Predicting Student Success Using Big Data and Machine Learning Algorithms

https://doi.org/10.3991/ijet.v17i12.30259

Farouk Ouatik[1](✉), Mohammed Erritali[1], Fahd Ouatik[2], Mostafa Jourhmane[1]
[1] Sultan Moulay Slimane University, Beni Mellal, Morocco
[2] Cady Ayyad University, Marrakech, Morocco
farouk.ouatike@gmail.com

Abstract—The prediction of student performance, allows teachers to track student results to react and make decisions that affect their learning and performance, given the importance of monitoring students to fight against academic failure. We realized a system of the prediction of academic success and failure of the students, which is the overall result and the goal of the educational system. We used the personal information of the students, the academic evaluation, the activities of the students in VLE, Psychological, the student environment, and we added practical work and homework, mini projects, and the number of student absences which gives a vision of the quality of the student. Then we applied the methods of artificial intelligence and educational Data mining such as KNN, C4.5 and SVM for the prediction of the academic success of students, but these methods are not sufficient given the progressive number of students, specialties, learning methods and the diversity of data sources as well as student data processing time. To solve this problem, Big Data technology was used to distribute the processing in order to minimize the execution time without losing the efficiency of the algorithms used. In this system we cleaned the data and then applied the property selection algorithms to find the useful properties in order to improve the algorithm prediction rate and also to reduce the execution time. Finally, we stored the data in HDFS and we applied the classification algorithms for the prediction of student success using MAPREDUCE. We compared the results before and after the use of big data and we found that the results after the use of Big Data are very good at execution time and we arrived at a recognition rate of 87.32% by the SVM algorithm.

Keywords—MapReduce, HDFS, SVM, KNN, C4.5

1 Literature review

In the literature we find several studies which concern the prediction of student performance which is linked to several factors [1]. The authors in [2] use neural networks, the decision tree and SVM, to make the prediction of academic performance of students from the behavior of students to the internet. They found that behavior of students to the internet is important for the prediction of student success, for example, academic performance is highly correlated with the frequency of internet connection.

On the other hand the volume of Internet traffic is negatively correlated. Other properties are used by [3] for the predictive of the performance of the students, they are exploited the data of the registration form, then they created a prediction model based on neural networks for classification. There is other research on the same topics [4] [5].

In our article we focused on predicting student success, using their performance. The research that exists to build a model for predicting student success is divided into two axis, the first axis concerns the research of the factors that influence the prediction of the academic success of students and the second axis concerns the data mining methods that are used for the construction of the model for predicting the success of the students and also for the verification of the models built.

Researchers use several factors that influence the prediction of students' academic success and also different data mining and machine learning methods.

Recently [6] used students' GPA for the first, second and the third year as features to predict student class of grade in the final CGPA. Then they tried to compare several classification algorithms to take the decision, the Probabilistic Neural Network, Decision Tree, Random Forest, Naïve Bayes, Logistic Regression and Tree Ensemble. In the meanwhile, [7] use also as features, student family background information, previous academic records and students' demographics information. In order to get the best model of student academic performance prediction, they applied Rule Based classification techniques, Naïve Bayes and Decision Tree. The maximum accuracy that they arrived was 71.3%, by using Rule Based classification techniques. Another work [8] tried to predict students' final GPA using their grades in previous courses, major, semester of graduation, campus and nationality. The classification algorithm used was J48 decision tree. Another model created by [9], compares five classifiers, ID3, j48, Naïve Bayes, Neural Network and Bayesian Network. But it used another features, Activity, Student Attendance, Midterms, Seminar, Lab Experiment, Office Automation, Final Examination, Project, Workshop and Previous Semester Final Mark. Based on accuracy, Bayesian Network has the highest accuracy.in the same way the authors in [10] adopted support vector machine and Logistic regression to select students who had the capability to succeed based only on prior academic achievement of student. The results show that the SVM model had the better prediction rate. And for the same purpose ,the authors in [11] built two model to predict student success , the first model based on naïve Bayes and the second model based on Bayes network , but this last, used students' questionnaire answers as features and department, Stage , Age, Gender, Status, Address, Work, Parent-Alive, Live-With-Parent, Father-Work, number of fail Courses, Mother-Work, Absence-Days Number, GPA, Credits Number, completed Credits Number, Years-Of-Study Number, Write-Notes During lectures, quiet During Exam and Prep-Study-Schedule . They found that the best model was the model based on naïve Bayes algorithm. There are others researches that concern the prediction of student success, in [12] the authors tried to predict student GPA score by using four classifiers, ID3, K-nearest neighbors, C4.5 and Naïve Bayes. They used as factors of student success, student's gender, previous student educational background and province of student's high school. The results revealed that Naïve Bayes gave the best accuracy (43.18%). More recently, [13] Construct a classifier using three random forests version to Predict student academic Success, they used the first two semesters of courses that are completed by a student, the data used are as follow: the student ID, the course's department, the course title, the semester, the grade and the credit value of the course but, the accuracy was low. The results of [14], show that course program and course block, course program and ethnicity are the best factors that separate successful student and unsuccessful student. They use CART algorithm and CHAID tree algorithm. The obtained results are as follow, the CHAID tree accuracy was 59.4% and the CART tree accuracy was 60.5%.

The number of features and factors that influence student academic success are in progress and also the number of students is increasing, what makes this process very difficult, because of the inability of these tools to analyze data and make decisions in a short time and more effectively. That is why we have created a system to predict the success of students based on the tools provided by big data technology that make predicting the success of the student more effective and in the shortest time by relying on a large set of data related to the student, which provides us with a greater ability to predict the success of the student and also gain a lot of time by recording information in a distributed way and analyzing it in a parallel way.

2 Introduction

Data mining is an analytical tool which contains a complex and also sophisticated set of methods and algorithms which are used to extract useful information from the analyzed data. It is used almost in all fields, Health, Marketing, Astronomy, Insurance, economy, finance, Human Resources, Pharmaceuticals, industry and in recent years, it is applied in education, it took the name of Educational Data Mining by what it is applied to data educational. It made a quantum leap and revolution in the field of education.

EDM methods allow prediction, clustering, relationship extraction, and model discovery and data visualization. It is used for distinct purposes, to assess learners, develop learner models and detect learner disengagement.

In this article we have focused on the academic success of the students. Student Academic success is an important indicator to measure the quality of education and the etablissement success. Successed student means that the student has completed their program and validates every semester. Student academic success is defined as a group of metrics that measure engagement, course completion and also learning. There are also several definitions of student success, which are presented in [30], they met that Student success is related to engagement in school activity, skills, academic achievement, skills, satisfaction, acquisition of knowledge and persistence. But to measure student academic success, we use either the Grade Point Average GPA or Cumulative Grade Point Average CGPA, which expresses and measures the academic performance of the student. In this article we will present a method for predicting the success of university students, because the student faces many changes, both in teaching methods and in evaluation methods. While this student needs help to be successful.

This article is divided into five parts, the first part contains the attribute selection methods, the second part concerns big data technology, the third part presents the classification algorithms used, the fourth part presents the model created and the fifth part contains a description of the data used and the results obtained, finally the conclusion.

There are several factors that influence the academic success of students. We have grouped them into five groups, the personal information of the students, the academic evaluation, the activities of the students in VLE, Psychological, the environment of the students. And then we applied the methods of selection of the properties in order to choose properties useful for predicting student success.

3 Feature selection

In order to increase the recognition rate and also to improve the classification of the data, we applied the feature selection [15] methods to select the significant properties that carry useful information for our prediction. These methods reduce the number of properties by removing properties which do not have a great effect on the results of the prediction. With the reduction of properties, the size of the data will also be reduced, which minimizes the execution time.

There are three types of feature selection methods. These methods differ from the point of view of their interactions with the classification algorithm. We distinguish the filtering method which removes redundant features and also searches for irrelevant properties using a metric statistic, and then calculate a score for each characteristic and the ordinates according to this score. The wrapper methods search for a subset in the property space. These methods are based on classification algorithms to select the best subset of properties. Finally, the embedded methods, which combine the two previous types.the selection of the sub-set of properties is done during the learning phase and also use a classification algorithm without the validation phase.

In this paper we compared three variable selection methods, MRMR, and the classification algorithms, j48 and SMO.

3.1 The minimal Redundancy Maximal Relevance algorithm

The minimal Redundancy Maximal Relevance algorithm [16] is a filtering algorithm based on the computation of correlation and mutual information to minimize redundance between features and also maximize relevance.

$$R(m)=\frac{1}{|P|^2}\sum_{m,n\epsilon P} I(m,n) \quad (1)$$

$$P(i)=\frac{1}{|P|^2}\sum_{m,n\epsilon P} I(m,X)$$

|P| : feature size

I(m, n) : mutual information between the same characteristic and the nth characteristic.

I(m, X) : the mutual information between the same characteristic and the labels of class X.

To calculate the score of a characteristic we use this formula:

$$\text{Score (j) =Relevance (j) - Redondancy (j).} \tag{2}$$

The second method is a Feature selection method based on j48 algorithm, that is a method for feature selection used j48 algorithm to validate the subset of properties.

The third method is a Feature selection method based on SMO algorithm, that is a method for feature selection used a new version of SVM algorithm to validate the subset of properties.Considering the progressive size of student data and the need for real-time processing, the methods and also the storage of this data, the basic methods of storage and processing are not sufficient. Pushed to use big data technology, to make storage and processing distributed.

Most of the styles are intuitive. However, we invite you to read carefully the brief description below.

4 Big data

Big Data [17] refers to large data that requires complex processing and which also poses a problem in terms of storage. Faced with the problem of massive data storage and processing, new technologies are needed. Hadoop framework makes it possible to respond to these constraints. Because it does the calculations in a way distributed over the nodes of the cluster is also parallel. The strong point of this platform is that it is open source, it offers a high level of data availability and durability because, it does not depend on the hardware, the data is automatically copied to different nodes of the cluster. Hadoop tolerant to errors and also to failures. And also, it is flexible. It allows you to store unstructured data of all kinds such as symbols, texts, videos or images. Unlike the relational database, data can be stored as is. Another advantage of hadoop is elasticity. It can change the number of nodes in a cluster either to reduce or expand the system.

Hadoop contains several modules and technology namely HDFS, Mapreduce, Yarn, Pig, Phoenix, HBase, Zookeeper, Impala, Hama, Hawq, Spark, Hive, ElasticSearch, Lucene, Sqoop, Mahout, Oozie, Storm and TEZ.

5 Classification

To predict student success, there are many data mining methods available to predict student success based on Classification Algorithms, K-Nearest Neighbors (KNN), C4.5, and the support Vector Machine algorithm (SVM). These algorithms are executed under hadoop by map reduce.

5.1 Knn

KNN [18] is a non-parametric machine learning algorithm and very simple and also easy to implement, it belongs to the class of supervised learning algorithms. The KNN algorithm is used to classify an unknown example based on the classes of its neighbors using distance calculator with its neighbors then it takes the majority class of its K neighbors. The algorithm of the KNN is that it calculates the distance of all the examples of the training data which increases the execution time, for this reason we used the parallel version of the KNN in order to minimize its execution time.

```
KNN algorithm
map
Input D,T
D: training data
T: testing data
For each line of training data one at time then Compute
the distance whith each element of T
Distance (D,T)
 save all of the test data distance with the training
 data and their class labels in order of distance (as-
 cending order)
reduce
k: number of neighbors to take
take K
take the k minimum distance and their class
take the most frequent class
end
```

5.2 C4.5 algorithm

C4.5 [19] algorithm is a directed classification algorithm, because it uses training samples, it allows the generation of decision trees to make the decision. The C4.5 algorithm uses the discrete data and also the continuous data unlike the ID3 algorithm which uses the discrete data only. The C4.5 algorithm prunes the decision trees after they have been built. It is often used in case of incomplete data.

```
C4.5 algorithm
MR1
Map
Input:<p1, v1>
 -  p1=line number;
 -  v1= the records;
 -  Extract class label and attribute
 -  p2= attribute and class label
 -  v2= 1
Output(p2, v2)
```

```
Reduce1
input<p2, List<v2>>
 -  Counts frequency of attribute with class Label
 -  P3=class label with attribute
 -  V3= frequency
Output (p3, v3)
MR2
Input p3 , V3
Map
 -  Calculate entropy (p3) and information gain (p3)
    and split-info (p3).
 -  P4: attributes
 -  V4: information gain, entropy, split-info
output (P4 , V4).
Reduce
 -  Calculate Information Gain Ratio for each attribute
 -  P5 = find the decision node
 -  V5 = the Information Gain Ratio
output (P5 , V5)
MR3
Input P5 , V5
Map
 -  Compute the id of the node for highest attribute
 -  P6=id of the node
 -  V6= Elements (attribute values)
 -  until all this data are classified recall this pro-
    cess to create non leaf branches
    for creating non leaf branches
Output (P6,v6)
Reduce
 -  create the tree
output (tree)
```

C4.5 algorithm starts with the selection of the root then the branches are divided for each case of the attribute and so on until the cases of the branch have the same class.

To find the root attribute, the c4.5 algorithm uses the gain ratio of all attributes and then takes the attribute with the highest gain ratio.

```
Gain-ratio(T,F)=entropy(T)-$\sum_{j=1}^{m}\frac{|T_j|}{T}$ *entropy(Tj)
T: training data.
F: attribute.
m: F elements number.
| Tj |: the cases number in the jth partition.
|T|: number of cases in T.
```

```
  entropy(T)=$\sum_{j=1}^{m} -di * \log_2 di$.
  di: the proportion of Ti to T.
```

5.3 SVM algorithm

SVMs [20] are a collection of machine learning algorithms and methods. they are used for regression, classification and anomaly detection. SVMs are very flexible and easy to use, its principle is to separate the training data, according to their classes by borders in a way to maximize the distance (the margin) between the border and the different groups of data. In case of non-linearly separable data, SVMs use the kernel. This method projects the data into a large-dimensional vector space, to separate the data.

SVMs are very efficient in case of binary classification, as in our case. They are based on a solid mathematical theory.

```
  SVM Pseudo code
  Map
  Initialization of support vectors. vector support is
empty at iteration i = 0. VSg is empty
  Until hi = hi-1 do
    for m ∈ M // for any mapreduce function m in
m=1,2,…,M
  $S_m^i \leftarrow S_m^i \cup VS_g^i$
    End for
    i←i+1
  End until

  Reduce
  While hi ≠ hi-1 do
    for m ∈ M
  $VS_m$,hi←binary_svm(Sm)
    end for
    for m ∈ M
  $VS_g \leftarrow VS_g \cup VS_m$
    end for
    i←i+1
  End while
```

hi: the best hypothesis at i iteration.
M: number of mapreduce function or number of computers.
Sm: sub data at computer m.
VSm: support vector at mapreduce function m.
VSg: the global support vector.

6 Dataset

There are a number of factors that influence a student's prediction of success. When we counted the factors according to their occurrence in the literature, in the first order, we found, academic achievement, E-learning academic activity, student demographique, psychological attributes and also student environment. These factors contain a combination of attributes, as shown in this figure.

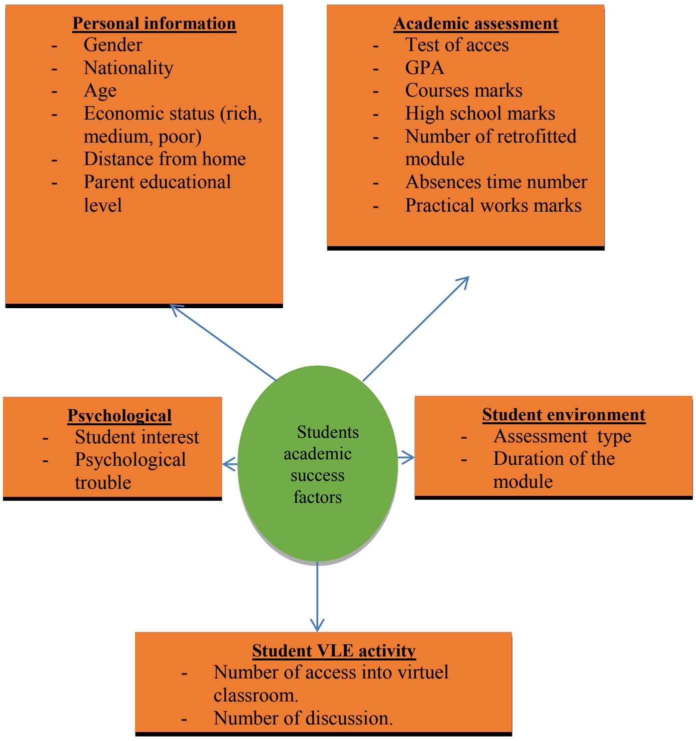

Fig. 1. Student success prediction attributes

We added the data of the practical work, Absences time number, distance from home and also Number of retrofitted module as attributes that will give a great value to the prediction of student success.

The model used.

Student data source: LMS, student university folder.

Data extraction, preprocessing (data cleaning (), discritisation), feature selection, storage in hdfs, classification, model creating, evaluation.

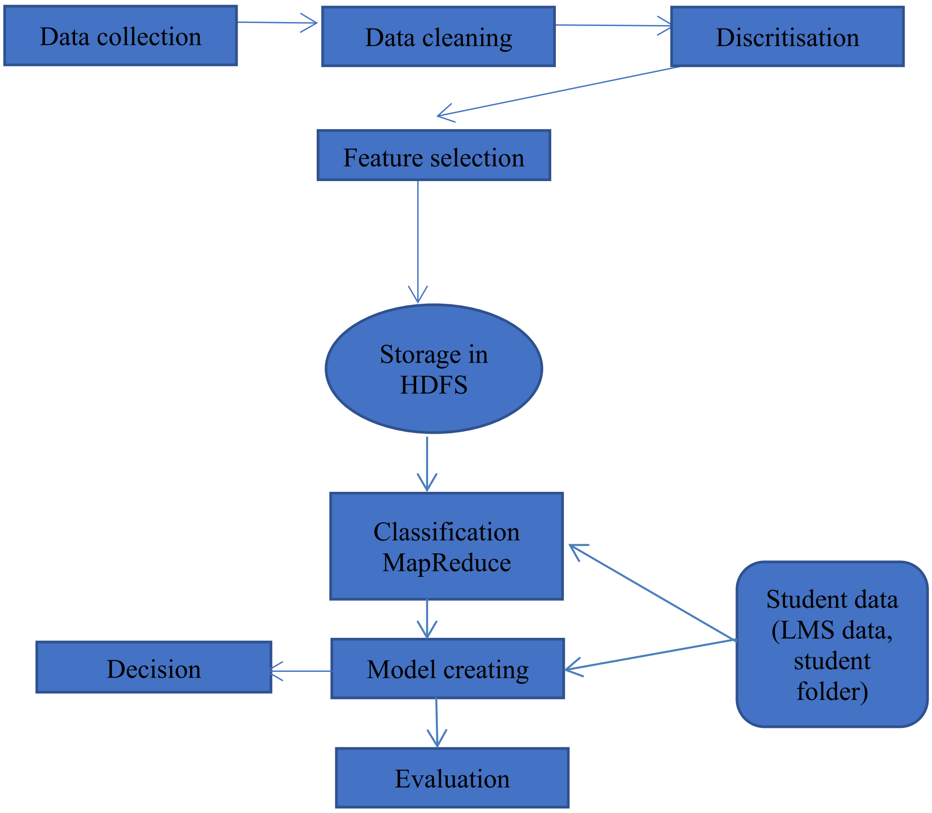

Fig. 2. System architecture used for student prediction

This model helps administrators and teachers predict student success or failure, and also intervene, to help students who are threatened with academic failure by possible tools.

This system begins with the collection of student data to build a database for the learning and testing phase. These data are passed to the cleaning phase to remove redundant data and incomplete data, and after discretization of continuous data, in order to make them usable by classification algorithms. The last three phases constitute the preprocessing step. The next step is the selection of attributes or the selection of variables. This step makes it possible to reduce the size of the data and to remove non-significant or redundant or irrelevant attributes. This increases the quality of prediction and minimizes the execution time. In this step we compared four attribute selection methods, the MRMR method and two envelope category attribute selection methods which are based on the j48 and SVM classification algorithms. Now the data is ready to use, it is stored in a distributed manner in HDFS to be used for the classification step. In this step, we compared three binary classification algorithms, because we have two classes, the "successful" class and the "Failure" class. After the classification we took the most powerful algorithm to use it in the model. Finally, we built a recommendation system that allows the prediction of student success.

7 Results

To choose the appropriate model for predicting student success, we tested several models that differ according to the attributes used, the attribute selection methods and the classification algorithms used.

We went through the confusion matrix, to calculate performance measures, which allowed us to evaluate the models for predicting student success or failure.

The measurements used are recall, precision, F-Measure, accuracy and Specificity.

- Recall, it is also called sensitivity, it represents the ability to detect student success by this model.

$$Recall=\frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{3}$$

- Precision, it represents the capacity of this model to detect only really successful students.

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{4}$$

- F-Measure is an evaluation criterion which is linked to both recall and precision, it is the harmonic mean.

$$\mathrm{F-mesure}=2\mathrm{x}\frac{\mathrm{PrecisionXRecall}}{\mathrm{Precision+Recall}}=\frac{\mathrm{2TP}}{\mathrm{2TP+FP+FN}} \tag{5}$$

- Accuracy, it is the proportion of correctly predicted student success.

$$Accuracy=\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{6}$$

- Specificity, it represents the ability of this model to detect all failed students.

$$Specificity=\frac{\mathrm{TN}}{\mathrm{FP+TN}} \tag{7}$$

TP: true positive
TN: true negative
FP: false positive
FN: false negative

Positive means class successful and negative means class failure.

True represents a correct classification, and false represents an incorrect classification.

The prediction quality increases as the previous measurements increase or close to 1.

These measurements were applied to models that use the MRMR method for the selection of attributes.

The following table shows the results obtained of MRMR algorithm.

Table 1. The results obtained by MRMR method

Algorithm	Recall	*Precision*	F-Mesure	Accuracy	*Specificity*
SVM	0.82.73	0.83,65	0.83,19	0.83	0.83.26
KNN	0.80.08	0.82,82	0.81,42	0.80.99	0.81.99
C4.5	0.78.35	0.83,24	0.80,72	0.80	0.81.89

This table presents recall, precision, F-Measure, Accuracy and Specificity of each student success prediction model with the classification algorithms used. According to the results of the table, we note that all the measurements are between 0.80 and 0.84 except the recall of the C4.5 algorithm which is less than 0.79. The accuracy of the SVM and also the other measurements of the same algorithm are the highest, followed by the KNN and finally C4.5. This makes the SVM algorithm the most efficient using the MRMR method for the selection of attributes.

For the results of the performance measurements of the models that use the j48 algorithm for the selection of attributes, the results are obtained which are presented in the following table.

Table 2. The results obtained by *the method based J48 algorithm*

Algorithm	Recall	*precision*	F-Mesure	Accuracy	*Specificity*
SVM	81.33	88,51	84,76	84	87.23
KNN	82.13	84,60	83,34	83	83.92
C4.5	83.08	90,63	86,69	86	89.55

According to the previous table, the model which uses the C4.5 algorithm has a very high classification rate (86%) compared to the other models. This is due to the use of the j48 algorithm [21] which is an implementation of the C4.5 algorithm at the attribute selection phase, with a precision exceeding 90%. And in second place is the model based on the SVM algorithm at the classification phase with a classification rate of 84%, and in third place we have the model which uses the KNN algorithm at the classification phase with a classification rate of 83%.

After the results of the performance measurement of the models which are based on the algorithm j48. The following table shows the results of the performance criteria that correspond to the models that are based on the SMO algorithms in the classification phase.

Table 3. The results obtained by the method based SMO algorithm

Algorithm	Recall	*Precision*	F-Mesure	Accuracy	*Specificity*
SVM	87.09	87,82	87,45	87.32	87.57
KNN	83	88,07	85,45	84.92	87.12
C4.5	85	87,87	86,41	86.1	87.29

When using the SMO algorithm in the attribute selection phase it was noticed that the model based on the SVM algorithm has the highest classification rate (87.32%) followed by the model based on the algorithm C4.5 then the model based on the KNN algorithm. The increase in the classification rate of the model based on the SVM algorithm is due to the use of the SMO algorithm in the attribute selection phase, because the SMO algorithm [22] is a simpler implementation of the SVM algorithm.

According to the results of the previous tables, we noticed that the model based on the SVM algorithm and the SMO algorithm is the most efficient because, the classification rate of this model and also its capacity to detect the success or the failure of students are the highest. Depending on the classification rate, we also notice that the models which are based on the SMO algorithm have the maximum classification rate compared to other algorithms (KNN and C4.5).

These performance measures alone are not enough. Because the choice of model also depends on the execution time. The following graph shows the execution time for each model. We took the time to run models based on the SMO algorithm because it gave maximum precision for all classification algorithms.

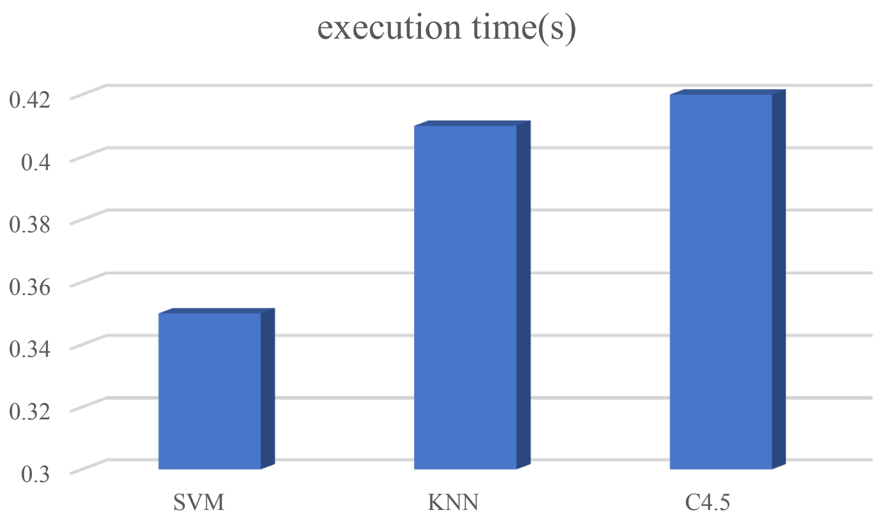

Fig. 3. SVM, KNN, C4.5 execution time

The preceding graph shows the execution time of the classification algorithms SVM, KNN and C4.5.

From the graph, we notice that the execution time of the SVM algorithm is the minimum followed by the KNN and finally the C4.5 algorithm.

8 Conclusion

The prediction of the success of the students is a very useful method because, it allows to decrease the rate of the academic failure by the intervention of the professors and all the actors in the teaching in the event that a student is (discovered) triggered by the system as a student is threatened by academic failure. The results of the comparison of models by performance measures and by execution time prove that the model based on the SMO algorithm at the attribute selection phase and the SVM algorithm at the classification phase is the best model among the models built. It has the highest classification rate (87.32%) and the lowest execution time.

After the selection of the attributes, we found that the factors that most influence the prediction of student success are firstly all the attributes of Academic assessment followed by economic status, parent educational level, distance from home, student interest, psychological disorder and number of access into virtual classroom. The other attributes have no value to add for predicting student success.

9 References

[1] Hussain, S., Khan, M.Q. Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning. Ann. Data. Sci. (2021). https://doi.org/10.1007/s40745-021-00341-0

[2] Xing Xu, Jianzhong Wang, Hao Peng, Ruilin Wu, »Prediction of academic performance associated with internet usage behaviors using machine learning algorithms», Computers in Human Behavior, 98 (2019) 166–173. https://doi.org/10.1016/j.chb.2019.04.015

[3] Alisa Bilal Zorić, «Predicting Students' Academic Performance Based on Enrolment Data», International Journal of Innovation and Economic Development, Volume 6 Issue 4 October 2020 Pages 54-61.

[4] Ahmed Mueen , Bassam Zafar,Umar Manzoor, «Modeling and Predicting Students' Academic Performance Using Data Mining Techniques«, International Journal of Modern Education and Computer Science 11(11):36-42 , 2016. https://doi.org/10.5815/ijmecs.2016.11.05

[5] Yassein NA, Helali RGM, Mohomad SB (2017) Predicting Student Academic Performance in KSA using Data Mining Techniques. J Inform Tech Softw Eng 7: 213. https://doi.org/10.4172/2165-7866.1000213

[6] A.I. Adekitan , Odunayo Paul Salau, «The impact of engineering students' performance in the first three years on their graduation result using educational data mining», Heliyon Volume 5, Issue 2, 2019. https://doi.org/10.1016/j.heliyon.2019.e01250

[7] Fadhilah Ahmad*, Nur Hafieza Ismail and Azwa Abdul Aziz,The prediction of students' academic performance using classification data mining techniques, Applied Mathematical Sciences ,Volume 9, Issue 129, 6415-6426 , 2015. https://doi.org/10.12988/ams.2015.53289

[8] Mashael A. Al-Barrak and Muna Al-Razgan «Predicting Students Final GPA Using Decision Trees: A Case Study», International Journal of Information and Education Technology, Vol. 6, No. 7, July 2016. https://doi.org/10.7763/IJIET.2016.V6.745

[9] Hilal Almarabeh, «Analysis of Students' Performance by Using Different Data Mining Classifiers«, International Journal of Modern Education and Computer Science · ,8, 9-15 August 2017. https://doi.org/10.5815/ijmecs.2017.08.02

[10] Ralph Olusola Aluko, Emmanuel Itodo Daniel, Olalekan Oshodi and Clinton Ohis Aigbavboa, Abiodun Olatunji Abisuga « Towards reliable prediction of academic performance of architecture students using data mining techniques » Journal of Engineering, Design and Technology,16(3), pp. 385-397,2018. https://doi.org/10.1108/JEDT-08-2017-0081

[11] Alaa Hamoud , Aqeel Humadi , Wid Akeel Awadh , Ali Salah Hashim, Students' Success Prediction Based on Bayes Algorithms International Journal of Computer Applications 178(7):6-12, November 2017. https://doi.org/10.2139/ssrn.3080633

[12] N. Putpuek, N. Rojanaprasert, K. Atchariyachanvanich and T. Thamrongthanyawong, "Comparative Study of Prediction Models for Final GPA Score: A Case Study of Rajabhat Rajanagarindra University," *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, Singapore, 2018, pp. 92-97. https://doi.org/10.1109/ICIS.2018.8466475

[13] Predicting University Students' Academic Success and Major Using Random Forests Cédric Beaulac, Jefrey S. Rosenthal, Res High Educ 60, 1048–1064 (2019). https://doi.org/10.1007/s11162-019-09546-y

[14] Zlatko J. Kovačić, »Early Prediction of Student Success: Mining Students Enrolment Data«, Proceedings of Informing Science & IT Education Conference (InSITE) 2010. https://doi.org/10.28945/1281

[15] Lu, H., & Yuan, J. (2018). Student Performance Prediction Model Based on Discriminative Feature Selection. International Journal of Emerging Technologies in Learning (iJET), 13(10), pp. 55–68. https://doi.org/10.3991/ijet.v13i10.9451

[16] Radovic, M., Ghalwash, M., Filipovic, N. et al. Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC Bioinformatics 18, 9 (2017). https://doi.org/10.1186/s12859-016-1423-9

[17] Khan, S. & Alqahtani, S. (2020). Big Data Application and its Impact on Education*. International Journal of Emerging Technologies in Learning (iJET), 15(17)*, 36-46. Kassel, Germany: International Journal of Emerging Technology in Learning. https://doi.org/10.3991/ijet.v15i17.14459

[18] Arowolo, M.O., Adebiyi, M.O., Adebiyi, A.A. et al. Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier. J Big Data 8, 29 (2021). https://doi.org/10.1186/s40537-021-00415-z

[19] Wu, Q. (2019). MOOC Learning Behavior Analysis and Teaching Intelligent Decision Support Method Based on Improved Decision Tree C4.5 Algorithm. International Journal of Emerging Technologies in Learning (iJET), 14(12), pp. 29–41. https://doi.org/10.3991/ijet.v14i12.10810

[20] Hu, Haifeng & Zheng, Junhui. (2016). Application of Teaching Quality Assessment Based on Parallel Genetic Support Vector Algorithm in the Cloud Computing Teaching System. International Journal of Emerging Technologies in Learning (iJET). 11. 16. https://doi.org/10.3991/ijet.v11i08.6040

[21] Singh, J., Singh, G. & Singh, R. Optimization of sentiment analysis using machine learning classifiers. Hum. Cent. Comput. Inf. Sci. 7, 32 (2017). https://doi.org/10.1186/s13673-017-0116-3

[22] Ghosh, M., Sanyal, G. An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning. J Big Data 5, 44 (2018). https://doi.org/10.1186/s40537-018-0152-5

10 Authors

Farouk Ouatik is member of Information processing and decision support laboratory. Faculty of sciences and Technics, Sultan Moulay Slimane University, Beni Mellal. Morocco (email: farouk.ouatike@gmail.com).

Mohammed Erritali is a Professor at Faculty of sciences and Technics, Sultan Moulay Slimane University. Morocco (email: m.erritali@usms.ma).

Fahd Ouatik is member of Physics High Energy and Astrophysics Laboratory. Faculty of sciences Semlalia, Cady Ayyad University, Marrakech. Morocco (email: fahd.ouatike@gmail.com).

Mostafa Jourhmane is a Professor at Faculty of sciences and Technics, Sultan Moulay Slimane University, Beni Mellal. Morocco (email: jourhmane@hotmail.com).

Article submitted 2022-02-16. Resubmitted 2022-03-22. Final acceptance 2022-03-23. Final version published as submitted by the authors.