

Predicting baccalaureate student result to prevent failure: a hybrid model approach

Abdesslam Essayad, Kassimi Moulay Abdellah

IMA Laboratory, ENSA, Université Ibn Zohr (UIZ) University, Agadir, Morocco

Article Info

Article history:

Received Apr 3, 2023

Revised Aug 18, 2023

Accepted Oct 21, 2023

Keywords:

Baccalaureate

Classification

Linear regression

Machine-learning

Student performance

ABSTRACT

The Moroccan Ministry of National Education has seen substantial modifications over the previous ten years, which have contributed to improving the quality of education. However, there is a discrepancy in the percentage of academic achievement between the regional directorates and educational institutions. Machine learning techniques have become a powerful tool for proactively predicting student admission. The goal of our paper is to build machine learning models using various algorithms to predict the final baccalaureate school year outcomes. We compare regression and classification to find the reasons behind students' failure and to choose an appropriate model for predicting the results. This helps decision-makers make appropriate interventions.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Abdesslam Essayad

IMA Laboratory, ENSA, UIZ University

Agadir, Morocco

Email: abdesslam.essayad@edu.uiz.ac.ma

1. INTRODUCTION

The Moroccan Ministry of National Education has implemented significant reforms over the past twenty years, including implementing the educational information system (MASSAR) for the year 2012, organizing and controlling the absences of teachers and students, the emergency program that extended from 2009 to 2013, and framework law 51.17, which established the strategic vision for the years 2015–2030. According to the Program for International Student Assessment (PISA) 2018 [1], Morocco came in 75th out of 79 countries in the ranking factors, which evaluated the ability to read, mathematics, and science skills of 15-year-old pupils around the world [2].

Ahajjam *et al.* [3] predict the performance of Moroccan students at the baccalaureate level using an intelligent system that relies on machine learning techniques such as neural networks to extract data to determine the student's basic average. It also aims to design and implement a system to direct students from common streams to one of the technical, literary, or scientific baccalaureate divisions. In their study, the researchers created a methodology to predict student performance using a machine learning algorithm. Student data was exploited from the MASSAR system for a secondary school in Morocco between 2016 and 2018. The researchers concluded that the chosen model could provide good accuracy for students' performance. Qazdar *et al.* [4] built a model using multivariate regression to predict the performance of students in a Moroccan secondary school over three years of study. Their goal was to combine their model with the "MASAR" system to predict students' early results for the purpose of increasing their performance. In contrast to the percentage obtained, which amounted to 37% according to the results of the predictive model, 42% of the students were at risk of failing. The system indicated that 29% of students completed the academic year successfully, although the actual number was closer to 21%. In order to support students in achieving better

academic achievements and raise the level of system outcomes generally, Alamri *et al.* [5] have attempted to anticipate students' academic success, identify obstacles and problems that affect student outcomes, and give ways to improve the educational system. The research we are presenting is focused in the same direction as previous research, except that it compares the efficiency of the selected models between linear regression and selection, as well as investigates the reasons that prevent achieving good results for success in the baccalaureate. We seek to forecast the results of the current academic year using the baccalaureate results from the previous two academic years. The goal is to predict what will happen at the end of the current academic year immediately after the results of the first semester have been released, exactly six months before the final exam results. The respective departments will receive the anticipated results in order to identify students who require immediate school support. Additionally, pre-intervention can be used to help students who are anticipated to fail by identifying them and letting the administration and teachers know who they are.

The remainder of this paper is structured as follows: in the second section, we discuss the justification for the study. The third topic we'll cover is comparable work. In the fourth section, we will focus on data processing and modeling by creating a solid model and attempting to use it on a newly created dataset. The resulting outcome is covered and discussed in Section 5. We will conclude the article with some concepts and recommendations that may be useful in order to improve the success rate of baccalaureate students.

2. SEARCH GROUNDS

Obtaining a baccalaureate degree is considered the end of secondary education and the gateway to higher education, with credibility at the national and international levels. According to the Moroccan Finance Law of 2021 [6], the budget allocated to the public education sector has increased; it is almost 22% of its budget and seven percent of gross domestic product (GDP). Additionally, Moroccan families have doubled their expenses for their children's education [7]. These numbers indicate that students' failure when the academic year has ended is costly for families and the state in terms of financial and human resources.

In Table 1, we note that the baccalaureate pass rate has clearly increased, even though the last few years have been punctuated by the COVID-19 pandemic. The concerned ministry has intensified its efforts to maintain the rhythm of the baccalaureate level. The ministry has allocated lessons on platforms including a public television channel as well as a free-access website.

Table 1. The percentage of successful Moroccan students at the baccalaureate [8]

Year	2017	2018	2019	2020	2021
Number of candidates attending the baccalaureate exam	352,303	387,933	323,668	311,758	323,022
Overall success rate %	65.20	71.91	77.96	79.62	81.83

The strengthened regional structure expanded since 2016, which has given regional education and training academies more authority, is one reason for this rise in the success rate. The Supreme Council for Education and Training has created a strategic vision for the 2015–2030 reform of a school of quality and equity to address weaknesses in the Moroccan educational system. Additionally, of the 18 projects of the Framework Laws (51.17), nine were adopted with the aim of improving educational work. The establishment of an information system called "MASSAR" also contributed to the management and investment of school results [2].

Early detection of student failure enables management to offer early coaching and counseling to students in order to improve success rates and student retention. Academic administrators will find it easy to arrange some important interventions to enhance the performance of students who are expected to fail in the middle of the academic year, thus helping to open up prospects for successful completion of their studies [9]. This research aims to predict the final results of the baccalaureate in order to enable early decision-making by the administration in cooperation with teachers and to provide school support, especially for students who are close to success, and increase the success rate.

3. RELATED WORKS

Many studies are concerned with students' performance. In the paper "Prediction of Student Success Using Enrollment Data" explained by Cengiz and Uka [10]. According to the researchers, early detection of students who are at a high risk of failing will allow for immediate intervention by educators with the appropriate measures, which will raise the graduation rate. Strecht *et al.* [11] use socioeconomic characteristics such as age, sex, marital status, nationality, displacement, scholarship, special needs, and kind of admission to forecast students' performance and grade in a class. Mohamed and Husein [12] employed machine learning algorithms

for learning management, academics, socioeconomics, and student demographics. Kraft *et al.* [13] discovered that the school climate can affect students' performances. Mesarić *et al.* [14] the objectives of this study are to develop a model that efficiently divides students into two groups based on how well they performed at the end of the first school year and to identify the important variables that have a major impact on performance. Xu *et al.* [15] research was done to determine how to forecast students' future success based on their current academic records. Kavitha *et al.* [16] attempted to use the analysis's findings to pinpoint at-risk students and offer suggestions for improvement. Berens *et al.* [17], in their research, they discover student traits that set probable dropouts apart from graduates using regression analyses, neural networks, decision tree methods, and the AdaBoost algorithm. Griffioen *et al.* [18] examine student satisfaction and performance. Kumari *et al.* [19] investigates student behavioral features that play an important role in showing the student's interactivity with the e-learning system. Aysha and Khan [20], create a model that admission-seeking students can employ to evaluate their performance in the chosen course, claiming that educational systems require creative approaches to raise educational quality to get the best outcomes and lower the failure rate. Sedova *et al.* [21], according to their paper, there is a substantial correlation between a student's achievement and the amount of conversation time they had in class, the number of times they used reasoning, and other factors. Suhaimi *et al.* [22], in their paper, compared interests among those who were interested in studying the performance of students by gender; they discovered that the methodology used in studies involving men and women differs. Imran *et al.* [23] examines the identification of students at risk based on socioeconomic and cultural factors. Lorenz and Wigger [24] utilizes two methods, including decision trees and logistic regressions, to predict dropping out of school using exam data.

4. METHOD

4.1. Data collection

Data processing is essential for academics to carry out initial survey screening, followed by editing and coding of the data [25]. Data processing allows for the measurement of the data's accuracy, completeness, consistency, timeliness, credibility, and interpretability [26]. The framework in Figure 1 is adapted from the cross-industry standard practice for data mining (CRISP-DM) approach. CRISP-DM is a project funded by the European Commission that creates a common paradigm for conducting data mining projects [27].

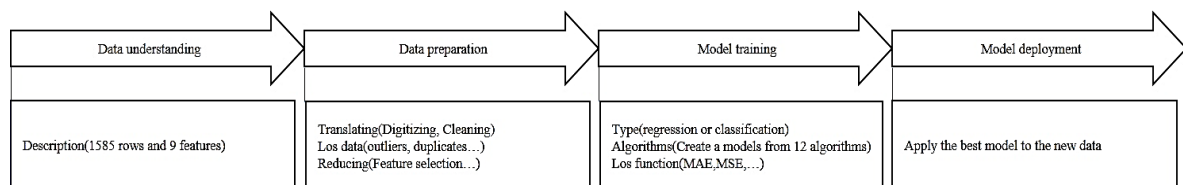


Figure 1. Framework of data processing and modelization

The data was approved and obtained from El Hajeb's Ministry of National Education Directorate in accordance with the Right to Information Law (No. 31-13) through the Transparency Portal and Access to Information (Chafafiya) [28]. The dataset includes two school years' worth of baccalaureate results. It was processed and purified, and only the correct data was kept. This data is described in Table 2 and contains 1,585 rows, 8 features, and 1 target.

Table 2. Dataset description

Column	Type	Description
id school	int64	Institution ID
id center	int64	Center (urban, rural)
id level	int64	Level ID
Gender	int64	Gender (male or female)
reg average	float64	Regional exam rate represents 25% of the pass rate
cc average	float64	Continuous monitoring rate represents 25% of the pass rates
ga average	float64	The general average that combines the two previous averages and the final exams
admis	int64	The result of admission

4.2. Data cleaning

Some of the goals of data cleaning include detecting anomalies and resolving the problem of missing and duplicate values. To ensure that the dataset is reliable and produces accurate predictions. The following tasks were performed prior to using the models: i) Removing unknown or null values from the data; ii) Converting the data in each column from categorical to numerical values; and iii) Selecting only the columns that have a high degree of influence or a strong correlation with the input value (feature selection).

4.3. Data analysis

Data analysis aims to gain an in-depth understanding of the data before it can be utilized. It also saves a significant amount of time, especially in determining whether the data is relevant to the problem we are trying to solve. Data analysis is just one of the many steps that must be taken when conducting research; nevertheless, it takes on special relevance when it is used in many different fields to guide businesses and organizations to make wiser financial decisions [29]. Researchers and companies are urged to adopt algorithms that process real-time data, analyze it, and deliver highly accurate analytics conclusions due to the complexity of data analysis and interpretation [30]. Machine learning, data science, and predictive modeling have become widely distributed in any sector where analyzing data is crucial [31]. During the data analysis process, we examine the correlations between different variables to determine which of these correlations is the strongest and has the greatest impact. According to Table 3, there is a strong relationship between the general average and the continuous monitoring rate, which is 81%. However, the relationship between the regional average and the general average is 71%.

Figure 2 shows that the relationship between continuous monitoring and the general average is good; it is stacked between 12 and 15, and most of the points found in the linear regression. Nevertheless, the relationship between the regional average and the general average is not perfect, and we find it stacked between six and 12. In Figure 2, there is a good regression between the cc_average and ga_average, although there are many points that deviate from the regression line between the reg_average and ga_average.

Table 3. Correlation between the input and output variables

	cc_average	reg_average	ga_average
cc_average	1.000000	0.484635	0.810675
reg_average	0.484635	1.000000	0.709361
ga_average	0.810675	0.709361	1.000000

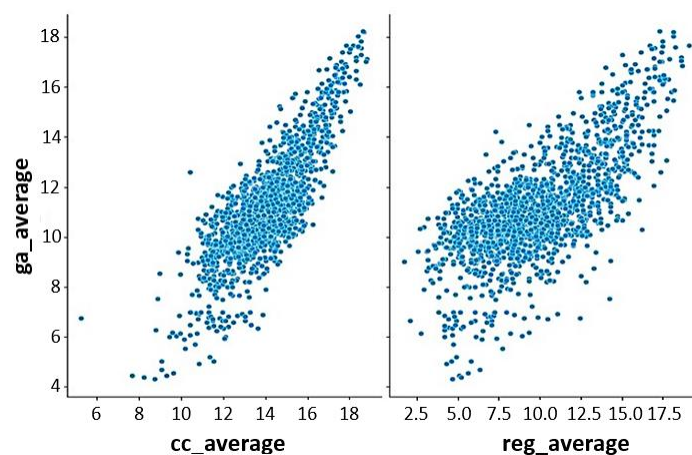


Figure 2. Regression between the input and the output variables

According to Table 4, the mean of the continuous monitoring rates is 14.03, whereas the mean of the regional average is 9.73, which has a favorable impact on the baccalaureate success rate. It is noted that the regional average has a negative impact on the success rate. Therefore, educational officials should reconsider passing the regional exam one year before passing the final baccalaureate exam. The second problem is passing this exam in specializations that will not be part of their future orientations. There is no doubt that this will lead to a decrease in their success rate at the end of the second year of the baccalaureate.

Table 4. Features description

	count	mean	std	min	25%	50%	75%	max
reg_average	1,585	9.73	3.40	3.40	1.70	7.08	9.37	12.16
cc_average	1,585	14.04	1.67	1.67	7.67	12.95	14.06	15.10
ga_average	1,585	11.12	2.10	2.10	4.30	10.04	10.78	12.06

In Figure 3, the rate of continuous observation has a positive effect on the pass rate, while the rate of regional examination has a negative effect for the aforementioned reasons. However, the coefficient of continuous monitoring and the regional examination have the same value. This shows that the negative impact should be the subject of discussion with officials to avoid a negative impact on the baccalaureate results.

To succeed at the baccalaureate level, an average of 10 or more is required. Figure 4 shows that about 15% are close to the success rate. This statistic informs the concerned administration about targeting this sample of students with academic support to raise the success rate by 15% to more than 90%.

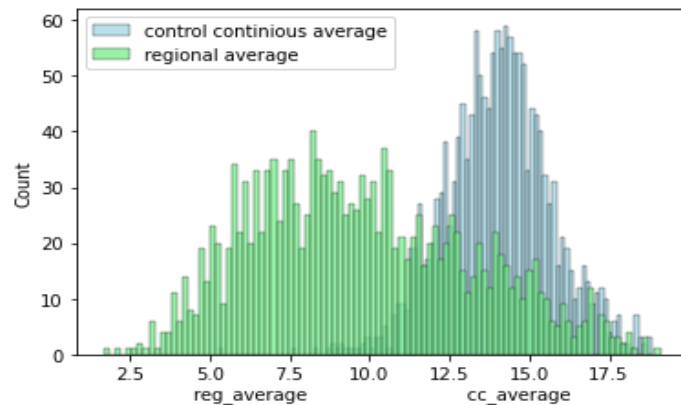


Figure 3. Difference between the effect of the reg average and the cc average

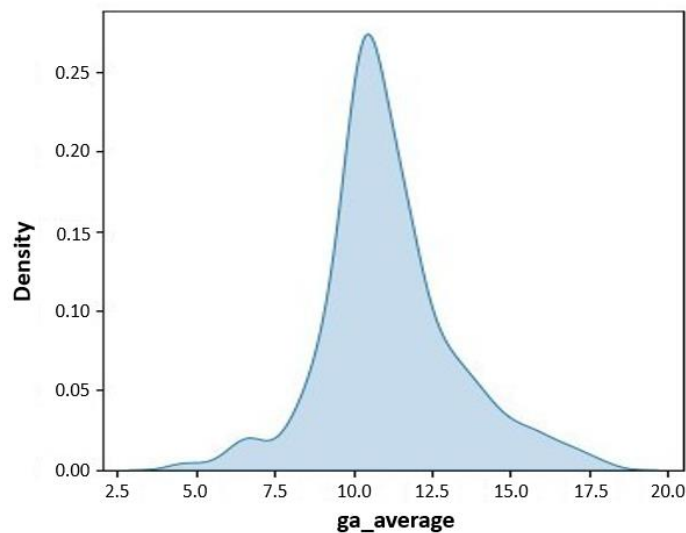


Figure 4. Distribution of general average

4.4. Modelization

Machine learning is a group of techniques that use data, such as classification or regression, to build models for performance prediction and decision-making. We will use a variety of machine-learning algorithms in our search for the most accurate indicators of the results once the academic year is over. Figure 5 shows the life cycle of managing machine learning models.

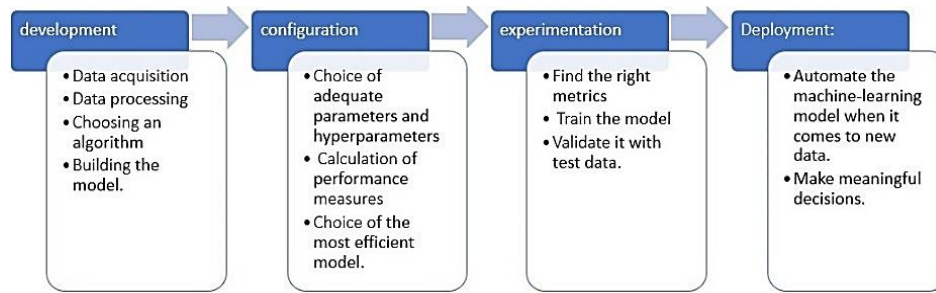


Figure 5. Lifecycle of model management

Recognizing students who are at risk of failure to take preventative measures is one of our objectives, such as providing educational support to help them succeed when the school year comes to a close. Therefore, we need a dataset for the first semester of each school year and work on preparing it to apply the choice model, which, at the end of the school year, displays the students' results. This problem falls under supervised learning and can be represented as a regression or classification model. To achieve this, there are many machine-learning algorithms, each with many hyperparameters to define.

5. RESULT

5.1. Classification using various evaluation matrix

The categorization problem is a common topic of study for experts in data mining (DM) and machine learning (ML) [32]. Based on the values of other attributes, the category attribute's value will be predicted [32]. It is important to construct a set of models that explain and classify data categories and concepts in order to use the model for predicting the value of an anonymous class [26].

As shown in Table 5, we compared 11 models using the Pycaret [33] library to determine their performance differences. By analyzing each model's performance, we can select the faster model based on the test results and use it to achieve the best predictive results. The gradient boosting method uses binary decision trees as its main predictors, and after selecting the hyperparameters, the CatBoost Classifier was shown to be the most effective model [34].

Table 5. Models' performance using Pycaret library

	Model	Accuracy	AUC	Recall	Prec.
catboost	Catboost classifier	0.8547	0.9013	0.9263	0.8907
lda	Linear discriminant analysis	0.8538	0.9077	0.9460	0.8759
et	Extra trees classifier	0.8538	0.9004	0.9226	0.8925
ridge	Ridge classifier	0.8528	0.0000	0.9693	0.8595
gbc	Gradient boosting classifier	0.8481	0.8984	0.9202	0.8882
ada	Ada boost classifier	0.8462	0.8958	0.9190	0.8868
lr	Logistic regression	0.8443	0.8968	0.9338	0.8746
rf	Random forest classifier	0.8433	0.9004	0.9202	0.8827
qda	Quadratic discriminant analysis	0.8396	0.8953	0.8993	0.8940
xgboost	Extreme gradient boosting	0.8376	0.8948	0.9128	0.8828

5.1.1. Confusion metrics

The efficacy of predictions in a classification model can be evaluated using a confusion matrix [35]. They are frequently employed to evaluate the accuracy of several machine learning models. After using the test data with the CatBoost Classifier model, we obtained the following confusion matrix in Figure 6.

The confusion matrix table serves as an example of a tabular display that evaluates the model's accuracy in predicting. The matrix compares the predicted values by machine learning models to the real values. The matrix shows how each model categorizes errors and whether it is better at predicting the proportion of students who succeed or fail.

5.1.2. Evaluation metrics using CatBoost classifier model

The three different forms of evaluation metrics are threshold measurement, likelihood, and ranking metrics [36]. As Table 6 shows, the accuracy of the CatBoost Classifier model is high, at 85%, and it predicts the percentage of students who pass well (Precision=89%). The model also has a high rate of sensitivity (Recall), which is about 93%, which means that our model can identify most of the number of students with success. Also, the model predicts the success of students well but predicts less well the percentage of students

who fail (specificity=73%). We can increase the percentage of specificity by reducing the outliers, or we can choose another model, like Xgboost, that predicts well the failure of students; in this case, the administration can focus only on the students who the model predicts to fail.

Precision and recall are often used in combination to evaluate the performance of a classifier, and a balance between precision and recall is desired. High precision means that the model accurately predicts the student's success. However, a low recall means that the model might miss many student successes.

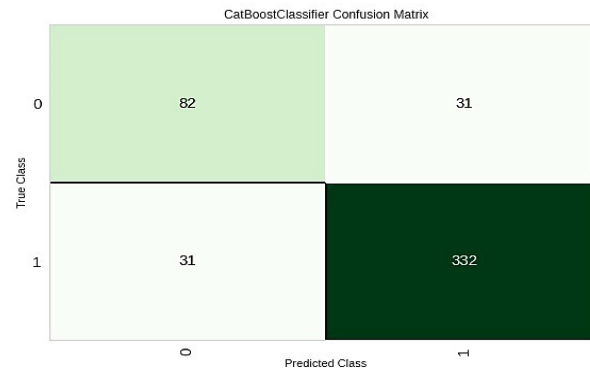


Figure 6. Confusion matrix result using CatBoost classifier

Table 6. Description of the different evaluation metrics using CatBoost Classifier model

Metrics name	Description	Formula	Result
Accuracy	Means how often is the model correct.	$\frac{TP + TN}{TP + TN + FP + FN}$	85%
Precision	Determines how well a model predicted positive results, such as student achievement in terms of accuracy.	$\frac{TP}{TP + FP}$	89%
Specificity	Means the percent of students whose failure was accurately predicted by the model.	$\frac{TN}{TN + FP}$	73%
Sensitivity	Measure the true positive rate or recall (TPR), often known as the precision of a model to recognize student success.	$\frac{TP}{TP + FN}$	93%

5.1.3. Learning curve

The model has an accuracy of 85%, a recall of more than 90%, and a high precision of 89%. This indicates that the model can accurately predict which students will be accepted at the end of the academic year. A ROC curve, which plots the true positive rates (TPR) and false positive rates (FPR) versus both of them, is a visual representation of the efficacy of a binary classification system.

Figure 7 displays for us the ROC, which is a metric that measures the ability of a binary classifier to distinguish well between successful and failing students. The area under the receiver (AUC) metric is yet another popular way to measure how well a binary classifier performs. The model's accuracy in identifying student outcomes increases with the AUC. We have an AUC of 0.92, which is around the ideal level of 1 for a classifier.

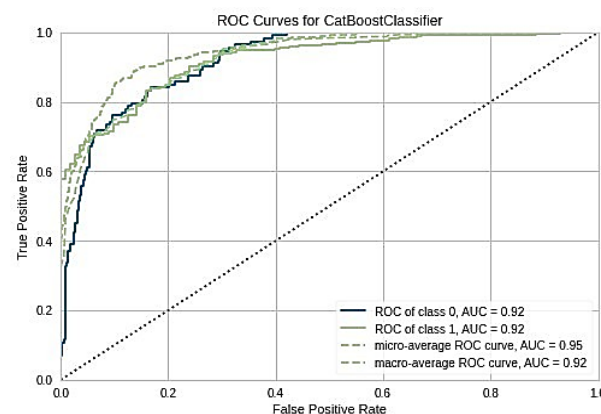


Figure 7. Receiver operating characteristic curves (ROC)

5.2. Regression using many functions of evaluation matrix

Regression was primarily used to determine which independent factors were significant predictors of the outcome and whether the independent variables were correct in predicting the result variable [37]. The mathematical formula $y = mx + c$ is used in linear regression analysis to determine the line of greatest fitting for the correlation between the dependent variable (y) and the independent variable (x) [38]. Gradient Boosting algorithms can optimize different functions, which makes them flexible, and they often provide predictive accuracy [39]. After training and testing the data with the Gradient Boosting model using many cost functions, as shown in Table 7, we got MSE 0.7686. Gradient boosting, a powerful machine learning model, successfully handles categorical data and takes advantage of handling it during training rather than during pre-processing time to achieve the greatest results on a variety of practical issues [40].

Table 7. Models performance using Pycaret library

	Model	MAE	MSE	RMSE	R2
gbr	Gradient boosting regressor	0.6769	0.7686	0.8744	0.7898
catboost	Catboost regressor	0.6833	0.7887	0.8844	0.7842
br	Bayesian ridge	0.7263	0.8550	0.9190	0.7638
ridge	Ridge regression	0.7269	0.8553	0.9191	0.7637
lr	Linear regression	0.7284	0.8578	0.9204	0.7628
lar	Least angle regression	0.7284	0.8578	0.9204	0.7628
huber	Huber regressor	0.7296	0.8686	0.9270	0.7616
lightgbm	Light gradient boosting machine	0.7174	0.8590	0.9240	0.7595
et	Extra trees regressor	0.7224	0.8867	0.9385	0.7554
rf	Random forest regressor	0.7219	0.8892	0.9407	0.7532
xgboost	Extreme gradient boosting	0.7322	0.9118	0.9518	0.7477

5.2.1. Evaluation metrics for gradient boosting regressor model

Utilizing evaluation measures, we may gauge the model's efficacy. The evaluation scale that will be utilized will be determined by the unique tasks and objectives of the model. Error measurements can be used to examine the remaining differences between actual and predicted values. The minimal mean residual suggests that the model is successful in predicting our situation.

Outliers are observations in a dataset that are significantly different from most of the data. They can have a significant impact on the results predicted by the model, as they can greatly affect the relationships between variables and the overall fit of the model. Excluding outliers can be a useful approach if they are truly anomalous and do not represent the underlying pattern in the data. As shown in Table 8, the outliers can affect the efficiency of the model. The explanation for this is that the students in these cases succeeded by cheating on the exam, failed due to their absence, or obtained a zero in one of the subjects.

Table 8. Evaluation metrics description

Metrics name	Outliers' impact	Formula	Value without outliers	Value with outliers
Mean squared error (MSE)	Outliers can greatly increase the MSE	$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	0.7686	1.9561
Root mean square error (RMSE)	RMSE is sensitive to outliers	$\sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}}$	0.8744	1.3882
Mean absolute error (MAE)	MAE is less to be influenced by extreme values	$\frac{1}{n} \sum_{i=1}^n Y_i - \hat{Y}_i $	0.6769	1.0458
R-square (R ²)	Outliers can have an effect on the R ² value	$1 - \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2}$	0.7898	0.5460

The MSE (0.7686 vs. 1.9561) and RMSE (0.8744 vs. 1.3882) are sensitive to outliers, as they involve squaring the variation from the expected and actual values. This means that outliers in the data can significantly affect the value of these metrics. MAE (0.6769 vs. 1.0458), which is the absolute value of the differences between expected and actual, could have a significant impact on outliers if not reduced.

An indicator of how well a model fits the training data is its R-squared value. It represents the percentage of `ga_average` that the regression model's `cc_average` and `reg_average` are able to predict. R² calculates the proportion of the change in the target variable that can be assigned to the model's predictor variables. A high R² (0.7898) value indicates that the model fits the training data well and explains a large proportion of the variation in the target variable, while a low R² (0.5460) value indicates that the model does not fit the data well and explains that the outliers can have a negative effect on the R². The prediction error R², also known as the cross-validated R², is a metric used to assess how well a model predicts fresh data that isn't

included in the training set. The prediction error R^2 and the R-squared R^2 are both measures of regression model performance, but they have different interpretations and applications. The prediction error R^2 (0.823) for a gradient boosting regressor, as shown in figure 8, is a measure of how well the model predicts new data that is not included in the training set, while the R-squared R^2 (0.7898) determines exactly how the model fits the training collected data and how much of the variance in the target variable is explained by the predictor variables. It is also valuable for predicting how well the model will perform on new data.

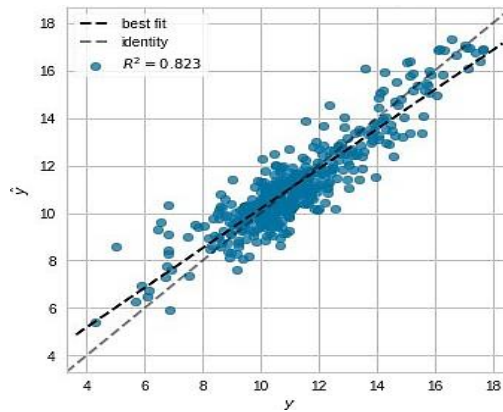


Figure 8. Prediction error for gradient boosting regressor model

6. DISCUSSION

In this paper, we employed several classification and regression models; however, we still need to compare them to determine which model is most suitable for making accurate predictions. For classification, we didn't take into account how close students were to the success criteria; instead, we solely considered forecasting whether a student's admission would be successful or unsuccessful. However, using linear regression, we were able to quantify each level of participation in the classroom. Students' learning was directly impacted by classroom management and interactions between students and teachers on both an academic and social level [41].

The regression model is useful in comparing the competencies of students and teachers and pinpointing where a large percentage of at-risk students, which is 15%, are located between nine and 10, as shown in Figure 9. Using the regression model, students can be grouped according to their needs for school support in specific subjects. This allows them to receive the appropriate level of support, whether full or partial, to meet their needs.

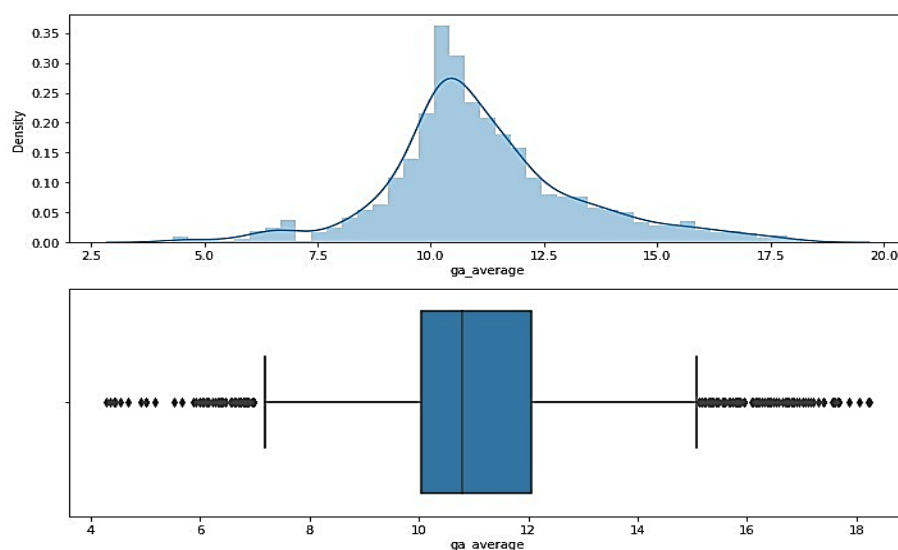


Figure 9. Breakdown of pass averages

The model's accuracy in predicting academic success rises as the number of outliers decreases. The presence of these values indicates that the results the concerned students obtained were far from those predicted by the model. The explanation for this is either that the students made more efforts in the second semester, which led to their success (these cases are rare), or they cheated on the final exam, which helped them succeed.

Choosing the ideal model that can be applied may not be an easy process, and that model can be applied across different types of models of the same type configured with different hyperparameters. We leave the choice to officials according to whether they should consider the outliers or not. As shown in Table 7, it is better to adopt the models (gradient boosting regressor, CatBoost regressor) in regression or, as shown in Table 5, (CatBoost classifier, linear discriminant analysis) in classification.

We intend to create a web application using the lightweight web framework Flask in order to deploy our model to production. We will extract predictive results from the model through the web application and generate a list of students whose model predicts they are at risk of failing. Officials at the educational institution will be able to identify students who need additional help and reinforcement on specific subjects as a result of this prediction.

7. CONCLUSION

This article explains how predicting baccalaureate graduation rates based on their academic abilities. To evaluate the efficacy of regression and classification models, we compared them. Informed decisions regarding how to support challenging students can be made by school administrators with the help of our methodology, which has been put into practice. In order to find solutions to lower failure rates, future studies can examine students' socioeconomic and health statuses, as well as the educational procedures that affect their academic achievement.




REFERENCES

- [1] M. Bouzahzah, "Quality of the education system and economic growth. Projections in the case of Morocco," *WSEAS Transactions on Business and Economics*, vol. 18, pp. 949–961, Jun. 2021, doi: 10.37394/23207.2021.18.90.
- [2] A. Essayad and M. A. Kassimi, "Predicting the baccalaureate students admission: the influence of teacher and administration," *ITM Web of Conferences*, vol. 43, p. 01013, Mar. 2022, doi: 10.1051/itmconf/20224301013.
- [3] T. Ahajjam, M. Moutaib, H. Aissa, M. Azrour, Y. Farhaoui, and M. Fattah, "Predicting students' final performance using artificial neural networks," *Big Data Mining and Analytics*, vol. 5, no. 4, pp. 294–301, Dec. 2022, doi: 10.26599/BDMA.2021.9020030.
- [4] A. Qazdar, B. Er-Raha, C. Cherkaoui, and D. Mammass, "A machine learning algorithm framework for predicting students performance: a case study of baccalaureate students in Morocco," *Education and Information Technologies*, vol. 24, no. 6, pp. 3577–3589, Nov. 2019, doi: 10.1007/s10639-019-09946-8.
- [5] L. H. Alamri, R. S. Almuslim, M. S. Alotibi, D. K. Alkadi, I. Ullah Khan, and N. Aslam, "Predicting student academic performance using support vector machine and random forest," in *2020 3rd International Conference on Education Technology Management*, Dec. 2020, pp. 100–107, doi: 10.1145/3446590.3446607.
- [6] "Moroccan Finance Law of 2022," *Bulletin Officiel Cent-dixième année*, vol. 7049, 2022.
- [7] S. Maghnouj, J. Bélanger, M. Clarke, E. Fordham, H. Kitchen, and I. McGregor, "OECD review of the education assessment framework-Morocco (In French: examen de l'OCDE du cadre d'évaluation de l'éducation-Maroc)," *Résumé le*, vol. 7, p. 2020, 2018.
- [8] M. M. of National Education, "Compendium of education statistics (In French: recueil statistiques de l'éducation)," 2019.
- [9] T. Thilagaraj and N. Sengottaiyan, "A review of educational data mining in higher education system," in *Proceedings of the Second International Conference on Research in Intelligent and Computing in Engineering*, Jun. 2017, vol. 10, pp. 349–358, doi: 10.15439/2017R87.
- [10] N. Cengiz and A. Uka, "Prediction of student success using enrolment data," *KOS*, vol. 14, no. 17, pp. 42–45, 2014.
- [11] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, and R. Abreu, "A comparative study of classification and regression algorithms for modelling students' academic performance," *International educational data mining society*, 2015.
- [12] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.
- [13] M. A. Kraft, W. H. Marinell, and D. Yee, "Schools as organizations," *New York*, 2016.
- [14] J. Mesarić and D. Šebalj, "Decision trees for predicting the academic success of students," *Croatian Operational Research Review*, vol. 7, no. 2, pp. 367–388, Dec. 2016, doi: 10.17535/corr.2016.0025.
- [15] J. Xu, K. H. Moon, and M. van der Schaar, "A machine learning approach for tracking and predicting student performance in degree programs," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 5, pp. 742–753, Aug. 2017, doi: 10.1109/JSTSP.2017.2692560.
- [16] G. Kavitha and L. Raj, "Educational data mining and learning analytics-educational assistance for teaching and learning," no. 1, Mar. 2017.
- [17] J. Berens, K. Schneider, S. Görtz, S. Oster, and J. Burghoff, "Early detection of students at risk—predicting student dropouts using administrative student data and machine learning methods,"
- [18] D. M. E. Griffioen, J. J. Döppenberg, and R. J. Oostdam, "Are more able students in higher education less easy to satisfy?," *Higher Education*, vol. 75, no. 5, pp. 891–907, May 2018, doi: 10.1007/s10734-017-0176-3.
- [19] P. Kumari, P. K. Jain, and R. Pamula, "An efficient use of ensemble methods to predict students academic performance," in *Proceedings of the 4th IEEE International Conference on Recent Advances in Information Technology, RAIT 2018*, 2018, pp. 1–6, doi: 10.1109/RAIT.2018.8389056.
- [20] A. Ashraf, S. Anwer, and M. G. Khan, "A comparative study of predicting student's performance by use of data mining techniques," *American Scientific Research Journal for Engineering, Technology and Sciences*, vol. 44, no. 1, pp. 122–136, 2018.




- [21] K. Sedova *et al.*, "Do those who talk more learn more? the relationship between student classroom talk and student achievement," *Learning and Instruction*, vol. 63, p. 101217, Oct. 2019, doi: 10.1016/j.learninstruc.2019.101217.
- [22] N. M. Suhaimi, S. Abdul-Rahman, S. Mutalib, N. A. Hamid, and A. Hamid, "Review on predicting students' graduation time using machine learning algorithms," *International Journal of Modern Education and Computer Science*, vol. 11, no. 7, pp. 1–13, Jul. 2019, doi: 10.5815/ijmecs.2019.07.01.
- [23] M. Imran, S. Latif, D. Mehmood, and M. S. Shah, "Student academic performance prediction using supervised learning techniques," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 14, no. 14, p. 92, Jul. 2019, doi: 10.3991/ijet.v14i14.10310.
- [24] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting student dropout: a machine learning approach," *European Journal of Higher Education*, vol. 10, no. 1, pp. 28–47, Jan. 2020, doi: 10.1080/21568235.2020.1718520.
- [25] T. Shukla, "Data Processing," May 2018.
- [26] S. Singhal and M. Jena, "A study on WEKA tool for data preprocessing, classification and clustering," *International Journal of Innovative Technology and Exploring Engineering*, vol. 2, no. 6, pp. 250–253, 2013.
- [27] P. Chapman *et al.*, "CRISP-DM 1.0: Step-by-step data mining guide," *SPSS inc*, vol. 78, no. 13, pp. 1–78, 2000.
- [28] "Transparency portal and access to information (Chafafiya)," Ministry of Digital Transition and Administration Reform, 2020.
- [29] N. Singh and A. K. Singh, "Data analysis in business research: key Concepts," *International Journal of Research in Management & Business Studies*, vol. 2, no. 1, p. 1, 2015.
- [30] S. S. Abdul-Jabbar and A. K. Farhan, "Data analytics and techniques," *ARO-The Scientific Journal Of Koya University*, vol. 10, no. 2, pp. 45–55, Oct. 2022, doi: 10.14500/aro.10975.
- [31] B. Butcher and B. J. Smith, "Feature engineering and selection: a practical approach for predictive models," *The American Statistician*, vol. 74, no. 3, pp. 308–309, Jul. 2020, doi: 10.1080/00031305.2020.1790217.
- [32] R. S. J. de Baker, T. Barnes, and J. E. Beck, "Educational data mining 2008," 2008.
- [33] M. Ali, "PyCaret: an open source, low-code machine learning library in Python," 2020.
- [34] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [35] J. Ortega, A. C. Lagman, E. T. Natividad, Lizel Rose Q Bantug, and L. Resurreccion, Michael R Manalo, "Analysis of performance of classification algorithms in mushroom poisonous detection using confusion matrix analysis," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.3, Jun. 2020, doi: 10.30534/ijatcse/2020/7191.32020.
- [36] R. Caruana and A. Niculescu-Mizil, "Data mining in metric space: an empirical analysis of supervised learning performance criteria," in *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 69–78.
- [37] Y. Koloğlu, H. Birinci, S. I. Kanalmaz, and B. Özyılmaz, "A multiple linear regression approach for estimating the market value of football players in forward position," *arXiv*, 2018.
- [38] K. Kumari and S. Yadav, "Linear regression analysis study," *Journal of the Practice of Cardiovascular Sciences*, vol. 4, no. 1, p. 33, 2018, doi: 10.4103/jpcs.jpcs_8_18.
- [39] P. Kulkarni, S. Karwande, R. Keskar, P. Kale, and S. Iyer, "Fake news detection using machine learning," in *ITM Web of Conferences*, 2021, vol. 40.
- [40] V. Dorogush, Anna Veronika Ershov and A. Gulin, "CatBoost: gradient boosting with categorical features support," *arXiv preprint arXiv:1810.11363*, 2018.
- [41] E. A. O'Connor and M. C. Fish, "Diferences in the clasrom systems of expert and novice teachers," 1998, p. 27.

BIOGRAPHIES OF AUTHORS



Abdesslam Essayad    holds a bachelor in electronics and a master of science in computer science. He is currently employed with the Directorate of the National Ministry of Education in El Hajeb, Morocco, where he is responsible for Program Genie. His research interests are in the fields of education and artificial intelligence. His interdisciplinary research program's objective is to examine how machine learning and deep learning techniques might influence instructional strategies and lead to creative administration-level educational solutions. He can be contacted at email: abdesslam.essayad@edu.uiz.ac.ma.



Kassimi Moulay Abdellah    received the D.Sc. degree (Doctor Habilitatus D.Sc.) in computer science from Sidi Mohamed Ben Abdellah University. He is also a professor of Computer Science at the University of Charia Ait Meloul, Morocco, since 2012. He is a member of the Information Technology, Data, Mathematics, and Applications Science Team at ENSA Agadir. His research interests are in natural language processing (NLP), semantic search engines, including word embeddings, classification, question-answering, and other related topics. He is also interested in information retrieval and machine translation. He can be contacted at email: m.kassimi@uiz.ac.ma.