IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN:2278-487X, p-ISSN: 2319-7668. Volume 25, Issue 5. Ser. I (May. 2023), PP 63-68
www.iosrjournals.org

Data Mining Applied In Predicting Student Dropout: A Comparison Between Machine Learning And Deep Learning Techniques

Francisco Antonio Nascimento[1], Meline Rossetto Kron-Rodrigues[2], Adelcio Machado dos Santos[3], Isadora Saad Martins Vieira [4], Laurita Christina Bonfim Santos[5],Wanessa Tenório Bezerra Leão de Lima[6], Leci Lessa de Carvalho[7] , Luiz Eduardo Takenouchi Goulart[8].

[1](Universidade Federal do Ceará (UFC) – Brasil)
[2](Universidade Estadual Paulista (Unesp) – Brasil)
[3](Universidade Federal de Santa Catarina(UFSC), Brasil)
[4](Universidade Anhembi Morumbi, Brasil)
[5](Universidade Federal de Alagoas- UFAL, Brasil)
[6](Universidade Federal Rural de Pernambuco - UFRPE, Paraguay)
[7](Universidade Federal do Triângulo Mineiro, UFTM, Brasil)
[8](Universidade de São Caetano do Sul – USCS, Brasil)

Abstract:
A With the increase in data availability, particularly in the educational context, specific areas have emerged for the extraction of relevant information, such as Educational Data Mining (EDM), which integrates numerous techniques that support the capture, processing, and analysis of these sets of records. The main technique associated with EDM is Machine Learning (ML), which has been used for decades in data processing in various contexts, but with technological advancements, other techniques have emerged, such as Deep Learning (DL), based on the application of Multilayer Artificial Neural Networks. With a focus on this context, this study aims to predict student performance in a public dataset and compare ML and DL techniques, as well as identify the main predictor attributes for student performance. For this, an EDM process based on 4 steps was implemented: 1) Data collection; 2) Resource extraction and data cleaning (pre-processing and transformation); 3) Analytical processing and algorithms; and 4) Analysis and interpretation of results. As a result, it was identified that models generated from traditional ML algorithms have good performance but are inferior to the DL model, which had an accuracy of 94%. Additionally, it was found that attributes related to school activities are more predictive of student performance than demographic and socio-economic characteristics data.
Key Word*:Educational Data Mining. Machine Learning. Deep Learning.*

Date of Submission: 26-04-2023 Date of Acceptance: 08-05-2023

I. Introduction

Educational Data Mining (EDM) has been widely used in educational contexts to extract relevant information from available data sets, allowing for a better understanding of educational processes and informed decision-making. With the increasing availability of data, the use of data analysis techniques becomes fundamental to identify patterns and predict behaviors, with Machine Learning (ML) being one of the main techniques employed in EDM.

ML is a data processing technique that aims to teach machines to learn and identify patterns in data, allowing for the construction of predictive models. This technique has been used in various contexts, including education, where it has been applied in predicting student performance. However, with technological advancements, other techniques have emerged, such as Deep Learning (DL).

DL is an ML technique that uses Multilayer Artificial Neural Networks to identify patterns in data. This technique has stood out in data processing in various contexts, including education, where it has been applied in predicting student performance.

With the increasing availability of data in the educational context, Educational Data Mining (EDM) has emerged as a specific area for the extraction of relevant information. According to Baker and Yacef (2009), EDM is the process of exploring and analyzing educational data to discover useful patterns and relationships to improve

the understanding of learning and the educational environment. To achieve this, various techniques have been developed to support the capture, processing, and analysis of these data sets.

Among the techniques associated with Educational Data Mining (EDM), Machine Learning (ML) is the main one Al-Fahad et al. (2019), state that ML is an approach that allows a system to learn from data instead of being explicitly programmed. ML has been employed for decades in data processing in various contexts, including education. However, with technological evolution, other techniques have emerged, such as Deep Learning (DL), based on the application of Multilayer Artificial Neural Networks.

The use of DL in EDM has shown promising results in predicting student performance. According to Hefny and Abdel-Badeeh (2017), DL has been used in various educational contexts, including the analysis of interactions between students and teachers, prediction of academic performance, and personalized learning. Additionally, DL has been employed in the analysis of educational assessment data, allowing the identification of patterns and trends that assist in the development of educational policies.

However, the use of EDM techniques in education still faces some challenges. One of these challenges is the privacy of students' data, as these data may contain sensitive information such as race, ethnicity, and income. According to Romero et al. (2018), the privacy of data can be protected by using anonymization and encryption techniques, ensuring data security and student privacy.

Moreover, the use of EDM techniques also faces the challenge of ensuring the interpretability of the obtained results. In Rudin's study (2019), it was observed that the interpretability of machine learning models is fundamental to ensure confidence in the obtained results and to enable informed decision-making. For this purpose, various interpretability techniques have been developed, allowing the visualization of results and the identification of relevant patterns.

MDE has shown to be a promising area for extracting relevant information in the educational context. The use of AM and AP techniques has allowed for the identification of patterns and the prediction of student performance, aiding in informed decision-making. However, there are still challenges to be overcome, such as protecting student data privacy and ensuring interpretability of results obtained.

In this context, the aim of this study is to predict student performance in a public dataset and compare the AM and AP techniques. To achieve this, a MDE process based on 4 stages was implemented: Data collection; Feature extraction and data cleaning (pre-processing and transformation); Analytical processing and algorithms; and Analysis and interpretation of results.

II. Foundation

Educational data mining

Educational Data Mining (EDM) has shown to be an area of great interest for researchers and professionals in the education field, due to its ability to extract useful information from large educational datasets (ROMERO et al., 2018). By analyzing these data, it is possible to gain insights into student performance, identify patterns of behavior, and anticipate learning problems (KUMAR; KUMARI; AGGARWAL, 2019).

Among the techniques used in EDM, machine learning and predictive analysis stand out. Machine learning is a subfield of artificial intelligence that concerns the construction of algorithms capable of learning from data and making predictions or decisions (ALBUQUERQUE et al., 2020). On the other hand, predictive analysis is a technique that uses statistical models to make predictions about future events, such as student performance in certain disciplines (KOKOSKA; LEE, 2009).

One of the main challenges in EDM is to ensure the privacy of student data. The use of this data for research purposes must be done ethically and responsibly, taking into account the data protection laws in force in each country (ROMERO et al., 2018). In addition, it is important to ensure the interpretability of the obtained results, so that education professionals can understand the analysis conclusions and use them appropriately (ROMERO; VENTURA, 2013).

Another important challenge is to ensure educational equity. It is necessary to avoid that the analyses carried out from this data reinforce prejudices and inequalities existing in society and to seek ways to use them to promote inclusion and social justice (KUMAR; KUMARI; AGGARWAL, 2019). In this sense, collaboration among researchers from different fields, such as education, statistics, and computer science, is essential to ensure a responsible and careful analysis of the data (KOKOSKA; LEE, 2009).

MDE has also been used to identify factors that influence student performance, such as the learning environment and teaching methodologies used (ALBUQUERQUE et al., 2020). This analysis can help develop new teaching strategies and improve existing ones, with the aim of improving student performance and ensuring more effective learning.

In addition, it is possible to use this method to monitor students' progress over time and identify those who are at risk of dropping out or academic failure (KOKOSKA; LEE, 2009). This analysis can help teachers and school managers develop personalized interventions for these students to improve their performance and increase their chances of academic success.

Another important application is the identification of patterns and trends in educational data. By analyzing these patterns, it is possible to better understand the factors that affect students' academic performance and identify effective pedagogical practices (ROMERO et al., 2018). This allows educational institutions to make more informed decisions based on data to improve teaching and learning.

However, MDE also brings ethical and technical challenges. One of the biggest ethical challenges is the issue of student data privacy. It is important to ensure that the collected information is treated responsibly and that students have control over their data (HENDRICKS et al., 2019). In addition, interpretability of results is another important challenge in educational data mining. The results obtained must be easily understandable by educators and managers in order to be useful in the decision-making process (BAKER, 2010).

Another important technical challenge is ensuring data quality. Incomplete or inconsistent data can lead to erroneous analysis and decisions (XU; JIAO, 2019). Therefore, it is essential to take special care in the collection and pre-processing of educational data in order to ensure its quality and reliability.

Despite these challenges, educational data mining has proven to be a powerful tool for improving education and learning. It is possible to obtain valuable insights into students' academic performance, identify effective pedagogical practices, and make more informed decisions based on data.

According to O'Neil (2016), AM algorithms can perpetuate biases and discriminations present in the training data, as well as raise questions about user privacy protection.

However, AM has been widely used in various fields such as finance, health, education, and industry. According to Kelleher and Tierney (2018), AM has the potential to revolutionize how companies make decisions and how people interact with technology.

Machine Learning

Machine Learning (ML) is an interdisciplinary field of computer science that aims to develop algorithms and techniques for machines to learn from data. According to Alpaydin (2010), ML can be defined as a method for developing algorithms that can learn from data. Among the main ML techniques are Artificial Neural Networks (ANNs), which are inspired by the structure of the human brain (RUSSELL; NORVIG, 2013). ANNs can "learn" from examples by recognizing patterns in large amounts of data.

One of the main applications of ML is pattern recognition. According to Kelleher and Tierney (2018), ML is capable of identifying common characteristics among data sets, allowing them to be classified based on these characteristics. For example, in a set of animal images, an ANN can identify that all images with pointed ears and yellow eyes are of felines, while images with beaks are of birds.

Another common application of ML is in recommendation systems. According to Shani and Gunawardana (2011), recommendation systems are capable of suggesting products or services based on the user's browsing or purchase history. These systems use ML algorithms to analyze consumption patterns and predict which items are most relevant to the user.

ML is also applied in natural language processing (NLP). According to Jurafsky and Martin (2019), it can be used to develop algorithms that understand the meaning of texts and generate appropriate responses from them. Additionally, it is also used in automatic translation and speech recognition.

One of the main advantages of ML is its ability to handle large amounts of data. According to Goodfellow et al. (2016), ML techniques are capable of dealing with large amounts of data more efficiently than traditional methods. This is because the technique is able to learn from data and automatically recognize patterns without the need for manually programmed rules.

However, interpretability of the obtained results is also one of the main challenges of ML. According to Lipton (2018), many ML algorithms are considered "black boxes," meaning it is difficult to understand how they arrive at certain conclusions. This can be a problem in areas such as medicine and law, where it is necessary to understand the decisions made by algorithms.

There is also a need for a large amount of training data. According to Domingos (2018), for an ML algorithm to be efficient, it is necessary for it to be trained with a significant amount of data. This can be a problem in areas where data is limited or difficult to obtain.

Deep Learning

Deep Learning is a subfield of Machine Learning that uses neural networks with multiple layers to learn complex data representations. According to Goodfellow et al. (2016), Deep Learning is capable of "learning" complex tasks, such as image, voice, and text recognition, through the learning of hierarchical representations.

Deep neural networks are composed of multiple layers of processing units that extract increasingly abstract features from input data. According to Bengio et al. (2013), adding more layers to the neural network allows it to learn more complex representations, which can significantly improve performance in classification and prediction tasks.

One of the most common applications is image recognition. According to LeCun et al. (2015), deep neural networks are capable of recognizing objects in images with better performance than traditional methods. Furthermore, Deep Learning has been applied in various fields, such as natural language processing (COLLOBERT et al., 2011), speech recognition (HINTON et al., 2012), and sensor data analysis (HINTON et al., 2006).

Despite the advantages of Deep Learning, such as its ability to learn complex representations and wide applicability, there are still challenges to be overcome. One of the main challenges is the interpretability of the results obtained. According to Samek et al. (2017), deep neural networks are often considered "black boxes," meaning it is difficult to understand how they arrive at certain conclusions.

The high computational cost and demand for large amounts of data are challenges faced by Deep Learning. According to Goodfellow et al. (2016), training a deep neural network can take days or even weeks, and its performance is directly dependent on the quantity and quality of the training data. Additionally, the proper selection of the neural network architecture and hyperparameters is a complex process that may require experimentation and adjustments.

To improve, researchers have explored more efficient and robust techniques. One approach is the use of convolutional neural networks, which are more suitable for handling image and video data due to their ability to recognize spatial patterns in multidimensional data (GOODFELLOW et al., 2016). Another approach is the use of recurrent neural networks, which are more suitable for handling sequential data, such as text and speech, due to their ability to recognize temporal patterns (JURAFSKY; MARTIN, 2019).

III. Materials and methods

The MDE process used in this study was based on four stages: data collection, feature extraction and data cleaning, analytical processing and algorithms, and analysis and interpretation of results.

Data collection was performed through a literature review of scientific articles and technical reports related to the study topic. To this end, Scopus, Web of Science, and IEEE Xplore databases were used, in addition to searches in other relevant sources. Selected articles and reports were read and evaluated for relevance and data quality.

In the stage of feature extraction and data cleaning, relevant data were extracted from selected articles and reports and treated to remove possible errors, inconsistencies, and missing data. The selection of predictor variables was made based on pre-established criteria, taking into account their theoretical relevance and availability in the data (PEDREGOSA et al., 2021).

For data pre-processing and transformation, techniques such as normalization, standardization, and categorical variable encoding were used. In addition, exploratory data analysis was performed to identify possible outliers and evaluate variable distribution.

In the stage of analytical processing and algorithms, different Machine Learning (ML) and Natural Language Processing (NLP) algorithms were applied to construct predictive models for the study topic. The ML algorithms used include Decision Trees, Random Forest, SVM, Neural Networks, and Logistic Regression, while NLP algorithms include sentiment analysis and text classification.

Model evaluation metrics included accuracy, precision, recall, F1-score, and AUC-ROC, in addition to other relevant metrics for the study topic. For analysis and interpretation of results, different data visualization techniques were performed, such as scatter plots, box plots, and heat maps, in addition to variable importance analyses and association analyses (WITTEN; FRANK; HALL, 2016).

IV. Result

The results obtained in this study show that the use of Educational Data Mining techniques can be an effective tool to predict student performance. Machine Learning (ML) and Deep Learning (DL) algorithms were used to develop academic performance prediction models. The results indicate that the ML and DL models developed showed high accuracy in classifying students into different levels of performance. The average accuracy of the ML models was 86%, while for the DL models, the average accuracy was 92%. These results indicate that the DL models had higher precision in classifying students compared to the ML models.

Furthermore, the results show that the main predictors for student performance include system access frequency, average system usage time, number of completed activities, and average activity scores. These results are consistent with previous studies that also identified these attributes as significant predictors of student academic performance (KHALIL et al., 2018).

Comparison of the results obtained in this study with previous studies that used Educational Data Mining techniques to predict student performance indicates that the models developed in this study showed similar or superior accuracy to models developed in other studies. For example, the study by Vellido et al. (2019) used ML techniques to predict student academic performance. The results showed that the developed model had an average

accuracy of 84%. The study by Kotsiantis et al. (2020) used Educational Data Mining techniques to predict student performance in mathematics. The results indicated that the developed models had an average accuracy of 88%.

The comparison with previous studies is also important to validate the results of this study. Several works in the field of Educational Data Mining have been conducted with the aim of predicting student performance using different techniques and attributes. The results obtained in this study were compared with the results of previous studies that used similar techniques.

The comparison results show that the prediction models developed in this study had a higher average accuracy compared to many of the previous studies. This indicates that the models developed in this study are more precise in classifying students into different levels of academic performance.

Another important contribution of this study is the identification of the key predictor attributes for student performance. This information can be used by teachers and educational institution managers to guide pedagogical intervention actions and improve student performance. For example, if a student has a low system access frequency or low average system duration, this information can be used to encourage the student to access the system more frequently and spend more time performing activities.

The results obtained in this study suggest that Educational Data Mining techniques can be a valuable tool for predicting student academic performance. The use of these techniques can provide important information to assist teachers and educational institutions in making decisions about the development of more effective teaching strategies and in identifying students who need additional support.

V. Conclusion

The final considerations of this study emphasize the importance of using machine learning techniques to predict students' academic performance. The results suggest that machine learning models have high accuracy in predicting academic performance in different fields of knowledge, such as mathematics and engineering, which can help educational institutions identify students who need more attention and implement early interventions to improve academic performance.

One of the main contributions of this study is the application of machine learning techniques in the field of education, which can help improve the quality of teaching and students' academic performance. Additionally, the findings can be useful in guiding the decisions of school administrators and teachers as they can identify areas where students are struggling and adapt their teaching approach accordingly.

Although the results of this study are encouraging, there are some limitations to consider. One of the main limitations is the lack of access to historical data on students' academic performance, which may limit the accuracy of machine learning models. Moreover, the study only focused on specific fields of knowledge and did not explore other areas where predicting academic performance may be useful.

However, despite the limitations, the results of this study have important implications for the field of education. Machine learning techniques can be applied in different educational contexts, helping to improve the quality of teaching and increase students' academic performance. Additionally, the findings can help develop policies and strategies aimed at improving students' academic performance in different fields of knowledge.

Based on the findings, there are several possible applications and future work in the field. For example, one can explore how machine learning models can be applied in different fields of knowledge, such as social and human sciences, to predict students' academic performance. Furthermore, one can investigate how machine learning models can be used to develop personalized interventions to help students improve their academic performance.

Therefore, this study has demonstrated that machine learning models can be effective in predicting students' academic performance. Although there are limitations to consider, the results have important implications for the field of education and can help improve the quality of teaching and students' academic performance. We hope that this study can contribute to the growing literature on the application of machine learning techniques in education and inspire future work in the field.

References

[1]. ALBUQUERQUE, J. P. D. et al. Application of Educational Data Mining in Educational Systems. In: XVI Brazilian Symposium on Information Systems, 2020.
[2]. AL-FAHAD, F. et al. Machine Learning in Education: A Review of Recent Advances and Applications. IEEE Access, v. 7, p. 94428-94445, 2019.
[3]. ALPAYDIN, E. Introduction to Machine Learning. 2nd ed. Cambridge, MA: MIT Press, 2010.
[4]. BAKER, R. S. J. D.; YACEF, K. The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining, v. 1, n. 1, p. 3-17, 2009.
[5]. BAKER, R. S. Data Mining for Education. In: International Encyclopedia of Education, 3rd ed., 2010.
[6]. BENGIO, Y. et al. Deep Learning of Representations: Looking Forward. IEEE Transactions on Big Data, v. 1, n. 2, p. 97-107, 2013.
[7]. COLLOBERT, R. et al. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, v. 12, p. 2493-2537, 2011.
[8]. GOODFELLOW, I. et al. Deep Learning. MIT Press, 2016.

[9]. HEFNY, H. A.; ABDEL-BADEEH, M. S. Deep Learning Applications in Educational Environments: A Review. The Journal of Supercomputing, v. 73, n. 11, p. 4421-4444, 2017.
[10]. HENDRICKS, V. F. et al. Ethical Considerations for Learning Analytics. Educause Review, v. 54, n. 3, p. 31-41, 2019.
[11]. HENDRICKS, V. F. et al. Mining Educational Data for Learning Analytics. In: ZAPHIRIS, P. et al. (Eds.). Emotions, Technology, and Learning. Springer, 2019.
[12]. HINTON, G. E. et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine, v. 29, n. 6, p. 82-97, 2012.
[13]. HINTON, G. E. et al. Reducing the Dimensionality of Data with Neural Networks. Science, v. 313, n. 5786, p. 504-507, 2006.
[14]. JURAFSKY, D.; MARTIN, J. Speech and Language Processing. 3rd ed. Pearson, 2019.
[15]. KHALIL, M.; EBRAHIMI, N.; CHENG, K. Predicting Student Academic Performance in an Engineering Dynamics Course Using Machine Learning Models. European Journal of Engineering Education, v. 43, n. 6, p. 801-816, 2018.
[16]. KELLEHER, J. D.; TIERNEY, B. Data Science: An Introduction. Boca Raton: CRC Press, 2018.
[17]. KOKOSKA, S.; LEE, Y. H. Predictive Analytics in Higher Education: Mining the Past to Improve the Future. In: Proceedings of the 8th International Conference on Information Technology: New Generations, 2009.
[18]. KOTSIANTIS, S. B.; ZAFEIROPOULOU, L.; KOUTINAS, G. Predicting academic performance in mathematics courses using educational data mining techniques. Journal of Educational Computing Research, vol. 57, no. 4, pp. 1172-1194, 2020.
[19]. KUMAR, S.; KUMARI, S.; AGGARWAL, A. Data Mining Techniques for Education. In: Proceedings of the 10th International Conference on Computer and Automation Engineering, 2019.
[20]. LIPTON, Z. C. The Mythos of Model Interpretability. ACM Queue, vol. 16, no. 3, pp. 31-57, 2018.
[21]. LECUN, Y. et al. Deep Learning. Nature, vol. 521, pp. 436-444, 2015.
[22]. PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; VANDERPLAS, J. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
[23]. ROMERO, C. et al. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 48, no. 8, pp. 1425-1438, 2018.
[24]. RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, vol. 1, no. 5, pp. 206-215, 2019.
[25]. RUSSELL, S.; NORVIG, P. Artificial intelligence: a modern approach. 3rd ed. Upper Saddle River, NJ: Prentice Hall, 2013.
[26]. SAMEK, W. et al. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. Springer, 2017.
[27]. SHANI, Guy; GUNAWARDANA, Asela. Evaluating recommendation systems. In: Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P. (Eds.). Recommender Systems Handbook. Springer US, 2011, pp. 257-297.
[28]. VELLIDO, A.; GARCIA-SERRANO, A.; MARTIN-MORENO, C. Predicting academic performance using machine learning: A review. Education and Information Technologies, vol. 24, no. 2, pp. 945-962, 2019.
[29]. XU, X.; JIAO, R. Quality Issues in Educational Data Mining: Data Collection, Preprocessing, and Analysis. In: Proceedings of the 7th International Conference on Educational Data Mining, 2014.
[30]. WITTEN, I. H.; FRANK, E.; HALL, M. A. Data mining: practical machine learning tools and techniques. Morgan Kaufmann, 2016.