Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Early Prediction of University Student Dropout Using Machine Learning Models

José Orlando Quintana Quispe[1], Osmar Cuentas Toledo[1], Maryluz Cuentas Toledo[2], Elmer Elio Calizaya Llatasi[3], Elizabeth Marina Ramos Saira[1]

[1]Universidad Nacional de Moquegua - Perú
[2]Universidad Nacional Micaela Bastidas de Apurímac - Perú
[3]Universidad Nacional del Altiplano – Perú
Email: jquintanaq@unam.edu.pe

The study analyzed and validated the best machine learning model to predict dropout among students at the National University of Moquegua, Mariscal Nieto campus, between 2009 and 2019. It also identified the characteristics influencing dropout. The models used were logistic regression, decision trees, support vector machines, naive Bayes, the CRISP-DM methodology, and metrics such as the confusion matrix and K-fold cross-validation. We selected important features using two techniques, random forests and Featurewiz, employing Python and its libraries for the task. The research type is descriptive correlational, and the design is observational using secondary data sources. The sample was selected through stratified random probability sampling, resulting in a sample of 109 students who dropped out without completing their studies and 220 who did not, totalling 329 data points. Cross-validation indicated that decision trees achieved 76% accuracy, followed by logistic regression at 73%, support vector machines at 71%, and naive Bayes at 62%. The characteristics influencing dropout include general data such as cycle, age, and zonal address; economic aspects such as total income, family load, children in higher education, household components, student's housing, and head of the household; and housing aspects such as type of construction and number of bedrooms.

Keywords: University dropout, machine learning, university student, predictive models.

1. Introduction

University dropout is a problem of great concern in all higher education institutions [1] because university dropout rates are very high [2], in addition to the reasons why there are many dropouts: students, higher education institutions and the government entities in charge and their implemented policies [3]. However, dropout has a wide variety of causes, all from the researcher's perspective [4]; we determined 112 factors to predict university student dropout, identifying them from the five dimensions. (personal, academic, economic, social and

institutional) [5]. It is essential to find the variables that prevent dropout and be able to predict it [4]; knowing this, the institutions in charge could take action to prevent students from abandoning university studies [6]. Likewise, there are tools based on artificial intelligence and machine learning that allow, based on known information, to detect patterns to make predictions about new information [7], which could allow the development of a prediction tool that would be useful for use in desertion universities [6].

Research by [8] analyzes the problem of university dropout in Peru, presenting the factors that influence it and proposing strategies to address the issue. The study used Spearman's rank correlation analysis technique to determine the relevance of the different criteria and categories influencing dropout. Among the main conclusions, the importance of implementing strategies to promote the vocation, the level of motivation and the development of study techniques, and the need to improve the economic and family conditions of the students stand out. Educational institutions must address these challenges to retain students in their academic training.

At the National University of Moquegua, a young institution with about 14 years of history, between 2009 and 2019, 2,305 students enrolled, of which 765 dropped out of the university, representing close to 30% of the students who dropped out due to the high dropout rate, it is crucial to find methods to reduce these numbers. Anticipating student attrition is now a valuable tool for university administrators, allowing them to identify at-risk students and promptly address the situation. [6]. Supports the efforts of the groups responsible for identifying, preventing, and controlling psychosocial risks. It was proven that the J48 methodology is the most effective, supporting the work of the groups in charge of work. Recent studies have confirmed that educators use machine learning methods to predict failure and dropout rates of at-risk students, aiming to improve their performance during their studies [9].

A research article by [8] presents a predictive analytical model to predict students' final grades using supervised machine learning methods. The study used data from 489 students from a polytechnic in northwest Malaysia, applying algorithms such as Decision Tree (J48), Random Forest (RF), Support Vector Machines (SVM) and Logistic Regression (LR). The results showed that J48 was the most accurate model, with an accuracy rate of 99.6%, which could help in the early detection of student dropout and improve academic performance in higher education. The study highlights the importance of predictive analytics in education and suggests future research to expand the data set and improve student achievement.

Another study compares machine learning and deep learning classifiers for specific classification tasks. Classic classifiers such as Random Forest, XGBoost, GMM and SVM, and deep learning classifiers such as CNN and LSTM stand out. The results show that the CNN-based classifier is the most effective, especially in image classification. The conclusion is that the complexity of the classification task significantly influences the performance of classifiers, offering a valuable reference for selecting the appropriate classifier in applied classification tasks. Furthermore, this highlights the importance of continuing research to optimize the selection and application of classifiers in various fields.[10]

A study based on a dataset of 489 students from a polytechnic in northwest Malaysia used supervised machine learning methods to predict students' final grades. The objectives were to detect student dropout and improve academic performance. Researchers applied algorithms such as Decision Tree (J48), Random Forest (RF), Support Vector Machines (SVM), and

Logistic Regression (LR). The results indicated that J48 was the most accurate model, with an accuracy rate of 99.6%. Furthermore, the study suggests expanding the dataset and continuing research to improve student performance [11].

The research aims to select the most significant variables that explain the National University of Moquegua student dropout rates. It also seeks to obtain the best logistic regression model, decision trees, support vector machines and naive Bayes for student dropout at the National University of Moquegua. Finally, by comparing their performance metrics, you will determine the best model among the four machine learning algorithms/models.

2. Materials and Methods

2.1. Study Area

The study focused on students enrolled at the Mariscal Nieto campus of the National University of Moquegua from 2009 to 2019. The university is located in the city of Moquegua, within the Moquegua region of Peru. The coordinates are 17°11'24.02" South Latitude and 70°56'8.88" West Longitude of the Greenwich Meridian, 1410 meters above sea level.

2.2. Methodology

We used the software Python 3.9.12 (main, Apr 4 2022, 05:22:27) [MSC v.1916 64 bit (AMD64)] Anaconda, Inc. on win32. It is a powerful tool for processing and forecasting time series through its libraries, available in the Repository (CRAN).

CRISP-DM (Cross-Industry Standard Process for Data Mining) is a methodology widely used to develop data mining and machine learning projects. It comprises six interrelated phases: business understanding, data understanding, data preparation, modelling, evaluation and deployment. These stages guide teams through the development process, from understanding business objectives and collecting data to implementing and maintaining predictive models. The CRISP-DM methodology provides a solid and systematic structure to address machine learning projects, ensuring quality and effectiveness at each process stage [12].

2.2.1 Supervised machine learning method and binary classification

The study aims to employ supervised machine learning and binary classification techniques using four classifiers: binary logistic regression, decision trees, support vector machines and naïve Bayes [13]. We will use the collected information to conduct data modelling and identify the best model for predicting student dropout at the National University of Moquegua from 2009 to 2019. Additionally, we aim to determine the characteristics that influence this dropout [14].

2.2.2 Focus and type of research

The research adopts a quantitative approach, classified as applied research, and is descriptive and transectional, utilizing a non-experimental design. We will use the hypothetical deductive method, and the technique applied is the student's socioeconomic file when enrolling for the first time at the university.

2.2.3 Application to 2 Python methods

We applied two methods to select characteristics or attributes. The first method, based on random forests, measures the importance of each attribute while reducing data with low correlation. The second method uses Featurewiz, a new Python library that creates high-performance models by selecting the best attributes. Featurewiz examines numerous variables and identifies only the least correlated and most relevant features for your model, using the MRMR (Minimum Redundancy Maximum Relevance) algorithm as the basis for its feature selection (AutoViML/featurewiz: Use advanced feature engineering strategies and select the best features from your dataset with a single line of code) [15], [16].

2.2.4 Comparison of models and Algorithm

The reduction of characteristics is to increase the predictive accuracy and reduce the complexity of the results; likewise, to compare the Logistics, Decision Trees, Support Vector Machine and Naive Bayes classification models, 40 variables were considered (8 numerical, 31 categorical and one objective) and 329 data (109 dropouts and 220 non-dropouts), we will take into account using the algorithms configuration the option class_weight='balanced' for unbalanced classes to avoid bias in the prediction [17], [18] and subsequent coding of numerical variables considering 94 variables (8 numerical, 86 categorical and one objective) and the objective variable is dropout (1 if they drop out and 0 if they do not drop out), in the evaluation of the models and to avoid overtraining is used, the retention technique that consists of dividing the data into 70% data for training and 30% retention for evaluation, and the cross-validation technique with K=5 and K=10 folds; additionally, we use the Area of the ROC curve. We subsequently evaluated four algorithms: Logistic Regression, Decision Trees, Support Vector Machine, and Naive Bayes. We used convenient tuning parameters, such as increasing the number of iterations from 100 to 600 and setting the class weight option to balance for unbalanced classes. For Support Vector Machines, we used a flexible core. Finally, we presented the evaluation results obtained from these algorithms using the confusion matrix, its metrics, and cross-validation with K = 5 and K = 10 folds [19] [20].

2.2.5 Data

This research will focus on a population consisting of three professional schools: two in the engineering area and one in the science area of the National University of Moquegua. Our population includes incoming enrolled students and those who failed or withdrew for any reason in previous semesters between 2009 and 2019. The Professional Schools in question are Professional Schools of Mining Engineering, Professional Schools of Agroindustrial Engineering, and Professional Schools of Public Management and Social Development. The main characteristic of the population is that all Professional Schools develop all semesters from the first to the tenth.

.

3. Results and Discussion

We identified the variables and excluded some characteristics due to incomplete information. These excluded characteristics include code, surnames and names, date of birth, home address, cell phone, family composition, foreign student status, health, and social diagnosis. The socioeconomic sheets reveal that many students did not answer or provide the requested data,

possibly because they needed to understand or consider the requests important (Table 1).

Table 1. Characteristics and dimensions for the study

ATTRIBUTE	DESCRIPTION
DROPOUT	1= Yes dropout; 0= No dropout, binary
PROFESSIONAL SCHOOL	EP_MINAS.EP_AGRO, EP_GPDS, binary 0=No, 1=Yes
CYCLE	CYCLE12, CYCLE34, CYCLE56, CYCLE78, CYCLE910, binary 0=No, 1=Yes
GENERAL DATA	
SEX	1: female and 2: male
AGE	Normalized variable AGE_N
MARITAL STATUS	ESTCIV_S, ESTCIV_C, ESTCIV_OT binaries 0=No, 1=Yes
PLACE OF BIRTH	LNAC_MOQ, LNAC_ILO, LNAC_SC, LNAC_OTRO binaries 0=No, 1=Yes
ZONE ADDRESS	DIRZON_URB, DIRZON_RUR, DIRZON_OTR binaries 0=No,1=Yes
PROVINCE	PROV_MCLN, PROV_ILO, PROV_SC binaries 0=No, 1=Yes
ACADEMIC BACKGROUND	
TYPE OF SCHOOL	TIPO_PRIM, TIPO_SEC, 1=National, 2=Private
PRE-UNIVERSITY PREPARATION	PRE_PP, PREU_AC, PREU_CEPRE, PREU_SOL, binaries 0=No, 1=Yes
MODE OF ENTRY	MD_ORDIN, MD_CEPRE, MD_EXTR, binaries 0=No, 1=Yes
ECONOMIC ASPECT	
FAMILY COMPOSITION	COM_HOGAR number of people in the household, numeric
PARENTS' EDUCATION LEVEL	1=NO EDUCATION, 2=PRIMARY, 3=SECONDARY, 4=HIGHER
HEAD OF HOUSEHOLD	1=PARENTS, 2=FATHER, 3=MOTHER, 4=FAMILY TUTOR, 5=STUDENT
ECONOMIC INCOME MODE	1=MONTHLY, 2=BIWEEKLY, 3=WEEKLY, 4=DAILY
STUDENT'S EMPLOYMENT STATUS	1=WORKS, 0=DOES NOT WORK
FAMILY ECONOMIC INCOME	TOTAL_INGRE
STUDENT'S HOUSING	VIV_CPADRES, VIV_C1PAD, VIV_ALOJ, VIV_CUID, VIVCALQ
FAMILY LOAD	CARGA_FAM, NUMBER OF MEMBERS
CHILDREN IN HIGHER EDUCATION	HIJOS_SUP, NUMBER OF CHILDREN IN HIGHER EDUCATION
STUDENT ASPECT	

ECONOMIC DEPENDENCE	1=BOTH PARENTS, 2= ONLY FATHER, 3=ONLY MOTHER, 4=A RELATIVE
FAMILY RISK	1=CHILD OF LIVING PARENTS, 2=ORPHAN OF MOTHER, 3=ORPHAN OF FATHER, 4=ORPHAN OF BOTH PARENTS, 5=LIVES ALONE
HOUSING ASPECT	
HOUSING TENURE	1=OWN, 2=RENTED, 3=SQUATTER, 4=OTHER
TYPE OF CONSTRUCTION	1=SOLID, 2=MIXED, 3=RUSTIC, 4=PRECARIOUS
TYPE OF HOUSING	1=INDEPENDENT, 2=APARTMENT, 3=TENEMENT, 4=OTHER
SERVICES	SERV_AGUA, SERV_DESAG, SERV_FONO, binaries 0=No, 1=Yes
NUMBER OF FLOORS	N_PISOS
NUMBER OF BEDROOMS	N_DORM
NUMBER OF KITCHENS	NCOCINA
NUMBER OF BATHROOMS	NBANO
NUMBER OF LIVING ROOMS	NSALA
NUMBER OF DINING ROOMS	NCOMEDOR
COLOR TV	binary, 0=No, 1=Yes
RADIO	binary, 0=No, 1=Yes
SOUND EQUIPMENT	binary, 0=No, 1=Yes
IRON	binary, 0=No, 1=Yes
MOBILE PHONE	binary, 0=No, 1=Yes
LAPTOP	binary, 0=No, 1=Yes
CABLE	binary, 0=No, 1=Yes
WARDROBE	binary, 0=No, 1=Yes
REFRIGERATOR	binary, 0=No, 1=Yes
INTERNET	binary, 0=No, 1=Yes
PERSONAL LIBRARY	binary, 0=No, 1=Yes
COMPUTER	binary, 0=No, 1=Yes

We compare dimensions encompassing various causes, all viewed from the researcher's perspective, and identify 112 factors used to predict university student dropout.

We obtained the variables by applying two feature reduction techniques and using binary variables for the applicable categorical ones [22]. When comparing models, those built from complete data performed better than those using reduction techniques (Table 2).

Table 2. Lists the characteristics to be utilized.

	Encoded Data	Encoded Data Dummy
No Feature Reduction	40	94
Random Forests	12	9
Featurewiz	19	24

Below is the summary of the cross-validation accuracies with 5 and 10 sheets. For the selection of the best model, the precision metrics, confusion matrix and additional metrics are used by adjusting the parameters using cross-validation with K=5 and 10 folds (Table 3).

Table 3. Summary of cross-validation accuracies.

Characteristics	Logistic Regression Average K=5 and K=10	Decision Trees Average K=5 and K=10	Support Vector Machine Average K=5 and K=10	Naive Bayes Average K=5 and K=10
40	74%	76%	70%	55%
12	72%	80%	73%	74%
19	73%	79%	70%	74%
94	73%	71%	70%	55%
9	72%	74%	72%	72%
24	74%	77%	72%	43%
Average	73%	76%	71%	62%

The results obtained from data that have no pedagogical or didactic value are a tool to reduce dropout [9]; additionally, to select the best model, precision metrics, confusion matrix and additional metrics are used, adjusting the parameters using cross-validation with K=5 and 10 Folds [9]. Finally, according to [23], you can use the Scikit-Learn cross-validation resource to evaluate optimally. The result is an array that contains the 5 or 10 evaluation scores.

The figure shows that decision trees stand out among the other techniques and highlights that the most incredible precision is achieved when 12 features are used. Most studies incorporate more than three machine learning techniques to compare the models. Decision trees are the most common, used in over 50% of cases. These studies evaluate multiple metrics, including accuracy, precision, recall, and F1 score.

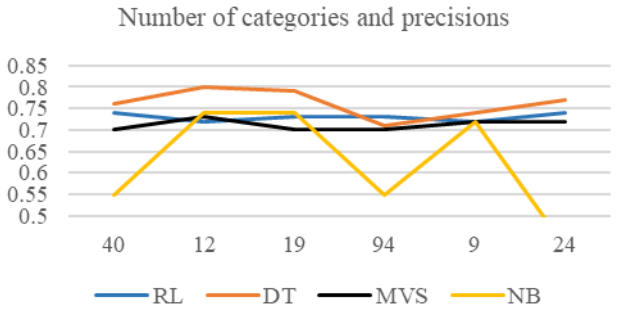

Fig. 1. Logistic Regression (RL), decision trees (DT), support vector machine (MVS) and

naïve Bayes (NB).

In k-fold cross-validation, we randomly split the training data set into k-folds without replacement. Here, k – 1 folds, the so-called training folds, are used for model training, and one fold, the so-called test fold, is used for performance evaluation [9]. This procedure is repeated k times to obtain k models and performance estimates [9] [24]. They consider the performance of each model according to its characteristics and reduction using random forest and Featurewiz techniques. The performance averages indicate that the decision tree model performs best, followed by Logistic regression, support vector machine, and naive Bayes [9].

Table 4 presents the results of the cross-validation with K = 5 and K = 10 Folds, considering the performance of each model according to the characteristics and their reduction using the Random Forest and Featurewiz techniques. The performance averages indicate that the Decision Trees model performs best, followed by Logistic Regression, Support Vector Machine, and Naive Bayes.

Table 4. Average cross-validation K=5 and K=10.

	Cross-Validation K Folds			
	Logistic Regression	Decision Trees	Support Vector Machine	Naive Bayes
Characteristics	K=5, K=10	K=5, K=10	K=5, K=10	K=5, K=10
40	0.7413	0.7608	0.6957	0.5543
12	0.7239	0.8022	0.7261	0.7369
19	0.7347	0.7891	0.6957	0.7435
94	0.7260	0.7086	0.7043	0.5522
9	0.7239	0.7435	0.7117	0.7239
24	0.7391	0.7739	0.7195	0.4347
Average	0.7315	0.7631	0.7105	0.6243

Table 5 shows the attributes that influence each classification model and that have the best performance, noting that in the general aspects, they are cycle and age; in the economic aspect, the total income, family burden and children who study in higher; in the aspect of the student's economic dependence and the family risk for the element of the housing, the type of construction and number of bedrooms.

Table 5. Summary of influencing attributes

ASPECT	12 ATTRIBUTES	9 ATTRIBUTES	24 ATTRIBUTES
GENERAL DATA	CICLO, EDADN, DIR_ZONA	CICLO12, CICLO34, EDADN	CICLO12, EP_MINAS
ECONOMIC ASPECT	TOTAL_INGRE, CARGA_FAM, COMP_HOGAR,	TOTAL_INGRE, CARGA_FAM, COMP_HOGAR,	CARGA_FAM, HI-JOS_SUP, VIV_CUID, VIV_ALOJ

	HIJOS_SUP, VI-VIENDA_EST, SOST_HOGAR	HIJOS_SUP.	
STUDENT ASPECT			DEPEC_SIMIS, HUERF_MAD, HUERF_PAD.
HOUSING ASPECT	TIPO_CONSTR, N_DORM.	N_DORM	TEVIV_PRO, TEVIV_INV, TPC_NOBLE, TPC_MIX, TPC_PREC, TIV_DEPA, TIV_CONV, TIV-OTR, N_BANO, TV_COLOR, CELULAR, INTERNET, COMPUTADORA, BIBLIO_PERS.

4. Conclusion

The research determines the best Machine Learning model to analyze and predict student dropout of students at the National University of Moquegua, Mariscal Nieto headquarters, and concludes:

First, The initial assumption that logistic regression would be the best supervised classification model to predict student dropout has yet to be confirmed according to cross-validation. According to the results, the best classifier is Decision Trees, with its highest precision achieved by reducing features using Random Forest, reducing from 40 to 12 features and reaching an accuracy of 80%. When reducing features from 94 to 24 using Featurewiz twice, Decision Trees achieve an average accuracy of 77%. The evaluation metrics confirm that Decision Trees is the best model for this study of student dropout at the National University of Moquegua. (see Table 4)

Second: While selecting features to predict student dropout, key features were found that formed the most effective model, achieving an accuracy of 80% with only 12 features. These include academic data such as cycle and age and economic information such as total income and family burden. Housing factors, such as construction type and number of bedrooms, were also considered. This study identified the main characteristics influencing student dropout at the National University of Moquegua. (See Table 5)

Third: According to cross-validation, logistic regression has its best precision with 40 characteristics, achieving 74% precision; we correctly classified 74% of the students as

dropouts. The support vector machine reaches its best accuracy when the number of features is reduced from 40 to 12 using Random Forest, achieving 73% accuracy; this means we correctly classified 73% of the students as dropouts. Finally, Naive Bayes achieves its best accuracy when the number of features is reduced from 40 to 12 using Featurewiz, achieving 74% accuracy, indicating that 74% of the students have been classified correctly (See Table 4).

Fourth: According to cross-validation, the Decision Tree model shows the best performance with approximately 77%, followed by Logistic Regression with 73%, then Support Vector Machine with 71%, and finally Naive Bayes with 62%. Cross-validation ensures that the classifier's performance is independent of the data partition, meaning the highest performance is observed in the Decision Tree classifier. (See Table 4)

Fifth, Including dummy variables improves the performance of the four classification models. This improvement is less notable in the first two classifiers, Logistic Regression and Decision Trees, but is compensated when the number of features is reduced to 24 in both cases. In the other two classifiers, Support Vector Machine and Naive Bayes, the improvement is more significant, significantly, when the number of features is reduced to 9 in both cases. (see Table 4)

References

1. V. Tito, “DESERCIÓN ESTUDIANTIL UNIVERSITARIA,” 2020.
2. J. I. Escalante López, C. J. Medina Valderrama, and A. Vásquez Muñoz, “La deserción universitaria: un problema no resuelto en el Perú,” Hacedor - AIAPÆC, vol. 7, no. 1, pp. 60–72, 2023, doi: 10.26495/rch.v7i1.2421.
3. H. Y. Ayala-yaguara and G. M. Valenzuela-sabogal, “Obtaining a data mining model to be applied to university desertion from the Systems Engineering program of the University of Cundinamarca,” vol. 7, 2020.
4. J. G. Preciado-León, J. Huerta-Hernández, J. Á. Vera-Noriega, and R. A. Corral-Guerrero, “Causas asociadas a la deserción escolar en educación superior. Una revisión sistemática del 2010 al 2020,” Ra Ximhai, no. February, pp. 83–101, 2022, doi: 10.35197/rx.18.01.2022.04.jp.
5. M. S. Albán Taipe, “Universidad Nacional Mayor de San Marcos Facultad de Ingeniería de Sistemas e Informática Unidad de Posgrado Contribuciones en el proceso de elicitación de requisitos : factores , actividades y cualidades TESIS Para optar el Grado Académico de Doctor en I,” p. 133, 2019.
6. M. Alban and D. Mauricio, "Neural networks to predict university dropout," Int. J. Mach. Learn. Comput., vol. 9, no. 2, pp. 149–153, 2019, doi: 10.18178/ijmlc.2019.9.2.779.
7. C. Aspilcueta, M. Ingrid, M. Rodriguez, and J. Christal, La inteligencia artificial en la operación del negocio. 2022.
8. C. Perez and L. Rojas, “Diseño de un sistema para predecir la deserción de los alumnos mediante Machine learning en la Universidad Tecnológica del Perú,” 2020.
9. B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student’ performance prediction using machine learning techniques,” Educ. Sci., vol. 11, no. 9, 2021, doi: 10.3390/educsci11090552.
10. H. Gu and J. Jiao, "Comparison of classifiers for different data in classification application," J. Phys. Conf. Ser., vol. 1994, no. 1, 2021, doi: 10.1088/1742-6596/1994/1/012015.
11. M. Solis, T. Moreira, R. Gonzalez, T. Fernandez, and M. Hernandez, “Perspectives to Predict

Dropout in University Students with Machine Learning," 2018 IEEE Int. Work Conf. Bioinspired Intell. IWOBI 2018 - Proc., 2018, doi: 10.1109/IWOBI.2018.8464191.
12. Ignacio G.R. Gavilán, "Metodología para Machine Learning (I): CRISP-DM | Ignacio G.R. Gavilán," 2021.
13. L. Carnevale, A. Celesti, G. Fiumara, A. Galletta, and M. Villari, "Investigating classification supervised learning approaches for the identification of critical patients' posts in a healthcare social network," Appl. Soft Comput. J., vol. 90, p. 106155, 2020, doi: 10.1016/j.asoc.2020.106155.
14. A. Holzinger, "Introduction to MAchine Learning & Knowledge Extraction (MAKE)," Mach. Learn. Knowl. Extr., vol. 1, no. 1, pp. 1–20, 2019, doi: 10.3390/make1010001.
15. A. Nayak, B. Božić, and L. Longo, "Data Quality Assessment and Recommendation of Feature Selection Algorithms: An Ontological Approach," J. Web Eng., vol. 22, no. 1, pp. 175–196, 2023, doi: 10.13052/jwe1540-9589.2219.
16. G. Chao, Y. Luo, and W. Ding, "Recent Advances in Supervised Dimension Reduction: A Survey," Mach. Learn. Knowl. Extr., vol. 1, no. 1, pp. 341–358, 2019, doi: 10.3390/make1010020.
17. D. Dablain, C. Bellinger, B. Krawczyk, D. W. Aha, and N. Chawla, "Understanding imbalanced data: XAI & interpretable ML framework," Mach. Learn., no. Iml, 2024, doi: 10.1007/s10994-023-06414-w.
18. V. García, R. A. Mollineda, and J. S. Sánchez, "A bias correction function for classification performance assessment in two-class imbalanced problems," Knowledge-Based Syst., vol. 59, pp. 66–74, 2014, doi: 10.1016/j.knosys.2014.01.021.
19. D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," Electronics (Switzerland), vol. 8, no. 8. 2019. doi: 10.3390/electronics8080832.
20. B. Wang, J. Zhou, Y. Li, and F. Chen, "Impact of Fidelity and Robustness of Machine Learning Explanations on User Trust," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 14472 LNAI, no. Ml, pp. 209–220, 2024, doi: 10.1007/978-981-99-8391-9_17.
21. P. Marcillo, Á. L. Valdivieso Caraguay, and M. Hernández-álvarez, "A Systematic Literature Review of Learning-Based Traffic Accident Prediction Models Based on Heterogeneous Sources," Appl. Sci., vol. 12, no. 9, 2022, doi: 10.3390/app12094529.
22. A. M. Anter, A. T. Azar, and K. M. Fouad, "Intelligent Hybrid Approach for Feature Selection BT - The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)," 2020, pp. 71–79.
23. A. Géron, Mãos à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow: Conceitos, Ferramentas e Técnicas para a Construção de Sistemas Inteligentes. 2019.
24. S. Raschka, Y. (Hayden). Liu, V. Mirjalili, and D. Dzhulgakov, Machine Learning with Pytorch and Scikit-Learn Develop Machine Learning and Deep Learning Models with Python. 2022.