Comparative Analysis of Oversampling and Undersampling Techniques in Predicting Customer Churn for Dqlab Telco

Bima Pramudya Asaddulloh
Department of Informatics, Postgraduate Program, Universitas Amikom Yogyakarta, Sleman, 55283, Indonesia
bima@students.amikom.ac.id

Kusrini
Department of Informatics, Postgraduate Program, Universitas Amikom Yogyakarta, Sleman, 55283, Indonesia
kusrini@amikom.ac.id (corresponding author)

Dhani Ariatmanto
Department of Informatics, Postgraduate Program, Universitas Amikom Yogyakarta, Sleman, 55283, Indonesia
dhaniari@amikom.ac.id

Received: 1 February 2025 | Revised: 6 March 2025 | Accepted: 9 March 2025

 DOI: https://doi.org/10.48084/etasr.10396

ABSTRACT

Customer churn prediction is a critical task in the telecommunications (telecom) industry for optimizing retention efforts and reducing customer attrition. This paper presents a churn prediction model using Machine Learning (ML) techniques, focusing on handling imbalanced data through resampling methods. A novel approach is proposed combining Gradient Boosting (GB) with Random Undersampling (RUS), (GB+RUS), and Random Forest (RF) with Synthetic Minority Oversampling Technique (SMOTE). Model performance is evaluated on a real-world telecom dataset, achieving significant results. The RF+SMOTE method outperforms existing models, obtaining an accuracy of 79.23%, precision of 79.32%, recall of 80.15%, F1-score of 79.73%, and AUC of 87.25%, outperforming traditional approaches, such as RF and Support Vector Machines (SVM). The importance of using advanced resampling techniques to address data imbalance and improve churn prediction models is highlighted.

Keywords-machine learning; customer churn classification; resampling method

I. INTRODUCTION

Predicting customer turnover is a vital component of corporate strategies within the telecom sector [1-4]. As rivalry in telecom business escalates, customer retention becomes increasingly vital compared to new client acquisition [5, 6]. Forecasting customer attrition allows organizations to formulate focused retention strategies, ultimately reducing client loss and enhancing profitability [7]. In this regard, numerous ML algorithms have been utilized to forecast churn; nonetheless, managing imbalanced datasets continues to provide a considerable issue [8].

Previous churn classification models, like RF, Gradient-boosted trees (GBT), SVM, and Decision Tree (DT) have failed to overcome class imbalance. Oversampling and undersampling methods are frequently employed to mitigate this issue in churn prediction tasks; however their efficacy differs among various models and datasets [12]. The current study seeks to assess and compare the effectiveness of oversampling and undersampling methods in forecasting customer attrition for Dqlab Telco. The employed dataset encompasses a diverse array of client factors, including demographic data, subscribed services, and account information. A variety of ML algorithms, such as RF, K-Nearest Neighbors (KNN), GB, and DT, were trained on the dataset deploying both undersampling and oversampling techniques. The primary aim of this research is to detect the most efficacious resampling approaches to enhance the aforementioned models' predictive accuracy, specifically regarding customer churn prediction. Authors in [9] introduced a churn prediction model for the telecom industry based on feature extraction. RF and Logistic Regression (LR) were employed to determine the most pertinent features, resulting in

a classification accuracy of 62.2% and 58.8%, respectively. It was observed that RF exhibited superior differentiation between churners and non-churners. The dataset was diminished due to absent values and extraneous variables, enhancing model efficacy. Authors in [11] introduced a churn detection model employing Essence RF to address redundancy in detailed mobile data. The model was assessed using data from a European telecom firm, revealing superior accuracy and expedited convergence relative to conventional RF, with an AUC performance of 62.56%. The research stressed the significance of combining web browsing, application utilization, and geospatial mobility data for enhanced churn detection. Authors in [7] introduced a churn prediction model employing Automated Machine Learning (AutoML) in conjunction with Social Network Analysis (SNA) for the telecom sector. The model was assessed using actual data from telecom operators in Bosnia and Herzegovina, resulting in higher precision with SNA than with conventional models. The research indicated that the GBT algorithm utilizing SNA characteristics attained optimal performance, achieving a precision of 76.00%. Authors in [13] presented a sampling technique utilizing an enhanced C4.5 DT to forecast telecommunication customer attrition. The approach was assessed utilizing UCI public data and telecom operator data, resulting in an accuracy of 58.36% on significantly skewed data. The research indicated that the proposed approach proficiently managed unbalanced data and enhanced predictive accuracy. Authors in [10] carried out a study on churn prediction within the telecom sector utilizing diverse ML methodologies. The models were assessed using data from a European mobile operator, attaining a classification accuracy of 73.1% when combining SVM with Radial Basis Function (RBF) and RF. It was noted that the proposed methodologies accurately forecasted client attrition and offered insights for its mitigation.

It should be noted that, despite proven efficient, the models investigated in [7, 9-11, 13] demonstrated complexity, which rendered them challenging to analyze, and hence significant processing resources were required. Moreover, model performance decreased when the former were implemented in more intricate datasets, stressing the need for improved preparation methods. Consequently, a more balanced, interpretable, and computationally efficient methodology for churn prediction is required.

Although prior studies have investigated numerous churn prediction ML models, they have not rigorously examined resampling strategy impact on them. Furthermore, the existing literature has not provided a thorough comparison of oversampling and undersampling techniques in the telecom churn prediction field. The current study attempts to address these issues by systematically evaluating several resampling strategies, emphasizing their effects on model performance, interpretability, and computing economy. The results promote the creation of more efficient churn prediction systems customized for actual telecom datasets. The proposed method seeks to enhance predictive performance and provide actionable insights for churn management in the telecom sector by addressing these deficiencies.

II. PROPOSED METHOD

This section provides a comprehensive description of the methodology employed for data analysis and modeling. The followed procedure is segmented into four principal phases: Data Loading and Exploratory Data Analysis (EDA), Data Preprocessing, Data Modeling, and Evaluation. Each stage is detailed, elucidating the processes and reasoning of the undertaken actions, as shown in Figure 1.

A. Dataset

This research utilizes the Telco Customer Churn dataset [14], designed to facilitate the creation of targeted customer retention initiatives. The former offers an extensive overview of customer behavior and attributes to forecast churn, enabling firms to examine patterns and implement preventive strategies. Each row signifies a distinct customer, whilst the columns delineate aspects pertaining to services, billing, and demographics. The dataset contains information on customer churn within the past month, types of subscribed services (phone, internet, and streaming TV), account details (tenure, monthly charges, and total charges), and demographic characteristics (gender, senior citizen status, and the presence of dependents or partners). Table I delineates the structure and attributes of the employed dataset. The dataset contains 7,049 categorical and numerical entries requiring pre-processing and analysis. The binary variable "Churn" signifies if a customer has departed. "Monthly charges" and "Total charges" yield consistent numerical data, whilst "Gender" and "Internet service" offer categorical insights. The dataset's varied data kinds render it suitable and reliable for many ML methodologies and classification models.

TABLE I. DATASET SAMPLE

No	Class	Gender	Partner	…	Tenure	Phone service	Monthly charges	Senior citizen
1	No	Female	Yes	…	1	No	29.85	No
2	No	Male	Yes	…	60	Yes	20.50	No
3	Yes	Male	No	…	5	Yes	104.10	No
4	No	Female	Yes	…	72	Yes	115.50	No
5	No	Female	Yes	…	56	Yes	81.25	No
…	…	…	…	…	…	…	…	…
319790	No	Male	No	…	1	Yes	44.75	No
319791	Yes	Female	No	…	1	Yes	70.15	Yes
319792	Yes	Female	No	…	1	Yes	85.55	No
319793	No	Female	Yes	…	72	Yes	117.15	No
319794	No	Male	No	…	64	Yes	99.25	No

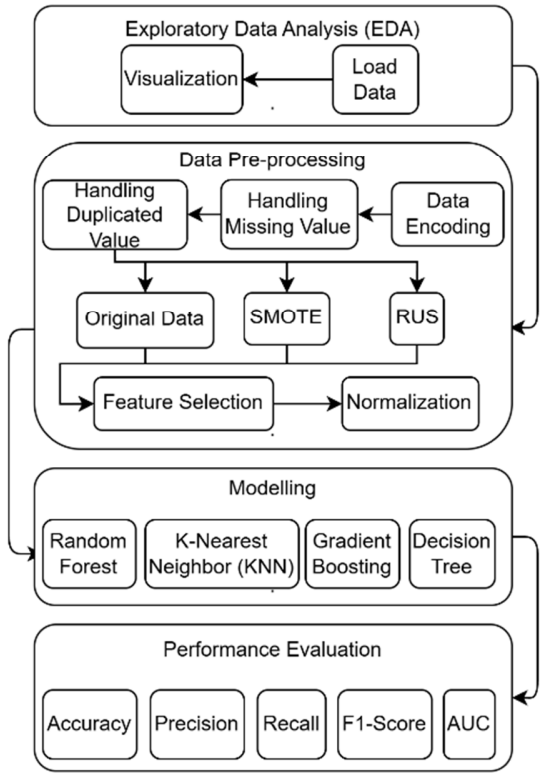

Fig. 1. Research flow diagram.

B. Exploratory Data Analysis

Histograms were constructed to elucidate the numerical properties, uncovering patterns, variability, and possible outliers in the data [15]. The class distribution of the target variable, Churn, was analyzed using a count plot and a pie chart, which offered a clear depiction of class proportions, as depicted in Figure 2.

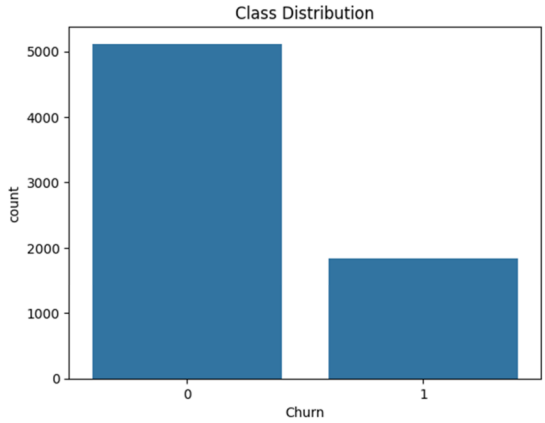

Fig. 2. Target variable distribution.

C. Data Pre-Processing

Data preprocessing commenced with label encoding, wherein categorical variables were converted into numerical representations via LabelEncoder, which allows ML models to understand and process categorical variables numerically, thereby improving prediction accuracy without losing important information [16]. Missing values were replaced with the average value of each column to maintain the distribution of the data without changing the general trend of the features, reduce potential distortions in the analysis, and ensure that the ML model can still utilize the entire dataset without losing significant information [17]. Duplicated data were removed to prevent bias in analysis, improve computational efficiency, and certify that the ML models are not affected by repetitive information that may cause overfitting or distortion in pattern learning [18]. To rectify class imbalance in the target variable, RUS was employed to diminish the majority class size, while SMOTE was utilized to provide synthetic samples for the minority class, and thus ascertain equitable representation across both classes and improve model generalization without causing overfitting or important information loss [12].

A correlation matrix was subsequently generated to examine the correlations between characteristics and the target variable, facilitating the selection of the top five features with the greatest predictive ability [19]. Implementing feature selection enhances model performance by eliminating irrelevant or redundant features, thus decreasing complexity, augmenting interpretability, and mitigating overfitting [20]. The selected features were normalized with StandardScaler to achieve a zero mean value and unit variance. This is essential for equalizing the scale of features, preventing the dominance of features with a large range of values, maximizing the performance and convergence of distance-based models, and improving the stability of ML algorithms [21].

D. Modeling

The preprocessed dataset was split into training and testing sets using an 80-20 split [22]. Four ML models, namely RF, KNN, GB, and DT, were configured with default parameters and trained on the training set to identify patterns and relationships in the data [11, 23-25]. After training, predictions were generated for the testing set, providing a basis for evaluating model performance and their ability to generalize to new, unseen instances.

E. Performance Evaluation

Confusion matrix, accuracy, precision, recall, and F1-score, were the considered performance metrics for the trained model assessment [26]. Accuracy quantifies the ratio of right predictions, whereas precision and recall assess the ratio of genuine positives among predicted positives and real positives, respectively. The F1-score equilibrates precision and recall. The confusion matrix categorizes predictions into True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) to illustrate model errors. The models' discriminative capacity is evaluated by ROC curves and AUC scores, which graph the TP rate versus the FP rate at different thresholds. They also measure overall model efficacy, with elevated AUC values signifying superior class differentiation.

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

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

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

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision+Recall}} \quad (4)$$

$$\text{AUC} = \int_0^1 \text{ROC}(t)\ dt \quad (5)$$

III. RESULTS AND DISCUSSION

The outcomes for of model (RF, KNN, GB, and DT) were examined, with specific focus having been placed on the effects of data balancing methods (undersampling and oversampling). The following section compares model performance, examines model strengths and limitations, and emphasizes crucial insights obtained from the review.

A. Model Performance Comparison

Table II displays the performance metrics of the four classifiers assessed on datasets comprising standard, undersampled, and oversampled data. RF outperformed the other models across all measures, attaining the greatest accuracy, precision, recall, F1-score, and AUC, especially in the oversampled context, where it achieved an AUC of 87.25. KNN exhibited subpar performance across all parameters, with a notable decline in precision and recall, particularly in the standard dataset, where it attained an AUC of merely 50.38. GB demonstrated robust performance, especially in the oversampled scenario, with an AUC of 86.70 and enhanced recall and F1-score relative to the reference dataset. DT displayed adequate performance, reaching an AUC of 71.76 in the oversampled scenario. However, it underperformed in comparison to RF and GB across most other metrics. The data resampling strategies significantly enhanced model performance, with oversampling yielding the most favorable outcomes for the majority of classifiers, particularly RF and GB.

TABLE II. PERFORMANCE METRICS OF MODELS TRAINED ON DATASET WITH VARIOUS RESAMPLING METHODS

Classifier	**Standard**					**Undersampled**					**Oversampled**				
	Acc	**Prec**	**Rec**	**F1**	**AUC**	***Acc***	***Prec***	***Rec***	***F1***	***AUC***	***Acc***	***Prec***	***Rec***	***F1***	***AUC***
RF	78.13	61.65	44.81	51.90	80.44	71.70	66.39	73.31	69.68	78.34	**79.23**	**79.32**	**80.15**	**79.73**	**87.25**
KNN	68.70	27.15	11.20	15.86	50.38	50.20	44.92	54.29	49.17	51.62	61.68	61.98	64.24	63.09	63.62
GB	**79.21**	**64.86**	**45.90**	**53.76**	**82.23**	**74.01**	**68.29**	**77.30**	**72.52**	**81.04**	78.15	76.99	81.50	79.18	86.70
DT	70.36	43.35	40.98	42.13	60.92	67.62	62.43	67.79	65.00	67.64	71.75	72.68	71.43	72.05	71.76

B. Comparison between the Proposed and Existing Methods

To evaluate the effectiveness of the proposed technique, its performance was compared against many established churn prediction models from the literature in terms of critical performance measures, including accuracy, precision, recall, F1-score, and AUC, as portrayed in Table III.

TABLE III. COMPARISON BETWEEN THE PROPOSED AND THE EXISTING METHODS IN CUSTOMER CHURN CLASSIFICATIONS

Scheme	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
[9]	RF	62.20	60.38	65.54	62.85	66.00
[11]	ERF	**-**	-	-	**-**	62.56%
[7]	GBT+SNA	-	76.00	68.89	72.31	-
[13]	C4.5+DT	58.36	-	**-**	-	-
[10]	SVM rbf + RF	73.1	50.00	64.00	-	-
Proposed	GB+RUS	74.01	68.29	77.30	72.52	81.04
	RF+SMOTE	**79.23**	**79.32**	**80.15**	**79.73**	**87.25**

Table III shows that the proposed approach outperformed the existing schemes. The RF+SMOTE combination achieved the highest accuracy (79.23%), precision (79.32%), recall (80.15%), F1-score (79.73%), and AUC (87.25%) among all the methods being assessed. Combining ensemble learning (RF and GB) with resampling techniques (SMOTE and RUS), effectively handles class imbalance in customer churn classification. RF+SMOTE improved recall by generating synthetic minority samples, ensuring better detection of churners, while GB+RUS reduced bias by balancing the dataset. In contrast, previous studies relied on standalone models, such as C4.5, SVM, or ensemble methods without resampling, which led to lower accuracy, recall, and AUC. Thus, the proposed method's ability to capture complex patterns while reducing overfitting and class imbalance issues was demonstrated.

IV. CONCLUSION

This paper presents a robust churn prediction model that utilizes advanced resampling techniques, namely Random Undersampling (RUS) and Synthetic Minority Oversampling Technique (SMOTE), to address the class imbalance prevalent in telecommunications (telecom) datasets. The proposed methodology, employing Gradient Boosting (GB) along with Random Under-Sampling (RUS) and Random Forest (RF) with SMOTE, surpassed current churn prediction models across multiple critical performance criteria. The RF+SMOTE technique attained the highest accuracy (79.23%), precision (79.32%), recall (80.15%), F1-score (79.73%), and AUC (87.25%). The results confirm the efficacy of the proposed model in accurately identifying churners with high precision and recall, rendering it a potent instrument for telecommunication firms aiming to enhance their customer retention tactics. Future research may investigate the optimization of resampling approaches and the integration of other information, such as customer interaction data, to augment model prediction efficacy.

ACKNOWLEDGMENT

This research is supported and fully funded by Universitas Amikom Yogyakarta.

REFERENCES

[1] K. Kusnawi, J. Ipmawati, B. P. Asadulloh, A. Aminuddin, F. F. Abdulloh, and M. Rahardi, "Leveraging Various Feature Selection Methods for Churn Prediction Using Various Machine Learning Algorithms," *JOIV : International Journal on Informatics Visualization*, vol. 8, no. 2, pp. 897–905, May 2024, https://doi.org/10.62527/joiv.8.2.2453.

[2] S. Ouf, K. T. Mahmoud, and M. A. Abdel-Fattah, "A proposed hybrid framework to improve the accuracy of customer churn prediction in telecom industry," *Journal of Big Data*, vol. 11, no. 1, May 2024, Art. no. 70, https://doi.org/10.1186/s40537-024-00922-9.

[3] A. Sikri, R. Jameel, S. M. Idrees, and H. Kaur, "Enhancing customer retention in telecom industry with machine learning driven churn prediction," *Scientific Reports*, vol. 14, no. 1, Jun. 2024, Art. no. 13097, https://doi.org/10.1038/s41598-024-63750-0.

[4] O. Soleiman-garmabaki and M. H. Rezvani, "Ensemble classification using balanced data to predict customer churn: a case study on the telecom industry," *Multimedia Tools and Applications*, vol. 83, no. 15, pp. 44799–44831, May 2024, https://doi.org/10.1007/s11042-023-17267-9.

[5] R. Chinnaraj, "Bio-Inspired Approach to Extend Customer Churn Prediction for the Telecom Industry in Efficient Way," *Wireless Personal Communications*, vol. 133, no. 1, pp. 15–29, Nov. 2023, https://doi.org/10.1007/s11277-023-10697-6.

[6] M. Z. Alotaibi and M. A. Haq, "Customer Churn Prediction for Telecommunication Companies using Machine Learning and Ensemble Methods," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14572–14578, Jun. 2024, https://doi.org/10.48084/etasr.7480.

[7] M. Mandic and G. Kraljevic, "Churn Prediction Model Improvement Using Automated Machine Learning with Social Network Parameters," *Revue d'Intelligence Artificielle*, vol. 36, no. 3, pp. 373–379, Jun. 2022, https://doi.org/10.18280/ria.360304.

[8] N. Siddiqui, M. A. Haque, S. M. S. Khan, M. Adil, and H. Shoaib, "Different ML-based strategies for customer churn prediction in banking sector," *Journal of Data, Information and Management*, vol. 6, no. 3, pp. 217–234, Sep. 2024, https://doi.org/10.1007/s42488-024-00126-z.

[9] K. Eria and B. P. Marikannan, "Significance-Based Feature Extraction for Customer Churn Prediction Data in the Telecom Sector," *Journal of Computational and Theoretical Nanoscience*, vol. 16, no. 8, pp. 3428–3431, Aug. 2019, https://doi.org/10.1166/jctn.2019.8303.

[10] S. Brmez and M. Znidarsic, "A Case of Churn Prediction in Telecommunications Industry," *IPSI Transactions on Internet Research*, vol. 15, no. 2, pp. 3–9, 2019.

[11] C. Colot, P. Baecke, and I. Linden, "Leveraging fine-grained mobile data for churn detection through Essence Random Forest," *Journal of Big Data*, vol. 8, no. 1, Apr. 2021, Art. no. 63, https://doi.org/10.1186/s40537-021-00451-9.

[12] I. N. M. Adiputra and P. Wanchai, "CTGAN-ENN: a tabular GAN-based hybrid sampling method for imbalanced and overlapped data in customer churn prediction," *Journal of Big Data*, vol. 11, no. 1, Sep. 2024, Art. no. 121, https://doi.org/10.1186/s40537-024-00982-x.

[13] W. Deng, L. Deng, J. Liu, and J. Qi, "Sampling method based on improved C4.5 decision tree and its application in prediction of telecom customer churn," *International Journal of Information Technology and Management*, vol. 18, no. 1, pp. 93–109, Jan. 2019, https://doi.org/10.1504/IJITM.2019.097887.

[14] "Telco Customer Churn." https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

[15] G. A. Lopez-Ramirez, A. Aragon-Zavala, and C. Vargas-Rosales, "Exploratory Data Analysis for Path Loss Measurements: Unveiling Patterns and Insights Before Machine Learning," *IEEE Access*, vol. 12, pp. 62279–62295, Jan. 2024, https://doi.org/10.1109/ACCESS.2024.3394904.

[16] M. Rath and H. Date, "Quantum data encoding: a comparative analysis of classical-to-quantum mapping techniques and their impact on machine learning accuracy," *EPJ Quantum Technology*, vol. 11, no. 1, Dec. 2024, Art. no. 72, https://doi.org/10.1140/epjqt/s40507-024-00285-3.

[17] A. Ali, N. A. Emran, and S. A. Asmai, "Missing values compensation in duplicates detection using hot deck method," *Journal of Big Data*, vol. 8, no. 1, Aug. 2021, Art. no. 112, https://doi.org/10.1186/s40537-021-00502-1.

[18] C. Ma, H. Wang, O. O. Odegbile, S. Chen, and D. Melissourgos, "Virtual Filter for Non-Duplicate Sampling With Network Applications," *IEEE/ACM Transactions on Networking*, vol. 30, no. 6, pp. 2818–2833, Sep. 2022, https://doi.org/10.1109/TNET.2022.3182694.

[19] M. S. Pathan, A. Nag, M. M. Pathan, and S. Dev, "Analyzing the impact of feature selection on the accuracy of heart disease prediction," *Healthcare Analytics*, vol. 2, Nov. 2022, Art. no. 100060, https://doi.org/10.1016/j.health.2022.100060.

[20] P. Kumari, "A fast feature selection technique in multi modal biometrics using cloud framework," *Microprocessors and Microsystems*, vol. 79, Nov. 2020, Art. no. 103277, https://doi.org/10.1016/j.micpro.2020.103277.

[21] B. Biswas, N. Kumar, Md. A. Hoque, and Md. A. Alam, "Weighted scaling approach for metabolomics data analysis," *Japanese Journal of Statistics and Data Science*, vol. 6, no. 2, pp. 785–802, Nov. 2023, https://doi.org/10.1007/s42081-023-00205-2.

[22] D. Medyakov, G. Molodtsov, A. Beznosikov, and A. Gasnikov, "Optimal Data Splitting in Distributed Optimization for Machine Learning," *Doklady Mathematics*, vol. 108, no. 2, pp. S465–S475, Dec. 2023, https://doi.org/10.1134/S1064562423701600.

[23] R. K. Halder, M. N. Uddin, Md. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *Journal of Big Data*, vol. 11, no. 1, Aug. 2024, Art. no. 113, https://doi.org/10.1186/s40537-024-00973-y.

[24] I. AlShourbaji, N. Helian, Y. Sun, A. G. Hussien, L. Abualigah, and B. Elnaim, "An efficient churn prediction model using gradient boosting machine and metaheuristic optimization," *Scientific Reports*, vol. 13, no. 1, Sep. 2023, Art. no. 14441, https://doi.org/10.1038/s41598-023-41093-6.

[25] T. Pitka *et al.*, "Time analysis of online consumer behavior by decision trees, GUHA association rules, and formal concept analysis," *Journal of Marketing Analytics*, vol. 13, no. 1, pp. 29–52, Mar. 2025, https://doi.org/10.1057/s41270-023-00274-y.

[26] A. Vanacore, M. S. Pellegrino, and A. Ciardiello, "Fair evaluation of classifier predictive performance based on binary confusion matrix," *Computational Statistics*, vol. 39, no. 1, pp. 363–383, Feb. 2024, https://doi.org/10.1007/s00180-022-01301-9.