TECHNO REVIEW | Vol. 11, No. 1, 2022 | ISSN 2695-9933
International Technology Science and Society Review / Revista Internacional de Tecnología Ciencia y Sociedad
https://doi.org/10.37467/gkarevtechno.v11.3324

ACADEMIC PERFORMANCE EVALUATION USING DATA MINING IN TIMES OF PANDEMIC

Relationship between Access to the Virtual Classroom and Grades of University Students

EDGAR AURELIO TAYA-ACOSTA, HUGO MANUEL BARRAZA-VIZCARRA, RUTH DE JESUS RAMIREZ-REJAS, EDGAR FAUSTINO TAYA-OSORIO

Universidad Nacional Jorge Basadre Grohmann, Peru

KEYWORDS

Learning analytics
Data analysis
Machine learning
Academic performance
Virtual education

ABSTRACT

This work focuses on studying the relationship that existed between the use of the learning management system (LMS) and the academic performance of the students of the Jorge Basadre Grohmann National University of Tacna-Perú. For this, we use the data provided by the LMS (access virtual classroom) and the university's academic management system (grades). For that, we perform various classification machine learning algorithms to predict academic performance with two classes SATISFACTORY or POOR where Gradient Boosted Trees algorithm had the best accuracy 91.79%. However, with three classes, SATISFACTORY, REGULAR AND POOR, Random Forest algorithm had the best accuracy of 89.26%.

Received: 30/ 04 / 2022
Accepted: 13/ 07 / 2022

1. Introduction

In this context, the generation of transactional data has increased exponentially in the different platforms or virtual learning environments. (Shah et al., 2021; Devasia et al., 2016) that are now used for non-face-to-face education imposed by restrictions imposed by governments in search of reducing the overload of the health sector (World Bank, 2020).

In addition, this increase in the transactional values of accesses to the EVAs provides us with a lot of data that is important to analyze and relate in order to learn from them and predict future behavior, thanks to data mining (Hardman et al., 2013).

On the other hand, we believe that the prediction of academic performance is one of the most important tasks within the field of Learning Analytics (LA) and Educational Data Mining (EDM) (Chatti et al., 2012; Peña-Ayala, 2014). Since it will allow us to know the academic behavior of students and in this way, they can become early warning tools and be able to make adjustment and strengthening decisions to reduce the academic failure of students in times of pandemic.

Machine learning algorithms allow intensive data processing, in addition to learning from them in order to later predict future behavior, especially variables related to access, permanence and interaction with virtual learning environments (Ramírez & Grandón, 2018).

In this we will use classification algorithms such as Naive Bayes, Decision Trees, Random Forest, Gradient Boosted Trees, logistic regression and SVM, which will allow us to propose a model with the highest precision to automatically classify the academic progress of university students.

In the present work we will use the KDD (Knowledge Discovery from Data) (Moulet & Kodratoff, 1995) as a framework to design a data mining model that allows us to capture the data, analyze it, clean it, transform it, relate it, visualize it. We will also use Machine Learning algorithms to be able to predict the behavior of academic performance based on a set of variables related to access to the virtual classroom and the socioeconomic conditions of the students of the Jorge Basadre Grohmann National University (Taya, 2021).

2. Related works

2.1. Learning concepts

In this section of the research work, a review of the most important research in the world on the subject was carried out, so it has collected very interesting works from the best indexed repositories that could be accessed such as: IEEE Xplore, ACM Digital Library, Springer Link among others.

In the work titled “Clarify of the Random Forest Algorithm in an Educational Field”, Ahmed & Hikmat (2018) focus on using the Random Forest classification algorithm to extract useful information from a student dataset and predict their academic progress. For this, the dataset was worked with Weka Tools and the model proposed by the authors. The results show that the accuracy of the proposed method is 83.56% and the accuracy of the technique in WEKA Tool obtained an accuracy of 80.82%. In addition, the proposed method can be executed several times with different results due to the nature of the Random Forest, which each time obtains a random sample of the dataset, this allows choosing the best execution.

In the work titled “Predicting Student Academic Performance using Support Vector Machine and Random Forest”, Alamri et al. (2020) made an interesting comparison of their binary classification and regression models in the prediction of the academic performance of students in the subjects of mathematics and Portuguese (secondary education level) of the model proposed by them with respect to the technique presented in the work "Using Data Mining to predict secondary school student performance" (Cortez & Silva, 2008) obtaining with the Support Vector Machine (SVM) technique 92.43% accuracy with respect to Cortéz's work that had an accuracy of 86.3% and with the Random Forest technique they obtained 91.59% compared to 91.2% in the other study.

Likewise, in the work entitled “Application of machine learning on student data for the appraisal of academic performance”, Alloghani et al. (2019) focused on applying data mining techniques to analyze the academic progress of students from 14 Middle Eastern European countries, including three North African countries. The dataset contains sixteen attributes, twelve of which are categorical and the remaining four numerical and 10 groups of 10 different grades were formed. Three predictive algorithms were used: decision trees, neural networks and naive bayes. As a result, it was obtained

that the CART classifier 98.6% of the students as members of grade G-2. Of the groups G-07 and G-08, the model predicted with an accuracy between 96% and 96.6%, respectively. The work also shows that the accuracy of the Naves Bayes algorithm is 87.1%, that of the neural networks is 93.1%, while the decision tree algorithm had an accuracy of 92.7%.

In the work titled "Student Performance Prediction using Multi-Layers Artificial Neural Networks: A Case Study on Educational Data Mining", Altaf et al. (2019) use neural networks to classify academic behavior from data obtained from the LMS Moodle of around 900 students from 10 university classes. What is interesting about this study is that it trains each neural network with each class, which demonstrates the importance of individual prediction with respect to prediction accuracy. It was also determined that the most important characteristics in the classification were: the degree and the total number of learning sessions. Regarding the performance of predictions with neural networks, an accuracy of: 74.3%, 75.8%, 97.1%, 80%, 90%, 65.2%, 95%, 92, was observed. 1%, 83.1% and 90% for the courses of Introduction to Computing, Fundamentals of Programming, Professional Ethics, Communication Skills, Multivariable Calculus, Numerical Analysis, Artificial Intelligence, Web Engineering, Systems Programming, Visual Programming respectively.

In the work titled "Student Performance Predictor using Multiclass Support Vector Classification Algorithm" Athani et al. (2018) took data from Portuguese schools and grouped into five levels from A to F, where A is the group of students with the best grades and F the group of students who failed. Several algorithms were implemented to classify as: Multiclass Support Vector Machines and neural networks using Weka Tools. It was possible to show that the Multiclass Support Vector Machines were the ones with the best precision using cross validation with a value of 89%.

In the work titled "Interpretable Deep Learning for University Dropout Prediction", Baranyi et al. (2020) design a prediction model based on deep neural networks to predict the final academic performance of students at the Budapest University of Technology and Economics in order to identify students at risk of dropping out, the accuracy they achieved was 72.4% (AUC=0.771) which showed that deep learning is suitable for predicting dropouts.

In the work titled "Features Exploration for Grades Prediction using Machine Learning", Bouchard et al. (2020) present the processing of a very large data of students from the Quebec School Board and applies classification algorithms to predict the final grade of the students, they were tested with several characteristics. On average there was around 75% accuracy.

In the work titled "Learning Models for Student Performance Prediction", Cavazos & Garza (2013) analyzed the Records of Mexican schools are analyzed in three periods: 2014-2017, 2015-2018 and 2016-2019 in 24 subjects, showing that family and motivational aspects are important characteristics when access to historical grades is available. The following results were obtained for the prediction of qualifications when historical information gives qualifications: with linear regression an MAE of 8.9 is obtained in the 3rd bimester, 8.01 in the fourth bimester and 10.24 in the 5th bimester , with neural networks a MAE of 13.99 is obtained in the 3rd bimester, 9.32 in the fourth bimester, 11.85 in the 5th bimester and with SVM a MAE of 8.19 in the 3rd bimester, 6.15 in the 4th bimester and 8.71 in the 5th bimester.

In the work titled "Predicting academic performance of university students from multi-sources data in blended learning", Chango et al., (2019) used different classification algorithms to predict the academic performance of engineering students in mixed learning environments (face-to-face and non-face-to-face) of 65 first-year students of the Electrical Engineering degree at the University of Córdoba (Spain). In general, good algorithm performances were obtained: accuracy between 73% and 82%, F measure between 0.72 and 0.82, and ROC between 0.80 and 0.97, but they find that PART (Partial Decision Tree Algorithm) is definitely the best algorithm.

In the work titled "Student Performance Prediction Model for Early-Identification of At-risk Students in Traditional Classroom Settings", Chanlekha & Niramitranon (2018) focus on comparing the performance of prediction models to identify which students have a tendency to obtain low grades. The data was obtained from the Kasetsart University central registry office containing grades and demographic information for the 10-year Faculty of Engineering between 2008 and 2017. The following results were obtained in terms of precision and r (correlation) value, for the Computing and programming course with the decision tree algorithm it was 48.56% and 56.92% respectively, with the Naive bayes algorithm it was 35.21% and 20.16% respectively, for the algorithm for Random

forest it was 48.77% and 62.06% respectively, for the SVM algorithm it was 48.15% and 47.43% respectively. For the Mathematics for Engineering I course with the decision tree algorithm it was 54.71% and 83% respectively, for the Naive Bayes algorithm it was 50.79% and 53.71% respectively, for the Random Forest 54.89% and 81.92% respectively, for the SVM 55.27% and 85.35% respectively, for neural networks 56.20% and 83.18% respectively. For the Mechanical Engineering I course with the decision tree algorithm it was 61.01% and 59.8% respectively, for the Naive Bayes algorithm it was 58% and 62.81% respectively, for the Random Forest algorithm 60.81% and 59.8% respectively, for the SVM 60.4% and 65.33% respectively, for neural networks 60.81% and 64.32% respectively. For the Mechanical Engineering II course with the decision tree algorithm it was 53.02% and 78.05% respectively, for the Naive Bayes algorithm it was 48.66% and 0% respectively, for the Random Forest algorithm 52.68% and 76.83% respectively, for the SVM 51.68% and 80.49% respectively, for neural networks 56.04% and 76.83% respectively and for the Algorithm Analysis and Design course with the decision tree algorithm was 65% and 100% respectively, for the Naive Bayes algorithm it was 57.5% and 100% respectively, for the Random Forest algorithm 62.5% and 100% respectively, for the SVM 57.5% and 1% respectively, for neural networks 62.5% and 100% respectively.

In the work titled "Predicting student performance using data from an Auto-grading system", Chen & Ward (2020) experiment building linear regression models and decision trees with characteristics of the data obtained from the Marmoset self-assessment system of the University of Waterloo, which includes approval ratios, test results, number of tasks submitted and intervals between submissions.

2.2. Theoretical bases

2.2.1 Academic performance

The academic performance in our research refers to the performance of the student in the term of grades of the different subjects, averaged in a single final grade for each student in the corresponding academic period.

2.2.2. Learning analytics

It refers to the interpretation of a large amount of data produced and obtained from various sources in order to evaluate academic performance, predict performance and detect possible problems. Since its first mention, Learning Analytics (Johnson et al., 2012) has gained increasing relevance. Learning analytics has been identified in various publications as one of the most important trends in technology-enhanced learning and teaching. Therefore, it is not surprising that Learning Analytics is the subject of many scientific articles. Learning Analytics research and improvement involves the development, use, and integration of new processes and tools to improve the teaching and learning performance of individual students and teachers. "Learning Analytics focuses specifically on the learning process" (Siemens & Long, 2014, p. 134). "Due to its connections to digital teaching and learning, Learning Analytics is an interdisciplinary research field with connections to the teaching and learning research field, computer science, and statistics" (Johnson et al., 2013, p. 5). "Available data is collected, analyzed, and the insights gained are used to understand student behavior and provide additional support" (Gašević et al., 2015, p. 64).

2.2.3. University and Pandemic

Higher education faces a scenario of uncertainty and change. In addition to the changes caused by the global pandemic caused by the Sars-Cov2 virus, as well as political and social changes, competition at the university level is increasing.

> Universities share the same challenges as businesses: the need to increase financial and operational efficiency, expand local and global impact, establish new funding models during a changing economic climate, and respond to demands for greater accountability to ensure success organization at all levels. (Jones, 2019, p. 11)

"Higher Education must overcome these external burdens efficiently and dynamically, but also understand the needs of the student body, which represents both the contributor and the donor of this system" (Siemens & Long, 2014, p. 135).

2.2.4. DM (Data Mining)

Data mining is the area of scientific research focused on developing models for discovering within data that comes from educational settings, and using those methods to better understand the behavior of students, teachers, and the environments in which they learn. The recent advent of public educational data repositories has made it possible for researchers to investigate a wide variety of scientific questions using data mining (Baker, 2010; Han et al., 2012).

2.2.5. EDM (Educational Data Mining)

When the data comes from an educational setting, we are dealing with a data mining subdomain called Educational Data Mining or EDM. This is a field of research that applies data mining, statistics, and machine learning to data derived from educational settings. It seeks to extract meaningful information from large amounts of raw data that can be used to improve and understand learning processes (Asif et al., 2017; Baker & Yacef, 2009).

Among the top five EDM approaches: "prediction, clustering, data mining, discovery within models, and data discovery" (Baker, 2010, p. 112). This research work combines two approaches: prediction and classification.

2.2.6. Prediction

In prediction, the goal is to predict the class or label of a data set. An important key application area of prediction in EDM is predicting student academic outcomes.

> Research within this area has been carried out at different levels of granularity: at the tutoring system level, at the course level, at the grade level, etc. At the intelligent tutoring system level, for example, EDM predicts students' exam results by integrating time information and the amount of help a student needs to solve problems. (Feng et al., 2006, p. 32)

"Systems also exist to predict whether a student is likely to perform the next item on their exam correctly and, if so, should be skipped by the tutoring system" (Pardos et al., 2007, p. 435).

> At the subject level, there are proposals that predict the success/failure and grade of students in a course using socioeconomic variables such as age, sex, marital status, nationality, address, scholarship, different abilities, type of university access, type of student (regular, mobility, extraordinary), situation of the student (ordinary, employee, athlete, etc.), years of enrollment, delayed courses, type of dedication (full time, part time) and debt situation. (Strecht et al., 2015, p. 394)

We also found papers that predict students' grades in a programming course by considering different factors such as students' math background, programming aptitude, problem-solving skills, gender, previous experience, math grade of the high school, locality, previous experience in computer programming and use of e-learning (ElGamal, 2013); on the other hand, there are research initiatives that predict course performance based on student performance in prerequisite courses and midterms (Huang & Fang, 2013); others investigated the adequacy of quantitative, qualitative, and social media information on forum use, as well as the adequacy of classical ranking algorithms and clustering algorithms to predict student success or failure in a course (Romero et al., 2013); Some other researchers provide an early intervention solution for difficult courses based on student activity in a learning management system (Arnold & Pistilli, 2012). Various studies predict the pass/fail or overall academic performance of students (full/or partial grades) at the end of a degree program.

2.2.7. Clustering

In clustering, the goal is to group objects into similar object classes. Although clustering has been used in data mining for a wide variety of tasks, an interesting sub-area is grouping students together to study typical behavior patterns. For example, there is a study that finds typical behaviors in forums such as high-level workers, that is, students who read all messages and post many messages in the forum, or curious, that is, students who read all messages without posting any (Cobo et al., 2012, p. 249).

Other works identify groups of students with similar performance from kindergarten to the end of high school (Bowers, 2010); while other works group student interaction data to build student profiles (Talavera & Gaudioso, 2004).

2.2.8. Database Knowledge Discovery (KDD)

"Database Knowledge Discovery (KDD) is developed from a research approach that involves various areas such as databases, machine learning, pattern recognition, statistics, information theory, artificial intelligence, data visualization" (Riquelme et al., 2006, p. 13). The main feature of KDD is to extract knowledge from the database and data warehouse. This knowledge is implicit information, previously unknown, potentially useful and easily understood. KDD is a branch of computer science gradually developed in recent years and a new attempt in the field of artificial intelligence. KDD has been successfully used in industrial, agricultural, military, financial, and commercial aspects, and has become one of the current approaches to computing. Currently, KDD is frequently used as a framework in research tasks description, knowledge evaluation and knowledge representation.

> Effective knowledge discovery algorithm is the key. Specifically, that is mining knowledge such as association rules, data grouping, classification rules, sequential pattern, similar pattern, chaotic pattern, etc. with the previous method and its integration technologies in all types of databases in the real world (relationships, interpretation, temporal, spatial, distributed, object-oriented).(Chen et al., 2011, p. 1285)

The sequence of KDD processes is the one displayed in figure 2.

3. Methodology

3.1. Research type and design

The present study is of an explanatory Correlational type. A description of the access data and academic performance of the academic periods 2019-I and 2020-I was made, in order to compare them and determine differences.

This research describes the differences between the accesses to the virtual classroom of the 2019-I and 2020-I periods, in the same way the academic performance data of the 2019-I and 2020-I academic periods were described.

Correlational research is responsible for determining the degree of association between two or more variables in the same sample of subjects, in our case of students.

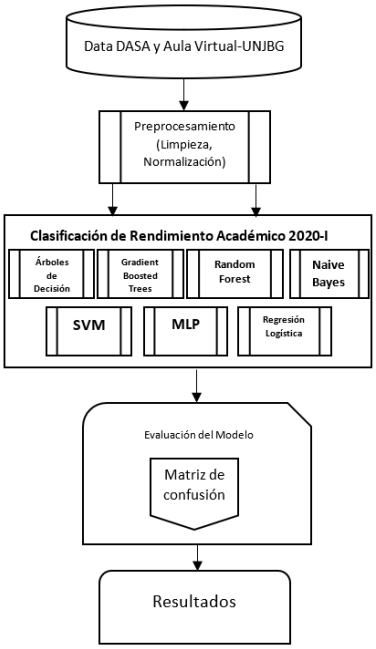
In the present study, we want to know if there is an association between access to the virtual classroom and academic performance in the academic period 2020-I. In simple terms, the methodology of this research work is summarized in figure 1.

3.2. Population and sample

The study population had several components distributed among the records of all the qualifications corresponding to all the students enrolled at UNJBG in the academic periods 2019-I and 2020-I, which make a total of 130,981 records, another component of our population was the records corresponding to the accesses to the Virtual Classroom of all UNJBG students in the academic periods 2019-I (1,148,710 records) and 2020-I (19,116,793 records) that make a total of 20,265,503 records, another component of the population were the 14,430 records of the socioeconomic data of the UNJBG students, data that were provided by the Central Registry Office of the UNJBG, the Virtual Classroom

Committee and the Academic Directorate of University Welfare of the same University corresponding to the academic period 2020-I.

Figure 1. Proposed methodology.



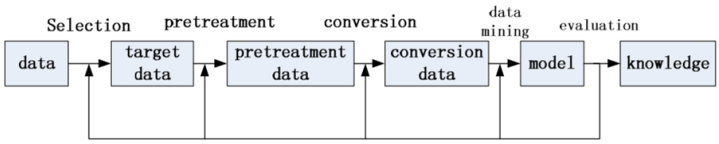
Source: Taya, 2021, p. 36

4. Results

4.1. Application of KDD

In order to order and have a better idea of our results, we will align them to the KDD format whose stages are defined in Figure 2.

Figure 2. KDD framework.



Source: Chen, et al., 2011, p. 1285

Once the necessary adjustments have been made to the data, we can summarize the results of the prediction of the target variable (OBSERVATION=Satisfactory, Poor) in Table 1.

Table 1. Classification results with different algorithms with two class values (Satisfactory and Poor)

	Overall accuracy	Overall mistake	Cohen's kappa (k)	Correctly classified
Gradient Boosted Trees	91.11 %	8.89 %	0.722	41
Random Forest	86.67 %	13.33 %	0.444	39
Decision Tree	84.44 %	15.56 %	0.493	38
Naive Bayes	84.44 %	15.56 %	0.493	38
Fuzzy Rule	82.93 %	17.07%	0.429	34
Tree Ensemble	88.89%	11.11%	0.561	40

Source: Taya, 2021, p. 58

The same objective variable was also predicted with three values (Satisfactory, poor and very poor), obtaining the results shown in Table 2.

4.2. Descriptive results

In this section, the descriptive results of the variables are plotted using Tableau 2021.2.0 as a tool to plot the data.

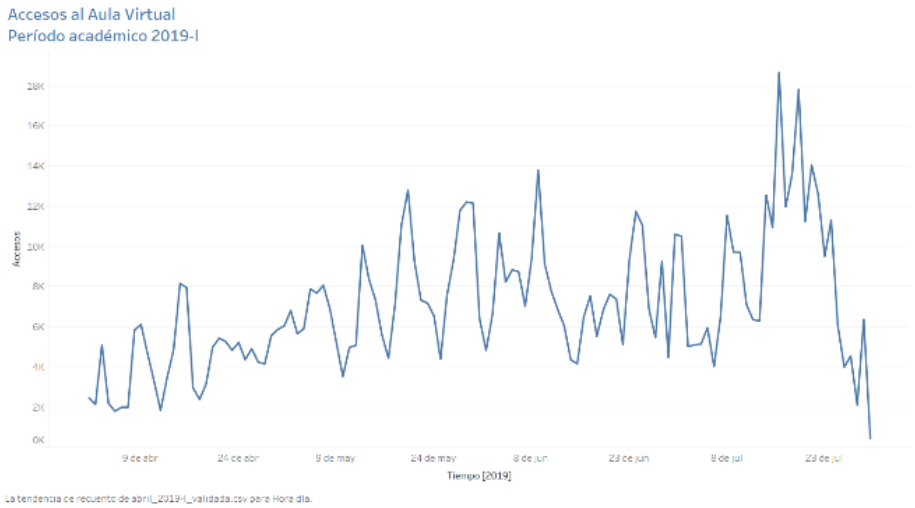
Table 2. Classification results with three class values (Satisfactory, Poor, Very Poor)

	Overall accuracy	Overall mistake	Cohen's kappa (k)	Correctly classified	Misclassified
Gradient Boosted Trees	88.68%	11.32%	0.632	47	6
Random Forest	83.02%	16.98%	0.297	44	9
Decision Tree	88.46%	11.54%	0.631	46	6
Naive Bayes	49.06 %	50.94%	0.110	26	27
Fuzzy Rule	78.26 %	21.74%	0.204	36	10
Tree Ensemble	83.02%	16.98%	0.267	44	9

Source: Taya, 2021, p. 58

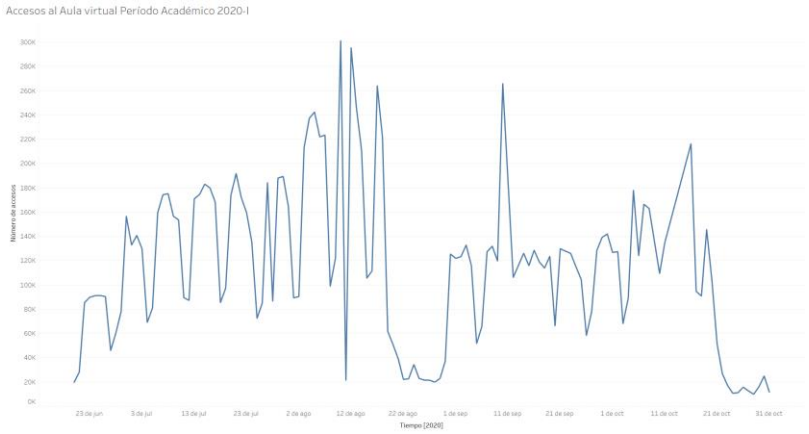
According to what figure 3 shows, we can notice that in the period from May 16 to 22, there is a slight increase in accesses, unlike the previous months, this is explained by the period of the first evaluation of the I semester of 2019, we also notice that there is a greater increase in the period from July 12 to 24, which corresponds to the second evaluation and substitute exam, we can affirm based on this evidence that those dates must reinforce the security measures that ensure the availability of virtual classroom services.

Figure 3. Total accesses to the virtual classroom academic period 2019-I.



Source: Taya, 2021, p. 59

Figure 4. Total accesses to the virtual classroom academic period 2020-I.



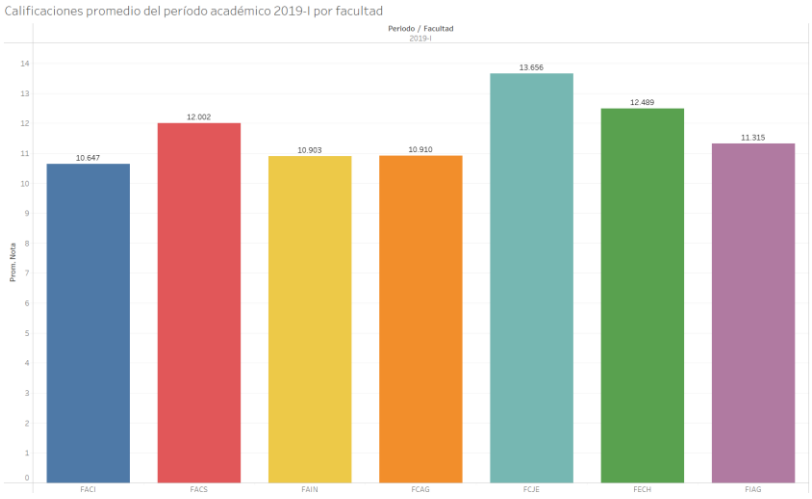
Source: Taya, 2021, p. 60

Figure 4 clearly shows that there is a greater use from Monday to Friday, decreasing on weekends. A significant increase in accesses between August 7 and 12 is also clearly noticeable, which coincides, according to the UNJBG Academic Calendar for the years 2020, endorsed by University Council Resolution No. 16862-2020-UN/JBG, with the first evaluation, except on August 11, which coincides with a total crash of the virtual classroom system.

We also noticed a considerable decrease in accesses between August 19 and 30, this is associated with the suspension of academic work due to a high rate of contagion among teachers and students, in accordance with University Council resolution No. 16784-2020-UN/JBG. We can also appreciate an increase in accesses in the second week of October that is clearly associated with the second evaluation of subjects. Finally, we noticed a final and drastic decrease in the week of October 20 to 31, which is the week of delivery of final grades and registration for the next cycle.

In figure 5, it can be seen that in the academic period 2019-I, the Faculty of Engineering has a weighted average of 10.903, which positions it as the penultimate Faculty in terms of student grades.

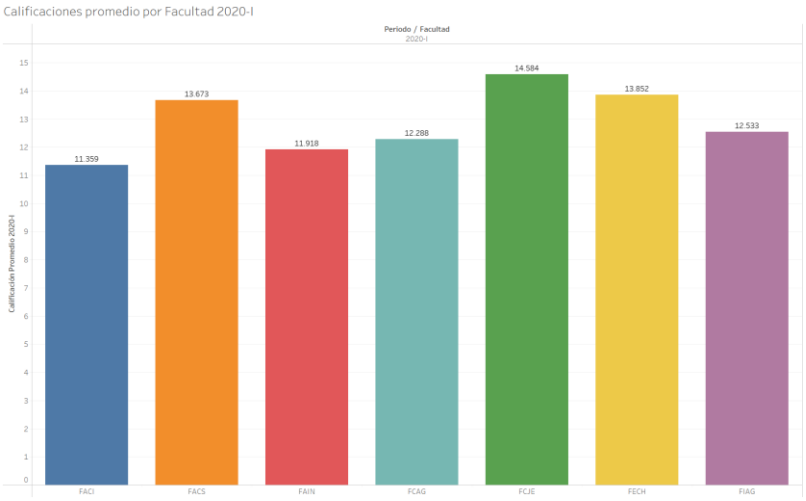
Figure 5. Academic performance by Faculty in the period 2019-I.



Source: Taya, 2021, p. 61

On the other hand, according to figure 6, in the academic period 2020-I we can see that the Faculty of Engineering raises its general grade point average reaching 11.918, but remains the penultimate Faculty in terms of academic performance.

Figure 6. Academic performance by Faculty in the period 2020-I.



Source: Taya, 2021, p. 62

4.3. Results of the applied models

In this section of the research work we will describe in detail each of the results of the application of the academic performance classification algorithms in the academic period 2020-I, which is the subject of our research, all this in search of the most accurate and appropriate model for the characteristics of our students, and the conditions of non-attendance at the Jorge Basadre Grohmann National University of Tacna.

Before showing the results, it is appropriate to remember that, in order to apply the algorithms, we partitioned the data into two groups, the first consisting of 80%; that is to say around 2774 registers that are used for the training of the classifier and the second one composed by 20%; that is, around 682 records for testing, approximately.

As we can see in Figure 7, the application of the Random Forest algorithm on the preprocessed data can be seen. If we observe the confusion matrix, this algorithm had an accuracy of 25.37 classifying "DEFICIENT" returns and 97.72 classifying "SATISFACTORY" returns. Of a total of 682 records tested, it classified 618 correctly and 64 incorrectly, achieving a general accuracy of 90.62%.

Figure 7. Random Forest confusion matrix.

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	SATISFACTORIO (Predicted)	
DEFICIENTE (Actual)	17	50	25.37%
SATISFACTORIO (Actual)	14	601	97.72%
	54.84%	92.32%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
90.62%	9.38%	0.304	618	64

Source: Taya, 2021, p. 77

Likewise, as we can see in figure 8, we appreciate the confusion matrix of the application of the Gradient Boosted Trees algorithm where it shows a 31.34% accuracy in the classification of "DEFICIENT" academic performance and a 98.37% in the classification of academic performance "SATISFACTORY", of a total of 682 records tested, correctly classified 626 and incorrectly 56, achieving a general accuracy of 91.79%.

Figure 8. Gradient Boosted Trees confusion matrix.

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	SATISFACTORIO (Predicted)	
DEFICIENTE (Actual)	21	46	31.34%
SATISFACTORIO (Actual)	10	605	98.37%
	67.74%	92.93%	

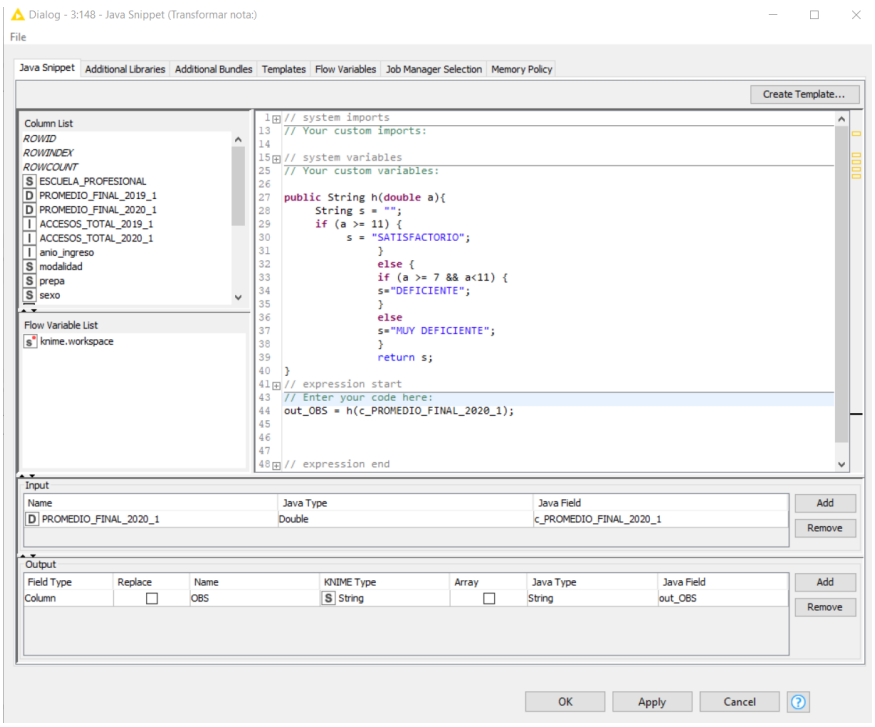
Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
91.79%	8.21%	0.391	626	56

Source: Taya, 2021, p. 78

Three grade levels were also experimented with: SATISFACTORY (final grade greater than or equal to 11), DEFICIENT (final average greater than or equal to 7 and less than 11) and VERY POOR (final average less than 7) as shown in the figure 9.

Figure 9. Determination of observation values according to final average.



Source: Taya, 2021, p. 80

This was done in order to determine if there is any variation in the precision of the algorithms, by expanding the class values to 3 and we obtained the following results:

For example, for the case of the decision tree algorithm according to Figure 10, we could see that there were 26.15% of successes classifying POOR ratings, 35.14% of successes in classifying VERY POOR ratings and 92.56% of correct answers classifying SATISFACTORY qualifications. On the other hand, out of a total of 908 records used for testing the generated model, 776 records could be classified correctly and 132 incorrectly, finally achieving a general accuracy of 85.46%.

Figure 10. Confusion matrix of the decision tree algorithm with 3 class values.

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	MUY DEFICIENTE (Pr...	SATISFACTORIO (Pre...	
DEFICIENTE (Actual)	17	3	45	26.15%
MUY DEFICIENTE (Ac...	7	13	17	35.14%
SATISFACTORIO (Act...	44	16	746	92.56%
	25.00%	40.63%	92.33%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
85.46%	14.54%	0.285	776	132

Source: Taya, 2021, p. 81

Likewise, for the case of the Random Forest algorithm according to figure 11, it was possible to verify that it had a success rate of 4.48% classifying ratings "POOR", a success rate of 30% classifying ratings "VERY POOR" and a 99.03% classifying qualifications "SATISFACTORY". In addition to a total of 931 records used for testing the model, it was able to correctly classify 831 and erroneously 100, obtaining a total accuracy of 89.26%.

Figure 11. Random Forest algorithm confusion matrix with 3 class values.

Scorer View

Confusion Matrix

	DEFICIENTE (Predicted)	MUY DEFICIENTE (Pr...	SATISFACTORIO (Pre...	
DEFICIENTE (Actual)	3	0	64	4.48%
MUY DEFICIENTE (Ac...	0	12	28	30.00%
SATISFACTORIO (Act...	4	4	816	99.03%
	42.86%	75.00%	89.87%	

Overall Statistics

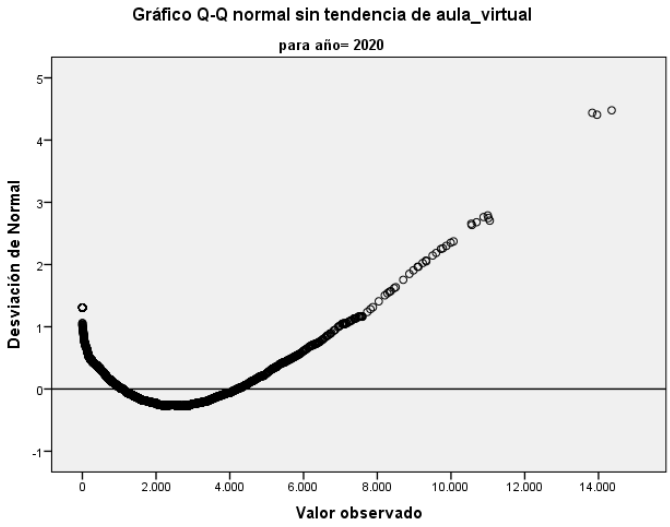
Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
89.26%	10.74%	0.207	831	100

Source: Taya, 2021, p. 82

4.4. Results at the inferential level

Regarding access to the virtual classroom in the academic period 2020-I, we determined that it does not have a normal distribution after applying the K-S test, as shown in Figure 12.

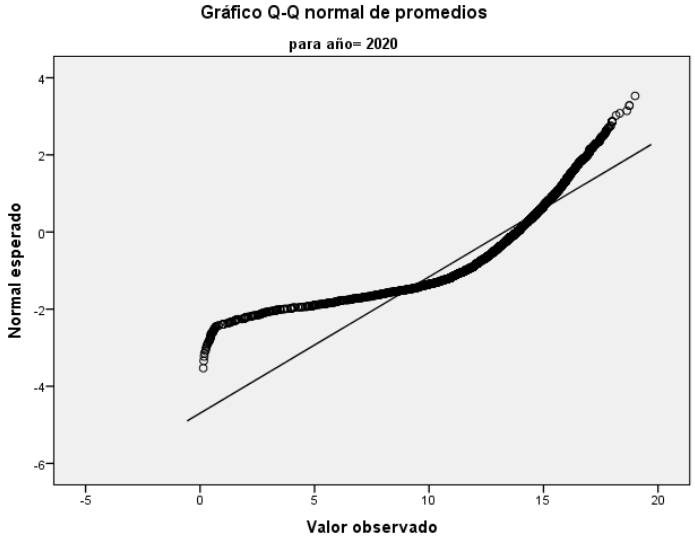
Figure 12. Q-Q graph of the test of normality of access to the virtual classroom in the academic period 2020 - I.



Source: Taya, 2021, p. 82

In the case of the final average values regarding student performance in the 2020-I academic period, we can see in Figure 13 that they do not behave normally.

Figure 13. Q-Q graph of the normality test of the final average of the grades in the academic period 2020 - I



Source: Taya, 2021, p. 91

5. Conclusions

Regarding the data analysis using data mining regarding the use of the virtual classroom of the students of the Faculty of Engineering-FAIN of the UNJBG in the academic period 2020-I, it is possible to determine that the number of accesses to the virtual classroom of the students of the Faculty of Engineering is: 2,063,047 and they constitute the fourth faculty with the most accesses, after the Faculty of Agricultural Sciences-FCAG with 3,072,320 accesses, Faculty of Health Sciences-FACS with 2,897,235 accesses and Faculty of Legal and Business Sciences-FCJE with 2,402,332 accesses, unlike the period 2019-I where the Faculty of Engineering barely had 126,193 accesses, showing an exponential increase of 1,534.83% between 2019- I and 2020-I in terms of access to the virtual classroom.

On the other hand, in the analysis we used various data mining techniques using various tools such as: Python, KNIME, Tableau, RapidMiner, which were very helpful in facilitating the work. Regarding the academic performance of the 2020-I period, we have been able to show that the Faculty of Engineering has the penultimate average in grades reaching 11.918, unlike the Faculty of Legal and Business Sciences, which with an average of 14.584 that is in the first UNJBG qualification site. And in the period 2019-I we were able to show that the Faculty of Engineering reached an average of 10.903 and the Faculty of Legal and Business Sciences reached 14,584, noting an increase in the average of the Faculty of Engineering in the academic period 2020-I of 1.015 points, which represents an increase of 9.31% compared to 2019-I and a decrease in the grades in the Faculty of Legal and Business Sciences. We believe that it is due to the fact that the use of the virtual classroom did not have a greater effect on the grades in the Letters and Business Degrees of the Faculty of Law and Business Sciences, however in the Faculty of Engineering it did, since most of the courses use laboratories, we think that it restricted the professors to be more demanding in the evaluations and in any case they were more condescending when it came to grading.

Regarding the prediction model based on data mining, it was possible to implement a model based on the Gradient Boosted Trees algorithm with the highest general accuracy, correctly predicting student grades based mainly on access to the classroom virtual, with an accuracy of 91.79% for two class values (SATISFACTORY AND POOR) and the model that produced the highest precision in the classification of grades with three class values (SATISFACTORY, POOR AND VERY POOR) was the model generated Based on the Random Forest algorithm with an accuracy of 89.26%, we believe that this difference in accuracy is due to class value segmentation, since the VERY POOR and POOR classes are considerably smaller than the SATISFACTORY class.

Finally, from the inferential point of view, it has been statistically shown that access to the virtual classroom in the academic period 2019-I and 2020-I do not follow a normal distribution and applying

the Mann Whitney test, we have shown that access to the virtual classroom in academic period 2020 - I was greater than access to the virtual classroom in the academic period 2019 - I. In the same way, using the same test, we showed that the grades for both the 2019-I and 2020-I periods did not fit a normal distribution and we were able to show that the grades in the 2020-I academic period were higher than the grades in the 2019-I academic period, which allows us to affirm that the level of access to the virtual classroom did have a significant positive effect on the students' grades.

This research has the particularity of showing us different results when applying different algorithms, since the models obtained are applicable to our reality. It was possible to train with 2,724 student records and test 682 student records for the classification of academic performance based on various variables, but mainly access to the virtual classroom in the 2020-I period, managing to determine that access to the virtual classroom had a significantly high effect on improving academic performance.

In contrast to the research entitled "Clarify of the Random Forest Algorithm in an Educational Field" (Ahmed & Hikmat, 2018, p. 183) reviewed in our background, where they manage to generate a model from the Random Forest algorithm obtaining an accuracy of 83.56% , in our research we managed to reach 89.26% accuracy, which we consider to be due to the most thorough analysis of the data, consideration of the previous ratings specifically for 2019-I as well, and prior cleaning of them.

On the other hand, according to the work entitled "Predicting Student Academic Performance using Support Vector Machine and Random Forest" (Alamri et al., 2020, p. 106) also referenced in our background, the researchers manage to obtain an accuracy of 94.43% using a model based on in the SVM algorithm and an accuracy of 91.59% with a model based on the Random Forest algorithm in contrast to our results of 90.47% using SVM and 90.62% using Random Forest. In this case, we consider that this difference is due to the fact that Alamri focused his study on only two subjects (Mathematics and Portuguese) unlike us, which we apply to all subjects; consequently, we work with much more.

References

Ahmed, N., & Hikmat, M. (2018). Clarify of the Random Forest Algorithm in an Educational Field. *ICOASE 2018 - International Conference on Advanced Science and Engineering*, 179–184. https://doi.org/10.1109/ICOASE.2018.8548804

Alamri, L., Almuslim, R., Alotibi, M., Alkadi, D., Ullah, I., & Aslam, N. (2020). Predicting Student Academic Performance using Support Vector Machine and Random Forest. *ACM International Conference Proceeding Series*, *PartF16898*, 100–107. https://doi.org/10.1145/3446590.3446607

Alloghani, M., Al-Jumeily, D., Hussain, A., Aljaaf, A., Mustafina, J., & Petrov, E. (2019). Application of machine learning on student data for the appraisal of academic performance. *Proceedings - International Conference on Developments in ESystems Engineering, DeSE*, *2018-Septe*, 157–162. https://doi.org/10.1109/DeSE.2018.00038

Altaf, S., Soomro, W., & Rawi, M. (2019). Student Performance Prediction using Multi-Layers Artificial Neural Networks: A case study on educational data mining. *ACM International Conference Proceeding Series*, 59–64. https://doi.org/10.1145/3325917.3325919

Arnold, K., & Pistilli, M. (2012). Learning Analytics to Increase Student Success. *2nd International Conference on Learning Analytics and Knowledge*, *May*, 267–270.

Athani, S., Kodli, S., Banavasi, M., & Hiremath, P. (2018). Student performance predictor using multiclass support vector classification algorithm. *Proceedings of IEEE International Conference on Signal Processing and Communication, ICSPC 2017*, *2018-Janua*(July), 341–346. https://doi.org/10.1109/CSPC.2017.8305866

Baker, R. (2010). Data Mining. *International Encyclopedia of Education*, 112–118. https://doi.org/10.1016/B978-0-08-044894-7.01318-X

Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3–16.

Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable Deep Learning for University Dropout Prediction. *SIGITE 2020 - Proceedings of the 21st Annual Conference on Information Technology Education*, 13–19. https://doi.org/10.1145/3368308.3415382

Bouchard, K., Gonzales, L., Maitre, J., & Gaboury, S. (2020). Features Exploration for Grades Prediction using Machine Learning. *ACM International Conference Proceeding Series*, 78–83. https://doi.org/10.1145/3411170.3411232

Bowers, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment, Research and Evaluation*, *15*(7).

Cavazos, R., & Garza, S. (2013). Learning Models for Student Performance Prediction. In *Springer International Publishing, 17*(12). https://doi.org/10.1007/s00500-013-1159-7

Chango, W., Cerezo, R., & Romero, C. (2019). Predicting academic performance of university students from multi-sources data in blended learning. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3368691.3368694

Chanlekha, H., & Niramitranon, J. (2018). Student performance prediction model for early-identification of at-risk students in traditional classroom settings. *MEDES 2018 - 10th International Conference on Management of Digital EcoSystems*, 239–245. https://doi.org/10.1145/3281375.3281403

Chatti, M., Dyckhoff, A., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5–6), 318–331. https://doi.org/10.1504/IJTEL.2012.051815

Chen, H., & Ward, P. (2020). Predicting student performance using data from an auto-grading system. *CASCON 2019 Proceedings - Conference of the Centre for Advanced Studies on Collaborative Research - Proceedings of the 29th Annual International Conference on Computer Science and Software Engineering*, 234–243. https://doi.org/10.5555/3370272.3370297

Chen, Y., Wang, Y., Xiao, X., & Shi, M. (2011). Knowledge discovery technology based on access information mining on knowledge warehouse. *Proceedings - 2011 8th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2011*, *2*, 1285–1288. https://doi.org/10.1109/FSKD.2011.6019701

Cobo, G., García-Solórzano, D., Morán, J., Santamaría, E., Monzo, C., & Melenchón, J. (2012). Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. *ACM International Conference Proceeding Series*, *May*, 248–251. https://doi.org/10.1145/2330601.2330660

Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *15th European Concurrent Engineering Conference 2008, ECEC 2008 - 5th Future Business Technology Conference, FUBUTEC 2008*, *2003*(2000), 5–12.

Devasia, T., Vinushree, T., & Hegde, V. (2016). Prediction of students performance using Educational Data Mining. *Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016*, 91–95. https://doi.org/10.1109/SAPIENCE.2016.7684167

ElGamal, A. (2013). An Educational Data Mining Model for Predicting Student Performance in Programming Course. *International Journal of Computer Applications*, *70*(17), 22–28. https://doi.org/10.5120/12160-8163

Feng, M., Heffernan, N., & Koedinger, K. (2006). Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required. In *In M. Ikeda, K. D. Ashley, y T.-W. Chan (Eds.), 8th International Conference, ITS 2006* (Vol. 3744).

Gašević, D., Dawson, S., Siemens, G., Gašević, B., & Dawson, S. (2015). art%3A10.1007%2Fs11528-014-0822-x. *TechTrends*, *59*(1), 64–71. https://bit.ly/3zyx4qP

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm. *Systems Research and Behavioral Science*, *30*(2), 194–203. https://doi.org/10.1002/sres.2130

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*, *61*(1), 133–145. https://doi.org/10.1016/j.compedu.2012.08.015

Johnson, L., Adams, S., & Cummins, M. (2012). *The NMC Horizon Report: 2012 Higher Education Edition*.

Johnson, L., Adams-Becker, S., Cummins, M., Freeman, A., Ifenthaler, D., & Vardaxis, N. (2013). Technology Outlook for Australian Tertiary Education 2013-2018: An NMC Horizon Project Regional Analysis. *In New Media Consortium.* https://eric.ed.gov/?id=ED559378

Jones, K. (2019). Learning analytics and higher education: a proposed model for establishing informed consent mechanisms to promote student privacy and autonomy. *International Journal of Educational Technology in Higher Education*, *16*(1). https://bit.ly/3b9IJ6b

Moulet, M., & Kodratoff, Y. (1995). From machine learning towards knowledge discovery in databases. *IEE Colloquium on Knowledge Discovery in Databases (Digest No. 1995/021 (A))*, 1–3.

Pardos, Z., Heffernan, N., Anderson, B., & Heffernan, C. (2007). The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks. In *In C. Conati, K. McCoy, y G. Paliouras (Eds.), User Modeling 2007: Vol. 4511 LNCS*. https://bit.ly/3zv9prp

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4 PART 1), 1432–1462. https://doi.org/10.1016/j.eswa.2013.08.042

Ramírez, P., & Grandón, E. (2018). Prediction of student dropout in a Chilean public university through classification based on decision trees with optimized parameters. *Formacion Universitaria*, *11*(3), 3–10. https://doi.org/10.4067/S0718-50062018000300003

Riquelme, J., Ruiz, R., & Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial. *Revista Iberoamericana de Inteligencia Artificial,* 11–18.

Romero, C., López, M., Luna, J., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers and Education*, *68*, 458–472. https://doi.org/10.1016/j.compedu.2013.06.009

Shah, S., Shah, A., Memon, F., Kemal, A., & Soomro, A. (2021). Online learning during the COVID-19 pandemic: Applying the self-determination theory in the 'new normal.' *Revista de Psicodidactica*, *26*(2), 169–178. https://doi.org/10.1016/j.psicod.2020.12.004

Siemens, G., & Long, P. (2014). Penetrare la nebbia: tecniche di analisi per l'apprendimento. *Italian Journal of Educational Technology*, *22*(3), 132–137. https://ijet.itd.cnr.it/article/view/195

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *Proceedings of the 8th International Conference on Educational Data Mining*, 392–395. http://www.educationaldatamining.org/EDM2015/proceedings/short392-395.pdf

Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. *Proceedings of Workshop on Artificial Intelligence in CSCL*, 17–23.

Taya, E. (2021). Modelo de minería y datos para evaluar el efecto del uso del aula virtual sobre el rendimiento académico de los estudiantes de la Facultad de Ingeniería de la Universidad Nacional Jorge Basadre Grohmann de Tacna, en tiempos de pandemia, 2020. *Repositorio Institucional de La Universidad Nacional Jorge Basadre Grohmann*.

World Bank. (2020). The COVID-19 Crisis Response: Supporting tertiary education for continuity , adaptation , and innovation. *World Bank*, *April*, 1–9. https://bit.ly/3PYp2x4