How Far Can We Trust the Predictions of Learning Analytics Systems?

Amal Ben Soussia and Anne Boyer
Université de Lorraine, LORIA, France

Keywords: Learning Analytics, Prediction Systems, Online Learning, Trust Granularities, Trust Index, k-12 Learners.

Abstract: Prediction systems based on Machine Learning (ML) models for teachers are widely used in the Learning Analytics (LA) field to address the problem of high failure rates in online learning. One objective of these systems is to identify at-risk of failure learners so that teachers can intervene effectively with them. Therefore, teachers' trust in the reliability of the predictive performance of these systems is of great importance. However, despite the relevance of this notion of trust, the literature does not propose particular methods to measure the trust to be granted to the system results. In this paper, we develop an approach to measure a teacher's trust in the prediction accuracy of an LA system. For this aim, we define three trust granularities, including: the overall trust, trust per class label and trust per prediction. For each trust granularity, we proceed to the calculation of a Trust Index (TI) using the concepts of confidence level and confidence interval of statistics. As a proof of concept, we apply this approach on a system using the Random Forest (RF) model and real data of online k-12 learners.

1 INTRODUCTION

Prediction systems based on Machine Learning (ML) models are a widespread solution in the Learning Analytics (LA) literature to identify at-risk of failure learners (López Zambrano et al., 2021).

When they are for teachers, these systems are intended to enable effective and accurate instructional intervention with at-risk learners. Indeed, teachers refer to the prediction outcomes of these systems in their pedagogical monitoring and in taking specific corrective actions with the less performing learners. Thus, teachers' trust in the reliability of the accuracy of an LA system's predictions is of great importance. In other words, after identifying the learner's academic situation, teachers also need to know how far they can trust the system's prediction results. This notion of trust is interesting as it ensures the teachers' acceptability of the prediction results for an effective and accurate pedagogical interventions. For example, for a teacher, not correctly identifying an at-risk learner is worse than identifying a successful learner as at-risk. And since prediction systems are often characterized by the instability and the oscillation of their results (Ben Soussia et al., 2022), such an example is quite common. In such a situation, teachers' trust in the system's performance comes into play to give leeway to the predictive outcomes. Indeed, (Qin et al., 2020) defines trust in Artificial Intelligence-based educational systems as the willingness of users to receive knowledge, provide personal information and follow suggestions based on the belief that these systems and their managers or developers will act responsibly. However, the LA literature does not address the problem of how far teachers can trust the predictive performance of an educational system. Furthermore, the LA often discusses trust from an ethical and black-box, in relation to the nature and type(s) of used data, point of view. Given the importance of the trust notion in LA, the main question is: ***how to measure the trust index to be granted to the performance of a prediction system?***

To answer this question, we focus in this paper on developing an approach to measure a Trust Index (TI) of teachers in the prediction accuracy of an LA system. First, we define three trust granularities in an LA system, including : *(1) the overall trust*, *(2) trust per class label of the system* and *(3) trust per prediction made by the system*. Then, for each trust granularity, we compute a TI using the concepts of confidence level and confidence interval of the statistics. For the TI of each trust granularity, we propose an algorithm to compute the confidence level to be granted to the system performance. For trust granularities (1) and (2), we proceed to the computation of the confidence intervals using the most popular statistical method called : *Normal Approximation Interval*

Ben Soussia, A. and Boyer, A.
How Far Can We Trust the Predictions of Learning Analytics Systems?.
DOI: 10.5220/0012057800003470
In *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023) - Volume 2*, pages 150-157
ISBN: 978-989-758-641-5; ISSN: 2184-5026

Based on a Test Set (Raschka, 2018).

To validate this approach, we apply these algorithms on an LA system using Random Forest (RF) model and real data of k-12 learners enrolled online within a French distance education center (CNED[1]).

The rest of this paper is organized as follows. The Section 2 presents the related work and discusses on our contribution with respect to the literature. Section 3 presents statistical information. Section 4 formalizes the problem and introduces the trust granularities and their TI algorithms. Section 5 describes our case study. Section 6 presents the experimental part and the obtained results. Section 7 concludes on the results and introduces the perspectives.

2 RELATED WORK

Artificial intelligence (AI) is an emerging science of dealing with the simulation of intelligent behavior in computers (Bitkina et al., 2020). However, while very promising, AI has been implicated in trust issues, and concerns have been raised about the use of AI in various initiatives and technologies (Lockey et al., 2021). For these reasons, trust in AI is gaining so much interest lately. (Siau and Wang, 2018) confirms that the level of trust a person has in someone or something can determine that person's behavior. (Stanton et al., 2021) defines user trust in AI application as based on the perception of its reliability. The actual reliability of the AI is influential to the extent that it is perceived by the user. Trust is a function of the user's perceptions of technical reliability characteristics. In this context, (Lockey et al., 2021) introduces concepts related to five central AI trust challenges including the ability to know and explain AI, accountability for accuracy and fairness of systems outcomes, systems automation and minimization of direct human involvement, the inclusion of human-like features into an AI's design and accountability for data privacy.

The literature of LA has also taken advantage of AI to widely propose systems and dashboards to solve learning issues and monitor learners' behaviors and their academic situations. The issue of trust is also addressed in LA and is gaining the interest of the actors of the field. (Tsai et al., 2021) discusses on the trust factors and threats in LA among : data accuracy, equity of treatment, potential misuse of LA.... To convince stakeholders that LA dashboards and systems work for their best interest, building trust, according to (Biedermann et al., 2018), involves several layers, from the integrity and quality of data sources, over secure storage and processing to the effectiveness of the analytics results. (Baneres et al., 2021) proposes a trustworthy early warning system to detect at-risk learners early to help them to pass the course. The infrastructure of this system is built on the basis of the requirements of the European Assessment List for trustworthy artificial intelligence including : human agency and oversight, robustness and safety, privacy and data governance, transparency, diversity, fairness, accountability.... In order to monitor students learning, (Susnjak et al., 2022) proposes a dashboard that incorporates data-driven prescriptive capabilities involving counterfactuals into its display. In addition, the proposed dashboard has a high degree of transparency as it communicates to the learners the reliability of the predictive models, the key factors driving the predictions, and the conversion of a black-box predictive model to a human-interpretable model so that learners can understand how their prediction is derived. This paper has demonstrated the importance of predictive models interpretability in building trust with dashboard users through transparency of evolution beyond black-box predictive models and, in doing so, to meet new regulatory requirements.

In summary, the trust notion is gaining interest in LA. However, most of the work is limited (1) to the ethical aspects of AI (e.g data protection and privacy) to increase the trust of stakeholders in the systems, (2) to the trust in the data, their quality and relevance as well as (3) to the transparency of the used models. Despite the importance of accuracy and fairness of systems outcomes as an AI trust challenge, to our knowledge, no work in the LA highlights the importance of measuring trust in the systems predictions. Several research works emphasize the importance of measuring confidence intervals of statistics of AI-based system predictions as a broader mean of validation than measuring only accuracy. In this paper, we build an approach to measure the trust in the predictions of an LA system for online teachers. We define three trust granularities. For each one of them, we compute a Trust Index (TI) using the concepts of confidence intervals and confidence levels of statistics.

3 STATISTICAL BACKGROUND

Confidence intervals are interesting in modeling and simulation as they are commonly used for model validation. A confidence interval is a range of values estimate of a parameter of a population calculated from a sample drawn from the population. A confidence interval has an associated confidence level, which is a percentage between 0% and 100% (Petty, 2012).

[1]Centre National d'Enseignement à Distance

In this paper, to create confidence intervals, we use the most popular method and that guarantees good results : *Normal Approximation Interval Based on a Test Set* (Raschka, 2018). Using this method, the confidence interval is calculated from a single train-test split and follows this structure :

$$\text{confidence interval} = [\text{estimated parameter} - \text{margin}, \text{estimated parameter} + \text{margin}]$$

where the *margin* is the standard error of the corresponding estimated parameter. In our context, the estimated parameter is the accuracy metric of ML, and the margin is calculated as follows:

$$margin = z_c \times \sqrt{\frac{accuracy_{test} \times (1 - accuracy_{test})}{n}} \quad (1)$$

Where $accuracy_{test}$ is the accuracy of the predictions and n is the size of the test dataset. z_c is the critical value for the normal distribution for a given confidence level (Petty, 2012).

In our work, each confidence interval is equal to:

$$\text{confidence interval} = [\text{accuracy} - \text{margin} , \text{accuracy} + \text{margin}]$$

In the literature, to create confidence intervals, some confidence level values and their corresponding z_c values are commonly used. In this work, we consider that the confidence level in a system depends on its predictions. Therefore, for each trust granularity, we propose an algorithm to compute the confidence level. Then, we can proceed to the creation of confidence intervals following the Normal Approximation Interval Based on a Test Set method.

4 TRUST INDEX OF EACH TRUST GRANULARITY

In this section, we formalize the problem and introduce the three trust granularities : *(1) the overall trust*, *(2) trust per class label* and *(3) trust per prediction*. Then, we present the TI of each trust granularity.

4.1 Problem Formalization

The use of predictive systems by teachers is widespread in LA. Through the identification of at-risk learners, these systems allow teachers to monitor the activity of their learners and intervene with those in critical academic situations. Therefore, a teacher's trust in the prediction results is important for the system acceptability and thus for the best follow-up.

Assume that $Y = \{C_1, C_2, .., C_m\}$ is the set of class labels. Let $S = \{S_1, S_2, .., S_q\}$ be the set of students in the test dataset and $T = \{t_1, t_2, .., t_k\}$ be the set of prediction times. At each $t_i \in T$, each $S_p \in S$ is represented by a vector $X_{p_i} =< f_1, f_2.., f_z, C_j >_{p_i}$ where $f_n \in \mathbb{R}$ represents the learning features of S_p and $C_j \in Y$ his class label. Let $y_{test} = \{y_1, .., y_q\}$ is the set of real class labels of each leaner, where $y_l \in Y$. Let $y_{prediction} = \{y1_{pred}, .., y_{q_{pred}}\}$ is the set of predicted class labels of each learner of S, where $y_{l_{pred}} \in Y$. The objective is to have at each prediction time, $y_l = y_{l_{pred}}$, which is not always obvious in a real context.

4.2 Overall Trust Granularity

In this section, we introduce the overall trust granularity. Then, we present the algorithm for its TI.

4.2.1 Definition

The objective of LA systems is to accurately predict at-risk learners so that teachers can intervene effectively. Therefore, it is necessary for teachers to show trust in the overall prediction performance. In this context, we define the overall trust as a TI computed from the confidence level and then the confidence interval that can be granted to the accuracy of all the predictions made by the system. Based on this definition, this overall trust granularity tells the teacher how far he can trust the overall performance of the system independently of the existing class labels and the individual predictions.

4.2.2 Overall Confidence Level Algorithm

In this section, we propose the Algorithm 1 to compute the overall confidence level of an LA system.The Algorithm 1 takes as input y_{test}, $y_{prediction}$ which are respectively the sets of real and predicted class labels and Y which is the set of class labels of the system. This Algorithm returns $confidence_{level}$ (a percentage) which is the confidence level to give to the overall performance of the system. The Algorithm starts by initializing to 0 the variable $confidence_g$ (Line 1) which is the general confidence of all predictions. Then, the Algorithm iterates over the predicted class labels set $y_{prediction}$ (Line 2). The Algorithm initializes to 0 the variable $confidence_i$, which is the confidence value to give to the i^{th} prediction (Line 3). The Algorithm verifies if the prediction at the i^{th} index is equal to the value of the y_{test} at the same index (Line 4). If so, the Algorithm assigns 1 to $confidence_i$ (Line 5). Else, it assigns to $confidence_i$ the value of $(1/size(Y))$, which refers to the number of class labels in Y (Line 7). At Line 9, the value of $confidence_i$ is added to $confidence_g$. At Line 11, the overall confidence level

given by $confidence_{level}$ is calculated by dividing the value of $confidence_g$ by the size of y_{test}.

Algorithm 1: Overall Confidence Level.

Require: $y_{test}, y_{prediction}, Y$
Ensure: $confidence_{level}$
1: $confidence_g \leftarrow 0$
2: **for each** i in $y_{prediction}$ **do**
3: $confidence_i \leftarrow 0$
4: **if** $(y_{prediction}[i] == y_{test}[i])$ **then**
5: $confidence_i \leftarrow 1$
6: **else**
7: $confidence_i \leftarrow 1/size(Y)$
8: **end if**
9: $confidence_g \leftarrow confidence_g + confidence_i$
10: **end for**
11: $confidence_{level} = confidence_g/size(y_{test})$

Once, the $confidence_{level}$ is calculated, we compute the value of z_c and then the overall margin ($margin_{overall}$) of the confidence interval:

$$margin_{overall} = z_c \times \sqrt{\frac{accuracy_{test} \times (1 - accuracy_{test})}{size(y_{test})}} \quad (2)$$

$size(y_{test})$ corresponds to the number of data points in test dataset.

4.3 Trust per Class Label Granularity

In this section, we introduce the trust per class label granularity. Then we present the algorithm for its TI.

4.3.1 Definition

In LA systems for class labels prediction, learners are usually classified into more than one class. Thus, the system could potentially perform differently with each of these classes. Therefore, it is of great interest for the teacher to know how far she/he can trust the performance of the system when it comes to the predictions of a particular class. In this perspective, we define the trust per class label as TI computed from the confidence level and the confidence interval to be granted to the accuracy of predictions of a given class label.Such a definition allows for a more thorough examination of the reliability of predictions. The purpose of such a TI is to enable effective intervention with learners of each of the system's class labels.

4.3.2 Confidence Level per Class Label Algorithm

In this section, we propose the Algorithm 2 to compute the confidence level of each class label for the TI of each class label's predictions. The Algorithm 2 takes as input the set of class labels Y and T which is the set of class probabilities tables for each data point of the test dataset. This Algorithm returns $confidence_Y$ which is a list of class labels and their corresponding confidence levels. This Algorithm starts by iterating over the class labels in Y (Line 1). For each $C_j \in Y$, the Algorithm initializes to 0 the variable $probability_{C_j}$, which corresponds to the sum of the prediction probabilities of each data point in C_j (Line 2). Then, the Algorithm iterates over T_{C_j}, which is the prediction probabilities table of C_j (Line 3). At Line 4, the value of T_{C_j} at index i is added to $probability_{C_j}$. At Line 6, the confidence level of the class label C_j given by $confidence_{level_{C_j}}$ is calculated by dividing the value of $probability_{C_j}$ by the size of T_{C_j}. Then, the measured confidence level $confidence_{level_{C_j}}$ is saved in $confidence_Y$ with its corresponding class label C_j.

Algorithm 2: Confidence Level of each class label - TI_class(Y, T) -.

Require: Y, T
Ensure: $confidence_Y$
1: **for each** C_j in Y **do**
2: $probability_{C_j} \leftarrow 0$
3: **for each** i in T_{C_j} **do**
4: $probability_{C_j} \leftarrow probability_{C_j} + T_{C_j}[i]$
5: **end for**
6: $confidence_{level_{C_j}} == probability_{C_j}/size(T_{C_j})$
7: $confidence_Y \leftarrow put(C_j, confidence_{level_{C_j}})$
8: **end for**

After applying the Algorithm 2, we can compute for each class label C_j the value of z_{C_j} and then of the $margin_{C_j}$ of the confidence interval of C_j as follows:

$$margin_{C_j} = z_{C_j} \times \sqrt{\frac{accuracy_{C_j} \times (1 - accuracy_{C_j})}{card_{C_j}}} \quad (3)$$

Where $accuracy_{C_j}$ is the accuracy of the predictions of this particular class label C_j and $card_{C_j}$ is the number of learners who are really labeled as in C_j.

4.4 Trust per Prediction Granularity

In this section, we introduce the trust per prediction granularity. Then, we present the algorithm for its TI.

4.4.1 Definition

The objective of LA systems is to predict at-risk learners so that teachers can intervene with them. Accu-

rate results are important for effective and personalized interventions. Therefore, it is pertinent for the teacher to know how far he can trust the reliability of each single prediction of the system. We define the trust per prediction as the TI computed from the confidence level to be granted to the accuracy of that prediction independently of the performance of the system with the rest of data points. Such a TI allows the teacher to have the reliability of the system performance with each prediction. This trust granularity fits in with the goal of personalized pedagogical intervention with each learner of the educational system.

4.4.2 Confidence Level per Prediction Algorithm

In this section, we propose the Algorithm 3 to calculate the TI of this trust granularity which is the confidence level of each prediction made by the system. The Algorithm 3 takes as input y_{test} and $y_{prediction}$ which are respectively the sets of real and predicted class labels. It requires also the set of class labels Y and T which is the set of class probabilities tables for each data point of the test dataset. This Algorithm returns $C_{y_{prediction}}$, which is a list of prediction indexes and their corresponding confidence levels. This Algorithm starts by iterating over the predictions of $y_{prediction}$ (Line 1). For each prediction, the Algorithm initializes the variable $confidence_i$ to 0 (Line 2). Then, it verifies if the prediction at the i index is the same of which of the test set at the same index (Line 3). If so, the Algorithm assigns 1 to $confidence_i$ (Line 4). Else, the prediction at the i index corresponds to a class label C_j among Y (Line 6). At Line 7 and 8, the Algorithm extracts the confidence level of C_j by calling the Algorithm 2 and it assigns it to $confidence_i$. At Line 10, the measured confidence $confidence_i$ is saved along with its corresponding index i in $C_{y_{prediction}}$.

Algorithm 3: Confidence Level per each prediction.

Require: $y_{test}, y_{prediction}, Y, T$
Ensure: $C_{y_{prediction}}$
1: **for each** i **in** $y_{prediction}$ **do**
2: $confidence_i \leftarrow 0$
3: **if** $(y_{prediction}[i] == y_{test}[i])$ **then**
4: $confidence_i \leftarrow 1$
5: **else**
6: $C_j \leftarrow y_{prediction}[i]$
7: $confidence_{C_j} \leftarrow extract(TI\_class(Y,T))$
8: $confidence_i \leftarrow confidence_{C_j}$
9: **end if**
10: $C_{y_{prediction}} \leftarrow put(i, confidence_i)$
11: **end for**

For this trust granularity, the TI of a prediction is the confidence level calculated using the Algorithm 3.

5 CASE STUDY

Our case study is the k-12 learners enrolled online within CNED. CNED offers multiple fully distance courses to a large number of heterogeneous and physically dispersed learners. In addition, learning is quite specific; for example, learners of same cohorts do not necessarily start their school year at the same time and everyone of them studies on her/his own pace. Given these learning particularities, CNED records high failure rates among its learners every year. In order to minimize the failure risk and improve the pedagogical monitoring of teachers, CNED aims to provide teachers with system based on LA technologies to help them in identifying accurately at-risk of failure learners. For the CNED, gaining the trust of teachers in the reliability of the results of this system is paramount to the success of this LA-based strategy.

Based on the grades average and according to the French system where marks are out of 20, learners of each module are classified into 3 classes as follows :

- success (C_1): when the average is higher than 12
- medium risk of failure (C_2): when the average is between 8 and 12
- high risk of failure (C_3): when the average is lower than 8

This system tracks learners activity on a weekly basis to allow teachers to regularly monitor their learners. Thus, on each prediction week, each learner is represented by a vector composed of learning features and a class label among $\{C_1, C_2, C_3\}$.

Indeed, this system uses learning traces of 647 learners enrolled in the physics-chemistry module for 35 weeks during 2017-2018 school year and is modeled with the Random Forest (RF) model.

6 EXPERIMENTS AND RESULTS

In this section, we analyze the results of applying our approach.

6.1 Overall TI Results

The overall trust granularity TI is important to analyze the reliability of the whole system's performance.

The Figures 1 and 2 represent the confidence intervals of the predictions overall accuracy using 90% and

95% confidence levels respectively, which are commonly used by default in the literature. Thus, at each prediction time, the margins of the confidence intervals of these Figures are calculated based on these confidence levels and on the accuracy of the predictions at this time point. We notice that the confidence intervals of these two Figures are quite wide. The margins of the confidence intervals of the 90% confidence level vary between 0.06 and 0.04. While the margins of the confidence intervals of the 95% confidence level are between 0.08 and 0.04.

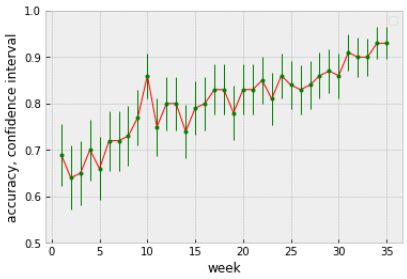

Figure 1: Overall Confidence intervals with a confidence level= 90%.

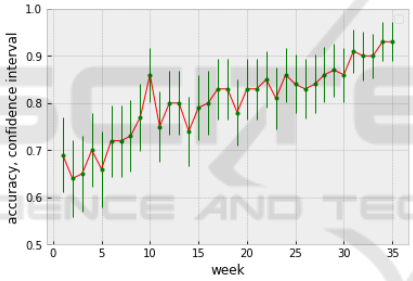

Figure 2: Overall Confidence intervals with a confidence level=95%.

To calculate the TI of the overall trust granularity following our approach, we first calculate the overall confidence levels at each prediction time based on the Algorithm 1 and which the evolution is presented by Figure 3. We can observe that both curves of this Figure have the same shape. In fact, the values of the confidence levels of the system performance evolve according to the system accuracy values. In other words, at each prediction time, we have a confidence level that is more correlated to the performance of the system, rather than assigning a default confidence level. Given the yielded confidence levels, we compute the confidence intervals to be granted to the overall accuracy of the system following the Equation 2. The evolution of the system accuracy values and their corresponding confidence intervals is given by the Figure 4. We remark that during all prediction times, the confidence intervals of Figure 4 are narrower than those of Figures 1 and 2. In fact, the margins of the intervals with respect to the accuracy values are smaller as they vary between 0.05 and 0.04. In statistics, it is better to have narrow confidence intervals as it proves the accuracy and the relevance of the used confidence level. A large confidence interval suggests that the sample does not provide a precise representation of the estimate, whereas a narrow confidence interval demonstrates a greater precision.

Instead of using default confidence levels, using the Algorithm 1 to calculate the confidence level enables to create more accurate and narrower confidence intervals for the accuracy of the predictions. Thus, this approach enables to measure a more precise TI for this trust granularity to give the teacher a finer idea of the reliability of the system's predictions.

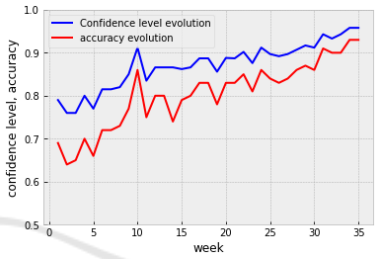

Figure 3: Overall confidence level and accuracy evolution.

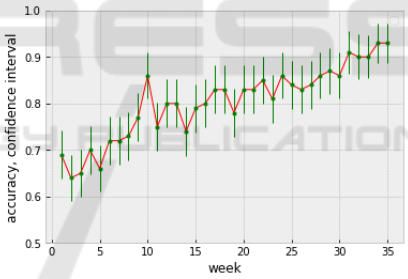

Figure 4: Overall Confidence intervals with a calculated overall confidence level.

6.2 TI per Class Label Results

The class label trust granularity TI is interesting as it allows to follow the reliability of the system performance with each class label.

The Figure 5 presents the confidence levels evolution of the system performance with each class label. For each class C_1, C_2 and C_3, the confidence levels are calculated following the Algorithm 2. This figure shows that the confidence levels of class labels of the same system evolve very differently over time. Indeed, the confidence levels of each class label depend on its population and the ability of the system to correctly predict the data points of that class. In fact, the highest confidence levels are recorded with the class of successful learners C_1. While the low-

est confidence levels are recorded with the two risk classes, especially C_3.

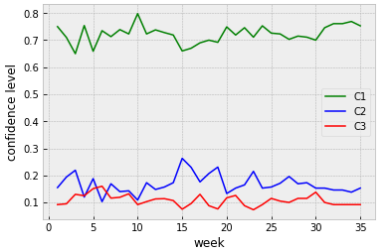

Figure 5: Evolution of all confidence levels of all class labels of the system.

Given these confidence levels, the Figures 6, 7 and 8 present the evolution over time of the confidence intervals to be granted to the prediction accuracy of the class labels C_1, C_2 and C_3 respectively. From these figures, we notice that the confidence intervals for the accuracy of C_1 predictions are narrow. That said, the margins of the intervals with respect to the accuracy values are small and vary between 0.03 and 0.01. We can highly trust the system when it comes to the prediction of C_1 as it gets rarely wrong with this class label. However, if we compare the results of Figure 7 and 8, we remark that the confidence intervals for the accuracy of C_3 are narrower than those for C_2 especially at some time points such as the prediction weeks 10, 22 and 23. Such a result shows that the accuracy of the system predictions with C_3 is more reliable than with the medium risk class C_2. Therefore, the system's performance is more trustworthy when it comes to the prediction of the risk class C_3 than the risk class C_2, which is the class of learners in a fuzzy learning situation. The obtained results show the relevance of the measurement of TI of the trust granularity per class as it allows to follow the reliability of the performance of the system with each of the class labels. The TI is different from on class label to another.

6.3 TI per Prediction Results

The TI of the trust per prediction granularity is interesting for tracking the reliability of the system performance with each single data point.

Applying the Algorithm 3 on the results of our system, the Figure 9 presents the number of data points by each confidence level among 1 which corresponds to the total trust in the prediction result and the confidence levels corresponding to the C_1, C_2 and C_3 class labels. For simplification reasons, the results are presented every 5 weeks from week 5 to week 35.

This Figure shows that as time progresses, the system predicts better and more correctly each data point.

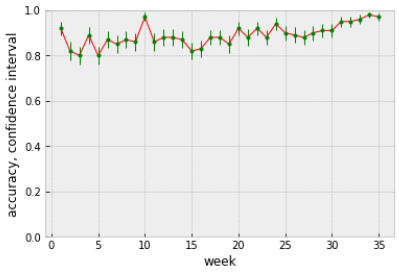

Figure 6: Confidence intervals of the class label C_1.

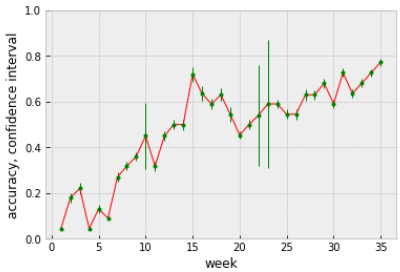

Figure 7: Confidence intervals of the class label C_2.

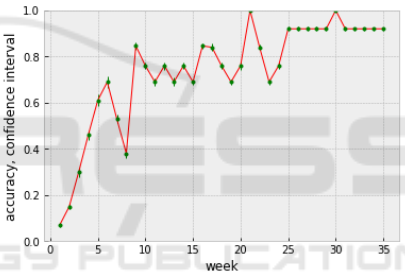

Figure 8: Confidence intervals of the class label C_3.

Indeed, we notice that the number of data points with total trust level gradually increases from week 5 to week 35. Until week 10, the Algorithm assigns the confidence level of the prediction accuracy of the class label C_1 to a fairly large number of data points. We explain this result by the fact that the learners of this class are the most numerous. Indeed, the predictions made by the same system at a given time have different levels of confidence from one another. And as time progresses and the system acquires new information, the confidence level of the prediction of a data point changes and can take several values. Indeed, at each prediction time and based on its TI, we can know if a single prediction is trustworthy or not.

These results show the importance of going through the measurement of TI for this fine granularity of trust to be able to accurately track each learner's prediction and to provide personalized interventions with the less performing learners.

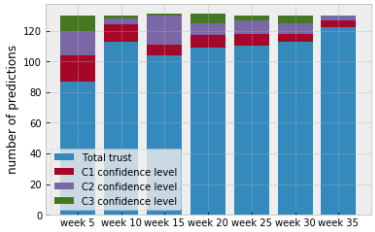

Figure 9: Number of prediction per each confidence level.

6.4 Discussion

In summary, the experimental study yielded the following results.

Following the Algorithm 1, we obtain confidence levels that are more correlated with the overall accuracy of the system. As a result, we have narrow confidence intervals and a more accurate TI of the overall trust granularity that is more representative of the reliability of the system performance. For the same system and at the same prediction time, the TI of the prediction reliability is different from one class label to another. This TI depends on the population of that class and the ability of the system to correctly detect the learners of that class. For the same system and at the same prediction time, the TI of the prediction reliability is different from a single data point to another. In addition, the TI of a same data point changes over time and can take on several values.

7 CONCLUSION

In this paper, we established an approach to measure how far a teacher can trust the predictive performance of an LA system. We started by defining three trust granularities in an LA system, including : the overall trust, trust per class label and trust per each prediction made by the system. Then, for each trust granularity, we proceeded to the calculation of a TI using the confidence level and confidence intervals of statistics. The experimental results show the importance of going through all these trust granularities to get a detailed idea of the reliability of the system performance. In fact, The TI of the overall trust granularity shows the teacher how much she/he can trust the overall performance of a predictive system. Both other trust granularities give a finer idea of how much trust to be granted to the system when it comes to a particular class or prediction. Indeed, at the same prediction time, the TI is different from one class label to another and also from one single prediction to another. This notion of trust comes into play in order to ensure an effective intervention adapted to the situation of each learner or group of learners.

As a perspective of this work we intend to propose a global trust index for the whole system computed from the TI of the three trust granularities.

REFERENCES

Baneres, D., Guerrero-Roldán, A. E., Rodríguez-González, M. E., and Karadeniz, A. (2021). A predictive analytics infrastructure to support a trustworthy early warning system. *Applied Sciences*, 11(13):5781.

Ben Soussia, A., Labba, C., Roussanaly, A., and Boyer, A. (2022). Time-dependent metrics to assess performance prediction systems. *The International Journal of Information and Learning Technology*, 39(5):451–465.

Biedermann, D., Schneider, J., and Drachsler, H. (2018). Implementation and evaluation of a trusted learning analytics dashboard. In *EC-TEL (Doctoral Consortium)*.

Bitkina, O. V., Jeong, H., Lee, B. C., Park, J., Park, J., and Kim, H. K. (2020). Perceived trust in artificial intelligence technologies: A preliminary study. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 30(4):282–290.

Lockey, S., Gillespie, N., Holm, D., and Someh, I. A. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions.

López Zambrano, J., Lara Torralbo, J. A., Romero Morales, C., et al. (2021). Early prediction of student learning performance through data mining: A systematic review. *Psicothema*.

Petty, M. D. (2012). Calculating and using confidence intervals for model validation. In *Proceedings of the Fall 2012 Simulation Interoperability Workshop*, pages 10–14.

Qin, F., Li, K., and Yan, J. (2020). Understanding user trust in artificial intelligence-based educational systems: Evidence from china. *British Journal of Educational Technology*, 51(5):1693–1710.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Siau, K. and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, 31(2):47–53.

Stanton, B., Jensen, T., et al. (2021). Trust and artificial intelligence. *preprint*.

Susnjak, T., Ramaswami, G. S., and Mathrani, A. (2022). Learning analytics dashboard: a tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, 19(1):12.

Tsai, Y.-S., Whitelock-Wainwright, A., and Gašević, D. (2021). More than figures on your laptop:(dis) trustful implementation of learning analytics. *Journal of Learning Analytics*, 8(3):81–100.