Assess Performance Prediction Systems: Beyond Precision Indicators

Amal Ben Soussia, Chahrazed Labba, Azim Roussanaly and Anne Boyer
Université de Lorraine, LORIA, France

Keywords: Earliness, Stability, Indicators, Learning Analytics, Machine Learning, k-12 Learners.

Abstract: The high failure rate is a major concern in distance online education. In recent years, Performance Prediction Systems (PPS) based on different analytical methods have been proposed to predict at-risk of failure learners. One of the main studied characteristics of these systems is its ability to provide accurate early predictions. However, these systems are usually assessed using a set of evaluation measures (e.g. accuracy, precision) that do not reflect the precocity, continuity and evolution of the predictions over time. In this paper, we propose to enrich the existing indicators with time-dependent ones including earliness and stability. Further, we use the Harmonic Mean to illustrate the trade-off between the predictions earliness and the accuracy. In order to validate the relevance of our indicators, we used them to compare four different PPS for predicting at-risk of failure learners. These systems are applied on real data of K-12 learners enrolled in an online physics-chemistry module.

1 INTRODUCTION

The use of distance online education has evolved over the last few years, and has exploded further with the recent covid-19 pandemic. It presents an effective way to maintain the continuity of the learning process by allowing access to school from anywhere at any time.

However, the major concern of this learning mode is the high failure rate among its learners. In order to meet this issue, Performance Prediction Systems (PPS) based on Machine Learning (ML) models have been proposed (Wang et al., 2017; Iqbal et al., 2017; Zhang et al., 2017; Chui et al., 2020). The main objective of this type of systems is to predict accurately and at the earliest at-risk of failure learners using real-time flow data.

The data generated by online education platforms are available progressively over time, as they are highly dependent on the time at which learners interact with the educational content. Thus, these data are called time series, as they consist of a set of sequences of events that occur over time. The time reference can be a timestamp or any other finite-grain time interval (e.g week). To assess the effectiveness of the PPS in fulfilling their objectives, common indicators are being used. On the one hand, performance measures are used to assess the temporal and space complexity of the systems in respect to the used data. On the other hand, ML indicators such as precision, recall and F1 measure are used to qualify the ability of the system to predict correctly. Despite the diversity of existing evaluation indicators, none of them is dedicated to assess the precocity, continuity and evolution of predictions over time. However, time is an important dimension that needs to be considered while assessing a PPS as both learning and prediction evolve.

In this paper, we focus on assessing PPS based on time-series classifiers. To overcome the limitations of existing common indicators, we focus on providing new time-dependent ones that can be used to assess these systems over time. Two indicators including earliness and stability are proposed. Indeed, the temporal stability is the ability of a PPS to provide the longest sequences of correct predictions over the prediction times. Whereas, the earliness indicates the first time that the PPS correctly predicted a given class label. A PPS is said to be effective if it predicts accurately and as early as possible. To this end, we used the Harmonic Mean (HM) to study the trade-off between the accuracy and earliness indicators since they are proportionally inverted.

To validate the relevance of these indicators, we applied them to assess four different PPS for predicting at-risk of failure learners. These systems use real data of k-12 learners enrolled in a physics-chemistry module within a French distance learning

Ben Soussia, A., Labba, C., Roussanaly, A. and Boyer, A.
Assess Performance Prediction Systems: Beyond Precision Indicators.
DOI: 10.5220/0011124300003182
In *Proceedings of the 14th International Conference on Computer Supported Education (CSEDU 2022) - Volume 1*, pages 489-496
ISBN: 978-989-758-562-3; ISSN: 2184-5026

center (CNED [1]).

To summarize, our contribution is twofold: 1) new indicators including earliness and stability

; and 2) a real case study to support the use of our indicators.

The rest of the paper is organized as follows: the Section 2 presents the related work and discusses our contribution with respect to the state of the art. The Section 3 introduces the problem formalization as well as the definitions of the proposed indicators. The Section 4 describes our context and the used PPS. The Section 5 presents the conducted experiments and the results. The Section 7 concludes on the results and introduces the work's perspectives.

2 RELATED WORK

ML-based education systems and especially the detection of learners at-risk of failure and dropout are gaining momentum in recent years.

Static ML precision indicators are the most used to evaluate the performance of these systems (Hu et al., 2014). (Bañeres et al., 2020) proposed a model based on students' grades to predict the likelihood to fail a course. Authors of this paper evaluated the performance of the model using the accuracy metric. The main goal of (Lee and Chung, 2019) was to improve the performance of a dropout early warning system. For this aim, the trained classifiers, including Random Forest and boosted Decision Tree, were evaluated with both the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. Based on an ensemble model using a combination of relevant ML algorithms, (Karalar et al., 2021) aimed to identify students at-risk of academic failure during the pandemic. In order to make a classification in which students with academic risks can be predicted more accurately, authors of this paper relied on the results of the specificity measure to evaluate the performance of the ensemble method. The goal of (Adnan et al., 2021) was to identify the best model that analyzes the problems faced by at-risk learners enrolled in online university. The performance of the various trained ML algorithms was evaluated by using accuracy, precision, recall, support and f-score metrics. The Random Forest was the model with the best results.

Predicting at-risk learners at the earliest is one of the main topics in the Learning Analytics (LA) field. (Hlosta et al., 2017) introduced a novel approach, based on the importance of the first assessment, for the early identification of at-risk learners. The key idea of this approach is that the learning patterns can be extracted from the behavior of learners who have submitted their assessment earlier. For the earliest possible identification of students who are at-risk of dropout during a course, (Adnan et al., 2021) divided the course into 6 periods and then trained and tested the performance of ML algorithms at different percentages of the course length. Results showed that at 20% of the course length, the RF model was producing promising results with 79% average precision score. At 60% of the course length, the performance of RF improved significantly. (Figueroa-Cañas and Sancho-Vinuesa, 2020) present a study for a simple and interpretable procedure to identify dropout-prone and fail-prone students before the halfway point of the semester. The results showed that the main factor to the final exam performance is continued learning acquired during at least the first half of the course. The work conducted within the Open University (OU) by (Wolff et al., 2014) has proven that the first assessment is a good predictor of a student's final outcome.

To summarize, the existing research works rely mainly on ML precision indicators to evaluate the performance of PPS. Although the results given by these indicators are important to obtain an overall assessment of ML projects, their use alone is not sufficient for evolving systems over time. In fact, the common indicators do not consider the importance of the temporal evolution of the predictions. When dealing with a time-continuous process, such as learning, the regular tracking of prediction results reveals the need for other time-dependent indicators. Thus, in this work, we consider earliness and stability indicators to provide a deeper evaluation of the PPS. Further, we propose to use the HM measure to establish a compromise between both time-dependent and precision indicators.

3 TIME-DEPENDENT INDICATORS

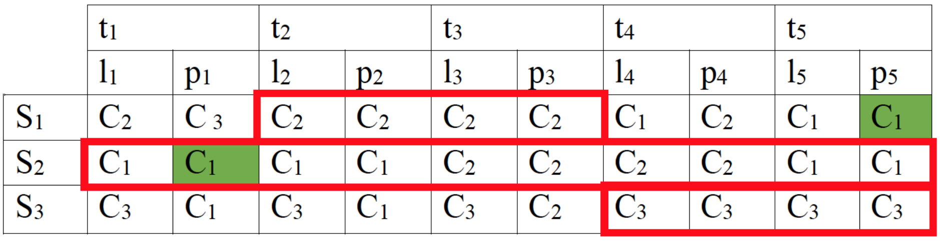
In this section, we formally present the problem of time series classification (Section 3.1) as well as the new proposed indicators, including earliness (Section 3.2) and stability (Section 3.3).

3.1 Problem Formalization

The objective is to predict the class of the students as early and accurate as possible.

Assume Y=$\{C_1, C_2, .., C_m\}$ is the set of predefined class labels that is determined using an existing training data. Let S=(S_1,S_2,...,S_k) be the set of the students

[1] Centre National d'Enseignement à Distance



	t1		t2		t3		t4		t5	
	l1	p1	l2	p2	l3	p3	l4	p4	l5	p5
S1	C2	C 3	C2	C2	C2	C2	C1	C2	C1	C1
S2	C1	C1	C1	C1	C2	C2	C2	C2	C1	C1
S3	C3	C1	C3	C1	C3	C2	C3	C3	C3	C3

Figure 1: An Example of a Regular tracking prediction.

in the test dataset D_{test} and T=$\{t_1$,t_2,..,$t_c\}$ be the set of the prediction times.

Each student $S_p \in S$, at a prediction time $t_i \in T$, is presented by a vector X_{t_i}=$\{f_1$, f_2,...,f_n, $C_j\}$ where the $f_k \in \mathbb{R}$ presents the k^{th} feature, and $C_j \in Y$ is the j^{th} class label. The main objective is to determine at each prediction time t_i the right class for the student.

As shown in Fig. 1, on each $t_i \in T$, each learner $S_p \in S$ is predicted to belong to a class $C_j \in Y$.

The Fig. 1 presents an example of a regular tracking prediction of a set of students $S = (S_1, S_2, S_3)$ over a time interval T=$\{t_1$,t_2, t_3,t_4,$t_5\}$. Each student is assigned a single class label in Y=$\{C_1$, C_2,$C_3\}$.

The quality of PPS based on classifiers can be measured by different indicators, but none of them determines their effectiveness with respect to the temporal dimension. Indeed, the regular tracking of prediction results is at the origin of identifying new time-dependent indicators including *earliness* and *stability*. Further, we need an additional measure that shows the trade-off between our time-dependent indicators and the accuracy. Indeed, the assignment of a class label should be as early and accurate as possible.

3.2 Earliness

The objective of PPS has always been the early prediction of the less performing learners. Early prediction is commonly defined as the adequate and relevant time allowing both an accurate prediction of learners performances and effective tutor interventions with at-risk learners (Bañeres et al., 2020).

Indeed, learners' behavior is not stable and may vary continuously over time, therefore the performance of the model may change and evolve from one prediction time to another. For these reasons, the earliness measurement is significant in the evaluation of an educational PPS as both learning and prediction are time-evolving.

We propose the following Algorithm (see Algorithm.1) to compute the earliness by class label. It takes as input the list of the students (S), the set of class labels (Y) as well as the test data (D_{test}) and provides as output the earliness measures by class label (E_Y). The Algorithm starts by iterating over the class labels ($C_j \in Y$) (Line 1). For each C_j, it initializes two variables ($Early_{tot}$ and $count_S$), which will respectively contain the sum of the first correct prediction times and the number of students who were assigned at least once to the class C_j (Lines 2-3). Then, the Algorithm iterates over the set of students ($S_k \in S$) (Line 4). It verifies first if S_k has been assigned at least once to the class C_j in question (line 5). If so, the Algorithm searches for the first time S_k has been in C_j (Line 6) as well as the first time S_k is predicted correctly in C_j (Line 7). The first correct prediction time is then calculated (Line 8) and both ($Early_{tot}$ and $count_S$) are updated (Lines 9-10). The earliness ($Early_{C_j}$) that corresponds to the class label C_j is determined at line 13. Then before iterating on the next class label, the measured earliness for the C_j in question is saved in E_y (Line 14).

As an example of calculation of earliness, we refer to the Fig. 1. The set S is composed of three students S_1, S_2 and S_3. At each t_i, a student belongs to a class label (l_i) and has a predicted label (p_i). Assume t_i represents a week, and we need to calculate the earliness with respect to class C_1. The first predicted right label for each of the three students is represented by the green box. The student S_3 is not considered, since he/she never been labeled as C_1. By applying the Algorithm.1, the earliness measures for S_1 and S_2 are equal respectively to 2 and 1. Thus, the earliness for class label $C_1 = 1.5$ (3/2). In other words, the system is able to predict correctly the class label C_1 after 1.5 weeks a student is labeled as C_1.

For a PPS, the earlier the predictions are correct, the better the system is.

However, while improving the earliness indicator outcomes, the results of ML precision indicators including the accuracy have to remain high enough to provide stakeholders with as accurate predictions as possible. For this aim, we propose to follow, along with the earliness indicator, the HM, which is a measure of central tendency and used when an average ratio is needed. The HM highlights the reverse proportionality between two variables. This measure has

Algorithm 1: Earliness Algorithm.

```
Require: $S, Y, D_{test}$
Ensure: $E_Y$
for each $C_j$ in $Y$ do
    $Early_{tot} \leftarrow 0$
    $count_S \leftarrow 0$
    for each $S_k$ in $S$ do
        if (assigned$(S_k, C_j)$ ==True) then
            $t_0 \leftarrow$ First_Labeled$(S_k, C_j)$
            $t_1 \leftarrow$ First_Predicted$(S_k, C_j)$
            $Early_{S_k C_j} \leftarrow t_1 - t_0 + 1$
            $Early_{tot} \leftarrow Early_{tot} + Early_{S_k C_j}$
            $count_S \leftarrow count_S + 1$
        end if
    end for
    $Early_{C_j} \leftarrow Early_{tot} / count_S$
    $E_Y \leftarrow put(C_j, Early_{C_j})$
end for
```

been already used in (Schäfer and Leser, 2020) to investigate the relation between accuracy and earliness for early classification of electronic signals. Like them, we used the same HM measure but this time to determine the relationship between earliness and accuracy in a completely different domain. To the best of our knowledge, the HM has never been used in the education domain to express the ability of PPS to provide accurate predictions at the earliest. The application of the HM measure is as follows:

$$HM = \frac{2 * (1 - earliness) * accuracy}{(1 - earliness) + accuracy} \tag{1}$$

The higher HM is, the more the system is qualified to be able to provide accurate early predictions.

3.3 Stability

Evaluating the performance of PPS based on the earliness and stability indicators is of high interest to show the evolution of predictions.

Given the changes of learners behaviors through the learning period, the model could have at each prediction time a different performance which makes the system unstable.

In the state of the art, stability is usually related to small changes in system output when changing the training set (Philipp et al., 2018). However, in our context, we are more interested in temporal stability which refers to the capacity of a model to give the correct output over time when training the same dataset (Teinemaa et al., 2018). The temporal stability characterizes the ability of the system to maintain the same performance throughout the prediction times.

In the frame of this work, we define the temporal stability as the average of the longest sequences of successive correct predictions over time. The stability is calculated using equation 2.

$$Stability = \frac{\sum_{p=1}^{k} |h(S_p)|}{|D|} \tag{2}$$

Where $h : S \rightarrow T^n$ is a function that associates to each student in $D \subseteq D_{test}$ the longest sequence of **successive correct predictions**. The given equation allows to calculate the stability on a class label or on the whole D_{test} at a given time interval. As an example of calculation of stability, we refer to the Fig. 1. For each student, the longest correct prediction sequence is presented by a red line.

The overall stability value till the time prediction t_3 is calculated as follows:

$$Stability = \frac{2+3}{3} = 1.66 \tag{3}$$

Whereas, the stability value at the time prediction t_5 is calculated as follows:

$$Stability = \frac{2+5+2}{3} = 3 \tag{4}$$

Thus, the evaluated PPS shows an ascending stability over time, which allows it to be qualified as a stable system.

4 PROOF OF CONCEPT: COMPARISON OF FOUR PPS

To prove the effectiveness of our temporal indicators, we used them to compare four existing PPS. The same dataset has been used for the four systems. This section presents the case study and introduces the evaluated PPS.

4.1 Context and Dataset Description

Our case study is the k-12 learners enrolled in the physics-chemistry module during the 2017-2018 school year within the French largest center for distance education (CNED [2]). CNED offers multiple fully distance courses to a large number of physically dispersed learners. In addition to the heterogeneity of learners, learning is also quite specific as the registration remains open during the school year. Subsequently, the start activity date t_0 could be different from one learner to another.

The objective is to track students performance on a weekly basis to identify those at-risk of failure. Thus, the prediction time $t_i \in T$ corresponds to a week.

[2] https://www.cned.fr/

Our dataset is composed of learning traces of 647 learners who followed the physics-chemistry module for 37 weeks.
Learners are classified into three classes based on the obtained grades average: Y={C_1, C_2, C_3}

- Success (C_1): when the marks average is strictly superior to 12.
- Medium risk of failure (C_2): when the marks average is between 8 and 12.
- High risk of failure (C_3): when the marks average is strictly inferior to 8.

Each week, a student is represented by a set of learning features and a class label. Based on a previous work (Ben Soussia et al., 2021), these learning features include performance, engagement, regularity and reactivity.

The systems are compared using accuracy and stability over the entire learning period. However, for earliness, it is evaluated in relation to the first 12 weeks. The choice of this period is not arbitrary. According to the existing work on earliness (Figueroa-Cañas and Sancho-Vinuesa, 2020), it can be deduced that earliness is always targeted on the first weeks.

4.2 Overview of the Evaluated PPS

To provide an example of application of the proposed time-dependent indicators, we used them to compare four different PPS (PPS_1, PPS_2 , PPS_3, PPS_4). The first two systems (PPS_1, PPS_2) are based on the Random Forest (RF) model, while the last two (PPS_3, PPS_4) use the Artificial Neural Network (ANN) model.

PPS_1 and PPS_3 use all the learning features including performance, engagement, regularity and reactivity, in addition to demographic data to make weekly basis predictions. While, PPS_2 and PPS_4 use only the engagement features to define students at-risk of failure. An additional difference between the systems is presented by the way the class is assigned in time. Indeed, for PPS_1 and PPS_2, the predictions are made with respect to the learner final class at the end of the year. In other words, each learner belongs to one single class over the year and the model tries to predict that final class as early as possible. Whereas, for PPS_3 and PPS_4, the class label is dynamic and may change based on the student performance. For example, a student may be in the successful class for 3 successive weeks, but in the 4th week he/she may be assigned to a different class label due to fluctuations in performance. The model must therefore capture these changes from one week to another to predict correctly the student's class label.

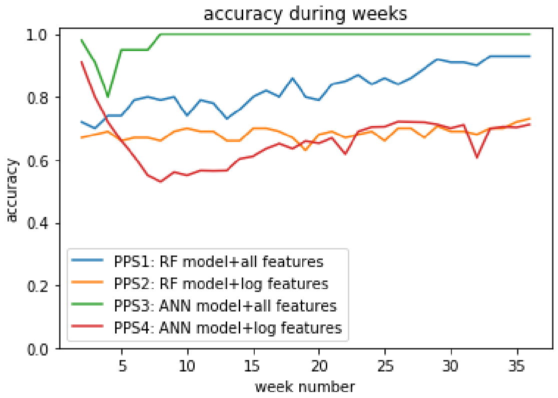


Figure 2: PPS_1 VS PPS_2 VS PPS_3 in terms of accuracy.

The Fig. 2 presents the accuracy of the four systems PPS_1, PPS_2, PPS_3 and PPS_4 over the whole learning weeks. As shown, PPS_1 and PPS_3 that use all the features perform much better in terms of weekly accuracy than PPS_2 and PPS_4 that use only the engagement features. Indeed, PPS_4 has almost $\approx 91\%$ of accuracy at t_1. From week 1 to 6, the accuracy of this system decreases, then, it increases again starting from week 7 to reach a value of $\approx 72\%$ by the end. While, the accuracy of PPS_2 does not present variations and it is almost stable over the weeks and it reaches a max value of $\approx 73\%$. The accuracy of these systems (PPS_2, PPS_4) is not poor and shows that it performs quite well in general.

However, the precision indicators do not reflect the earliness of systems with respect to all the classes and particularly the high and medium risk.

Thus, in the Section.5, the four systems are assessed beyond the use of precision indicators by using the time-dependent ones defined in Section.3.

5 EXPERIMENTAL RESULTS

In this section, we interpret the earliness, HM and stability results for the four systems PPS_1, PPS_2, PPS_3 and PPS_4.

5.1 Earliness and HM Results

Earliness is relevant, especially for detecting high and medium risk students (C_2 and C_3 classes). Thus, we present the results of earliness and HM by class label (See Table 1, and Table 2).

Comparing PPS_1 and PPS_2: For both systems PPS_1 and PPS_2, the dominant class is C_1 (success), which is predicted respectively at 9.5% ($\approx$ 1.14 week) and 8.83% ($\approx$ 1 week) of the fixed prediction time interval (12 weeks). The earliness and accuracy are

Table 1: Earliness and HM measurements-PPS_1 VS PPS_2.

	PPS_1			PPS_2		
	Earliness	Accuracy	HM	Earliness	Accuracy	HM
C_1	9.5%	92.63%	90.95%	8.83%	91%	91.29%
C_2	26.6%	9%	16.03%	0%	0%	0%
C_3	44.1%	46%	50.46%	0%	0%	0%

Table 2: Earliness and HM measurements-PPS_3 VS PPS_4.

	PPS_3			PPS_4		
	Earliness	Accuracy	HM	Earliness	Accuracy	HM
C_1	10.75%	78.94%	83.83%	37.85%	21.62%	25%
C_2	17.66%	55.75%	66.52%	0%	0%	0%
C_3	9.33%	100%	95.12%	8.5%	92.37%	90.65%

slightly different. By referring to the HM measure the PPS_2 is better in predicting at the earliest the learners at success class. However, PPS_2 is worst when it comes to detecting both classes C_2 and C_3. Indeed, over the 12 first weeks the system has 0% HM measures for C_2 and C_3.

Comparing PPS_3 and PPS_4: For both systems PPS_3 and PPS_4, the dominant class, at the beginning of the school year, is C_3 which is predicted respectively at 9.33% ($\approx$ 1.11 week) and 8.5% ($\approx$ 1 week). The earliness values are slightly different, but PPS_3 has higher accuracy and higher HM measures. Conventionally, for a good model, the accuracy increases with time. However, PPS_4 predicted C_1 latter and less accurately than PPS_3. Further, despite of predicting C_3 accurately and quite early, PPS_4 has never predicted learners belonging to C_2 during the first 12 weeks. According to the HM measures, PPS_3 outperforms PPS_4 in predicting accurately at the earliest all of the three classes.

To summarize, for the four systems, the dominant class is always predicted at the earliest with respect to the rest of the class labels. If we interpret, for example, the earliness rate in relation with the class label C_2 for PPS_2 and PPS_4, one can think that the system was able to predict the class at the earliest. Theoretically 0% is the best earliness value with an accuracy of 100%. However, this is not true since the accuracy and the HM are equal to zero. Thus, 0% does not explain anything unless we investigate accuracy and consequently the HM measure. These results prove the importance of feature selection in predicting at the earliest students at risk of failure. PPS_2 and PPS_4, which use only engagement features, are the best examples, where one or two of the medium and high risk classes are not detected. These results prove also that evaluating a PPS based on either accuracy or earliness is not pertinent enough to conclude on the performance of a classifier. The trade-off between both indicators through the measurement of HM gives a more precise insight about the PPS performance.

The earliness and HM can be used also to evaluate PPS adopting different classification approaches such as PPS_1 and PPS_3. The first one performs the predictions with respect to the final class at the end of the school year. While the second one performs the predictions at a time t with respect to the class label at the time t+1. As shown in Tables 1 and 2, the HM measures show that PPS_1 is less early and less accurate than PPS_3 in predicting C_2 and C_3, knowing that they use the same test data. Determining which system is better than the other is beyond the scope of this paper, but we can conclude that the prediction approach can have an impact on the earliness of the system.

5.2 Stability Results

Stability is complementary to HM. The Table. 3 presents the stability measures for the four systems over the first 12 weeks. The stability per class is more pertinent for PPS_1 and PPS_2 than for PPS_3 and PPS_4. This can be explained by the fact that the former predict with respect to the final class labels while the latter predict with respect to the t+1 class labels. For this reason, we consider to follow also the predictions stability on the entire test dataset D_{test}. The results of D_{test} stability are more coherent with the HM ones and the overall stability is more pertinent as it reflects the true stability of the PPS. As shown in Table 3, until week 12, PPS_1 is more stable than PPS_2 in terms of class stability and overall stability. Indeed, PPS_1 succeeded in predicting the class labels correctly and build longer sequences of correct predictions. While, when it comes to PPS_3 and PPS_4, the overall stability is more relevant, and it shows that PPS_3 is more stable.

Unlike earliness, the stability of PPS is more interesting when it is tracked throughout the learning period. The Figure. 3 shows the D_{test} stability evolution of the four systems throughout the school year.

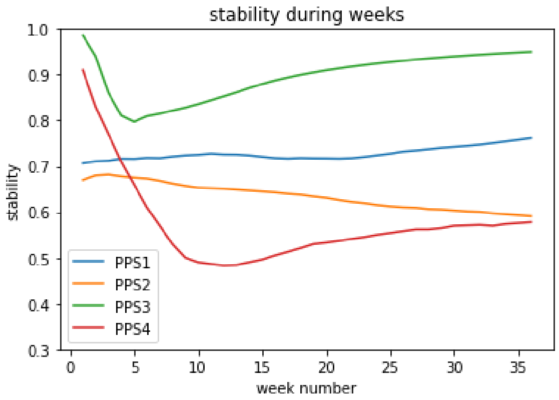


Figure 3: Comparing PPS_1 V, PPS_2 , PPS_3 and PPS_4 in terms of temporal stability.

The stability of PPS_1 increases slightly over time. It is at $\approx 70\%$ and $\approx 76\%$ respectively at the first and last prediction weeks. While the stability of PPS_2 is decreasing over the weeks. It started with a value of $\approx 64\%$ and ended with $\approx 59\%$.

Both systems PPS_3 and PPS_4 start with high stability values ($\approx 100\%$ and $\approx 92\%$ respectively). This can be explained by the dominance of the class C_3, since at the beginning all learners are assigned to this class by default. However, these high values decrease rapidly over the following weeks. For PPS_3, around the week 4, it starts to correctly assign each learner to the suitable class among C_1, C_2 and C_3. Then, from week 5, the stability of PPS_3 increases continuously and reaches a rate of $\approx 96\%$ at the last prediction time. For PPS_4, the shape of the curve is identical to this of PPS_3 with a downward shift. Until week 13, the overall stability decreases, then from week 14, it starts to increase again to reach $\approx 57\%$. We notice either a partial or a total drop in stability for the systems PPS_2, PPS_3 and PPS_4. Although the stability of PPS_1 has never decreased over time, PPS_3 remains the most stable.

Table 3: Stability Measurements.

	PPS_1	PPS_2	PPS_3	$PPS4$
C_1	92.26%	89.12%	58.79%	18.24%
C_2	11.36%	0%	35.85%	0%
C_3	27.56%	0%	47.70%	39.86%
D_{test}	72.72%	65.12%	88.10%	48.48%

Yet, stability and accuracy are proportional, however, from the accuracy graphs of PPS_2 and PPS_4, we cannot conclude on the effectiveness of the systems in maintaining correct prediction sequences over time.

6 THREATS TO VALIDITY

The current work presents some limitations that we tried to mitigate when possible: i) the earliness algorithm returns a single value which corresponds to the mean of the first correct predictions. However, a high associated HM is partially reliable as the accuracy of the system may decrease after the identified earliness. We intend to consider several earliness points with respect to the one returned by our algorithm. The objective is to study the variations of the HM measures between these earliness points. ii) in this work, we have adopted a weekly-basis prediction approach. However, we believe that the proposed indicators can be also used to define the appropriate temporal granularity that provides better prediction results. iii) in the frame of this work, we only considered classification problems. To prove the generic use of our indicators, we aim to apply them on other systems that use regression-based analytical models.

7 CONCLUSION AND PERSPECTIVES

In this paper, we introduced time-dependent indicators, namely the earliness and stability to assess PPS used in online distance education. Further, a trade-off between the earliness and accuracy is important to assess the ability of a PPS to predict learners at-risk of failure. The use of HM measure serves to illustrate this balance. The new indicators along with the accuracy are used to compare four different PPS. The experimental results prove that the accuracy is not sufficient to evaluate a PPS within a context of time-series data. The experimental results show that the HM measure is relevant in identifying the earliest and accurate system and that the stability is more pertinent when the whole test dataset is considered. Further, a system is considered better when its stability increases over time.

As perspectives for this work, we intend to improve the proposed earliness algorithm so that it returns a set of different earliness values. Such a result will enable us to study more deeply the behavior of the PPS. In addition, our next goal is to study the trade-off between the stability and earliness indicators and conclude on the relation between both of them.

REFERENCES

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., Bashir, M., and Khan, S. U. (2021).

Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *IEEE Access*, 9:7519–7539.

Bañeres, D., Rodríguez, M. E., Guerrero-Roldán, A. E., and Karadeniz, A. (2020). An early warning system to detect at-risk students in online higher education. *Applied Sciences*, 10(13):4427.

Ben Soussia, A., Roussanaly, A., and Boyer, A. (2021). An in-depth methodology to predict at-risk learners. In *European Conference on Technology Enhanced Learning*, pages 193–206. Springer.

Chui, K. T., Fung, D. C. L., Lytras, M. D., and Lam, T. M. (2020). Predicting at-risk university students in a virtual learning environment via a machine learning algorithm. *Computers in Human Behavior*, 107:105584.

Figueroa-Cañas, J. and Sancho-Vinuesa, T. (2020). Early prediction of dropout and final exam performance in an online statistics course. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 15(2):86–94.

Hlosta, M., Zdrahal, Z., and Zendulka, J. (2017). Ouroboros: early identification of at-risk students without models based on legacy data. In *Proceedings of the seventh international learning analytics & knowledge conference*, pages 6–15.

Hu, Y.-H., Lo, C.-L., and Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, 36:469–478.

Iqbal, Z., Qadir, J., Mian, A. N., and Kamiran, F. (2017). Machine learning based student grade prediction: A case study. *arXiv preprint arXiv:1708.08744*.

Karalar, H., Kapucu, C., and Gürüler, H. (2021). Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. *International Journal of Educational Technology in Higher Education*, 18(1):1–18.

Lee, S. and Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, 9(15):3093.

Philipp, M., Rusch, T., Hornik, K., and Strobl, C. (2018). Measuring the stability of results from supervised statistical learning. *Journal of Computational and Graphical Statistics*, 27(4):685–700.

Schäfer, P. and Leser, U. (2020). Teaser: early and accurate time series classification. *Data mining and knowledge discovery*, 34(5):1336–1362.

Teinemaa, I., Dumas, M., Leontjeva, A., and Maggi, F. M. (2018). Temporal stability in predictive process monitoring. *Data Mining and Knowledge Discovery*, 32(5):1306–1338.

Wang, W., Yu, H., and Miao, C. (2017). Deep model for dropout prediction in moocs. In *Proceedings of the 2nd international conference on crowd science and engineering*, pages 26–32.

Wolff, A., Zdrahal, Z., Herrmannova, D., Kuzilek, J., and Hlosta, M. (2014). Developing predictive models for early detection of at-risk students on distance learning modules.

Zhang, W., Huang, X., Wang, S., Shu, J., Liu, H., and Chen, H. (2017). Student performance prediction via online learning behavior analytics. In *2017 International Symposium on Educational Technology (ISET)*, pages 153–157. IEEE.