ISSN: 2146-1961

Bezek Güre, Ö. (2024). Classification of Students' Dropout Status using the Random Forest Method, *International Journal of Eurasia Social Sciences (IJOESS)*, 15(57), 1401-1411.

DOI: http://dx.doi.org/10.35826/ijoess.4507

ArticleType (Makale Türü): Research Article

CLASSIFICATION OF STUDENTS' DROPOUT STATUS USING THE RANDOM FOREST METHOD

Özlem BEZEK GÜRE

Assist. Professor, Batman University, Batman, Türkiye, ozlem.bezekgure@batman.edu.tr

ORCID: 0000-0002-5272-4639

Received: 10.06.2024 *Accepted: 16.08.2024* *Published: 01.09.2024*

ABSTRACT

Higher education institutions are regarded as indicators of countries' economic and social development. The dropout or failure of students from higher education institutions due to various reasons not only affects the reputation of these institutions but also poses significant problems for students, their families, and society in general. Therefore, predicting students at risk of dropping out is considered crucial. This study primarily aims to predict students at risk of dropping out from higher education institutions using the Random Forest method, which is among the educational data mining techniques. Secondly, it aims to compare the classification performance of the method based on sample size. For this purpose, a dataset from the Kaggle database, created to reduce academic failure and dropout rates in higher education, was used. This dataset includes data on students' enrollment information and their demographic and socioeconomic status. The dataset comprises 4424 samples with 37 variables, one of which is the dependent variable. Random samples of 500, 1000, 2000, 3000, and 4000 were drawn from the dataset. Analyses were conducted using an open-source Python-based program. The area under the ROC curve (AUC), accuracy, F1 score, precision, and recall metrics were used to measure the classification performance of the method. The performance criteria were found as follows: AUC: 0.961, accuracy: 0.881, F1: 0.878, precision: 0.879, and recall: 0.881. The analysis results indicate that the method shows better classification performance with a sample size of 4000. Additionally, it was determined that classification success increases with the sample size. The most significant variable across all sample sizes is Curricular units 2nd sem (approved). It is recommended to examine the conditions of different data mining methods concerning student dropout and failure under varying conditions.

Keywords: Dropout, higher education, sample size, random forest, educational data mining.

INTRODUCTION

Dropping out of school poses significant problems for educational institutions both in terms of reputation and financially. This situation can be attributed to various factors, including economic and social reasons as well as academic failure (El Aouifi, El Hajji & Es-Saady, 2024). Students who face difficulties are more likely to exhibit antisocial personality disorders and have lower chances of participating in the labor market (Chung & Lee, 2019). In other words, the likelihood of students enhancing their skills and seizing job opportunities decreases, leading to a decline in their standard of living (Cuevas-Chávez et al., 2023). Therefore, identifying the factors that cause dropout and failure is crucial for the entire society, particularly for higher education institutions (El Aouifi, El Hajji & Es-Saady, 2024). Educational data mining (EDM) methods can be used to predict dropout risk with high accuracy. EDM is a collection of methods that combine databases, machine learning, and statistics (Anuradha & Velmurugan, 2016; Güre, 2023). Educational data mining explores large data sets in the educational field using various methods to examine student behaviors and learning environments (Sachin & Vijay, 2012). This allows for the discovery of patterns in student behaviors and the prediction of their future states (Ayala, 2014; Kayri, 2023).

This study aims to predict the dropout risk of students in higher education using the Random Forest method, one of the educational data mining techniques. Additionally, it seeks to examine the performance of the method across different sample sizes. In line with this goal, the following questions were addressed:

1. What is the classification performance of the Random Forest method in predicting dropout status?
2. Does the classification performance of the Random Forest method vary according to sample sizes?
3. Which are the most important variables affecting dropout status according to the Random Forest method?

Related Studies

There are numerous studies in the literature that investigate student dropout status using machine learning methods. Some of these studies have also evaluated the dataset used in the current study. For instance, Martins et al. (2021) conducted research using Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) methods. Realinho et al. (2022) examined student dropout status using RF, Extreme Gradient Boosting (XGBOOST), Light Gradient Boosting Machine (LGBM), and CatBoost (CB) methods. Similarly, Tang et al. (2024) evaluated RF, Bagging, SVM, DT, Extra Tree, K-Neighbors (KNN), Linear Discriminant Analysis (LDA), LR, Ridge Classifier (RC), and Stochastic Gradient Descent (SGD) methods. Villar and Andrade (2024) conducted their research using DT, SVM, RF, Gradient Boosting (GB), XGBoost, CB, and LGBM methods. Bhurre and Prajapat (2023) preferred Multilayer Perceptron (MLP), Simple Logistic (SL), DT, RF, and Reduced Error Pruning Tree (REPTree) methods. Sabbir et al. (2024) predicted student dropout status using XGBoost, RF, KNN, and DT methods. Additionally, Cuevas-Chávez et al. (2023) used RF, SVM, and XGBoost methods, along with three different resampling techniques in their research. Oludipe, Saeed, and Mohammed (2023) examined LR, XGBoost, and GB methods. El Aouifi, El Hajji, and Es-Saady (2024) conducted similar

studies using KNN, Naive Bayes (NB), MLP, DT, SVM, RF, and Deep Neural Network (DNN) methods. Sharan, Ghosh, and Chhabra (2024) investigated Gaussian Naive Bayes (GNB), LR, RF, and SVM methods. Strohmier et al. (2024) evaluated DT, RF, Multinomial Naive Bayes (MNB), and GNB methods. Zhu, Wang, and Fan (2024) used a genetic algorithm method to predict student dropout status. Drousiotis et al. (2023) applied the Sequential Monte Carlo method. Beyond the dataset used in the current study, there are other studies predicting college student dropout status. For example, Chung and Lee (2019) predicted the dropout risk of college students using the Random Forest method.

The current study employed the Random Forest method, one of the data mining techniques, to predict students at risk of dropping out. Unlike other studies, this research examined the performance of the method across different sample sizes. A review of the literature revealed no similar study addressing this specific aspect.

METHOD

Research Model

The present study is designed as a quantitative research study. A correlational survey model was employed in this research. This model aims to determine the existence of a covariation between two or more variables (Karasar, 2011).

Employed dataset

A dataset prepared to predict student dropout status and academic success available in the Kaggle and UCI database. The data file was obtained from https://www.kaggle.com/datasets/waleedejaz/predictstudents-dropout-and-academic-success (Realinho et al., 2021). The dataset consists of 37 variables and 4424 samples. The dependent variable, school status, has three categories: Dropout, Enrolled, and Graduate. The dataset includes variables such as gender, age, nationality, and marital status, as well as academic information for the first and second semesters and economic variables such as unemployment rate and inflation rate.

Data Analysis

Data analysis was conducted using the Orange program, a Python-based application. Orange is a user-friendly program used for the application of statistical and data mining methods. This program allows for interactive workflows and visualizations, which makes it easier to understand complex data patterns. The Random Forest method, which does not require any assumptions and has a high accuracy rate in classifying school dropout situations in large datasets, was preferred in the study.

RF method, an ensemble method, combines multiple decision trees (Wang et al., 2016). This method, used for both classification and regression purposes, was developed by Breiman (Biau & Scornet, 2016). Breiman combined bagging, the random subspace method, and the Classification and Regression Tree (CART) method (Sun et al., 2024). The RF method is considered non-parametric because it does not rely on any assumptions

(Geneur et al., 2017; Şevgin & Eranıl, 2023). With the RF method, variable importance and proximity between samples can be measured, missing data imputation can be performed, and outliers can be detected (Breiman, 2001; Cutler et al., 2012). It can be applied to both large and small datasets and allows for the use of any desired number of trees (Archer & Kimes, 2008; Biau & Scornet, 2016). The method exhibits high performance even in the presence of missing data, outliers, and overfitting (Liu, Wang & Zhang, 2012).

In the RF method, decision trees are created using clustering and bootstrap techniques (Bezek Güre, Şevgin & Kayri, 2023). Each tree in the method is grown iteratively, starting from the root tree. Because it combines many weak learners, its predictive power is high. The splitting of nodes is determined by the Gini index for categorical dependent variables and by variance for continuous dependent variables. The final layer of the tree is called the leaf node. The node splitting is done by selecting the best predictors randomly chosen from the node (Akar & Güngör, 2012). These leaves are used for predicting new observations (Hu & Szymczak, 2023). For the result, the new dataset is predicted based on the results of the decision trees in the forest, using the average for regression problems and the majority vote for classification problems (Akman et al., 2011; Liaw & Wiener, 2002).

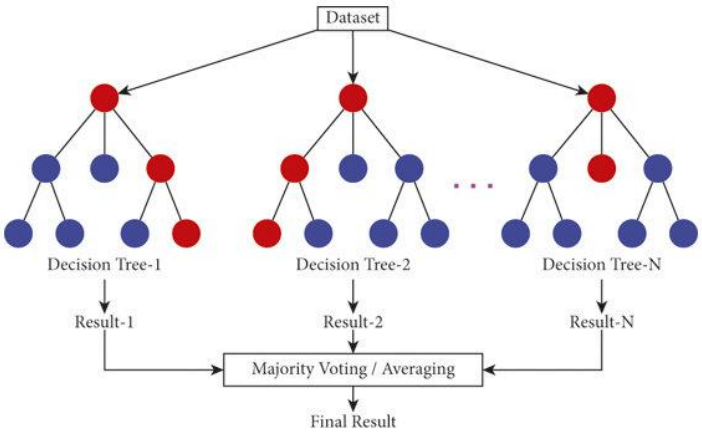


Figure 1. Structure of the Random Forest (Khan et al., 2021)

FINDINGS

Evaluation Metrics

The study used AUC, accuracy, precision, F1 score, and recall as performance metrics.

Table 1. Confusion Matrix

		Estimated Class		
		No	Yes	Total
Real Class	No	TN	FP	TN+FP
	Yes	FN	TP	FN+TP
	Total	TN+FN	FP+TP	TN+FN+FP+TP

AUC: It is also known as the area under the ROC curve, indicates the performance of the model (Şevgin & Önen, 2022).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ score = 2x\ \frac{(Precision\ x\ Recall)}{(Precision + Recall)} \quad (4)$$

The student's dropout status was taken as the dependent variable. The dataset comprised information from 1421 (32.1%) dropout students, 2209 (49.9%) graduates, and 794 (17.9%) enrolled students. Before applying the method, the dataset was split into 70% training data and 30% test data. Then, using the Orange program, random samples of sizes 500, 1000, 2000, 3000, and 4000 were drawn. Subsequently, analyses were conducted using the Random Forest method in the same program. The results of the analyses are presented in Table 2.

Table 2. Performance of the Method by Sample Sizes

Sample size	AUC	Accuracy	F1 score	Precision	Recall
500	0,871	0,738	0,724	0,722	0,738
1000	0,892	0,778	0,766	0,766	0,778
2000	0,921	0,820	0,813	0,812	0,820
3000	0,943	0,857	0,853	0,854	0,857
4000	0,961	0,881	0,878	0,879	0,881

Table 2 shows that the highest classification performance was obtained with a sample size of 4000 according to the performance metrics. Additionally, it was determined that classification success increased with sample size.

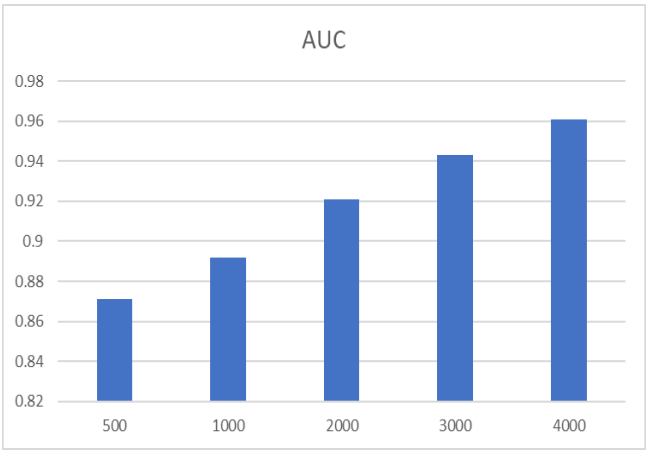


Figure 2. AUC Values by Sample Sizes

Figure 2 shows that as the sample size increases, the AUC values also increase.

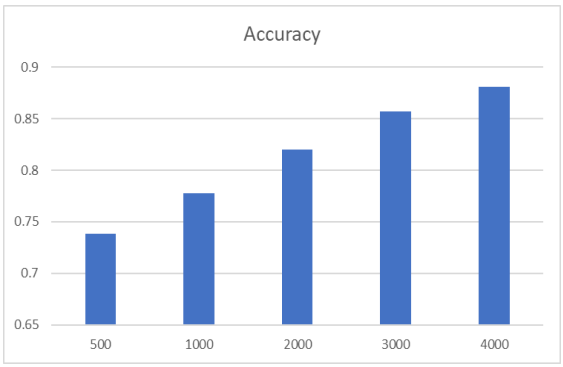


Figure 3. Accuracy Values by Sample Sizes

As seen in Figure 3, the accuracy values increase as the sample sizes increase.

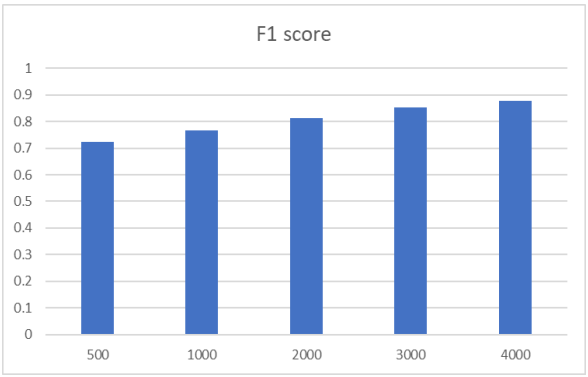


Figure 4. F1 Values by Sample Sizes

Figure 4 indicates that F1 values increase as the sample sizes increase.

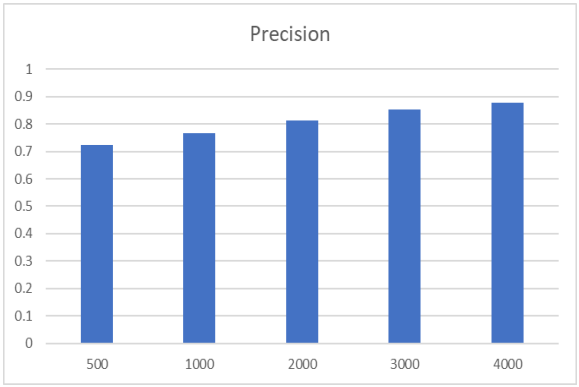


Figure 5. Precision Values by Sample Sizes

Figure 5 indicates that precision values increase as the sample sizes increase.

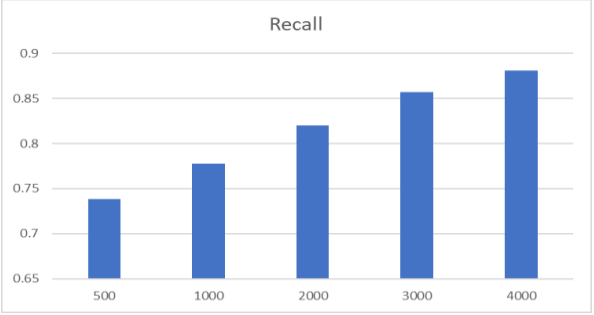
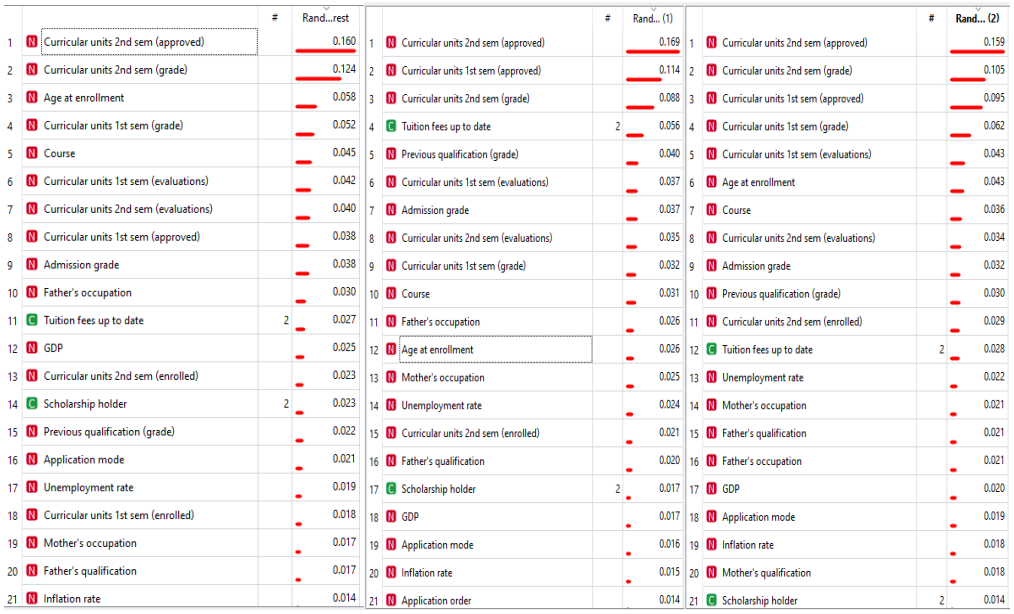


Figure 6. Recall Values by Sample Sizes

As seen in Figure 6, the recall values increase as the sample sizes increase.

Figure 7 and Figure 8 presents the visualizations of variable importance rankings according to different sample sizes.



		#	Rand...rest
1	Curricular units 2nd sem (approved)		0.160
2	Curricular units 2nd sem (grade)		0.124
3	Age at enrollment		0.058
4	Curricular units 1st sem (grade)		0.052
5	Course		0.045
6	Curricular units 1st sem (evaluations)		0.042
7	Curricular units 2nd sem (evaluations)		0.040
8	Curricular units 1st sem (approved)		0.038
9	Admission grade		0.038
10	Father's occupation		0.030
11	Tuition fees up to date	2	0.027
12	GDP		0.025
13	Curricular units 2nd sem (enrolled)		0.023
14	Scholarship holder	2	0.023
15	Previous qualification (grade)		0.022
16	Application mode		0.021
17	Unemployment rate		0.019
18	Curricular units 1st sem (enrolled)		0.018
19	Mother's occupation		0.017
20	Father's qualification		0.017
21	Inflation rate		0.014

		#	Rand... (1)
1	Curricular units 2nd sem (approved)		0.169
2	Curricular units 1st sem (approved)		0.114
3	Curricular units 2nd sem (grade)		0.088
4	Tuition fees up to date	2	0.056
5	Previous qualification (grade)		0.040
6	Curricular units 1st sem (evaluations)		0.037
7	Admission grade		0.037
8	Curricular units 2nd sem (evaluations)		0.035
9	Curricular units 1st sem (grade)		0.032
10	Course		0.031
11	Father's occupation		0.026
12	Age at enrollment		0.026
13	Mother's occupation		0.025
14	Unemployment rate		0.024
15	Curricular units 2nd sem (enrolled)		0.021
16	Father's qualification		0.020
17	Scholarship holder	2	0.017
18	GDP		0.017
19	Application mode		0.016
20	Inflation rate		0.015
21	Application order		0.014

		#	Rand... (2)
1	Curricular units 2nd sem (approved)		0.159
2	Curricular units 2nd sem (grade)		0.105
3	Curricular units 1st sem (approved)		0.095
4	Curricular units 1st sem (grade)		0.062
5	Curricular units 1st sem (evaluations)		0.043
6	Age at enrollment		0.043
7	Course		0.036
8	Curricular units 2nd sem (evaluations)		0.034
9	Admission grade		0.032
10	Previous qualification (grade)		0.030
11	Curricular units 2nd sem (enrolled)		0.029
12	Tuition fees up to date	2	0.028
13	Unemployment rate		0.022
14	Mother's occupation		0.021
15	Father's qualification		0.021
16	Father's occupation		0.021
17	GDP		0.020
18	Application mode		0.019
19	Inflation rate		0.018
20	Mother's qualification		0.018
21	Scholarship holder	2	0.014

n=500 n=1000 n=2000

Figure 7. Importance Levels of Variables by Sample Sizes (500, 1000, 2000)

		#	Rand... (3)
1	Curricular units 2nd sem (approved)		0.207
2	Curricular units 2nd sem (grade)		0.147
3	Curricular units 1st sem (approved)		0.088
4	Admission grade		0.040
5	Curricular units 1st sem (grade)		0.039
6	Curricular units 2nd sem (evaluations)		0.033
7	Previous qualification (grade)		0.032
8	Course		0.028
9	Age at enrollment		0.027
10	Curricular units 1st sem (evaluations)		0.027
11	Father's occupation		0.027
12	Tuition fees up to date	2	0.025
13	Mother's occupation		0.020
14	Curricular units 2nd sem (enrolled)		0.020
15	Father's qualification		0.019
16	Application mode		0.018
17	Curricular units 1st sem (enrolled)		0.017
18	GDP		0.017
19	Inflation rate		0.017
20	Unemployment rate		0.015
21	Mother's qualification		0.015

		#	Rand... (4)
1	Curricular units 2nd sem (approved)		0.121
2	Curricular units 2nd sem (grade)		0.109
3	Curricular units 1st sem (approved)		0.099
4	Curricular units 1st sem (grade)		0.078
5	Curricular units 2nd sem (evaluations)		0.041
6	Admission grade		0.037
7	Curricular units 1st sem (evaluations)		0.036
8	Course		0.035
9	Previous qualification (grade)		0.033
10	Age at enrollment		0.032
11	Curricular units 2nd sem (enrolled)		0.030
12	Tuition fees up to date	2	0.029
13	Father's occupation		0.022
14	Application mode		0.022
15	Mother's occupation		0.022
16	GDP		0.021
17	Inflation rate		0.021
18	Unemployment rate		0.020
19	Debtor	2	0.020
20	Father's qualification		0.019
21	Scholarship holder	2	0.016

n=3000 n=4000

Figure 8. Importance Levels of Variables by Sample Sizes (3000 and 4000)

Figure 7 and Figure 8 shows that the most significant variable affecting students' dropout status, across all sample sizes, is Curricular units 2nd sem (approved).

CONCLUSION and DISCUSSION

This study aimed to predict students' dropout status using the Random Forest method, one of the educational data mining techniques. Additionally, it examined the classification performance of the methods used in both small and large datasets based on sample size. The analysis results indicated that the most successful performance across all performance metrics was achieved with a sample size of 4000. The performance criteria were found as follows: AUC: 0.961, accuracy: 0.881, F1: 0.878, precision: 0.879, and recall: 0.881. It was also observed that classification performance increased with sample size. The most significant variable across all sample sizes was Curricular units 2nd sem (approved). Like the results of this study, Realinho et al. (2022) found Curricular units 2nd sem (approved) to be the most important variable in their study using RF, XGBoost, LightGBM, and CatBoost methods.

There are studies in the literature that use the same dataset. Tang et al. (2024) used RF, Bagging, SVM, DT, Extra Tree, KNN, LDA, LR, RC, and SGD methods and found that the RF method performed better with a classification accuracy of 0.781. Strohmier et al. (2024) examined DT, RF, Multinomial Naive Bayes (MNB), and GNB methods and found that the RF method had the best performance with an accuracy of 77%. Cuevas-Chávez et al. (2023) determined that the SVM method outperformed RF and XGBoost methods with a classification accuracy of 93.55%. Oludipe, Saeed, and Mohammed (2023) evaluated LR, XGBoost, and GB methods, reporting that the LR method had the best performance with an accuracy of 90.91%. Sharan, Ghosh, and Chhabra (2024) analyzed Gaussian Naive Bayes (GNB), LR, RF, and SVM methods, indicating that the LR method was the most successful with a classification accuracy of 92.1%. In the study conducted by El Aouifi, El Hajji, and Es-Saady (2024), KNN, Naive Bayes (NB), MLP, DT, SVM, RF, and Deep Neural Network (DNN) methods were used, concluding that the DNN method was superior with an accuracy of 0.92. Drousiotis et al. (2023) used the Sequential Monte Carlo (SMC) method and achieved a classification accuracy of 71.48%. The present study achieved better results when compared to these other studies.

On the other hand, Villar and Andrade (2024) used DT, SVM, RF, GB, XGBoost, CB, and LGBM methods. They found that the LightGBM and CatBoost methods performed better, with AUC values around 0.9. The present study achieved AUC values better than 0.9 for all sample sizes except for the 500 and 1000 sample sizes. Therefore, the present study obtained better results compared to the study by Villar and Andrade.

There are also studies in the literature that have obtained better results than the present study. Sabbir et al. (2024) found an AUC value of 0.99, achieving more successful results. Similarly, Zhu, Wang, and Fan (2024) applied a genetic algorithm method to predict student dropout status, achieving a precision value of 0.99 and an F1 value of 0.98.

In addition to these studies, Martins et al. (2021) worked with LR, SVM, DT, and RF methods, concluding that the RF method made more accurate predictions. Bhurre and Prajapat (2023) tested MLP, SL, DT, RF, and REPTree methods, finding that the RF method performed better than the other methods. Using a different dataset, Chung and Lee (2019) predicted the dropout risk of college students with the Random Forest method,

achieving a classification accuracy of 0.95, sensitivity of 0.85, specificity of 0.95, and an AUC value of 0.97. They identified unauthorized absenteeism as the most significant variable affecting dropout risk.

Students' school attendance contributes to their personal development and academic growth. Therefore, it is crucial to accurately predict the factors causing absenteeism with high accuracy. The present study is expected to be beneficial in identifying the dropout tendencies of college students and in planning preventive measures accordingly.

SUGGESTIONS

Using this dataset, which is suitable for the application of data mining, further studies can be conducted with different methods. Additionally, studies can be performed to reveal the performance of methods based on the number of variables and sample sizes. Comparisons can also be made according to different test and training set ratios.

ETHICAL TEXT

"In this article, the journal writing rules, publication principles, research and publication ethics, and journal ethical rules were followed. The responsibility belongs to the author (s) for any violations that may arise regarding the article. The dataset used in the present study was sourced from the publicly accessible Kaggle database; therefore, the research did not necessitate ethical committee approval."

Author(s) Contribution Rate: In this study, the contribution rate of the first author is 100%.

REFERENCES

Akar, Ö., & Güngör, O. (2012). Classification of multispectral images using Random Forest algorithm. *Journal of Geodesy and Geoinformation,* 1(2), 105-112. https://doi.org/10.9733/jgg.241212.1

Akman, M., Genç, Y. & Ankaralı, H. (2011). Random Forests Yöntemi ve Saglik alaninda Bir Uygulama/Random Forests Methods and an Application in Health Science. *Türkiye Klinikleri Biyoistatistik*, 3(1), 36.

Bhurre, S., & Prajapat, S. (2023, September). *Analyzing Supervised Learning Models for Predicting Student Dropout and Success in Higher Education*. In UK Workshop on Computational Intelligence (pp. 234-248). Cham: Springer Nature Switzerland.

Biau, G. & Scornet, E. (2016). *A random forest guided tour*. An Official Journal of the Spanish Society of Statistics and Operations Research, ISSN 1133-0686 25(2), 197—227. https://doi.org/10.1007/s11749-016-0481-7

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

Chung JY, & Lee S., (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review,* 96, 346–53 https://doi.org/10.1016/j.childyouth.2018.11.030

Cuevas-Chávez, P. A., Narciso, S., Sánchez-Jiménez, E., Pérez, I. C., Hernández, Y., & Ortiz-Hernandez, J. (2023). *School Dropout Prediction with Class Balancing and Hyperparameter Configuration*. In Mexican International Conference on Artificial Intelligence (pp. 12-20). Cham: Springer Nature Switzerland.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). *Random forests*. Ensemble machine learning: Methods and applications, 157-175. In Zhang ve Ma (Ed.) (pp.157-175).

Drousiotis, E., Varsi, A., Spirakis, P. G., & Maskell, S. (2023, October). A Shared Memory SMC Sampler for Decision Trees. In 2023 IEEE 35th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD) (pp. 209-218). IEEE.

El Aouifi, H., El Hajji, M., & Es-Saady, Y. (2024). A hybrid approach for early-identification of at-risk dropout students using LSTM-DNN networks. *Education and Information Technologies*, 1-19. https://doi.org/10.1007/s10639-024-12588-0

Geneur, R., Poggi, J.M., Tuleao Malot, C., Villa-Vialaneix, N., (2017). Random Forest for big data. *Big Data Research.* 9, 28-46. https://doi.org/10.1016/j.bdr.2017.07.003

Güre, Ö. B. (2023). Investigating the Performance of Feature Selection Methods in Classifying Student Success. *International Journal of Education Technology and Scientific Researches*, 8(24), 2695-2728. https://doi.org/10.35826/ijetsar.668

Hu, J., & Szymczak, S. (2023). A review on longitudinal data analysis with random forest. Briefings in Bioinformatics, 24(2),1-11. https://doi.org/10.1093/bib/bbad002

Kaggle Homepage. https://www.kaggle.com/datasets/waleedejaz/predict-students-dropout-and-academic-success

Karasar, N. (2011). *Bilimsel Araştırma Yöntemleri*. Ankara: Nobel Yayınları.

Kayri, M. (2023). *Eğitimde Veri Madenciliği ve Bilgisayar Uygulamaları içinde (Eğitimde Veri Madenciliği. (s.1-11) Pegem yayınevi.*

Khan, M. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S., & Raazi, S. M. K. U. R. (2021). Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. Complexity, 2021(1), 1-18. https://doi.org/10.1155/2021/2553199

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18- 22. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3 (pp. 246-252). Springer Berlin Heidelberg.

Martins, M.V., Tolledo, D., Machado, J., Baptista, L.M., & Realinho, V. (2021). Early prediction of student's performance in higher education: A case study. Paper presented at the World Conference on Information Systems and Technologies, pp. 166–175. https://doi.org/10.1007/978-3-030-72657-7_16

Oludipe, J., Saeed, F., & Mohammed, R. (2023, December). Machine Learning Techniques for Evaluating Student Performance. In International Conference of Reliable Information and Communication Technology (pp. 306-317). Cham: Springer Nature Switzerland.

Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict Students'Dropout and Academic Success, *UCI Machine Learning Repository*. doi:10.24432/C5MC89.

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, 7(11), 146. https://doi.org/10.3390/data7110146

Sabbir, W., Abdullah-Al-Kafi, M., Afridi, A. S., Rahman, M. S., & Karmakar, M. (2024, February). Improving Predictive Analytics for Student Dropout: A Comprehensive Analysis and Model Evaluation. In 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 951-956). IEEE.

Sachin, R. B., & Vijay, M. S. (2012, January). A survey and future vision of data mining in educational field. In 2012 second international conference on advanced computing & communication Technologies, IEEE, 96-100

Şevgin, H. & Önen, E. (2022). "Comparison of Classification Performances of MARS and BRT Data Mining Methods: ABİDE- 2016 Case", *Education & Science/Egitim ve Bilim*, 47(211),195-222, 2022. https://doi.org/ 10.15390/eb.2022.10575

Şevgin, H., & Eranıl, A. K. (2023). Investigation of Turkish Students' School Engagement through Random Forest Methods Applied to TIMSS 2019: A Problem of School Psychology. *International Journal of Psychology and Educational Studies,* 10(4), 896-909. https://doi.org/10.52380/ijpes.2023.10.4.1260

Sharan, B., Ghosh, S., & Chhabra, M. (2024). *The Application of AI for Automated Education System*. In Innovation in the University 4.0 System based on Smart Technologies (pp. 211-226). Chapman and Hall/CRC.

Strohmier, H., Langner, V., Mohamed, F., & Wood, E. (2024, April). Examination of Artificial Intelligence Integration and Impact on Higher Education. In 2024 12th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1-6). IEEE.

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., & Liang, X. (2024). An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications,* 237(121549), 1-19. https://doi.org/10.1016/j.eswa.2023.121549

Tang, Z., Jain, A., & Colina, F. E. (2024). A Comparative Study of Machine Learning Techniques for College Student Success Prediction. *Journal of Higher Education Theory & Practice*, 24(1), 101-116.

Wang, F., Li, Z., He, F., Wang, R., Yu, W., & Nie, F. (2019). Feature learning viewpoint of AdaBoost and a new algorithm. IEEE Access, 7, 149890-149899.

Zhu, B., Wang, H., & Fan, M. (2024). Constructing small sample datasets with game mixed sampling and improved genetic algorithm. *The Journal of Supercomputing*, 80, 20891–20922. https://doi.org/10.1007/s11227-024-06263