PRISMA-Guided Systematic Review on Machine Learning for University Student Dropout Prediction

Sari Fauzia Elza[1,a)], Yohana Dewi Lulu Widyasari [1)]

[1]*Magister Terapan Teknik Komputer, Politeknik Caltex Riau, Pekanbaru, Indonesia*

[a)]Corresponding author: sari23mttk@mahasiswa.pcr.ac.id

Abstract. This systematic review examines the application of machine learning techniques to predict students dropout. The prisma 2020 guidelines were followed to ensure a comprehensive and transparent review process. As the behaviour of students who drop out becomes increasingly complex due to factors such as academic performance, personal characteristics and socio-economic conditions, machine learning offers promising solutions for the early identification of students at risk. This review summarises findings from peer-reviewed studies published between 2014 and 2024 and indexed in the scopus database. The focus is on the performance, strengths and limitations of different machine learning models such as decision trees, support vector machines and neural networks. The selection of the 2014-2024 timeframe reflects the significant advances in machine learning technologies, the improved quality and availability of educational data, and the evolving research trends in education. This timeframe also coincides with changes in education policy and ensures that the study captures current and relevant findings. The report concludes with recommendations for future research, including the integration of complex data characteristics and the development of universal models that can be adapted to different student populations.

Keywords: Systematic review, machine learning, prediction, prisma

INTRODUCTION

The dropout rate among university students has become a significant problem affecting many aspects of higher education on a global scale. This phenomenon not only affects the reputation of educational institutions, which tends to decrease when the dropout rate increases, but also has profound consequences for the students themselves, as it affects their academic progress, their emotional well-being and their financial situation. Failure to complete their studies on time often results in reduced future employment opportunities, increased financial burdens from unfinished educational commitments and feelings of frustration and hopelessness. Given the far-reaching implications, it is crucial for educational institutions to continually understand and address the factors that contribute to dropout in higher education

With the rapid advancement of technology, data-driven approaches are increasingly being used to help educational institutions identify students at high risk of dropping out of school. Among the best-known technological approaches with significant potential in this context is machine learning, a tool that enables the processing and analysis of highly complex data to uncover patterns that may not be immediately apparent. By using machine learning, various factors such as students' academic performance, demographic background, social engagement and behavioural patterns can be analysed simultaneously to predict the likelihood of dropping out. This approach not only provides institutions with new insights to monitor their students more effectively, but also creates opportunities to develop timely interventions aimed at preventing dropout before it occurs.

Although machine learning has shown great potential for predicting school dropout, research in this area continues to face several complex challenges. One of the main problems is the different algorithms used in different studies, the variety of data sources analysed and the very different methodological approaches, which often make a direct comparison of research results difficult. For example, some studies may conclude that a particular algorithm works exceptionally well in predicting early school leaving, while other studies may find that the same algorithm is less effective due to differences in the predictor variables considered. Furthermore, the lack of standardised reporting practises in these studies is a significant obstacle to gaining a clear picture of the overall effectiveness of machine learning in predicting dropout in higher education.

In response to these challenges, this study aims to conduct a systematic review guided by PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), a widely recognised framework designed to improve transparency and consistency in the reporting of systematic reviews. Using PRISMA as a guide, this study will review, collate and critically appraise relevant literature on the application of machine learning in the prediction of study attrition. In addition, the study will examine the algorithms used in previous research, identify key predictor variables and identify research gaps that still exist in this area. It is expected that the results of this systematic review will provide a more comprehensive overview of the current research landscape and offer clearer guidelines for the development of more accurate and practical dropout prediction models in the future.

It is expected that universities and educational institutions will gain deeper insights into the implementation of effective data-driven solutions to prevent dropout through this research. In addition, the results of this study can serve as a solid foundation for more targeted, evidence-based research in the future, ultimately aimed at improving the overall quality of education and ensuring that students achieve academic success by completing their studies on time.

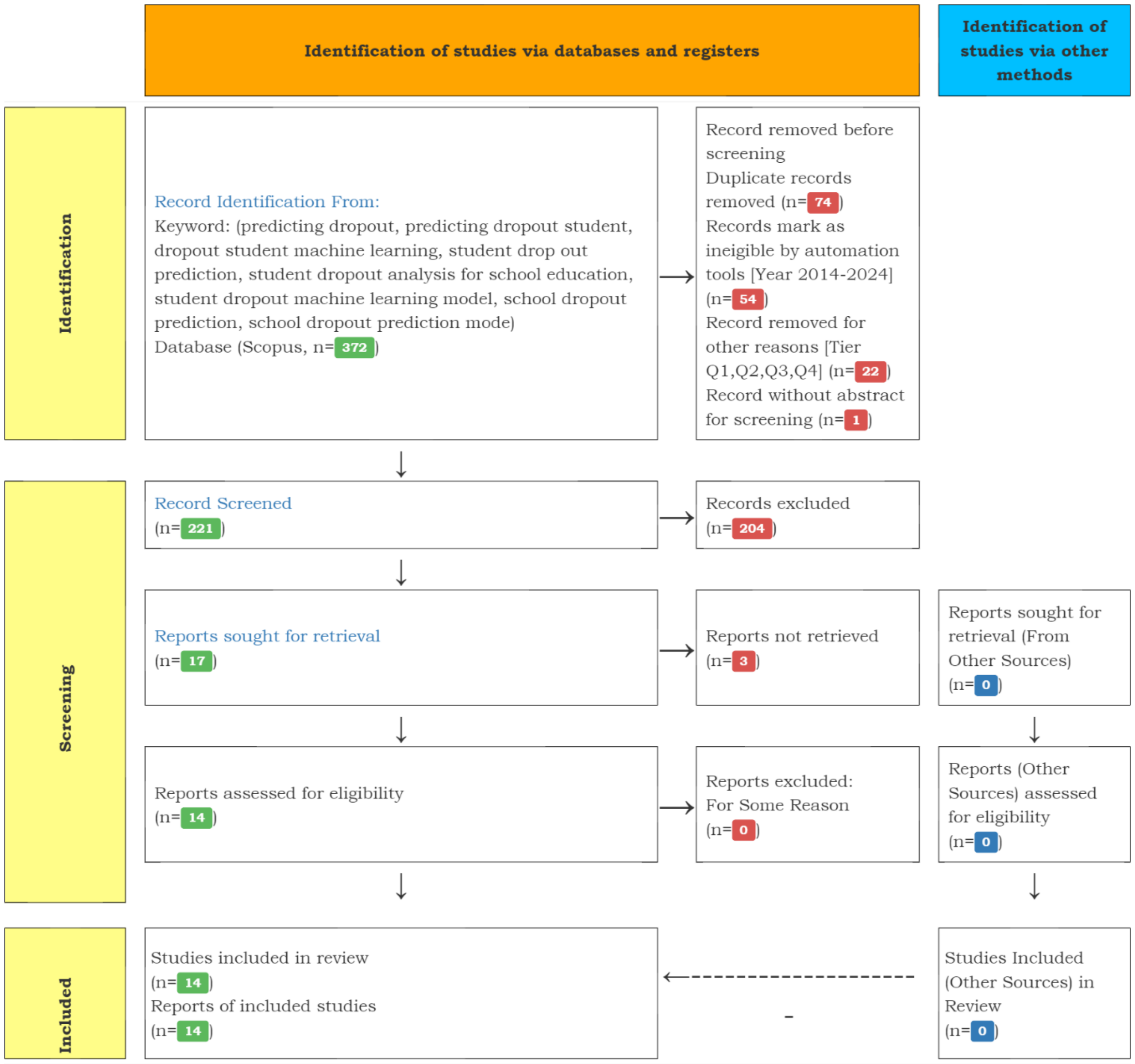
METHODS

A systematic review is a rigorous methodological approach that aims to collect, evaluate and summarise research findings that are relevant to a particular research question or topic of interest. The main aim of this approach is to minimise bias by applying well-defined strategies throughout the review process [1]. Systematic reviews are considered the "gold standard" for consolidating research knowledge and are an important tool for evidence-based research. Emphasise the importance of systematic reviews in capturing existing literature and guiding new research investigations, ensuring that researchers build on a solid foundation of evidence.

In the context of the social sciences, as emphasised in Research Synthesis and Meta-Analysis: A Step-by-Step Approach [2], systematic reviews play a crucial role in understanding complex human behaviours and social dynamics. These reviews facilitate the identification of trends and gaps in various disciplines, including sociology, psychology, economics and anthropology

In addition, systematic reviews allow researchers to comprehensively capture the theoretical developments around a particular topic by examining studies from multiple digital libraries. This comprehensive examination contributes to a better understanding of the topics in question [1].

This study uses a systematic review approach based on PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). In this study, PRISMA is used to systematically collate and assess studies on the prediction of dropout using machine learning to ensure that the review process is comprehensive and methodologically sound. To improve the efficiency of the literature search, Watase UEKA technology is used in this SLR. Keywords such as "predicting dropout student"," "student dropout machine learning models"," "machine learning for dropout prediction"," "student dropout"," "higher education retention" and "systematic review in education" are used in the search.



Prisma Reporting: Student

Identification of studies via databases and registers
Identification of studies via other methods
Identification
Record Identification From:
Keyword: (predicting dropout, predicting dropout student, dropout student machine learning, student drop out prediction, student dropout analysis for school education, student dropout machine learning model, school dropout prediction, school dropout prediction mode)
Database (Scopus, n=372)
Record removed before screening
Duplicate records removed (n=74)
Records mark as ineigible by automation tools [Year 2014-2024] (n=54)
Record removed for other reasons [Tier Q1,Q2,Q3,Q4] (n=22)
Record without abstract for screening (n=1)
Screening
Record Screened (n=221)
Records excluded (n=204)
Reports sought for retrieval (n=17)
Reports not retrieved (n=3)
Reports sought for retrieval (From Other Sources) (n=0)
Reports assessed for eligibility (n=14)
Reports excluded: For Some Reason (n=0)
Reports (Other Sources) assessed for eligibility (n=0)
Included
Studies included in review (n=14)
Reports of included studies (n=14)
-
Studies Included (Other Sources) in Review (n=0)

Generate From Watase Uake Tools, based on Prisma 2020 Reporting

Figure 1. PRISMA Diagram (Preferred Reporting Items for Systematic Reviews and Meta-Analysis)

The search was performed in the Scopus database and initially returned 372 entries. Prior to screening, 74 duplicate records were removed as well as one record that was deemed unsuitable by the automation tools as it did not meet the year criteria (2014–2024). A further 22 records were removed based on classification by level (Q1, Q2, Q3, Q4), leaving 221 records for the first screening phase. During the screening process, 204 records were excluded, reducing the pool to 17 reports that were eligible for the search.

Of the 17 journals originally identified as relevant, 3 journals could not be accessed or retrieved despite being requested. The reasons for not finding these journals can vary, e.g. the possibility that the journals are not available in the databases used, that access is restricted by the institution or that the articles are no longer published. Therefore, after checking the accessibility and completeness of the data, we decided to continue the analysis with the 14 journals that we were able to successfully access and that are relevant to the research topic. This restriction ensures that the studies analysed can be verified and supports the validity of the results obtained

RESULTS AND DISCUSSION

The following are the results of PRISMA (Preferred Reporting Items for Systematic Review and Meta-Analysis). These results were obtained through a systematic and transparent review process that ensures a rigorous assessment of the relevant studies. Established guidelines were followed at every step, which strengthens the credibility and reproducibility of the results presented

Table 1. List of Journals Results Of The Prisma WITH WATASE Method

No.	Topic	Author	Year	Algorithm	Outcome Result
1	A Study on Dropout Prediction for University Students Using Machine Learning	Choong Hee Cho, Yang Woo Yu, Hyeon Gyu Kim	2023	Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Deep Neural Network, LightGBM	LightGBM model provided the best performance with an F1-score of 0.840
2	Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students	Norka Bedregal-Alpaca, Víctor Cornejo-Aparicio, Joshua Zárate-Valderrama, Pedro Yanque-Churo	2020	Artificial neural networks, ID3 algorithm, and C4.5 algorithm	The C4.5 algorithm presented the best performance metrics, with correctness, accuracy, and sensitivity of 0.83, 0.87, and 0.90, respectively. The ratio of credits approved by a student to the credits they should have taken was the most significant variable for predicting dropout, while the number of abandoned subjects was the most significant for classifying academic risk.
3	Comparative analysis of Machine Learning Techniques for the prediction of cases of university dropout	Anthony Edwin Aco Tito, Bryan Orlando Hancco Condori, Yasiel Pérez Vera	2023	Logistic Regression, Naive Bayes, Multilayer Perceptron Neural Network, Decision Tree, Support Vector Machine, and Random Forest.	The study concluded that Logistic Regression is the technique that yields the best results for predicting university dropout in the considered dataset.
4	Early prediction models and crucial factor extraction for first-year undergraduate student dropouts	Thao-Trang Huynh-Cam, Long-Sheng Chen, Tzu-Chuen Lu	2024	Decision trees, multilayer perceptron, and logistic regression	Decision trees outperformed multilayer perceptron and logistic regression with accuracy (97.59%), precision (98%), recall (97%), F1-score (97%), and ROC-AUC (98%). The top-ranking factors were "student loan," "dad occupations," "mom educational level," "department," "mom

					occupations," "admission type," "school fee waiver," and "main sources of living."
5	Early Prediction of University Dropouts – A Random Forest Approach	Andreas Behr, Marco Giese, Herve D. Teguim K. and Katja Theune	2020	Using random forests	sing a Random Forest algorithm, the model achieved an AUC (Area Under the Curve) of 0.86, signifying high accuracy. Key predictors include final secondary school grades, student satisfaction, as well as students' academic self-concept and self-assessment.
6	Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Rol	Marina Segura, Jorge Mello, and Adolfo Hernández	2022	SVM, Decision Trees, ANN, Logistic Regression	Dropout detection does not work only with enrollment variables, but improves after the first semester results. Academic performance is always a relevant variable, but there are others, such as the level of preference that the student had over the course that he or she was finally able to study. The success of the techniques depends on the program areas. Machine Learning obtains the best results, but a simple Logistic Regression model can be used as a reasonable baseline.
7	Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education	Mónica V. Martins, Luís Baptista, Jorge Machado, Valentim Realinho	2023	Random Forest, other machine learning algorithms	Best models use Random Forest with strategies to deal with imbalanced data, best results obtained at the end of the first semester
8	Predicting student dropout : A machine learning approach	Lorenz Kemper, Gerrit Vorhoff, Berthold U. Wigger	2020	Logistic regression and decision tree	Both logistic regression and decision tree models yielded high prediction accuracies of up to 95% after three semesters, with decision trees performing slightly better. A classification with more than 83% accuracy was already possible after the first semester.
9	Analysis of First-Year University Student	Diego Opazo, Sebastián	2023	KKN,SVM,Decision Tree,Random	Among the models tested, the Gradient Boosting

	Dropout through Machine Learning Models: A Comparison between Universities	Moreno, Eduardo Álvarez-Miranda, Jordi Pereira		Forest,Gradient Boosting, Decision Tree,Naive Bayes,Logistic Regression,Neural Network	Decision Tree had the highest average score across the combined dataset (students from both universities) and the Universidad Adolfo Ibáñez (UAI) dataset. This model also performed well on the Universidad de Talca (U Talca) dataset and the U Talca All dataset (which included additional non-shared variables).
10	Prediction of Student Dropout in a Chilean Public University through Classification based on Decision Trees with Optimized Parameters	Patricio E. Ramírez and Elizabeth E. Grandón	2018	Classification based on decision trees with parameter optimization	The optimized CBAD technique achieved a precision rate of 87.27% in predicting student dropout
11	Student Dropout Prediction for University with High Precision and Recall	Sangyun Kim, Euteum Choi, Yong-Kee Jun, Seongjin Lee	2023	Hybrid model using machine learning techniques	Achieved a precision of 0.963, recall of 0.766, and F1-score of 0.808, outperforming existing models
12	Towards Predicting Student's Dropout in University Courses	Janka Kabathova and Martin Drlik	2021	Decision tree, machine learning classifiers	Prediction accuracy varied between 77 and 93%
13	A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study	Roderick Lottering, Robert Hans, Manoj Lall	2020	Machine learning classification algorithms	The overall accuracy rate of Random Forest (94.14%) was better than the other algorithms in identifying students at risk of dropout.
14	A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data	Antonio Jesús Fernández-García, Juan Carlos Preciado, Fran Melchor, Roberto Rodriguez-Echeverria, José María Conejero, Fernando Sánchez-Figueroa	2021	Gradient Boosting, Random Forest, Support Vector Machine	Accurate and reliable predictive models to serve as a decision-making system for effective dropout prevention policies

The first study, titled "A Study on Dropout Prediction for University Students Using Machine Learning", was conducted by Choong Hee Cho, Yang Woo Yu, and Hyeon Gyu Kim in 2023. The researchers explored multiple machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Deep Neural Network, and LightGBM to predict student dropouts. The LightGBM model stood out as the best performer, achieving an F1-score of 0.840, making it the most effective in identifying students at risk of dropping out [3].

The second study, "Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students", was authored by Norka Bedregal-Alpaca, Víctor Cornejo-Aparicio, Joshua Zárate-Valderrama, and Pedro Yanque-Churo in 2020. This study utilized Artificial Neural Networks, the ID3 algorithm, and the C4.5 algorithm. Among these, the C4.5 algorithm delivered the best results, with correctness, accuracy, and sensitivity values of 0.83, 0.87, and 0.90, respectively. The ratio of approved credits compared to the credits that should have been taken was the most significant variable for predicting dropout, while the number of abandoned subjects was crucial for classifying academic risk [4].

The third study, titled "Comparative Analysis of Machine Learning Techniques for the Prediction of Cases of University Dropout", was conducted by Anthony Edwin Aco Tito, Bryan Orlando Hancco Condori, and Yasiel Pérez Vera in 2023. It applied a variety of algorithms, including Logistic Regression, Naive Bayes, Multilayer Perceptron Neural Network, Decision Tree, Support Vector Machine, and Random Forest. The study found that Logistic Regression produced the best predictive results for university dropout cases in the dataset they analyzed [5].

In the fourth study, "Early Prediction Models and Crucial Factor Extraction for First-Year Undergraduate Student Dropouts", authored by Thao-Trang Huynh-Cam, Long-Sheng Chen, and Tzu-Chuen Lu in 2024, the researchers employed Decision Trees, Multilayer Perceptron, and Logistic Regression. Decision Trees outperformed the other models, achieving an accuracy of 97.59%, along with a precision of 98%, recall of 97%, F1-score of 97%, and ROC-AUC of 98%. Key factors identified included "student loan," "parental occupation and education," and "school fee waivers." [6]

The fifth study, "Early Prediction of University Dropouts – A Random Forest Approach", conducted by Andreas Behr, Marco Giese, Herve D. Teguim K., and Katja Theune in 2020, applied the Random Forest algorithm to predict university dropouts. This approach achieved a strong AUC score of 0.86, indicating high predictive accuracy. The most influential predictors included students' final grades in secondary school, their satisfaction with their studies, and their academic self-concept and self-assessment [7].

The sixth study, "Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?", was authored by Marina Segura, Jorge Mello, and Adolfo Hernández in 2022. This research used algorithms such as SVM, Decision Trees, Artificial Neural Networks, and Logistic Regression. It revealed that dropout predictions improved significantly after the first semester when students' academic performance and their preference for their chosen courses were considered. Although machine learning provided the best results, a simple Logistic Regression model was also a reliable baseline [8].

In the seventh study, "Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education", conducted by Mónica V. Martins, Luís Baptista, Jorge Machado, and Valentim Realinho in 2023, the researchers focused on using Random Forest and other machine learning algorithms. Random Forest models that accounted for imbalanced data gave the best results, particularly at the end of the first semester, when predicting dropouts and academic performance [9].

The eighth study, titled "Predicting Student Dropout: A Machine Learning Approach", was carried out by Lorenz Kemper, Gerrit Vorhoff, and Berthold U. Wigger in 2020. The researchers employed Logistic Regression and Decision Tree models. Both models achieved high prediction accuracy, reaching up to 95% after three semesters. Decision Trees performed slightly better, and classification with more than 83% accuracy was possible by the end of the first semester [10].

The ninth study, "Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities", was authored by Diego Opazo, Sebastián Moreno, Eduardo Álvarez-Miranda, and Jordi Pereira in 2023. This study utilized K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, Random Forest, Gradient Boosting Decision Tree, Naive Bayes, Logistic Regression, and Neural Network algorithms. Among these, the Gradient Boosting Decision Tree delivered the best results, achieving the highest average score across both the combined dataset (students from both universities) and the Universidad Adolfo Ibáñez (UAI) dataset. Additionally, this model performed effectively on the Universidad de Talca (U Talca) dataset, as well as the U Talca All dataset, which incorporated additional non-shared variables. [11].

The tenth study, "Prediction of Student Dropout in a Chilean Public University through Classification Based on Decision Trees with Optimized Parameters", conducted by Patricio E. Ramírez and Elizabeth E. Grandón in 2018, focused on decision trees with optimized parameters. Their approach, referred to as Classification Based on Decision Trees (CBAD), achieved a precision rate of 87.27% in predicting university dropout, demonstrating its effectiveness in Chilean public university settings [12].

The eleventh study, "Student Dropout Prediction for University with High Precision and Recall", authored by Sangyun Kim, Euteum Choi, Yong-Kee Jun, and Seongjin Lee in 2023, developed a hybrid model that combined multiple machine learning techniques. The model outperformed existing approaches, achieving a high precision of 0.963, recall of 0.766, and F1-score of 0.808, showing significant improvements in predicting dropouts [13].

The twelfth study, "Towards Predicting Student's Dropout in University Courses", conducted by Janka Kabathova and Martin Drlik in 2021, applied decision trees and other machine learning classifiers. Their results showed prediction accuracy varying between 77% and 93%, depending on the specific dataset and configuration of the models, highlighting the effectiveness of decision tree-based approaches in university dropout prediction [14].

The thirteenth study, "A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study", conducted by Roderick Lottering, Robert Hans, and Manoj Lall in 2020, employed a range of machine learning classification algorithms. The Random Forest model performed the best, with an accuracy rate of 94.14% in identifying students who were at risk of dropping out from their courses [15].

The fourteenth study, "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data", was authored by Antonio Jesús Fernández-García, Juan Carlos Preciado, Fran Melchor, Roberto Rodriguez-Echeverria, José María Conejero, and Fernando Sánchez-Figueroa in 2021. This study used Gradient Boosting, Random Forest, and Support Vector Machine models to develop accurate and reliable predictive models. These models were intended to assist universities in implementing effective dropout prevention policies based on data collected at different stages of students' academic journeys [16].

CONCLUSIONS

This study successfully conducted a systematic review using the PRISMA guidelines to analyze the use of machine learning to predict dropout. By examining peer-reviewed literature published between 2014 and 2024, the study highlights the strengths and limitations of various algorithms, including decision trees, support vector machines and neural networks. The report confirms that machine learning can provide valuable insights for predicting early dropout by identifying students at risk based on a combination of academic performance, personal characteristics and socio-economic factors. However, it is also emphasized that the diversity of data sources, algorithms and reporting standards across studies poses a challenge to a consistent understanding of the most effective predictive models. Future research should focus on integrating more complex data features and developing universally adaptable models for different student populations. This would increase the accuracy of dropout predictions and support the development of effective intervention strategies by educational institutions.

REFERENCES

[1] B. Kitchenham, Procedures for Performing Systematic Review, 2004.

[2] H. Cooper, Research Synthesis and Meta-Analysis, 2017.

[3] C. H. Cho, Y. W. Yu and H. G. Kim, "A Study on Dropout Prediction for University Students Using Machine Learning," *MDPI,* 2023.

[4] N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama and P. Yanque-Churo, "Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students," *(IJACSA) International Journal of Advanced Computer Science and Applications,*, 2020.

[5] A. E. A. Tito, B. O. H. Condori and Y. P. Vera, "Comparative analysis of Machine Learning Techniques for the prediction of cases of university dropout," *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação,* 2023.

[6] T.-T. Huynh-Cam and L.-S. C. a. Tzu-Chuen, "Early prediction models and crucial factor extraction for first-year undergraduate student dropouts," *Journal of Applied Research Emerald Publishing,* 2024.

[7] A. Behr, M. Giese, H. D, T. K and K. Theune, "Early Prediction of University Dropouts – A Random Forest Approach," *Journal of Economics and Statistics,* 2020.

[8] M. Segura, J. Mello and A. Hernández, "Machine Learning Prediction of University Student Dropout:Does Preference Play a Key Role?," *MDPI Mathematics,* 2022.

[9] M. V. Martins, L. Baptista, J. Machado and V. Realinho, "Multi-Class Phased Prediction of Academic Performance and Dropout in Higher Education," *MDPI,* 2023.

[10] L. Kemper, G. Vorhoff and B. U. Wigger, "Predicting student dropout: A machine learning approach," *European Journal of Higher Education,* 202.

[11] D. Opazo, S. Moreno, E. Álvarez-Miranda and J. Pereira, "Analysis of First-Year University Student Dropout through Machine Learning Models: A Comparison between Universities," *MDPI,* 2021.

[12] P. E. Ramírez and y. E. E. Grandón, "Prediction of Student Dropout in a Chilean Public University through Classification based on Decision Trees with Optimized Parameters," *Formación Universitaria,* p. 2018.

[13] S. Kim, E. Choi, Y.-K. Jun and S. Lee, "Student Dropout Prediction for University with High Precision and Recall," *MDPI,* 2023.

[14] J. Kabathova and M. Drlik, "Towards Predicting Student's Dropout in University Courses Using Different Machine Learning Techniques," *MDPI,* 2021.

[15] R. Lottering, R. Hans and M. Lall, "A Machine Learning Approach to Identifying Students at Risk of Dropout: A Case Study," *(IJACSA) International Journal of Advanced Computer Science and Applications,* 2020.

[16] J. C. P. Antonio Jesús Fernández-García, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero and F. Sánchez-Figueroa, "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data," *IEEEAcces,* 2021.