”. “Visited Resources”. “Parent Responsible”. “Parent Answering Survey” and “Student Absence Days” allow us to explain 73% of the variability of the performance. Further analysis would be necessary because an amount of information is not explained by our model.

9 Conclusion and Future Work

In recent years. predicting a student’s academic performance is the main objective of all educational institutions. The numerous studies demonstrate that machine learning can be an efficient technology to meet this objective. In this research work. our first aim was to compare several machine learning algorithms for predicting student’s academic performance. Therefore. we apply and evaluate several algorithms which are ANCOVA. Logit-R. SVR. Log-LR. DTR. RFR and PLS-R. Our second aim was to determine the relationships between the features and the student’s academic performance. As a result of our experimental study. we can conclude that “Raised Hand”. “Visited Resources”. “Parent Responsible”. “Parent Answering Survey” and “Student Absence Days” provide a significant amount of information for predicting student’s academic performance. Certainly, this research work has some limitations. That’s why the major directions for future work could focus on the following: Firstly. applying techniques such as clustering and artificial neural networks to have better predicting. Secondly. utilizing dataset with massive size and diverse features to tackle the issue of

scalability. The final area that can be improved is exploiting few hybrid feature selection algorithms.

10 References

[1] Kalaivani. S. Priyadharshini. B. &Nalini. B. S. (2017). Analyzing student's academic performance based on data mining approach. International Journal of Innovative Research in Computer Science and Technology. 5(1). 194-197. https://doi.org/10.21276/ijircst.2017.5.1.4

[2] Moubayed. A. Injadat. M. Shami. A. &Lutfiyya. H. (2020). Student engagement level in e-learning environment: Clustering using k-means. American Journal of Distance Education. 1-20. https://doi.org/10.1080/08923647.2020.1696140

[3] Alenezi. H. S. &Faisal. M. H. (2020). Utilizing crowdsourcing and machine learning in education: Literature review. Education and Information Technologies. 1-16. https://doi.org/10.1007/s10639-020-10102-w

[4] El Guabassi. I. Al Achhab. M. Jellouli. I. & El Mohajir. B. E. (2016. October). Recommender system for ubiquitous learning based on decision tree. In 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt) (pp. 535-540). IEEE. https://doi.org/10.1109/cist.2016.7805107

[5] 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. https://doi.org/10.1016/j.compedu.2019.103724

[6] Qazdar. A. Er-Raha. B.. Cherkaoui. C. &Mammass. D. (2019). A machine learning algorithm framework for predicting students' performance: A case study of baccalaureate students in Morocco. Education and Information Technologies. 24(6). 3577-3589. https://doi.org/10.1007/s10639-019-09946-8

[7] Huang. A. Y. Lu. O. H. Huang. J. C. Yin. C. J. & Yang. S. J. (2020). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. Interactive Learning Environments. 28(2). 206-230. https://doi.org/10.1080/10494820.2019.1636086

[8] Waheed. H. Hassan. S. U. Aljohani. N. R. Hardman. J. Alelyani. S. & Nawaz. R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior. 104. 106189. https://doi.org/10.1016/j.chb.2019.106189

[9] El Guabassi. I. Al Achhab. M. Jellouli. I. &Mohajir. B. E. E. (2018). Personalized ubiquitous learning via an adaptive engine. International Journal of Emerging Technologies in Learning (iJET). 13(12). 177-190. https://doi.org/10.3991/ijet.v13i12.7918

[10] Syed. A. M. Ahmad. S. Alaraifi. A. & Rafi. W. (2020). Identification of operational risks impeding the implementation of eLearning in higher education system. Education and Information Technologies. 1-17. https://doi.org/10.1007/s10639-020-10281-6

[11] Bousalem. Z. El Guabassi. I. &Cherti. I. (2018. July). Toward adaptive and reusable learning content using XML dynamic labeling schemes and relational databases. In International Conference on Advanced Intelligent Systems for Sustainable Development (pp. 787-799). Springer. Cham. https://doi.org/10.1007/978-3-030-11928-7_71

[12] El Guabassi. I. Bousalem. Z. Al Achhab. M.. & EL Mohajir. B. E. (2019). Identifying learning style through eye tracking technology in adaptive learning systems. International Journal

of Electrical & Computer Engineering (2088-8708). 9. https://doi.org/10.11591/ijece.v9i5.pp4408-4416
[13] Bravo-Agapito. J. Romero. S. J. & Pamplona. S. (2020). Early Prediction of Undergraduate Student's Academic Performance in Completely Online Learning: A Five-Year Study. Computers in Human Behavior. 106595. https://doi.org/10.1016/j.chb.2020.106595
[14] Gray. C. C. & Perkins. D. (2019). Utilizing early engagement and machine learning to predict student outcomes. Computers & Education. 131. 22-32. https://doi.org/10.1016/j.compedu.2018.12.006
[15] Hamsa. H. Indiradevi. S. &Kizhakkethottam. J. J. (2016). Student academic performance prediction model using decision tree and fuzzy genetic algorithm. Procedia Technology. 25. 326-332. https://doi.org/10.1016/j.protcy.2016.08.114
[16] Hussain. M. Zhu. W. Zhang. W. Abidi. S. M. R. & Ali. S. (2019). Using machine learning to predict student difficulties from learning session data. Artificial Intelligence Review. 52(1). 381-407. https://doi.org/10.1007/s10462-018-9620-8
[17] Karthikeyan. V. G. Thangaraj. P. & Karthik. S. (2020). Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation. Soft Computing. 1-11. https://doi.org/10.1007/s00500-020-05075-4
[18] 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. https://doi.org/10.14257/ijdta.2016.9.8.13
[19] Amrieh. E. A. Hamtini. T. &Aljarah. I. (2015. November). 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. https://doi.org/10.1109/aeect.2015.7360581
[20] "Kalboard360-E-learning system". http://kalboard360.com/
[21] Rutherford. A. (2001). Introducing ANOVA and ANCOVA: a GLM approach. Sage.
[22] Kleinbaum. D. G.. Dietz. K. Gail. M.. Klein. M. & Klein. M. (2002). Logistic regression. New York: Springer-Verlag.
[23] Smola. A. J. &Schölkopf. B. (2004). A tutorial on support vector regression. Statistics and computing. 14(3). 199-222. https://doi.org/10.1023/b:stco.0000035301.49549.88
[24] Heien. D. M. (1968). A note on log-linear regression. Journal of the American Statistical Association. 63(323). 1034-1038. https://doi.org/10.2307/2283895
[25] Safavian. S. R. &Landgrebe. D. (1991). A survey of decision tree classifier methodology. IEEE transactions on systems. man. and cybernetics. 21(3). 660-674. https://doi.org/10.1109/21.97458
[26] Liaw. A. & Wiener. M. (2002). Classification and regression by randomForest. R news. 2(3). 18-22.
[27] Geladi. P.. & Kowalski. B. R. (1986). Partial least-squares regression: a tutorial. Analytica chimica acta. 185. 1-17. https://doi.org/10.1016/0003-2670(86)80028-9
[28] Swalin. A. (2018). Choosing the right metric for evaluating machine learning models.
[29] Miles. J. (2014). R squared. adjusted R squared. Wiley StatsRef: Statistics Reference Online. https://doi.org/10.1002/9781118445112.stat06627
[30] Willmott. C. J. & Matsuura. K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research. 30(1). 79-82. https://doi.org/10.3354/cr030079
[31] Addinsoft. X. (2015). Data analysis and statistics with MS Excel. Addinsoft. NY. USA. xlstat available at http://www.xlstat.com/en/home

11 Authors

Inssaf El Guabassi received his PhD in February 2019 from the Abdelmalek Essaadi University, Faculty of Sciences, Morocco, Tetouan.

Zakaria Bousalem is a Phd student at the Faculty of Sciences and Technology, Settat Morocco.

Rim Marah received his PhD in July 2018 from the Abdelmalek Essaadi University, Faculty of Sciences, Morocco, Tetouan.

Aimad Qazdar is an Assistant Professor at the Faculty of Sciences Semlalia, ISI Laboratory – Cadi Ayyad University in Marrakech, Morocco.

Article submitted 2020-11-24. Resubmitted 2020-12-08. Final acceptance 2020-12-11. Final version published as submitted by the authors.