A Predictive Model Implemented in KNIME Based on Learning Analytics for Timely Decision Making in Virtual Learning Environments

Benjamín Maraza-Quispe, Enrique Damián Valderrama-Chauca, Lenin Henry Cari-Mogrovejo, Jorge Milton Apaza-Huanca, and Jaime Sanchez-Ilabaca

***Abstract*—The present research aims to implement a predictive model in the KNIME platform to analyze and compare the prediction of academic performance using data from a Learning Management System (LMS), identifying students at academic risk in order to generate timely and timely interventions. The CRISP-DM methodology was used, structured in six phases: Problem analysis, data analysis, data understanding, data preparation, modeling, evaluation and implementation. Based on the analysis of online learning behavior through 22 behavioral indicators observed in the LMS of the Faculty of Educational Sciences of the National University of San Agustin. These indicators are distributed in five dimensions: Academic Performance, Access, Homework, Social Aspects and Quizzes. The model has been implemented in the KNIME platform using the Simple Regression Tree Learner training algorithm. The total population consists of 30,000 student records from which a sample of 1,000 records has been taken by simple random sampling. The accuracy of the model for early prediction of students' academic performance is evaluated, the 22 observed behavioral indicators are compared with the means of academic performance in three courses. The prediction results of the implemented model are satisfactory where the mean absolute error compared to the mean of the first course was 3. 813 and with an accuracy of 89.7%, the mean absolute error compared to the mean of the second course was 2.809 with an accuracy of 94.2% and the mean absolute error compared to the mean of the third course was 2.779 with an accuracy of 93.8%. These results demonstrate that the proposed model can be used to predict students' future academic performance from an LMS data set.**

***Index Terms*—Model, prediction, learning analytics, performance, academic, environments, virtual, learning, KNIME.**

I. Introduction

Currently Learning Management Systems (LMS) store a large amount of data about student interactions in log files, these files generally contain variables in the data, such as the number of logins, the number of accesses to elements of an online course, the number of completed assignments, the number of days in the online course, the activity grades, the period grade, the course grade etc. This data could be interesting for online instructors as it could contain information about the behavior of students that could influence their academic performance [1].

There is a growing opportunity in the educational field, where international efforts seek to improve the quality of education, for which it is necessary to have decision-making support systems that deliver quality and timely information. As mentioned in ref. [2], in a sphere of institutional practice, academic directors and managers are often content only with knowledge derived from their own practice. They tend to be self-sufficient with studies that respond to their purposes and short-term needs. It is here where we seek to intervene, providing tools that allow to contribute and contribute efficiently, with relevant information for these decision-making processes, in this case, making an application limited in scope only to variables that directly impact the retention of the students, basing the analysis on these factors and correlations that are assumed, which affect retention, this in order to be able to intervene in a timely manner with preventive rather than corrective actions [3].

Much of the research in the field of learning analytics has used LMS data to model student performance to predict student grades and to predict which students are at risk of failing a course [4], [5]. This is an important step in learning analytics, as it informs the implementation of interventions, such as personalized feedback. In addition, the question is whether there really is a single best way to predict student performance in a diverse set of courses [6].

However, studies that have used similar methods and predictors have found different results in correlational analyzes and prediction models. Additionally, most studies focus on predicting student performance after completion of a course, establishing how well student performance could have been predicted with LMS usage data, but at a time when the findings can no longer be used for timely intervention [7]. Since LMS data provides information throughout the course, it seems useful to determine whether data from just the first few weeks of a course is sufficient for an accurate prediction of student performance [8]. Therefore, the authors argued that the best time to predict student performance is as soon as possible after the first assessment, as this would be the best compromise between early feedback and sufficient predictive power [9].

Early warning systems use data mining methods to detect students at risk of failing courses at different educational

Manuscript received May 11, 2021; revised August 17, 2021. This research was supported by Universidad Nacional de San Agustín de Arequipa through UNSA INVESTIGA, Contract IBA-CS-07-2020-UNSA.

Benjamín Maraza-Quispe, Enrique Damián Valderrama-Chauca, Lenin Henry Cari-Mogrovejo, and Jorge Milton Apaza-Huanca are with Facultad de Ciencias de la Educación, Universidad Nacional de San Agustín de Arequipa, Arequipa-Perú, Peru (e-mail: bmaraza@unsa.edu.pe, evalderramach@unsa.edu.pe, lunincari@unsa.edu.pe, japazahuan@unsa.edu.pe).

Jaime Sanchez-Ilabaca is with Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Chile (e-mail: jsanchez@unsa.edu.pe).

doi: 10.18178/ijiet.2022.12.2.1591

levels and contexts [10]. According to Knowles [11] early warning indicators provide instructors with an advanced warning that students need help in their learning process. These systems contain predictive models with a collection of variables related to early warning indicators. These variables generally contain information on demographic and institutional data, student characteristics, term or mid-term grades, and LMS interaction data.

Most student achievement prediction models only focus on the accuracy of the prediction results, for this reason, getting an interpretable prediction model can be as important as getting high accuracy in student learning prediction research.

The increase of e-learning resources, instrumental educational software, the use of the Internet in education, and the establishment of state databases of student information has created large repositories of educational data. Traditional educational institutions have used for many year information systems that store plenty of interesting information. Nowadays, web-based educational systems have been rising exponentially and they led us to store a huge amount of potential data from multiple sources with different formats and with different granularity levels [12]. New types of educational environments such as blended learning (BL), virtual/enhanced environments, mobile/ubiquitous learning, game learning, etc. also gather huge amount of data about students. All these systems produce huge amount of information of high educational value, but it is impossible to analyze it manually. So, tools to automatically analyze this kind of data are needed because of all this information provides a goldmine of educational data that can be explored and exploited to understand how students learn. In fact, today, one of the biggest challenges that educational institutions face is the exponential growth of educational data and the transformation of this data into new insights that can benefit students, teachers, and administrators [13].

With the growth of data in virtual online learning environments, researchers and academics are beginning to find ways to make this data understandable and meaningful [14]. Therefore, to analyze and unearth more potential educational information, researchers have further explored learning theory and analysis, frameworks, tools, and practices [15], [16]. In recent years, people have increasingly studied the analytics of learning behaviors and the prediction of student performance has attracted the attention of academics. Since 2013, with the continued development of research and learning analysis, researchers have begun to use machine learning to study predictions of learning [17]. Of course, these benefits from the development of online learning platforms such as MOOCs, and a large number of platform users generate educational data.

Regarding the phenomenon that the number of registered users on the MOOC platform is high and the dropout rate is extremely high, researchers have begun to explore the relationship between user behavior and whether they have dropped out of the course (or whether they can get a certificate). By analyzing user behavior information and predicting learning outcomes, they hope to discover the relationship, in order to take early action to reduce the dropout rate from the MOOC platform [18], [19].

In actual prediction studies, most studies used models of incomprehensible algorithms to predict learning outcomes [20], such as logistic and Bayesian networks. Although such models can accurately predict learning outcomes, they cannot be interpreted. This will undoubtedly have an impact on the implementation of specific interventions. Therefore, to promote the construction of a prediction model and improve the teaching quality of online learning, from the perspective of high interpretation of the prediction results, it is very necessary to build a prediction model of student performance based on online learning behavior analytics.

The learning analysis model is the theoretical basis for the analysis of online learning behavior in the context of Big Data in education. Currently, learning analytics is still in its infancy phase. However, the existing representative learning analysis models have the common characteristics: Data cycle. From the perspective of systems approach analytics. George Siemens provides a cyclical learning analytics model, which includes seven components: data collection, storage, cleansing, integration, analysis, representation and visualization, and action [21]; From the angle of improving teaching and learning, Siemens [21] presents a cyclical model of continuous improvement for learning analytics, which consists of three parts: data collection, information processing and application of knowledge, the whole process is supported by four types of technological resources: computers, theory, people, organizations. In order to explore different approaches to data analysis, Ifenthaler and Widanapathirana [22] proposed a learning analysis framework, which includes ten parts, and the relationship between each part became bidirectional.

With the continued development of learning and analysis technology [23], more new research on slant prediction has emerged in recent years. From the existing research, the learning prediction model can be divided into two categories, one belongs to the black box model, that is, for the prediction result, the reason cannot be seen directly; the other belongs to the white box model, that is, there is a direct explanation of the result of the prediction.

In related learning prediction studies, researchers generally believe that black box prediction is more accurate. Especially when it comes to complex relationships. Black box prediction algorithms often used for research include logistic regression, support vector machines (SVM), and random forest (RF) [18]. Use logistic regression algorithms to predict whether students register courses in intelligent assisted systems (ITS) [18]. Considering the complexity of the research and the difficult data collected on emotional state, motivation, and prior knowledge, the accuracy of the final prediction is close to 70% and the prediction performance is not bad. In the study, Harwati [24] implemented two data mining techniques, clustering analysis and classification analysis. First, they clustered the data of students from the Department of Industry, Universities Islam Indonesia. Next, they used the K-means algorithm, since it is a popular and easy to apply algorithm. Once the optimal number of clusters was defined, they applied classification. They used linear regression and support vector machine (SVM) because all the attributes used in that study were numerical data. The clustered and uncluttered data were evaluated at the classification stage, and then compared with the results based

on root mean squared error (RMSE).

The white box prediction has a higher degree of interpretation, that is, there is a specific reason for the result of the prediction. Of course, when the interpretation is high, the precision of the prediction can be reduced. In the field of education, the white box prediction algorithms that are often used for research include decision trees and random trees. They developed a white box prediction system that predicts student performance in a learning management system (LMS) using the learning time spent on the activity module and the frequency of use of the module [25]. The decision tree was used to develop early prediction systems using four eigenvalues and classifying them into four categories: acceptance behavior, use of online course materials, assignment status, and participation in a discussion forum. The goal of the prediction is the student's score, and the overall prediction accuracy reached 95% [26], after which they combined these techniques with the ADABOOST algorithm. Greater accuracy at 98% (Freund). In this context, classification trees are being used to predict students' academic performance. A decision tree is a flowchart-like structure in which leaf nodes represent class labels and non-leaf nodes represent attributes [27].

The selection of appropriate learning behavior indicators is an important part of prediction. At present, there are many theoretical studies on the selection of learning behavior indicators [28]. These studies cover indicators that may be related to the effect of learning from different perspectives. For example, Brown summarized three main predictor indicators: student characteristics, learning behavior indicators, and student work. He discussed the related predictive capabilities and cases for different types of indicators. He summarized several important indicators of students' prior academic performance, learning history, class participation, and social performance.

Romero and López *et al.* [29] used four indicators of the cumulative percentage of video lectures viewable, the number of forum posts, and the number of users relying on the forum, and the number of views of course progress as predictors; Romero and López *et al.* [29] directly predicted student performance from forum participation. The indicators included the number of messages from students, the number of students creating new topics, the number of students reading stickers, the concentration of students and students, persistence and other indicators.

Writing analytics, as mentioned above, are most efficiently collected by computers. Once stored as data, they may be applied in a variety of circumstances, depending on the complexity and relevance of the metrics. The simplest examples of analytics use are present in Microsoft Word's inline spelling and grammar checking, which simply adds wavy lines beneath misspelled words or improper grammar to point it out. The 'word count' tool of Word is another example of a simple analytical metric [30]. Additionally, LMSs allow teachers to provide and manage these resources in a relatively easy and integrated way.

As every action in an LMS is monitored and stored, insight can be gained into students' online behavior, which in turn can be used to improve learning and teaching. The analysis of LMS data is often referred to as learning analytics [30].

Coronavirus (COVID-19) pandemic has imposed a complete shut-down of face-to-face teaching to universities and schools, forcing a crash course for online learning plans and technology for students and faculty. In the midst of this unprecedented crisis, video conferencing platforms (e.g., Zoom, WebEx, MS Teams) and learning management systems (LMSs), like Moodle, Blackboard and Google Classroom, are being adopted and heavily used as online learning environments (OLEs). However, as such media solely provide the platform for e-interaction, effective methods that can be used to predict the learner's behavior in the OLEs, which should be available as supportive tools to educators and metacognitive triggers to learners [31].

II. Methodology

A. Description of the Context and the Participants

The research has been developed at the Faculty of Education of the National University of San Agustín de Arequipa. The university uses a virtual support platform based on the LMS Moodle. Under this platform, the subjects taught in virtual mode allow, on the one hand, teachers to maintain a repository of information and record of academic activities; and on the other hand, for students this platform allows them to have a practical vision of the learning activities that are programmed in the syllables of the subjects. The training of the model implemented in the KNIME platform has been trained with 1000 student records and 22 behavioral indicators observed in the LMS in three general courses during the first semester of 2020, which were selected through a simple random sampling of a total of 30,000 tested records of general subject data.

B. Instruments and Procedures

The CRISP-DM methodology has been followed, whose standard includes a model and a guide, structured in six phases, and some of these phases are bidirectional, which means that some phases will allow a partial or total review of the previous phases [32]. The phases or levels that are identified in the CRISP-DM methodology are Business understanding, Data understanding, Data preparation, Modeling, Evaluation and implementation, as shown in Fig. 1.

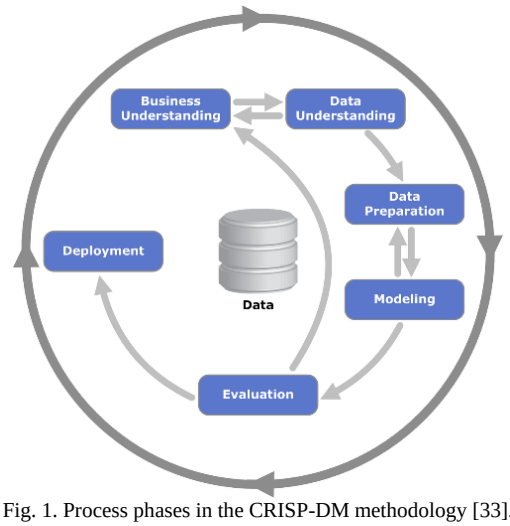

Fig. 1. Process phases in the CRISP-DM methodology [33].

Step 1. Business understanding

The present research aims to implement a predictive model to analyze and compare the prediction of academic performance using LMS data, we analyze whether it is possible to identify students at risk at the beginning of a course and to what extent the model can be used to generate specific interventions.

TABLE I: ONLINE LEARNING BEHAVIOR INDICATORS

Dimensions	Indicators
Prediction	(1) Academic performance
Access	(2) Attendance rate (3) # of records in LMS (4) # of activities carried out at night (5) # of logins (6) Time spent in session (7) Frequency of access to forums (8) # forum posts (9) Frequency of access to resources (10) Frequency of access to glossaries (11) # of days elapsed since last Access Total (12) # of lesson starts
Chores	(13) # of lessons completed (14) # of records in LMS (15) Preparation for evaluation course (16) # of shipments completed total (17) Job consultation frequency (18) # of submitted jobs
Social aspects	(19) Gender (20) Age (21) Parent education (22) Food quality (23) Race
Questionnaires	(24) Average 1 course (25) Average 2 course (26) Average 3 course

Indicators of learning behaviors in virtual learning environments directly affect the accuracy and credibility of predicting student performance. Therefore, the scientific selection of effective learning behavior indicators is an important part of predicting student academic performance [34]. Due to the diversity of online learning behaviors, and the complexity of the correlation between behaviors, not all indicators of learning behaviors that can affect the learning effect can be collected quantitatively. Therefore, based on existing research results, based on the research developed by Zhang and Zhou *et al.* [35], five dimensions are taken as a basis: Prediction, accesses, tasks, social aspects and questionnaires. The 26 learning behavior indicators required for the study were selected as shown in Table I.

Step 2. Data preparation

The IntelliBoard plugins were installed on the university's Moodle platform. IntelliBoard provides analysis and reporting services on the university server, the statistical data collected within the LMS was extracted, the data was stored in spreadsheets, then it was necessary to perform a data pre-processing, as shown in Table II where A general description is made of all the predictor variables and some descriptive statistical data, where when doing a correlation analysis for the three courses, it is shown that 15 of the 16 predictor variables had a statistically significant correlation, with the final grade of the three courses, the exception was the age variable (See Fig. 2).

TABLE II: PREDICTOR VARIABLES AND DESCRIPTIVE STATISTICS SAMPLE

Column	Min	Max	Mean	Std. Deviation	Variance	Skewness	Kurtosis	Overall sum	Row count
Attendance Rate	12	100	56.084	20.606	424.608	-0.007	-0.966	56084	1000
# of logins	8	100	51.668	18.643	347.557	0.297	-0.555	51668	1000
# of activities carried out at right	1	30	15.896	8.355	69.801	0.084	-1.243	15896	1000
# of lessons starts	1	30	16.077	8.289	68.708	-0.003	-1.164	16077	1000
Time Spent in session	11	50	35.155	9.143	83.597	-0.091	-1.114	35155	1000
Frequency of access to forums	2	30	17.074	7.983	63.728	0	-1.242	17074	1000
# forum posts	4	30	17.15	7.761	60.238	-0.012	-1.196	17150	1000
Frequency of access to resources	4	100	50.713	27.663	765.236	0.044	-1.184	50713	1000
Frequency of access to glossaries	4	40	21.678	10.59	112.146	0.011	-1.208	21678	1000
# of days elapsed since last access	0	10	4.915	3.141	9.866	0.049	-1.202	4915	1000
# of records in LMS	0	10	5.406	2.851	8.129	0.058	-1.199	5406	1000
# of lessons completed	5	60	39.735	11.931	142.351	-0.003	-1.11	39735	1000
# of shipments completed	7	60	40.787	11.664	136.05	-0.097	-1.036	40787	1000
# of submitted jobs	10	60	38.875	11.76	138.306	0.069	-1.175	38875	1000
Job consultation frequency	7	100	58.922	23.399	547.527	0.04	-1.137	58922	1000
Age	17	22	19.534	1.717	2.95	0.006	-1.307	19534	1000
Average 1 course	0	100	66.089	15.163	229.919	-0.279	0.275	66089	1000
Average 2 course	17	100	69.169	14.6	213.166	-0.259	-0.068	69169	1000
Average 3 course	7	100	67.987	15.317	234.599	-0.333	0.122	67687	1000

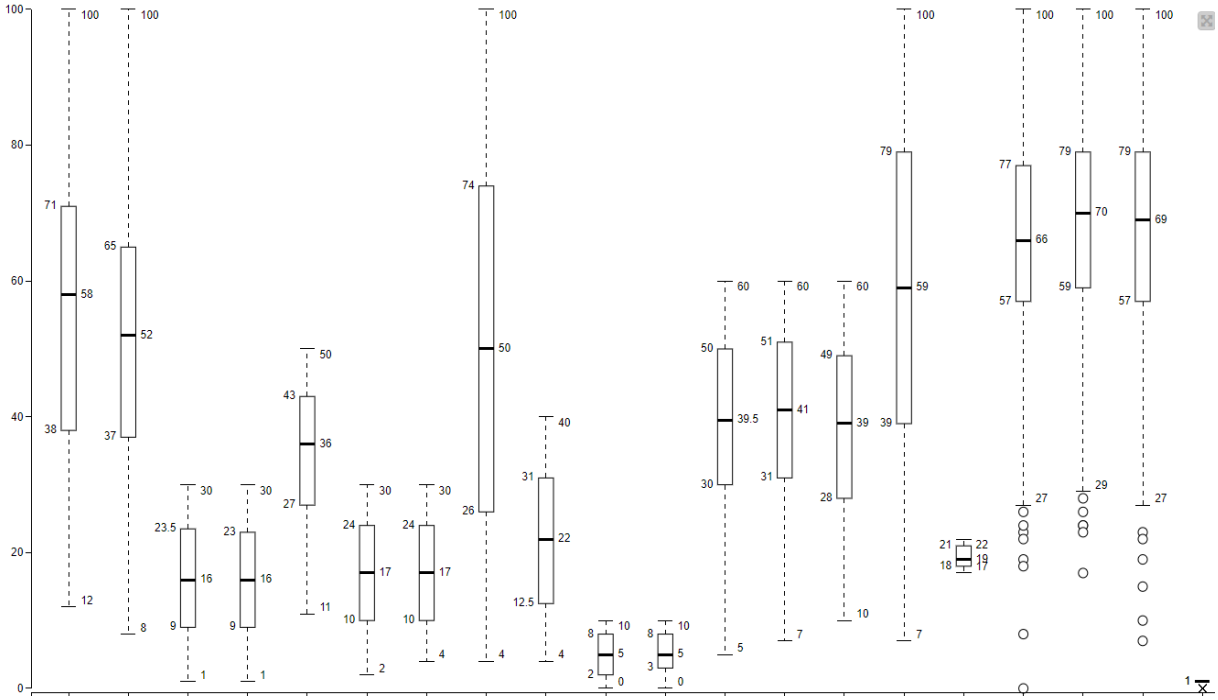

Fig. 2. Data preparation through box plot analysis.

Step 3. Data modeling

- Modeling Technique

Because the KNIME software will be used to perform data mining, some of the modeling techniques offered by this tool are used in accordance with our objectives. Of the models that KNIME offers us, the one that best suits our objectives is a Simple Regression Tree Learner model, since the problems we want to solve are prediction problems and the fields to be predicted contain continuous values (See Fig. 3).

- Generate the Test Plan

The procedure that has been used to test the quality and validity of the model is to use the measures of the "R squared" (R^2), the mean absolute error and the "percentage of the mean absolute error" (mean absolute percentage error). These error measures are automatically calculated by KNIME when you run the model. A previous partition of the 1000 records of the students is made, on the one hand, there is the set of data that will be used to generate the model, called training data, and a second set of data that will be used to perform the tests and measure the quality of the model, called test or evaluation data. The 70% of the data is used for the training data and the remaining 30% for the test data.

- Build the Model

Next, the chosen model is run on the training data. The parameter settings of the model that were chosen in the data-mining tool were described, as well as the output of said model and its description (See Fig. 3).

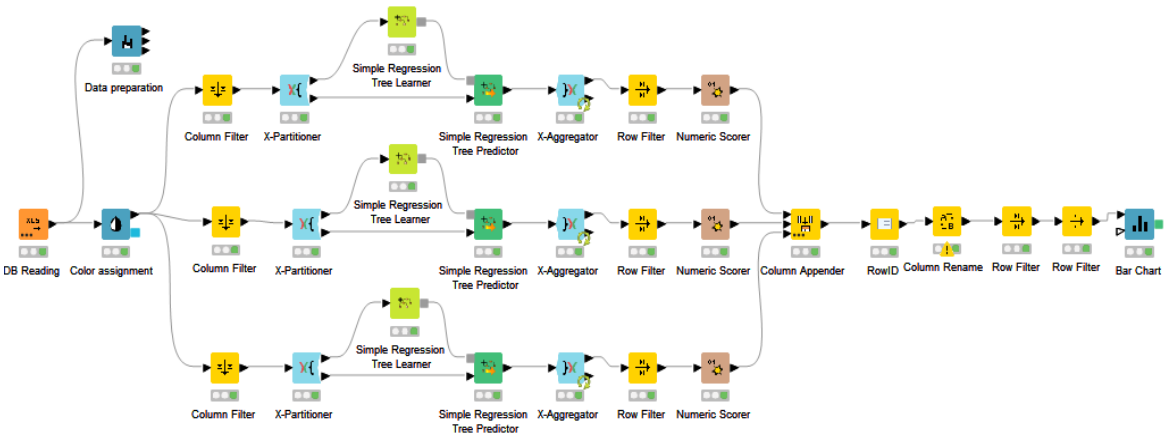

Fig. 3. Predictive model implemented in KNIME.

III. Evaluations of Results

A. Prediction Results

The results showed that the precision of the prediction model presents a mean absolute error compared to the Science area was 3813 and a precision of 89.7%, with the Mathematics area the mean absolute error of 2809 and a precision of 94.2% and with the area of letters the mean absolute error of 2779 and a precision of 93.8%, these results show us that based on the 16 predictor variables considered it is possible to make a prediction of the academic performance of the students, as shown in Fig. 4.

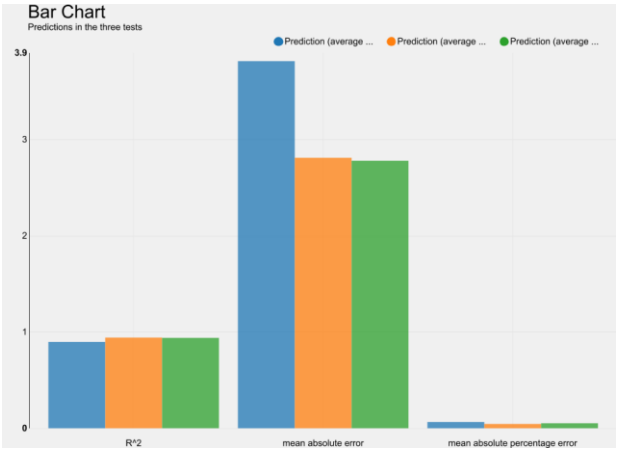

Fig. 4. Precision results in the predictions made by the model.

According to the results shown in the decision tree in Fig. 4, the prediction is made based on the 16 variables and compared with the average of the first course, 900 records are taken from the students where the algorithm takes as a condition the variable access frequency to resources and if this is less than or equal to 47.5, the number of instances correctly classified is 424 records, which corresponds to 47.1%; while if the frequency of access to resources is greater than 47.5, the correct number of instances correctly classified is 476 records, which represents 52.9%; In general, the 6-level decision tree considers the 16 variables to make prediction decisions.

The results of the prediction with the average of the second course implemented through the decision tree is made based on the 16 variables and compared with the average of course two, 900 records are taken from the students where the algorithm takes as a condition the variable number of lessons completed and if this is less than or equal to 145, the number of correctly classified instances is 393 records, which corresponds to 43.7%; while if the number of completed lessons is greater than 14.5, the number of correctly classified instances is 507 records, corresponding to 56.3%; In general, the 6-level decision tree considers the 16 variables to make prediction decisions (See Fig. 5).

The results of the prediction with the average of the third course implemented through the decision tree is performed based on the 16 variables and compared with the average of the third course, 900 records are taken from the students where the algorithm takes as a condition the variable attendance rate and if this is less than or equal to 50.5, the number of instances correctly classified is 371 records, which corresponds to 41.2%; while if the attendance rate is greater than 50.5, the correct number of instances correctly classified is 529 records, which represents 58.8%; In general, the 6-level decision tree considers the 16 variables to make prediction decisions.

TABLE III: RESULT OF SOME PREDICTIONS

ID	Average of the first course	Model prediction	ID	Average of the second course	Model prediction	ID	Average of the second course	Model prediction
16	88	89.25	10	54	52.11	4	75	73.39
53	88	84.80	14	53	56.44	6	92	94.16
61	39	19.17	17	32	24.43	34	82	83.84
64	59	47.33	19	59	61.20	46	62	64.34
65	67	65.15	20	69	63.88	49	82	79.29
72	41	39.00	23	73	74.94	51	68	69.30
111	62	60.22	24	71	76.66	67	74	74.47
137	70	65.15	28	70	67.94	71	63	58.45
139	71	70.36	34	87	86.57	74	41	45.7
148	68	67.55	35	81	76.66	78	72	74.47
160	82	82.19	46	65	63.15	79	68	70.30
175	81	81.13	48	74	71.52	81	45	45.70
176	46	47.77	89	86	82.35	83	63	65.53
180	62	70.73	97	72	71.52	106	100	94.16
183	65	63.58	99	67	67.14	114	100	99.73
201	65	70.79	122	93	90.13	124	73	73.39
210	80	82.19	123	57	57.08	128	67	67.00
Mean	66.71	65.08	Mean	68.41	67.28	Mean	72.18	72.54

TABLE IV: COMPARATIVE ANALYSIS OF PREDICTION RESULTS

Coefficient of determination	Prediction accuracy compared to the first course average	Prediction accuracy compared to the average of the second course	Prediction accuracy compared to the third course average
R^2	0.897	0.942	0.938
Mean absolute error	3.813	2.809	2.779

Table III shows 17 records, a part of the 1000 records that constitute the total sample taken to train the predictive model, the results are satisfactory, where the weighted mean of academic performance in the 22 behavioral indicators is 66.71 and the weighted mean of the prediction made is 65.08.

B. Model Evaluation

The implemented model shows a good accuracy in terms of the predictions made comparing them with the average of the first course presents 89.7% of correct predictions, comparing them with the average of the second course

presents 94.2% of correct predictions and comparing them with the average of the third course presents 93.8% of correct predictions, it can also be appreciated the lowest mean absolute error in the first prediction of 3.813, second prediction of 2.809 and third prediction of 2.779 showing that the proposed model can make efficient predictions (See Table IV).

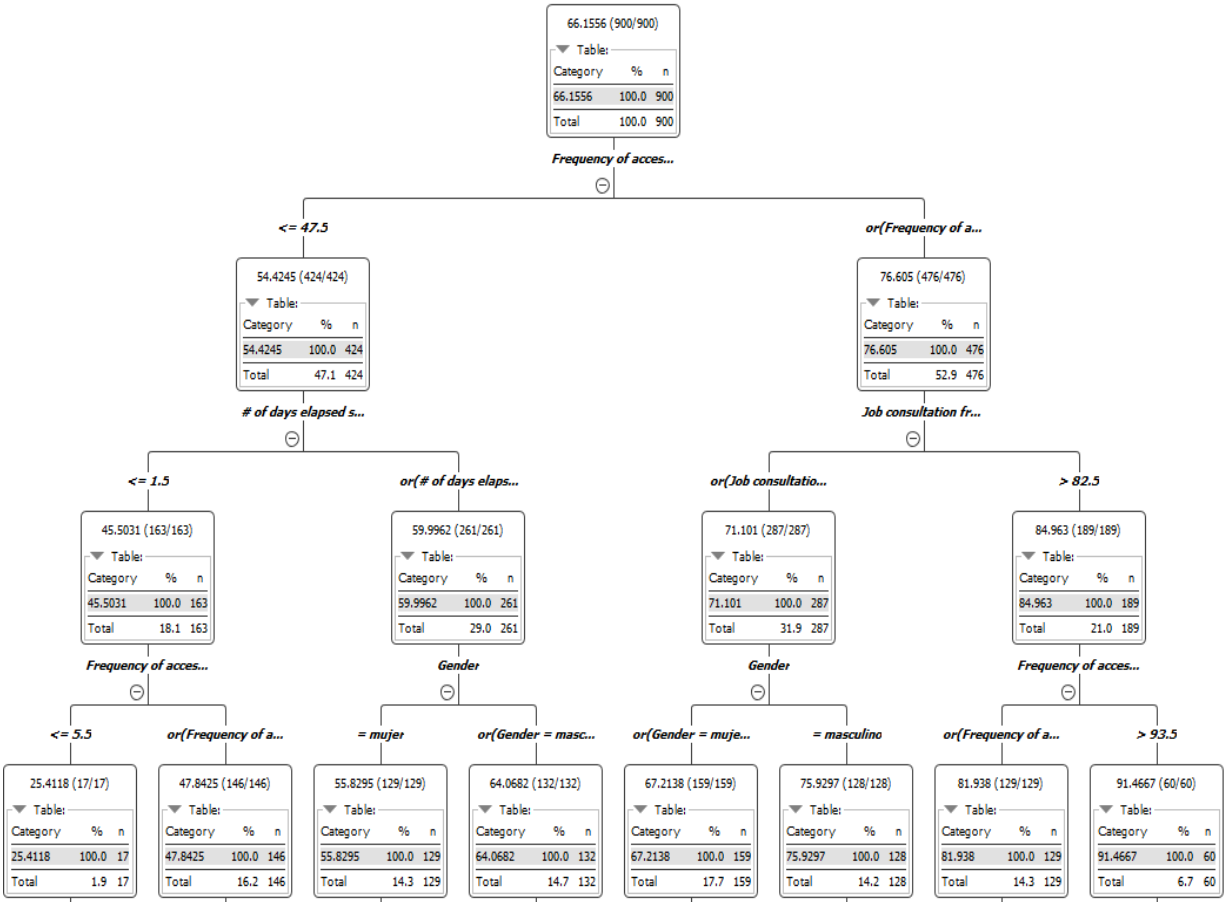

Fig. 5. Tree diagram of decisions compared to the average one.

C. Implementation

In order to implement this project in real business, it would first be necessary to have access to the university's real database, that is, the database that contains all the information related to the university's students. From there, the steps to follow would be the same as those followed in the research from understanding the business to implementation. Although, it must be said that there will be some phases, such as the understanding and preparation of the data, which in the real business will probably be more complex and will take more time than in this project since it can be expected that the real database will have many more records and they contain more noise than in our fictitious database created specifically for this use.

As a supervision and maintenance plan, the following processes could be established:

- Quarterly data extraction and storage, saving the information obtained in spreadsheet format
- Distribution of data according to the data mining software models to work with.
- The results obtained in each data exploitation should be taken into a spreadsheet format and generate graphs of different types for a better visualization and interpretation of the results obtained in each period.

IV. Conclusions

The research implements a predictive model to analyze and compare the prediction of academic performance using LMS data, analyzing whether it is possible to identify at-risk students at the beginning of a course and to what extent the model can be used to generate specific interventions. To make the predictions, the Simple Regression Tree Learner algorithm is implemented in the model based on 1000 real registration data taken from an LMS, working with data from 22 behavioral indicators observed and used in the LMS, which when compared with the averages of academic performance in three courses the results of the predictions are satisfactory, where the mean absolute error compared with the average of the first course was 3.813 and with an accuracy of 89.7%, the mean absolute error compared to the average of the second course was 2.809 with an accuracy of 94.2% and the mean absolute error compared to the average of the third course was 2.779 with an accuracy of 93.8%. These results demonstrate that the proposed model can be used to predict future academic performance of students based on a dataset from an LMS.

The results add to the empirical basis of learning analytics findings and corroborate previous studies on predicting student success, which have also shown different results in

correlations and prediction models, albeit for more varied contexts than our research.

A very important contribution of the proposed model is that it can be scalable and applicable to large databases according to user requirements.

V. Limitations and Future Work

Despite the promising results of the proposed model for user LMS-based prediction, there are certain limitations. In particular, no correlation analysis with the content evaluation result of, for example, questionnaires, intermediate / final exams were carried out. In fact, this was left for a future effort, as the focus here was explore the predictive performance of the model.

Furthermore, the data used here refer to one semester; therefore, data from two or more semesters is required for a better prediction of the academic performance, as well as more thorough data preprocessing.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Benjamin Maraza Quispe designed and implemented the proposal, conducted the experiments, and wrote the paper as the main author. Jaime Sánchez Ilabaca gave the idea of the proposal and supervised the whole activities including the experiments and the paper writing. Enrique Valderrama Chauca and Jorge Apaza Huanca collected the data. Lenin Cari Mogrovejo improved the paper writing. All the authors had approved the final version.

References

[1] J. Bravo-Agapito. S. Romero, and S. Pamplona, "Early prediction of undergraduate student's academic performance in completely online learning: A five-year study," *Computers in Human Behavior*, vol. 5, no. 12, doi.org/10.1016/j. chb.2020.106595, 2020.

[2] C. Guadilla, "Research and decision making in higher Education. Challenges and transformations of education in Latin America," *Nueva Sociedad*, vol. 165, pp. 97-168, 2000.

[3] B. Maraza-Quispe, M. Alejandro-Oviedo *et al*., "Towards a standardization of learning behavior indicators in virtual environments," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, doi: 10.14569, ISSN.2156-5570, 2020.

[4] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 40, no. 6, pp. 601–618, 2010.

[5] S. Shum and R. Ferguson, "Social learning analytics," *Educ. Technol. Soc*., vol. 15, no. 3, pp. 3–26, 2012.

[6] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–17, 2009.

[7] J. Campbell and D. G. Oblinger, "Academic analytics," *Educause*, vol. 42, pp. 40–42, 2007.

[8] D. Tempelaar, B. Rienties, and B. Giesbers, "In search for the most informative data for feedback generation: Learning analytics in a data-rich context," *Computing. Human Behavior*, vol. 47, pp. 157-167, 2015.

[9] M. Richardson, C. Abraham, and R. Bond, "Psychological correlates of university students' academic performance: A systematic review and meta-analysis," *Psychol. Bull*., vol. 138, no. 2, pp. 353-387, 2012.

[10] E. Howard, M. Meehan, and A. Parnell, "Contrasting prediction methods for early warning systems at undergraduate level," *The Internet and Higher Education*, vol. 37, pp. 66-75, 2018.

[11] J. Knowles, "Of needles and haystacks: Building an accurate statewide dropout early warning system in Wisconsin," *Journal of Educational Data Mining*, vol. 7, no. 3, pp. 18-67, 2015.

[12] C. Romero and S. Ventura, "Educational data science in massive open online courses," *WIREs: Data Mining and Knowledge Discovery*, vol. 7, no. 1, p. e1187, 2017.

[13] R. S. Baker, *Big Data and Education*, New York, NY: Teachers College, Columbia University, 2015.

[14] L. De-Marcos, E. García-López, and A. García-Cabot, "Social network analysis of a gamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance," *Computers in Human Behavior*, vol. 60, pp. 312-321, 2016.

[15] J. A. Ruipérez-Valiente, P. J. Muñoz-Merino, and D. Leony, "ALAS-KA: A learning analytics extension for better understanding the learning process in the Khan Academy platform," *Computers in Human Behavior*, vol. 47, pp. 139-148, 2015.

[16] W. D. Greller and M. Hendrik, "Translating learning into numbers: A generic framework for learning analytics," *Educational Technology & Society*, vol. 15, no. 3, pp. 42-57, 2012.

[17] P. Baepler and C. J. Murdoch, "Academic analytics and data mining in higher education," *International Journal for the Scholarship of Teaching & Learning*, vol. 4, no. 2, pp. 267-281, 2008.

[18] P. Mozs, R. Baker, and A. Bowers, "Predicting college enrollment from student interaction with an intelligent tutoring system in middle school," *Langmuir the Acs Journal of Surfaces & Colloids*, vol. 27, no. 11, pp. 6897-6904, 2013.

[19] C. Burgos, M. Campanario, and D. Peña, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," *Computers & Electrical Engineering*, vol. 5, no. 11, 2017.

[20] S. Iglesias-Pradas, "Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning," *Computers in Human Behavior*, vol. 31, no. 2, pp. 542-550, 2014.

[21] G. Siemens, "Learning analytics. The emergence of a discipline," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380-1400, 2013.

[22] D. Ifenthaler and C. Widanapathirana, "Development and validation of a learning analytics framework: Two case studies using support vector machines," *Technology, Knowledge and Learning*, vol. 19, no. 1, pp. 221-240, 2014.

[23] A. Rupp and J. Leighton, *Educational Data Mining and Learning Analytics*, Springer New York, pp. 379-396, 2014.

[24] A. Harwati, "Educational data mining techniques approach to predict student's performance," *International Journal of Information and Education Technology*, vol. 9, no. 2, p. 115, 2019.

[25] L. Macfadyen and S. Dawson, "Mining LMS data to develop an "early warning system" for educators: A proof of concept," *Computers & Education*, vol. 54, no. 2, pp. 588-599, 2010.

[26] Y. Hu, C. Lo, and S. Shihp, "Developing early warning systems to predict students' online learning performance," *Computers in Human Behavior*, vol. 36, pp. 469-478, 2014.

[27] J. Chak and C. Chak, "Measuring students' academic performance through educational data mining," *International Journal of Information and Education Technology*, vol. 10, no. 11, p. 797, 2020.

[28] P. Sinclair, A. Kable, and T. Levett-Jones, "The effectiveness of Internet-based e learning on clinician behavior and patient outcomes: A systematic review," *International Journal of Nursing Studies*, vol. 57, pp. 70-81, 2016.

[29] C. Romero, M. López, and J. Luna, "Predicting students' final performance from participation in on-line discussion forums," *Computers & Education*, vol. 68, pp. 458-472, 2013.

[30] B. Maraza, "Towards personalized learning in virtual environments," *Virtual Campus*, vol. 5, no. 1, pp. 20-29.

[31] G. Siemens and R. S. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in *Proc. 2nd Int. Conf. Learn. Analytics Knowl.*, 2019, pp. 252–254.

[32] S. B. Dias *et al*., "DeepLMS: A deep learning predictive model for supporting online learning in the Covid-19 era," *Sci Rep*, vol. 10, p. 19888, 2020.

[33] P. Chapman, T. Khabaza, and C. Shearer, *CRISP-DM 1.0, Step by Step Data Mining Guide*, Netherlands: SPSS Inc, 2000.

[34] C. Villagrá-Arnedo, F. Gallego-Durán, and F. Llorens-Largo, "Improving the expressiveness of black-box models for predicting student performance," *Computers in Human Behavior*. vol. 1, no. 5, 2016.

[35] W. Zhang, Y. Zhou, and B. Yi, "An interpretable online learner's performance prediction model based on learning analytics," in *Proc. 11th International Conference on Education Technology and Computers*, pp. 148-154, 2019.

Benjamín Maraza-Quispe is a doctor in computer science. He currently works as a research professor at the Faculty of Education Sciences of the National University of San Agustín de Arequipa-Peru.

He is a consultant and lecturer on educational technology issues, as a researcher in this area he has published many research articles in journals indexed to databases such as Web of Science and SCOPUS. He has been recognized by the government of his country with the awards: "Magisterial Palmas in the degree of Teacher in 2016" and "Teacher who leaves a mark in 2018" among other recognitions.

He has also been the winner of several international educational innovation competitions, such as: "The Microsoft World Forum", held in Barcelona in 2014; the "International Science, Technology and Engineering Fair, INTEL-ISEF-2017-2018" held in the USA.

Enrique Damian Valderrama-Chauca is a research professor at the Faculty of Education Sciences of the National University of San Agustín de Arequipa. He is a doctor of education and develops topics of interest for the improvement of teaching-learning processes.

With experience in pre-professional practice and research in the classroom, advisor of different undergraduate and postgraduate research projects, with a quantitative and qualitative approach to research.

Lenin Henry Cari-Mogrovejo is a professor of the Academic Department of Educational Sciences of the National University of San Agustín where he currently works.

He is teacher graduated from the specialty of Physics-Mathematics, postgraduate in systems engineering and second specialty in applied statistics, with a master's degree in accounting and administrative sciences MBA, Doctor in educational sciences from the National University of San Agustín, student at the Universidad del United Nations in agreement with the University of São Paulo, specializing in technologies, São Paulo Brazil.

Jorge Milton Apaza-Huanca is a professor of the Academic Department of Educational Sciences of the National University of San Agustín.

He is a doctor in educational sciences, a proactive person who assumes challenges within the organization with criteria and decision-making. He is a teacher in the area of education administration.

With experience in pre-professional practice and research in the classroom, advisor to different undergraduate and graduate research projects, with a quantitative and qualitative approach to research.

Jaime Hernán Sánchez Ilabaca is an appointed assistant professor of human-computer interaction (1994), associate professor of human-computer interaction (1998) and professor of human-computer interaction (2010) in the Department of Computer Science of the University of Chile.

He works in the HCI area, having introduced new courses and integrating this new area of research and teaching in the Department of Computer Science. As a result, he renew and update the human-computer interaction course (1994-2009) taken by the majority of computer engineering students, as well as computer science master's and doctoral students. Within the human-computer interaction research and teaching area of the Department of Computer Science.