



# A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco

Aimad Qazdar<sup>1</sup>  · Brahim Er-Raha<sup>1</sup> · Chihab Cherkaoui<sup>2</sup> · Driss Mammass<sup>3</sup>

Received: 29 March 2019 / Accepted: 7 June 2019 / Published online: 25 June 2019

© Springer Science+Business Media, LLC, part of Springer Nature 2019

## Abstract

The use of machine learning with educational data mining (EDM) to predict learner performance has always been an important research area. Predicting academic results is one of the solutions that aims to monitor the progress of students and anticipates students at risk of failing the academic pathways. In this paper, we present a framework for predicting student performance based on Machine Learning algorithm at H.E.K high school in Morocco from 2016 to 2018. The proposed model was analyzed and tested using student's data collected from The School Management System "MASSAR" (SMS-MASSAR). The dataset used in this study concerns 478 Physics students during the school years: 2015–2016, 2016–2017 and 2017–2018. The predictive performance results showed that our model can make more precise predictions of student's performance.

**Keywords** Machine learning · Educational data mining · Decision support tools · Predictive model · Academic performance

---

<https://massar.men.gov.ma/>

---

✉ Aimad Qazdar  
aimad.qazdar@gmail.com

Brahim Er-Raha  
erraha@gmail.com

Chihab Cherkaoui  
ccherkaoui@gmail.com

Driss Mammass  
mammass@uiz.ac.ma

<sup>1</sup> IMA Laboratory, ENSA / IRF-SIC Laboratory, FSA, Ibn Zohr University, Agadir, Morocco

<sup>2</sup> IRF-SIC Laboratory, ENCG, Ibn Zohr University, Agadir, Morocco

<sup>3</sup> IRF-SIC Laboratory, FS, Ibn Zohr University, Agadir, Morocco

## 1 Introduction

Since the launch of the first initiatives to integrate information and communication technologies in education in Morocco, schools and universities have been generating and storing a huge amount of student data in electronic form. This massive amount of student data can be more useful and can guide the decision-making at the level of ministry, regional academies, provincial directorates, and institutions if it is analyzed and transformed effectively into knowledge. Such data can also improve the quality of teaching and learning, and hence help students succeed in their academic pathways.

According to the work of (Han & Kamber, 2006), the main aim of data mining methods is to extract meaningful knowledge from data. Some authors like in (Bakhshinatogh et al. 2018) and (Fernandes 2019) affirms that, the Educational Data Mining (EDM) is a field of research that aims to apply data mining methods to educational data. It aims at developing methods and applying techniques for data mining, statistics and machine learning to analyze the data collected for each learner during his/her academic career (Niemi et al. 2018) (Anand 2019).

Five different approaches in EDM are proposed in (Baker 2010); namely, relationship mining, discovery within models, distillation of data for human judgment, clustering and prediction. For prediction, the goal is to predict the label or class of a data object. Several research projects for application of EDM predicting have been carried out, while the prediction of student performance is one of its main areas of application.

Indeed, prediction of student performance can help teachers identify students who need extra assistance and prevent students from dropping out before final exams. Many Research in this area has been conducted using different algorithms such as machine learning algorithm (Hussain et al. 2019). In this perspective, this study used machine learning algorithms to predict the individual performance of H.E.K high school students in Morocco. Our main objective is to identify at the end of the first semester the students enrolled in the Baccalaureate (Bac) who are at risk of failing the year before the final exams.

The study purpose are collecting student data from the Massar platform, annotating our data set and then analyzing it using machine learning algorithms to create a model able to predict student performance at the end of the first semester, to determine for each student which subjects need reinforcement and tutoring, and to predict and improve school success rates.

This paper is outlined as follows: section 2 presents a literature review of the educational data mining, machine learning and related works. Section 3 describes the proposed methodology, the description of the students sampled for this study and depicts the proposed framework as well as the capability of this work to make an accurate prediction of the students' performance. In section 4, we provide the discussion of the obtained results. Finally, the conclusion and the planned future work are presented in the last section.

## 2 Literature review

This part presents a general overview of EDM and machine learning and lists some research work in this context.

## 2.1 Educational data mining and machine learning

According to (Manjarres et al. 2018), Educational Data Mining is a rising discipline that aims to develop methods to explore data from educational contexts, in order to understand students' behavior, interests and results in a better way.

The EDM can be considered as the intersection of three main areas: education, statistics and informatics. This intersection among these three areas also generates other sub-fields, narrowly related to EDM, such as computer-based education, learning analysis, data mining (DM) and machine learning (ML). Fig. 1.

The coupling of machine learning with EDM has received much more attention in recent years. According to the literature of the field, machine learning is an axis of study of artificial intelligence. It brings together all the techniques that allow machines to learn and make accurate predictions from past observations. There are four main types of machine learning: supervised, unsupervised, semi-supervised and reinforcement.

The supervised learning is the most popular paradigm for machine learning, it aims to build a model based on observation data and desired results; This model allows the best approach to the relationship between input and output observable in the data. Supervised learning problems can be grouped into regression and classification problems. While, unsupervised learning uses only input observations data without having corresponding outputs. So, the goal of unsupervised learning is to determine the patterns or clusters hidden in the data from unlabeled data. Unsupervised learning problems can be seen as a problem of clustering or association. The semi-supervised learning falls between the two previous types of learning. Indeed, data labeling is a very expensive operation and requires presence of human experts. However, the availability of labels in some observations, even if they are missing in the most cases, gives to the semi-supervised algorithms the best chance for model construction. The semi-supervised learning algorithms exploit the idea that, even if group memberships of unlabeled data are unknown, this data has important information about the parameters

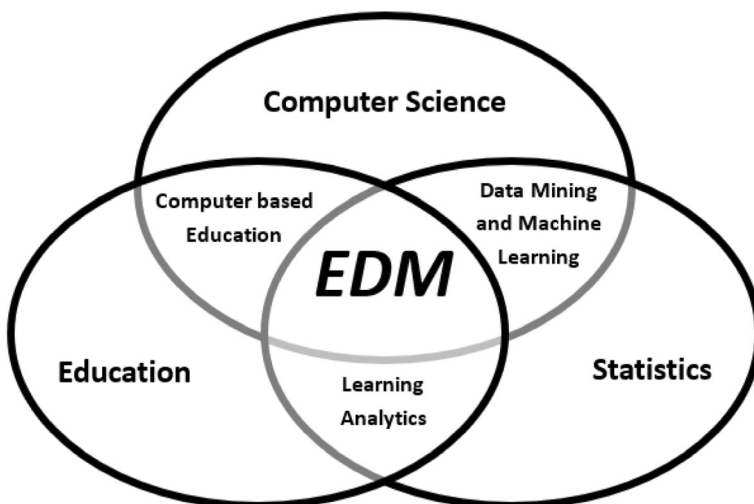


Fig. 1 Educational data mining (Romero and Ventura 2013)

of the group. Finally, the reinforcement learning is a relatively new approach to machine learning, and it is quite different from the previous types of learning. The Reinforcement algorithm works in an iterative fashion by using the observations collected from the interaction and taking actions to minimize the risk and maximize the benefits. The choice of machine learning algorithm strongly depends on the data used (continuous or categorical) and the outcome of the model (c.f.: cluster or numerical values, etc.). Among the most used machine learning algorithms, we quote: Linear Regression, Neural Networks, Decision Trees, Naive Bayes, Nearest Neighbor, Vector Support Machines (SVM), K-means, etc. We note that all these algorithms are available in the *skit-learn* python library. In this work, we used the multiple regression algorithm to predict students' performance as we will present later.

## 2.2 Related works

Many research studies have been conducted on predicting the performance or result of students based on certain parameters using different machine learning algorithms. Khan (Khan et al. 2019) has developed a model that allows students enrolled in an introductory programming course to identify their likely final grades using the Decision Tree algorithm. Iqbal (Iqbal et al. 2017) used a Restricted Boltzmann Machines (RBM) model to predict the grade of students who have been admitted to the ITU Electrical Engineering Department Degree Program. Elbadrawy (Elbadrawy et al. 2015) developed a personalized multiple linear regression model to predict student performance enrolled in Moodle at the University of Minnesota's. Using two different datasets, Pojon have used naïve Bayes classification and decision trees model to predict the academic student's success (Pojon 2017). (Hamsa et al. 2016) used a fuzzy genetic algorithm and decision tree model to predict academic performance of students based on their session exam scores. Applying the neural network algorithm, the author Puziah Mohd and al. have developed a model to predict students' final grades based on theirs first year's grades (Arsad et al. 2012).

Unlike other studies that focused mainly on data extracted from questionnaires and self report surveys, the predictions model used in this study is built on and tested on authentic data of students generated by the official Massar platform of the national ministry of education. In addition to this, the predictions model was tested on a large number of students and was validated by both statistical and methodological measures throughout three different school years.

## 3 Proposed methodology

The methodology used in this work is based on the Standard Interprofessional Standard Data Mining Process (CRISP-DM) (Chapman et al. 2000). The Cross Industry Standard Process for Data Mining (CRISP-DM) model is composed of six steps as shown in Fig. 2:

1. **Business Understanding:** The first phase of this model consists of understanding the Business in all the different aspects: the field of application, the project objectives, the requirements and the management rules: e.g. rules for calculating of the grade of Bac, the score of CA, etc.

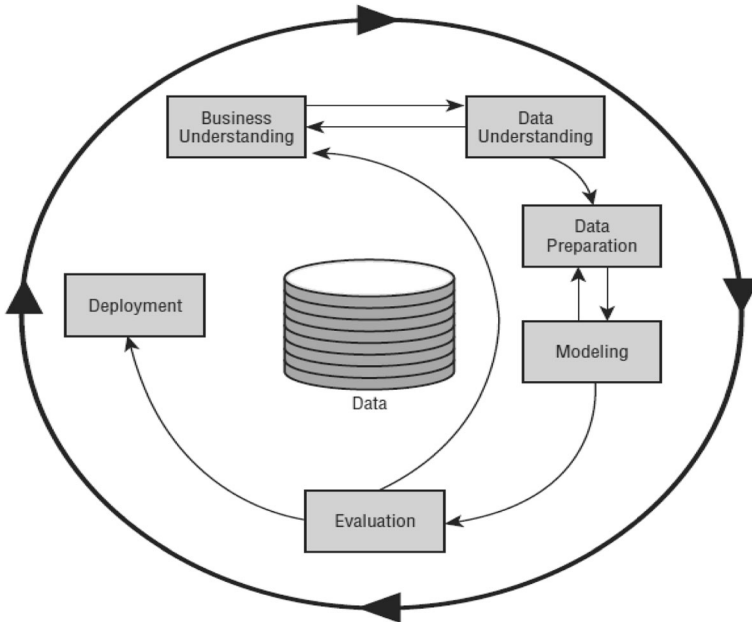


Fig. 2 CRISP-DM (Chapman et al. 2000)

2. **Data Understanding:** The second phase of CRISP DM Framework focuses on collecting the data, describing, exploring and manipulation data using techniques designed to acquaint users with the collected data.
3. **Data preparation:** This phase covers all activities required to build the final dataset, which is manipulated in the modeling phase.
4. **Modeling:** This step involves the implementation of different machine learning algorithms (regression, classification, clustering, recommendation). The ML algorithm to choose depends on the needs of the project, the input dataset and the output result.
5. **Evaluation of the Model:** Different techniques are available to evaluate the model. Among these techniques, we quote Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) and Relative Squared Error (RSE)., Accuracy, Sensitivity, etc. The choice of evaluation measures depends entirely on the project requirements, the algorithm used, the desired outcome, and so on.
6. **Deployment:** Finally, once the model is created, tested, and evaluated on the test and validation data, the deployment can be a report generation or an implementation of the data mining process.

### 3.1 Sample

The number of the student sampled for this study was 478. They all were second-year baccalaureate students studying at H.E.K high school in the school years 2015–2016,

2016–2017 and 2017–2018. Two hundred eighty-two of the students were females, and 196 were males. Fifty two of the students in the sample are repeaters, 25 were females, and 27 were males. All students in the sample are ranged in age from 16 to 19.

Two reasons determined the choice of H.E.K high school. First, the researcher has been a practitioner at the same school, and this has facilitated access to data. Second, the school did not meet the target baccalaureate success rate ( $\geq 40\%$ ) expected by the national ministry of education<sup>1</sup> in the 2016–2017 school year and therefore it incorporated a reinforcement program to boost the baccalaureate success rate. For this reason, predicting the success rate thus was important for the reinforcement program implementation.

### 3.2 Business and data understanding

The proposed framework is designed for Moroccan physics Bac classes. The Bac is a certificate issued by the state through the Ministry of Education in Morocco for different majors. This certificate is prepared in 2 years. At the end of the first year, students are required to pass a regional examination to assess their knowledge of secondary subjects (for Science e.g. French, Arabic, Islamic Education, History and Geography). Then, at the end of the second year, a national examination on the main subjects of the subject stream must be carried out.

During the Business and Data Understanding phase of the problem, we presented studies and analysis of student achievement at the end of each first semester of the 2016–2017, 2017–2018 and 2018–2019 school years. This work allowed us to better understand the problem and to discover the different calculation formula used.

The formula for calculating the Grade of the Bac (GB) in Morocco is as follows: CA: Continuous Assessment score (25%), R: Regional Examination Score (25%), NE: National Examination score (50%)

$$GB = \frac{(CS + RE + (NE \times 2))}{4}$$

The formula for calculating the CA is as follows:

$$CA = \frac{(CA1 + CA2)}{2}$$

CA1 and CA2 are successively the average of the continuous assessment of the 1st semester and the 2nd semester of the second year Bac.

The average of the continuous assessment (CA) for each semester is equal to the sum of the score of each subject multiplied by its coefficient, and the total is divided by the sum of the coefficients.

$$CA = \frac{\sum(\text{Score}_{\text{Subject}} \times \text{Coef}_{\text{Subject}})}{\sum \text{Coef}_{\text{Subject}}}$$

<sup>1</sup> Ministry memo N° 17/778 date June 22, 2017

As mentioned in the following formula, the National Examination Score (*NE*) is calculated by summing the score of the subjects examined multiplied by their coefficients, then the total is divided by the sum of the coefficients.

$$NE = \frac{\sum(\text{Score}_{\text{Subject}} \times \text{Coef}_{\text{Subject}})}{\sum \text{Coef}_{\text{Subject}}}$$

According to the Grade of the Bac, a merit is attributed to each student according to the following table.

Interval of the grade of the Bac	Merits ( <i>Mention</i> )
[10–12[	Passed, Average (A) – ( <i>Mention Passable</i> )
[12–14[	Passed, Quite Good (QG) – ( <i>Mention Assez bien</i> )
[14–16[	Passed, Good (G) – ( <i>Mention Bien</i> )
[16–20]	Passed, Very Good (VG) – ( <i>Mention Tres bien</i> )

### 3.3 Data preparation

The dataset used to design this framework was based on data of real-world students. The data concerns 478 students in Physics (P) stream for the 2015–2016, 2016–2017 and 2017–2018 school years. The data was collected from the SMS-MASSAR. We note that the data of students used in this study were kept anonymous and they were assigned codes.

Since the 2013–2014 school year, the Moroccan Ministry of National Education has set up the school management system “MASSAR”. Among the main objectives of this system, we note:

- Allowing students as well as their parents to keep track of their performance at school.
- Establishing transparency and equal opportunities for all students;
- Ensuring autonomy of schools’ management;
- To have a record of the school process of students from their first school enrollment until they obtain their Bac.
- etc.

The introduced variables are described in the following Table 1.

### 3.4 The proposed framework

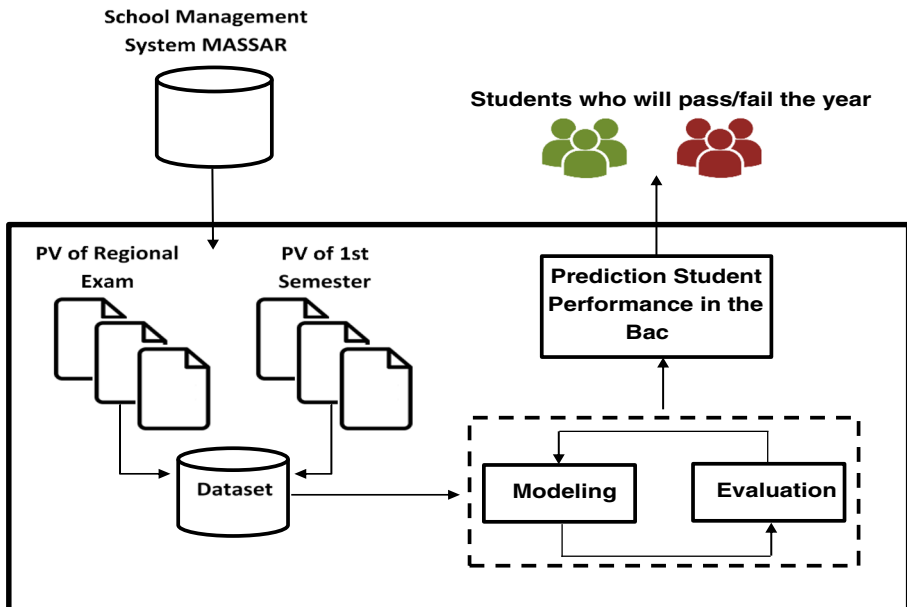
Figure 3 illustrates the proposed framework; it starts with the collection and preparation of the data. Student data is collected from the SMS-Massar. Then, the models are generated by the machine learning algorithm, the multiple regression. The modeling phase was developed in two stages: (1) model for forecasting the second semester results and (2) model for forecasting the results of the national exam. The results of these two models are used to predict the grade of the Bac for each student.

To choose the variables of the dataset to build our model, we used a correlation matrix between the variables. This allowed us to define, for each subject, the variable

**Table 1** Used variables

Column	Description	Type
Num	Student ID	Nominal
Gender	Student's gender	Nominal
RE	Score of the regional exam	Quantitative
Math_1	Math score of the first semester	Quantitative
PC_1	Chemistry & Physics score of the first semester	Quantitative
SVT_1	Science score of the first semester	Quantitative
Philo_1	Philosophy score of the first semester	Quantitative
Ang_1	English language score of the first semester	Quantitative
Trad_1	Translation score of the first semester	Quantitative
Ed_Ph_1	Physical education score of the first semester	Quantitative
Ed_Is_1	Islamic Education score of the first semester	Quantitative
Fr_1	French language score of the first semester	Quantitative
Arab_1	Arabic language score of the first semester	Quantitative
Disc_1	Attendance and behavior score of the first semester	Quantitative
CA1	Continuous Assessment score of the first semester	Quantitative

(subjects) with which the correlation is highest and to determine the relationship of influence between the different subjects. Figure 4 presents an example of the relationship of influence for 2nd semester Math (*Math\_2*) Physics (*PC\_2*) and national exam Math (*Math\_NE*).



**Fig. 3** The proposed framework



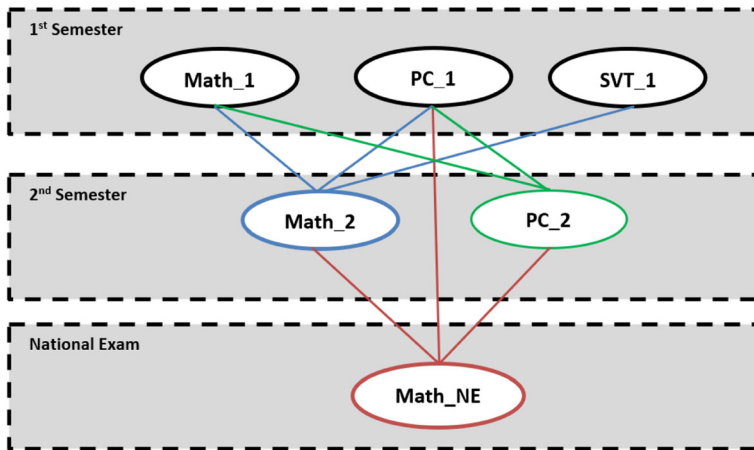


Fig. 4 An example of relationship of influence between subjects

We emphasize that this relationship helps to determine student performance for a given subject, as well as the topics (subjects) that influence it in order to determine the type of reinforcement and additional assistance that the student needs to improve her/his performance.

The performance of the model was validated on two levels. The first level was by a various performance indicator. MAE, RMSE, RAE and RSE were calculated, analyzed and interpreted. And the second level was by a pedagogical committee set by the school administration. The committee was headed by the school principal and included 11 teachers representing all the baccalaureate school subjects. The committee compared the predictions of the model to the real grades of students at the end of each school year.

## 4 Results and discussions

Several methods of measurement to evaluate the success of models exist. Nevertheless, the evaluation of each model is extremely dependent on the domain of study and the expected results of the system. For our system, the goal is to predict the grade of students' Bac and make decisions if a student needs reinforcement courses, supports and has to work hard to pass. These decisions are relevant if the predictions are accurate. We recall that our model is generated by the Multiple regression algorithm using the skit-learn python library. The model is composed of two sub-models: model of second semester (*Model S2*) and Model National Examination (*Model NE*). Model S2 predicts the results of the second semester while the National Examination model function is to predict the results of each student's national exam. Finally, using the results of model S2, the results of the National Examination model, we can predict the final grade of Bac of each student as mentioned in the formula of calculation of GB (formula GB). To this end, we compared the actual GB against the prediction GB for each student. Some of the most used metrics for evaluation of the models are the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE) and Relative Squared Error (RSE). The model's predictions were evaluated by

**Table 2** The used metrics for evaluation of the models

	Model S2	Model NE	Calcul of GB
MAE	0.28	0	0
RMSE	0.94	1.27	1.25
RAE	0.66	0.62	0.51
RSE	0.6	0.66	0.75

ten repetitions random cross-validation. In each run, we selected randomly 80% of students' data into the training set and 20% of students' data into the testing set.

The following Table 2 presents the models evaluation by comparing the real values with the predicted ones for each students.

#### 4.1 Results of the model of prediction of the 2nd semester (model S2)

The following graph Fig. 4 illustrates a comparison between the real values of the continuous assessments score of the second semester and the predicted ones. The model predicted that 10% of students would fail in the second semester and 90% would succeed with the following merits (P: 22%, QG: 30%, G: 21% and VG 30%); whereas, analysis of students results at the end of the semester showed that 17% of students did not really succeed in the second semester and 83% of students succeeded in the semester (P: 18%, QG: 27%, G: 25%, VG: 19%). (Fig. 5).

#### 4.2 Results of the model of prediction of the national exam (model NE)

Regarding the evaluation of the predictions of the national examination results, our system predicted that 52% of the students will fail in the national exam while 50% of the students have failed in the national exam with a negligible difference equal to 2%. (Fig. 6).

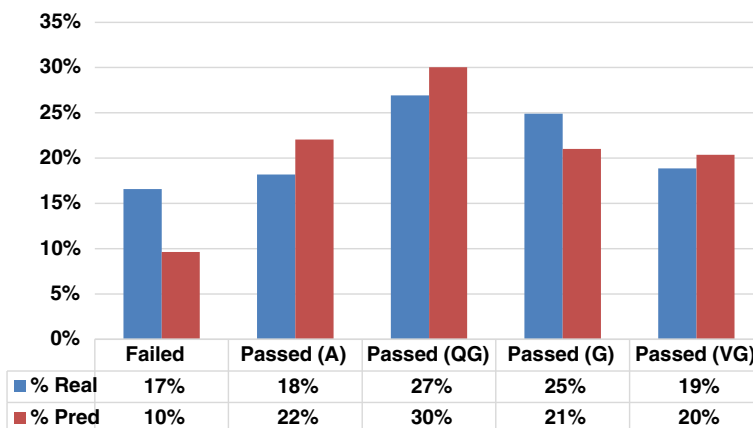


Fig. 5 Results of the model of prediction of the 2nd semester

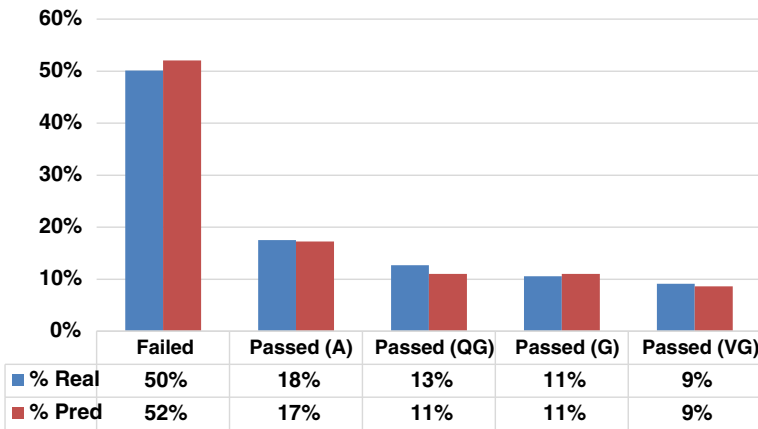


Fig. 6 Results of the model of prediction of the National Exam

### 4.3 Results of the prediction of the grade of the Bac

As we mentioned before, the prediction of the second semester and that of the national exam are used in addition to results of 1st semester and regional exam to calculate the grade of the bac and to attribute a merit to each student. As illustrated in the following graphic Fig. 7, the prediction of our system showed that 42% of students are at risk of failing the year with a 5% difference from to the real percentage (37%). Also, the system showed that 29% of students would pass with an Average merit compared to 21% of students who actually did.

For the merits Quite Good (QG), our model predicted the same percentage as the real one (14%) with 0% of the difference. 14% represents the percentage of predicted students who would have their Bac with good merits while 11% represents the percentage of students who have actually obtained this merit (a difference of 3%). For the Very Good (VG) merits, the real percentage was the same as the predicted one (8%) (Fig. 7).

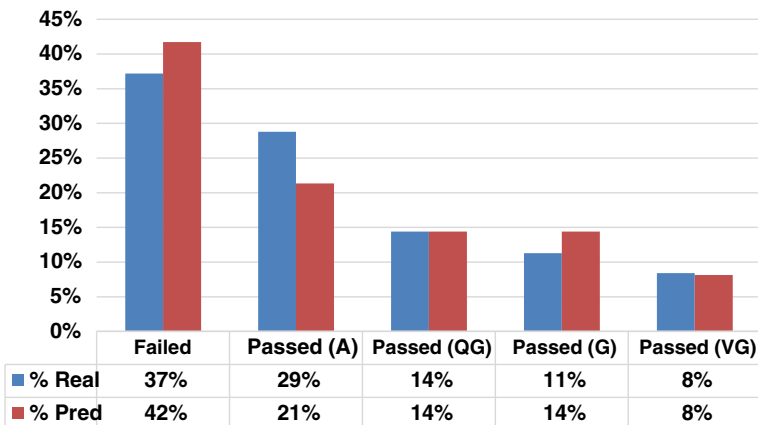


Fig. 7 Results of the prediction of the Grade of the Bac

## 5 Conclusion and perspective

Early prediction of the general Bac average is an effective way to determine student performance and make decisions if a student needs reinforcement courses, supports and has to work harder to be pass. To this end, we present a predictive analysis methodology of performance of students enrolled in Bac at H.E.K high school in Morocco based on the machine learning algorithm.

The proposed methodology is founded on the CRISP-DM and used a dataset obtained from the Scholar Management System MASSAR. We started by understanding the problem and the data. This phase is an important step and it lasted almost 2 years. In this phase, we presented studies and analyses of students results for the 2016–2017, 2017–2018 and 2018–2019 school years. Subsequently, we constructed our dataset from real data of students enrolled in Bac of the last 3 years. Concerning the modeling step, we used the multivariate Regression machine learning algorithm available in the library scikit-learn v0.20.3 of python. We have developed two models: *Model S2* and *Model NE* that cooperate. The first model predicts the results of the second semester, while the second model allows to predict the results of the national exam. To choose variables from dataset to build our model, we used a correlation matrix between the variables in the dataset. This allowed us to define for each subject the variables with which the correlation is highest and to determine the relationship of influence between the different subjects.

The predictions of the models (*Model S2* and *Model NE*), the grades of students in regional exam and the first semester grades are used to calculate the Grade of Bac. The results of this system have been evaluated in two levels. First level, by calculating the most frequently used metrics such as MAE, RMSE, RAE and RSE and the second level, by a pedagogical committee headed by the school principal. The results showed that our system can make more accurate predictions of student performance.

In future research, we intend to test our model on other school majors, apply another machine learning algorithm of regression and classification and evaluate our model at the provincial, regional and national level. In the long term, we aim that our model will be incorporated into the School Management System -MASSAR to allow all schools to predict the results of their students and plan actions to improve them.

**Acknowledgements** I would like to express my sincere gratitude to the H.E.K school director (Morocco), for his collaboration, particularly in the collection of data, as well as for his suggestions and encouragements made to the development of this model.

## References

- Anand, M. (2019). Advances in EDM: A state of the art. *Software Engineering Springer*, 193–201.
- Arsad, P. M., Buniyamin, N., & Ab Manan, J.-L. (2012). Neural network model to predict electrical students' academic performance. In *4th International Congress on Engineering Education (ICEED)*.
- Baker, R.S.J.D. Data mining for education. *International encyclopedia of education* McGaw, B., Peterson, P., Baker, E. (Eds.) 7, 3 (2010).
- Bakhshinategh, B., Zaiane, O. R., ElAtia, S., & Ipperciel, D. (2018). Educational data mining applications and tasks: A survey of the last 10 years. *Education and Information Technologies*, 23(1), 537–553.

- Chapman, P., Clinton, J., Kerber, R., Khabza, T., Reinartz, T., Shearer, C., & Wilrth, R. (2000). *CRISP-DM 1.0 step-by-step data mining guide*. The CRISP-DM consortium.
- Elbadrawy, A., Studham, S., Karypis, G. (2015). Personalized multi-regression models for predicting students performance in course activities. Paper presented at the 5th international conference on learning analytics and knowledge, March 2015: 16–20.
- 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. <https://doi.org/10.1016/j.jbusres.2018.02.012>.
- 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> (<http://www.sciencedirect.com/science/article/pii/S2212017316304613>)
- Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques*, 2nd ed. Morgan Kaufman.
- Hussain, M., Zhu, W., Zhang, W., Abidi, R., & Ali, S. (2019). Using machine learning to predict student difficulties from learning session data. *Artificial Intelligence Review*, 52(1), 381–407.
- Iqbal, Z., Qadir, J., Mian, A.N., Kamiran, F. (2017). Machine learning based student grade prediction: A case study. *arXiv preprint arXiv:1708.08744*.
- Khan, I., Al Sadiri, A., Ahmad, A. R., & Jabeur, N. (2019). Tracking student performance in introductory programming by means of machine learning. In *4th MEC International Conference on Big Data and Smart City (ICBDSC)* (pp. 1–6).
- Manjarres, A.V., Sandovalm, L.G. M., Suárez, M.S. (2018). Data mining techniques applied in educational environments: Literature review. *Digital Education*, n° 33.
- Niemi, D., Pea, R.D., Saxberg, B., Clark, R.E. (2018). *Learning analytics in education*. Information Age Pub Inc (August 1, 2018).
- Pojon, M. (2017). Using machine learning to predict student performance. MS thesis, University of Tampere, Faculty of Natural Sciences, 35.
- Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12–27.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.