EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Development of a Classification Model for Predicting Student Payment Behavior Using Artificial Intelligence and Data Science Techniques

Henry Villarreal-Torres [1], Julio Ángeles-Morales [1], William Marín-Rodriguez [2,*], Daniel Andrade-Girón[2], Edgardo Carreño-Cisneros [2], Jenny Cano-Mejía [1], Carmen Mejía-Murillo [1], Mariby C. Boscán-Carroz [3], Gumercindo Flores-Reyes [1], Oscar Cruz-Cruz [1],

[1] Universidad San Pedro. Chimbote, Perú.
[2] Universidad Nacional José Faustino Sánchez Carrión. Huacho, Perú.
[3] Universidad del Zulia. Maracaibo, Venezuela.

Abstract

Artificial intelligence today has become a valuable tool for decision-making, where universities have to adapt and optimize their processes, improving the quality of their services. In this context, the economic income from collections is vital for sustainability. There are several problems that can contribute to student delinquency, such as economic, financial, academic, family, and personal. For this reason, the study aimed to develop a classification model to predict the payment behavior of enrolled students. The methodology is a proactive, technological study of incremental innovation with a synchronous temporal scope. The study population consisted of 8,495 undergraduate students enrolled in the 2022 - II academic semester, containing information on academic performance, financial situation, and personal factors. The result is a classification model using the H2O.ai platform, discretization algorithms, data balancing, and the R language. Data science algorithms obtained the base from the institution's computer system. The data sets for training and testing correspond to 70% and 30%, obtaining the GBM Grid model whose performance metrics are AUC of 0.905, AUCPR of 0.926, and logLoss equivalent to 0.311; that is, the model efficiently complies with the classification of student debtors to provide them with early intervention service and help them complete their studies.

Keywords: Automated Machine Learning, Higher Education, Data Mining, Delinquency.

Received on 12 December 2022, accepted on 9 June 2023, published on 26 June 2023

doi: 10.4108/eetsis.3489 _____

*Corresponding author. Email: wmarin@unjfsc.edu.pe

1. Introduction

Financing in Education

Total public spending on education, according to OECD (2022), the average is 10.6%, where most of it is destined for the primary and secondary levels; in the case of the tertiary level, the contribution of private sources gives it; Likewise, there is a difference between 2015 and 2019 where the proportion of public spending allocated to education had a slight decrease among the countries belonging to the OECD of around 1%, this figure grew due to the global crisis caused by the COVID-19 pandemic. 19 that prompted governments to spend more to reactivate their economies. Regarding the average private spending on educational institutions, it remained stable. On the other hand, regardless of the educational level, the remuneration of personnel and other current expenses represent an average of 90% of the expenditure in educational institutions. In the

case of tertiary education, it corresponds to 67%, including expenses for the support of research and development among the countries belonging to the OECD, where they had an expense in tertiary education of 1.5% of the gross domestic product on average; that is, if an increase of US$1.0000 in GDP per capita is associated on average with an increase of US$200 in student spending; Such is the case that in 2019 an average of US$17,560 per student was spent at the tertiary level, of which the majority were allocated to essential services (salaries, buildings, didactic materials, and administration); followed by research and development activities; and finally, in auxiliary activities (meals and transportation).

Considering the III Biennial Report on the University Reality in Peru of the National Superintendence of Higher University Education (SUNEDU, 2021), the financing refers to resources and investments aimed at the operation of universities and graduate schools; in addition to continuous improvement of the educational quality and well-being of its university community. In this regard, finances are mainly from the state budget allocation, mining canon and own collection, for public universities; and in the case of private universities, it is given by income, pensions and the investment of educational promoters; During the period 2010 to 2020, the public budget for higher education started at S/. 1,489 million, increasing periodically until 2015, reaching S/. 3,000 million, for 2016, it had a decrease of S/. 2,776 million; then it gradually increased until reaching an amount of S/. 3,570 million; another decrease in 2020, reaching S/. 3,263 million. In the case of private universities, from 2014 to 2018, they presented an annual income growth rate. 2014 there was S /. 4,590.8 million, maintaining an increase until 2017 with income corresponding to S/. 6,704.8 million, and in 2018 reached S/. 6,918.2 million. From 2014 to 2020, scholarships and educational loans presented a favorable evolution for students; Regarding the scholarships, the initial amount was S/. 406.5 million ending in 2020 with S/. 759 million, presenting a decrease in the years 2018 and 2019 with amounts of S/. 599 and S/. 576.8 million, respectively. Regarding educational loans, the amount started with S/. 7.4 million in 2014. By 2020 it had S/. 6.6 million; the maximum and minimum peaks were in 2018 with S/. 12.4 million and S/. 1.9 million in 2019, respectively.

Artificial intelligence

Artificial intelligence has been studied since the 1940s, including machine learning, artificial neural networks, and deep learning; its growth has been possible due to digital transformation, a phenomenon of greater importance; artificial intelligence can be explained as the imitation of human thought and actions or the human brain for the generation of similar responses to investigate real-world problems by programming through machine learning, logic, perception, and reasoning; considering a combination of information technologies and physiological intelligence to process a large amount of data helping in the creation of functional tools for complex decision making; thus giving effective exploitation of its potential due to its maturity and affordability (Wiesmüller, 2023; Chatterjee et. al., 2023; Vähäkainu & Lehto, 2023; Chryssolouris et. al., 2023).

To date, there is a wide catalog of artificial intelligence solutions, with the arrival of OpenAI ChatGPT, Google Bard, Microsoft Bind, and Adobe Firefly, among other solutions, they have impacted a substantial change in people, businesses, and organizations; all these innovations have been given by the computing capacity and the enormous amount of data where artificial intelligence has played a relevant role to be adopted in our daily lives and in many sectors, revolutionizing society; One of the barriers or problems encountered is explainability, generating a new field, explainable artificial intelligence, which provides explanations for the predictions, recommendations, and decisions of machine learning and deep learning systems, providing greater transparency, interpretability, and explainability to its users. algorithms (Angelov et. al., 2022; Andrade-Girón et al., 2023; Ridley, 2022; Minh et. al., 2021; Arrieta et. al., 2019; Miller, 2017). In education, the development and use of information and communication technologies it has given rise to the rise of artificial intelligence, where institutions have adopted and incorporated integrated systems to perform functions similar to that of teachers or instructors, substantially improving efficiency. and efficiency of their teachers, giving a better educational quality to their students (Chen et. al., 2020; Mejías et al., 2022; Subbarayan & Gunaseelan, 2022).

Data mining

Currently, in a digital world, there are huge volumes of data provided by IoT, cybersecurity, mobility, social networks, commerce, health, etc., which must be intelligently analyzed to build intelligent applications where artificial intelligence and, specifically, machine learning have the leading role. The classification of machine learning algorithms comprises supervised, unsupervised, semi-supervised, and reinforcement learning (Silva Coimbra & Rodrigues Dias, 2022; Marinho de Sousa & Shintaku, 2022; Takaki & Dutra, 2022; Sarker, 2021), with many software tools (Bartschat et al., 2019). Garg & Sharma (2013) and Francis et al. (2019) argue that data mining is proliferating and is defined as an interdisciplinary field derived from computer science, considered a process to discover patterns in vast volumes of data by applying one or more algorithms or techniques. solid according to the types of data and the proposed objective. Also, it is considered as the process of extracting significant information from a database, becoming an effective tool through its phases: cleaning, integration, selection, transformation, and data visualization, help the development of applications for different areas of knowledge (Tan et al., 2022; Yağcı, 2022). On the other hand, educational mining can explore and analyze data that comes from multiple sources originated by educational services with the support of data mining; In this sense, higher education institutions, with the use of educational mining techniques and tools, can make better decision-making and explain educational phenomena (Sharma & Sharma, 2018; Romero & Ventura, 2012); Educational mining has the support of multiple disciplines, including cognitive science, computer science, cognitive psychology, education, and statistics, leading to a

greater challenge to improve the quality of educational processes with the implementation of strategies and plans based on the information extracted (Koedinger, 2015; do Carmo & da Silva Lemos, 2022; Sumitha & Vinothkumar, 2016). The importance of data mining lies in: a) the process of collecting large amounts of data to extract information; b) interpretation of the data for its subsequent transformation into information; and c) evaluate the behaviors and ideas of consumers resulting in organizational growth based on data (Bolaño García et al., 2023; Olufemi, 2021).

Machine Learning

Machine learning has become an indispensable tool with an advanced approach to data analysis. It is defined as the ability of systems to learn from the training data of the problem to generate analytical models and solve the respective tasks; On the other hand, deep learning is based on neural networks from the concept of machine learning; It is highlighted that deep learning models are superior to machine learning models in most applications (Janiesch et al., 2021; Mirande & Martínez Debat, 2023; Catrambone & Ledwith, 2023; Zhong, et al., 2021). Artificial intelligence and machine learning are not synonymous but different concepts; The first is a machine's ability to develop cognitive functions: perception, reasoning, learning, interaction with the environment, problem-solving, decision-making, and creativity. while machine learning is a technique used in artificial intelligence to allow machines to learn from data and improve their performance in a specific task (Benito, 2022; Corrêa da Silva, 2022; Kühl eet al., 2022; Silva, 2022).

Considering machine learning models as successful and can solve problems and make decisions, they come to make a combination of complexity (black box) to understand its internal functioning by a person; This is where interpretability plays a vitally important role due to the lack of transparency and thus avoiding negative consequences, producing correct answers for the wrong reasons in high-risk areas (Rudin et al., 2021; Zhou et al., 2021). The three different types of machine learning are a) supervised learning, which refers to a type of learning where labeled data is provided to the model to train it; the model learns to make predictions based on the training data and the corresponding labels; b) unsupervised learning, on the other hand, refers to a type of learning in which no labels are provided to the model. Instead, the model must find patterns and structures in the data on its own; and c) reinforcement learning refers to learning in which the model learns to make decisions in a given environment to maximize a reward. Depending on its decisions, the model receives feedback through rewards or punishments (Kühl et al., 2022).

Data Balancing

For Chawla (2005), the data set is unbalanced or unbalanced if the classification categories are not equally represented, which would influence the training data, erroneous predictions, and poor performance of the resulting models. Thus, the study by Ghanem, Venkatesh & West (2008) describes machine learning methods that have structured data files as their main input under the assumption that the classes of categorical variables are similar in quantity; However, in reality, the data is stored primarily on relational database systems and is unbalanced; that is, a class of data contains a greater proportion compared to the other classes; There are also studies on the development of machine learning with unbalanced data as a research area in order to find efficient methods for solving real problems, which requires a broad vision to understand the nature of learning (Chawla, 2009; Lali et al., 2023a; Lali et al., 2023b). In this sense, the necessary attention has been received to improve the performance of classification models, where various components or factors have been considered, such as distribution, and cost-sensitive learning, among others (Yin et. al., 2020).

In this sense, to obtain relevant data in higher education institutions, there is a proposal for the management model of information and communication technologies for higher education institutions (Villarreal et al., 2021). Furthermore, the analysis performed by Liu et al. (2020) expresses the existence of degradation and low performance in the classification algorithms when the data set is unbalanced with a minority class. Likewise, using inappropriate or incorrect metrics to evaluate the performance of the algorithms can affect the experimental results in a classification model with highly unbalanced data (Hancock, Khoshgoftaar & Johnson, 2023; Martín Ferron, 2022; de Araújo Telmo et al., 2021).

Feature Selection

A characteristic is conceived as an individual measurable property of the observed process. Being immersed in the digital age, data is generated from different information systems, leading to an increase in the dimensions of the data; having more characteristics should result in more discrimination. However, practice indicates that this is not always the case. Some factors affect the success of machine learning, such as the quality of the data set; feature selection is selecting a specific subset of variables from the original set, which can efficiently describe the input data while reducing the effects of irrelevant variables providing good prediction results. That is, identify and eliminate irrelevant and redundant information to reduce dimensionality and algorithms to be faster and more efficient, obtaining optimal or desired performance (Shi, 2022; Zebari et. al., 2020; Venkatesh & Anuradha, 2019; Chandrashekar & Sahin, 2014; Hall, 1999). Dimensionality reduction contemplates two main methods: a) feature selection, it is an important method that effectively solves dimensionality problems, such as decreasing redundancy and improving the understandability of the results, and b) feature extraction, which searches for the most distinctive, informative and reduced subset of features, improving data processing and storage (Driss Hanafi et al., 2023; Macea-Anaya et al., 2023; Olusegun Oyetola et al., 2023; Zebari et al., 2020).

According to Chandrashekar & Sahin (2014), the methods for the selection of characteristics: a) filtering methods, which use various classification techniques, the criterion used for the selection is ordering; b) wrapping methods, which use the predictor as a black box, the evaluation of the subset of variables is carried out using the performance

metrics as an objective function; c) integrated methods, it has the particularity of reducing the calculation period for the reclassification of different subsets, the criterion is to incorporate the selection of functions as part of the training process. Li et al. (2016) state that is selecting characteristics as a strategy is effective and efficient for creating more straightforward and more understandable models, improving performance, and preparing clean and understandable data. Saeys et al. (2007) argue that the most critical objectives of feature selection are: a) avoid overfitting and improve model performance; b) generate faster and more reliable models; c) gain insight into the underlying processes that generated the data.

Problem

In the United States, in 2012, the outstanding balance of student loans exceeded one trillion dollars; between 2005 and 2012, the delinquency rate on student loans increased by 77%. This figure was negatively associated with the suicide rate among people between 20 and 34 (Jones, 2019). On the other hand, in South Africa, one of the first challenges faced by future university students is the need to obtain financing for their studies, so universities are being pressured to grant scholarships and ensure that students are not excluded. of the university system (López Pérez et al., 2022; McKay et al., 2021). Higher education in Latin America and the Caribbean, according to Gazzola (2021), presents disruptions and instability in the region, which try to stop the transformative incidence given by the interest of the elites in betting on the privatization of higher education institutions; Another of the indicated factors is corruption, which has been impacting resources and generating an ethical and moral crisis; in addition to corporatism, by not defending the necessary changes to give new meaning to higher education; and finally, the discontinuity since each government in power has impacts that do not guarantee stable regulatory frameworks and resources.

In Peru, the Federation of Private Institutions of Higher Education (FIPES), through its president Juan Ostopa indicates that 15% of students dropped out of the university during the state of emergency, and they also estimate that in the following semester, the university desertion would arrive at 35%, there will be approximately 350,000 students who will stop studying; In addition, payment delinquency reaches 50%, making it difficult or even impossible to sustain the universities, which would allow going back on the university reform, but would definitely not have qualified personnel (FIPES cited by Quinto, 2020). SUNEDU (2021) argues that 28% of young Peruvians had access to university, and only 10.3% of young adults had access to postgraduate studies. This is due to the health crisis that has significantly impacted master's programs and doctorate. Likewise, one in five students did not have a computer at home, and 22% did not have an Internet connection at home, mitigating these inconveniences with mobile Internet devices. The interruption of studies increased significantly, affecting private universities from 6% to 18% from 2019 to 2020.

Higher education institutions, throughout their institutional life, according to Mense et al. (2020), have been strategically looking for reliable methods and means to improve the learning process. It has been considered one of the current challenges due to the growth of educational data and how to use it to improve the quality of decision-making associated with efficiency, objectivity, transparency, and innovation of organizations. On the other hand, machine learning has been achieving a significant impact on society due to the diversity of solutions to solve complicated problems of reality such as classification, prediction, and grouping occurring during the pandemic (Abdul et al., 2022; Cárdenas Espinosa et al., 2023; Correa Moreno & González Castro, 2023; Junco Luna, 2023; Silva-Sánchez, 2022); Thus, academic and industrial areas have included recognition of patterns and trends, computer vision and natural language processing have demonstrated the capacity of deep networks influencing performance and improving results, achieving the development of the sector (Wang et al., 2016; Albarracín Vanoy, 2022; Khalaf, 2021).

The private university under study is an academic community with social responsibility, aligns its research, teaching, cultural outreach, and social projection activities to provide comprehensive education with a clear awareness of the country as a multicultural reality; according to its internal regulations, it is adequate to University Law No. 30220; Likewise, its economic and financial resources derived from the commitments of the students, generated by tuition fees, tuition fees, educational fees, debts receivable among other income or contributions, the same that are breached considering the established payment schedule, caused by several factors.

First of all, we can indicate that students have generalized an inadequate culture of payment, the same that is carried out at the end of the academic semester, protected by the validity of Law 29947 dated November 28, 2012, Law for the Protection of the Economy Family, which allows students to continue their studies without fulfilling the economic commitments generated, causing high delinquency rates affecting the economic flow of the organization. Secondly, the collection strategies and policies in the organization are inadequate, the same ones that generate reluctance to pay pensions on the part of the students, even generating complaints to the regulatory and consumer protection entity, impacting fines and sanctions savings towards the university. Thirdly, the deceleration of macroeconomic variables, corruption, insecurity, the COVID-19 pandemic, social upheaval, and natural disasters converge in this problem, influencing the drop in employment rates. All this contributes to an irresponsible culture on the part of the users of the university service; That is why it is necessary to know the current situation of payment behavior in university students through a system based on data mining and artificial intelligence. For these reasons, we ask ourselves: What is the classification model based on machine learning algorithms and data mining to predict the payment behavior of students in a private university?

Objective

Develop a classification model to predict the payment behavior of students in a private university by applying machine learning algorithms and data mining techniques.

2. Methodology

The research was focused on developing a classification model for payment behavior to predict non-payment students. 8495 students have been considered as participants, and a detailed analysis of the data was carried out using techniques and methods of automatic learning, techniques for unbalanced data and feature selection being fundamental tools that help find patterns in the data; In addition, the help of the H2O platform, it was possible to understand the patterns and predict the students who have late or owed commitments. This tool made it possible to identify the model with greater precision and better performance in quality metrics. In this way, it will be possible to have a prior classification of students who are unpaid and thus develop assertive strategies to counteract the situation opportunely.

The systems used as tools in the generation of the classification model for payment behavior were the R Statistical Software language (v4.2.2; R Core Team 2022) together with the R Studio development environment (v2023.03.1 Build 446; Posit Team 2023) installed on the Windows 11 desktop operating system (x64 build 22621); Regarding the packages used, there is H2O (v 3.40.0.4; Fryda et al. 2023) proposed by the H20.ai platform for the generation of the classification model. For dimensionality reduction through feature selection, the following packages were used: familiar (v1.4.1; Zwanenburg & Löck 2021), Information (v0.0.9; Kim 2016), Boruta (v8.0.0; Kursa & Rudnicki 2010); for Tidyverse data exploration and analysis (v2.0.0; Yutani 2019); The control of unbalanced data was used imbalance (v 1.0.2.1; Cordón et al. 2018), ROSE (v0.0.4; Lunardon et al. 2014) and in the variable discretization process using fastDummies (v1.6.3; Kaplan 2020).

3. Results

The data set was extracted from the computer system of the higher education institution. The characteristics or variables are detailed in Table 1.

Table 1. Description of data set variables.

N	Description	Type
01	Sex	Dichotomic
02	School type	Polytomous
03	Work	Dichotomic
04	Economic status	Ordinal
05	Disability	Dichotomic
06	Family support	Dichotomic
07	Marital status	Dichotomic
08	Scholarship	Dichotomic
09	Affiliation	Polytomous
10	Faculty	Polytomous
11	Regular enrollment	Dichotomic
12	Approved	Dichotomic
13	Debt	Dichotomic

The data analysis process was developed in several stages. First, data cleaning and preparation was performed by removing outliers, coding categorical variables, and creating additional variables, also called variable discretization. Regarding the discretization process of the variables, the result was a dichotomous data set comprising 33 predictor or exogenous variables or characteristics and one response or endogenous variable, allowing better performance in machine learning algorithms.

Subsequently, the resampling was carried out, that is, sub-sampling and over-sampling in the participants, to guarantee adequate proportions of the objective or response variable.

Table 2. Description of the data set with its respective characteristics

N	Sample method	Variables	Cases			Characteristics selection	Discretization
			Total	Training	Test		
01	Under	34	4276	2992	1284	SI	SI
02	Over	34	10813	7628	3185	SI	SI
03	Both	34	8495	5987	2508	SI	SI
04	**MWSmote**	**34**	**12606**	**8799**	**3807**	**SI**	**SI**
05	Smote	34	12606	8817	3789	SI	SI
06	MWSmote	13	12606	8820	3786	SI	NO
07	Smote	13	12606	8847	3759	SI	NO

Next, we reduced the number of variables using the feature selection method, reducing five independent variables. Then the automatic learning algorithms were trained, where 70% and 30% were considered for the sizes of the data sets, and

the H2O.automl method was executed, allowing us to find the best prediction models for each case with their respective metrics.

Table 3. Description of quality metrics of models with training data.

N	Model description	AUC	LOGLOSS	AUCPR
01	GBM_grid	0.615	0.625	0.739
02	StackedEnsemble_AllModels	0.763	0.567	0.632
03	StackedEnsemble_AllModels	0.771	0.540	0.647
04	**GBM_grid**	**0.894**	**0.329**	**0.914**
05	StackedEnsemble_AllModels	0.824	0.486	0.820
06	StackedEnsemble_AllModels	0.727	0.601	0.680
07	StackedEnsemble_AllModels	0.718	0.608	0.672

Table 3 contains the models with the highest performance of each data set, considering the predefined parameters; the machine learning model with the best performance is the Gradient Bosting Machine (GBM Grid) compared to the other models obtained. Finally, the results of the models were compared with the test data set to select the best classification model. The main performance metrics were evaluated according to Table 4.

Table 3. Description of quality metrics of models with test data.

N	Model description	GINI	AUC	AUCPR	LOGLOSS
01	GBM_grid	0.253	0.626	0.732	0.630
02	StackedEnsemble_AllModels	0.552	0.776	0.639	0.563
03	StackedEnsemble_AllModels	0.554	0.777	0.658	0.530
04	**GBM_grid**	**0.810**	**0.905**	**0.926**	**0.311**
05	StackedEnsemble_AllModels	0.663	0.831	0.830	0.489
06	StackedEnsemble_AllModels	0.449	0.724	0.674	0.603
07	StackedEnsemble_AllModels	0.318	0.659	0.628	0.641

The H2O.ai platform helps identify the most appropriate prediction model to predict the delinquency behavior of students through performance metrics, which is vital in predicting delinquent students to carry out corrective actions allowing adequate and timely management of risks due to non-payment.

Table 4. GBM Grid classification model confusion matrix.

Prediction values	Real values		Error	Rate
	Yes	No		
Si	4870	40	0.008147	= 40 / 4910
No	1052	2837	0.270507	= 1052 / 3889
Total	5922	2877	0.124105	= 1092 / 8799

Among the most reliable metrics, we have the accuracy, which allows for determining the correct predictions. It is defined as the total proportion of predictions made. A score close to unity represents optimal performance. Table 4, we can obtain an accuracy equivalent to 87.59% (true positives and false negatives divided by the total). That is, the model has a predictive capacity of 100 observations, and it can successfully predict 87 cases; for sensitivity, there is 82.24%, indicating a prediction of 100 cases out of 82 are successful for the positive class; finally, for specificity, we identified 98.61% of the cases to predict the negative class. Likewise, the ROC curve is shown. It is a graph representing the relationship between true positives (sensitivity) and false positives (specificity). A curve close to the upper left corner is demonstrated, thus indicating optimal performance. It should be noted that the lower left side of the graph represents a lower tolerance for false positives. In contrast, the upper right side represents a higher false-positive tolerance.

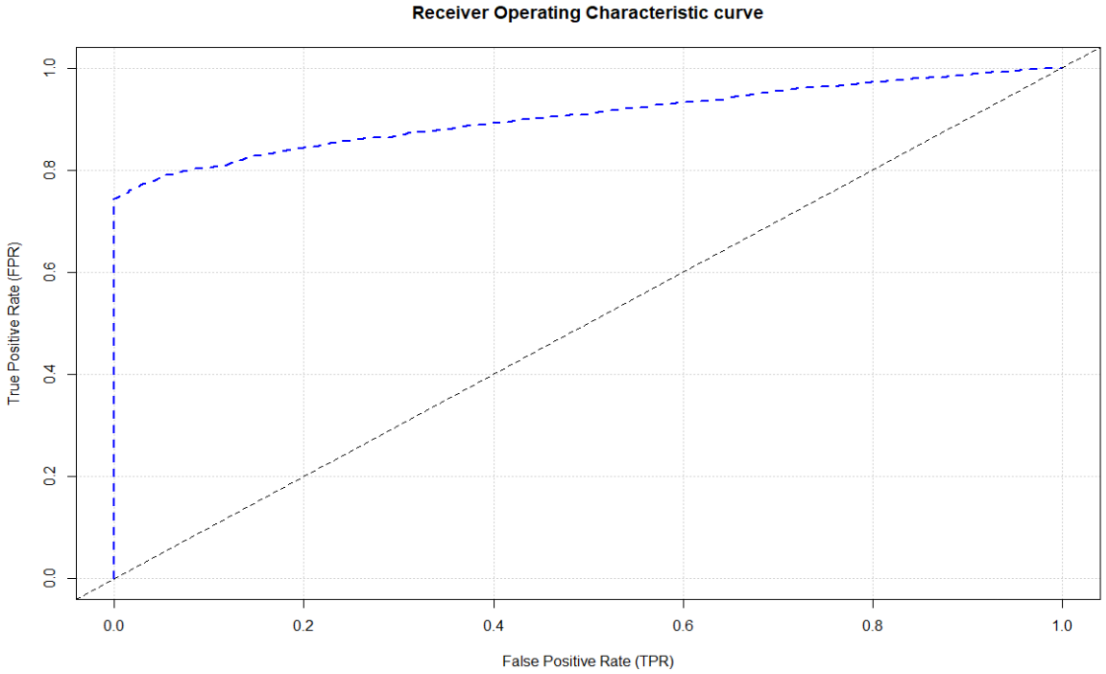

Figure 1. ROC curve of the GMB Grid model.

4. Conclusions

Having developed the GBM Grid classification model for the prediction of payment behavior in students of a private university in Peru, we can conclude that it meets the performance metrics, such as a GINI of 0.810, AUC of 0.905, AUCPR of 0.926 and LOGLOSS of 0.311. In addition, the precision, sensitivity, and specificity demonstrated a high success rate in obtaining satisfactory results. Likewise, the model supports unbalanced data and multiclass characteristics.

The results were improved by adjusting the model parameters. The cross-validation allowed evaluating the model's accuracy and making predictions in real-time. These results show an efficient classification model, having the ability to use algorithms to infer the knowledge obtained from a data set, which allows rigorous monitoring and analysis to detect delinquency in university students, achieving greater financial control in the student body, helping the authorities to apply tools to promote financial responsibility. The research contributes significantly to knowledge by providing a tool for decision-making in finance to predict student delinquency and, at the same time, to understand in depth the causes that lead to this phenomenon. Regarding the practical implications, a model has been satisfactorily established to identify students with delinquent tendencies, which allows the university to take precautions and measures to avoid non-compliance with tuition and tuition payments so that the impact is reduced.

The generated classification model can be used by those responsible for the financial administration of the institution to apply policies and improvements in collection procedures to prevent future inconveniences related to delinquency; that is, having the possibility of early intervention in potentially problematic situations through the implementation of strategies to grant benefits to students in financial matters such as scholarships, discounts, incentives, promotions, educational loans, among others; in order to achieve greater satisfaction among them and, at the same time, reduce administrative costs. In this way, it improves the university community's quality of services and financial security. As a suggestion, the classification model for predicting student payment behavior should be implemented in the university institution for timely intervention by the responsible personnel. Later, measuring the follow-up of the students who have received the intervention programs will be possible. Finally, it will be possible to evaluate the effectiveness of the intervention programs offered by the institution. On the other hand, the functionality and processing capacity provided by the H2O.ai platform for the automatic generation of learning models saves time and resources, allowing users to perform data preprocessing, model generation using training data easily; and, later, the evaluation of the metrics of the model with test data allowing H2O to identify the optimal model according to the defined parameters.

References

[1] Abdul, M., Yusoff, M. & Mohamed, A. (2022). Survey on Highly Imbalanced Multi-class Data, International Journal of Advanced Computer Science and Applications (IJACSA), 13(6). http://dx.doi.org/10.14569/IJACSA.2022.0130627

[2] Albarracín Vanoy, R. J. (2022). STEM Education as a Teaching Method for the Development of XXI Century

Competencies. *Metaverse Basic and Applied Research*, *1*, 21. https://doi.org/10.56294/mr202221

[3] Andrade-Girón, D., Carreño-Cisneros, E., Mejía-Dominguez, C., Marín-Rodriguez, W., & Villarreal-Torres, H. (2023). Comparison of Machine Learning Algorithms for Predicting Patients with Suspected COVID-19. *Salud, Ciencia Y Tecnología*, *3*, 336. https://doi.org/10.56294/saludcyt2023336

[4] Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., & Atkinson, P.M. (2021). Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11. DOI:10.1002/widm.1424

[5] Arrieta, A.B., Rodríguez, N.D., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. ArXiv, abs/1910.10045.

[6] Bartschat, A., Reischl, M., & Mikut, R. (2019). Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1309. doi:10.1002/widm.1309

[7] Benito, P. V. (2022). Contemporary art and networks: Analysis of the Venus Project using the UCINET software. *AWARI*, *3*. https://doi.org/10.47909/awari.166

[8] Bolaño García, M., Duarte Acosta, N., & González Castro, K. (2023). Scientific production on the use of ICT as a tool for social inclusion for deaf people: a bibliometric analysis. *Salud, Ciencia Y Tecnología*, *3*, 318. https://doi.org/10.56294/saludcyt2023318

[9] Cárdenas Espinosa, R. D., Caicedo-Erazo, J. C., Arbeláez Londoño, M., & Jimenez Pitre, I. (2023). Inclusive Innovation through Arduino Embedded Systems and ChatGPT. *Metaverse Basic and Applied Research*, *2*, 52. https://doi.org/10.56294/mr202352

[10] Catrambone, A. R., & Ledwith, A. S. (2023). Acompañamiento interdisciplinar de las trayectorias académicas, en formación docente y psicopedagógica. *Salud, Ciencia Y Tecnología - Serie De Conferencias*, *2*(1), 186. https://doi.org/10.56294/sctconf2023186

[11] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Comput. Electr. Eng., 40, 16-28. DOI: 10.1016/j.compeleceng.2013.11.024

[12] Chatterjee, J., Garg, H. & Thakur, R.N. (2023). A Roadmap for Enabling Industry 4.0 by Artificial Intelligence. Wiley. ISBN 978-1-119-90485-4

[13] Chawla, N.V. (2005). Data Mining for Imbalanced Datasets: An Overview. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/0-387-25465-X_40

[14] Chawla, N.V. (2009). Data Mining for Imbalanced Datasets: An Overview. In: Maimon, O., Rokach, L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-09823-4_45

[15] Chen, L., Chen, P. & Lin, Z. (2020). Artificial Intelligence in Education: A Review, in IEEE Access, vol. 8, pp. 75264-75278, doi: 10.1109/ACCESS.2020.2988510.

[16] Chryssolouris, G., Alexopoulos, K. & Arkouli, Z. (2023). Perspective on Artificial Intelligence in Manufacturing. Springer. https://doi.org/10.1007/978-3-031-21828-6

[17] Cordón, I., García, S., Fernández, A. & Herrera, F. (2018). Imbalance: Oversampling algorithms for imbalanced classification in R. Knowledge-Based Systems, 161, 329-341. https://doi.org/10.1016/j.knosys.2018.07.035.

[18] Corrêa da Silva, F. C. (2022). The value of information in the face of new global disorder. *AWARI*, *3*. https://doi.org/10.47909/awari.165

[19] Correa Moreno, M. C., & González Castro, G. L. (2023). Unveiling Public Information in the Metaverse and AI Era: Challenges and Opportunities. *Metaverse Basic and Applied Research*, *2*, 35. https://doi.org/10.56294/mr202335

[20] de Araújo Telmo, F., Matos Autran, M. de M., & Araújo da Silva, A. K. (2021). Scientific production on open science in Information Science: a study based on the ENANCIB event. *AWARI*, *2*, e027. https://doi.org/10.47909/awari.127

[21] de Araújo Telmo, F., Matos Autran, M. de M., & Araújo da Silva, A. K. (2021). Scientific production on open science in Information Science: a study based on the ENANCIB event. *AWARI*, *2*, e027. https://doi.org/10.47909/awari.127

[22] do Carmo, D., & da Silva Lemos, D. L. (2022). Quality standards for data and metadata addressed to data science applications. *Advanced Notes in Information Science*, *2*, 161–170. https://doi.org/10.47909/anis.978-9916-9760-3-6.116

[23] Driss Hanafi, M., Lali, K., Kably, H., & Chakor, A. (2023). The English Proficiency and the Inevitable Resort to Digitalization: A Direction to Follow and Adopt to Guarantee the Success of Women Entrepreneurs in the World of Business and Enterprises. *Data & Metadata*, *2*, 42. https://doi.org/10.56294/dm202342

[24] Francis, B.K., Babu, S.S. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. J Med Syst 43, 162 (2019). https://doi.org/10.1007/s10916-019-1295-4

[25] Fryda, T., LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., Malohlava, M., Poirier, S., Wong, W. (2023). h2o: R Interface for the 'H2O' Scalable Machine Learning Platform. R package version 3.40.0.4, https://CRAN.R-project.org/package=h2o

[26] Garg, S.K., & Sharma, A.K. (2013). Comparative Analysis of Various Data Mining Techniques on Educational Datasets. International Journal of Computer Applications, 74, 1-5. https://research.ijcaonline.org/volume74/number5/pxc3889673.pdf

[27] Gazzola, A. (18 de octubre de 2021). Educación superior en América Latina y Caribe, presente y futuro. UNESCO. https://www.iesalc.unesco.org/2021/10/18/educacion-superior-en-america-latina-y-caribe-presente-y-futuro/

[28] Ghanem, A. S., Venkatesh, S., & West, G. (2008). Learning in imbalanced relational data. 2008 19th International Conference on Pattern Recognition. doi:10.1109/icpr.2008.4761095

[29] Hall, M. (1999). Correlation-based Feature Selection for Machine Learning [Tesis doctoral, Universidad de Waikato]. Repositorio institucional de la Universidad Waikato https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

[30] Hancock, J.T., Khoshgoftaar, T.M. & Johnson, J.M. Evaluating classifier performance with highly imbalanced Big Data. J Big Data 10, 42 (2023). https://doi.org/10.1186/s40537-023-00724-5

[31] Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31, 685-695. https://doi.org/10.1007/s12525-021-00475-2

[32] Jones, R. W. (2019). The Impact of Student Loan Debt and Student Loan Delinquency on Total, Sex-, and Age-specific Suicide Rates during the Great Recession. Sociological Inquiry, 89(4), 677–702. doi:10.1111/soin.12278

[33] Junco Luna, G. J. (2023). Study on the impact of artificial intelligence tools in the development of university classes at the school of communication of the Universidad Nacional José Faustino Sánchez Carrión. *Metaverse Basic and Applied Research*, *2*, 51. https://doi.org/10.56294/mr202351

[34] Kaplan, J. (2020). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. R package version 1.6.3, https://CRAN.R-project.org/package=fastDummies.

[35] Khalaf, A.S., Dahr, J.M., Najim, I.A., Kamel, M.B., Hashim, A.S., Awadh, W.A., & Humadi, A.M. (2021). Supervised Learning Algorithms in Educational Data Mining: A Systematic Review.

[36] Kim, L. (2016). _Information: Data Exploration with Information Theory (Weight-of-Evidence and Information Value). R package version 0.0.9, https://CRAN.R-project.org/package=Information.

[37] Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. Wiley Interdisciplinary Reviews: Cognitive Science, 6(4), 333–353. doi:10.1002/wcs.1350

[38] Kühl, N., Schemmer, M., & Goutier, M. (2022). Satzger, G. Artificial intelligence and machine learning. Electron Markets 32, 2235–2244. https://doi.org/10.1007/s12525-022-00598-0

[39] Kursa, M.B. & Rudnicki, W.R. (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1-13. https://doi.org/10.18637/jss.v036.i11.

[40] Lali, K., & Chakor, A. (2023a). Improving the Security and Reliability of a Quality Marketing Information System: A Priority Prerequisite for Good Strategic Management of a Successful Entrepreneurial Project. *Data & Metadata*, *2*, 40. https://doi.org/10.56294/dm202340

[41] Lali, K., Chakor, A., & El Boukhari, H. (2023b). The Digitalization of Production Processes : A Priority Condition for the Success of an Efficient Marketing Information System. Case of the Swimwear Anywhere Company. *Data & Metadata*, *2*, 41. https://doi.org/10.56294/dm202341

[42] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., & Liu, H. (2016). Feature Selection. ACM Computing Surveys (CSUR), 50, 1 - 45. DOI:10.1145/3136625

[43] Liu, C., Jin, S., Wang, D., Luo, Z., Yu, J., Zhou, B., & Yang, C. (2020). Constrained Oversampling: An Oversampling Approach to Reduce Noise Generation in Imbalanced Datasets with Class Overlapping. IEEE Access, 1–1. doi:10.1109/access.2020.3018911

[44] López Pérez, T. E., Manzano Pérez, R. S., Manzano Pérez, R. J., & Zumbana Herrera, L. F. (2022). Methodological strategies to strengthen the teaching-learning process in basic education children. *Salud, Ciencia Y Tecnología*, *2*(S1), 254. https://doi.org/10.56294/saludcyt2022254

[45] Lunardon, N., Menardi, G., Torelli, N. (2014). ROSE: a Package for Binary Imbalanced Learning. R Journal, 6(1), 82-92.

[46] Macea-Anaya, M., Baena-Navarro, R., Carriazo-Regino, Y., Alvarez-Castillo, J., & Contreras-Florez, J. (2023). Designing a Framework for the Appropriation of Information Technologies in University Teachers: A Four-Phase Approach. *Data & Metadata*, *2*, 53. https://doi.org/10.56294/dm202353

[47] Marinho de Sousa, R. P., & Shintaku, M. (2022). Data privacy policy: relevant observations for its implementation. *Advanced Notes in Information Science*, *2*, 82–91. https://doi.org/10.47909/anis.978-9916-9760-3-6.112

[48] Martín Ferron, L. (2022). Jumping the Gap: developing an innovative product from a Social Network Analysis perspective. *AWARI*, *2*, e026. https://doi.org/10.47909/awari.128

[49] McKay, T., Naidoo, A. & Simpson, Z. (2021). Exploring the Challenges of First-Year Student Funding: An Intra-Institutional Case Study. DOI: 10.24085/jsaa.v6i1.3063

[50] Mejías, M., Guarate Coronado, Y. C., & Jiménez Peralta, A. L. (2022). Inteligencia artificial en el campo de la enfermería. Implicaciones en la asistencia, administración y educación. *Salud, Ciencia Y Tecnología*, *2*, 88. https://doi.org/10.56294/saludcyt202288

[51] Mense, E. G., Lemoine, P. A., & Richardson, M. D. (2020). Data Mining in Global Higher Education: Opportunities and Challenges for Learning. In C. Bhatt, P. Sajja, & S. Liyanage (Eds.), Utilizing Educational Data Mining Techniques for Improved Learning: Emerging Research and Opportunities (pp. 86-120). IGI Global. https://doi.org/10.4018/978-1-7998-0010-1.ch005

[52] Miller, T. (2017). Explanation in Artificial Intelligence: Insights from the Social Sciences. Artif. Intell., 267, 1-38. DOI: 10.1016/J.ARTINT.2018.07.007

[53] Minh, D., Wang, H.X., Li, Y.F., & Nguyen, T.N. (2021). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 55, 3503 - 3568. DOI: 10.1007/s10462-021-10088-y

[54] Mirande, S. N., & Martínez Debat, C. (2023). Conflictos de Intereses, Ghostwriting, Invasiones Epistémicas, Principio Precautorio y un Análisis de Riesgo de las vacunas de ARNm modificado. *Salud, Ciencia Y Tecnología - Serie De Conferencias*, *2*(1), 105. https://doi.org/10.56294/sctconf2023105

[55] OECD (2022), Education at a Glance 2022: OECD Indicators, OECD Publishing, Paris, https://doi.org/10.1787/3197152b-en.

[56] Olufemi, J. (2021). The Concept of Data Mining. Artificial Intelligence. DOI:10.5772/intechopen.99417

[57] Olusegun Oyetola, S., Oladokun, B. D., Ezinne Maxwell, C., & Obotu Akor, S. (2023). Artificial intelligence in the library: Gauging the potential application and implications for contemporary library services in Nigeria. *Data & Metadata*, *2*, 36. https://doi.org/10.56294/dm202336

[58] Posit Team (2023). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

[59] Quinto, C. (6 de agosto de 2020). El 15% de estudiantes abandonó la universidad durante el estado de emergencia, según gremio de instituciones privadas. RPP. https://rpp.pe/peru/actualidad/covid-19-el-15-de-estudiantes-abandono-la-universidad-durante-el-estado-de-emergencia-segun-gremio-de-instituciones-privadas-noticia-1283361?ref=rpp

[60] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[61] Ridley, M. (2022). Explainable Artificial Intelligence (XAI). Information Technology and Libraries.

[62] Romero, C. & Ventura, S. (2012). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12–27. doi:10.1002/widm.1075

[63] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2021). Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. ArXiv, abs/2103.11251.

[64] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics, 23 19, 2507-17. DOI:10.1093/bioinformatics/btm344

[65] Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. Sn Computer Science, 2. 160. https://doi.org/10.1007/s42979-021-00592-x

[66] Sharma, P., & Sharma, D. S. (2018). DATA MINING TECHNIQUES FOR EDUCATIONAL DATA: A REVIEW. International Journal of Engineering Technologies and Management Research, 5(2), 166–177. https://doi.org/10.29121/ijetmr.v5.i2.2018.641

[67] Shi, Y. (2022). Feature Selection. In: Advances in Big Data Analytics. Springer, Singapore. https://doi.org/10.1007/978-981-16-3607-3_4

[68] Silva Coimbra, F., & Rodrigues Dias, T. M. (2022). A process for the identification and analysis of scientific articles in conference proceedings. *Advanced Notes in Information Science*, *2*, 74–81. https://doi.org/10.47909/anis.978-9916-9760-3-6.93

[69] Silva, E. (2022). Digital transformation and knowledge management: relationships in scientific production. *Advanced Notes in Information Science*, *2*, 43–52. https://doi.org/10.47909/anis.978-9916-9760-3-6.107

[70] Silva-Sánchez, C. A. (2022). Psychometric properties of an instrument to assess the level of knowledge about artificial intelligence in university professors. *Metaverse Basic and Applied Research*, *1*, 14. https://doi.org/10.56294/mr202214

[71] Subbarayan, S., & Gunaseelan, H. G. (2022). A Review of Data and Document Clustering pertaining to various Distance Measures. *Salud, Ciencia Y Tecnología*, *2*(S2), 194. https://doi.org/10.56294/saludcyt2022194

[72] Sumitha, R., & Vinothkumar, E. (2016). Prediction of Students Outcome Using Data Mining Techniques.

[73] Superintendencia Nacional de Educación Superior (2021). III Informe Bienal sobre la Realidad Universitaria en el Perú. https://www.gob.pe/institucion/sunedu/informes-publicaciones/2824150-iii-informe-bienal-sobre-la-realidad-universitaria-en-el-peru.

[74] Takaki, P., & Dutra, M. (2022). Data science in education: interdisciplinary contributions. *Advanced Notes in Information Science*, *2*, 149–160. https://doi.org/10.47909/anis.978-9916-9760-3-6.94

[75] Tan, P., Steinbach, M.S., & Kumar, V. (2022). Introduction to Data Mining. Data Mining and Machine Learning Applications. https://doi.org/10.1002/9781119792529.ch1

[76] Vähäkainu, P. & Lehto, M. (2023). Use of Artificial Intelligence in a Cybersecurity Environment. En T. Sipola, T. Kokkonen & M. Karjalainen (Eds.). Artificial Intelligence and Cybersecurity: Theory and Applications (pp. 3 - 27). Springer. https://doi.org/10.1007/978-3-031-15030-2

[77] Venkatesh, B. & Anuradha, J. (2019). A Review of Feature Selection and Its Methods. Cybernetics and Information Technologies, 19(1) 3-26. DOI: https://doi.org/10.2478/cait-2019-0001

[78] Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., & Kennedy, P. J. (2016). Training deep neural networks on imbalanced data sets. 2016 International Joint Conference on Neural Networks (IJCNN). doi:10.1109/ijcnn.2016.7727770

[79] Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. doi:10.21105/joss.01686 <https://doi.org/10.21105/joss.01686.

[80] Wiesmüller, S. (2023). The Relational Governance of Artificial Intelligence, Forms and Interactions. Springer. https://doi.org/10.1007/978-3-031-25023-1

[81] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. 9, 11. https://doi.org/10.1186/s40561-022-00192-z

[82] Yin, J., Gan, C., Zhao, K., Lin, X., Quan, Z., & Wang, Z.-J. (2020). A Novel Model for Imbalanced Data Classification. Proceedings of the AAAI Conference on Artificial Intelligence, 34(04), 6680-6687. https://doi.org/10.1609/aaai.v34i04.6145

[83] Zebari, R., Abdulazeez, A., Zeebaree, D., Zebari, D., & Saeed, J. (2020). A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. Journal of Applied Science and Technology Trends, 1(2), 56 - 70. https://doi.org/10.38094/jastt1224

[84] Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., Ma, X., Marrone, B. L., Ren, Z. J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B. M., Xiao, X., Yu, X., Zhu, J. J., & Zhang, H. (2021). Machine Learning: New Ideas and Tools in Environmental Science and Engineering. Environmental science & technology, 55(19), 12741–12754. https://doi.org/10.1021/acs.est.1c01339

[85] Zhou, J., Gandomi, A.H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics, 10, 593. https://doi.org/10.3390/electronics10050593

[86] Zwanenburg, A. (2021). familiar: Vignettes and Documentation. https://github.com/alexzwanenburg/familiar.

[87] Zwanenburg, A., & Löck, S. (2021). familiar: End-to-End Automated Machine Learning and Model Evaluation. https://github.com/alexzwanenburg/familiar.