PREDICTING FACTORS AFFECTING STUDENT'S PERFORMANCE IN A LEARNING MANAGEMENT SYSTEM

Majjate Hajar

LISAC Laboratory,Faculty Of Science Dhar El Mahraz,Department of Computer science, Sidi Mohamed Ben Abdellah University,Fez 30030,Morocco

E-mail : hajar.majjate@gmail.com

Jeghal Adil

LISAC Laboratory, National School of Applied Sciences Fes (ENSA), Sidi Mohamed Ben Abdellah University,Fez 30030, Morocco.
E-mail : adil.jeghal@usmba.ac.ma

Yahyaoui Ali

LISAC Laboratory, Faculty Of Science Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, Fez 30030,Morocco
E-mail : ali.yahyaouy@usmba.ac.ma

Abstract

Throughout the last years, the Online Learning Platforms have experienced a great success due to their power to offer accessible content and innovative Learning Strategies who Respect the rhythm of learning of each learners. These educational environments handle a large amount of information, which generates a large amount of data to be explored, in order to increase the performance of the various educational actors. This article discusses the evolution of Educational Data Mining and reviews different exploration tools and powerful prediction algorithms that will help in a pertinent Data visualisation and Analysis in the way to Predict and classify the most important Factors Affecting student's performance and success in a Learning management system.

Keywords: **Logistic Regression's; XGBoost Algorithm; Educational Data Mining; E-learning; Student's Performance.**

1. Introduction

The evolution of the internet has influenced all areas of life, it has marked new generations of people, their lifestyles and their learning methods, and the internet has conquered the world of education and training by imposing new learning technologies, technologies that offer affordable, quality education for all knowledge seekers even in the most remote places on the planet. In this context, many organizations around the world have started to use online training platforms, for the creation of virtual classes, design of educational content, the sharing of educational resources and the monitoring of distance learners, such as they are in a real classroom. It is undeniable that the digitization of education offers flexible content and accessible knowledge for all educational actors. However, it also has some drawbacks, such as lack of interaction with teachers, lack of commitment, lack of parental guidance. What are the factors behind this lack of interactions among learners? What are the forms of learner engagement in a distance-learning platform? Does parental guidance really affect learner outcomes? Can we predict other factors influencing student performance in a distance-learning platform?

In this work, we will therefore attempt to analyse the main factors that influence student success in a Learning management system, by exploring a suitable data set containing a variety of information that deserve to be carefully studied.

2. Related Work

The emergence of information and communication technologies has changed not only people's daily lives but also their way of learning; students today are giving an increasingly important place to the Internet and to digital media, which has prompted educators and computer scientists to invent new learning methods that pass through the Internet such as distance training platforms (MOOCS, LMS Platforms). These Platforms now have a reservoir of data flooded with a wide variety of information (images, videos, sounds, etc.), requiring researchers to invent tools capable of managing, analysing and visualizing clearly and efficiently. These large masses of data for the

discovery of knowledge and improvement of the education system based on these data. This gave birth to a new discipline called "Educational Data Mining".

Data Mining Or Educational Data Mining is a new field of research developed using the methods and tools of data Mining in smart education that create new opportunities to collect, analyse, visualize and present student data. This data can be a crucial tool for making the right decision in the direction of improving the quality of training, the performance of learners, as well as for modifying assessment criteria and training outcomes, including forecasting and monitoring students' academic progress.

We focused in the present research. On building a machine-learning model with a high accuracy, using the gradient boosting decision tree algorithm (XGBoost), to study the most important features affecting student success on an learning management system. Under this context, various researches have been conducted exploring Data mining techniques on the way to study and evaluate student's performance:

- [Khanna *et al* (2017)] Conducted a Systematic Review exposing various techniques and tools available in educational data mining and explaining their Roles and their importance in Determining Factors Affecting Students Academic Performance and teaching-Learning process.
- [Surjeet *et al (2012)*] conducted a Related research experimenting a multiple decision tree classifiers on an educational dataset to classify the educational performance of students. They found out that the CART (Classification and Regression Tree) decision tree classification method is the best classifier to predict the student's performance, this selection was based on the produced accuracy and precision using 10-fold cross validations.
- [Osmanbegović *et al (2015)*] Conducted a related research exposing the data mining possibilities using WEKA Software containing a collection of classification algorithms, the Results indicate that All algorithms obtained a good accuracy butt the best results are obtained by J48 decision trees classifier, when using only nine attributes and Random Forest when using all the attributes. With an accuracy of 74%.

In our way to properly study and analyse the role of educational data mining, a relevant educational database has been selected to be studied which contains a variety of information on students from an LMS platform.

3. Methodology and Analysis

3.1 *Source of Data*

we was looking for an educational dataset suitable and rich in information about students, such the gender, the Nationality, the place of birth, containing a large number of learners with a multitude details, such how many times the student visit a course and how many times the student participate on a discussion also the number of absence days. Therefore, we founded an appropriate dataset that will help us to build a good model to predict the most important Factors that affect student's performance.

The data is collecting from a learning management system (LMS), called Kalboard 360, endowed with a Learning Record tool called experience API "xAPI" that record student's actions and make the learning activity something trackable and sharable.

3.2 *Population of the Study*

The dataset include 480 student from 13 different nationalities. Mentioning 179 students from Kuwait, 172 students from Jordan, 28 students from Palestine, 22 students from Iraq, 17 students from Lebanon, 12 students from Tunis, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from USA, Iran and Libya, 4 students from Morocco and 1 student from Venezuela.

3.3 *Data description*

Columns/ Variable name	Type
gender	Categorical
NationalITy	Categorical
PlaceofBirth	Categorical
StageID	Categorical
GradeID	Categorical
SectionID	Categorical
Topic	Categorical
Semester	Categorical
Relation	Categorical
raisedhands	Numeric
VisITedResources	Numeric
AnnouncementsView	Numeric
Discussion	Numeric
ParentAnsweringSurvey	Boolean

Table 1. Description of the dataset

Table 1 (Continued)

ParentschoolSatisfaction	Categorical
StudentAbsenceDays	Categorical
Class	Categorical

Table 2. Description of the dataset

This dataset includes 480 student's records, and 16 columns describing student's reactions during their learning processes such as raised hand on class and visited Ressources and other features describing their personal informations such their gender and nationality,we found also some features describing their Academic background such as educational stage, grade Level and section.

The students are classified into three numerical intervals based on their total grade:

(1) High-Level: 142 students in the dataset Belong to the High-level interval which includes values from 90-100.
(2) Middle-Level: 211 students in the dataset Belong to the Middle-level interval, which includes values from 70 to 89.
(3) Low-Level: 127 students in the dataset Belong to the Low-level interval, which includes values from zero to 69.

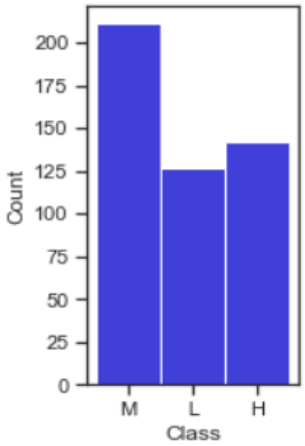

Fig. 1. Description of the Classes

3.4 *Data visualization and Analysis*

3.4.1. Data visualizations

In this section, we will Identify The Most Informative Features Using pandas's plot.

- ***Gender :***

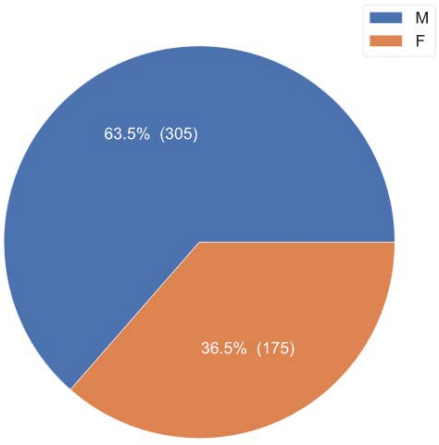

Fig. 2. Sum of each Gender

The total number of students are 480 including 305 Males and 175 females.

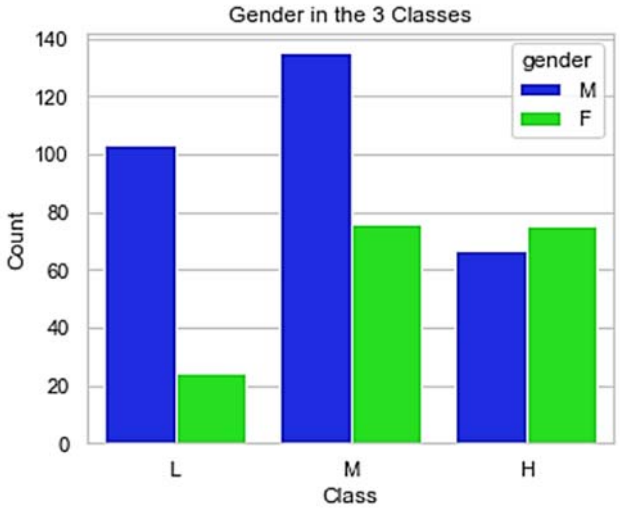

Fig. 2. Gender repartition

The numbers of girl are higher than boys in the high-level class; it seems like women performed better than men.

- ***Nationalities :***

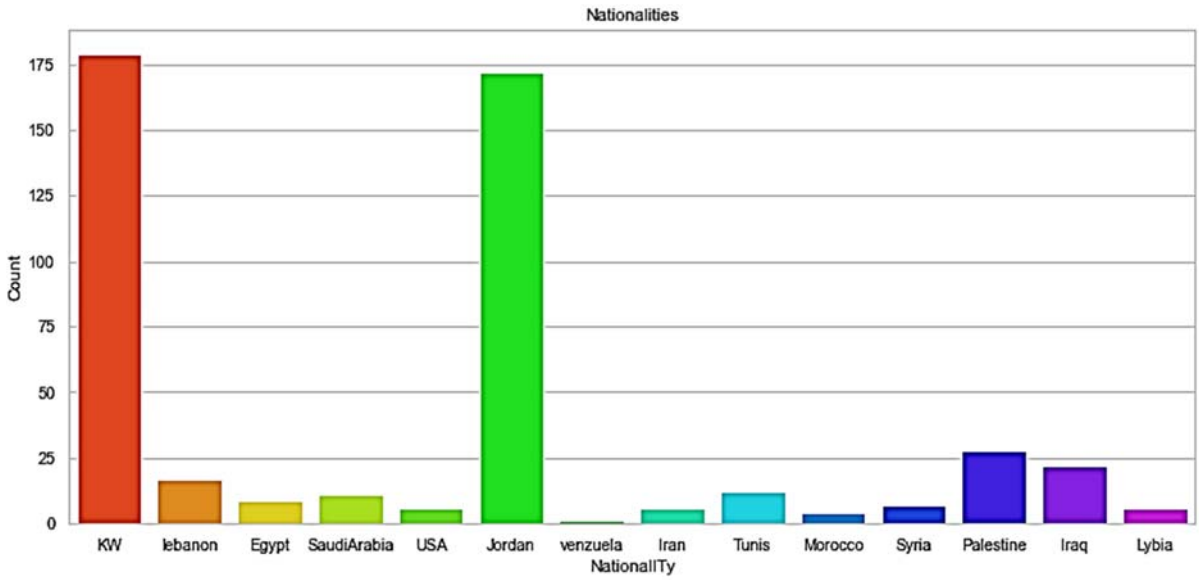

Fig. 2. Nationalities

We can see 13 different nationalities, but the majority numbers of student are from Kuwait and Jordan.

3.4.2. Exploratory Analysis

- ***Students Behaviors***

Dealing with the subject of understanding the Classification of students in the tree Classes, we made a swarm plot with the seaborn library[a] including all the factors we need to analyse:

[a] Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn, which helps in exploration, and understanding of data. [Katari (2020)]

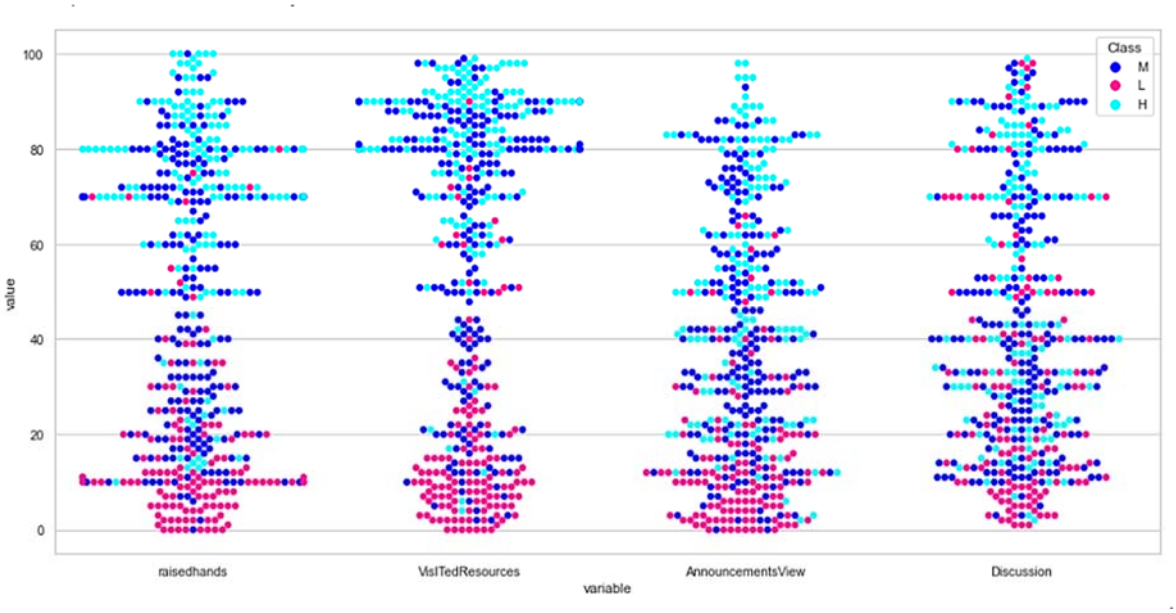

Fig. 3. Student Behaviors swarm plot

As expected, we can notice that students who have higher values of raising hands, visited resources, viewing announcements and actively participate in-group discussions are usually classified in High-Level Class.

Nevertheless, we can see some students classified in Low-level Class although they actively participate in-group discussions and have a higher value of raising hands, visiting resources and viewing announcements. So are there any other factors that appear to have influenced the student scoring? We will study the others given features to respond to this question.

- ***Student engagement***

In the next step, we will try to understand the results from the involvement of student in the course using pivot table function from Pandas library.

Class	Absence rate
H	0.028169
L	0.913386
M	0.336493

Table 2. Student Absence day frequency

$$table0 = pd.pivot\_table(data, index = ['Class'], aggfunc = \{'StudentAbsenceDays': np.mean\}) \quad (1)$$

$$table0.plot()$$

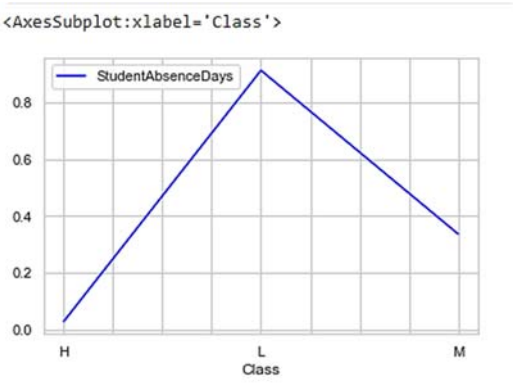

Fig. 4 Student Absence days plot

We notice that the students who are classified in the Low-level have a higher rate of absence, on the other hand the students who are classified in the high-level have a low rate of absence.

Therefore, the commitment and the presence in the sessions of the course is a very important and determining factor in the success of the student.

- ***Parent's involvement***

We used the same function pivot_table as seen in the previous section to present clearly features having relation with parent involvement.

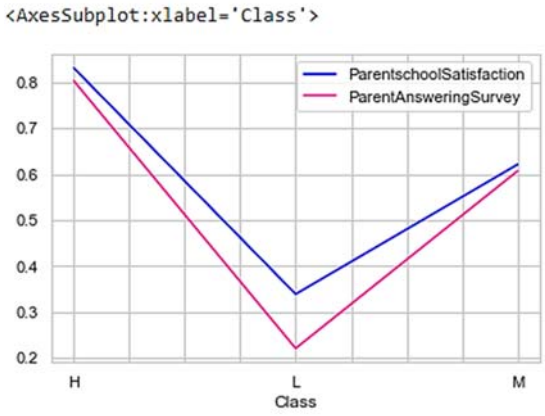

Fig. 5 Parents involvement plot

As we can see, the parents who are involved with their children in the course and follow up their activities, their children are classified in the high level, so the engagement of the parents is an important factor in academic success.

3.5 *Spotting Most Important Features*

After creating various visualizations and exploring of the dataset, we will work to reveal the most Important Features from 7 features in dataset.

3.5.1. *Label Selection*

After creating a new dataframe called df that include 7 chosen features copied from the original dataframe, we select our categorical variables:

$$obj\_df = data.select\_dtypes(include = ['object']).copy() \quad (2)$$
$$obj\_df.head()$$

Class	StudentAbsenceDays	ParentAnsweringSurvey	ParentschoolSatisfaction
M	Under-7	Yes	Good
M	Under-7	Yes	Good
L	Above-7	No	Bad
L	Above-7	No	Bad

Table 3. Selection of categorical variables

3.5.2. *Label Encoding*

In label encoding in Python, we replace the categorical value with a numeric value between 0 and the number of classes minus 1. If the categorical variable value contains 5 distinct classes, we use (0, 1, 2, 3, and 4). [Great Learning Team (2020)]

Next, we encode our 4 categorical variables with the next code :

$$Features = data.drop('Categoricalvariable', axis = 1)$$
$$Target = data['Categoricalvariable']$$
$$label = LabelEncoder()$$
$$Cat\_Colums = Features.dtypes.pipe(lambda\ Features: Features[Features == 'object']).index \quad (3)$$
$$for\ col\ in\ Cat\_Colums:$$
$$Features[col] = label.fit\_transform(Features[col])$$

3.5.3. *Splitting the dataset*

In the next step, we start to split the dataset into a training set and a testing set :

$$X\_train, X\_test, y\_train, y\_test = train\_test\_split(Features, Target, test\_size = 0.2, random\_state = 123) \quad (4)$$

3.5.4. *Implementing a Logistic Regression model*

We will implement our model and check the accuracy and performance of our model:

$$\begin{gathered} Log\_Model = LogisticRegression(solver =' lbfgs', max\_iter = 1000) \\ Log\_Model.fit(X\_train, y\_train) \\ Prediction = Log\_Model.predict(X\_test) \\ print("the\ accuracy\ is" + str(accuracy\_score(y\_test, Prediction))) \\ print(classification\_report(y\_test, Prediction)) \end{gathered} \quad (5)$$

```
the accuracy is 0.8229166666666666

              precision    recall  f1-score   support

           H       0.70      0.81      0.75        26
           L       0.93      0.97      0.95        29
           M       0.83      0.73      0.78        41

    accuracy                           0.82        96
   macro avg       0.82      0.83      0.83        96
weighted avg       0.83      0.82      0.82        96
```

Fig. 6 Logistic Regression model

We obtained an accuracy of 0.8229166666.

3.5.5. *Feature Importance with XGBoost*

A benefit of using ensembles of decision tree methods like gradient boosting is that they can automatically provide estimates of feature importance from a trained predictive model. [Brownlee (2020)]

So we will work to improve our Gradient Boosting model, by using the plot_importance() method to classify and represent the most important features from our selective features.

$$\begin{gathered} \boldsymbol{xgb = XGBClassifier(max\_depth = 10, learning\_rate = 0.1, n\_estimators = 100, seed = 10)} \\ \boldsymbol{xgb\_pred = xgb.fit(X\_train, y\_train).predict(X\_test)} \\ \boldsymbol{print\ (classification\_report(y\_test, xgb\_pred))} \end{gathered} \quad (6)$$

```
              precision    recall  f1-score   support

           H       0.75      0.81      0.78        26
           L       0.93      0.97      0.95        29
           M       0.84      0.78      0.81        41

    accuracy                           0.84        96
   macro avg       0.84      0.85      0.85        96
weighted avg       0.84      0.84      0.84        96
```

Fig. 7 Calculing the Best Accuracy

We can notice that using a learning_rate of 0.1, a max_depth of 10 and 100 estimators in our XGB classifier provides an accuracy of 0.84375.

we use the feature importance plot :

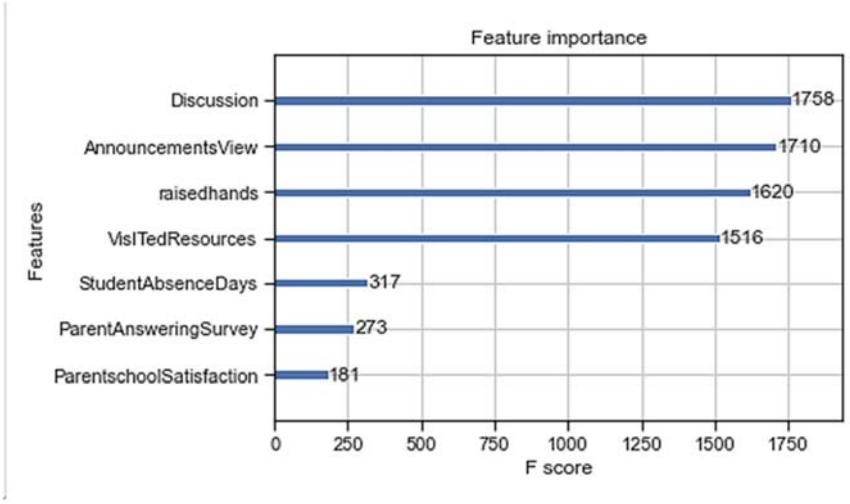

Fig. 5 plot of feature importance

the most two important features which represent the XGBoost plot is the number of times the student participate on discussion group And the number of viewing announcements. Moreover, this is what we saw in the first swarm plot, that there was a big difference between the L, M and H classes in the score of student's participation and viewing announcements.

4. Results and Discussion :

The presented research focused in data analytics and interpreting a performant machine-learning model using the gradient boosting decision tree algorithm (XGBoost), to represent the importance of a selective features well chosen from our dataset in an attractively simple bar-chart.

The results of this research concluded some points:

- Student who actively participate on group discussion in the platform are most likely to have a higher classification and good performance. So we can says that exchanging information, skills or ideas with other learners is the key of success of any student.
- Students who view and read announcements, have taken too high-level values mostly. Therefore, we can says that reading the recent information and following the recommendations, make the student always connected with every new ideas.
- The factors of the raising hands and visiting resources are too a required factors that influence student's Performance.
- Actions related to the absence days are less likely to improve student's performance. Because it is could be related to a poor Internet connection, or the student chose to Download courses and work without connection.
- The parents who are involved with their children in the course and follow up their activities, their children are have to a good classification but it is not a determinant factors in student success.

5. Conclusion

Educational Data Mining is vital discipline, equipped with a variety of tools their most important goal is improving Online Courses experience, Developing the performance of student and make the learning process as much attractive and productive, because the online learning it is an area which imposes its importance in our days especially in these conditions that live the world.

Decision making in the elaboration of Online Courses platform is a complex job but the variety of decision-making tools existing in Educational Data Mining with help to analyse student behaviour in an E-learning platform or predict the most important factors that affect student performance, or predict the student success. All that became something possible and easy to be developed with the variety and the efficiency the Educational Data Mining tools.

References

[1]. 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.

[2]. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

[3]. Brownlee, J. (2020, August 27). Feature Importance and Feature Selection With XGBoost in Python. Retrieved from machinelearningmastery: www.machinelearningmastery.com.

[4]. Brownlee, J. (27. August 2020). Feature Importance and Feature Selection With XGBoost in Python. Von machinelearningmastery: www.machinelearningmastery.com abgerufen
[5]. Great Learning Team. (22. August 2020). Label Encoding in Python Explained. Von mygreatlearning: https://www.mygreatlearning.com/blog/label-encoding-in-python/ abgerufen
[6]. Katari, K. (11. August 2020). Seaborn: Python. Von towards data science: https://towardsdatascience.com/seaborn-python-8563c3d0ad41 abgerufen
[7]. Khanna, L., Singh, S. N., & Alam, M. (13. July 2017). Educational data mining and its role in determining factors affecting students academic performance: A systematic review. 2016 1st India International Conference on Information Processing (IICIP).
[8]. Osmanbegović, E., Suljic, M., & Agić, H. (February 2015). Determing dominant factors for students performance prediction by using data mining classification algorithms.
[9]. Surjeet, Kumar, & Yadav. (Décembre 2012). A Comparative Study for Predicting Student's Performance. International Journal of Innovative Technology & Creative Engineering .

Authors Profile

Hajar Majjate**,** currently works at the Regional Center for Careers Education and Training (CRMEF) of Morroco, as a teacher of computer science,she is also a Ph.D student in the faculty of science of Dhar elmahraz, Sidi Mohamed Ben Abdellah University of morocco.she received the "Innovative Teachers" award for the 2018/2019 school year.

Adil Jeghal, presently working as an Associate Professor of computer sciences in the National School of Applied Sciences ,Sidi Mohamed Ben Abdellah University of Morocco. His research interest includes distance education, interactive learning system, educational technology, adaptive learning and access control.

Ali yahyaouy, presently working as a Professor of computer science in Sidi Mohamed Ben Abdellah University, Faculty of Sciences: Fez,Morocco.he is also a member of the scientific community in the Faculty of Sciences: Fez,Morocco.