Survey of Machine Learning Techniques for Student Profile Modelling

https://doi.org/10.3991/ijet.v16i04.18643

Touria Hamim (✉), Faouzia Benabbou, Nawal Sael
Faculty of sciences Ben M'SIK, University Hassan II,
Casablanca, Morocco
hamimxtouria@gmail.com

Abstract—Developments in information technology have led to the emergence of several online platforms for educational purposes, such as e-learning platforms, e-recommendation systems, e-recruitment system, etc. These systems exploit advances in Machine Learning to provide services tailored to the needs and profile of students. In this paper, we propose a state of art on student profile modelling using machine learning techniques during last four years. We aim to analyse the most used and most efficient machine learning techniques in both online and face-to-face education context, for different objectives such as failure, dropout, orientation, academic performance, etc. and also analyse the dominant features used for each objective in order to achieve a global view of the student profile model. Decision Tree is the most used and the most efficient by most of research studies. And academic, personal identity and online behaviour are the top characteristics used for the student profile. To strengthen the survey results, an experiment was carried out, based on the application of machine learning techniques extracted from the state of art analysis, on the same datasets. Decision tree gave the highest performance, which confirms the survey results.

Keywords—Profile modelling, Student profile, Machine learning

1 Introduction

Student profile modelling relies on a profile representation that captures the main characteristics and gives the most coherent, complete and operational representation of the student. The student's characteristics include background knowledge, learning preference, behaviours, skills, goals, etc. The student profile model can be constructed through the analysis of data from different sources as student records, social networks, learning platforms, web form, etc. Indeed, several works used the student profile either to propose an adaptive learning, to guide him in his/her academic choices, or to make recommendations about his/her future career. Machine Learning (ML) is one of the used methods used for student profile modelling, that aims to create knowledge automatically from data. Mainly, they are used in classification, prediction and in decision support domains [1]. Machine Learning techniques can be usually classified into three categories: Supervised learning, unsupervised learning, and semi-supervised; but with the development in artificial intelligence others classes were introduced as deep, transfer and reinforcement learning. This techniques were applied in several levels to achieve many academic objectives like predicting failure or drop-out, orientation and academic decision making [2]–[4].

In this paper, we propose a survey of research studies on student profile modelling using machine learning techniques during last four years (2016-2019). We focus on the student's features categorization, machine learning techniques based-on and the context of the research study, to be able in the future, to develop a profile model and to generalize on the conclusions obtained in our previous work which has shown that decision trees are the most efficient based on academic data [5].

This article is organized as follows. In section II we present the related works dealing with student profile modelling using machine learning. In section III, we give a comparative and statistical analysis of the different research studied. The last section presents a case of study where we applied different machine learning techniques on two online datasets and at last, we give a conclusion and some perspectives.

2 Related Works

With the availability of data on the e-learning platforms, several studies have been carried out in order to adapt the training and to personalize the contents to the learner expectations. A. Topîrceanu et al. [6] proposed to optimize the way e-learning systems are developed using Decision Tree (DT) technique. S. V. Kolekar et al. [7], proposed to identify the way students learn to customize the resources delivery through the use of FCM clustering and NN-based classification techniques. M. Abdullah et al. [8], have performed multiple classifiers as J48, NBTree (Naive Bayes (NB) in association with DT) and NB in association with Sequential Minimal Optimization (SMO) to carry out the relationship between educators and student's learning style. R. Cerezo et al. [9] propose a student's classification based on their behavior to predict their achievement. Other models for student's performance were proposed based on Expectation-Maximization (EM) and K-means techniques [10] and SVM and Association Rules [11].

Failure and dropping out of school are serious educational systm challenges, that is why several studies have been devoted to these topics. Based on academic students' data, a system for an early detection of students with difficulties [12] was proposed using three classification techniques : Random Forest (thresholds and leaves), Logistic Regression (LR) and Artificial Neural Network (ANN). They all gave good accuracy and Random Forest TH (thresholds) was the most efficient. A. U. Khasanah et al. [13] proposed a system to prevent student failure, Bayesian Network and DT techniques were implemented and compared. The most significant prediction was provided by NB. Likewise, in the context of failure prediction, other classifier models were proposed such as SVM (RTV-SVM) [14] and ANN [15]. Several researches have been conducted to understand students' reasons for dropping and some reported were economic situation, social status, drugs and motivation. In the context of MOOC platforms, M. Khalil et al. [16] used K-means clustering to split the students according to their engagement and behaviors. C. Burgos et al. [17], proposed a predictive model of courses' student drop-out using a system based on LR. Another approach [18] based on big-data was conducted to minimize drop-out by enhancing online learning courses to satisfy the learner objectives, the SVM, NB and K-NN techniques were compared and SVM was the most efficient. In [19], S. Kai et al. proposed a model using J-Rip classifier and J-48 Decision Trees to predict the potential students that would continue participation in the online college program. In the traditional education, random forests was used [20] to predict students at risk of dropping out. Using data gathered from a large dataset, L. Aulck et al. [21], proposed to seek determinant features in prediction of dropout students and make recommendation to reduce it. Some authors tried to identify relevant student' attributes to predict student dropout rate with ID3 DT [22]. C. Márquez-Vera et al. [23], proposed a methodology and a specific classification algorithm ICRM2, a variant of GP known as grammar-based genetic programming (GBGP), to discover comprehensible prediction models of student dropout earlier.

The recommendation systems rely on the students' features to propose them the appropriate pedagogical contents or the most relevant educational pathways. Making a recommendation for students, teachers, educators and administration was the objective of the study in [24], and C4.5 algorithm was used to improve learning outcome, by detecting automatically the students' learning styles and recommend them the better aligned contents. Linear Discriminant Analysis (LDA), LR, and the Linear SVM (LSVM) were exploited to improve the e-learning platform by offering a personalized learning situation to guide student behavior, and LSVM made the best accuracy [25]. Based on students' academic history analysis, P. Dash et al. [26] proposed a decision support system that helps the student to select a particular subject to read and RF algorithm gave 99% for accuracy.

Other researches have been conducted to improve the student performance. S. Bharara et al. [27], aim to find the features that directly influence student performance using the K-means algorithm. A student's performance prediction model based on the interactivity with the e-learning management system is proposed by using the three classifiers ANN, NB and J48 which was the most efficient using three class labels [28]. A. Mueen et al. [29] proposed prediction model for students' academic performance based on their academic records and forum participation by using the three classifiers Multiple Layer Perceptron, DT(C4.5) and NB that was the most powerful. The impact of using social networks was analyzed to predict the academic results by using CART method [30]. To enhance the quality of the higher education system by evaluating student performance in courses, A. El-Halees [31], applied data mining techniques to discover knowledge based on association rules, classification (J48), also they clustered students into groups using EM clustering. The paper [32], proposed to make early intervention to improve the module results and enhance student's experience by using RF and SMO. Another work proposed by T. Mahboob et al. [33] suggest to help the students to improve their performance by evaluating themselves on the basis of their prior records and act as a guide for future evaluations on performance. In the context of traditional education, C. Masci et al. [34] aim to identify student characteristics that have a direct impact on his or her results using regression trees. Understanding issues and problems students encounter in their learning experience with the goal of minimizing the student's educational problems is the objective of S. Patil et al. [35], by using NB, DT(ID3) and Memetic Algorithm (MA) which was the most efficient algorithm. In order to build an interpretable student performance prediction model, comment data mining with DT(C4.5) and RF have been performed and RF was the most efficient [36]. Studying the relationship between the cognitive admission entry requirement and the academic performance of students in their first year, using NN was the objective of [37]. A. Abu [38] explored multiple factors that affect students' performance in higher education to predict their performance, and four DT algorithms have been implemented as well as NB and the variables 'neighborhood'(student's residence) and 'school' were the main factors that affect student's performance. A new model that enhances the DT accuracy in identifying student's performance was presented in [39], four DT algorithms were applied and BFTree shown more accuracy than other classifiers. K. Karthikeyan et al. [40], predict students' performance and give them a chance to improve it in the future. The research work combines two data mining techniques, namely, Clustering (Enhanced K-Means) and Classification (SVM) named as CESVM-SPPS and it gave successful results. To identify students with special need attention from the beginning of the course at the right time, the authors used K-means clustering to concentrate students in groups of similar characteristics [41]. Association rules was performed to help educators understanding the learning and psychological states of students in different grades, so as to formulate teaching plans and improve their academic performance [42]. R. Asif et al. [43] analyze the performance of students and study the directors program which could help them improving the program, and the NB was the most efficient technique used to predict the graduation performance in a four-year university program.

Some researches studies focused on improvements that can be done on learning platforms. The objective is to help administration to improve the learning environment by analyzing different students' opinions by using BN [44]. In [45], the authors examined the variation in students' confidence and engagement with digital technologies in learning and considered possible implications for teacher's learning design using association rules. S. K. Howard et al. [18] proposed a big-data driven approach for online learning evolution to discover students' learning patterns to guide courses improvement and satisfy the learner by comparing three machine learning techniques (SVM,NB and K-NN), and SVM was the most efficient technique in this case.

3 Comparative Study

3.1 Criteria

The state of the art addresses various academic problems where providing a model of student profile is essential to give effective solutions. Sometimes the same techniques are used in different contexts. The discovery of student patterns is often based on clustering techniques and a set of characteristics have been identified in order to model the student profile accurately. In this comparative study, we rely on set of criteria as described in table 1, and we focus on: publishing year, Objective: the subject covered by the paper, Context: the circumstances surrounding learning which can be either distance learning or e-learning or both.

Dataset source and size, student features : each feature of each state-of-the-art paper is assigned to a category of features according to a categorization defined in [46], and the techniques ML applied and their performance. For the performance value we gave only the best result and the belonging technique is given in bold font.

Table 1. Criteria

Attribute	Explanation
Year	Publishing year
Context	E-learning (EL), traditional Learning (TL), E-learning & traditional education
Objective	Adaptive learning, E-learning performance, failure prediction, dropout prediction, recommendation, enhance academic performance and improve learning environment, etc.
Dataset	Source (Questionnaire: Q or Dataset: D) and size of data
Student features' categories	Personal Identity (PI), Social Identity (SI), Academic (AC), Online Behaviour (OB), Learning Behaviour (LB), Language (L), Psychological (PSY), Physical Conditions (PHC), Skills/Interests (SK), Learning Goal (LG), Learning Style (LS)
Technique	NB, NN, SVM, K-NN, DT (J48, ID3, C4.5, J-Rip, CART, etc.), MA, RF, K-means, FCM, EM, LR, AR, etc.
Performance metrics	Accuracy (A), Precision (P), Recall (R), F-Score (F) and (O) for other performance metrics (Confidence (C), Area Under Curve (AUC), Average Silhouette (S))

3.2 Statistical analysis and discussion

In Table 2, we present a comparison of researches dealing with student profile modeling using machine learning during last four years. The purpose of most papers is to improve student academic performance, to understand difficulties encountered in the learning process and how to enhance competencies. For other papers, to identify parameters that influence failure or dropout was a real challenge. These studies analyze several context-dependent factors in order to predict a student's outcomes and propose solutions to deal with educational challenges. The third most common objective is the adaptive learning, to improve the quality of the learning environment and finally to improve the students’ outcomes. The data source used varies according to three types: questionnaires, databases (from the university or drawn online), or both of them. Most of them uses a data size of the hundreds scale and researches that use a large data size are quite rare.

Ref	Year	Context			Objective	Dataset			Student features' cathegories	Technique	Performance (%)				
		EL	TL	EL & TE		Source		Size			A	P	R	F	O
						Q	D								
[45]	2016	✓			Learning environment	✓	✓	21795	AC - OB	AP					C 93
[28]	2016			✓	Academic performance	✓		500	PI – SI –AC – OB - LB	ANN – NB – J48	82.2	85	82.2	81.8	
[36]	2016		✓		Academic performance	✓		89	AC	C4.5 - RF		89	87.8	88.3	
[29]	2016			✓	Academic performance		✓	60	PI – SI – AC - OB - LB	NB - MLP – C4.5	86	88.4	85.8		
[21]	2016		✓		Dropout		✓	32538	PI – SI - AC	LR – RF – K-NN	66.59				
[33]	2016			✓	Academic performance			60	PI – SI - AC	J48 – NB - RF		100	100	100	
[22]	2016		✓		Dropout		✓	240	PI – SI - AC	ID3	97.5	91.4	99	95	
[38]	2016		✓		Academic performance	✓		270	PI – SI – L – AC – PSY - LB	C4.5 – ID3 –CART – CHAID	40				
[30]	2016			✓	Academic performance	✓		139	PI – AC - OB	CART			92.27	83.76	
[9]	2016	✓			E-learning performance		✓	140	OB	EM K-means					
[23]	2016		✓		Dropout	✓		419	PI – SI –PHC – LB - AC – PSY – L – SK	NB – SMO – K-NN – JRip – J48 – ICRM2	99.8				
[31]	2016			✓	Academic performance		✓	151	PI – AC – OB	AR – J48 - EM					C 62.5
[15]	2017			✓	Failure		✓	EL:262 TE:161	PI – SI – AC – OB - LB	SVM – J48 - NN	92				
[16]	2017		✓		Failure		✓	6845	PI - AC	RF(TH) – RF(L) – LR - NN	100				
[25]	2017	✓			Recommendation and adaptive learning		✓		OB	LDA – LR - LSVM	99.7	64.3	81.8	72	
[6]	2017	✓			Adaptive learning	✓		632	PI – AC - OB	DT					
[16]	2017	✓			E-learning retention	✓	✓	838	PI – OB	K-means					
[18]	2017	✓			E-learning retention and learning environment		✓		OB	SVM – NB –K-NN	80-90				
[44]	2017		✓		Learning environment	✓		250		NB					
[10]	2017	✓			E-learning performance		✓	336	OB	SVM	74.1				
[26]	2017		✓		Recommendation	✓		324	PI – AC – SI	NB – J48 – K-NN – RF – PART	90.07				
[7]	2017	✓			Adaptive learning		✓	108	LS	FCM – GSBPNN - BPNN	95.93	96.33	99.05	97.67	
[13]	2017		✓		Failure		✓	90	PI – SI - AC	BN - DT	98.08				
[8]	2017			✓	Adaptive learning	✓	✓	48	LS - AC	J48 – NBTree(NB+DT) – NB – NB+SMO – SMO	100				
[19]	2017	✓			E-learning retention		✓	151	PI – SI –OB – SK – PSY – LB - PHC	J48 – JRIP					AUC 93.7

[39]	2017		✓		Academic performance	✓	✓	450	PI – SI – LB – AC – SK – PSY – PHC - LG	BFTree – J48 – RepTree – Simple Cart	80.22				
[43]	2017		✓		Academic performance		✓	210	AC	DT – RI – 1-NN – NB – NN – RF – X-means	83.65				
[41]	2017		✓		Academic performance		✓	662	PI - AC	K-means					
[40]	2017		✓		Academic performance	✓	✓	1045	PI – SI - AC	SVM – BPNN – KNN – CESVM – 2SC – 2BC – 2KC – 2CC					
[17]	2017	✓			E-learning retention		✓	100	OB	LOGIT_Act – SEDM – FFNN – PESFAM - SVM	97.13	98.95	96.73		
[14]	2018			✓	Failure		✓	32593	PI – OB - AC	RTV-SVM	93.8		94		
[34]	2018		✓		Academic performance	✓			PI – SI – AC – LB – SK - PSY	RT	≈ 90				
[27]	2018			✓	Academic performance		✓	500	PI – SI – AC – OB - LB	K-means					S 62.4
[35]	2018		✓		Academic performance		✓	2785	PSY - AC	MA – ID3 - NB	90.94				
[42]	2018		✓		Academic performance		✓		PI – AC - PSY	AP					
[24]	2018		✓		Recommendation	✓		700	PI – AC - LS	C4.5 – CART – BN - NB	95.7	98.6	72	83	
[37]	2018		✓		Academic performance		✓	1445	PI - AC	Tree – RF – NN – NB – LR	51.9	48.6	50	49.4	
[32]	2018			✓	Academic performance		✓	22	AC – LB - OB	J48 – RT –RF – LMT(LR+DT) - HT – DS – NB - SMO	100				
[20]	2019		✓		Dropout		✓	165715	AC – SK - LB	RF	95		85		
[11]	2019	✓			E-learning performance		✓	76268	PI - OB	AP					

In our analysis, we are interested in studying the objectives that have been most addressed in the research studies to understand the major concerns of the academic community. As shown in figure 1, the student academic performance is the greatest goal of the community, followed by adaptive learning, failure and dropout and different machine learning techniques have been investigated to improve performance, and explain the reasons for failure and dropout. The survey shows that the most prevalent context is traditional education, followed by e-learning and then 25% of papers are related to both contexts.

Data quality and size are important factors in the field of learning machines. The databases from learning management system (LMS) are generally large, but the size of university databases generally varies between the hundreds and thousand scale. From our analysis, we notice that most authors used databases from academic systems or online e-learning systems, and the use of questionnaires represents only 27% of the researchers and 13% use the data from both.

Figure 2 presents the distribution of the student features' categories used in the research studied. The academic information (grades, major, diploma, etc.), are the most used for different contexts with a percentage of 75%, followed by personal identity characteristics (65%) (Gender, age, nationality, etc.), online behavior (45%) (Comments, navigation, quizzes, etc.), and then social identity (37.5%) (Marital status, parents' education, parents' job, address, etc.). We can say that academic and personal information are unavoidable regardless of the purpose of the research study.

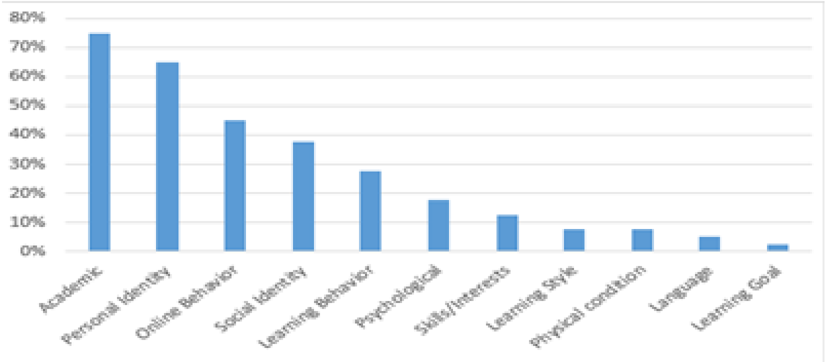


Fig. 1. Student features' categories distribution

To explain relation between the student features' categories and the context, the figure 3 reveals that the distribution of the top five student's characteristics depends on the context, and that online behaviour is the most used in the context of e-learning, whereas the academic performance and the social identity characteristics are the most relevant for traditional education.

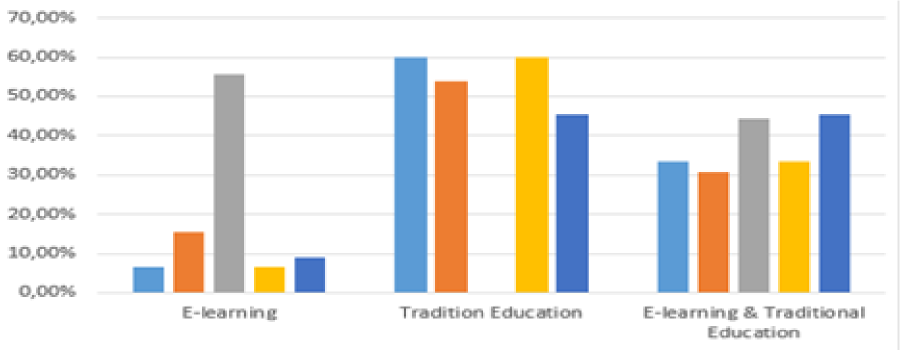
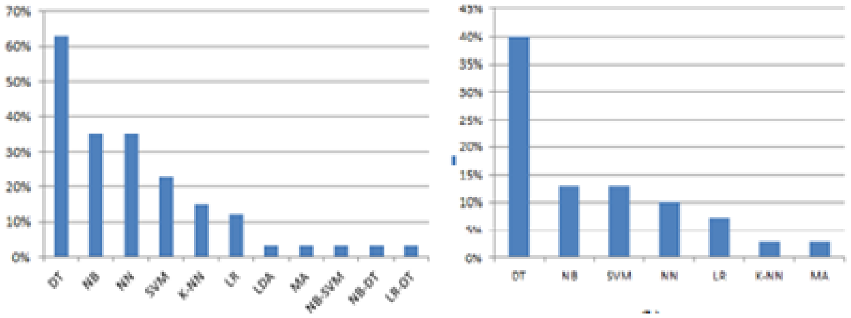


Fig. 2. Student features' categories distribution based on studies contexts

At last, figure 4 highlights the most common machine learning techniques used (a) , and we can see that Decision Trees and its variants is the most used technique (63%) with a very wide margin compared to other techniques (NB: 35%,NN:35%, SVM: 23%). In figure (b), we notice that Decision Trees technique is always in first level compared to other techniques and it is most efficient in 40% of the research studies, followed by NB and SVM (13% for each) and neural network with 10%.



(a) Distribution of the most used classification algorithms

(b) Distribution of the most performant classification algorithms

Fig. 3.

4 Experiments

From the comparative study we can notice that Decision Tree performs better than others machine learning algorithms in the educational field, but as these latter have been applied to different datasets, it is difficult to confirm these result. Therefore, our objective in this experimental study, is to apply these techniques on two datasets in order to confirm the results of our study by applying Machine Learning algorithms cited on state of art on two datasets from two contexts : TE and EL & TE.

4.1 Background of machine learning techniques used

Decision Tree (DT) is multistage decision making technique able to break down a complex decision making process into collection of simpler decision providing an easier solution to interpret [47]. The most used implementations today are: ID3, C4.5, C50 and CART. ID3 (Iterative Dichotomiser 3) builds the decision tree recursively, at each step of the recursion, it calculates from among the attributes remaining for the current branch, the one that will maximize the information gain using Shanon's entropy [48]. C4.5 is an extension of ID3, to overcome the limitations of ID3 and one of its implementation is J48. CART (Classification and Regression Trees) is a binary tree construction algorithm [49]. **K-Nearest Neighbors** (K-NN) is also used in our comparison process, which is one of the oldest and simplest methods for pattern classification, it is a very intuitive method that classifies untagged examples on the basis of their similarity with the examples of the learning dataset [50]. **Support Vector Machine** (SVM) is a discriminant model that attempts to minimize learning errors while maximizing the margin between class data [51], SVM is particularly effective at handling high-dimensional data. **Neural Networks** (NN) have been developed as generalization of mathematical models of biological nervous systems and they have shown their effectiveness in several fields [52], [53]. Naïve Bayes (NB) is a process that estimates the probability of a new observation belonging to a predefined category [54]. Logistic Regression, is a predictive technique that aims to build a model to predict / explain the values taken by a qualitative target variable (most often binary, we then speak of binary LR; if it has more than 2 modalities, we speak of polytomous LR) from a set of quantitative or qualitative explanatory variables [54].

4.2 Dataset

Two types of dataset with two contexts (traditional education and e-learning & traditional education) were used. Traditional Education dataset (TE) [55], is a two classes dataset (Pass and Fail based on the final grade), consists of 395 student records and 33 features, collected from student achievement in secondary education of two Portuguese schools, and we focused on Mathematics subject to predict student performance. We chose this dataset because it purely reflects traditional education and because it mainly contains the same features' categories that we extracted during our analysis for this type of context, which is based on academic data, student behaviour features such as: raised hand on class, opening resources, answering survey by parents, social data, etc. The second dataset (EL & TE), is a three-class dataset, where students are classified into three classes based on their total grade marks (Low level, Middle level and High level), it contains, in addition to the data from the traditional context, data reflecting the online behaviour of the student, namely visited online resources, discussion groups, etc. This dataset was collected from learning management system (LMS) called Kalboard 360 [28, 56], and consists of 480 student records and 16 features including personal identity, social identity, online behaviour, learning behaviour and especially the important characteristics that we extracted from our analysis for this type of context.

4.3 Process and result

To lead our case study, a well-defined process is adopted as shown in figure 5, to detect the most efficient ML techniques in the classification of each dataset, we proceeded to a feature selection by information gain method is used to keep the important attributes, and a partitioning of the data using the split method (30% for testing and 70% for training) was carried out. The chosen ML techniques were applied to the two types of dataset to see the most efficient technique (with the highest accuracy). The algorithms used are: C4.5, J48, ID3, CART, NB, SVM, NN, LR and K-NN, and the evaluation is done by the accuracy performance metric.

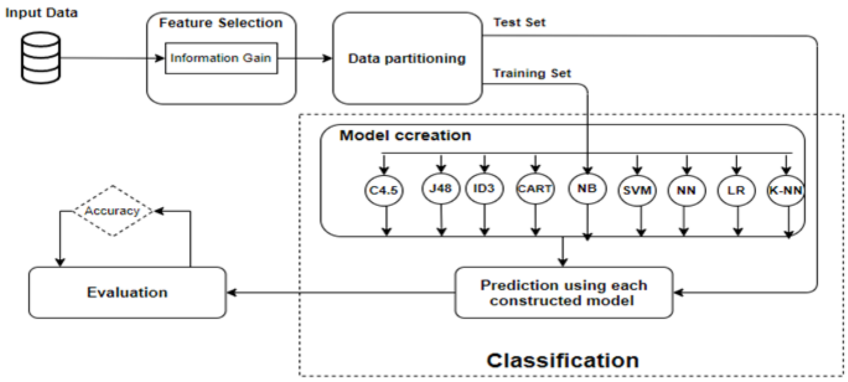


Fig. 4. Comparison methodology

Table 3 shows the performance results (Accuracy) of each algorithm applied on each dataset. From de Table 3, we notice that for TE dataset, the best performing techniques are techniques derived from the Decision Tree algorithm (C4.5, J48 and CART), the algorithm which gave the lowest accuracy is SVM (78.81%). For the EL&TE dataset, neural network gave the best performance, Logistic Regression and C4.5 also gave a good performance around 70%.

Technique ***Data***		**C4.5**	**J48**	**ID3**	**CART**	**NB**	**SVM**	**NN**	**LR**	**K-NN**
A (%)	TE	94.07	93.22	88.98	93.22	83.90	78.81	86.44	87.25	81.36
	EL & TE	70.14	51.39	56.94	63.19	65.97	52.78	77.78	72.92	55.56

We notice that the decision trees with its variants, gave good performance results for the traditional dataset, where there was no online data, and for the other which actually includes this type of data, the neural network was the most efficient, which, according to our survey, was ranked third level for the most used algorithms and fourth level among the algorithms that were used and performed the best.

5 Conclusion

In this study, we present a survey on student profile modeling using machine learning in order to give the most efficient machine learning techniques used and the overall description of the student profile used in different fields. The study shows that the Decision Tree algorithms are the most used and the most efficients among all research studies cited in this comparative study. In addition, the main student features used in the profile modeling was academic information by more 70%, followed by personal identity and online behavior, which shows that the combination between academic and online behaviors when modeling a student profile using machine learning has become more important. An experiment was carried out on two datasets, the results approved that the Decision Tree algorithm gives a good performance for both datasets and especially for traditional education context. In our future works, we attempt to propose a generic student profile model that can be exploited in many situations such as: prediction, classification, adaptive learning, and e-recommendation.

6 References

[1] A. Abyaa, M. K. Idrissi, and S. Bennani, 'Towards an adult learner model in an online learning environment', in 2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET), Istanbul, Turkey, Sep. 2016, pp. 1–5, https://doi.org/10.1109/ithet.2016.7760735.

[2] R. L. S. do Nascimento, R. B. das Neves Junior, M. A. de Almeida Neto, and R. A. de Araújo Fagundes, 'Educational Data Mining: An Application of Regressors in Predicting School Dropout', in Machine Learning and Data Mining in Pattern Recognition, vol. 10935, P. Perner, Ed. Cham: Springer International Publishing, 2018, pp. 246–257. https://doi.org/10.1007/978-3-319-96133-0_19

[3] Y. Nieto, V. Gacia-Diaz, C. Montenegro, C. C. Gonzalez, and R. Gonzalez Crespo, 'Usage of Machine Learning for Strategic Decision Making at Higher Educational Institutions', IEEE Access, vol. 7, pp. 75007–75017, 2019, https://doi.org/10.1109/access.2019.2919343.

[4] S. B. Kotsiantis, 'Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades', Artificial Intelligence Review, vol. 37, no. 4, pp. 331–344, Apr. 2012, https://doi.org/10.1007/s10462-011-9234-x.

[5] N. Sael, T. Hamim and F. Benabbou, 'Multilevel Hybrid System based on machine learning and AHP for student failure prediction', International Journal of Computer and Network Security, vol. 19, no. 9,pp. 103-112, 2019.

[6] A. Topîrceanu and G. Grosseck, 'Decision tree learning used for the classification of student archetypes in online courses', Procedia Computer Science, vol. 112, pp. 51–60, 2017, https://doi.org/10.1016/j.procs.2017.08.021.

[7] S. V. Kolekar, R. M. Pai, and M. P. M M, 'Prediction of Learner's Profile based on Learning Styles in Adaptive E-learning System', International Journal of Emerging Technologies in Learning (iJET), vol. 12, no. 06, p. 31, Jun. 2017, https://doi.org/10.3991/ijet.v12i06.6579.

[8] M. Abdullah, A. Y. Bayahya, E. S. Shammakh, K. A. Altuwairqi, and A. A. Alsaadi, 'A novel adaptive e-learning model matching educator-student learning styles based on machine learning', p. 11.

[9] R. Cerezo, M. Sánchez-Santillán, M. P. Paule-Ruiz, and J. C. Núñez, 'Students' LMS interaction patterns and their relationship with achievement: A case study in higher education', Computers & Education, vol. 96, pp. 42–54, May 2016, https://doi.org/10.1016/j.compedu.2016.02.006.

[10] C. J. Villagrá-Arnedo, F. J. Gallego-Durán, F. Llorens-Largo, P. Compañ-Rosique, R. Satorre-Cuerda, and R. Molina-Carmona, 'Improving the expressiveness of black-box models for predicting student performance', Computers in Human Behavior, vol. 72, pp. 621–631, Jul. 2017, https://doi.org/10.1016/j.chb.2016.09.001.

[11] M. Cantabella, R. Martínez-España, B. Ayuso, J. A. Yáñez, and A. Muñoz, 'Analysis of student behavior in learning management systems through a Big Data framework', Future Generation Computer Systems, vol. 90, pp. 262–272, Jan. 2019, https://doi.org/10.1016/j.future.2018.08.003.

[12] A. S. Hoffait and M. Schyns, 'Early detection of university students with potential difficulties', Decision Support Systems, vol. 101, pp. 1–11, Sep. 2017, https://doi.org/10.1016/j.dss.2017.05.003.

[13] A. U. Khasanah and Harwati, 'A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques', IOP Conference Series: Materials Science and Engineering, vol. 215, p. 012036, Jun. 2017, https://doi.org/10.1088/1757-899x/215/1/012036.

[14] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, 'Predicting at-risk university students in a virtual learning environment via a machine learning algorithm', Computers in Human Behavior, Jun. 2018, https://doi.org/10.1016/j.chb.2018.06.032.

[15] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, 'Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses', Computers in Human Behavior, vol. 73, pp. 247–256, Aug. 2017, https://doi.org/10.1016/j.chb.2017.01.047.

[16] M. Khalil and M. Ebner, 'Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories', Journal of Computing in Higher Education, vol. 29, no. 1, pp. 114–132, Apr. 2017, https://doi.org/10.1007/s12528-016-9126-9.

[17] C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, 'Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout', Computers & Electrical Engineering, vol. 66, pp. 541–556, Feb. 2018, https://doi.org/10.1016/j.compeleceng.2017.03.005.

[18] J. Song, Y. Zhang, K. Duan, M. Shamim Hossain, and S. M. M. Rahman, 'TOLA: Topic-oriented learning assistance based on cyber-physical system and big data', Future Generation Computer Systems, vol. 75, pp. 200–205, Oct. 2017, https://doi.org/10.1016/j.future.2016.05.040.

[19] S. Kai, J. M. L. Andres, L. Paquette, R. S. Baker, K. Molnar, and M. Moore, 'Predicting Student Retention from Behavior in an Online Orientation Course', p. 6.

[20] J. Y. Chung and S. Lee, 'Dropout early warning systems for high school students using machine learning', Children and Youth Services Review, vol. 96, pp. 346–353, Jan. 2019, https://doi.org/10.1016/j.childyouth.2018.11.030.

[21] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, 'Predicting Student Dropout in Higher Education', arXiv:1606.06364 [cs, stat], Jun. 2016, Accessed: Jun. 30, 2019. [Online]. Available: http://arxiv.org/abs/1606.06364.

[22] S. Sivakumar, S. Venkataraman, and R. Selvaraj, 'Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree', Indian Journal of Science and Technology, vol. 9, no. 4, Jan. 2016, https://doi.org/10.17485/ijst/2016/v9i4/87032.

[23] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, 'Early dropout prediction using data mining: a case study with high school students', Expert Systems, vol. 33, no. 1, pp. 107–124, Feb. 2016, https://doi.org/10.1111/exsy.12135.

[24] R. K. Jena, 'Predicting students' learning style using learning analytics: a case study of business management students from India', Behaviour & Information Technology, vol. 37, no. 10–11, pp. 978–992, Nov. 2018, https://doi.org/10.1080/0144929x.2018.1482369.

[25] K. Liang, Y. Zhang, Y. He, Y. Zhou, W. Tan, and X. Li, 'Online Behavior Analysis-Based Student Profile for Intelligent E-Learning', Journal of Electrical and Computer Engineering, vol. 2017, pp. 1–7, 2017, https://doi.org/10.1155/2017/9720396.

[26] P. Dash, Department of Computer Science, Christ University, Hosur Road, Bengaluru – 560029, Karnataka, India;, V. Vaidhehi, and Department of Computer Science, Christ University, Hosur Road, Bengaluru – 560029, Karnataka, India;, 'Enhanced Elective Subject Selection for ICSE School Students using Machine Learning Algorithms', Indian

Journal of Science and Technology, vol. 10, no. 21, pp. 1–10, Feb. 2017, https://doi.org/10.
17485/ijst/2017/v10i21/109551.

[27] S. Bharara, S. Sabitha, and A. Bansal, 'Application of learning analytics using clustering data Mining for Students' disposition analysis', Education and Information Technologies, vol. 23, no. 2, pp. 957–984, Mar. 2018, https://doi.org/10.1007/s10639-017-9645-7.

[28] E. A. Amrieh, T. Hamtini, and I. Aljarah, 'Mining Educational Data to Predict Student's academic Performance using Ensemble Methods', International Journal of Database Theory and Application, vol. 9, no. 8, pp. 119–136, Aug. 2016, https://doi.org/10.14257/ijdta.2016.9.8.13.

[29] King Abdulaziz University, Saudi Arabia, Jeddah, A. Mueen, B. Zafar, and U. Manzoor, 'Modeling and Predicting Students' Academic Performance Using Data Mining Techniques', International Journal of Modern Education and Computer Science, vol. 8, no. 11, pp. 36–42, Nov. 2016, https://doi.org/10.5815/ijmecs.2016.11.05.

[30] Y. Mastoory, S. R. Harandi, and N. Abdolvand, 'The Effects of Communication Networks on Students' Academic Performance: The Synthetic Approach of Social Network Analysis and Data Mining for Education', International Journal on Integrating Technology in Education, vol. 5, no. 4, pp. 23–34, Dec. 2016, https://doi.org/10.5121/ijite.2016.5403.

[31] A. El-Halees, 'Mining students data to analyze learning behavior: A case study', p. 4.

[32] R. Hasan, S. Palaniappan, A. R. A. Raziff, S. Mahmood, and K. U. Sarker, 'Student Academic Performance Prediction by using Decision Tree Algorithm', in 2018 4th International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Aug. 2018, pp. 1–5, https://doi.org/10.1109/iccoins.2018.8510600.

[33] T. Mahboob, S. Irfan, and A. Karamat, 'A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms', in 2016 19th International Multi-Topic Conference (INMIC), Islamabad, Pakistan, Dec. 2016, pp. 1–8, https://doi.org/10.1109/inmic.2016.7840094.

[34] C. Masci, G. Johnes, and T. Agasisti, 'Student and school performance across countries: A machine learning approach', European Journal of Operational Research, vol. 269, no. 3, pp. 1072–1085, Sep. 2018, https://doi.org/10.1016/j.ejor.2018.02.031.

[35] S. Patil and S. Kulkarni, 'Mining Social Media Data for Understanding Students' Learning Experiences using Memetic algorithm', Materials Today: Proceedings, vol. 5, no. 1, pp. 693–699, 2018, https://doi.org/10.1016/j.matpr.2017.11.135.

[36] S. E. Sorour and T. Mine, 'Building an Interpretable Model of Predicting Student Performance Using Comment Data Mining', in 2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI), Kumamoto, Japan, Jul. 2016, pp. 285–291, https://doi.org/10.1109/iiai-aai.2016.114.

[37] A. I. Adekitan and E. Noma-Osaghae, 'Data mining approach to predicting the performance of first year student in a university using the admission requirements', Education and Information Technologies, vol. 24, no. 2, pp. 1527–1543, Mar. 2019, https://doi.org/10.1007/s10639-018-9839-7.

[38] A. Abu, 'Educational Data Mining & Students' Performance Prediction', International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, 2016, https://doi.org/10.14569/ijacsa.2016.070531.

[39] S. Pal and V. Chaurasia, 'Is Alcohol Affect Higher Education Students Performance: Searching and Predicting Pattern Using Data Mining Algorithms', SSRN Electronic Journal, 2017, https://doi.org/10.2139/ssrn.2991214.

[40] epartment of Computer Science, Government College of Arts and Science, Karamakudi, India, K. Karthikeyan, P. Kavipriya, and Department of computer Science, Sri Krishna Arts and Science College, Coimbatore, India, 'On Improving Student Performance Prediction in Education Systems using Enhanced Data Mining Techniques', International Journal

of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 5, pp. 935–941, May 2017, https://doi.org/10.23956/ijarcsse/sv7i5/0348.

[41] D. B. Fernandez and S. Lujan-Mora, 'Comparison of applications for educational data mining in Engineering Education', in 2017 IEEE World Engineering Education Conference (EDUNINE), Santos, Brazil, Mar. 2017, pp. 81–85, https://doi.org/10.1109/edunine.2017.7918187.

[42] J. Kong, J. Han, J. Ding, H. Xia, and X. Han, 'Analysis of students' learning and psychological features by contrast frequent patterns mining on academic performance', Neural Computing and Applications, Oct. 2018, https://doi.org/10.1007/s00521-018-3802-9.

[43] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, 'Analyzing undergraduate students' performance using educational data mining', Computers & Education, vol. 113, pp. 177–194, Oct. 2017, https://doi.org/10.1016/j.compedu.2017.05.007.

[44] N. Tanwani, S. Kumar, A. H. Jalbani, S. Soomro, M. I. Channa, and Z. Nizamani, 'Student opinion mining regarding educational system using facebook group', in 2017 First International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT), Karachi, Nov. 2017, pp. 1–5, https://doi.org/10.1109/intellect.2017.8277622.

[45] S. K. Howard, J. Ma, and J. Yang, 'Student rules: Exploring patterns of students' computer-efficacy and engagement with digital technologies in learning', Computers & Education, vol. 101, pp. 29–42, Oct. 2016, https://doi.org/10.1016/j.compedu.2016.05.008.

[46] T. Hamim, F. Benabbou, and N. Sael, 'Toward a Generic Student Profile Model', in The Proceedings of the Third International Conference on Smart City Applications. Springer, Cham 2019, pp. 200–214. https://doi.org/10.1145/3368756.3369075

[47] V. K. Ayyadevara, 'Decision Tree', in Pro Machine Learning Algorithms, Berkeley, CA: Apress, 2018, pp. 71–103. https://doi.org/10.1007/978-1-4842-3564-5_4

[48] J. R. Quinlan, 'Induction of decision trees', Machine Learning, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[49] N. Patil, R. Lathi, and V. Chitre, 'Comparison of C5.0 & CART Classification algorithms using pruning technique', International Journal of Engineering Research, vol. 1, no. 4, p. 6, 2012.

[50] S. Sun and R. Huang, 'An adaptive k-nearest neighbor algorithm', in 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, China, Aug. 2010, pp. 91–94, https://doi.org/10.1109/fskd.2010.5569740.

[51] T.-K. Wu, S.-C. Huang, and Y.-R. Meng, 'Identifying and Diagnosing Students with Learning Disabilities Using ANN and SVM', p. 8.

[52] P. H. Sydenham and R. Thorn, Eds., Handbook of measuring system design. Chichester, England: Wiley, 2005.

[53] M. D. Calvo-Flores, E. G. Galindo, M. C. P. Jiménez, and O. Pérez, 'Predicting students' marks from Moodle logs using neural network models', p. 5, 2006.

[54] P. Tsangaratos and I. Ilia, 'Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size', CATENA, vol. 145, pp. 164–179, Oct. 2016, https://doi.org/10.1016/j.catena.2016.06.004.

[55] P. Cortez and A. Silva, 'Using data mining to predict secondary school student performance', p. 9.

[56] E. A. Amrieh, T. Hamtini, and I. Aljarah, 'Preprocessing and analyzing educational data set using X-API for improving student's performance', in 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), Amman, Jordan, Nov. 2015, pp. 1–5, https://doi.org/10.1109/aeect.2015.7360581.

7 Authors

Touria Hamim received the engineer degree in Information and Communication Systems Engineering from ENSAJ, Morocco, in 2017. Currently she is preparing her Ph.D. in Computer Science in Faculty of Science Ben M'sik. Her research interests profile modeling using machine learning. Email: hamimxtouria@gmail.com

Faouzia Benabbou is a professor of Computer Science and member of Compute Science and Information Processing laboratory. She is Head of the team "Cloud Computing, Network and Systems Engineering (CCNSE)". She received his Ph.D. in Computer Science from the Faculty of Sciences, University Mohamed V, Morocco, 1997. His research areas include cloud Computing, data mining, machine learning, and Natural Language Processing. Email: faouzia.benabbou@univh2c.ma

Nawal Sael is a professor of Computer Science and member of Computer Science and Information Processing laboratory at faculty of science Ben M'sik (Casablanca, Morocco). She received her Ph.D. in Computer Science from the Faculty of Sciences, University Hassan II Casablanca, Morocco, 2013 and her engineer degree in software engineering from ENSIAS, Morocco, in 2002. Here research interests include data mining, educational data mining, machine learning, deep learning and Internet of things. Email: saelnawal@hotmail.com

Article submitted 2020-09-17. Resubmitted 2020-11-04. Final acceptance 2020-11-05. Final version published as submitted by the authors.