Using Machine Learning for Prediction Students Failure in Morocco: An Application of the CRISP-DM Methodology

[1]Nada Lebkiri, [1]Mohamed Daoudi, [2]Zakaria Abidli, [2]Joumana Elturk, [1]Abdelmajid Soulaymani,
[1]Youssef Khatori, [3]Youssef El Madhi, [1]Mohammed Benattou,
[1]Ibn tofail University, Faculty of Sciences, Kénitra, Morocco
[2]Faculty of Health Sciences International University of Casablanca, Casablanca, Morocco
[3]Regional Center for Education and Training Professions, Rabat, Morocco

Received: March 12, 2021. Revised: September 14, 2021. Accepted: October 4, 2021. Published: October 22, 2021.

Abstract—Student failure prediction is one of the main topics in university learning contexts, as it helps to avoid failure in higher education institutions and provides a basis to make the teaching and learning process more effective, efficient and reliable. The overall aim of this study is to identify students who are susceptible to fail a given university course. This research paper reports the implementation of an Educational Data Mining project based on the CRISP-DM methodology. The data was collected from the APOGEE system of Ibn Tofail University, a form and specifications of the tested courses. The business goal of this paper is to develop a model that can identify students who are susceptible to failure in a given academic course. Such a model helps prevent failure in higher education institutions and provides a basis for making the teaching and learning process more effective, efficient and reliable. Most common machine learning algorithms in the field of Educational Data Mining were used. The results of our research showed that the proposed method was able to achieve an overall accuracy of 97% in predicting students at potential failure.

Keywords—Academic failure, Student failure prediction, Questionnaire validation, Educational Data Mining, Machine Learning, Decision Tree, Random forest, SVM, KNN, Adaptative Boosting, CRISP-DM, Ibn Tofail University, Morocco.

I. Introduction

UNIVERSITY failure in Morocco has become an alarming phenomenon since it affects a large number of students at all levels of university education. University failure is experienced in a dramatic way by both the student and their family, but it is also the failure of the educators who are in charge of training and who are very often powerless to deal with this phenomenon. The failure of students is also the failure of university institutions, which remain incapable of curbing this phenomenon despite the enormous investments made by the public authorities. In fact, university devote enormous human and material resources to teaching, while the return on investment remains largely insufficient. However, the rate of failure and university dropout is one of the major issues faced by the Moroccan Ministry of Education and Higher Education. In 2018, according to the Minister of Education in Morocco, 25% of new students in faculties that do not have a selection system do not pass the first semester. Worse still, 43% of students leave the university without obtaining a university diploma. This could be explained by several factors.

Predicting student performance has long been an important research topic in many academic disciplines[1]. According to the literature review in relation to academic failure [2] [3], predicting student failure helps to improve retention and graduation rates [4]. The automated prediction of student performance at an early stage is a useful perspective in the teaching and learning process, if students who are likely to fail are identified early enough, a series of corrective measures can be taken to improve their performance grades, providing additional support, i.e. if we predict that students who have a low GPA at graduation, additional efforts can be made to improve their results and therefore their grades[5]. As well as to select teaching materials that are appropriate to the capabilities of each student group. On the other hand, predicting students' academic performance helps higher education institutions to be able to plan a strategic intervention before students reach the last semester [6]. Thus prediction provides a basis for making the teaching and learning process more efficient and avoids the continuous waste of human and material resources on those unproductive students [7]. The application of data mining (DM) to education is an interdisciplinary research field, also known as educational data mining (EDM) [8].

Educational data mining is a field that exploits statistical, machine learning and data mining algorithms on different types of educational data. The educational data mining community website [9] defines educational data mining as follows: " Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in. " [10]. The EDM process therefore aims to convert raw data from education systems into useful information that could have a great impact on educational research and practice [11].

II. Literature review

A. EDM and Machine learning

The application of data mining to education is an interdisciplinary field of research, also known as Educational Data Mining (EDM) [8]. Educational data mining is a field that leverages statistical, machine learning, and data mining algorithms on different types of educational data. The EDM process therefore aims to convert raw data from educational systems into useful information that could have a great impact on educational research and practice [11]. The context of educational data mining can be seen as the intersection of three major fields: computer science, statistics and education. This intersection between these three fields also generates other subfields, closely related to EDM, such as computer-aided instruction, learning analytics (LA), data mining (DM) and machine learning (ML)[12]. To demonstrate the growing interest in EDM, Figure 1 shows the number of references returned by Scopus each year between 2000 and 2020 when searching “Educational Data Mining”. EDM has thus become a research area in recent years for researchers from all over the world from different and related research fields.

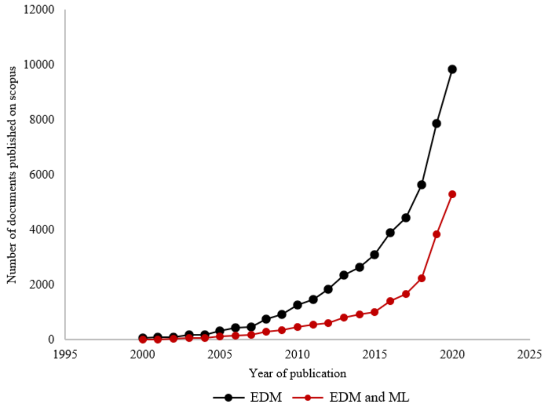

Fig. 1 Number of references in EDM and EDM ML in Scopus

B. Related work

The research by A. Saa, 2019 [13] was carried out at a private university in the United Arab Emirates. The overall size of the dataset extracted from the database includes 231,782 records related to the academic history of the students. The dataset in this study has a total of 34 attributes, divided into four types (demographic data, course and instructor information, general student information, and student past performance information). The main objective of this study was to predict students' academic performance from a new dataset extracted from a student information system. Six data mining algorithms were applied on the extracted dataset (decision tree, random forest, artificial neural network, Navie Bayes, logistic regression and generalized linear model. The results indicate that the Random Random Forest (RF) algorithm for predicting student performance. The empirical results of this research indicated that the Random Forest algorithm was the most appropriate data mining technique for predicting students' academic performance. The study also reveals that the most important attributes that have a direct effect on students' academic performance fall into four main categories, namely student demographics, information about students' past performance, information about courses and instructors, and general information about students.

Similarly, the use of random forests has shown excellent performance in predicting school dropout in terms of various performance measures for binary classification [2]. The data used in this study are the samples of 165,715 high school students from the 2014 National Education Information System (NEIS), which is a national education administration information system connected via the Internet to approximately 12,000 primary and secondary schools in Korea. 12 characteristics of the NEIS dataset for predicting school dropout: "unauthorised absence in the first 4 weeks", "unauthorised early leave in the first 4 weeks", "unauthorised absence from class in the first 4 weeks", "unauthorised tardiness in the first 4 weeks", "unauthorised absence", "unauthorised early leave", "unauthorised absence from class", "unauthorised tardiness", "self-regulated activity time", "club activity time", "volunteer work time", "career development time". The results showed that unauthorised absence was the most important variable in predicting student drop-out, followed by unauthorised lateness, self-regulated activity time, career development time and unauthorised early leave.

In order to examine the use of machine learning techniques in the field of dropout prediction, the six most common machine learning techniques were compared in the study of S. Kotsiantis, 2004 [14] these include decision trees, neural networks, the Naïve Bayes algorithm, instance-based learning algorithms, logistic regression and support vector machines. The Naïve Bayes algorithm showed the best behaviour. A prototype web-based support tool, which can automatically recognise students with a high probability of dropping out, was built using this algorithm.

In [15], the authors used rule induction when predicting the dropout of new nursing students. The main data set consisted of 3978 records of 528 nursing students. The source was standard university student records. The results of this study showed that the method achieved a sensitivity of 84%, a specificity of 70% and an accuracy of 94% on previously unseen cases.

III. Methodology

In this study, the steps of the predictive model development methodology using data mining are implemented according to the Cross Industry Standard Process for Data Mining (CRISP-DM) model [16]. The CRISP-DM process consists of a methodology and process model for data mining, which provides anyone with a comprehensive approach to conducting a data mining project. According to a survey published in KDnuggets [17] (a leading site in business analytics, big data, data mining, data science and machine learning), the CRISP-DM process is considered among the leading data mining methodologies [18] [19]. On the other hand, according to the work of [20] a systematic review of the literature on the application of the CRISP-DM process model, the authors listed the different areas in which CRISP-DM is applied. Most of the use cases are in the field of education [20].

The CRISP-DM process is an approach that consists of six steps (business understanding, data understanding, data preparation, modelling, evaluation and deployment) as shown in the Figure 2.

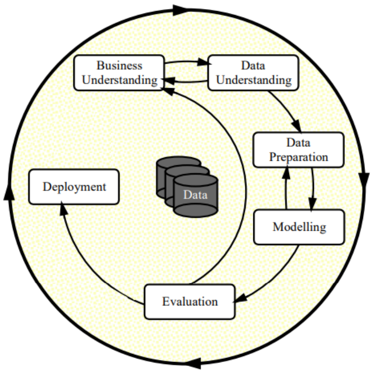

Fig. 2 Phases of the CRISP-DM process model for data extraction

CRISP-DM has the characteristic of being a cyclical process and the different stages communicate with each other, the process being able to return to previous stages to optimise its results. In Figure 2, the arrows indicate the most important and frequent dependencies between the phases, while the outer circle symbolises the cyclical nature of data mining itself and illustrates the fact that knowledge gained from the data mining process and the deployed solution can trigger new, often more focused, business questions. In the following, the main phases of the data mining process according to the CRISP-DM model will be explained [21] [22].

1) Business Understanding: This first phase is perhaps the most important phase of any data mining project, it focuses on understanding the objectives and requirements of the project from a business perspective and then converting this knowledge into a definition of the data mining problem and a preliminary plan designed to achieve the objectives. In other words, this initial phase therefore focuses on understanding the objectives and requirements of the project and then converting this information into a definition of the data mining problem and a preliminary plan designed to achieve the defined objectives.
2) Data Understanding: DU is the second phase of the CRISP-DM framework and aims to determine precisely what data to analyse, to identify the quality of the available data and to establish the link between the data and its meaning from a business perspective. The main objective of this stage is therefore to prepare for possible deficiencies, to guide the approach and to identify sub-categories on which hypotheses will be applied in order to discover imperceptible trends and patterns.
3) Data Preparation: This data preparation stage covers all activities necessary to construct the final dataset for analysis, made from the raw data. The data preparation tasks are likely to be repeated many times and in no prescribed order. They include the selection of tables, datasets, characteristic values, as well as the transformation, cleaning and recoding of the data to make it usable in the next phase.
4) Modelling: In this phase, a set of modelling techniques is selected and applied. Thus, their parameters are calibrated to optimal values. This step involves the implementation of different machine learning algorithms (regression, classification, clustering, recommendation) or statistical methods. Some techniques have specific requirements on the shape of the data. Therefore, it is often necessary to go back to the data preparation phase [22].
5) Evaluation: At this stage of the project, a model or models have been built that appear to be suitable for use in data analysis. Before the model is finalised, it must undergo a thorough evaluation and verification of the steps in its creation to ensure that the model is fit for purpose. A fundamental objective is to ensure that all the critical problems have been adequately addressed. The end of this phase should be marked by a decision on how to use the results of the extraction. The choice of evaluation measures depends entirely on the requirements of the project, the algorithm used, the desired outcome, etc.
6) Deployment: Once the model has been 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.

IV. The Achieved Results

A. Business Understanding

Essentially, the challenge in the presented data mining project is to predict the academic failure of students based on the collection of attributes providing information on the pre-academic characteristics of students. Our hypothesis is that a student can be characterised by some variables such as i. socio-demographic profile, ii. economic situation, iii. professional background, iv. academic background, v. non-academic background and vi. the knowledge of the students with respect to the prerequisites of a given course. And that a student's knowledge can be characterised by the grades obtained during the academic career. Based on this information, we can select students with similar characteristics and knowledge and then predict whether a particular student has sufficient skills to live in a non-failure situation. In this study, the dependent variable is therefore "academic failure", calculated on the basis of the student's master's degree averages:

- "No-failure"= If the student has validated all the modules without having a catch-up: all the marks obtained in the Master cycle are in the first session and are higher than 10 out of 20,

- "Failure"= If the student has validated his/her degree with at least one catch-up, or if he/she has validated his/her degree with one or more repeats, or if the student has dropped out of university: If the student obtains at least one mark below 10 out of 20.

The purpose of our study is to develop a Machine Learning algorithm that could predict when students are expected to fail or not to fail a given course of study. This is based on data from the university programme and the experiences of previous students (student's curriculum, professional experience, etc.).

This research focuses mainly on the grades obtained during the students' academic career before and after the integration of the formation (the students' marks data), and on the students' socio-demographic data. Therefore, these data were collected from the administration system of Ibn Tofail University (application for the organisation and management of training and students - APOGEE).

In the second part, the process of integration into the training courses (including the formal procedures and application documents), the requirement and the training modules. All this information has been extracted from the National Curriculum Standards booklet of each course - CNPN.

The first specific objective of the project is to propose a predictive tool that is sufficiently effective in predicting the potential for student's failure or success through the analysis of data collected at the level of university authorities (APOGEE, CNPN) and data collected from students in order to understand specific criteria that do not appear in the institutional data and that strongly condition student failure or success (according to the literature review). The objectives of student performance prediction can be classified into four main groups: i. dropout, ii. student performance, iii. recommended activities and resources and iv. student knowledge. In this research the dependent variable is academic failure and it was defined from the combination of the three objectives i. ii. and iv.

In our study we used five factors to predict student failure. These factors are: i) socio-demographic factors, ii) socio-economic factors, iii) academic factors, iv) non-academic factors, v) experience factors.

Fig. illustrates the proposed framework, it starts with data collection and preparation. The next step involves the preparation of the data set. Then, models are generated by learning algorithms. The results of these models are used to predict failure.

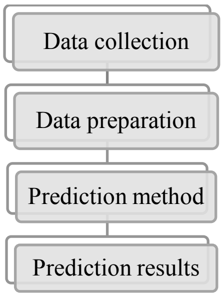

Fig. 3 The framework for the use of Artificial Intelligence to predict student failure

B. Data Understanding

The data used in our research comes from a public university (University Ibn Tofail - UIT). We have extracted it from three Master's courses namely: Master in education and teaching professions of the subject chemistry, mathematics and English of the Faculty of Science of Kenitra (chemistry, mathematics) and the Faculty of Letters and Humanities of Kenitra (English), from the academic year 2014-2015 to the academic year 2019-2020.

- **APOGEE data**

APOGEE is an Integrated Management Software Package (IMP) designed to manage student files and registrations in French and Moroccan universities [23]. The data collected from the APOGEE system are as follows: (i) Academic data - baccalaureate and license (data on the cursus of each student: baccalaureate data (the type of baccalaureate, the year of obtaining the baccalaureate, the grade of the baccalaureate), the title of the license, the university of the license, the year of obtaining the license, the number of years studied to obtain the license, the final grade of the license and the mention, the results of the module examinations of the students from the first semester to the sixth semester. (ii) The dependent variable "failure" (which has been calculated from the data obtained by the student during the whole licence training). (iii) Socio-demographic data of the student (date of birth, address...).

- **Documents used to apply for accreditation of the training courses according to the new pedagogical standards**

We analysed the documents used to apply for accreditation of the training courses according to the new pedagogical standards (CNPN) in order to extract the following data: (i) the required skills to integrate the training program, (ii) the program modules, and (iii) the skills that will be acquired during the program.

- **Data from the implemented questionnaire**

Presentation of the questionnaire

The literature review on the determinants of motivation [24] [25], satisfaction [26] [27] , the abandonment [28] and professional insertion [29] allowed us to retain several items that we used to construct our questionnaire. In total, 44 items were retained, divided into five parts: non-academic curriculum, professional curriculum, motivation, satisfaction, dropout, socio-demographic situation, and socio-economic situation. Thus, we distributed a survey form personally to the students in coordination with the pedagogical managers. 256 students participated in the survey.

Description and data mining

The data set was pre-processed, the incomplete or empty responses were removed to produce 184 responses used in the model. The average age during the first Master year is 26.59±5.65 years. In our study we noted that the age group 25-30 is the most represented. Regarding gender, we noted that men represent a percentage of 61.5% (n=131), while women represent 24.9% (n=53) with a sex ratio (m/f) of 2.4 ($p<0.005$). Students of urban origin represent 71.7% (n=132) and those from rural areas represent 28.3% (n=52). Concerning the socio-professional situation, we noted that before the Master's period, 23.4% (n=43) of the students were civil servants (work in the public sector of the kingdom) and 20.7% (n=38) of the students work in the private sector).

In our study we noted a significant differentiation between students having experienced failure and those not having experienced failure (p-value<0.0). With 16% of students not having experienced failure versus 84% of students having experienced failure. The failure rate is reflected by 55% having validated their modules in the catch-up session. On the other hand, we observed low percentages of students who repeated their academic years (4%), as well as students who reoriented themselves towards another course (3%). The dropout rate of university studies reached 22%.

Data quality validation

The quality of the predictions made by our method is totally dependent on the quality and quantity of the data collected on students. The measuring tools, as well as the data collected, must therefore be reliable so that the system can correctly identify students who are at risk of academic failure. The following section addresses this issue. In our study, the questionnaire validation process is based on two steps. Initially, we calculated the reliability of the entire questionnaire and the scale dimensions by Cronbach's alpha. Then, we performed principal component analyses to test its validity. All statistical analyses were performed using the Statistical Package for the Social Sciences (SPSS) statistical software, trial version. The data collected were subjected to exploratory analyses. In our study we calculated Cronbach's alpha to check the reliability and homogeneity between the items of the measurement scale. An alpha between 0.6 and 0.8 is acceptable for an exploratory study [30]. Similarly, we used the Kayser Meyer Olkin (KMO) test. A KMO index of less than 0.5 is unacceptable, 0.5 is poor, more than 0.6 is acceptable, 0.7 is average, 0.8 is meritorious and 0.9 is excellent [31], and Bartlett's test of sphericity to assess the potential effectiveness of the PCA studied. For a factor analysis to be feasible, Bartlett's test must be significant ($p<0.05$). Principal component analysis (PCA) is the most effective method for synthesizing information and uncovering the underlying structure of a construct since it is a multivariate data analysis method that allows for the simultaneous exploration of relationships that exist between several variables under study [32].

Internal item reliability

The Cronbach's alpha value of all items was 0.84, similarly, the calculation of Cronbach's alpha showed homogeneity of the different dimensions of the questionnaire used, both for non-academic background ($\alpha=0.87$), professional experiences and languages ($\alpha=0.93$), personal motivations ($\alpha=0.97$), satisfaction with training ($\alpha=0.93$), and reasons for dropping out ($\alpha = 0.96$).

Factor analysis

To perform the factor analysis of the questionnaire, we considered the value of the KMO index and Bartlett's sphericity test. For our study, the KMO index was 0.853>0.5 which shows an acceptable value to do the factor analysis, thus, Bartlett's sphericity test was highly significant. Factor analysis in the principal axes with varimax rotation showed that five factors with eigenvalues greater than 1 explaining 88.105% of the total variance.

The five dimensions are respectively well-defined and distinct, and each represents a dimension in its own right: the first factor, with four items (EX_1, EX_2, EX_3, EX_4) constituting the work experience dimension, explains 29.03% of the total variance. The second factor, with seven items (MOT_1, MOT_2, MOT_3, MOT_4, MOT_5, MOT_6, MOT_7) constituting the dimension of student motivation, explains 21.83% of the total variance. As for the third factor, with four items (SAT_1, SAT_2, SAT_3, SAT_4), it explains 14.87% of the total variance, constituting the dimension of students' satisfaction with the training. The fourth factor, which includes five items (ABD_1, ABD_2, ABD_3, ABD_4, ABD_5) constituting the dimension related to dropping out of school, explains 13.59% of the total variance. Table 1 represents the cleaned component matrix.

Table 1 The cleaned component matrix

	Component				
	1	**2**	**3**	**4**	**5**
EX_1				,724	
EX_2				,694	
EX_3				,723	
EX_4				,676	
MOT_1	,961				
MOT_2	,874				
MOT_3	,890				
MOT_4	,903				
MOT_5	,937				
MOT_6	,935				
MOT_7	,904				
SAT_1			,791		
SAT_2			,877		
SAT_3			,841		
SAT_4			,797		
ABD_1		,838			
ABD_2		,784			
ABD_3		,807			
ABD_4		,787			

ABD_5	,796	
NA_1		,628
NA_2		,764
NA_3		,769

The analyses that we carried out on the measurement tools were conclusive insofar as, on the one hand, the measurement scale proved to be reliable and valid and, the Bartlett analysis test proved to be significant (less than 0.001). The fulfilment of these two conditions allows us to carry out a causality analysis, which we did using the logistic regression method.

C. Data preparation

It is important to find the attributes that are most correlated to the final class (Failure) and how much they affect the final class. Significantly, this step shows the average correlation of the attributes to the final class. In turn, this average will help us find low correlation questions and remove them to improve the accuracy of the results. Items with high correlation can be considered as recommendation points for students and academic staff. This step is crucial because we want to find the variable(s) with which (i) the correlation is highest between the study variables and the dependent variable (ii) that affect the class (Failure) and (iii) eliminate the less related attributes from the model and (iv) improve the accuracy of the final model.

The statistical methodology used was based on two axes: descriptive statistics and analytical statistics. In the first part, we calculated the frequencies and characteristics of each variable studied that could give a general idea about the students. The results are expressed as a percentage. For qualitative variables and in mean ± standard deviation for quantitative variables. Then, in a second step, we used logistic regression with the objective of determining the factors that influence the dependent variable (academic failure). This predictive method aims to build a model that predicts and/or explains the values taken by a qualitative target variable. Technically, logistic regression proposes to test a regression model whose dependent variable is dichotomous (coded 0-1: Failure / Non-failure). The logistic regression model can also predict the probability of an event occurring (value of 1) or not (value of 0) from the optimization of the regression coefficients. When the predicted value is greater than 0.05, the event is likely to occur, while when the value is less than 0.05, it is not.

The chi-square test was conducted to determine the relationship between the variables studied (such as sociodemographic and economic variables, academic and non-academic variables, work experience, and foreign language) with the dependent variable explained by the student's failure.

On the multivariate level, the logistic regression shows that the variables : age of the student at the time of enrolment in the master's program, geographic location in relation to the university of training, number of children, economic situation, type of baccalaureate, option of the baccalaureate, year of obtaining the baccalaureate, field of study of the baccalaureate, grade of the baccalaureate, mention of the baccalaureate, pedagogical prerequisites, specialty prerequisites, number of years between the year of obtaining the baccalaureate and the license, work experience, duration of work experience, type of work experience, number of languages studied, non-academic degrees related to the training and number of years of study of these studies influence the dependent variable (failure) with p-value<0. 05.

In our study, we found a significant relationship between students' economic status and their academic performance (P-value<0.005). This result is in harmony with several studies that have shown that the economic situation plays a primary role on the success or failure of students, in addition, the lack of promising economic and professional motivation, the lack of respect and social status related to the field of study and the lack of job security in the future were mentioned as the most important factors by the students.

In the data preparation phase, we first applied a pre-processing technique to the collected data to prepare the data for extraction. The null value will cause problems for the classification model, such as loss of useful information and increasing uncertainty. The data was therefore studied for missing values.

In the data pre-processing phase, student and study course data from three databases are extracted and organized into a new flat file. The data provided undergoes numerous transformations. Some parameters are removed, for example, the last name, first name, and student number fields containing irrelevant data that are not of interest to the research. Some of the variables containing important data to the research are text fields in which free text is entered during data collection. Therefore, these variables are processed and transformed into categorical variables with a limited number of distinct values. Other values were added.

D. Modelling

The modelling phase consists of applying different machine learning techniques to the data set. The focus of this study, which is failure prediction, belongs to the problem of binary classification. Although there are many different techniques in the study of data mining, in this study we will focus on those used in the area of student performance prediction [33] [34] [35] [36]. We studied several machine learning classifiers and selected five classification methods that are commonly used to predict student performance. The five proposed algorithms are Decision Trees (DT) [37], Random Forests (RF) [38], Support Vector Machine (SVM) [39], K-Nearest Neighbor (KNN) [40] and Adaptive Boosting (AB) [41].

- A DT is one of the classification methods in data mining that is used to build a top-down tree model based on the attributes of a given data set. It is a predictive modeling technique used to predict, classify or categorize given data objects based on a previously generated model using a training data set with the same characteristics [37].

- The RF model is a popular model for prediction due to its high predictive accuracy,The RF algorithm is mainly the combination of Bagging and random subspace algorithms, and has been defined by Leo Breiman as "A combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [38].
- The SVM model was proposed by Vapnik in his book "The nature of statistical learning theory", and has been applied to many EDM classification problems in the literature. The mapping function in SVMs can be either a classification function (as is the case in this study) or a regression function [39].
- The KNN algorithm is one of the simplest and laziest supervised machine learning algorithms. It is a simple machine learning technique that approximates the decision boundary locally. The KNN classifier is an algorithm that ranks samples based on their proximity to the data used for training in a d-dimensional subspace [40].
- The AB algorithm generates a set of weak binary learners and combines them using a voting method; the t-th weak classifier takes into account the errors made by the previous weak classifier. This is done by assigning a weight D(i), to each learning example i: an example that is well ranked by the previous classifier will get a lower weight. A new training set is then sampled from the initial training set against these weights and a new classifier is trained [41].

The dataset used for this analysis was divided into two subsets: Trainset and Test set. The database has been divided so that the proportions of the target Failure are respected in both subsets, i.e. in the Trainset and in the Testset have the same proportions of profiles with the mention Failure and profiles with the notion No Failure. The tool used is the train-test division method in sklearn. 70% of the data points belong to the training set and the remaining 30% to the test set. The training set is used to fit the model of interest.In this study, we use four performance measures for binary classification to evaluate our trained model:

- Accuracy: measures the overall classification accuracy of the model; it compares predicted and actual values.
- Precision: Mathematically, precision is the ratio of correctly predicted positive observations to total predicted positive observations, and high precision generally corresponds to low false positive predictions.
- Recall: Recall is defined as the recall that quantifies the ratio of correctly labeled positive observations to all positive observations in the actual class. It measures the ability of the trained model to identify positive observations in all samples that should have been labelled as positive.
- F1 score: This is the weighted average of precision and recall. The F1 measure is the harmonic mean of precision and recall.

E. Evaluation

After the implementation of the models, we can observe from the table 2 That there was little difference in the performance of the prediction models: the decision tree classifier had the worst accuracy, at 83.33%. Followed by the Adaboost and KNN classifier, and were 94.44%. The accuracy of the model built by the random forest and SVM had the highest accuracy, which can reach 97.22%.

Table 2. Different algorithms used and the corresponding accuracy

Prediction model	Accuracy
DT	83%
RF	89%
AB	83%
SVM	97%
KNN	94%

In this study, several predictive modeling techniques of the data mining approach were applied to predict student performance. The accuracy results indicate, in terms of accuracy rate, that the random drill classifier and SVM outperform the other algorithms. The different evaluation methods previously discussed were used to evaluate the prediction results. The accuracy results indicate, in terms of accuracy rate, that the random forest classifier and SVM outperform the other algorithms. The different evaluation methods previously discussed were used to evaluate the prediction results. Based on the results presented in this study, the recall rate for the failure class shows some discrepancies between the five prediction models as a whole: SVM achieved the highest precision rate (99%), followed by KNN (96%), followed by RF (93%), while DT and AB achieved the lowest recall rate (82%). In contrast, the recall rate of the non-failure class: DT, AB, SVM, and KNN had the highest precision rate (88%), followed by RF (75%). The purpose of this analysis was to identify potential failures. The value of the F-measure targeted at the failure class reflects the overall effectiveness of the prediction models in predicting that class. The rankings of the models in ascending order are: SVM with 98%, KNN with a rate of 96%, RF with a rate of 93%, and finally DT and AD with 88%. The overall accuracy rate reflects the overall effectiveness of the prediction model. All three models had a relatively high overall accuracy rate, above 83%. The DT and AD classifier had the lower accuracy at 83%. Followed by the RF classifier which was 89%, followed by the KNN classifier with an overall accuracy of 94%. The accuracy of the model built by SVM had the highest accuracy, which can reach 97%.

Process review

Improving the techniques of machine learning algorithms can also help to increase the accuracy of prediction. We used only one model for prediction in this study, while using an integrated multi-model algorithm will help improve the accuracy to some extent. In this study we used stacking.

Stacking tries to achieve greater prediction accuracy by implementing different lower level learners and then combine them using a high-level meta-base learner [42].This technique has been used in many previous researches to increase the prediction accuracy and to reduce the prediction error [43] [44].

Following this, we experimented to increase the efficiency with the ensemble stacking technique. The results presented in Table 3 show that the stacking ensemble model can achieve a high overall accuracy of 98%.

Table 3. Stacking ensemble model accuracy`

	Accuracy
Stacking (DT, RF, AB, SVM, KNN)	0.98%

F. Deployment

In the present work, the implementation was deployed as a predictive web service based on prediction rules to focus on potential failures. Subsequently, feedback on predictions and processing should provide new data to improve the model.

V. Conclusion

This study proved that predicting student performance is important for the university to improve teaching performance. A variety of prediction models were applied. The ensemble learning technique resulted in the best overall performance of 98%. Moreover, our studies have demonstrated that there is a meaningful correlation between the validation of modules with no failures and the amount of professional experience of students. Therefore, we can conclude that it is advantageous to include the validation of acquired experience (VAE) in the pre-requisites for the integration of Moroccan university training. Our future work is now oriented towards the proposal of an intelligent experience validation process in Moroccan universities.

References

[1] S. Huang et N. Fang, « Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models », *Comput. Educ.*, vol. 61, p. 133‑ 145, févr. 2013, doi: 10.1016/j.compedu.2012.08.015.

[2] J. Y. Chung et S. Lee, « Dropout early warning systems for high school students using machine learning », *Child. Youth Serv. Rev.*, vol. 96, p. 346‑ 353, 2019, doi: https://doi.org/10.1016/j.childyouth.2018.11.030.

[3] M. A. Amoo, O. B. Alaba, et O. L. Usman, « Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach », *Int. J. Sci. Technol. Educ. Res.*, vol. 9, nº 1, p. 1‑ 8, mai 2018, doi: 10.5897/IJSTER2017.0415.

[4] A. Slim, G. L. Heileman, J. Kozlick, et C. T. Abdallah, « Predicting student success based on prior performance », in *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, Orlando, FL, USA, déc. 2014, p. 410‑ 415. doi: 10.1109/CIDM.2014.7008697.

[5] A. Tekin, « Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach », *Eurasian J. Educ. Res.*, vol. 14, nº 54, p. 207‑ 226, févr. 2014, doi: 10.14689/ejer.2014.54.12.

[6] P. M. Arsad, N. Buniyamin, et J. A. Manan, « A neural network students' performance prediction model (NNSPPM) », in *2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, Kuala Lumpur, Malaysia, nov. 2013, p. 1‑ 5. doi: 10.1109/ICSIMA.2013.6717966.

[7] J. Y. Chung et S. Lee, « Dropout early warning systems for high school students using machine learning », *Child. Youth Serv. Rev.*, vol. 96, p. 346‑ 353, 2019, doi: https://doi.org/10.1016/j.childyouth.2018.11.030.

[8] C. Romero et S. Ventura, « Data mining in education: Data mining in education », *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, nº 1, p. 12‑ 27, janv. 2013, doi: 10.1002/widm.1075.

[9] « Site web de la communauté l'exploration des données éducatives ». févr. 13, 2020. [En ligne]. Disponible sur: www.educationaldatamining.org

[10] I. H. Witten et E. Frank, « Data mining: practical machine learning tools and techniques with Java implementations », *Acm Sigmod Rec.*, vol. 31, nº 1, p. 76‑ 77, 2002.

[11] C. Romero et S. Ventura, « Educational data mining: A survey from 1995 to 2005 », *Expert Syst. Appl.*, vol. 33, nº 1, p. 135‑ 146, 2007.

[12] R. Cristobal et S. Ventura, « Data mining in education », *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 3, nº 1, p. 12‑ 27, 2013.

[13] A. A. Saa, M. Al-Emran, et K. Shaalan, « Mining student information system records to predict students' academic performance », in *International conference on advanced machine learning technologies and applications*, 2019, p. 229‑ 239.

[14] S. Kotsiantis, C. Pierrakeas, et P. Pintelas, « PREDICTING STUDENTS' PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES », *Appl. Artif. Intell.*, vol. 18, nº 5, p. 411‑ 426, mai 2004, doi: 10.1080/08839510490442058.

[15] L. G. Moseley et D. M. Mead, « Predicting who will drop out of nursing courses: A machine learning exercise », *Nurse Educ. Today*, vol. 28, nº 4, p. 469‑ 475, mai 2008, doi: 10.1016/j.nedt.2007.07.012.

[16] C. Shearer, « The CRISP-DM model: the new blueprint for data mining », *J. Data Warehous.*, vol. 5, nº 4, p. 13‑ 22, 2000.

[17] *kdnuggets*. 2020. [En ligne]. Disponible sur: https://www.kdnuggets.com/

[18] « G. Piatetsky-Shapiro. (2014). Data Mining Methodology ». [En ligne]. Disponible sur: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-miningdata-science-projects.html

[19] M. Naili, « Fouille de données dynamique basée sur la stabilité des modèles dynamiques », Université Mohamed Kheider-Biskra, 2019.

[20] C. Schröer, F. Kruse, et J. M. Gómez, « A Systematic Literature Review on Applying CRISP-DM Process Model », *Procedia Comput. Sci.*, vol. 181, p. 526‑534, 2021.

[21] O. Niaksu, « CRISP data mining methodology extension for medical domain », *Balt. J. Mod. Comput.*, vol. 3, nº 2, p. 92, 2015.

[22] P. Chapman *et al.*, « CRISP-DM 1.0: Step-by-step data mining guide », *SPSS Inc*, vol. 9, p. 13, 2000.

[23] « Arrêté du 26 janvier 1995 portant création d'une application informatique nationale de gestion des enseignements et des étudiants », avr. 25, 2020. [En ligne]. Disponible sur: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000186826

[24] R. Viau, « La motivation des étudiants à l'université: mieux comprendre pour mieux agir », 2006.

[25] Y. FORNER, « La motivation à la réussite scolaire dans les situations de formation: QMF Manuel », *Issy--Moulineaux Éditions EAP*, 1993.

[26] R. Kahombera et F. Duranton, « SATISFACTION DES ETUDIANTS DANS UNE INSTITUTION D'ENSEIGNEMENT UNIVERSITAIRE », 2018.

[27] S. Carayon et P.-Y. Gilles, « Développement du questionnaire d'adaptation des étudiants à l'université (QAEU) », *Orientat. Sc. Prof.*, nº 34/2, p. 165‑189, 2005.

[28] N. Beaupère, G. Boudesseul, et S. Macaire, « Sortir sans diplôme de l'Université », *Compr. Parcours*, 2009.

[29] F. Mourji et A. Gourch, « Modélisation de l'insertion professionnelle des diplômés de l'enseignement supérieur au Maroc », *Crit. Économique*, nº 22, 2008.

[30] R. A. Johnson et D. W. Wichern, *Applied multivariate statistical analysis*, vol. 5, nº 8. Prentice hall Upper Saddle River, NJ, 2002.

[31] D. W. Stewart, « The application and misapplication of factor analysis in marketing research », *J. Mark. Res.*, vol. 18, nº 1, p. 51‑62, 1981.

[32] A. Field, *Discovering statistics using IBM SPSS statistics*. sage, 2013.

[33] L. C. Borges, V. M. Marques, et J. Bernardino, « Comparison of data mining techniques and tools for data classification », in *Proceedings of the International C* Conference on Computer Science and Software Engineering - C3S2E '13*, Porto, Portugal, 2013, p. 113. doi: 10.1145/2494444.2494451.

[34] C. Romero et S. Ventura, « Educational Data Mining: A Review of the State of the Art », *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 40, nº 6, Art. nº 6, nov. 2010, doi: 10.1109/TSMCC.2010.2053532.

[35] R. B. Sachin et M. S. Vijay, « A Survey and Future Vision of Data Mining in Educational Field », in *2012 Second International Conference on Advanced Computing & Communication Technologies*, Rohtak,Haryana, India, janv. 2012, p. 96‑100. doi: 10.1109/ACCT.2012.14.

[36] M. Sanchez-Santillan, Mp. Paule-Ruiz, R. Cerezo, et Jc. Nuñez, « Predicting Students' Performance: Incremental Interaction Classifiers », in *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, Edinburgh Scotland UK, avr. 2016, p. 217‑220. doi: 10.1145/2876034.2893418.

[37] N. Bhargava, G. Sharma, R. Bhargava, et M. Mathuria, « Decision tree analysis on j48 algorithm for data mining », *Proc. Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, nº 6, 2013.

[38] L. Breiman, « Random forests », *Mach. Learn.*, vol. 45, nº 1, p. 5‑32, 2001.

[39] C. Cortes et V. Vapnik, « Support-vector networks », *Mach. Learn.*, vol. 20, nº 3, p. 273‑297, 1995.

[40] G. Guo, H. Wang, D. Bell, Y. Bi, et K. Greer, « KNN model-based approach in classification », in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 2003, p. 986‑996.

[41] Y. Freund et R. E. Schapire, « A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting », *J. Comput. Syst. Sci.*, vol. 55, nº 1, p. 119‑139, août 1997, doi: 10.1006/jcss.1997.1504.

[42] D. H. Wolpert, « Stacked generalization », *Neural Netw.*, vol. 5, nº 2, p. 241‑259, 1992.

[43] N. Chanamarn, K. Tamee, et P. Sittidech, « Stacking technique for academic achievement prediction », *Int Work Smart Info-Media Syst Asia SISA 2016*, p. 14‑17, 2016.

[44] Z. Kang, « Using machine learning algorithms to predict first-generation college students' six-year graduation: a case study », *Int. J. Inf. Technol. Comput. Sci.*, vol. 11, nº 9, p. 1‑8, 2019.