Intelligent Methods in Engineering Sciences 3(3): 091-098, 2024

IMIENS

INTELLIGENT METHODS
IN ENGINEERING SCIENCES

https://www.imiens.org

International
Open Access

September, 2024

e-ISSN 2979-9236

Research Article

https://doi.org/10.58190/imiens.2024.103

Predicting Student Dropout Using Machine Learning Algorithms

Suleyman Alpaslan SULAK [a,*], ***Nigmet KOKLU*** [b]

[a] *Ahmet Kelesoglu Educational Faculty, Necmettin Erbakan University, Konya, Türkiye*
[b] *Technical Science Vocational High School, Konya Technical University, Konya, Türkiye*

ARTICLE INFO

Article history:
Received 29 July 2024
Accepted 10 September 2024

Keywords:
Artificial Neural Network,
Decision Tree,
Machine Learning,
Random Forest,
Student Dropout

ABSTRACT

This article comprehensively examines the use of machine learning algorithms to predict and reduce student dropout rates. These methods, developed to monitor and support student achievement in education, also aimed to enhance success rates in education and ensure more effective student engagement in the learning process. Big data analysis and machine learning models provide important contributions to the development of strategic solutions to the problem of school dropout by predicting student movements and trends. This study uses a dataset consisting of 4424 student data and has 37 features. The dataset is divided into three classes: "Dropout", "Enrolled" and "Graduate" according to the students' school dropout status. Decision Tree (DT), Random Forest (RF) and Artificial Neural Network (ANN) competitions, which are frequently used in such training studies in the literature, are aimed at this dataset. According to the obtained operations, DT showed moderate performance with an accuracy rate of 70.1%. The RF algorithm showed higher success with an accuracy rate of 75.5%. The highest success was achieved by the ANN algorithm with an accuracy rate of 77.3%. ANN's flexible structure has produced superior results compared to other algorithms for this dataset, its ability provide successful classification in complex datasets. The article ultimately demonstrates how machine learning-based prediction models can have a significant impact on student achievement and offer a powerful tool for reducing school dropouts.

1. Introduction

School dropouts are a complex problem that deeply affects individuals and society. Factors such as socioeconomic status, psychological distress and educational background play a significant role in this process [1-2]. Students may withdraw from education due to economic difficulties [3]; child labor [4] and early marriage [5]. Dropout risk is also increased by academic failure and lack of motivation [6].

Demographic factors also play a significant role in school dropout. Gender, age [7], ethnicity and immigration status [8] influence students' retention. Especially female students are at risk of dropping out due to early marriages and societal pressures [9], whereas male students are at risk due to economic factors [10]. Language and cultural barriers [11] may disadvantage immigrant students.

To prevent school dropouts, flexible education models, guidance, and psychosocial support are essential. In addition, students' retention in education can be supported by ensuring the more active participation of families in the process through family education and awareness campaigns. The measures will increase social welfare and individuals' education [12-13].

Student dropout rates are a major problem globally. School dropout not only affects academic achievements, but also has serious consequences on social development and economic growth. This problem is especially common among disadvantaged groups such as children from low-income families, ethnic minorities and immigrant students. Identifying the factors that lead to school dropout is critical for the development of effective strategies to combat this problem. While traditional methods provide significant data on school dropout rates, machine learning algorithms allow us to understand this problem more comprehensively and in depth.

Machine learning (ML) has become a powerful tool for predicting school dropout rates by performing complex analyses on large datasets [14]. This study aims to reveal how machine learning algorithms can be used to determine school dropout rates. Traditional statistical methods often work with limited datasets and fixed modeling approaches, while machine learning (ML) algorithms can process large-scale datasets and make more accurate predictions [15]. Techniques such as machine learning, deep learning, and natural language processing provide effective solutions for identifying the risk of school dropout by

* **Corresponding Author:** sulak@erbakan.edu.tr

analyzing student performance, behaviors, and social interactions in the educational process [16-17].

Machine learning algorithms identify high-risk student groups by observing student behaviors and trends [18]. Supervised learning methods can provide significant insights into educational deficiencies and the factors that contribute to students' tendencies to drop out of school [19]. However, owing to deep learning algorithms, data can be processed to be analyzed in more depth, such as students' intra-school social interactions and psychological states. This, in turn, reveals the fact that not only academic failure, but also social and emotional factors can contribute to school dropout [20].

This article provides a comprehensive review of how machine learning algorithms can be used to predict and reduce student dropout rates. Machine learning algorithms will help students be more involved in educational processes by ensuring the more efficient use of student tracking and support mechanisms in education. Big data analysis and machine learning-based prediction models will contribute to the development of more effective strategies for solving school dropout problems in education.

2. Material and Method

In this section, the dataset used in the study, the machine learning algorithms applied and the performance metrics used to evaluate these models are mentioned. The general structure and functioning of the study are shown in a flow diagram in Figure 1.

2.1. Student Dropout Dataset

The student dropout dataset used in this study consists of 37 features and was created by obtaining 4424 students. In this dataset, students' drop-out status is divided into three classes: "Dropout", "Enrolled" and "Graduate". The dataset is used to better understand and predict the dropout status of students. Table 1 gives the dataset and value properties [21].

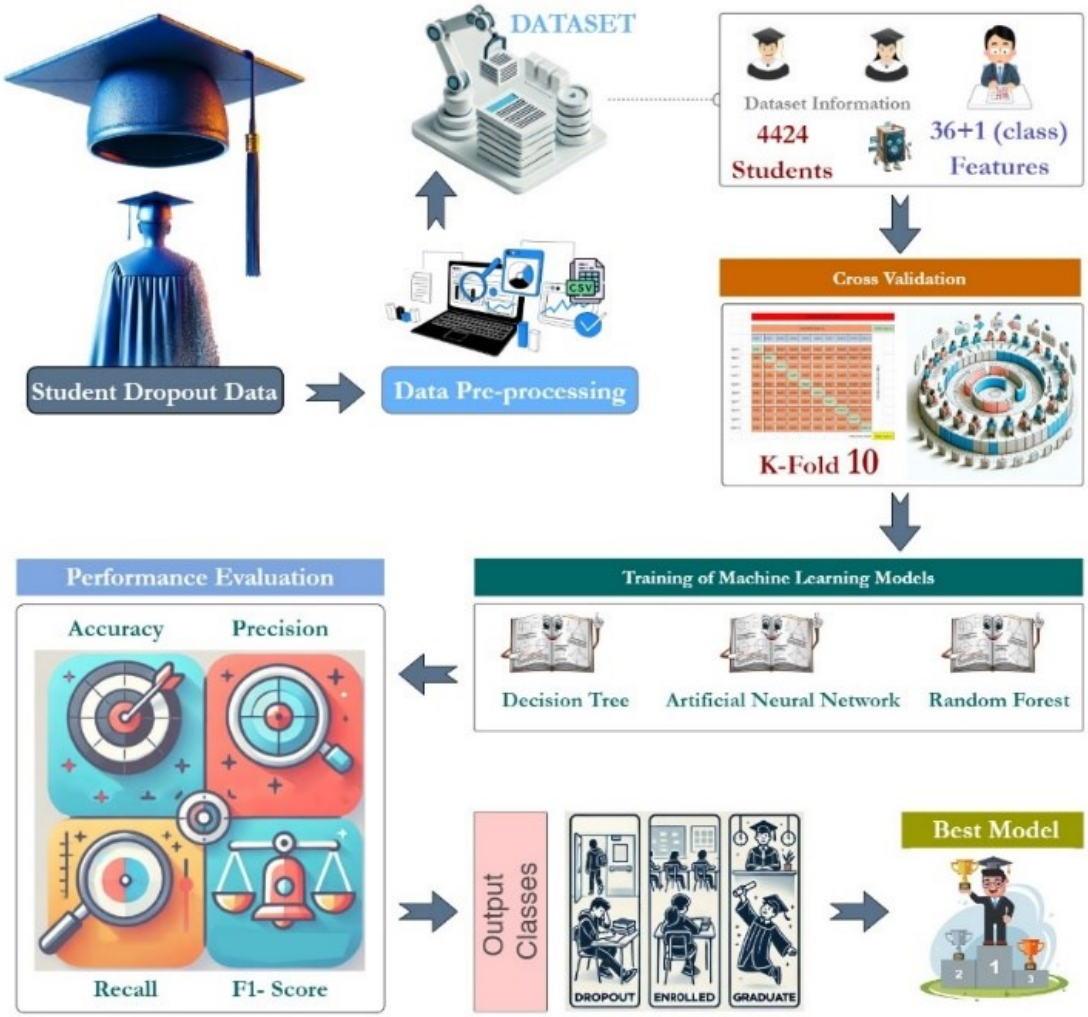

Figure 1. Flow diagram of the Study

Table 1. Student Dropout Dataset Features.

Attributes	Values
Marital status	1 - 6
Application mode	1 - 57
Application order	0 - 9
Course	33 - 9991
Daytime/evening attendance	0 – 1
Previous qualification	1 – 43
Previous qualification (grade)	95 – 190
Nacionality	1 – 109
Mother's qualification	1 – 44
Father's qualification	1 – 44
Mother's occupation	0 – 194
Father's occupation	0 – 195
Admission grade	95 – 190
Displaced	0 – 1
Educational special needs	0 – 1
Debtor	0 – 1
Tuition fees up to date	0 – 1
Gender	0 – 1
Scholarship holder	0 – 1
Age at enrollment	17 – 70
International	0 – 1
Curricular unit 1st sem. (credited)	0 – 20
Curricular unit 1st sem. (enrolled)	0 – 26
Curricular unit 1st sem. (evaluations)	0 – 45
Curricular unit 1st sem. (approved)	0 – 26
Curricular unit 1st sem. (grade)	0 – 18.875
Curricular unit 1st sem. (without evaluations)	0 – 12
Curricular unit 2nd sem. (credited)	0 – 19
Curricular unit 2nd sem. (enrolled)	0 – 23
Curricular unit 2nd sem. (evaluations)	0 – 33
Curricular unit 2nd sem. (approved)	0 – 20
Curricular unit 2nd sem. (grade)	0 – 18.571
Curricular unit 2nd sem. (without evaluations)	0 – 12
Unemployment rate	7.6 – 16.2
Inflation rate	-0.8 – 3.7
GDP	-4.06 – 3.51
Class	Dropout Enrolled Graduate

2.2. Performance Measure

The performance metrics used to evaluate the machine learning models employed in this study are discussed. In such studies, performance metrics are derived from the confusion matrix obtained for machine learning algorithms. Confusion matrices vary according to the nature of the data and the output features, categorized into two-class and multi-class outputs. Table 2 presents the confusion matrix and its explanations for the two-class output case.

Table 2. Two-Class Confusion Matrix and Explanations

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

True Positive (TP): The cases in which the model correctly predicts the positive class.
True Negative (TN): The cases in which the model correctly predicts the negative class.
False Positive (FP): The cases in which the model predicts the negative class as positive
False Negative (FN): The cases in which the model predicts the positive class as negative

The dataset used in this study consists of three classes: “Dropout”, “Enrolled”, and “Graduate”. The confusion matrix corresponding to these outputs is presented in Table 3. Using Table 3, the resulting confusion matrix and the values to be used in the calculations are provided in Table 4.

Table 3. Three-Class Student Dropout Dataset Confusion Matrix

		Predicted		
		Dropout	Enrolled	Graduate
Actual	Dropout	T_1	F_{12}	F_{13}
	Enrolled	F_{21}	T_2	F_{23}
	Graduate	F_{31}	F_{32}	T_3

Table 4. Multi-Class Confusion Matrix

Dropout	**Enrolled**	**Graduate**
$TP_1=T_1$	$TP_2 = T_2$	$TP_3 = T_3$
$TN_1=T_2+T_3+F_{23}+F_{32}$	$TN_2 = T_1 + T_3 + F_{21} + F_{23}$	$TN_3 = T_1 + T_2 + F_{12} + F_{21}$
$FP_1=F_{21}+F_{31}$	$FP_2 = F_{12} + F_{32}$	$FP_3 = F_{13} + F_{23}$
$FN_1=F_{12}+F_{13}$	$FN_2 = F_21 + F_{23}$	$FN_3 = F_{31} + F_{32}$

Performance metrics are criteria used to measure how successful a system, model, or process is. These metrics are critical for assessing how closely specific goals are approached and the effectiveness of the work done. Among the most fundamental performance metrics are criteria such as accuracy, precision, recall, and F-score [22]. Table 5 includes the formulas and descriptions of the performance metrics used.

Table 5. Formulas and Explanations of the Performance Metrics Used

Metric	Formula	Definition
Accuracy	$\frac{TP + TN}{TP + FP + FN + TN} \times 100$	It is a fundamental performance metric that expresses the ratio of correct predictions among all predictions made by a model.
Precision	$\frac{TP}{TP + FP} \times 100$	It is a performance metric that shows how many of a model's positive predictions are actually correct and critical in situations where false positives are significant.
Recall	$\frac{TP}{TP + FN} \times 100$	It is a performance metric that indicates how well a model captures true positives and is critical in situations where false negatives are significant.
F-score	$\frac{2 \times TP}{2 \times TP + FP + FN} \times 100$	The F-score is a performance measure that balances precision and recall metrics, considering the harmony between these two metrics when evaluating the overall performance of the model.

2.3. Cross Validation

It is an important method used to assess machine learning model generalization ability. It is designed to measure not only the model's performance on a training set but also how well it performs on newly collected, untrained data. Cross-validation divides the dataset into a specified number of subsets, and each subset is used as a test set in turn, while the remaining subsets are used to train the model. This process is repeated until each subset is selected as the test set. This way, each data point in the model is used for both training and testing. However, a different portion of the data is tested in each iteration. The most used cross-validation type is known as k-fold cross-validation. In this method, the dataset is divided into 'k' numbers of 'subsets,' and the model is tested with one of these subsets in each iteration while trained with the remaining 'k-1' subsets. The results are calculated by taking the average of the performance metrics obtained at the end of each iteration. This approach provides a more reliable assessment of the model's overall performance by reducing randomness and imbalance issues in the dataset [23-24]. Due to the dataset used in the study consisting of 4424 data points and the large number of data, the k-fold value was chosen and applied as 10. Figure 2 shows 10-fold cross-validation.

k Fold 10	STUDENT DROPOUT DATASET									
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
Split 1	Test 1	Train 2	Train 3	Train 4	Train 5	Train 6	Train 7	Train 8	Train 9	Train 10
Split 2	Train 1	Test 2	Train 3	Train 4	Train 5	Train 6	Train 7	Train 8	Train 9	Train 10
Split 3	Train 1	Train 2	Test 3	Train 4	Train 5	Train 6	Train 7	Train 8	Train 9	Train 10
Split 4	Train 1	Train 2	Train 3	Test 4	Train 5	Train 6	Train 7	Train 8	Train 9	Train 10
Split 5	Train 1	Train 2	Train 3	Train 4	Test 5	Train 6	Train 7	Train 8	Train 9	Train 10
Split 6	Train 1	Train 2	Train 3	Train 4	Train 5	Test 6	Train 7	Train 8	Train 9	Train 10
Split 7	Train 1	Train 2	Train 3	Train 4	Train 5	Train 6	Test 7	Train 8	Train 9	Train 10
Split 8	Train 1	Train 2	Train 3	Train 4	Train 5	Train 6	Train 7	Test 8	Train 9	Train 10
Split 9	Train 1	Train 2	Train 3	Train 4	Train 5	Train 6	Train 7	Train 8	Test 9	Train 10
Split 10	Train 1	Train 2	Train 3	Train 4	Train 5	Train 6	Train 7	Train 8	Train 9	Test 10

Figure 2. 10-fold Cross Validation

2.4. Development of Machine Learning Algorithms

In the literature, DT, RF, and ANN algorithms, commonly used in educational studies, have been applied to the student dropout dataset. These algorithms are among the powerful methods frequently used for classifying student achievement status and educational data, especially for evaluating classification performance in complex datasets. The study explains these algorithms.

2.4.1. Decision Tree (DT) Algorithm

DT is a popular artificial intelligence and machine learning model used for making decisions within a dataset. It resembles a tree structure, branching out from a root node. Each node makes a decision based on a specific feature, and the branches represent possible outcomes. When the final leaf nodes are reached, the model provides a classification or regression result. This model is quite useful for visualizing and understanding data, as each decision step can be clearly traced [25-27].

DT are an effective tool, especially in classification and regression problems. In classification problems, while dividing the data into predetermined classes, numerical values are estimated in regression problems. One of its advantages is that it can work quickly even on large and complex datasets [27]. However, over-branching can cause overfitting. Therefore, parameters such as tree depth should be carefully adjusted [28-29]. Table 6 shows the DT algorithm parameters and values.

Table 6. Parameters and Values of DT Algorithm

Parameters	Values
Minimum number of instances in leaves	2
Do not split subsets smaller than	5
Limit the maximal tree depth to	100

2.4.2. Random Forest (RF) Algorithm

RF is a powerful machine learning algorithm created by combining multiple DT. It was developed to reduce

overfitting, one of DTs' weaknesses. RF trains each DT on a different subset of data and a subset of features, and then aggregates the predictions from all the trees [30]. This approach helps achieve more balanced and accurate results by reducing individual trees' errors [31-33].

RFs have a wide range of applications in classification and regression problems. Each tree makes its own prediction, and the final prediction is obtained by majority voting in classification problems or by averaging in regression problems. One of the advantages of RF is its ability to provide high accuracy and robustness to outliers in the dataset. Additionally, it can indicate the importance of features, showing which factors contribute most to the outcome. However, as the model complexity increases, interpretability may decrease [34-35]. Table 7 presents the RF algorithm parameters and values.

Table 7. Parameters and Values of RF Algorithm

Parameters	Values
Number of trees	10
Number of attributes considered at each split	5
Limit depth of individual trees	3
Do not split subsets smaller than	5

2.4.3. Artificial Neural Network (ANN) Algorithm

ANNs are machine learning models inspired by biological neural systems. ANNs rely on the principle that a large number of simple computational units (neurons) come together to solve complex problems, similar to how neurons in the human brain operate [30]. An ANN typically consists of an input layer, one or more hidden layers, and an output layer. The neurons in each layer are interconnected through weights and activation functions, processing information through these connections [33, 36-37].

ANNs are utilized in various fields such as classification, regression, image recognition, audio processing, and natural language processing. These networks can learn complex relationships within data and make predictions about new data. One of the greatest advantages of ANNs is their ability to learn non-linear relationships, allowing them to solve many complex problems [38-39]. Table 8 presents the ANN algorithm parameters and values.

Table 8. Parameters and Values of ANN Algorithm

Parameters	Values
Neurons in hidden layers	100
Activation Fuction	Logistic
Solver	Adam
Regularization	a=0
Maximal number of iterations:	200

3. Result and Discussion

In this section, confusion matrices for the machine learning algorithms applied to the student dropout dataset, specifically for DT, RF and ANN have been obtained. The results of these matrices have been evaluated.

3.1. Classification Result Made with DT Algorithm

DT algorithm has been applied to the Student Dropout dataset, resulting in the confusion matrix shown in Table 9. Upon examining the confusion matrix:

- DT algorithm made 1050 correct predictions for the Dropout class. However, 174 students were misclassified as Enrolled, and 197 students were misclassified as Graduate. In total, there were 371 incorrect predictions for this class, resulting in a moderate accuracy rate.
- For the Enrolled class, 319 correct predictions were made, but 237 students were incorrectly classified as Dropout, and 238 students as Graduate. A total of 475 misclassifications occurred in this class, indicating that the model exhibits poor performance in the Enrolled class.
- In the Graduate class, the model achieved high accuracy with 1732 correct predictions. However, 206 students were incorrectly predicted as Dropout and 271 as Enrolled. There was a total of 477 misclassifications in the Graduate class, but the overall accuracy for this class can be considered high.

While the DT algorithm demonstrates strong performance in the Graduate class, it shows low performance in the Enrolled class. For the Dropout class, the accuracy rate is reasonable, although misclassifications remain present.

Table 9. Confusion Matrix of Classifications Performed by the DT Algorithm

DT		Predicted		
		Dropout	Enrolled	Graduate
Actual	**Dropout**	1050	174	197
	Enrolled	237	319	238
	Graduate	206	271	1732

3.2. Classification Result Made with RF Algorithm

RF algorithm has been applied to the Student Dropout dataset, resulting in the confusion matrix shown in Table 10. Upon examining the confusion matrix:

- In the Dropout class, the model correctly predicted 1079 students as "Dropout," but misclassified 123 students as "Enrolled" and 219 students as "Graduate." This indicates that while the overall performance for the Dropout class is good, 342 students were incorrectly classified.
- The Enrolled class appears to be the most challenging

for the model. There were 276 correct predictions for this class; however, 199 students were misclassified as "Dropout" and 319 students as "Graduate." This high error rate suggests that the model struggles to accurately learn the Enrolled class.

- In the Graduate class, the model made 1984 correct predictions. However, 89 students were incorrectly classified as "Dropout" and 136 as "Enrolled." This indicates that the model performs quite well in the Graduate class, although some misclassifications are still present.

RF algorithm generally demonstrates that the model is highly successful in the “Graduate” class, while there is a significant need for improvement in the “Enrolled” class. Although the overall performance in the “Dropout” class is acceptable, some misclassifications are noteworthy.

Table10. Confusion Matrix of Classifications Performed by the RF Algorithm

RF		Predicted		
		Dropout	Enrolled	Graduate
Actual	**Dropout**	1079	123	219
	Enrolled	199	276	319
	Graduate	89	136	1984

3.3. Classification Result Made with ANN Algorithm

ANN algorithm has been applied to the Student Dropout dataset, resulting in the confusion matrix shown in Table 11. Upon examining the confusion matrix:

- In the “Dropout” class, the model made 1101 correct predictions. However, 142 students have been misclassified as “Enrolled” and 178 students as “Graduate”. This indicates that there were 320 misclassifications in the “Dropout” class, suggesting that the model's performance in this class is at a moderate level. While the model learns the “Dropout” class well, there are still a significant number of incorrect predictions.
- In the “Enrolled” class, the model's performance is at its weakest level. Only 307 correct predictions were made, while 487 students were misclassified (179 predicted as “Dropout” and 308 as “Graduate”). This high error rate indicates that the model struggles to distinguish the “Enrolled” class. Improvement and optimization strategies are needed for the model to learn this class better.
- In the “Graduate” class, the model made 2012 correct predictions and only 197 misclassifications. This result indicates that the model is highly successful in the “Graduate” class, accurately classifying students in this category.

While the ANN algorithm has achieved high accuracy in the “Graduate” class, it has exhibited a noticeable failure in the “Enrolled” class. For the “Dropout” class, however, an acceptable performance level has been observed.

Table 11. Confusion Matrix of Classifications Performed by the ANN Algorithm

ANN		Predicted		
		Dropout	Enrolled	Graduate
Actual	**Dropout**	1101	142	178
	Enrolled	179	307	308
	Graduate	74	123	2012

4. Conclusion

Three machine learning algorithms were applied to the student dropout dataset, which consists of 4424 instances. These algorithms include DT, RF and ANN. The results of the applied algorithms in terms of accuracy, precision, recall, and F-score are presented in Table 12. Additionally, the graphical results of these machine learning algorithms are illustrated in Figure 3.

Table 12. Results of Machine Learning Algorithms

Model	Accuracy	Precision	Recall	F-score
DT	70.1	70.0	70.1	70.0
RF	75.5	73.9	75.5	74.2
ANN	77.3	76.0	77.3	76.2

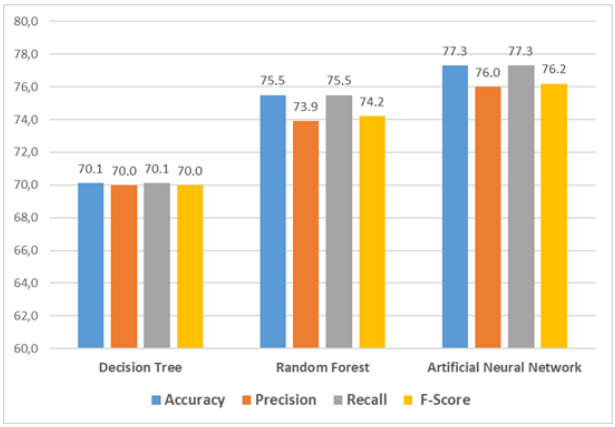

Figure 3. Graphical Results of Machine Learning Algorithms

DT algorithm has an accuracy rate of 70.1%. Its precision value is 70.0%, recall is 70.1%, and F-score is 70.0%. RF algorithm demonstrates a higher performance than DT, with an accuracy of 75.5%. Its precision value is 73.9%, recall is 75.5%, and F-score is 74.2%. ANN algorithm shows the highest performance, achieving an accuracy rate of 77.3%. The precision is recorded at 76.0%, recall at 77.3%, and F-score at 76.2%.

Given these results, it is evident that the performance of different machine learning algorithms can vary depending on the dataset complexity. DT a simple and interpretable algorithm, exhibits limited performance, especially in complex classification problems. In contrast, RF algorithm, by aggregating multiple decision trees, is more generalizable and achieves better results than DT. ANN algorithm has demonstrated the highest accuracy and F-

score, outperforming the other algorithms. This indicates ANN's flexible structure and superior ability to classify complex datasets effectively.

When considering how these algorithms perform across different classes, both ANN and RF provide more balanced results than DT. Metrics such as precision, recall, and F-score indicate that ANN and RF outclass at handling more complex data, thereby achieving higher performance.

These results indicate that machine learning-based predictive models can be a powerful tool in addressing student dropout issues in the education sector. It can be inferred that attention should be paid to the structure and complexity of the dataset in model selection and development, as more advanced algorithms may yield better results. ANN can better handle complex data and make more accurate predictions.

Considering this study, different machine learning algorithms can be employed to better analyze the student dropout dataset and enhance predictive power. Optimizing parameters for similar model types can significantly improve model performance. By employing hyperparameter tuning, techniques such as grid search or random search can be used to identify the most effective parameter combinations. Additionally, implementing hybrid methods can facilitate the combined use of various algorithms, leading to higher success rates. To enhance the dataset's effectiveness, improvements can be made during the data preprocessing stage, such as implementing data normalization and outlier analysis. Feature reduction techniques can reduce model complexity and shorten computation time by selecting the most significant and decisive variables. It is recommended that new studies be conducted considering these strategies to enable a more in-depth examination of the dataset and increase its predictive power.

Data availability

The data used to support the findings of this study are available on the https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could influence the work reported in this paper

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for profit sectors.

References

[1] Aina, C., Baici, E., Casalone, G., and Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, *79*, 101102. https://doi.org/10.1016/j.seps.2021.101102

[2] Domar, A. D. (2004). Impact of psychological factors on dropout rates in insured infertility patients. *Fertility and sterility*, *81*(2), 271-273. https://doi.org/10.1016/j.fertnstert.2003.08.013

[3] Bennett, R. (2003). Determinants of undergraduate student drop out rates in a university business studies department. Journal of Further and Higher Education, 27(2), 123-141. https://doi.org/10.1080/03098770320006515 4

[4] Tang, C., Zhao, L., and Zhao, Z. (2018). Child labor in China. *China Economic Review*, *51*, 149-166. https://doi.org/10.1016/j.chieco.2016.05.006

[5] Mehra, D., Sarkar, A., Sreenath, P., Behera, J., and Mehra, S. (2018). Effectiveness of a community based intervention to delay early marriage, early pregnancy and improve school retention among adolescents in India. *BMC public health*, *18*, 1-13. https://doi.org/10.1186/s12889-018-5586-3

[6] Kaplan, D. S., Peck, B. M., and Kaplan, H. B. (1997). Decomposing the academic failure–dropout relationship: A longitudinal analysis. *The Journal of Educational Research*, *90*(6), 331-343. https://doi.org/10.1080/00220671.1997.10544591

[7] Brorson, H. H., Arnevik, E. A., Rand-Hendriksen, K., and Duckert, F. (2013). Drop-out from addiction treatment: A systematic review of risk factors. *Clinical psychology review*, *33*(8), 1010-1024. https://doi.org/10.1016/j.cpr.2013.07.007

[8] Archambault, I., Janosz, M., Dupéré, V., Brault, M. C., and Andrew, M. M. (2017). Individual, social, and family factors associated with high school dropout among low-SES youth: Differential effects as a function of immigrant status. *British Journal of Educational Psychology*, *87*(3), 456-477. https://doi.org/10.1111/bjep.12159

[9] Stratton, L. S., O'Toole, D. M., and Wetzel, J. N. (2007). Are the factors affecting dropout behavior related to initial enrollment intensity for college undergraduates? *Research in Higher Education*, *48*(4), 453-485. https://doi.org/10.1007/s11162-006-9033-4

[10] Wood, L., Kiperman, S., Esch, R. C., Leroux, A. J., and Truscott, S. D. (2017). Predicting dropout using student-and school-level factors: An ecological perspective. *School Psychology Quarterly*, *32*(1), 35.

[11] Perreira, K. M., Harris, K. M., and Lee, D. (2006). Making it in America: High school completion by immigrant and native youth. *Demography*, *43*(3), 511-536. https://doi.org/10.1353/dem.2006.0026

[12] Christenson, S. L., and Thurlow, M. L. (2004). School dropouts: Prevention considerations, interventions, and challenges. *Current Directions in Psychological Science*, *13*(1), 36-39. https://doi.org/10.1111/j.0963-7214.2004.01301010.x

[13] Janosz, M., Le Blanc, M., Boulerice, B., and Tremblay, R. E. (2000). Predicting different types of school dropouts: A typological approach with two longitudinal samples. *Journal of educational psychology*, *92*(1), 171.

[14] Ameen, A. O., Alarape, M. A., and Adewole, K. S. (2019). Students' academic performance and dropout predictions: A review. *Malaysian Journal of Computing*, *4*(2), 278-303.

[15] Rahmani, A. M., Azhir, E., Ali, S., Mohammadi, M., Ahmed, O. H., Ghafour, M. Y., ... and Hosseinzadeh, M. (2021). Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. *PeerJ Computer Science*, *7*, e488. https://doi.org/10.7717/peerj-cs.488

[16] Gubbels, J., Van der Put, C. E., and Assink, M. (2019). Risk factors for school absenteeism and dropout: A meta-analytic review. *Journal of youth and adolescence*, *48*, 1637-1667. https://doi.org/10.1007/s10964-019-01072-5

[17] Sorensen, L. C. (2019). "Big Data" in educational administration: An application for predicting school dropout risk. *Educational Administration Quarterly*, *55*(3), 404-446. https://doi.org/10.1177/0013161X18799439

[18] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., and Addison, K. L. (2015, August). A machine learning framework to identify students at risk of adverse

academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1909-1918). https://doi.org/10.1145/2783258.2788620

[19] Rumberger, R. W., and Lim, S. A. (2008). Why students drop out of school: A review of 25 years of research.

[20] Becker, B. E., and Luthar, S. S. (2002). Social-emotional factors affecting achievement outcomes among disadvantaged students: Closing the achievement gap. *Educational psychologist*, *37*(4), 197-214. https://doi.org/10.1207/S15326985EP3704_1

[21] Realinho, V., Vieira Martins, M., Machado, J., and Baptista, L. (2021). Predict Students' Dropout and Academic Success [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.

[22] Koklu, N. and Sulak, S. A. (2024a). The Systematic Analysis of Adults' Environmental Sensory Tendencies Dataset. Data in Brief, Vol.55, 110640, https://doi.org/10.1016/j.dib.2024.110640

[23] Arlot, S., and Celisse, A. (2010). A survey of cross-validation procedures for model selection. https://doi.org/10.1214/09-SS054

[24] Kaya, I. and Cinar, I. (2024). Evaluation of Machine Learning and Deep Learning Approaches for Automatic Detection of Eye Diseases. Intelligent Methods In Engineering Sciences, 3(1), 37-45.

[25] Rana, K. K. (2014). A survey on decision tree algorithm for classification. *International journal of Engineering development and research*, *2*(1), 1-5.

[26] Charbuty, B., and Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20-28.

[27] Koklu N. and Sulak S.A., (2024b). "Classification of Environmental Attitudes with Artificial Intelligence Algorithms", Intell Methods Eng Sci, vol. 3, no. 2, pp. 54–62, Jun. 2024, https://doi.org/10.58190/imiens.2024.99

[28] Loh, W. Y. (2011). Classification and regression trees. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(1), 14-23. https://doi.org/10.1002/widm.8

[29] Xu, M., Watanachaturaporn, P., Varshney, P. K., and Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, *97*(3), 322-336. https://doi.org/10.1016/j.chieco.2016.05.006

[30] Sulak, S. A. and Koklu, N. (2024). Analysis of Depression, Anxiety, Stress Scale (DASS-42) With Methods of Data Mining. *European Journal of Education*, e12778. https://doi.org/10.1111/ejed.12778

[31] Biau, G., and Scornet, E. (2016). A random forest guided tour. *Test*, *25*, 197-227. https://doi.org/10.1007/s11749-016-0481-7

[32] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

[33] Koklu, N. and Sulak, S.A. (2024c). Using artificial intelligence techniques for the analysis of obesity status according to the individuals' social and physical activities. Sinop Üniversitesi Fen Bilimleri Dergisi, 9(1), 217-239. https://doi.org/10.33484/sinopfbd.1445215

[34] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958. https://doi.org/10.1021/ci034160g

[35] Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., ... and Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, *22*(16), 2028-2036. https://doi.org/10.1093/bioinformatics/btl344

[36] Agatonovic-Kustrin, S., and Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis*, *22*(5), 717-727. https://doi.org/10.1016/S0731-7085(99)00272-1

[37] Zurada, J. (1992). *Introduction to artificial neural systems*. West Publishing Co..

[38] Kumar, B. R., Vardhan, H., Govindaraj, M., and Vijay, G. S. (2013). Regression analysis and ANN models to predict rock properties from sound levels produced during drilling. *International Journal of Rock Mechanics and Mining Sciences*, *58*, 61-72. https://doi.org/10.1016/j.ijrmms.2012.10.002

[39] Abiodun O. I. et al., "Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition," in IEEE Access, vol. 7, pp. 158820-158846, 2019, doi: 10.1109/ACCESS.2019.2945545.