MULTIDISCIPLINARY SCIENCE JOURNAL

CONFERENCE PAPER
Published Online: December 15, 2023
https://doi.org/10.31893/multiscience.2024ss0103

Mathematics and its Applications in Technology (ICMAT23)

Student success prediction using a novel machine learning approach based on modified SVM

P. Umamaheswari[a] ✉ | M. Vanitha[a] | P. Vimala Devi[a] | J. Glory Thephoral[a] | Badal Rihan Basha[a]

[a]Department of Computer Science and Engineering, SASTRA Deemed University, SRC, Kumbakonam, Tamilnadu.

Abstract Education holds an indispensable place in society. The 2020 coronavirus outbreak, which wrought havoc worldwide, imparted varying ramifications on the educational landscape. Numerous studies underscored a decline in student performance, thereby accentuating the urgency of addressing this concern proactively and discerning the contributory factors. As a cornerstone of societal progress, education is universally championed by governments and nations alike. Recognizing the vital need to monitor students to avert academic derailment, the capacity to predict student performance equips educators to vigilantly track outcomes and make informed decisions that bolster both learning and achievement. The model proposed in this study emerges as a superior classifier, offering enhanced accuracy while concurrently mitigating risks of overfitting and underfitting, courtesy of sophisticated machine learning algorithms. This investigation delineates the primary drivers influencing student success. It undertakes student data-based classification and juxtaposes various classifiers. The efficacy of the proposed methodology was corroborated using metrics like accuracy, recall, and the F1 score, registering commendable values of 84%, 95%, and 82% respectively, outpacing traditional models. This innovative approach promises to be instrumental in forecasting students' scholastic trajectories, thereby empowering stakeholders to execute timely interventions.

Keywords: Student result prediction, Classification, Forecasting, KNN, SVM

1. Introduction

Predicting student performance accurately and early is a significant challenge faced by educational institutions. An early assessment of student performance allows them to identify their strengths and weaknesses, which can subsequently enhance their exam outcomes. Through data gathered from Learning Management Systems (LMS), it's possible to predict a student's performance using machine learning. Data from LMSs and social media can provide insights into student behaviors that correlate with either positive or negative exam results. By analyzing these insights, measures can be implemented to assist students who traditionally underperform, guiding them towards better outcomes. Educational institutions aim to assess student performance, pinpoint their strengths and weaknesses, and provide the necessary guidance for improved test scores. As a result, institutions are increasingly relying on LMSs to streamline operations, monitor student engagement, and manage educational content.

Student retention is a pressing concern for educational institutions worldwide due to financial constraints and limited resources. In countries within the Organization for Economic Co-operation and Development (OECD), the average dropout rate hovers around 45%. To combat this, many higher education institutions are developing and implementing intervention strategies. These strategies are most effective when applied during a student's first year. The early identification of at-risk students and subsequent interventions have been prioritized. Machine learning-based predictive analysis, which has revolutionized commercial decision-making in areas like risk management, finance, and operations, showcases the potential of applying such technology across different industries. As competition intensifies annually among educational institutions, every school, college, or university aims for its students to excel in every exam.

The outcomes that students achieve are vital both for the individuals and the educational institutions they attend. The advancement of e-learning technology has made it easier for educational institutions to equip their students with valuable learning tools. A Learning Management System (LMS) can provide insightful data and pertinent statistics regarding system usage and student behavior. Based on metrics such as the frequency of student engagement with course materials, the LMS can offer insights into a student's level of involvement in the course.

Several studies have explored various methods for predicting student performance. Initially, tools like artificial neural networks, logistic regression, and random forests were employed to assess the risk of student failure. These studies found that the rate of accurate predictions for this issue was relatively low, suggesting that 12.2% of students were at risk of failing (Hoffait

et al., 2017) Later research recommended the use of data mining techniques. It was posited that the overall academic performance of a student in their first year is sufficient to create a robust model. Such a model (Migueis et al 2018) can guide educational institution policies and strategies effectively, and subsequently be utilized to predict future student performance.

Data mining methods can be used to predict a student's performance with (Delen et al., 2010) approximately 80% accuracy. While algorithms such as decision trees, logistic regression (LR), and neural networks are beneficial, the Support Vector Machine (SVM) often outperforms them. However, some argue in favor of decision trees over SVMs due to their transparent model structure. Decision trees clearly illustrate various prediction scenarios and offer specific criteria for each test case. The use of machine learning algorithms like Random Forest, Support Vector Machine, Nearest Neighbor, and Logistic Regression (Yagci et al., 2022) can further enhance prediction accuracy. In the referenced paper, the authors used midterm results as a predictor, achieving accuracy rates between 70% and 75%. They concluded that midterm exam results are pivotal for accurately predicting student performance.

Earlier research explored the use of various analytical techniques to predict student achievement. Among classification methods, Amirah Mohamed Shahiria emphasized the effectiveness of Neural Networks and Decision Trees in forecasting student performance. Due to prior meta-analyses on predicting student performance, there has been increased interest from researchers in this area. Meta-analyses facilitate systematic monitoring of student performance within the educational system. When it comes to classification, Multi-labeled KNN demonstrates the highest time accuracy compared to C4.5, Naïve Bayes, and AODE. This is particularly useful in identifying slow learners, enabling early intervention to support and enhance their learning outcomes (Mayilvaganan et al., 2014). It's also suggested that future research could delve into identifying students' cognitive skills. Classification-based methods, which aim to predict if a student will excel or struggle in a course, showed promising results. However, the regression approach, which focuses on forecasting a student's course grade, yielded even better outcomes.

In future research, optimization techniques can be applied to fine-tune parameters. Moreover, feature selection and weighting might be undertaken, potentially yielding positive results in education (Strecht et al., 2015). The study's findings suggest that student interaction is the primary driver of academic success (Putnik et al., 2016). A student's grades improved when they actively interacted with their peers. Students who had connections with multiple other students, particularly those well-connected themselves, achieved higher final grades. Additionally, when these connections were strong, the academic outcomes were even more favorable. Notably, students with strong average tie strengths exhibited greater diversity in their work performance compared to their counterparts with weaker ties. Categorization models have proven reliable in predicting dropouts before a course reaches its midpoint. However, there's been limited research on dropout rates in compulsory schooling. The few findings available have been primarily the result of statistical analysis, not data mining (Márquez-Vera et al., 2016). When identifying students at risk, it's essential to minimize both false positive (type I) and false negative (type II) errors. (Marbouti et al., 2016) employed feature selection to enhance a model's generalization and improve prediction accuracy. In separate studies, researchers conducted three tests to gauge the significance of exam behavior patterns. They utilized four distinct machine learning models to ascertain which students successfully completed their first academic year and which ones didn't. It was found that exam-taking behaviors significantly boosted the predicted F-measure for the class of students at risk of failing (with an approximate increase of 0.3). Furthermore, the methodology centered on student behavior enabled the identification of essential exam-taking patterns. This insight aids educators in pinpointing students who are at risk, ultimately helping them improve their time management skills and increasing their likelihood of successfully completing the first academic year (Kuzilek et al., 2021).

In their quest to predict and intervene in real-world student success—the ultimate goal of such models—(Gardner et al., 2018) identified several critical methodological gaps. These include the comprehensive filtering of experimental subgroups, ineffective validation of student models, and reliance on experimental results that might not be readily available. They also highlighted potential areas for future research, such as temporal modeling, studies that combine explanatory and predictive student models, research bolstering learning theory, and analyses focused on the long-term success of MOOC learners.A student's prior achievements play a significant role in their academic success. The research underscores that past performance indeed exerts a profound influence on student outcomes. It has also been revealed that the efficacy of neural networks increases with dataset size (Agrawal et al., 2015). (Alyahyan et al., 2020) exhaustively explored all potential decisions and parameters, with the aim of making data mining techniques more accessible to educators, thereby unlocking their full potential in the educational domain. In another study, researchers assessed the predictive capability of the Big Five Questionnaire's domains and facets in the Five-Factor Model. They focused on academic performance and engagement, as well as the mediating role of participation in the correlation between personality and achievement (Serrano et al., 2022). The structure of the rest of this paper is as follows: Section 2 introduces the proposed work architecture and the dataset used for analysis. Section 3 delves into the results and discussion, while Section 4 wraps up with conclusions and suggestions for future enhancements.

2. Materials and Methods

There exists a plethora of machine learning methodologies tailored for predictive analyses. In our research, we juxtaposed traditional methods like K-Nearest Neighbour (K-NN), Logistic Regression, and Artificial Neural Networks against a nuanced SVM approach, leading to the examination of four distinct algorithms to gauge student success. The fundamental hypothesis of our approach hinges on the belief that new data points will naturally align with categories that mirror pre-existing ones. As the foundational dataset is established, K-NN comes into play, categorizing novel data entries based on their resemblance to pre-existing data sets. The K-NN technique allows for the rapid and precise categorization of fresh data, contingent upon the predetermined number, KK, of neighboring data points to be considered. The overarching framework is illustrated in Figure 1. The process commences with the extraction of the student dataset from the database, encompassing diverse attributes like academic institution, gender, extracurricular interests, and family background. Subsequent to the feature scaling phase, which transmutes categorical data points into their numerical counterparts, the data undergoes a visualization process before being segregated for training and testing purposes. The culmination involves a classification process pivoting around the previously outlined attributes.

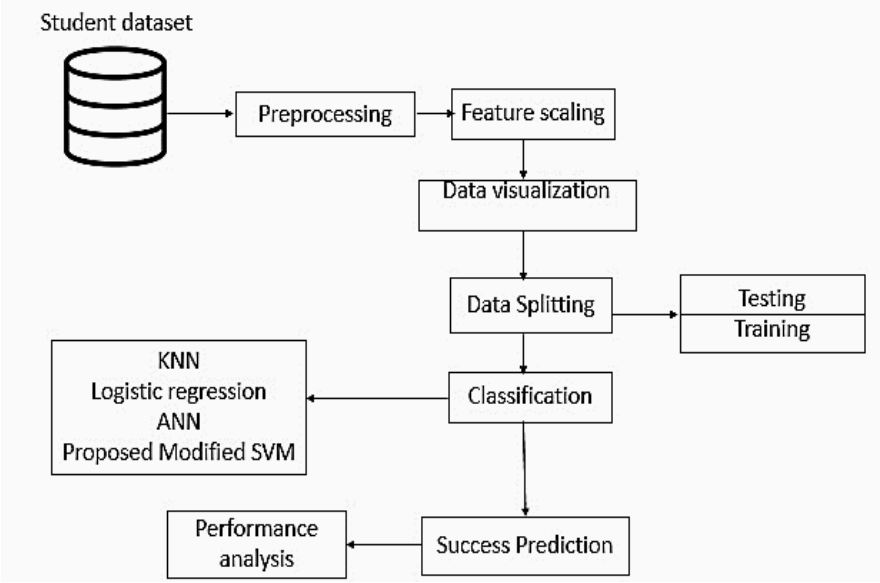

Figure 1 Architecture of the Proposed Work.

Logistic Regression is a statistical method used to predict binary outcomes. It predicts dependent variables by examining the relationship with independent variables. This method facilitates understanding the relationships between different variables and their influence on the output. It is commonly used in Business Analysis and Healthcare sectors due to its ease of setup and training.On the other hand, a Support Vector Machine (SVM) is a supervised learning algorithm that identifies the best decision boundary known as a hyperplane.An Artificial Neural Network (ANN) is a computational model comprised of numerous elements that process inputs and produce outputs based on specific activation functions. Structured to mimic the behavior of neurons in the human brain, an ANN employs mathematical functions, represented as F(x)F(x), simulating the neural connections and interactions found in the brain (as cited by G. James, D. Witten, et al.). By layering these functions in a network configuration, the ANN models the intricate connections between neurons.

The network is trained using historical data to determine the weights associated with each factor. The ANN operates as a multi-layered hierarchical model. Each layer contains several nodes (or neurons), which are connected to all nodes in the subsequent layer through unidirectional linkages. Nodes within the same layer or in preceding layers remain unconnected. The subsequent diagram illustrates the input layer, several intermediate layers, and the output layer (refer to Figure 2). In this specific ANN architecture, the input layer represents various categorical values of the data features under investigation. The network also includes two intermediary layers with 12 and 7 neurons respectively, and an output layer featuring a single neuron which denotes the binary outcome.An artificial neural network (often simply termed a neural network) emulates the functionalities of nerve cells in the human brain. ANNs utilize learning algorithms that allow them to autonomously adapt—or essentially "learn"—as they process new data. With nonlinear mapping, data that isn't linearly separable can be projected into a higher-dimensional feature space. In this transformed space, the network identifies an optimal hyper-plane for classification. The choice of an appropriate inner-product kernel, satisfying Mercer's conditions, enables linear classification post-mapping.

In the outlined procedure, the SVM technique was employed to categorize student data, leading to the identification of their statuses. Leveraging data from earlier phases, we executed the SVM algorithm on both training and test datasets for classification. The primary repository for data accumulation was our dedicated database. For validation, a portion of the training dataset was harnessed. Both datasets, alongside the classification framework, were processed using the SVM classifier.

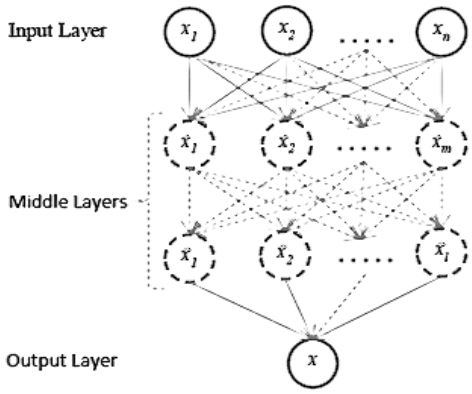

Figure 2 ANN Hierarchy.

3. Results and discussion

Data for the experiment was gathered using tools such as NUMPY, PANDAS, SEABORN, and SKLEARN. We collected various qualitative data points, including Student Identity, Gender, Age, Grade in Lower Class, Grade in Higher Class, Supplementary Skills, Resources Used, Attendance Records, Study Duration, Internal Class Test Grades, Seminar Performance, Lab Assignments, Quizzes, E-Exercises, and E-Homework. To gauge the performance of the trained model in predicting rainfall, we used certain evaluation metrics. These metrics were chosen based on the structure of the datasets used and the outcomes of our experiments. The methodology in our proposed approach serves to measure the efficacy of our prediction models. The specific metrics adopted for this approach are detailed in the following sections.

The confusion matrix offers a robust method for evaluating the performance of a classifier. It essentially tallies how often instances from one class (Class A, for example) are mistakenly classified as another class (Class B, for example). The number of neurons in the hidden layer of an ANN plays a significant role in its prediction accuracy. While increasing the number of neurons has often been observed to enhance the model's fidelity to the training data, it doesn't guarantee better performance in every scenario. Through learning algorithms, the ANN is trained to accurately represent a dataset.The confusion matrices corresponding to various machine learning algorithms are depicted in Figure 3.

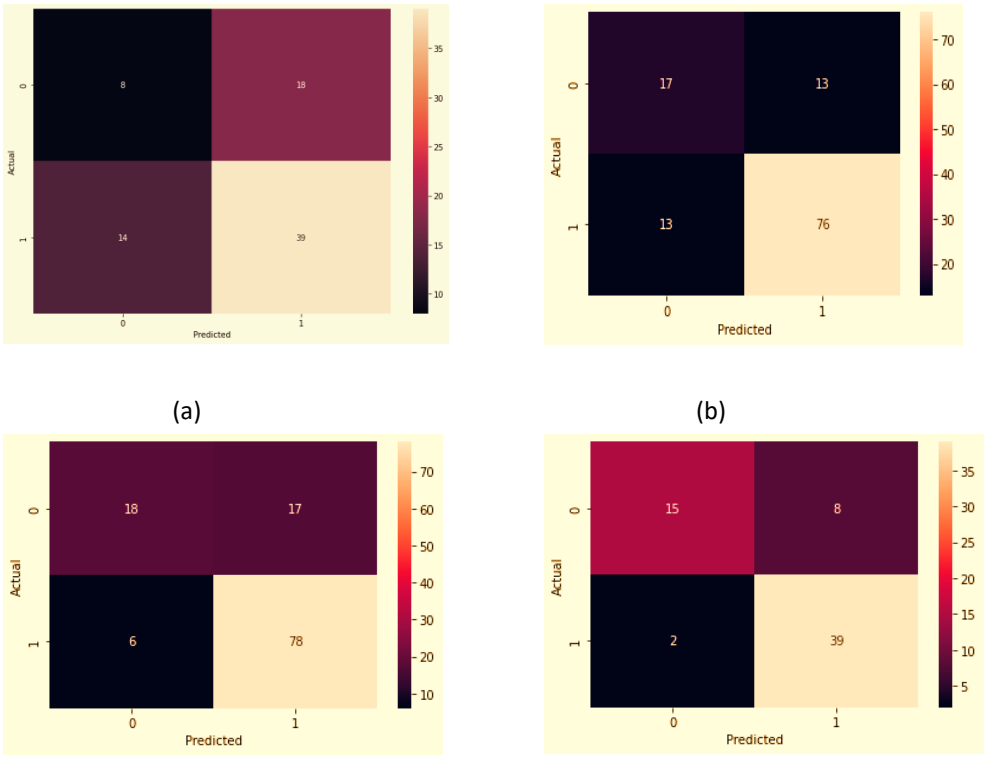

Figure 3 Confusion matrices for (a)ANN, (b)KNN, (c) LR & (d) Modified SVM.

TPRp = TPTp+FNp ---------- (1)
FPRp = FPFPp+TNp ---------- (2)

The ROC (Receiver Operating Characteristic) curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various classification thresholds. These rates can be determined using Equations 1 & 2. By reducing the classification threshold, more items are classified as positive, which in turn increases both the number of False Positives and True Positives. Figure 4 depicts a typical ROC curve.

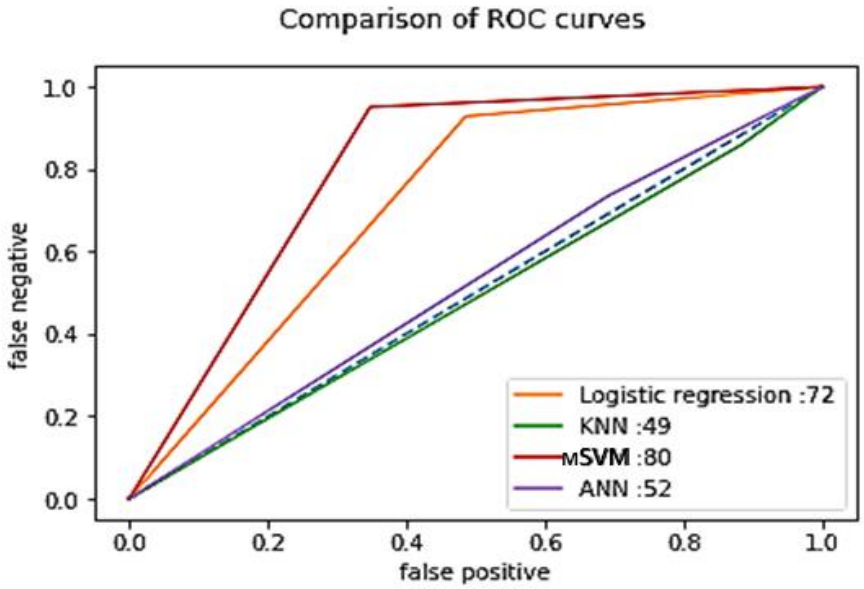

Figure 4 ROC curves for all the Algorithms.

Three primary metrics evaluate a classification model: accuracy, precision, and recall. Of these, accuracy is especially important. It is defined as the proportion of correct predictions in the experimental data. This metric is calculated by dividing the number of correct predictions by the total number of predictions made.

$$\text{Accuracy} = \text{Correct\_predictions} / \text{all\_predictions} \quad (3)$$

Dividing the number of true positives by the sum of true positives and false positives. This measure is known as precision, and it can be formally defined as:

$$\text{Precision} = (\text{true\_positive samples} / \text{true\_positive samples}) + \text{false\_positive samples} \quad (4)$$

Precision is especially crucial when the consequences of false positives are significant. It gives an insight into how many of the predicted positive cases are actually true positives.

Recall, often referred to as sensitivity or the true positive rate, is especially vital when the cost of missing a true positive is high. It indicates how many of the actual positive cases were correctly identified by the model.

Dividing the number of true positives by the sum of true positives and false negatives. This metric is termed recall, and it can be formally defined as:

$$\text{Recall} = \text{true\_positive samples} / \text{true\_positive samples} + \text{false\_negatives} \quad (5)$$

All the evaluation metrics mentioned above are utilized in both the conventional algorithms and the proposed modified Support Vector Machine (M-SVM). It's evident that the M-SVM exhibits superior performance in comparison to the other algorithms. A graphical representation comparing these evaluation measures is depicted in Figure 5.

For the proposed student academic dataset, the prediction model is evaluated using RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) metrics. A comparison of these error metrics is presented in Figure 6. From the illustration, it's evident that the M-SVM boasts the lowest values when compared to other metrics, indicating its superior performance.

$$\boldsymbol{MAE} = \frac{\sum_{i=1}^{J} |er_p - er_t|}{J} \quad (6)$$

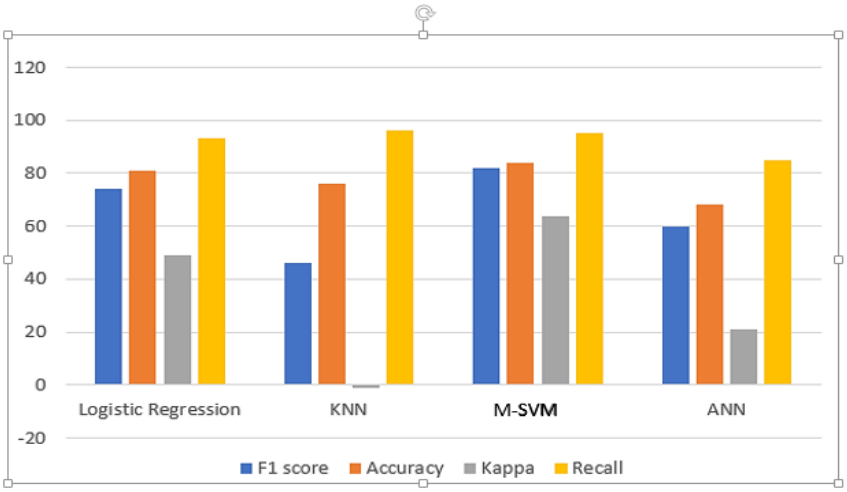

Figure 5 Comparison of evaluation measures.

The Root Mean Squared Error (RMSE) stands as a primary measure for determining the accuracy of predictions, particularly within regression models. It serves to measure the disparity between the values predicted by a model and the actual observed values. In essence, RMSE computes the square root of the mean of the squared discrepancies between forecasts and real-world outcomes. A model is considered to be performing well when the RMSE is low, signaling close alignment between predicted and observed data. On the other hand, a high RMSE indicates potential shortcomings in the model's ability to grasp the intrinsic patterns within the data. By aligning the error magnitude with the unit of the prediction, RMSE offers a direct, interpretable insight into prediction accuracy, making it a valuable metric in practical contexts.Root Mean Squared Error (RMSE) is one of the often-used measures for assessing prediction quality.

$$\boldsymbol{MSE} = \sqrt{\frac{\sum_{i=1}^{J} |er_p - er_t|^2}{J}} \qquad (7)$$

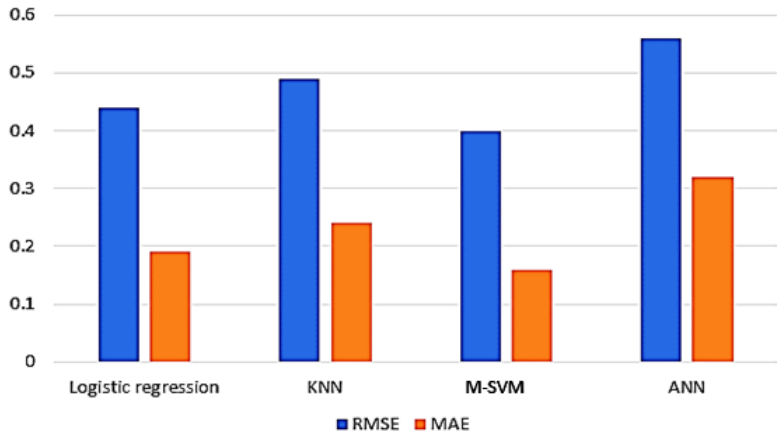

Figure 6 Comparison of Error metrics.

The Euclidean distance is employed to depict the magnitude of discrepancies between predicted outcomes and actual observed values. The computation for MAE and RMSE is provided in Equations 6 and 7. In this investigation, we evaluate the efficiency of the educational landscape using diverse machine learning methodologies. Among the methods explored, experimental results indicated that the Support Vector Machine (SVM) exhibited superior classification accuracy. This prominence was especially evident when weighing the significance of test scores and other facets influenced by the defined ruleset.

4. Conclusions

In this proposed approach, we identified various factors that influence student performance using machine learning algorithms. Implementing diverse machine learning methods to understand the impact on student performance necessitates extensive work and research within the broad realm of educational data mining. The proposed model achieved an accuracy of

80%, making it a viable tool for predicting student outcomes. Our research presents a valuable method in education to unearth patterns or problems in student progression that might be obscured within student data. Implementing early intervention strategies can greatly enhance student performance. Enriching the data with additional attributes, either at the onset of enrollment or at any stage during the academic journey, can further increase student success rates. Future work will seek to generalize this approach to larger student cohorts, incorporate more detailed programming assignments, and integrate additional relevant data, including intervention strategies and demographic details.

Ethical considerations

Not applicable.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research did not receive any financial support.

References

Agrawal, H., Mavani, H. (2015). Student performance prediction using machine learning. *International Journal of Engineering Research and Technology*, 4(03), 111-113.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education, 17*(1), 1-21.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems, 49*(4), 498-506.

Gardner, J., Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction, 28*(2), 127-203.

Gareth J., Daniela, W., Trevor, H., Robert, T. (2013). An introduction to statistical learning: with applications in R. Springer.

Hoffait, A. S., & Schyns, M. (2017). Early detection of university students with potential difficulties. *Decision Support Systems, 101,* 1-11.

Kuzilek, J., Zdrahal, Z., & Fuglik, V. (2021). Student success prediction using student exam behaviour. *Future Generation Computer Systems, 125,* 661-671.

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education, 103,* 1-15.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems, 33*(1), 107-112.

Mayilvaganan, M., & Kalpanadevi, D. (2014). Comparison of classification techniques for predicting the performance of student's academic environment. In *2014 International Conference on Communication and Network Technologies IEEE,* 113-118.

Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems, 115,* 36-51.

Putnik, G., Costa, E., Alves, C., Castro, H., Varela, L., & Shah, V. (2016). Analyzing the correlation between social network analysis measures and performance of students in social network-based engineering education. *International Journal of Technology and Design Education, 26*(3), 413-437.

Serrano, C., Murgui, S., & Andreu, Y. (2022). Improving the prediction and understanding of academic success: The role of personality facets and academic engagement. *Revista de Psicodidáctica (English ed.), 27*(1), 21-28.

Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72,* 414-422.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J., & Abreu, R. (2015). A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society.*

Yagci, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 1-19.