BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **23**, No 4
Sofia • 2023 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2023-0044

Comparing Different Oversampling Methods in Predicting Multi-Class Educational Datasets Using Machine Learning Techniques

Muhammad Arham Tariq[1], *Allah Bux Sargano*[2], *Muhammad Aksam Iftikhar*[2], *Zulfiqar Habib*[2]

[1]*University of Central Punjab, Department of Computer Science, Lahore, Pakistan*
[2]*COMSATS University Islamabad, Department of Computer Science, Lahore, Pakistan*
E-mails: arhamtariq99@gmail.com allahbux@cuilahore.edu.pk aksamiftikhar@cuilahore.edu.pk drzhabib@cuilahore.edu.pk

Abstract: *Predicting students' academic performance is a critical research area, yet imbalanced educational datasets, characterized by unequal academic-level representation, present challenges for classifiers. While prior research has addressed the imbalance in binary-class datasets, this study focuses on multi-class datasets. A comparison of ten resampling methods* (*SMOTE, Adasyn, Distance SMOTE, BorderLineSMOTE, KmeansSMOTE, SVMSMOTE, LN SMOTE, MWSMOTE, Safe Level SMOTE, and SMOTETomek*) *is conducted alongside nine classification models: K-Nearest Neighbors* (*KNN*)*, Linear Discriminant Analysis* (*LDA*)*, Quadratic Discriminant Analysis* (*QDA*)*, Support Vector Machine* (*SVM*)*, Logistic Regression* (*LR*)*, Extra Tree* (*ET*)*, Random Forest* (*RT*)*, Extreme Gradient Boosting* (*XGB*)*, and Ada Boost* (*AdaB*)*. Following a rigorous evaluation, including hyperparameter tuning and 10 fold cross-validations, KNN with SmoteTomek attains the highest accuracy of 83.7%, as demonstrated through an ablation study. These results emphasize SMOTETomek's effectiveness in mitigating class imbalance in educational datasets and highlight KNN's potential as an educational data mining classifier.*

Keywords: *Imbalance educational datasets, Students' academic performance, Educational data mining, Data re-sampling.*

1. Introduction

Educational data mining is the field of study that focuses on the development of methods for exploring and analyzing data from educational systems in order to gain insights and improve teaching and learning outcomes [1]. This includes techniques from machine learning, data mining, and statistics to uncover patterns and relationships in data from various sources, such as student assessments, learning management systems, and educational simulations for example. Educational data mining can be used to recognize students who have a high chance of falling behind so that educators can provide targeted support to help them succeed. It can also help educators understand how different teaching approaches and educational interventions affect student outcomes, which can be useful in designing more effective educational programs. Additionally, educational data mining can provide valuable insights into the effectiveness of online and blended learning environments, helping to guide the development of these educational formats. Overall, educational data mining has the potential to revolutionize the way we approach education and has the potential to have a significant impact on student outcomes and the future of education [2-4].

Imbalance distribution between classes is a common issue in educational data mining [5-8]. Imbalance classes where the distribution of classes in the data is skewed, with one or more classes having significantly fewer examples than others. This can create a number of challenges when building predictive models, as the models are often trained on imbalanced data and can lead to biased results [9, 10]. By Nature, most of the educational datasets do not contain an equal representation of every class. This tends to introduce class imbalance issues in the training phase of machine learning models. There are many approaches presented to tackle this issue. Data oversampling is one of the most appropriate solutions. In the past, only binary class-based datasets have been evaluated in depth. Keeping this thing in mind, a new approach that deals with a larger number of data oversampling algorithms comparative analysis on multi-class educational datasets can prove to be effective.

The main contribution of the paper can be understood by the fact that there are many re-sampling approaches configured in the previous decades. There were limited in-depth analyses performed for educational datasets, and most have been made for binary class-based data sets. Keeping this in mind new research is conducted that will not only use multi-class datasets, but a number of re-sampling methods evaluated using a number of classification approaches are greater than in previous research. The results of this research could provide insights into the effectiveness of different re-sampling methods in improving the prediction of student performance using machine learning algorithms. This information could be used to guide the selection of re-sampling methods for future studies in this field. Additionally, the results could help educators and researchers make more informed decisions about the use of machine-learning techniques in educational data analysis.

The paper is based on ten different Re-sampling methods comparison. The methods are SMOTE, Adayn, Distance SMOTE, BorderLineSMOTE, KmeansSMOTE, SvmSMOTE, LN SMOTE, MWSMOTE, Safe Level SMOTE, and SMOTETomek [11-20]. Each Re-sampling method is evaluated with Nine different classification techniques. The classification models are KNN, LDA, QDA, SVM, Logistic Regression, Extra Tree, Random Forest, Extreme Gradient Boosting, and Ada Boost. Section 2 is literature review, Section 3 is dataset, Section 4 is methodology, Section 5 is configured methodology, Section 6 is model evaluation methodologies, Section 7 is results, and Section 8 is conclusion.

2. Literature review

In [21] the authors have conducted a study focusing on rules to predict college students' academic overall performance and engagement in a digital gaining knowledge of surroundings. The research utilizes a balanced dataset, consisting of various features such as demographic information, learning behaviors, and interaction data. The results show that the Random Forest model outperforms other models in terms of accuracy and stability, demonstrating its effectiveness in predicting students' academic performance and engagement. The dataset used in this study includes both numerical and categorical variables, collected through online surveys and learning management systems. The data balancing approach has been applied to handle the imbalanced distribution of the target variable. In conclusion, the use of Random Forest with data balancing provides a valuable tool for educators to understand and improve students' academic performance and engagement in virtual learning environments.

The approach mentioned in [22] presents a study using machine learning techniques to predict students' academic performance. The authors describe the use of a real-world data set of students' academic performance and apply the K-Nearest Neighbor (KNN) and C4.5 algorithms to the dataset. They also use the Synthetic Minority Oversampling TEchnique (SMOTE) to balance the dataset and improve the performance of the algorithms. The authors evaluate the performance of the KNN and C4.5 algorithms with and without SMOTE balancing, and they compare the results using various evaluation metrics. They find that SMOTE balancing can significantly improve the performance of the KNN and C4.5 algorithms for predicting students' academic performance. Overall, the results of this study suggest that the use of SMOTE balancing can improve the performance of machine learning algorithms for predicting students' academic performance. The authors recommend further research to confirm these findings and to explore the potential for other techniques to improve the performance of machine learning models in this context. After comparing the performance of the Decision Tree method and the KNN Algorithm, the former exhibited better predictive accuracy, recall, and precision values, with scores of 71.09%, 71.63%, and 71.54%, respectively. In contrast, the KNN Algorithm demonstrated inferior performance in the same evaluation criteria. The authors in [23] have conducted research that aims to explore the application of these cutting-edge technologies in the innovation and reform of college English education. The use of 5G technology provides a stable and high-speed network for online education, enabling students to access high-quality English courses anytime and anywhere. The integration of AI technology, such as natural language processing and machine learning, can improve the efficiency and accuracy of English language assessment and provide personalized learning experiences for students. In addition, the integration of virtual reality and augmented reality technology can create a more immersive and interactive learning environment. This paper also includes the challenges faced in the application of these technologies, such as data privacy and security and the need for teacher training and professional development. Overall, the integration of 5G and AI technology holds great promise for the future of college English education and is worth further exploration and development. The dataset used in results generation is the Kalboard 360 LMS student's academic performance dataset.

The authors have presented a study in [24] that is based on using data mining techniques to predict students' academic performance and main behavioral features. The authors describe the use of a real-world dataset of students' academic performance and behavioral features, and they apply various data mining techniques to the dataset in order to build predictive models. The authors evaluate the performance of the different models using a range of evaluation metrics. They have found that the data mining techniques are able to achieve good performance in predicting both academic performance and behavioral features. The authors also discuss the potential applications of their approach, such as identifying students at risk of academic failure and providing targeted support to improve their performance. Overall, the results of this study suggest that data mining techniques can be effective for predicting both academic performance and behavioral features of students. The authors recommend further research to confirm these findings and to explore the potential for improving the performance of the predictive models. The study has employed six different machine learning methods to forecast student performance, including random forest, logistic regression, XGBoost, MLP, and two types of ensemble learning using bagging and voting techniques. Among these methods, random forest achieved the highest accuracy score of 77% when utilizing the top ten selected features.

The authors in [25] examine the significance of employing data re-sampling and dimensionality reduction methods in utilizing machine learning techniques for predicting the academic achievement of students. The authors first provide an overview of data re-sampling and dimensionality reduction and their importance in machine learning. They then describe the experimental setup, which involves using a real-world dataset of students' academic performance and applying various machine-learning algorithms to it. The authors evaluate the performance of the different machine learning algorithms with and without data re-sampling and dimensionality reduction, and they compare the results using various evaluation metrics, including accuracy, precision, recall, and F1 score. They have found that data re-sampling and dimensionality reduction can significantly improve the performance of the machine learning algorithms for predicting student success. Overall, the results of this study suggest that data re-sampling and dimensionality reduction are important considerations when using machine learning techniques to predict student success. The authors recommend further research to confirm these findings and to explore the potential for other techniques to improve the performance of machine learning models in this context. The best result achieved by this paper is 0.93% accuracy by using SVM and a combined approach of PCA and SMOTE.

The authors in [26] have presented an intelligent decision support system for predicting students' performance in an e-Learning environment using ensemble machine learning. The authors have developed the system using a real-world dataset of student e-Learning performance that included a variety of features, such as demographics, academic background, and e-Learning activity. The authors have applied several machine learning algorithms to the dataset and used ensemble methods to combine the predictions of these algorithms into a single, more accurate prediction. The results have shown that the ensemble approach outperforms the individual algorithms in terms of accuracy and F1 score. Overall, the results of this study suggest that an intelligent decision support system using ensemble machine learning can be an effective approach for predicting student performance in an e-Learning environment. The authors recommend further research to confirm these findings and to explore the potential for adapting the system to other learning contexts. The individual machine-learning models have been evaluated based on their F1 scores. The DT model achieved a score of 0.675, RF scored 0.777, GBT scored 0.714, NB scored 0.654, and KNN scored 0.664. The application of ensemble techniques has shown significant improvement in the overall model performance. The Final accuracy achieved was 81.95% after stacking.

In [27] the authors have conducted research aimed to compare the performance of different re-sampling methods for predicting student performance using machine learning techniques. The authors first provide an overview of the different re-sampling methods, including under-sampling, oversampling, and hybrid methods. They then describe the experimental setup, which involves using a real-world data set of students' academic performance and applying various machine-learning algorithms to it. The authors evaluate the performance of the different re-sampling methods. They find that the hybrid method (a combination of under-sampling and oversampling) performs the best, followed by oversampling and under-sampling. They also observe that the choice of re-sampling method can significantly impact the performance of the machine learning models for predicting student performance. Overall, the results of this study suggest that hybrid re-sampling methods may be the most effective approach for predicting student performance using machine learning techniques. The authors recommend further research to confirm these findings and to explore the potential for other re-sampling methods to improve the performance of machine learning models in this context. By utilizing SVM-SMOTE as a re-sampling technique, the Random forest model has delivered the most outstanding performance compared to all other models.

The authors in [28] have configured an approach for the prediction of small-sized educational datasets. The approach has been configured for a low-a-sized binary class-based dataset. A combined model based on SMOTE-IPF, CTGAN, and ENN has been developed for adding new data points. The model has achieved 97.7% accuracy by using a stacking-based classifier. The authors in the approach in [29] have utilized academic data of students by utilizing a benchmark dataset, having 1044 instances and 33 features. A model is based on the CHI-SQUARE feature selection algorithm to reduce features. After that, a Deep learning-based model named LSTM has been trained for prediction. The model has achieved 90.16% accuracy. The authors in [30] propose a study that investigates the effects of caffeine on memory consolidation in healthy young adults. Using a randomized, double-blind, placebo-controlled design, they administered caffeine or a placebo to 50 participants after they completed a memory task. The results have shown that caffeine significantly improved memory consolidation compared to placebo, as indicated by higher recall accuracy on a subsequent memory test. These findings suggest that caffeine may enhance memory consolidation and have potential implications for the use of caffeine as a cognitive enhancer. After the application of different supervised classifiers, the C4.5 classifier achieved 84% accuracy.

In the paper of [31], a new academic advising system called COHRS is introduced. COHRS is a Hybrid Recommender System that combines both Case-Based Reasoning (CBR) and Ontology. By utilizing a predefined set of rules and a knowledge base, COHRS generates recommendations based on a student's academic history and preferences. The system's ontology-based approach enables personalized recommendations based on a student's specific goals and interests, beyond their academic performance. Additionally, the case-based reasoning approach utilizes past experiences to provide tailored recommendations for each student. The study's findings demonstrate that COHRS is an effective academic advising approach that assists students in making more informed decisions about their academic paths.

The authors in [32] propose a Deep Neural Network-based (DNN-based) approach to predict student grades and provide recommendations for similar learning approaches. The approach uses a combination of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) architectures to analyze various features such as student demographics, prior academic performance, and course-specific data. To evaluate the effectiveness of the proposed approach, the study uses a dataset of over 40,000 student records from a university in Turkey. The results indicate that the DNN-based model outperforms traditional machine learning algorithms and can accurately predict student grades with an 89.7% accuracy rate. Additionally, the model can provide early warning alerts to identify students at risk of failing a course, enabling timely intervention and support. The proposed DNN-based approach has the potential to assist instructors and academic advisors in identifying students who need additional support and to tailor their teaching methods to individual student's needs. This approach may help to improve student outcomes and reduce dropout rates in higher education.

3. Dataset

The dataset [33] contains information about students who participated in an educational program. It includes 480 data points and 16 attributes. Each observation represents a student, and each column represents a characteristic or attribute of that student. It contains three classes which can be described as Low-Level, Middle-Level, and High-Level. The attributes include demographic information such as gender, nationality, and section of the class. It also includes information about the student's academic performance, including their grades in different subjects, attendance, and performance on quizzes and exams. Additionally, it includes information about the student's learning styles, such as their approach to learning, their preferred learning method, and their motivation level. Overall, this dataset could be useful for exploring relationships between academic performance and various demographic and learning style factors. It may also be useful for developing models to predict student performance or identifying areas where educational interventions may be needed. Fig 1 below represents the distribution of three classes among the dataset.

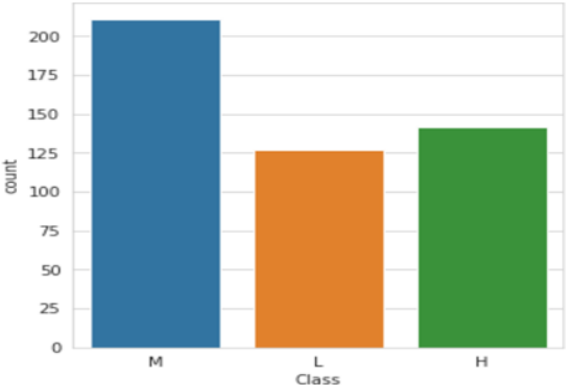

Fig. 1. Classes distribution in the dataset

4. Methodology

In this section, the paper's detailed proposed methodology is described. The Proposed methodology is divided into the following steps

- Data pre-processing;
- Data oversampling;
- Hyper-parameter tuning;
- Models evaluation.

4.1. Data pre-processing

Data pre-processing is a crucial procedure in machine learning that involves transforming raw data into a form. It includes a variety of techniques for cleaning, transforming, and preparing the data for analysis. The dataset used for the study contains no missing values or outliers. So, only raw inputs are converted into a numeric format with the help of feature encoding. Depending upon the values in the features, four different types of encoding have been adopted. These types are binary, Numeric, label, and ordinal which are included in the dataset.

4.2. Data oversampling

In this step, different chosen data oversampling algorithms have been applied respectively. Imbalanced data refers to a situation in which the number of examples in different classes of a classification problem is significantly different, such that one or more classes have very few examples compared to the others. Fig. 2 below represents the re-sampled dataset before and after the application of the data oversampling algorithm. Data oversampling is an effective technique to deal with imbalanced data. Data Oversampling algorithms are applied and configured in such a way that each algorithm is applied and then evaluated with eight different classifiers with different evaluation metrics.

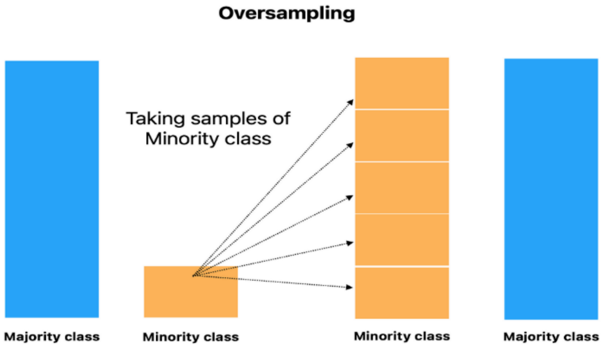

Fig. 2. Data balancing working demonstration

4.3. Hyper-parameter tuning

Hyper-parameters are parameters that are set before training a model, and they control the behavior of the learning algorithm, such as the learning rate, number of layers, and regularization strength. In this step of the study, eight different nature classifiers are evaluated with different parameter sets. Table 1 describes the optimal hyper parameters' setting for each classifier used in this research.

Table 1. Machine learning classifiers with their parameters set

No	Classifier	Parameter space
1	KNN	{1, 2, 3, ..., 20}
2	LDA	grid['solver'] = ['svd', 'lsqr', 'eigen']

3	QDA	reg param = {0.00001, 0.0001, 0.001,0.01, 0.1}
		store covariance = {True, False)},
		tol = {0.0001, 0.001,0.01, 0.1}

4	SVM	Kernel= rbf, linear, poly, C: [0.1, 0.5, 1, 2, 5]
5	Logistic regression	space['penalty'] = ['none', 'l1', 'l2', 'elasticnet'] space['C'] = [1×10^{-5}, 1×10^{-4}, 1×10^{-3}, 1×10^{-2}, 1×10^{-1}, 1, 10, 100]
6	Extra tree	grid['estimators'] = [10, 50, 100, 500,700,1000,1200]
7	Random tree	grid['estimators'] = [10, 50, 100, 500,700,1000,1200]
8	Extreme gradient boosting	grid['estimators'] = [10, 50, 100, 500,700,1000,1200] grid['learning rate'] = [0.0001, 0.001, 0.01, 0.1, 1.0]
9	Ada boost	grid['estimators'] = [10, 50, 100, 500,700,1000,1200]

5. Evaluation of models

Model evaluation is a crucial step in machine learning that helps assess the quality of a model's predictions. In essence, it involves measuring the model's performance on a set of data that has not been used during training. The evaluation process provides insights into how well the model is likely to perform in the real world and helps identify areas where the model can be improved. There are various methods for evaluating a machine learning model, but the most common ones are:

- **Train/Test split.** The dataset is split into a training set and a testing set. The model is trained on the training set, and its performance is evaluated on the testing set.

- **Cross-validation.** K-fold cross-validation is a technique utilized in machine learning to assess a model's performance. It involves dividing the available data set into k subsets of equal sizes. The model is then trained and tested *k* times, where each subset is utilized once as the testing set, and the remaining subsets are employed as the training set. By rotating through all the subsets, every data point is used for testing once, and an average measure of the model's performance is computed.
- **Evaluation metrics.** The model's performance can be measured using various evaluation metrics such as accuracy, precision, recall, F1 score, and AUC-ROC. These metrics help in assessing the model's performance on specific aspects of the data, such as identifying false positives or false negatives.

In the approach, being configured 10 fold cross-validation and 70/30 % split distribution is used to extract evaluation metrics. Cross-validation has been applied to each classifier combination with each oversampling algorithm. Four different evaluation metrics have been used for model performance evaluation. They can be described as:

the formula for accuracy –

(1) $$\text{Accuracy} = \text{TP} + \text{TN}/(\text{TP} + \text{TN} + \text{FP} + \text{FN});$$

the formula for precision –

(2) $$\text{Precision} = \text{TP}/\ (\text{TP} + \text{FP});$$

the formula for precision –

(3) $$\text{Recall} = \text{TP}/\ (\text{TP} + \text{FN});$$

the formula for the F1 score –

(4) $$\text{F1score} = 2 \ast (\text{precision} \ast \text{recall})/(\text{precision} + \text{recall}).$$

6. Results and discussion

In this section, detailed results of the paper are mentioned. Table 2 represents the accuracy of classifiers, attained with 10 fold cross-validation. Figs 4-6 represent the precision, recall, and F1 score of the classifiers by using the best parameter selected in hyper-parameter tuning and the best-performing data balancing algorithm (SMOTETomek) with an 80-20 split. Figs 3-4 are generated after the application of the best-performing data balancing algorithm (SMOTETomek) in terms of accuracy. Fig. 6 represents the confusion matrix of the classifiers generated from the chosen parameter with an 80-20 split.

Table 2. Classifiers Hyper-parameter tuning results with 10 fold cross-validation

Data balancing algorithm	KNN, %	LR, %	SVM, %	LDA, %	QDA, %	ET, %	RT, %	AdaB, %	XGB, %
SMOTE	0.751	0.797	0.744	0.787	**0.515**	0.680	0.696	0.717	0.687
ADASYN	0.745	0.795	0.741	0.783	0.519	0.658	0.713	0.664	0.682
Borderline SMOTE	0.750	0.809	0.740	0.785	0.536	0.673	0.695	0.684	0.682
KMeans SMOTE	0.744	0.820	0.750	0.809	0.554	0.695	0.705	0.717	0.715
SVM SMOTE	0.730	0.801	0.757	0.786	0.565	0.698	0.714	0.712	0.711
LN SMOTE	0.734	0.802	0.770	0.779	0.569	0.696	0.698	0.714	0.701
MWSMOTE	0.739	0.804	0.713	0.780	0.573	0.700	0.693	0.690	0.738
Safe level SMOTE	0.661	0.757	0.694	0.717	0.522	0.654	0.692	0.669	0.688
SMOTETomek	**0.837**	0.816	0.829	0.792	0.624	0.725	0.744	0.726	0.767
DistanceSMOTE	0.821	0.811	0.737	0.792	0.566	0.658	0.676	0.711	0.704

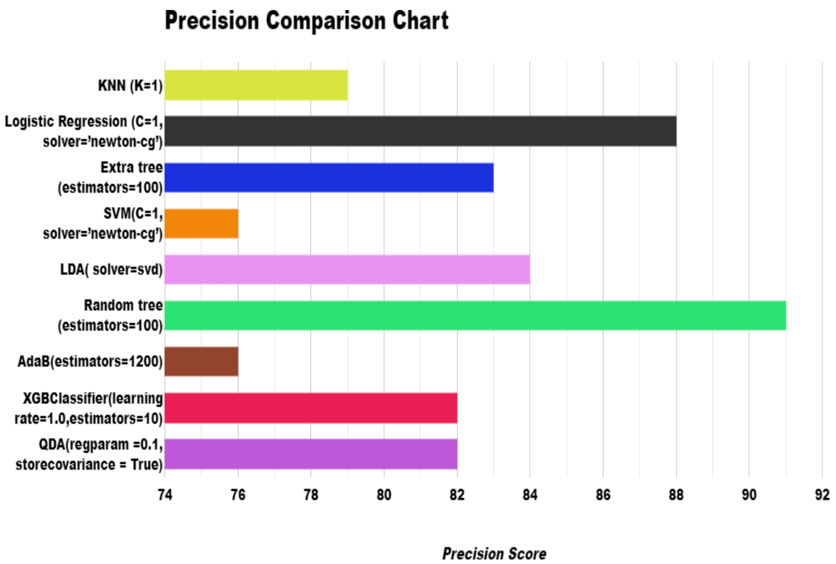

Fig. 3. Precision of different classifiers with their best parameters

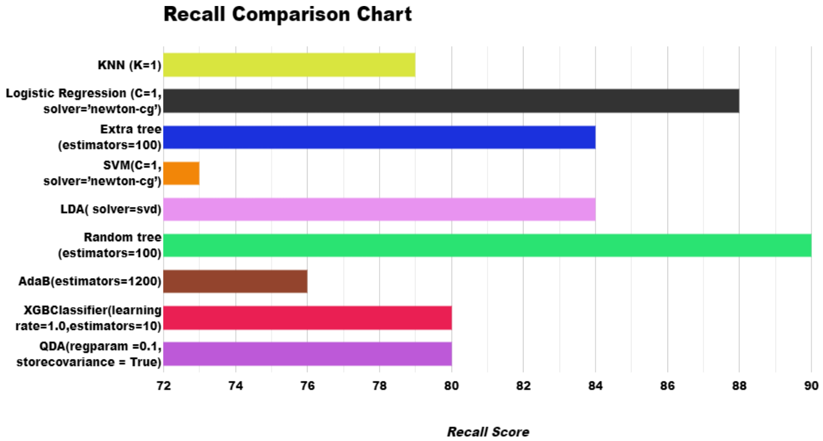

Fig. 4. Recall of different classifiers with their best parameters

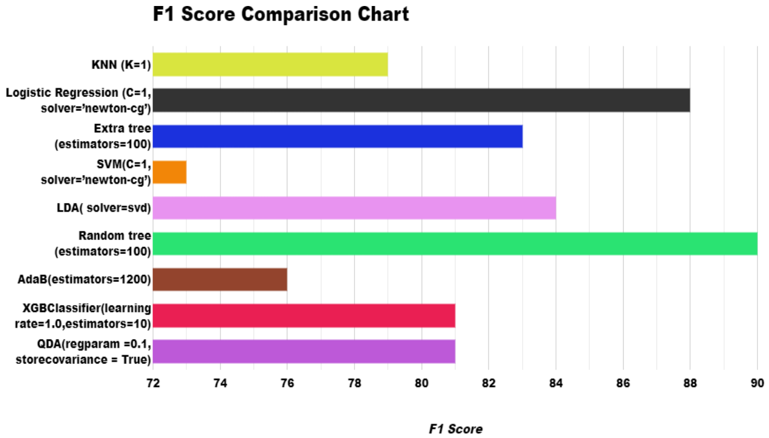

Fig. 5. F1 score of different classifiers with their best parameters

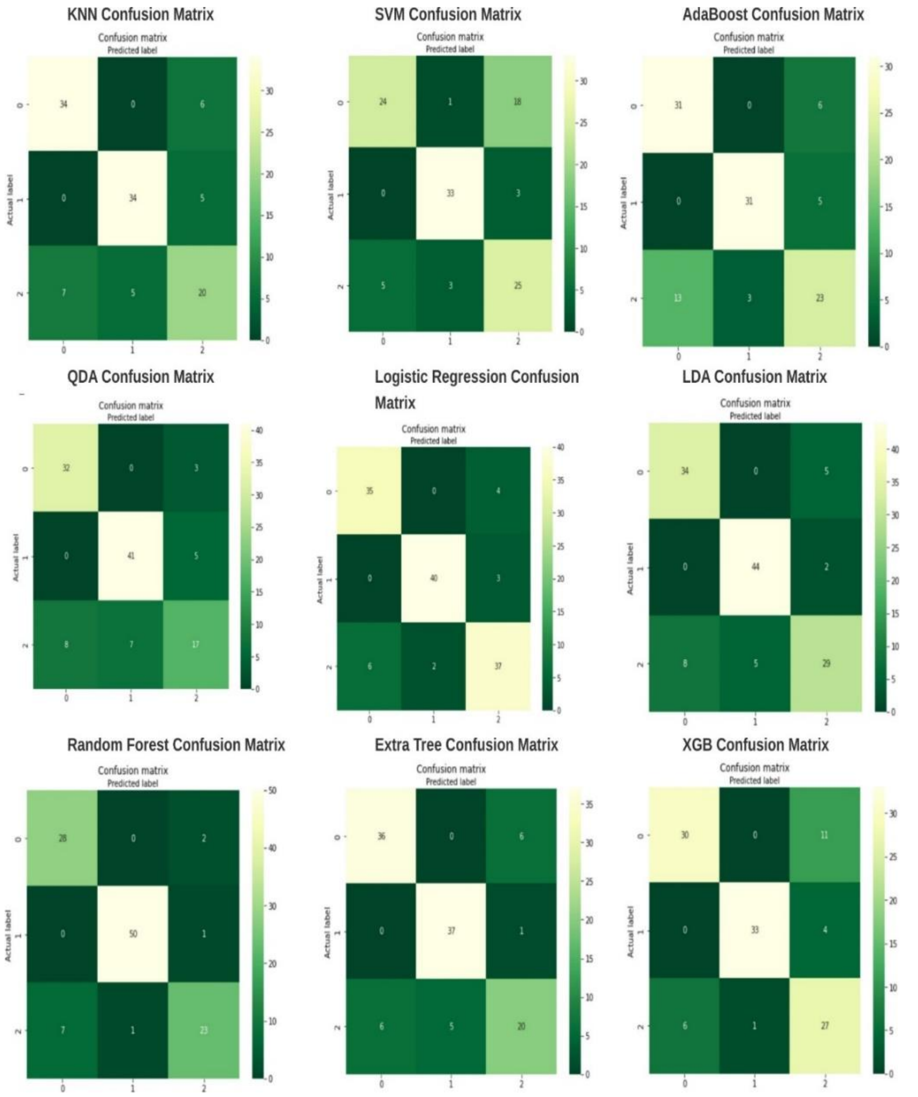

Fig. 6. Best performing classifiers confusion matrix

Keeping the above results in mind, the SMOTETomek combination with the KNN classifier performs best in terms of accuracy. These results are generated because SMOTETomek is a hybrid oversampling algorithm. It first reduces noise in the data set by applying the modified noise filtering approach Tomek and then generates required synthetic data within closed zones of actual data points. Due to SMOTETomek less noisy and balanced data points are generated near the actual data points. KNN basically works by calculating the nearest distance between data points, because SMOTETomek data points are now less overlapped with each other, which is why KNN performs well by building a more clear and more precise boundary between data points. There could be several reasons why QDA may have performed worse than other classification algorithms like KNN, SVM, and Decision tree on the educational domain dataset:

- **Assumptions of QDA not met.** QDA assumes that the data in each class are normally distributed and have different covariance matrices. If these assumptions are not met, then the performance of QDA may suffer.
- **Curse of dimensionality.** QDA can become computationally expensive when dealing with high-dimensional data, which can lead to overfitting. If the dataset has a large number of features or variables, then QDA may not perform as well as other algorithms like KNN or decision trees.

Here are some possible reasons why SMOTE might have produced worse results in the experiments

- **Incorrect assumption about the data distribution.** SMOTE assumes that the data in each class are distributed locally uniformly, which may not always be true. If the underlying distribution of the minority class is different from what SMOTE assumes, then the oversampled data may not be representative of the minority class, leading to worse results.
- **Noise sensitivity.** SMOTE can be sensitive to noise in the data, which can lead to the generation of unrealistic synthetic examples. This can reduce the quality of the oversampled data and lead to worse results.
- **Overgeneralization.** SMOTE can generate synthetic examples that are too similar to the original minority class, leading to overgeneralization and reduced diversity in the oversampled data. This can lead to worse results in some cases.

7. Conclusion and future work

The research of the paper has been conducted to study the effects of data oversampling algorithms and find the best performance among them on multi-class educational datasets. The previous approaches have had limitations in terms of the number of data-balancing algorithms considered for the study of multi-class educational datasets. To address this issue, a new approach is necessary that incorporates a wider range of data balancing algorithms to achieve better results. The study contains ten different data re-sampling methods SMOTE, Adasyn, Distance SMOTE, BorderLineSMOTE, KmeansSMOTE, SvmSMOTE, LN SMOTE, MWSMOTE, Safe Level SMOTE, SMOTETomek with the help of nine different classifications named KNN, LDA, QDA, SVM, Logistic Regression, Extra Tree, Random Forest, Extreme Gradient Boosting, Ada Boost. After hyper-parameter tuning, the best combination among them is KNN with SMOTETomek. This combination has achieved 83.72% accuracy. In the future, deep learning algorithms for evaluation and different models of GAN for data balancing can be utilized. Their combination can give many effective insights.

References

1. Kustitskaya, T. A., A. A. Kytmanov, M. V. Noskov. Early Student-at-Risk Detection by Current Learning Performance and Learning Behavior Indicators. – Cybernetics and Information Technologies, Vol. **22**, 2022, No 1, pp. 117-133. **https://doi.org/10.2478/cait-2022-0008**.

2. Atahua, A. S., J. V. Guerrero, L. Andrade-Arenas, C. M. Huerta. Data Mining: Application of Digital Marketing in Education. – Advances in Mobile Learning Educational Research, Vol. **3**, 2023, pp. 621-629.

3. Abouzinadah, E., O. Rabie, A. Bessadok. Exploring Students Digital Activities and Performances through Their Activities Logged in Learning Management System Using Educational Data Mining Approach. – Interactive Technology and Smart Education, Vol. **20**, 2023, pp. 58-72.

4. Asif, R., N. G. Haider, K. Mahboob. Quality Enhancement at Higher Education Institutions by Early Identifying Students at Risk Using Data Mining. – Mehran University Research Journal of Engineering and Technology, Vol. **42**, 2023, pp. 120-136.

5. Souza Neto, P. A., I. Silva, L. A. Guedes, T. M. Barros. Predictive Models for Imbalanced Data: A School Dropout Perspective. – Education Sciences, Vol. **9**, 2019.

6. Düsçtegör, D., E. Alyahyan. Predicting Academic Success in Higher Education: Literature Review and Best Practices. – International Journal of Educational Technology in Higher Education, Vol. **17**, 2020, pp. 1-21.

7. Lin, W. C., Y. H. Hu, G. T. Yao, C. F. Tsai. Under-Sampling Class Imbalanced Datasets by Combining Clustering Analysis and Instance Selection. – Information Sciences, Vol. **477**, 2019, pp. 47-54.

8. Kalegele, K., D. Machuve, N. Mduma. A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. – Data Science Journal, Vol. **18**, 2019, pp. 1-10.

9. Hammoud, S., F. Kamalov, Gonsalves, F. Thabtah. Data Imbalance in Classification: Experimental Evaluation. – Information Sciences, Vol. **513**, 2020, pp. 429-441.

10. Rawashdeh, J., M. Abdullah, R. Mohammed. Machine Learning with Oversampling and Under-Sampling Techniques: Overview Study and Experimental Results. – In: Proc. of 11th International Conference on Information and Communication Systems (ICICS'20), 2020, pp. 243-248.

11. Chawla, N. V., K. W. Bowyer, L. O. Hall, Kegelmeyer. SMOTE: Synthetic Minority Over-Sampling Technique. – Journal of Artificial Intelligence Research, Vol. **16**, 2002, pp. 321-357.

12. He, H., Y. Bai, E. A. Garcia, S. Li. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. – In: Proc. of IEEE International Joint Conference on Neural Networks, 2008, pp. 1322-1328.

13. Wang, W. Y., B. H. Mao, H. Han. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. – In: Proc. of International Conference on Advances in Intelligent Computing: Intelligent Computing, 2005, pp. 878-887.

14. De La Calleja, J., O. Fuentes. A Distance-Based Over-Sampling Method for Learning from Imbalanced Data Sets. – In: Proc. of 20th International Florida Artificial Intelligence, 2007, pp. 634-635.

15. Douzas, F. B. G., F. Last. Improving Imbalanced Learning through a Heuristic Oversampling Method Based on k-Means and SMOTE. – Information Sciences, 2018, pp. 1-20.

16. Zhang, Y. Q., N. V. Chawla, S. Krasser, Y. Tang. SVMS Modeling for Highly Imbalanced Classification. – IEEE Transactions on Systems, Vol. **39**, 2008, pp. 281-288.

17. Maciejewski, T., J. Stefanowski. Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data. – In: Proc. of IEEE Symposium on Computational Intelligence and Data Mining, 2011, pp. 104-111.

18. Barua, S., M. M. Islam, X. Yao, K. Murase. MWMOTE – Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning. – IEEE Transactions on Knowledge and Data Engineering, Vol. **26**, 2014, pp. 405-425.

19. Bunkhumpornpat, C., K. Sinapiromsaran, C. Lursinsap. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem. – In: Proc. of 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, 2009, pp. 475-482.

20. Prati, R. C., M. C. Monard, G. E. Batista. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. – ACM, Vol. **6**, 2004, pp. 20-29.

21. Tahir, M., K. Jawad, M. A. Shah. Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing. – Sustainability, Vol. **14**, 2022.

22. Prasetyo, W. A., A. R. Taufani, U. Pujianto. Students Academic Performance Prediction with k-Nearest Neighbor and C4.5 on Smote-Balanced Data. – In: Proc. of 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI'20), 2020, pp 348-353.

23. Kissoum, Y., A. Mouhssen, M. A. Karek, S, Mazouzi, M. L. Boughouas. Towards a Big Educational Data Analytics. – In: Proc. of International Conference on Advanced Aspects of Software Engineering (ICAASE'22), 2022, pp. 1-6.

24. Shaiba, H., M. Bezbradica, S. Almutairi. Predicting Students' Academic Performance and Main Behavioral Features Using Data Mining Techniques. – In: Proc. of 1st International Conference on Computing, in Advances in Data Science, Cyber Security and IT Applications, 2019, pp. 245-259.

25. Ajoodha, R., K. Padayachee, E. Buraimoh. Importance of Data Resampling and Dimensionality Reduction in Predicting Students' Success. – In: Proc. of International Conference on Electrical, Communication, and Computer Engineering (ICECCE'21), 2021, pp. 1-6.

26. Ullah, Z., B. Fakieh, F. Kateb, F. Saleem. Intelligent Decision Support System for Predicting Student's e-Learning Performance Using Ensemble Machine Learning. – Mathematics, Vol. **9**, 2022.

27. Ullah, Z., B. Fakieh, F. Kateb, F. Saleem. Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques. – IEEE Access, Vol. **8**, 2020, pp. 67899-67911.

28. Arham, T., Y. Niaz, A. Amin. Systematic Approach for Re-Sampling and Prediction of Low Sample Educational Datasets. – International Journal of Computing and Digital System, 2021.

29. Rahman, T., I. Khan, I. Ullah, A. Ur Rehman, M. Baz, H. Hamam, O. Cheikhrouhou, B. K. Yousafzai, S. A. Khan. Student-Performulator: Student Academic Performance Using Hybrid Deep Neural Network. – Sustainability, Vol. **13**, 2021.

30. Lin, J., J. Yu. Data Mining Technology in the Analysis of College Students' Psychological Problems. – Computer Science and Information Systems, Vol. **12**, 2022, pp. 1583-1596.

31. Lahoud, C., H. E. Khoury, P. Champin, C. Obeid. Novel Hybrid Recommender System Approach for Student Academic Advising Named Cohrs, Supported by Case-Based Reasoning and Ontology. – Computer Science and Information Systems, Vol. **19**, 2022, pp. 979-1005.

32. Sun, C., Z. Wu, J. Yang, J. Wang, T. Tao. Deep Neural Network-Based Prediction and Early Warning of Student Grades and Recommendations for Similar Learning Approaches. – Computer Science and Information Systems, Vol. **12**, 2022.

33. Hamtini, T., I. Aljarah, E. A. Amrieh. Preprocessing and Analyzing Educational Data Set Using x-Api for Improving Student's Performance. – In: Proc. of Applied Electrical Engineering and Computing Technologies (AEECT'15), 2015, pp. 1-5.

Received: 11.10.2023; Second Version:09.11.2023; Accepted:17.11.2023