An in-depth methodology to predict at-risk learners

Amal Ben Soussia, Azim Roussanaly, Anne Boyer

An in-depth methodology to predict at-risk learners

Amal BEN SOUSSIA[1], Azim ROUSSANALY[1], and Anne BOYER[1]

Lorraine University, LORIA, Nancy, France

Abstract. Nowadays, the concept of education for all is gaining momentum thanks to the widespread use of e-learning systems around the world. The use of e-learning systems consists in providing learning content via the Internet to physically dispersed learners. The main challenge in this regard is the high fail rate particularly among k-12 learners who are our case study. Therefore, we established an in-depth methodology based on machine learning models whose objectives are the early prediction of at-risk learners and the diagnosis of learning problems. Going through this methodology was of a great importance thus it started by identifying the most relevant learning indicators among performance, engagement, regularity and reactivity. This, then, led us to extract and select the adequate learning features that reflect the activity of an online learner. For the modeling part of this methodology, we apply machine learning models among k-nearest neighbors (K-nn), Support Vector Machine (SVM), Random Forest and Decision tree on a real data sample of 1361 k-12 learners. The evaluation step consists in comparing the ability of each model to correctly identify the class of learners at-risk of failure using both accuracy and False Positive Rate (FPR) measures.

Keywords: At-risk learners . Early prediction . Methodology . Learning indicators. Machine learning . Evaluation

1 General introduction

Many educational institutions are now opting for e-learning by offering their courses through their own private online Learning Management Systems (LMS). While adopting a technology-driven approach allows these institutions to maintain their competitiveness, it comes with many challenges. Indeed, the main issues detected in e-learning environments are the high number of no-shows, early dropouts and low completion rates which lead to a total failure of the learner [12]. In this paper, we are interested in systems designed for teachers to help them detecting the potential learning difficulties.
In the context of a fully distance learning institution, data is generally multi-source as we have more than one application, which may provide us with informative, heterogeneous and different types of data. The heterogeneity of data is explained by having administrative data describing the demographics of learners profiles, traces of use and interaction between learners and the learning environment as well as data about the academic performance and assessments. These

learning applications provide a time-independent data that is stable over time or a time-dependent data type that is evolutive over time. Given the volume and diversity of data, teachers are no longer able to assist all their learners at the same time with a pedagogical follow-up adapted to the situation of each of them. Therefore, teachers need a summary of how each learner's experience unfolds through four learning indicators: performance, engagement, regularity and reactivity. Each indicator is represented by features extracted and computed from learning data sources. The identification of these learning indicators has more than one intention. They are useful for the prediction of at-risk learners as well as for the diagnosis of learning gaps of each learner.
In this paper, we propose an in-depth methodology that exploits the numeric traces generated by learning applications. This methodology is based on machine learning (ML) models whose objective is the early and accurate prediction of learners at-risk of failure. The depth of this methodology comes from the fact that we first started with the identification of the most relevant learning indicators among performance, engagement, regularity and reactivity. Second, and based on these indicators, we were able to extract and select the adequate features representing the activity of an online learner. The last parts of this methodology are for modeling and evaluation. Using the False Positive Rate (FPR) measure, we concluded on the best ML model that correctly predict the class of at-risk learners. For this end, we build a real data sample of 1361 k-12 learners following the same module. We identify the learning indicators and extract features from two available applications. Then, we follow a weekly prediction approach and formalize the problem into a 3-class classification problem: success, medium risk of failure and high risk of failure. The trained and tested ML models in this paper are: k-nearest neighbours (k-nn), Support Vector Machine (SVM), Random forest and Decision tree. These models are the most used in literature and show a good predictive performance. Several techniques of filtering, wrapper and embedded methods for feature selection are applied. The techniques of filtering and wrapper methods give a very promising result. Also, the FPR evolution confirm that the Decision tree model has a good ability to predict at-risk learners on the first prediction weeks.
The paper is organized as follows. Section 2 presents the state of art projects related to our work. Section 3 explains the proposed methodology. Section 4 introduces the application of the methodology in our case study. Section 5 explains the experiments and the results. Section 6 concludes on the study.

2 Related work

The high dropout and failure rates registered in k-12 are rarely discussed in the literature especially when it comes to online education and when learners are in total autonomy. One of the main solutions to reduce failure is to predict correctly and at the earliest at-risk learners. Therefore, studies that are interested in solving this problem start generally by proposing working methodologies and frameworks. The major common point between all strategies and method-

ologies proposed is the importance of work done on the collection, extraction, engineering and selection of features to aliment the machine learning models. [9] proposes an integrated framework to predict the dropout in MOOCs. This framework includes three main steps: feature generation, feature selection and dropout prediction. They used an ensemble feature selection method as it does not depend on a specific learning algorithm for feature scoring. [5] proposes an analytics framework for Moodle that abstracts out the most relevant elements of prediction models. This framework goes through the steps of analysing raw data and dividing it into features and target variable, modelisation and insights given by some predictions about the learners potential difficulties. These later projects propose solutions for one specific online learning context which are Moodle and MOOCs respectively. In addition, they extract data from one application. The methodology we propose shows the importance of going through almost the same processes and phases but it is more general and emphasizes the use of heterogeneous and multi-source data. Other studies of the field focus more on the relevant and effectiveness of data for the prediction of at-risk learners in the context of online education. [7] reviews on the most used data types to discover at-risk students. The learning behaviour data including number of logs into the course, number of views, clicks and downloads, the time spent on teaching materials... is in the top list of the most used data. Learning network data such as the number of forums discussions posts, replies and comments is a very used data type in the state of the art. The third data type is related to the learning level. It includes data about tests and grades. One other used type is the learning emotional data which includes non-cognitive assessment, self-efficacy and self assessed level. Other common used data is related to learners demographics and characteristics. The Open University (OU) records also a high dropout rate. In order to solve this problem, the OU project interests in detecting as early as possible the students who are likely to dropout by identifying the less engaged ones at an early stage of a course [1]. In addition to demographic data, the models used features expressing the engagement of a learner and his interaction with the VLE [11]. Student Success System (S3) is an analytical system based on ensemble models to identify and treat at-risk students[3]. S3 is based on the calculation of a generic measure called the success index composed of five indicators: attendance, participation, preparation, completion and social learning[2]. A first step in the approach is developing basic models to predict each indicator. Thus, simple logistic regression was used for the prediction of presence while social network analysis (SNA) is more appropriate for the index of social learning. The methodology we propose and apply shows also the importance of the performance and engagement learning indicators, gives a new definition of the regularity indicator and defines the reactivity indicator. These later indicators are important to follow the learning rhythm of in total autonomy learner.

3 Methodology

Data is the fuel of ML projects and is the start point of the methodology we propose. As shown in Fig. 1, the first step in a learning analytics project is the collection of the different learning traces from the available data sources and cleaning it. This first phase allows us having the raw data ready to be used and analyzed in the next phase, which is feature extraction. The overall goal of the feature extraction is to prepare a new dataset composed of a set of computed features representing learning indicators. The third step is modeling using ML algorithms. Each model takes as input the set of features previously computed and gives as output the predicted class of each learner. The prediction results are then evaluated according to one general measure, which is the accuracy. Based on these results, we select for next experiments only algorithms with the highest prediction accuracy. Then, we go through the feature selection process to identify the most relevant features as well as learning indicators to predict the learner class with no accuracy degradation. The selected features are the input for the second modeling phase. To finally evaluate the ability of models to predict learners at-risk of failure, we use the FPR measure.

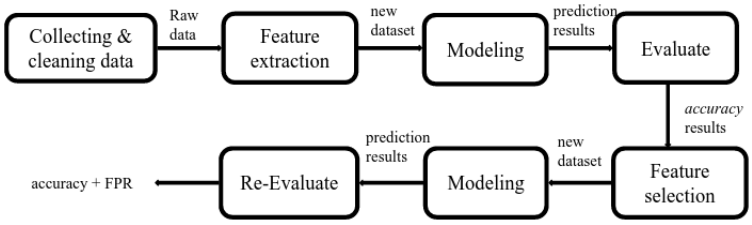

Fig. 1. The in-depth methodology phases

3.1 Feature extraction

A feature is a representation of raw data [4]. Input features are the most important factor for ML models. Therefore, feature extraction is a central task in every ML project workflow as illustrated in Fig. 1. The idea is to define a learner activity through learning indicators. First of all, we identify these indicators which are the basis when extracting features from raw data. An indicator is an observable that is pedagogically significant, computed or established with the help of observations, and testifying to the quality of interaction, activity and learning. It is defined according to an observation objective and motivated by an educational objective. According to this definition, each learning indicator is defined by a subset of features. The identification of the above learning indicators was established based on a deep study of the behavioural profile of an at-risk learner given by the education sciences as well as on the pertinent

results of multiple state-of-the-art projects which are interested in solving this issue. In addition, we have taken into consideration the specificity of distance learning available tools and data and the particular characteristics of an in total autonomy learner:

- performance: it is a very used learning indicator. It represents all features related to marks and exams that are highly correlated with a learner final result [8] [1] [3].
- engagement: it reflects the involvement of a learner toward his work. Features related to learners participation in the online platforms are frequently used in literature [11].
- regularity: in the state of the art, the learning regularity was proven to be highly correlated with the prediction of learners final results. Regularity can be defined in two domains: actions and time, or a combination of both. Regularity in actions is repeating patterns in user's actions sequence. Regularity in time corresponds to repeating patterns in timing of study sessions. Regularity in the combined domains is reflected by the dependencies between action types and their occurrence time [10]. As it is important to follow the regular progress made by an in total autonomy learner, we introduce the regularity of progress.
- reactivity: as far as we know, reactivity has not been used in the literature as a learning indicator. In fact, unlike face to face education, each online learner has its own learning rhythm. Reactivity in the context of an online learning corresponds to the time required to become active in the LMS and to respecting deadlines for exams submissions. This indicator serves to analyze the learner behaviour and to compare it to those of his peers.

For each indicator, we extract features from raw data. To obtain such features, we go through multiple computations of raw data such as composition and combinations.

3.2 Feature selection

ML models need relevant features to give accurate results. However, a high dimension set of input features could contain noisy, redundant and irrelevant data. Such a data weakens the predictive performance of the model, causes overfitting and increases the error rate. To handle this issue, the feature selection process aims at selecting a subset of relevant features from the initial set based on redundancy and relevance [13]. To this end, several techniques are used in classification problems that fall into three categories:

- Filtering: based on statistical tests, the model selects from the initial set a k-dimension subset of the most correlated features with the target variable[13].
- Wrapper methods: features subset is selected based on inductive algorithms.
- Embedded methods: they aim at selecting the best features during the training phase [13]. The embedded feature selection could use two methods: Regularization and tree-based methods.

3.3 Approach

For the early prediction of at-risk learners, the problem is generally formalized into a n-class classification problem. The classes of learners are usually identified based on the required results fixed by the grading system of each teaching institution. Depending on the needs of each project as well as on the frequency of learners activity follow-up required by teachers, we choose a period of time after which we make a regular prediction. To represent the activity of a learner during this learning period p_i, all features of learning indicators are grouped in the same vector X. Thus, on each prediction time p_i, a learner is represented by a vector X composed of features going from f_1 to f_n and the class y to which he belongs to. Each learner belongs to one and only class over the year.

$$X = < f_1, f_2, ..., f_n, y >$$

Each feature f_1 to f_n represents one learning activity till the prediction time p_i. For each prediction time p_i, the value of one feature is added to that of prediction time p_{i-1}: we proceed to an accumulation of values.

4 Case study: CNED

4.1 CNED presentation

The CNED[1] is the french largest national center for distance education. It offers multiple and fully distance courses to a very large number of physically dispersed learners. These learners are from different demographic profiles and cannot go to traditional schools for multiple reasons. Each learner is unique, in total autonomy and follows his own learning rhythm and schedule. The only information we have about him are the exams he submits and the traces of his activity within the LMS. Learning is also quite specific and provided through more than one application. It is multi-modal as the courses contents are available online and in printed papers. Moreover, by relying on traditional teaching methods, teachers monitor the progress of a large number of heterogeneous learners (up to thousands of learners) at the same time. These methods are no longer effective and teachers need help as well as new techniques and tools, which allow them a better tracking of learners performance and an early detection of their potential learning difficulties. In fact, CNED records among its k-12 learners a high failure rate every year. K-12 learners are the main focus of this study.

4.2 Data description

In this project, learning traces are collected from two data sources. The first one is the LMS, which generates the interaction traces between learners and learning environment. This data is related to learners actions within the platform and their use of its different components. The second one is the students

[1] Centre National d'Enseignement à Distance created in 1939

administrative management application GAEL [2]. This application provides two types of data. The first data type is demographic such as gender, age, native country, place of birth, city of residence, having or not a scholarship, repeating or not the year. The second type of data is related to modules, exams and their submission dates, marks, and correctors. The k-12 learners enrolled in the physical-chemistry module during the school years 2017-2018 and 2018-2019 are the case study of this paper. The school year starts on September 1 and ends on July 7. It is composed of 44 weeks. As the registration in CNED is open during the year, the start activity date t_0 of each learner is defined as the maximum date between the start school year date and the registration date. Depending on t_0, learners don't have the same number of activity weeks. In addition, study programs for learners who register after October 31 of each year go through adjustments. In this project, we focus on learners who subscribed before October 31. According this information, we collect the learning traces of 663 and 698 learners respectively from 2017-2018 and 2018-2019. All learners of 2017-2018 and 2018-2019 have respectively 37 and 35 activity weeks. From these two dates, we have a decrease in the number of learners per activity week.

4.3 Feature extraction

In the context of CNED, in addition to the demographic data provided by GAEL, the activity of a learner is represented by the four learning indicators introduced in the section 3.1. Based on the available and extracted features from both data sources, these indicators are defined as follows:

- performance: grades and exams are the current criteria for CNED tutors to evaluate their learners. The performance of a learner is represented by 3 features. These features are about the academic assessments and grades. They are evolutive over time.
- engagement: as CNED learners are in total autonomy, the only way to track their engagement is the online presence. In addition, CNED teachers push especially the k-12 learners to be more active on the LMS. The engagement is represented by 36 features. These features are time-dependent and are about the learner's use of the LMS and his interaction with its components.
- regularity: it is defined by the progress made by a learner in terms of number of actions within the LMS and number of submitted exams. The regularity is represented by 2 features. These features are also evolutive over time.
- reactivity: it is represented by features about the reactivity of a learner to submit an exam or to connect to the online course. The reactivity is represented by 7 features. These features are time-dependent, evolve over time and are computed based on the exams schedule calendar.

Thus, each learner is defined by 10 demographic features (which are constant) and learning indicators represented by the extracted features. In total, each learner is defined by 58 features.

[2] Gestion Administrative des ÉLèves

4.4 Application of the approach

CNED teachers need to have a regular and frequent tracking of their learners' activity. Therefore, the temporal granularity chosen here is the activity week as the period of time to apply the approach. This makes it possible to predict, for the context of CNED, learners in learning difficulties on a weekly basis and to compare their reliability over time. The prediction weeks, for each learner, depend on his start activity day t_0. More explicitly, the first prediction week of one learner is $w_1 = t_0 + 7days$, the second prediction week is $w_2 = w_1 + 7days$ and so on until w_n corresponds to the school year end date. With the exception of demographic data, which is of course time-independent, the rest of extracted learning features are therefore weekly updated. Demographic and features of learning indicators are grouped together in the same vector X to represent the weekly activity of a learner. The French system allows teachers to give marks between 0 and 20. The average of 10 in a module generally determines the success or failure of a learner. However, it is of great importance to have more focus on learners in the uncertainty zone with an average between 8 and 12. Therefore, for each module, learners are classified into three classes based on the obtained marks average by the end of the school year:

- success: when the marks average is superior to 12
- medium risk of failure: when the marks average is between 8 and 12
- high risk of failure learner: when the marks average is inferior to 8

The tab. 1 gives the number of learners belonging to each of the three classes during each of the school years. As the majority of state of art projects, most of learners are classified as successful.
The experimental part will focus on comparing the prediction performance of the following supervised machine learning models: Random Forest, Decision tree, K-nn and SVM. These models are frequently used and show good prediction results in the majority of the state of art projects [11] [7].

Table 1. Number of learners per class.

Learner class / School year	Success	Medium risk	High risk
2017-2018	488	111	64
2018-2019	538	101	59

5 Methodology implementation and results

5.1 Experimental protocol

The models are tested with 5-fold cross validation and have as input features those of the vector X. To evaluate the performance of the ML models to give

an output y_{pred} similar to the y_{test}, we followed a two-step method allowing the identification of the models with the best accuracy. First, we randomly select 80% of learners vectors from the 2017-2018 school year to train the models and use the remaining 20% for the test phase. Then, to be sure of the first obtained results, we train the models with the school year $n-1$ (2017-2018) learners vectors and test them with those of the school year n (2018-2019).

5.2 Accuracy results

Comparing the accuracy curves of Fig. 2, SVM, Random Forest and Decision tree are the most performing models and keep an increasing accuracy evolution. K-nn has not stable results throughout the school year. On the first prediction week, the accuracies of SVM, Random forest and Decision tree were respectively 0.729, 0.706 and 0.639. The highest accuracies obtained by SVM, Random forest and Decision tree were respectively on week 32, 36 and 35. The results of the

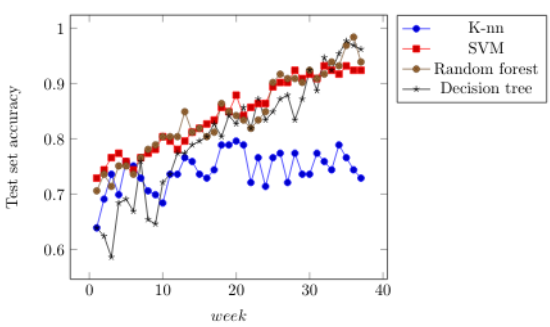

Fig. 2. First step of the accuracy evaluation

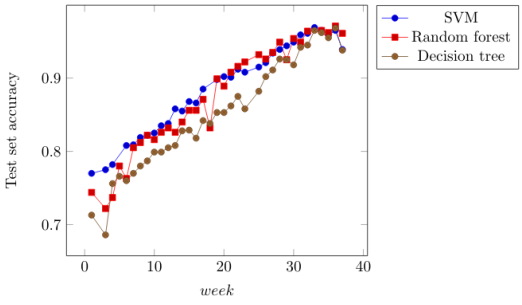

Fig. 3. Second step of the accuracy evaluation

Table 2. Models accuracy when using only demographic features.

Evaluation step \ Model	Random Forest	Decision tree	SVM
First step results	0.608	0.655	0.756
Second step results	0.713	0.614	0.770

second step of the experimental protocol presented in section 5.1 are shown in Fig. 3. Indeed, SVM, Random forest and Decision tree keep a high and increasing accuracy during the prediction dates. The selected models are pertinent.
The tab. 2 presents the models prediction accuracy when only using as input features the demographic ones. The first and second rows of the table are respectively the results of applying the two-step method of section 5.1. These results

show that the prediction performance of the algorithms during the early dates given in Fig. 2 and Fig. 3 is due to demographic features. This makes sense since during the first weeks we do not have enough data about learner's activity.
We want to gain in dimension, computing time and why not in accuracy. For these reasons, we proceed to the feature selection process.

5.3 Feature selection process

For this study, we follow the accuracy curves of Fig. 2. Then, we apply, for each model, the feature selection techniques with learners vectors that give the maximum accuracy. For the next experiments, we train and test models with the 2017-2018 learners vectors.

Filtering methods. Two statistical tests are applied as filtering methods: Chi-square and ANOVA[13]. We set the number of features to be selected to $k = 20$ which is the optimum value for k. We applied these two tests, on the same data, with other values for k that did not give better accuracy results. Therefore, on every prediction week, each learner is now represented by a vector X of 20 features and his success/risk class y. Most of the selected features by both tests belong to the engagement indicator. These features are about the learner activity within the LMS. Features related to the performance indicator such as number of exams and marks obtained up to the prediction week w_i are selected by both tests. ANOVA test selects features of regularity such as the progress in number of actions and submitted exams made by a learner comparing to the previous prediction date. The demographic features selected by Chi-square test are country of residence and city and those selected by ANOVA are having or not a scholarship and repeating or not the year. Applying the Chi-square and ANOVA tests, there is no degradation in the models accuracy. The curves shapes of Fig. 4 and Fig. 5 keep the same properties of those of Fig. 2 but with a faster accuracy evolution especially during the first prediction weeks. On some prediction weeks, the input features selected with the ANOVA test seems to give a higher prediction accuracy of y_{test} than with the Chi-square test. The selected learning features by the ANOVA test are relevant and independent of algorithms.

Wrapper methods. The Recursive Feature Elimination with Cross Validation (RFECV)[6] is used here as a technique of wrapper methods. The number of features selected by RFECV technique with SVM, Random forest and Decision tree are respectively 13, 10 and 29. With the three models (SVM, Random forest and Decision tree), RFECV selects features indicating the engagement of a learner such as the amount of logs. Features of the learner performance given by the number of submitted exams and marks are always selected. With the three models, features expressing the reactivity of a learner such as the number of days between the start of the activity date and the first connection to the LMS date are selected to confirm their high correlation with the prediction of y_{test}. Features of regularity of the progress are selected with SVM and Decision tree. As for demographic features, place of birth and city are selected by the three models. The feature concerning having or not a scholarship is selected by Random forest and Decision tree. Age and gender are selected by SVM and Decision

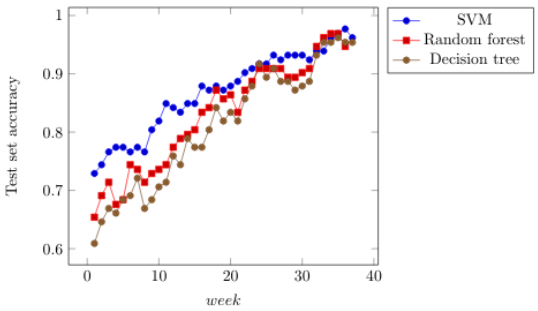

Fig. 4. Accuracy evolution-chi-square-

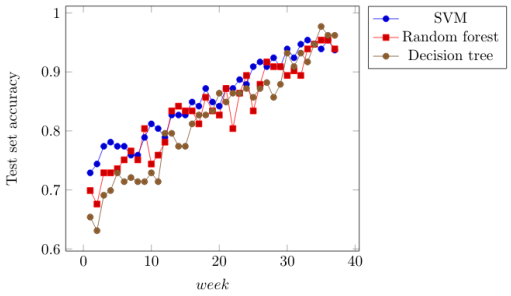

Fig. 5. Accuracy evolution- Anova-

tree. The curves shapes of Fig. 6 have the same properties of those of Fig. 2. In fact, there is no degradation in accuracy and the input features selected by the RFECV method seems to have a better impact on the prediction accuracy of the three models. The curves of the Fig. 6 have a faster accuracy evolution on the first prediction weeks. The RFECV technique gives good results with the three models. The selected features by the RFECV technique are relevant but are dependent to the models.

Embedded methods. Due to their powerful structure, tree-based algorithms,

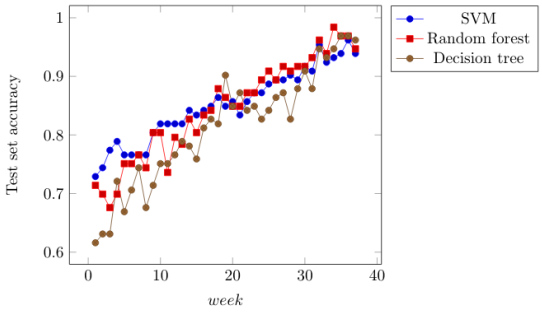

Fig. 6. The accuracy evolution-RFECV-

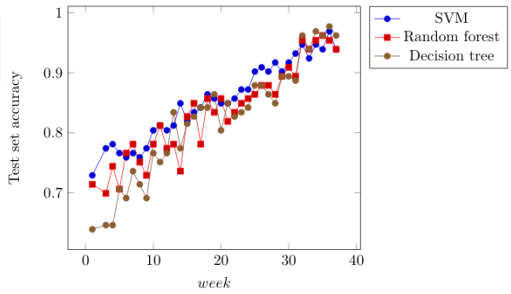

Fig. 7. The accuracy evolution-embedded method-

have the feature importance hyperparameter that serves to select the most important features to make an accurate prediction. In this experiment, the Random Forest is the tree-based algorithm used for feature selection. To train SVM, Random forest and Decision tree and test their prediction accuracy, the tree-based algorithm selects 8 relevant features. The main selected features in this case are about the learner performance given by marks and average. Features of the reactivity of a learner to connect to the LMS are selected too. The demographic features selected are city, place of birth and age. The curves of Fig. 7 still have

the same properties as those of Fig. 2. In fact, there is no degradation in accuracy results. The selected features here are also dependent of the tree-based model and have generally better accuracy with models of the same category. Features of the performance indicator are selected by all the feature selection techniques. Features about the engagement of a learner are selected by the ANOVA, Chi-square and RFECV techniques. Features expressing the regularity of progress are selected by the ANOVA and RFECV techniques. Features of the reactivity indicator are selected by the wrapper and embedded methods.

5.4 FPR results

Classification performance without focussing on a class is the most general way of comparing algorithms. Thus, the accuracy measure does not distinguish between the number of correct labels of different classes. Therefore, opting for a more specific performance metric is necessary to identify the model which correctly predicts learners at medium and high risk of failure (at-risk learners). It is with these learners that the educational interventions will take place. The aim is to minimize at-risk learners classified as successful. To this end, we propose to track the evolution of FPR measure during the learning period given by:

$$FPR = \frac{FP}{FP+TN}$$

The lower FPR is, the more the model is qualified to have a significant ability to predict at risk learners. The Fig. 8, Fig. 9, Fig. 10 and Fig. 11 show that SVM is the algorithm with the highest FPR during the first prediction dates. Despite having the highest overall accuracy, SVM doesn't correctly predict at-risk learners on the first prediction dates. Decision tree is the algorithm with the lowest FPR during the first weeks. Decison tree shows a better ability to correctly predict at-risk learners.

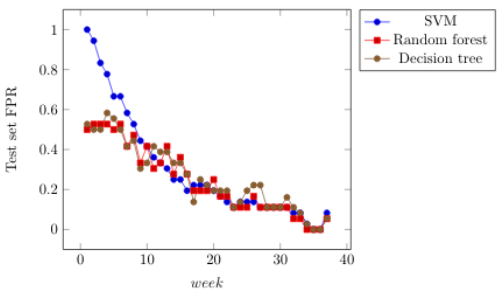

Fig. 8. The FPR evolution-Chi-square-

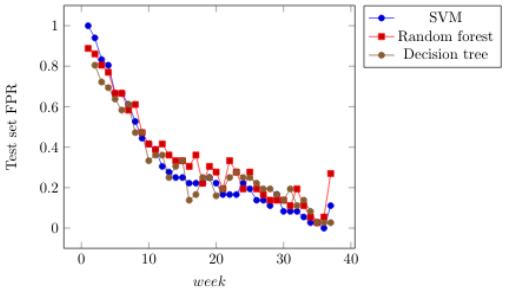

Fig. 9. The FPR evolution-ANOVA-

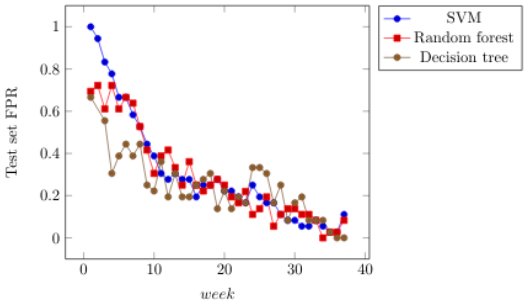

Fig. 10. The FPR evolution-RFECV-

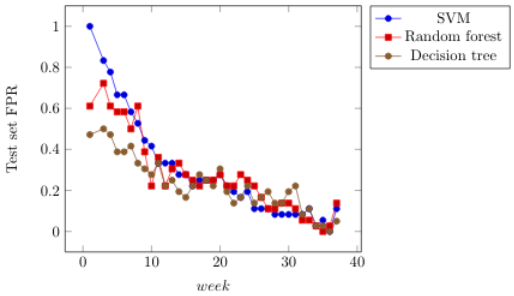

Fig. 11. The FPR evolution-embeded method-

5.5 Results analysis

Going through the feature selection process allows gaining in dimension and keeping a high prediction accuracy of models. In addition, it shows the pertinence of the identified learning indicators particularly the performance and engagement ones. These indicators serve for the diagnosis of learning problems. Some techniques are related to algorithms and others are independent of algorithms. The ANOVA test selects features which are correlated with the target variable independently from the models. SVM is the model with the highest accuracy and the highest FPR during the first weeks. These results come from the fact that this model predicts very well successful learners. On the other hand, Decision tree is the algorithm with the lowest accuracy and lowest FPR during the first weeks. Decision tree is the best to predict at-risk learners during the first weeks. Random forest performs slightly worse than decision tree in predicting at-risk learners but is still much better than SVM. From week 10, all algorithms show almost the same accuracy and FPR values. From week 20, we predict with the minimum of error the at-risk learners.

6 Conclusion

The early prediction of students with learning difficulties is one of the most popular studies in the literature. However, this issue is less discussed when it comes to k-12 online and in total autonomy learners. The CNED is not an exception and records a high failure rate every year. Thus, it aims at providing its instructors with a tool to identify correctly and at the earliest k-12 at-risk learners. In addition to the challenges of dealing with multi-source, heterogeneous and of different types data, we proposed an in-depth methodology which gives ML based solutions to early predict at-risk learners. This methodology started with the identification of learning indicators among: performance, engagement, regularity and reactivity. Then, we extracted features from raw data to define each indicator. The identification of learning indicators is of a great importance as it serves on one hand for the prediction of at-risk learners and on the other

hand for the diagnosis of each learner situation and learning gap. Then, we formalized the problem into a 3-class classification problem and followed a weekly prediction approach. For the evaluation phase of the methodology, we used the FPR measure to compare the ability of the used algorithms to well identify the classes of at-risk learners. The findings show that decision tree is the best model that correctly predicts at-risk learners especially on the first weeks. Through these experiments, we also affirmed that the prediction of at medium and high risk of failure learners is given with the minimum error starting from week 20. The perspectives of this study are numerous. We have to extend the application of the methodology on other learning levels and modules. We have also the intention to evaluate these findings with teachers and in a real learning situation. To make the methodology more generic and complete, we aim at adding a phase for the suggestion of academic actions for learners from their teachers.

7 Acknowledgements

This project is funded by the CNED which provides us with data for this work.

References

1. akub, K., Martin, H., Drahomira, H., Zdenek, Z., Jonas, V., Wolff, A.: Ou analyse: Analysing at-risk students at the open university. LAK (2015)
2. Alfred, E., Hanan, A.: Improving student success using predictive models and data visualisations. Research in Learning Technology (2012)
3. Alfred, E., Hanan, A.: Student success system: Risk analytics and data visualization using ensembles of predictive models. LAK (2012)
4. Alice, Z., Amanda, C.: Feature engineering for machine learning. O'REILLY (2018)
5. David, M.O., Du, Q.H., Mark, R., Martin, D., Damyon, W.: A supervised learning framework: using assessment to identify students at risk of dropping out of a mooc. Journal of Computing in Higher Education (2020)
6. Jiliang, T., Salem, A., Huan, L.: Feature selection for classification: A review
7. Kew, S.N., Zaidatun, T.: Identifying at-risk students in online learning by analysing learning behaviour: A systematic review. IEEE Conference on Big Data and Analytics (ICBDA) (2017)
8. Kimberly, E.A., Matthew, D.P.: Case study : A traffic lights and interventions: Signals at purdue university. LAK'12 (April 2012)
9. LIN, Q., YANSHEN, L., YI, L.: An integrated framework with feature selection for dropout prediction in massive open online courses. IEEE Access **6** (2018)
10. Mina Shirvani, B., Kshitij, S., Lukasz, K., Lorenzo, L., Pierre, D.: How to quantify student's regularity. European Conference on Technology Enhanced Learning (September 2016)
11. Mushtaq, H., Wenhao, Z., Wu, Z., Syed Muhammad Raza, A.: Student engagement predictions in an e-learning system and their impact on student course assessment scores. Computational Intelligence and Neuroscience (2018)
12. Papia, B.: Retention in online courses: Exploring issues and solutions —a literature review. SAGE Open pp. 1–11 (2016)
13. Venkatesh, B., Anuradha, J.: A review of feature selection and its methods. CYBERNETICS AND INFORMATION TECHNOLOGIES **19, No 1** (2019)