Lecture Notes in Networks and Systems 619

Miguel Botto-Tobar · Omar S. Gómez ·
Raul Rosero Miranda ·
Angela Díaz Cadena ·
Washington Luna-Encalada *Editors*

Trends in Artificial Intelligence and Computer Engineering

Proceedings of ICAETT 2022

Springer

Editors
Miguel Botto-Tobar
Eindhoven University of Technology
Eindhoven, The Netherlands

Omar S. Gómez
Escuela Superior Politécnica de Chimborazo
Riobamba, Ecuador

Raul Rosero Miranda
Escuela Superior Politécnica de Chimborazo
Riobamba, Ecuador

Angela Díaz Cadena
Universitat de Valencia
Valencia, Valencia, Spain

Washington Luna-Encalada
Escuela Superior Politécnica del Chimborazo
Riobamba, Ecuador

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-25941-8 ISBN 978-3-031-25942-5 (eBook)
https://doi.org/10.1007/978-3-031-25942-5

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Design of a Predictive Model to Evaluate Academic Risk Using Data Mining

Shirley Alarcón-Loza(✉), Diana Calderón-Onofre, Karen Mite-Baidal, and Mishel Macías-Plúas

Instituto Superior Tecnológico ARGOS, Daule, Ecuador
galarcon@tecnologicoargos.edu.ec

Abstract. The impact of academic risk in college can be anticipated, through data analysis, to minimize its impact on the educational community. This article seeks to establish a predictive model that evaluates the academic risk of students of the online modality of a technological institute, from the perspective of performance, for its timely detection and early actions. The database included the registration of demographic data and grades of various subjects taken in the year 2021, of a sample of 1023 students for the 2020 academic period. For this research, the factors Attendance and General Average were considered, to evaluate the performance that affects academic risk. The Cross Industry Standard Process for Data Mining methodology and the software Waikato Environment for Knowledge Analysis were used for evaluation algorithms and search methods to determine suitable predictive attributes for each factor, with attendance records and general averages being the most significant. The results showed that, for the Attendance variable, the best classification algorithm was Random Tree, whose precision value was 99.70% and for the area, under the curve (ROC Area) it was 0.992. Regarding the general average variable, the best classification algorithm was J48, with values of 98.50% and 0.937, respectively. It is suggested to develop research related to data mining that promotes improving the academic quality and services in the study modality.

Keywords: Data mining · Predictive model · Classification algorithm · Higher education · Academic risk

1 Introduction

Technology has brought with it a large number of significant changes that result in the generation of large volumes of data, which are stored without treatment and then lost without having applied actions. The current trend is to convert data into useful information that facilitates timely and effective decision-making for the common good of societies.

In this context, data mining emerges as an analysis tool that is constituted by a set of techniques that works the data for different purposes such as evidencing situations that, at first glance, are not detected, such as patterns and models; that make it possible

M. Botto-Tobar et al. (Eds.): ICAETT 2022, LNNS 619, pp. 221–235, 2023.
https://doi.org/10.1007/978-3-031-25942-5_18

to understand behaviors and predict future actions [8, 37]. Within the field of higher education, information related to students, teachers, academic management [3, 33], and other data resulting from the execution of additional processes such as extracurricular activities, qualifications, and others are stored.

The objective of the article is to establish a predictive model that evaluates the academic risk of students of the online modality of a Technological Institute in Guayaquil, Ecuador. For this purpose, the academic data of the participating sample was organized and clarified, the CRISP-DM methodology was selected, and the Waikato Environment for Knowledge Analysis (WEKA) program was aplied. The decision trees were generated using classification algorithms and the results obtained were interpreted.

It should be noted that academic risk [7, 10, 15, 29] is a state of the student that occurs at any time of the student's trajectory, placing it in any of the conditions described in Table 1. For that reason, data mining applied to higher education has significant advantages, in the prediction of possible unfavorable scenarios. On the other hand, in online higher education, there are many more possibilities to address such as evaluating the management and interactivity of the teacher [1] in the permanent search to improve educational quality.

Table 1. Conditions to be at academic risk

Conditions	Works
Academic backwardness	[11, 12, 42]
Poor academic performance	[9, 18–20, 22, 23, 40]
Low academic achievement	[14, 41]
Academic failure	[7, 30]

2 Materials y Methods

In the framework of the development of this article, the information was the main key from which the results emerged that, by themselves, did not constitute clear evidence of any situation associated with academic risk for the group of participating students. In this sense, the source of the information, the treatment given to the data, the applied data mining methodology, the processing in the corresponding software, and the interpretation of the results obtained are explained.

2.1 Database and Resources

The research was based on a higher education institution that offers the modality of online studies in technological careers. To start in this new environment, they entered into an inter-institutional agreement with the local government to grant scholarships to a significant number of its citizens. The characteristics of those awarded were: having a high school degree belonging to priority groups or in a situation of vulnerability, residents of the city, of low socioeconomic level, whose graduation qualification was high, who is in the age range from 16 to 45 years old and who does not have a registered university degree.

Table 2 shows the research population that corresponds to the 2400 students of the 2020 academic period, who have the aforementioned characteristics. This group was registered in the 17 technological careers offered by the institute. As part of the general data of the careers, the Broad field of knowledge was located, which constitutes the structure of the codification of professional titles and academic degrees. From this aspect, the sample was selected using the non-probabilistic technique for convenience, considering the broad Administration field, which registers the largest number of students among all the careers that comprise it.

Table 2. Population distribution and sample used

Population	Sample		
	Broad Field	Technological careers	N° of students
2,400 students from the 2020 academic period of the online study modality	Management	Human Talent Management	153
		Foreign commerce	152
		Sales	103
		Accounting	181
		Marketing	177
		Management	181
		Tourism Operations Management	76

For this research study, the database of 1023 students was considered, according to the sample described, with the characteristics reflected in Table 3. The demographic aspects resulted from the initial registration made by the students to apply for the scholarship; the selected subjects correspond to those that share all the careers in bimesters 2, 3, 4, 5, and 6, which were carried out during the year 2021; the grade records include the general activities that are recorded in the corresponding software and the attendance record includes the recurring states.

Table 3. Characteristics of the students of the 2020 academic period, online modality

Characteristics	Description	Characteristics	Description
Demographic Aspects	Name	Transcripts for Bimonthly 2–6	Forum
	Career		First Project Delivery
	Gender		Recovery 1
	Marital status		Participation
	Age		Second Project Delivery
	Disability		Recovery 2
	Sector		Final exam
	Parish		Supplementary Exam
	Graduation grade		Final average
	Quintile		
Subjects considered for Bimester 2–6	Management	Bimesters 2–6 Attendance Record	Present
	Calculus 1		Late
	Financial Accounting Statistics		Absent
	Business Ethics		

As a technological resource it was used the data mining open source software WEKA [28], based on the JAVA programming language. It supports different types of data format which includes ARFF, CSV, and LibSVM. It consists of many algorithms that can be applied to different data sets for promote future analysis. In it, the processing of the information was carried out to generate the results corresponding to data mining and due evaluation.

2.2 Methodology

The CRISP-DM methodology is a standardized methodology for the data mining process [21] and knowledge discovery in databases [39]. This methodology, according to [38], consists of six phases, and its application for this article is described below:

Business or Problem Understanding Stage

The academic data that is generated in each academic period in the online modality does not receive any treatment that allows showing situations of risk in this group of students. By identifying this scenario, the present investigation was proposed to determine a predictive classification model that allows recognizing if a student could be placed at academic risk, from the perspective of performance, to detect it in time and solve the situation opportunely. For this, the data mining classification technique was used and, the ideal decision algorithms were validated, based on the data set collected from the participating sample.

Data Understanding Stage

In this phase, the data was collected and filtered based on the characteristics of the students, shown in Table 2, organizing 1023 records from 75 different fields. These fields were standardized and stored in Microsoft Excel for their corresponding preparation.

Data Preparation Stage

Academic risk variables created in the database

To determine the academic risk, an edge called academic performance has been distinguished, and here, the variables Attendance [27, 35] and General Average have been selected, to evaluate the condition of the students regarding the risk. With this foundation, it is necessary to create a dependent variable, also known as a class variable, which is of dichotomous nominal type (YES/NO). From it, the predictive classification model is executed. The dependent variables for this study that directly affect academic performance [5, 6, 13, 26, 34] and, therefore, affect academic risk, are the Attendance class variable (RISK_ASSIST) and the General Average class variable (RISK_AVG).

Variables considered for the training data set in WEKA

To carry out the predictive model, each academic performance factor has independent variables and a dependent variable. Table 4 details them, which were entered into the WEKA program to consider the academic performance associated with the Attendance factor. Under the institutional policy, students' attendance must be a minimum of 80% of the total subject, otherwise, they fail. Students receive 8 class sessions for each subject in an academic period and it is considered that having 3 absences within the academic period, they are at risk. Table 3 also reflects the dependent variable RISK_ASSIST and its respective formula in the database.

Table 4. Description of the variables considered for the class attendance factor

Variables	Meaning	Type
ADM_PRESENT CALC_ PRESENT ACCOUNT_ PRESENT STA_ PRESENT ETHICS_ PRESENT	The number of times the student attends the Administration (ADM), Calculus (CALC), Accounting (ACCOUNT), Statistics (STA), and Ethics (ETHICS) class.	Numeric
ADM_LATE CALC_LATE ACCOUNT _LATE STA_LATE ETHICS_LATE	The number of times the student registered late in the Administration (ADM), Calculus (CALC), Accounting (ACCOUNT), Statistics (STA), and Ethics (ETHICS) class.	Numeric
ADM_ABSENT CALC_ABSENT ACCOUNT_ABSENT STA_ABSENT ETHICS_ABSENT	The number of times the student does not attend the Administration (ADM), Calculus (CALC), Accounting (ACCOUNT), Statistics (STA), and Ethics (ETHICS) class.	Numeric
RISK_ASSIST (Class variable)	Academic risk =YES.SET (ADM_PRESENT<=5; "YES"; CALC_PRESENT <=5; "YES"; ACCOUNT_PRESENT <=5; "YES"; STA_PRESENT <=5; "YES"; ETHICS_PRESENT <=5; "YES "; ADM_PRESENT >=6; "NO"; CALC_PRESENT >=6; "NO"; ACCOUNT_PRESENT >=6; "NO"; STA_PRESENT >=6; "NO"; ETHICS_PRESENT >=6; "NO")	Nominal dichotomous (YES/NO)

Table 5 details the independent and dependent variables that were entered into the WEKA program to consider the academic performance associated with the General Average factor. Within the framework of the scholarship, the student must maintain 80 points in the general average, including all the subjects that she is taking. You are at academic risk if your grade is below 80. Table 5 also reflects the dependent variable RISK_AVG and its respective formula in the database.

Table 5. Description of the variables considered for the general average factor

Variables	Meaning	Type
Career	Educational program of the broad field of Administration	Nominal
Age	Age at admission (2020)	Numeric
Gender	Student's gender	Nominal (Male / Female)
Disability	A disability that the student has	Nominal (YES / NO)
Quintile	Distribution of students according to monthly income	Nominal (Quintile1, Quintile2, Quintile3, Quintile4, Quintile5)
ADM_AVG	Administration final average	Numeric
CALC_AVG	Calculus final average	Numeric
ACCOUNT_AVG	Accounting final average	Numeric
STA_AVG	Statistics final average	Numeric
ETHICS_AVG	Ethics final average	Numeric
RISK_AVG (Class variable)	Academic risk =YES((AVERAGE (ADM_AVG; CALC_AVG; ACCOUNT_AVG; STA_AVG; ETHICS_AVG))>=80; "NO"; "YES")	Nominal dichotomy (YES / NO)

Modeling Stage

Aspects considered for the test data set in WEKA

Two experiments were considered about the dependent variables, RISK_ASSIST and RISK_AVG; considering before them, a selection of attributes, to then apply the classification algorithm, this will reduce the work of processing irrelevant attributes.

In the process of selecting attributes within the WEKA application, evaluator algorithms and search methods were applied. One of the evaluative algorithms used was the CfsSubsetEval [4, 17] with the BestFirst search method [16] to select independent variables that most closely affect the dependent variable. Another evaluator algorithm that was used was the CorrelationAttributeEval [2] with the Ranker search method [40] which allowed evaluating the value of a variable by determining the correlation with the class variable.

Table 6 reflects the results after applying the evaluator algorithms and the search methods concerning the variable RISK_ASSIST.

Table 6. Results of the application of the evaluator algorithm for the selection of attributes of the RISK_ASSIST variable

Attribute	CfsSubsetEval BestFirst	CorrelationAttributeEval/Ranker	
		Average Merit	Average Rank
ADM_PRESENT	0%	0.373 ± 0.015	1 ± 1.04
ADM_LATE	80%	0.073 ± 0.014	3 ± 0.46
ADM_ABSENT	100%	0.375 ± 0.014	4.9 ± 0.83
CALC_ PRESENT	10%	0.237 ± 0.012	1 ± 0.3
CALC_LATE	0%	0.137 ± 0.013	±0
CALC_ABSENT	50%	0.236 ± 0.011	9 ± 0.3
ACCOUNT_ PRESENT	80%	0.366 ± 0.008	4 ± 0.92
ACCOUNT_LATE	10%	0.137 ± 0.013	±0
ACCOUNT_ABSENT	0%	0.353 ± 0.008	6 ± 0.8
STA_ PRESENT	100%	0.805 ± 0.005	±0
STA_LATE	0%	0.02 ± 0.01	±0
STA_ABSENT	0%	0.582 ± 0.011	±0
ETHICS_ PRESENT	100%	0.689 ± 0.01	±0
ETHICS_LATE	0%	0.197 ± 0.018	±0
ETHICS_ABSENT	0%	0.305 ± 0.008	±0

Table 7 reflects the results after applying the evaluator algorithms and the search methods concerning the variable RISK_AVG.

Table 7. Results of the application of the evaluator algorithm for the selection of attributes of the variable RISK_AVG

Attribute	CfsSubsetEval BestFirst	CorrelationAttributeEval/Ranker	
		Average Merit	Average Merit
Career	0%	0.053 ± 0.005	±0
Age	0%	0.017 ± 0.001	8.4 ± 0.49
Gender	0%	0.095 ± 0.01	±0
Disability	0%	0.013 ± 0.007	9.3 ± 0.9
Quintile	0%	0.013 ± 0.003	9.3 ± 0.64
ADM_AVG	0%	0.446 ± 0.015	±0
CALC_AVG	100%	0.524 ± 0.02	±0

(*continued*)

Table 7. (*continued*)

Attribute	CfsSubsetEval BestFirst	CorrelationAttributeEval/Ranker	
		Average Merit	Average Merit
ACCOUNT_AVG	100%	0.582 ± 0.014	±0
STA_AVG	80%	0.865 ± 0.008	1.9 ± 0.3
ETHICS_AVG	100%	0.881 ± 0.007	1.1 ± 0.3

The classification algorithms [32] used to identify the causes of academic risk related to performance were the J48 and Random Tree decision trees. The J48 algorithm [6] shows the percentage of well-classified instances and those that are not correctly classified in the confusion matrix, and the precision detailed by the class variable. This algorithm allows to discover of specific relationships between instances and attributes, using the best attributes of the generated tree. The Random Tree classification algorithm [18, 19] builds a tree that considers a random number of attributes and instances for each node.

For the attribute selection process and application of classification algorithms, Evaluation and Validation mode of models called Cross-Validation Folds [4] was applied, since it provides an average precision per variable class based on K iterations that executes with the data.

Evaluation Stage

According to Table 8, two measures have been used for the evaluation of the quality of the classification, which is Area Under the Curve (ROC-AUC) [24] and Accuracy [25]. According to these measurements, the J48 and Random Tree algorithms provide a point in the space under the curve ROC, as they are binary classifiers. The ROC analysis is the ratio of true positives, and information retrieval, versus the ratio or proportion of false positives. When is near to unity, the classifier behavior approaches the perfect classifier. Regarding precision, the most accurate percentage of all the positives that have classified the test is obtained.

Table 8. Evaluation of the quality of the applied algorithms

Algorithm	RISK_ASSIST		RISK_AVG	
	Accuracy	ROC Area	Accuracy	ROC Area
J48	98,50%	0,952	98,50%	0,937
Random Tree	99,70%	0,992	97,30%	0,923

Implementation Stage

The predictive model [22, 36] was generated by applying classification algorithms in the WEKA program. Several tests were carried out, in such a way that two significant algorithms were chosen. According to Table 8, the Random Tree algorithm was applied to assess the academic risk to Attendance. Eight knowledge rules were generated with the SI value (see Fig. 1), indicating that there is a risk of failing two or more subjects.

The interpretation of the generated knowledge rules that declare that there is academic risk, are explained below, using literals: A) If the value of ACCOUNT_ABSENT < 1.5 & ETHICS_ABSENT >= 2, then if there is risk academic and 15 students were located; B) If the value of ACCOUNT_ABSENT < 1.5 & ETHICS_ABSENT < 2 & ETHICS_PRESENT < 6, then there is academic risk and 42 students were located; C) If the value of ACCOUNT_ABSENT < 1.5 & ETHICS_ABSENT < 2 & ETHICS_PRESENT >= 6 & STA_PRESENT < 6, then if there is academic risk and 52 were located students; D) If the value of ACCOUNT_ABSENT < 1.5 & ETHICS_ABSENT < 2 & ETHICS_PRESENT >= 6 & STA_PRESENT >= 6 & CALC_ABSENT >= 2.5, then if there is academic risk and 3 students were located; E) If the value of ACCOUNT_ABSENT < 1.5 & ETHICS_ABSENT < 2 & ETHICS_PRESENT >= 6 & STA_PRESENT >= 6 & CALC_ABSENT <

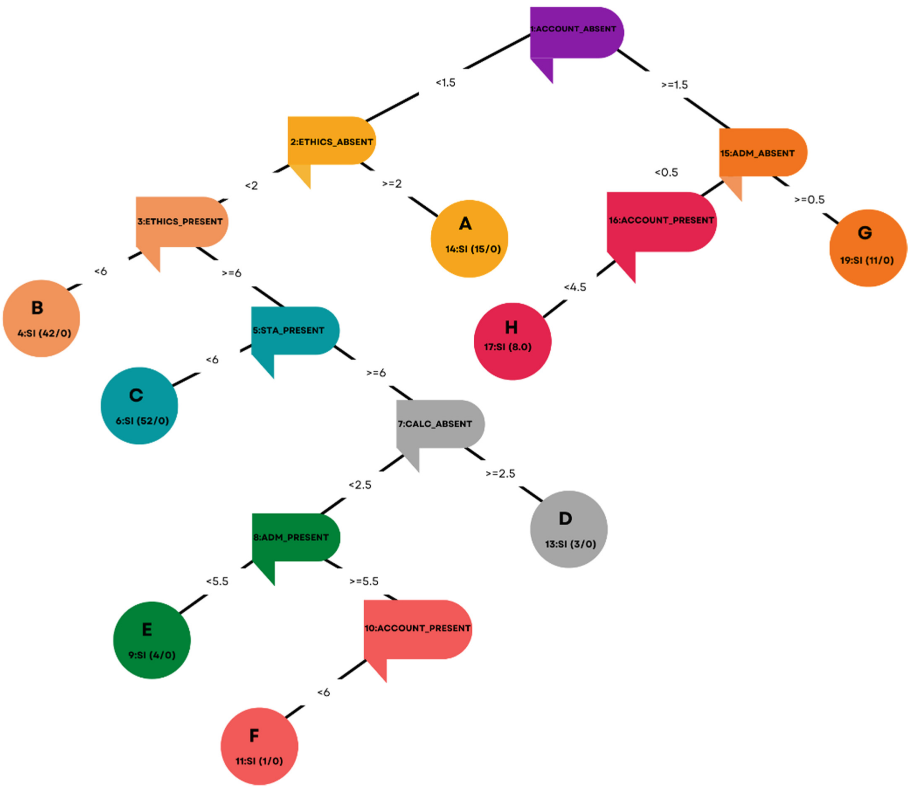

Fig. 1. Knowledge rules derived from the Random Tree algorithm for RISK_ASSIST

2.5 & ADM_PRESENT < 5.5, then if there is academic risk and 4 students were located; F) If the value of ACCOUNT_ABSENT < 1.5 & ETHICS_ABSENT < 2 & ETHICS_PRESENT >= 6 & STA_PRESENT >= 6 & CALC_ABSENT < 2.5 & ADM_PRESENT >= 5.5 & ACCOUNT_PRESENT < 6, then if there is academic risk and 1 student was located; G) If the value of ACCOUNT_ABSENT >= 1.5 & ADM_PRESENT >= 0.5, then if there is academic risk and 11 students were located and, H) If the value of ACCOUNT_ABSENT >= 1.5 & ADM_ABSENT < 0.5 & ACCOUNT_PRESENT < 4.5, then if there is academic risk and 8 students were located.

The J48 algorithm was applied to evaluate the academic risk to the general average of the subjects. There are a total of three knowledge rules with the SI value (see Fig. 2), indicating that there is a risk of failing one or more subjects. The interpretation of the generated knowledge rules that declare that there is an academic risk is explained below, using literal: A) If the value of ETHICS_AVG <= 50, then if there is academic risk and 67 were located students; B) If the value of ETHICS_AVG > 50 & ADM_AVG <= 71 & CALC_AVG <= 70, then if there is academic risk and 6 students were located and, C) If the ETHICS_AVG > 50 & ADM_AVG <=71 & CALC_AVG > 70 & STA_AVG <= 75, then if there is academic risk and 3 students were located.

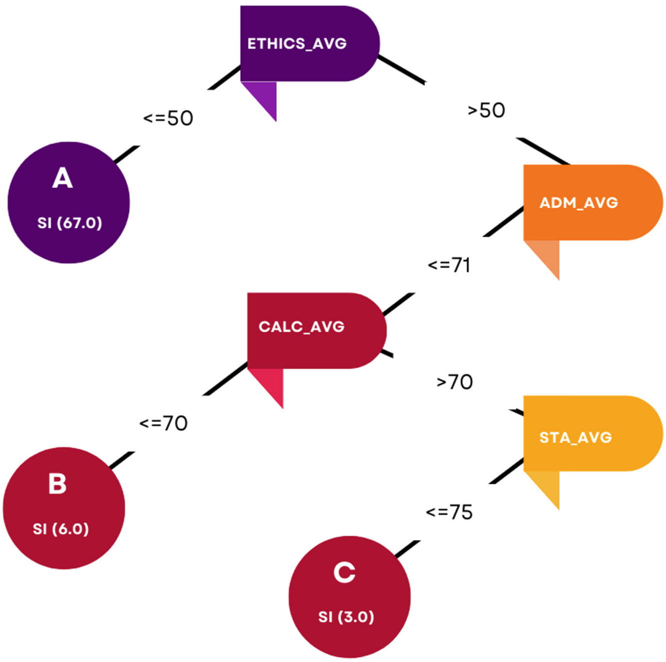

Fig. 2. Knowledge rules derived from the J48 decision tree for RISK_AVG

3 Results

For both variables, evaluator algorithms and search methods were executed to identify the most influential predictive attributes to generate higher precision decision trees. In the case of the Attendance variable (RISK_ASSIST), the attributes referring to the Attend record had a better rating, except for the Administration subject. In the case of the General Average variable (RISK_AVG), the best-valued attributes were the final averages of each subject.

For the Assistance variable (RISK_ASSIST), the Random Tree algorithm was applied, obtaining an excellent percentage in its precision and quality margin, according to the evaluation shown in Table 8. The quality of this algorithm is also reflected in the data set. Evaluated in the research carried out by [19, 31], where their results were the best. When generating it, it broke down ten rules that consider the existence or not of academic risk, evaluating the presence or absence of students in the selected subjects.

The results show that even though students record absences less than 2 or 1.5, and their presence in some subjects is greater than 6 or 5.5, they are at academic risk. On the other hand, if their absence in at least one subject is greater than 2, they would also be at academic risk.

For the General Average variable (RISK_AVG) the J48 algorithm was applied, whose precision percentage and quality margin were higher, according to the evaluation shown in Table 8; the quality and realism offered by this algorithm have been evaluated and demonstrated through studies carried out by [6, 18, 20]. In this case, five easy-to-interpret rules were obtained that consider a minimum average of 80 to avoid being placed in low performance, which leads to academic risk. The results consider that if the average is below 50, 70, 71, or 75, in two or more subjects, it is an academic risk. In another position, the fact that some subject obtains averages higher than 70, does not guarantee that it is not placed at academic risk.

4 Conclusions

Broadly speaking, data mining is a versatile tool that can address various aspects of the operation of higher education institutions, where large volumes of information are generated or integrated. Professional training activities have a high social impact because they deal with people and their future development, which they contribute to society. With this base, making better decisions, and using technological models that allow predicting future scenarios, is a success that contributes to the improvement of the service of this type of organization.

From the research, it is recognized the importance of ordering and clarifying the data before processing it in the selected data mining software. In the same order, verifying the precision and quality of the algorithms that are used makes a difference between the rules of behavior that are accepted, because, from them, possible changes and actions will be made in the aspects that have been evaluated.

In this case, it was shown that attendance records have a greater influence on academic performance and risk, so students should maintain a record higher than 5.5 to avoid it. Regarding the averages, all must stay above 71 to avoid academic risk; however, at least one subject with an average greater than 50 is accepted. In relation, using a predictive model that focuses on academic performance allows observing the students who continue in the various technological careers, to alert them promptly and prevent them from being at academic risk, which suggests other associated problems.

On this basis, it would be interesting to identify and evaluate the data recorded by the online study modality of the technological institute, within the framework of the characteristics of its student community, to generate new research where it is integrated into the whole, the behavior is observed of other subjects, the interactivity of the teacher and the student is considered in response to the asynchronous and synchronous activities that are offered, among other aspects.

With a technical approach, attribute evaluator algorithms and search methods can be used to find features closely related to the class variable that positively affect the success of a search. Consequently, it is suggested to carry out predictive models with data mining using unsupervised algorithms.

References

1. Ahmed, A.M., et al.: Using data mining to predict instructor performance. Procedia Comput. Sci. **102**, 137–142 (2016). https://doi.org/10.1016/j.procs.2016.09.380
2. Ahmed, R., Ahmed, H.: Determining the most effective filter feature selection algorithm for machine learning algorithm using WEKA version 3.8.1 and NSL-KDD dataset (2017)
3. Al-Twijri, M.I., Noaman, A.Y.: A new data mining model adopted for higher institutions. Procedia Comput. Sci. **65**, 836–844 (2015). https://doi.org/10.1016/j.procs.2015.09.037
4. Alabdulwahab, S.: Feature selection methods simultaneously improve the detection accuracy and model building time of machine learning classifiers, vol. 12, pp. 03–20 (2020)
5. Alexander, V., Hicks, R.E.: Does class attendance predict academic performance in first year psychology tutorials? Int. J. Psychol. Stud. **8**(1), 28 (2015). https://doi.org/10.5539/ijps.v8n1p28
6. Anoopkumar, M., Rahman, A.M.J.Z.: Model of tuned J48 classification and analysis of performance prediction in educational data mining. Int. J. Appl. Eng. Res. **13**(20), 14717–14727 (2018)
7. Baker, R.S., et al.: Analyzing early at-risk factors in higher education e-learning courses. In: Proceedings of the 8th International Conference on Educational Data Mining, pp. 150–155 (2015)
8. Bedregal-alpaca, N., et al.: Classification models for determining types of academic risk and predicting dropout in university students. Int. J. Adv. Comput. Sci. Appl. **11** (2020). https://doi.org/10.14569/IJACSA.2020.0110133
9. Bedregal-Alpaca, N., et al.: Análisis del rendimiento académico de los estudiantes de Ingeniería de Sistemas, posibilidades de deserción y propuestas para su retención. Ingeniare. Rev. Chil. Ing. **28**(4), 668–683 (2020). https://doi.org/10.4067/s0718-33052020000400668
10. Berens, J., et al.: Early detection of students at risk - predicting student dropouts using administrative student data from German universities and machine learning methods. J. Educ. Data Min. **11**(3), 1–41 (2019). https://doi.org/10.5281/zenodo.3594771
11. Bound, J., et al.: Why have college completion rates declined? An analysis of changing student preparation and collegiate resources. Am. Econ. J. Appl. Econ. **2**(3), 129–157 (2010). https://doi.org/10.1257/app.2.3.129

12. Cabrales, R.A., et al.: Dropout, student lag and successful completion in 40 cohorts of the Medicine Program of the Universidad Tecnológica de Pereira. Colombia; [Deserción, rezago estudiantil y egreso exitoso en 40 cohortes del Programa de Medicina de la Universidad Tecnológica. Iatreia. **35**(3), 239–248 (2022). https://doi.org/10.17533/udea.iatreia.133
13. Dey, I.: Class attendance and academic performance: a subgroup analysis. Int. Rev. Econ. Educ. **28**, 29–40 (2018). https://doi.org/10.1016/j.iree.2018.03.003
14. Duke, N.N.: Adolescent adversity, school attendance and academic achievement: school connection and the potential for mitigating risk. J. Sch. Health **90**(8), 618–629 (2020). https://doi.org/10.1111/josh.12910
15. El-zeweidy, A.P.M.: Academic educational data mining predictive model for early detection of students at academic risk. J. ACS **9**, 21–42 (2018)
16. Fong, S., et al.: Swarm search methods in Weka for data mining. In: Proceedings of the 2018 10th International Conference on Machine Learning and Computing, pp. 122–127 (2018)
17. Gnanambal, S.: Classification algorithms with attribute selection: an evaluation study using WEKA. Int. J. Adv. Netw. Appl. **3644**(6), 3640–3644 (2018)
18. Hamoud, A.K. et al.: Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. Int. J. Interact. Multimed. Artif. Intell. 5, 2, 26 (2018). https://doi.org/10.9781/ijimai.2018.02.004
19. Hasan, R., et al.: Student academic performance prediction by using decision tree algorithm. In: 2018 4th International Conference on Computer and Information Sciences Revolutionizing Digital Landscapes Sustainability Smart Society (ICCOINS) ICCOINS 2018 – Proceedings, pp. 1–5 (2018). https://doi.org/10.1109/ICCOINS.2018.8510600
20. Imran, M., et al.: Student academic performance prediction using supervised learning techniques. Int. J. Emerg. Technol. Learn. **14**(14), 92–104 (2019). https://doi.org/10.3991/ijet.v14i14.10310
21. International, A., et al.: Data mining: process, tools, techniques, and applications. J. Anal. Comput. 1–8 (2020). http://www.ijaconline.com/wp-content/uploads/2020/02/G.Abirami-Data-Mining-processtoolstechniques-and-applications.-1.pdf
22. Ismael, M.N.: Students performance prediction by using data mining algorithm techniques. Eurasian J. Eng. Technol. **6**, 11–25 (2022)
23. Jalota, C., Agrawal, R.: Analysis of educational data mining using classification. In: Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon) 2019, pp. 243–247 (2019). https://doi.org/10.1109/COMITCon.2019.8862214
24. Janssens, A.C.J.W., Martens, F.K.: Reflection on modern methods: revisiting the area under the ROC Curve. Int. J. Epidemiol. **49**(4), 1397–1403 (2020). https://doi.org/10.1093/ije/dyz274
25. Juba, B., Le, H.S.: Precision-recall versus accuracy and the role of large data sets. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI 2019), 31st Innovative Applications of Artificial Intelligence Conference, 9th AAAI Symposium on Educational Advances in Artificial Intelligence EAAI 2019, pp. 4039–4048 (2019). https://doi.org/10.1609/aaai.v33i01.33014039
26. Kassarnig, V., et al.: Class attendance, peer similarity, and academic performance in a large field study. PLoS ONE **12**(11), 1–15 (2017). https://doi.org/10.1371/journal.pone.0187078
27. Kauffman, C.A., et al.: Relationship between classroom attendance and examination performance in a second-year medical pathophysiology class. Adv. Physiol. Educ. **42**(4), 593–598 (2018). https://doi.org/10.1152/advan.00123.2018
28. Kotak, P., et al.: Enhancing the data mining tool WEKA. In: 2020 5th International Conference on Computing, Communication and Security (ICCCS), pp. 4–9 (2020)
29. Kouser, F., et al.: Early detection of failure risks from students' data. In: International Conference on Emerging Trends in Smart Technologies - ICETST 2020, pp. 1–6 (2020)

30. Márquez-Vera, C., et al.: Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Appl. Intell. **38**(3), 315–330 (2013). https://doi.org/10.1007/s10489-012-0374-8
31. Mhetre, V., Nagar, M.: Classification-based data mining algorithms to predict slow, average, and fast learners in an educational system using WEKA. In: Proceedings of the International Conference on Computing Methodologies and Communication, ICCMC 2017, pp. 475–479 (2018). https://doi.org/10.1109/ICCMC.2017.8282735
32. Mirza, H.B.: Classifier tools : a comparative study. In: 2018 Second International Conference Intelligent Computing and Control Systems ICICCS, pp. 1543–1547 (2018)
33. Mohamed, M.H., Waguih, H.M.: A proposed academic advisor model based on data mining classification techniques. Int. J. Adv. Comput. Res. **8**(36), 129–136 (2018). https://doi.org/10.19101/IJACR.2018.836003
34. Moore, R.: Attendance and performance. J. Coll. Sci. Teach. **6**, 367–371 (2003)
35. Nieuwoudt, J.E.: Investigating synchronous and asynchronous class attendance as predictors of academic success in online education. Australas. J. Educ. Technol. **36**(3), 15–25 (2020). https://doi.org/10.14742/AJET.5137
36. Preet, K., et al.: Exploring data mining tool - WEKA and using WEKA to build and evaluate predictive models. Adv. Appl. Math. Sci. **19**(6), 451–469 (2020)
37. Ramaswami, G., et al.: Using educational data mining techniques to increase the prediction accuracy of student academic performance. Inf. Learn. Sci. **120**(7–8), 451–467 (2019). https://doi.org/10.1108/ILS-03-2019-0017
38. Schröer, C., et al.: A systematic literature review on applying CRISP-DM process model. Procedia Comput. Sci. **181**(2019), 526–534 (2021). https://doi.org/10.1016/j.procs.2021.01.199
39. Vickery, B.: Knowledge discovery from databases: an introductory review. J. Doc. **53**(2), 107–122 (1997). https://doi.org/10.1108/EUM0000000007195
40. Walia, N., et al.: student's academic performance prediction in academic using data mining techniques. SSRN Electron. J. 1–5 (2020). https://doi.org/10.2139/ssrn.3565874
41. Yağcı, M.: Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. **9**(1), 1–19 (2022). https://doi.org/10.1186/s40561-022-00192-z
42. Zulfikar, W.B., et al.: Comparison of Naive Bayes Classifier and C4.5 in Predicting Student Study Period. Presented at the (2020)