SITEKIN: Jurnal Sains, Teknologi dan Industri, Vol. 22, No. 2, June 2025, pp.316 - 324
ISSN 2407-0939 print/ISSN 2721-2041 online

Enhancing Student Performance Prediction with Limited Data in Distance Learning Environments

Mulyanto Mulyanto[1*], Evizal Abdul Kadir[2]
[1,2] Department of Informatics Engineering, Universitas Islam Riau.
Jl. Kaharuddin Nasution 113, Pekanbaru, Riau, Indonesia, 28284.
Email Authors: mulyanto@eng.uir.ac.id, evizal@eng.uir.ac.id

Abstract

Providing early predictions of student performance assessments is an essential task in the educational system. Previous studies on predicting student performance assessments have traditionally relied on academic scores and test indicators. The utilization of assignments, grades, and exams has been an extensive and successful method for evaluating student performance. However, with the increasing popularity of distance learning, a new perspective has emerged. The Online Learning Management System (OLMS) provides a wide array of features that can be leveraged in various ways to predict student performance. This study aims to propose an alternative approach to predicting student performance assessments by utilizing student engagement in an online learning management system. The study strives to investigate and analyze prospective features based on student activity. Bagging ensemble learning methods are proposed to predict student performance assessments through oversampling datasets. The effectiveness of these prediction models is then compared with various machine-learning models, with the results indicating that the proposed model outperforms others at all comparison levels. Furthermore, the proposed model demonstrates the ability to discriminate and predict student performance assessments based on OLMS-related features.

Keywords: *Student Performance Assessment, Ensemble Learning, Machine Learning, Student Performance Prediction*

Introduction

Distance learning [1] is a process of learning activities that utilizes information and communication technology (ICT) as a means of online teaching and learning [2]. Distance learning introduces a new paradigm where student behavior and activities on the Online Learning Management System (OLMS) platform can be leveraged to predict student performance [3]. Additionally, student activities, such as engagement with content, learning modules, and communication within distance learning, can serve as valuable parameters for predicting student assessment [4]. However, the implementation of this strategy is challenging. Barriers, such as personal behavior, technical constraints, and financial limitations [5], can impede students from successfully adapting to distance learning. Therefore, there is a need for a proper analysis to predict the student's performance to identify potential challenges that may arise during the transition from a traditional learning system to distance learning.

In predicting student performance assessment, we observe two challenges. First, implementation is inhibited by the small volume and limited features of the dataset. Considering the limited size of the classroom, it was difficult to obtain a uniform dataset with consistent features due to the volume of the dataset. A second issue is that the existing dataset also displays uneven distribution towards the majority classes, while it is difficult to provide adequate data for the minority but important classes [6].

Considering this matter, we propose strategies for resolving the issue. In the first instance, we employ the oversampling on the limited dataset. To oversample the dataset, we utilize the synthetic minority oversampling technique (SMOTE). SMOTE has successfully demonstrated itself to produce reliable synthetic data [7]. Then, we propose an ensemble model to improve the student performance assessment prediction on the OLMS platform. Ensemble machine learning is expected to enhance the capability of a single machine-learning model and produce optimal results even when limited data is available [8].

This paper follows a structured outline. In section 2, we briefly described the related earlier studies on student performance prediction, particularly the ones related to the distance learning approach. Section 3 offered a detailed insight into our proposed methodology. The experimental design and evaluation metrics were

outlined in section 4. Section 5 detailed thorough experimental comparisons of various methods. The concluding results and challenges for future research were addressed in the final section.

Related Work

At present, the student's ability to adapt to the OLMS platform continues to be an issue and a question. An efficient system that predicts student performance is essential for helping the students who require assistance. By providing good predictions, it is expected that student failure will be detected as early as possible during the distance learning process. An overview of studies related to prediction methods will be presented. First, the sampling method implemented the use of data processing techniques to manipulate the skewed dataset. A study conducted by [9] employed the oversampling technique to address class imbalances. As a result of the use of this approach, several imbalanced datasets could be classified with improved accuracy. Similarly, in a study published in [10], the SMOTE oversampling method was applied to handle imbalanced data sets in the Internet of Things. By using SMOTE, classification problems with binary and multi-class categories could be improved in accuracy. The paper demonstrated an improvement in accuracy as well as precision, recall, and F-score for the imbalanced dataset.

Secondly, the prediction of student performance has received extensive attention in recent years. There have been numerous studies conducted on the assessment of student performance in the literature. A study conducted by [11] examined the potential attributes used to determine student performance indicators. As an algorithm, a Gradient Boosting Machine (GBM) was used for prediction. Next, the research conducted by [12], this paper assessed the key features and identified students' risks before enrolling in a course. To classify the data, several machine learning models were compared, including Decision Tree (DT), support vector machine (SVM), Random Forest (RF), and Gradient Boosting (GB). GB and DT showed a greater degree of detection accuracy than other algorithms. Furthermore, several researchers have combined multiple machine learning algorithms instead of repeating the training and reconstruction process using multiple shallow machine learning algorithms. Using ensemble learning could be a viable method for conducting the research conducted by [13]. It was found that an ensemble model proved to be more effective than other machine learning classifiers in identifying patterns and hidden knowledge in e-learning systems.

Research Methods

Research Framework

This research focuses on conducting a comprehensive study of the influence of student activity on online learning systems on student performance prediction. To accomplish our research objectives, we devised and implemented several steps within a research framework. Initially, we preprocessed the dataset to enhance readability and comprehension while retaining relevant information. This step involved employing feature engineering methods, such as data cleansing, feature encoding, and normalization, to amplify data visibility and significance for machine learning. We then applied the SMOTE technique to rectify an imbalanced dataset, synthesizing additional data from the minority class to achieve a balanced dataset and provide proportional information for more effective machine learning. Next, a bagging ensemble learning model was developed for predicting student performance. In addition, the ensemble model was utilized in order to achieve the best results by combining the basic classification algorithm. To summarize, Figure 1 displays the conceptual flow of the proposed method.

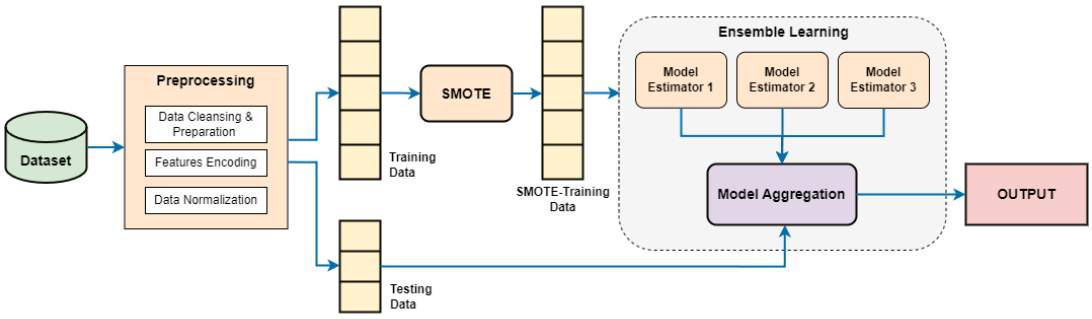

Figure 1. The proposed framework for predicting student performance assessment

Data Preprocessing

Data preprocessing is a critical step in improving dataset quality and extracting meaningful information to support effective model training. In this study, we initiated by performing data cleaning, during which we verified that there were no missing values to ensure data integrity and avoid potential bias during model training.

Following the cleaning process, feature scaling emerged as an essential preprocessing step to enhance the speed and efficiency of the machine learning algorithm. After converting all categorical features into numerical values, we applied standardization, which is particularly effective when the data approximately follow a Gaussian distribution.

Synthetic Minority Oversampling Technique (SMOTE)

The SMOTE method for oversampling was developed in 2022 and described in the paper [7]. SMOTE focuses on generating new instances based on the lines joining similar data points for minority classes. The algorithm generates artificial instances of minority classes instead of duplicating the data randomly. A detailed description of the SMOTE algorithm is shown in the *pseudo-code SMOTE*.

Algorithm SMOTE

Input: Initial $\mathcal{T}$ as number of minority class instance $\mathcal{T}$;
Amount of SMOTE percentage $\mathcal{N}$;
Set of nearest neighbours k

Process:

1. **for** i = $1, 2, \dots, \mathcal{T}$ **do**
2. Compute k nearest neighbors from a minority of class x_i
3. $\widehat{\mathcal{N}} = [\mathcal{N}/100]$
4. **while** $\widehat{\mathcal{N}} \neq 0$ **do**
5. Determine one of the k nearest neighbors, as $\bar{x}$
6. Determine random value of $\alpha = [0,1]$
7. $\hat{x} = x_i + \alpha\,(\bar{x} - x_i)$
8. Attach all values of $\hat{x}$ into $\mathcal{S}$
9. Repeat all processes until the finish, $\widehat{\mathcal{N}} = \widehat{\mathcal{N}} - 1$
10. **end while**
11. **end for**

Output: Return the result of artificial data $\mathcal{S}$

Bagging Ensemble Learning

In contrast to traditional machine learning that produces models based on a single training, ensemble learning excels at reconstructing some combinations of learners to achieve the best results [14][15]. In general, ensemble learning enhances prediction accuracy by balancing the bases that are overfitting or underfitting.

The bagging ensemble learning technique introduces the concept of bootstrap aggregation. This method involves bootstrapping the original training dataset to train several classifiers [16]. In order to train each classifier, instances from the original dataset are indiscriminately substituted for instances in a subset dataset. Finally, the algorithm applies the learning problems to every classifier, and the result is inferred based on a majority or weighted vote. The general algorithm of bagging ensemble learning is mentioned in the pseudo-code *BaggingEnsembelLearning*.

Algorithm BaggingEnsembelLearning

Input: Initial dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i)\}$
Initial base learning algorithm as $\mathcal{L}$;
Setup the number of learning rounds as $\mathcal{T}$

Process:

1. **for** $t = 1, 2, \dots, \mathcal{T}$
2. Generate bootstrap from $\mathcal{D}$: $\mathcal{D}_t = \text{Bootstrap}(\mathcal{D})$
3. Train individual learner h_t: $h_t = \mathcal{L}(\mathcal{D}_t)$
4. **End for**

Output: $H_{(x)} = \arg\max_{y} \sum_{t=1}^{T} \Pi\,(y = h_t\,(x))$

Metrics Evaluation

To evaluate the machine learning performance, we employed widely used metrics including accuracy, precision, recall, and F-score. We acquire accuracy, precision, recall, and f-score the following metrics:

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

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

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

$$f\text{-}score = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

However, predicting imbalanced data poses a challenging problem. The accuracy metrics generally can be used to explore the overall performance of the predictor, but accuracy is less than adequate to understand the more specific performance for each class [17]. To address this limitation, we utilized the precision-recall curve (PR-curve) to visualize the ensemble learning model [18]. The precision-recall curve mapped recall (R) on the x-axis and precision (P) on the y-axis.

Experimental Design

Environment

The experiment was carried out on a computer with GPU Nvidia GeForce RTX-2080 11 GB in the Intel® Xeon® CPU specifications E5-2630 and 128 GB of DDR4 RAM. Machine learning used TensorFlow 2.0 and Keras framework on Ubuntu 18.04.

Dataset Description

In this study, we utilized a secondary dataset originally collected at North American University, which documents the learning activities of 2nd-year undergraduate students enrolled in online courses through a Learning Management System (LMS) [19]. This dataset was not collected by the authors of the present study but was obtained from previously published research by Moubayed et al. [19]. Although subsequent studies such as [20] have also employed this dataset, the primary source remains the original work in [19]. In order to apply this framework, the selection features will be based on 50% coursework delivery. The detailed feature selection is presented in Table 1, and Figure 2 shows the probability of features for each class.

Table 1. Feature selection for 50% coursework delivery

No.	Feature Name	Data Type	No.	Feature Name	Data Type
1.	Logins	Discrete	10.	Assign_2_dur	Discrete
2.	ContentReads	Discrete	11.	Assign_3_dur	Discrete
3.	ForumReads	Discrete	12.	Avg_time_submit	Discrete
4.	ForumPosts	Categorical	13.	Quiz_01	Discrete
5.	QuizReviews	Categorical	14.	Assignment_01	Discrete
6.	Assign_1_lateness	Discrete	15.	Midterm_Exam	Discrete
7.	Assign_2_lateness	Discrete	16.	Assignment_02	Discrete
8.	Assign_3_lateness	Discrete	17.	Class	Categorical
9.	Assign_1_dur	Discrete			

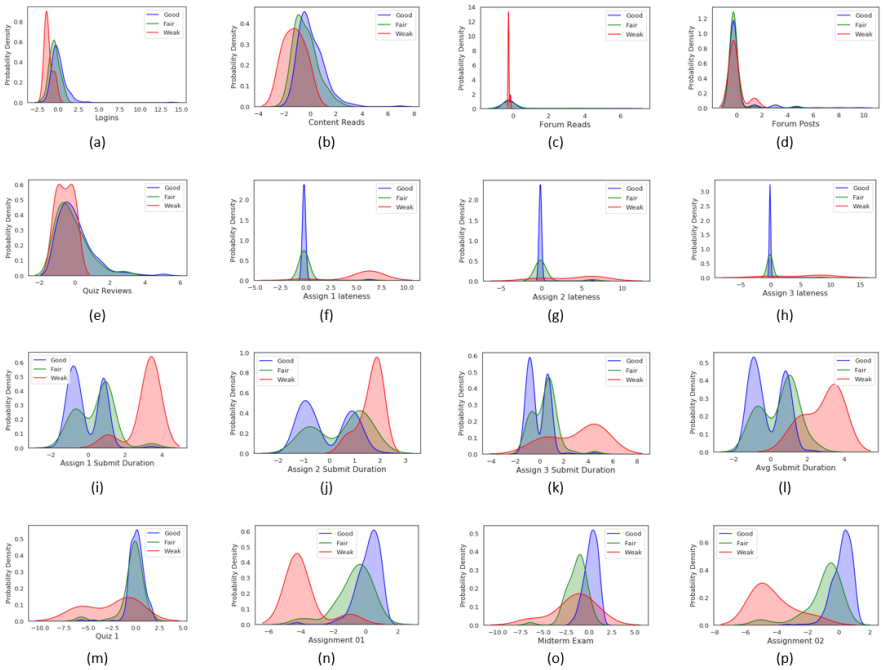

Figure 2. Visualization of the features probability in multi- class: Good, Fair and Weak, (a) Logins, (b) ContentReads, (c) ForumReads, (d) ForumPosts, (e) QuizReviews, (f) Assign_1_lateness, (g) Assign_2_lateness, (h) Assign_3_lateness, (i) Assign_1_dur, (j) Assign_2_dur, (k) Assign_3_dur, (l) Avg_time_submit, (n) Quiz_01, (m) Assignment_01, (o) Midterm_Exam, (p) Assignment_02.

Synthetizing Dataset using SMOTE

Similar to many real-world datasets, the student performance assessment dataset exhibited a class imbalance. To address this, we adopted a proportional oversampling strategy using SMOTE, in which 100% synthetic samples were generated for each minority class. Rather than aiming for a fully balanced (1:1 ratio) distribution, this approach aims to increase minority classes' representation proportionally. This strategy was adopted to improve the model's ability to recognize patterns in underrepresented classes while minimizing overfitting risk and maintaining generalization. Table 2 presents a comparison between the original training data and the dataset after applying the proportional SMOTE oversampling strategy.

Table 2. Comparison of student performance assessment dataset: original training data and after preprocessing using SMOTE

Dataset	Good	Weak	Fair
Original Data Training	319	**40**	**5**
SMOTE Data Training	319	**80**	**10**
Testing Data	100	19	3

Result and Discussion

In this section, we elucidate and compare the prediction results of student performance assessments. The evaluation of the performance of the proposed method will be presented in three parts: (i) a discussion on the impact of SMOTE on the performance of ensemble learning; (ii) a general comparison between the proposed ensemble method and related work; and (iii) a explanative observation of the effectiveness of online learning features on student performance assessments.

SMOTE Impact Performance Comparison

While ensemble learning offers notable improvements, it does not directly reveal whether SMOTE or ensemble learning significantly influences the performance results. To discern the impact of SMOTE on the classification results, we conducted training in bagging ensemble learning using pristine datasets. Ensemble learning was trained with the original training data and compared to the SMOTE training data. In comparison

to the ensemble model, SMOTE ensemble learning yielded superior results for most metrics. The comparisons of the ensemble model at each stage are presented in Table 3.

Table 3. Results of ensemble learning performance using pristine and SMOTE oversampling data

Matrics	**Bagging** [20]	**SMOTE-Bagging**
Accuracy	0.9180	**0.9262**
Precision	0.9148	**0.9283**
Recall	0.9180	**0.9262**
F-score	0.9108	**0.9264**

Performance Evaluation of Ensemble Learning

To achieve the research aim, we juxtaposed the performance of SMOTE ensemble learning with other single machine learning algorithms, including DT, SVM, k-Nearest Neighbor (KNN), and Gaussian naïve Bayes (GNB). Overall, the proposed ensemble model surpassed the single machine learning models in all matrices. SMOTE-Bagging exhibited superior performance compared to other algorithms, with its accuracy surpassing other machine learning models (DT: 0.8852, SVM: 0.9016, k-NN: 0.8770, GNB: 0.8689, Bagging: 0.9180) by 0.9262. Table 4 details the comparison of accuracy, precision, recall, and f-score between the SMOTE ensemble algorithm and the single machine learning models, and Figure 3 shows a visual comparison of every model's performance.

Table 4. Performance comparison of SMOTE ensemble learning versus single machine learning on 50% student activity

Model	Accuracy	Prec.	Recall	F-score
DT [12][21]	0.8852	0.8228	0.8022	0.8771
SVM [12][21]	0.9016	0.7730	0.7730	0.8857
K-NN [22]	0.8770	0.7742	0.7648	0.8544
GNB [23]	0.8689	0.7627	0.7905	0.8580
Bagging [20]	0.9180	0.9148	0.9180	0.9108
SMOTE-Bagging [our]	**0.9262**	**0.9283**	**0.9262**	**0.9264**

Evaluation of Classification Online Learning Feature

This section focuses on determining how well the features in the student activity on OLSM are used to predict student performance assessments. To address this question, we employed the PR-curve. The PR-curve revealed that the bagging ensemble's performance was relatively better than other models, particularly showcasing impressive results in the minority of classes. The SMOTE-bagging ensemble learning, in particular, achieved outstanding results with values of 0.9938, 0.8568, and 1.0 for the Good, Fair, and Weak classes, respectively. However, drawing conclusive insights remains challenging due to the scarcity of available testing data, leading to hesitations in providing a comprehensive result justification. PR-curve comparisons at each stage are detailed in Figure 4.

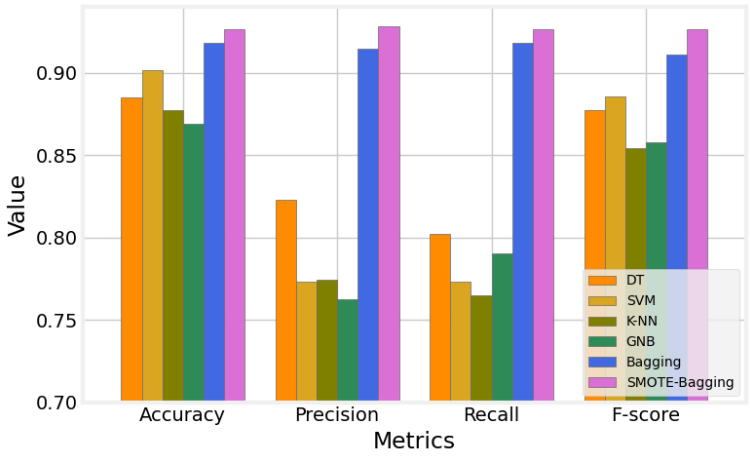

Figure 3. Performance comparison of SMOTE-Bagging and benchmark models

Evaluation of Classification Online Learning Feature

This section focuses on determining how well the features in the student activity on OLSM are used to predict student performance assessments. To address this question, we employed the PR-curve. The PR-curve revealed that the bagging ensemble's performance was relatively better than other models, particularly showcasing impressive results in the minority of classes. The SMOTE-bagging ensemble learning, in particular, achieved outstanding results with values of 0.9938, 0.8568, and 1.0 for the Good, Fair, and Weak classes, respectively. However, drawing conclusive insights remains challenging due to the scarcity of available testing data, leading to hesitations in providing a comprehensive result justification. PR-curve comparisons at each stage are detailed in Figure 4.

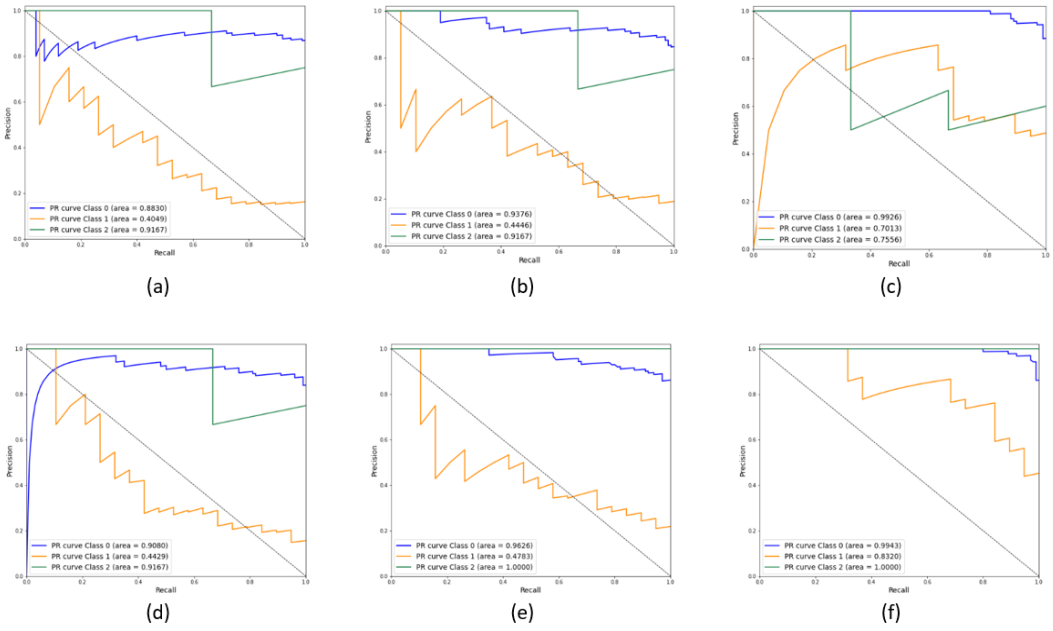

Figure 4. The PR curve SMOTE-boosting ensemble on multi-class classification (a) student activity stage, (b) 20% feature stage, (c) 50% feature stage, and the PR curve SMOTE-bagging ensemble on multi-class classification (d) student activity stage, (e) 20% feature stage, (f) 50% feature stage.

Conclusion

As a result of the findings of this study, distance learning with an OLMS is capable of providing features that can be used to predict a student's performance assessment. It is possible to predict student learning success by using the features available in the system, such as student activities or learning assessments. The combination of SMOTE and ensemble learning has proven effective in enhancing the prediction of student performance assessments compared to shallow machine learning. The ensemble model significantly improved overall classification performance, while SMOTE demonstrated its ability to offer more detailed attention to minority classes. Nevertheless, the primary challenge lies in the limitations of the dataset. This hinders an objective representation of how features in the online learning platform affect student performance assessment prediction. The limited data in this dataset may not fully capture the real-world scenario's nuances. Further research can explore additional features of student activity on OLSM. Distance learning offers a wealth of features and attributes for assessment, many of which remain untapped. Attributes associated with learning content, such as content type, length, number of comments, and number of likes, hold potential significance for future analysis. Meanwhile, deep learning stands out among various machine learning approaches for predicting student performance.

References

[1] F. V. Ferraro, F. I. Ambra, L. Aruta, and M. L. Iavarone, “Distance learning in the covid-19 era: Perceptions in Southern Italy,” *Educ. Sci.*, vol. 10, no. 12, pp. 1–10, 2020, doi: 10.3390/educsci10120355.

[2] A. Bozkurt, “From Distance Education to Open and Distance Learning,” in *Handbook of Research on Learning in the Age of Transhumanism*, no. April, 2019, pp. 252–273. doi: 10.4018/978-1-5225-8431-5.ch016.

[3] A. Moubayed, M. Injadat, A. B. Nassif, H. Lutfiyya, and A. Shami, “E-Learning: Challenges and Research Opportunities Using Machine Learning Data Analytics,” *IEEE Access*, vol. 6, pp. 39117–39138, 2018, doi: 10.1109/ACCESS.2018.2851790.

[4] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, “Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS,” *IEEE Trans. Learn. Technol.*, vol. 10, no. 1, pp. 17–29, 2017, doi: 10.1109/TLT.2016.2616312.

[5] S. Abuhammad, “Barriers to distance learning during the COVID-19 outbreak: A qualitative review from parents’ perspective,” *Heliyon*, vol. 6, no. 11, p. e05482, 2020, doi: 10.1016/j.heliyon.2020.e05482.

[6] K. T. Chui, R. W. Liu, M. Zhao, and P. Ordóñez de Pablos, “Predicting Students’ Performance with School and Family Tutoring Using Generative Adversarial Network-Based Deep Support Vector Machine,” *IEEE Access*, vol. 8, pp. 86745–86752, 2020, doi: 10.1109/ACCESS.2020.2992869.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *J. Artif. Intell. Res.*, vol. 16, no. January, pp. 321–357, 2002, doi: 10.1613/jair.953.

[8] I. Alazzam, I. Alsmadi, and M. Akour, “Software fault proneness prediction: a comparative study between bagging, boosting, and stacking ensemble and base learner methods,” *Int. J. Data Anal. Tech. Strateg.*, vol. 9, no. 1, p. 1, 2017, doi: 10.1504/ijdats.2017.10003991.

[9] A. Amin *et al.*, “Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study,” *IEEE Access*, vol. 4, no. Ml, pp. 7940–7957, 2016, doi: 10.1109/ACCESS.2016.2619719.

[10] H. Zhang, L. Huang, C. Q. Wu, and Z. Li, “An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset,” *Comput. Networks*, vol. 177, no. May, 2020, doi: 10.1016/j.comnet.2020.107315.

[11] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, “Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil,” *J. Bus. Res.*, vol. 94, no. August 2017, pp. 335–343, 2019, doi: 10.1016/j.jbusres.2018.02.012.

[12] A. Polyzou and G. Karypis, “Feature Extraction for Next-Term Prediction of Poor Student Performance,” *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 237–248, 2019, doi: 10.1109/TLT.2019.2913358.

[13] A. Kumar and M. Jain, *Ensemble Learning for AI Developers*. 2020. doi: 10.1007/978-1-4842-5940-5.

[14] Zhi-Hua Zhou, *Ensemble Methods, Foundations and Algorithms*. 2012.

[15] J. Beemer, K. Spoon, L. He, J. Fan, and R. A. Levine, "Ensemble Learning for Estimating Individualized Treatment Effects in Student Success Studies," *Int. J. Artif. Intell. Educ.*, vol. 28, no. 3, pp. 315–335, 2018, doi: 10.1007/s40593-017-0148-x.

[16] G. Fumera, F. Roli, and A. Serrau, "A theoretical analysis of bagging as a linear combination of classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1293–1299, 2008, doi: 10.1109/TPAMI.2008.30.

[17] M. Mulyanto, S. W. Prakosa, M. Faisal, and J.-S. Leu, "Using Optimized Focal Loss for Imbalanced Dataset on Network Intrusion Detection System," in *IEEE Vehicular Technology Conference*, 2022. doi: 10.1109/VTC2022-Spring54318.2022.9861034.

[18] J. Davis and M. Goadrich, "The Relationship between Precision-Recall and ROC Curves," in *Proceedings of the 23rd International Conference on Machine Learning*, in ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[19] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Relationship between student engagement and performance in e-learning environment using association rules," *EDUNINE 2018 - 2nd IEEE World Eng. Educ. Conf. Role Prof. Assoc. Contemp. Eng. Careers, Proc.*, 2018, doi: 10.1109/EDUNINE.2018.8451005.

[20] M. N. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-split optimized bagging ensemble model selection for multi-class educational data mining," *Appl. Intell.*, vol. 50, pp. 4506–4528, 2020.

[21] X. Xu, J. Wang, H. Peng, and R. Wu, "Prediction of academic performance associated with internet usage behaviors using machine learning algorithms," *Comput. Human Behav.*, vol. 98, no. April, pp. 166–173, 2019, doi: 10.1016/j.chb.2019.04.015.

[22] I. A. A. Amra and A. Maghari, "Students Performance Prediction Using KNN and Naïve Bayesian," *Int. Conf. Inf. Technol.*, pp. 909–913, 2017.

[23] M. Masud, J. Gao, L. Khan, J. Han, and B. M. Thuraisingham, "Classification and novel class detection in concept-drifting data streams under time constraints," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 6, pp. 859–874, 2011, doi: 10.1109/TKDE.2010.61.