

ADDRESSING CLASS AND DEMOGRAPHIC IMBALANCE IN E-COMMERCE BEHAVIOR PREDICTION: A CASE STUDY USING RESAMPLING TECHNIQUES

Nurul Ain Mustakim^{1*}, Maslina Abdul Aziz², Shuzlina Abdul Rahman³ and
Rahmiati Rahmiati⁴

^{1*} Faculty Business & Management, Universiti Teknologi MARA,
Melaka 75350 Malaysia

^{2,3} Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA,
40450 Shah Alam, Selangor Malaysia

⁴ Management Department, Universitas Negeri Padang,
Sumatera Barat 25171, Indonesia

^{1*}ainmustakim@uitm.edu.my, ²maslina_aziz@uitm.edu.my, ³shuzlina@uitm.edu.my,
⁴rahmiati@fe.unp.ac.id

(* indicates the corresponding author)

ABSTRACT

In e-commerce predictive modeling, imbalanced data remains a critical challenge, particularly when both class labels and demographic attributes are unequally distributed. This study investigates a combined approach of Synthetic Minority Oversampling Technique (SMOTE) and demographic resampling to improve the performance of models predicting online purchasing behavior in Malaysia. Using a dataset of 1,126 survey responses, six classifiers (J48, Random Tree, REPTree, JRip, PART, and OneR) were evaluated under three conditions: unbalanced, after SMOTE, and after SMOTE with demographic balancing. The results displayed clear improvements in model performance. For example, J48's accuracy increased from 62.85% (unbalanced) to 98.69% (fully balanced), while Random Tree achieved 99.29%. These results highlight the effectiveness of integrating class and demographic balancing, an approach seldom explored in e-commerce analytics. This study contributes by demonstrating how addressing both types of imbalances yields more reliable predictive model, offering practical insights for consumer segmentation, targeting, and personalization. Future work could extend this approach by balancing additional attributes and applying it to ensemble or deep learning models for improved robustness and interpretability.

Keywords: Classification, Consumer Behavior, Data Imbalance, Demographic Resampling, E-Commerce, SMOTE

Received for review: 19-07-2025; Accepted: 31-10-2025; Published: 01-04-2026
DOI: 10.24191/mjoc.vol11i1.8077



This is an open access article under the CC BY-SA license
(<https://creativecommons.org/licenses/by-sa/3.0/>).

1. Introduction

The world is now in a digital era, where technology, especially computers, supports many aspects of daily life. Organizations, groups, and individuals are constantly seeking the latest and most useful information. However, the methods used to collect and process information are not always efficient. To improve this, many researchers and practitioners have turned to machine learning (Ibrahim et al., 2024). Although machine learning has great potential, it relies heavily on the quality of the data. In many cases, especially with large datasets, problems such as class imbalance can occur. This happens when one group has much more data than the others, which can affect the model's performance. Due to its frequent occurrence and impact on results, class imbalance is often seen as one of the main challenges in data mining, as it can reduce the accuracy and fairness of predictive models across various applications (Cheng et al., 2019).

This imbalance can be a major issue in classification tasks such as predicting consumer behavior. For instance, if most consumers fall into one purchasing group, the model may become biased toward that group. According to (Carvalho et al., 2025), class imbalance is a common problem in machine learning, where one group contains more samples than the rest. This can lead the model to focus more on the majority group and overlook the smaller ones, especially when accuracy is the main metric. As a result, the model may not perform well for the minority group, reducing its fairness and general usefulness. A similar issue was observed by (Shafie et al., 2023) in their employee promotion prediction study, where they used a hybrid SMOTE-Tomek method to address the imbalance and improve model performance.

In e-commerce, this can lead to inaccurate customer insights, poor targeting, and weak personalization, particularly for niche or underrepresented consumer groups. This challenge becomes even more relevant in the context of Malaysia's fast-growing digital market, where consumer behavior is highly varied and influenced by diverse socio-demographic factors. Understanding the multifaceted nature of online purchasing decisions shaped by elements such as attitude, motivation, perceived risk, belief, and demographics requires robust predictive models capable of handling these imbalances effectively.

Past studies have shown the importance of solving class imbalance to improve prediction results. For example, (Fitriyani et al., 2022) found that class imbalance in e-commerce fraud detection led to biased outcomes. By using SMOTE, their model could better detect rare fraud cases, which is like identifying less common customer behaviors. (Rakib Mollah et al., 2024) also addressed imbalance in product review sentiment analysis. They used class re-weighting and a language model (RoBERTa) to better recognize rare or subtle sentiments. This helped improve personalization for specific customer groups. These studies show that balancing data can help models make fairer and more accurate predictions.

This study investigates the impact of combining the SMOTE algorithm with demographic resampling on the performance of machine learning models. Experiments were conducted using WEKA, a widely used data mining tool, and focused on three conditions: (1) using the original dataset without any balancing, (2) after applying class balancing with SMOTE, and (3) after applying both class and demographic balancing to address demographic imbalance, particularly in attributes such as gender and age. The primary objective is to enhance the fairness, predictive accuracy, and representativeness of classification models trained on Malaysian e-commerce datasets. By doing this, the study contributes to the development of more inclusive and reliable consumer behavior prediction systems capable of informing targeted marketing strategies and improving customer engagement.

This research also fills a critical gap in the literature by exploring the integrated use of SMOTE and demographic resampling in the Malaysian e-commerce context an area that remains underexplored despite the growing availability of consumer data. Unlike prior studies

that often examine either class imbalance or demographic imbalance in isolation, this study combines both dimensions into a single comprehensive model that incorporates psychological and behavioral factors alongside socio-demographic attributes. By addressing both class and demographic imbalances simultaneously, this study offers an initial step towards extending research in Malaysian e-commerce analytics and suggests direction for developing fairer, more reliable predictive models.

The rest of the paper is structured as follows: Section 2 reviews relevant literature on classification techniques and imbalanced data handling strategies in consumer behavior modeling. Section 3 details the methodology, including dataset characteristics, resampling procedures, and the supervised learning algorithms applied. Section 4 presents the experimental results and comparative analysis, highlighting the impact of different resampling strategies on model performance. Finally, Section 5 concludes the study by summarizing key findings and outlining implications for future research and practical applications in e-commerce.

2. Related Works

In machine learning, particularly in classification problems, class imbalance is a frequent occurrence. This phenomenon occurs when one class has an overwhelmingly larger sized dataset than other classes, leading to a model bias towards the majority class. This situation has created the need to develop strategies to remove or minimize its effects. These strategies include reducing the majority class, increasing the minority class, or creating synthetic data. One of the best-known strategies is the Synthetic Minority Over-sampling Technique (SMOTE), which builds new synthetic samples for the minority class. This helps to balance the dataset, thus improving the model's accuracy (Fitriyani et al., 2022).

To enhance model performance and data balance, several researchers have explored alternative versions of SMOTE and other methods of resampling. Several studies have looked at how SMOTE and resampling approaches can mitigate not only class imbalance, but also demographic imbalances, including those of gender and age. Zheng (2020) introduced improved versions of SMOTE called DSMOTE and ESMOTE, which focus on harder-to-classify regions of the data for improved accuracy. (Rachmatullah, 2022) demonstrated that the iterative application of SMOTE can mitigate multi-class problems with uneven class distribution. (Hairani et al., 2024) and others reviewed the focus on the enhancement of SMOTE to look beyond the inclusion of noisy data, sample clustering, the adjustment of distance measures, among others. Such adjustments with demographic variation proved to be very useful. (Husain et al., 2025) compared SMOTE with SMOTEENN, a version that also cleans noisy data, and found it produced more accurate and stable results.

Another method, SMOTE-BFT have been introduced by (Grina et al., 2020), uses belief function theory to eliminate poor quality synthetic samples, reducing the impact of noisy or overlapping data. While these techniques were initially developed to address class imbalance, many researchers highlight their usefulness in handling demographic imbalances as well, particularly in domains like healthcare and consumer behavior analysis. This study validates that applying multilayer SMOTE upsampling combined with stratified sampling significantly improves prediction accuracy and class balance in a highly imbalanced multiclass dataset, using Fantasy Premier League player price changes as a case study (Khamsan & Maskat, 2019).

In e-commerce, class imbalance issues typically arise because of infrequent users, which leads to a sparse cohort of users who are frequent buyers. This phenomenon is typical in session-based datasets, where a vast majority of user activity is not accompanied by any purchase, leading to a highly unbalanced distribution (Diwandari et al., 2024). Moreover, demographic imbalances, like the overrepresentation of a specific gender or age group, such as women between the age of 21 and 40, can add additional burdens to the equality and

efficiency of the predictive models. When not properly catered to, these imbalances can produce skewed results and compromise the accuracy of analytics generated from the analysis of consumer behavior. As class imbalance issues arise in e-commerce, practitioners often turn to SMOTE as a means of solution. However, the combination of SMOTE with random undersampling has shown great potential in the prediction of tourism behavior, (Zhang & Zeng, 2024). Nonetheless, in the context of Malaysia, e-commerce studies incorporating demographic SMOTE-based resampling techniques remain sparse, underscoring the need for further investigation.

In support of this, a study by (Selvalakshmi et al., 2025) demonstrates the predictive performance of underrepresented classes with the usage of ensemble techniques like Random Forest and Gradient Boosting and different SMOTE techniques like Borderline-SMOTE, ADASYN, and SMOTEENN. Their results indicated that the use of SMOTEENN in a voting ensemble produced a recall of 99% for the minority class, illustrating that a combination of resampling and powerful classifiers can enhance the accuracy and fairness of predictions.

Based on earlier work, this study investigates the impact of combining the SMOTE algorithm with demographic resampling on the performance of machine learning models. The work was carried out on WEKA, one of the most popular data mining software, and focused on three primary configurations: SMOTE only, the combination of SMOTE and downsampling, and SMOTE with upsampling. The focus was on evaluating the performance of the models with regard to their accuracy in predicting consumer behaviors. This study aims to improve the performance of consumer behavior models by addressing class imbalance and demographic imbalance. The findings of this study can enhance automated decision-making processes for e-commerce companies in the areas of customer segmentation, targeting, and personalization.

Numerous studies have used machine learning algorithms to predict consumer behavior in e-commerce. For example, in tree-based algorithms J48, Random Tree, and REPTree are commonly used because they can handle various types of data. J48 is known for its simple and understandable decision rules. Random Tree performs well with large datasets, while REPTree is fast and automatically removes unnecessary branches to make the model simpler. Meanwhile for rule-based algorithms such as JRip, OneR, and PART are also popular because they generate straightforward rules that are easier to explain to non-technical users. However, a common issue across these algorithms is data imbalance, where one group has more samples than others. This imbalance can affect the model's performance and reduce its ability to generalize. To address this, many researchers use techniques like SMOTE to balance the data before training the models. Some studies have shown that combining decision tree or rule-based models with resampling methods like SMOTE can improve prediction accuracy, especially for underrepresented groups (Abdul Wadud Soraker et al., 2022). Based on this, the current study applies J48, Random Tree, REPTree, JRip, OneR, and PART along with SMOTE and demographic resampling to assess their impact on prediction accuracy using a Malaysian e-commerce dataset.

3. Methodology

This section consists of the methodology of this study, including dataset overview in Section 3.1, addressing class imbalance with SMOTE in Section 3.2, addressing demographic imbalance in Section 3.3, and classifier and evaluation matrix in Section 3.4.

3.1 Dataset Overview

This study uses a dataset consisting of 1,126 instances, gathered through a survey focused on online purchasing behavior among Malaysian consumers (Aziz et al., 2024). The dataset includes a variety of attributes such as gender, age, employment status, psychological factors,

and behavioral aspects related to online shopping. The class attribute, labeled “Will the consumer purchase?”, is measured on a five-level ordinal scale reflecting purchase intention. The categories range from "definitely would not buy", "probably would not buy", "might buy", "probably would buy", to "definitely buy". This classification enables a more detailed analysis of consumer intent in online shopping contexts.

3.2 Addressing Class Imbalance with SMOTE

The initial dataset showed that the class attribute, which reflects consumer purchasing intention on a five-point scale, was imbalanced. Some categories, such as “definitely would not buy” or “definitely buy,” had fewer responses compared to others. This imbalance can affect how well a model learns to predict all classes fairly.

To improve this, the Synthetic Minority Over-sampling Technique (SMOTE) was used in WEKA. SMOTE generates new synthetic samples for underrepresented classes based on existing data patterns. This helped balance the number of instances across all five purchasing categories, making the dataset more suitable for training classification models.

The application of SMOTE resulted in a more even class distribution, reducing bias toward the majority class and improving the model’s ability to generalize across all consumer intention levels. As part of this study, model accuracy improved significantly after class balancing, with accuracy gains observed across all classifiers.

3.3 Addressing Demographic Imbalance

After applying SMOTE to handle class imbalance in the target variable, the dataset still showed uneven distribution in some demographic attributes. The most noticeable was **gender**, where female respondents were higher than males. Since this could affect prediction results, resampling was used by reducing female instances (downsampling) and increasing male instances (upsampling). After testing several settings, the final dataset contained **754 female and 366 male records**, giving a total of **1,120 records**. The total is slightly lower than the 1,126 records because some majority groups were reduced while minority groups were increased.

Other demographic attributes were also adjusted to reduce imbalance. For age, the 18–25 group was dominant, while the 36–45 and 46–60 groups were fewer; therefore, their proportions were rebalanced. In terms of employment status, most respondents were employed compared to unemployed, so the number of unemployed instances was increased to improve representation. For ethnicity, Malay respondents formed the majority, while Chinese, Indian, and other groups were underrepresented. To address this, the Malay group was slightly reduced, and the minority groups were increased. These adjustments helped create a dataset that is fairer across gender, age, employment, and ethnicity, supporting the development of models that are less likely to be biased.

3.4 Classifier and Evaluation Matrix

To evaluate the performance of the dataset after applying SMOTE and demographic resampling, six classification algorithms available in WEKA were selected. These classifiers include J48, Random Tree, REPTree, JRip, PART, and OneR. The selection was made to cover a range of tree-based and rule-based algorithms commonly used in classification tasks. J48, a well-known implementation of the C4.5 decision tree, was chosen for its interpretability and widespread use. Random Tree and REPTree provide alternative decision tree approaches, with Random Tree introducing randomness in attribute selection and REPTree optimizing through reduced error pruning. Meanwhile, JRip and PART represent rule-based classifiers that construct if-then rules, offering simplicity and transparency in prediction logic. OneR, although basic, was included for comparison as it builds rules based on a single attribute and serves as a simple baseline.

In this study, accuracy was used as the primary evaluation metric to measure the performance of each classifier under different data balancing conditions. Accuracy reflects the

proportion of correctly classified instances out of the total instances and provides a general overview of model effectiveness. While other metrics such as precision, recall, and F-measure are commonly used in imbalanced data scenarios, this study focuses specifically on accuracy to compare the impact of class and demographic balancing techniques across all classifiers.

All experiments were conducted using 10-fold cross-validation, a widely adopted validation technique that divides the dataset into ten subsets. The model is trained on nine subsets and tested on the remaining one, and the process is repeated ten times to ensure that each instance is used for both training and testing. This approach helps reduce overfitting and provides a more reliable estimate of model accuracy.

3.4.1 Decision Tree

Algorithms that popular in machine learning is decision tree, because they are simple and easy to understand. Three decision tree algorithms from WEKA were used in this study: J48, Random Tree, and REPTree. J48 is based on the C4.5 algorithm and works by choosing the best way to split the data using information gain. It also includes a process called pruning to avoid overfitting, which helps make the model more accurate. Random Tree builds several trees using random parts of the data and does not use pruning, which makes it more flexible. REPTree builds the tree using information gain or variance and uses pruning to make the model simpler and faster. These algorithms are useful because they give good results while being easy to use and understand.

3.4.2 Ruled Based

This study also used three rule-based algorithms: PART, JRip, and OneR. These methods build simple rules to help classify the data. JRip is based on the RIPPER algorithm and works by creating and adjusting rules to make the model more general and better at handling noisy data. OneR creates one rule for each attribute and picks the one with the fewest errors. It is the simplest method and is easy to understand. PART builds part of a decision tree and then chooses the best rule from it. This helps create models that are both accurate and clear. Rule-based methods are useful when we want models that are easy to explain and understand.

4. Results and Discussion

This section presents the results of the classification experiments using six selected classifiers tested in three different conditions: (1) using the original dataset without any balancing, (2) after applying class balancing with SMOTE, and (3) after applying both class and demographic balancing. These conditions were used to compare how each balancing method affects model performance. In this study, accuracy was used as the main evaluation metric. Accuracy was selected because it is straightforward, easy to interpret, and commonly applied in classification research. Since the main aim was to compare overall performance across classifiers under different balancing conditions, accuracy was considered sufficient for this purpose.

All experiments were conducted using 10-fold cross-validation to ensure reliability and reduce overfitting. The dataset itself was obtained through a structured survey on online purchasing behavior among Malaysian consumers. The survey instrument was developed with reference to existing literature, reviewed by subject experts, and pre-tested with a pilot group to confirm clarity and consistency. Responses were screened for completeness and accuracy before inclusion in the analysis. These steps help ensure that the dataset is valid and suitable for classification experiments.

Table 1 shows the accuracy performance of six classifiers in three situations: without balancing, with only class label balancing, and with both class label and input feature

balancing. When no balancing was applied, all classifiers gave low accuracy results, between 53% and 63%, showing that imbalance reduces predictive performance. When class labels were balanced using SMOTE, accuracy improved for most classifiers, with J48 and Random Tree showing the biggest gains. OneR, however, performed slightly worse, likely because it is a simple model that does not handle synthetic data well.

The best results were obtained when both the class labels and demographic features such as gender and age were balanced. Most classifiers then achieved accuracy above 98%, with Random Tree reaching 99.29%, followed by PART and J48. This indicates that combining class and demographic balancing produces fairer and more reliable models. The lowest accuracy (75.70%) was still recorded by OneR, confirming that this algorithm is better used as a baseline.

In summary, the results show that balancing both class labels and demographic features is important for improving model performance. Decision tree models like J48 and Random Tree are recommended because they performed consistently well under balanced conditions. Although OneR is easy to understand, its limitations suggest that it is less suitable for complex prediction tasks. These findings are useful for improving prediction in e-commerce, as they suggest that balanced models can help marketers build fairer consumer profiles and design marketing strategies that reach diverse customer groups more effectively.

Table 1. Classifier Accuracy Comparison Across Balancing Scenarios

Classifier	Accuracy (Without Balanced)	Accuracy (Class Label Balanced)	Accuracy (Input and Class Features Balanced)
J48	62.85%	86.41%	98.69%
Random Tree	53.75%	89.16%	99.29%
REPTree	62.67%	83.83%	98.30%
JRip	62.14%	84.90%	98.80%
OneR	62.32%	60.92%	75.70%
PART	54.07%	87.47%	98.72%

5. Conclusion and Recommendation

This research addresses the problem of both class imbalance and demographic imbalance in predicting online shopping behavior. To overcome these issues, SMOTE was applied for class label balancing, and resampling techniques were used to adjust demographic characteristics such as gender. The results verified that balancing both dimensions significantly improved classification performance, with J48 and Random Tree showing the strongest results. These findings underline the importance of addressing multiple types of imbalances to enhance the accuracy, fairness, and reliability of predictive models. Importantly, this study is among the first in Malaysia to demonstrate how combining class and demographic balancing can improve e-commerce prediction models, offering practical insights for fairer personalization and customer targeting.

For future research, several directions are recommended. Other demographic attributes such as education, employment status, and residential area could be examined for balancing, as they may also influence purchasing behavior. Researchers may also explore combining multiple balancing techniques or using automated balancing tools. In terms of modeling, more advanced approaches such as ensemble methods (e.g., Random Forest, AdaBoost) and deep learning models (e.g., neural networks) can be investigated to further improve robustness. Additionally, future studies could examine the long-term effects of balancing approaches on real marketing outcomes, including personalization strategies and customer satisfaction. Finally, integrating explainability techniques will help companies better understand the contribution of each attribute, fostering transparency and trust in AI-driven decision-making.

Acknowledgements

The researchers would like to thank UiTM for their support of academic research and funding possibilities to share research findings locally.

Funding

The author(s) received no specific funding for this work.

Author Contribution

Author 1 prepared the literature review, wrote the content, developed the research methodology, and conducted the results analysis and discussion. Author 2 and Author 3 reviewed the literature review and provided constructive feedback on the manuscript. All authors have reviewed and approved the final version of the manuscript.

Conflict of Interest

No conflicts of interest to declare.

References

- Abdul Wadud Sorker, M., Siddika, A., Abedin Titly, T., Jueal Mia, M., & Hasan Imam Bijoy, M. (2022). Online Consumer Alignment using Supervised Machine Learning Technique. *2022 13th International Conference on Computing Communication and Networking Technologies, ICCCNT 2022*, 1–7. <https://doi.org/10.1109/ICCCNT54827.2022.9984234>
- Aziz, M. A., Mustakim, N. A., & Rahman, S. A. (2024). *Decision Tree and Rule-Based Classification for Predicting Online Purchase Behavior in Malaysia*. 9(2), 1905–1915. <https://doi.org/10.24191/mjoc.v9i2.27130>
- Carvalho, M., Pinho, A. J., & Brás, S. (2025). Resampling approaches to handle class imbalance: a review from a data perspective. *Journal of Big Data*, 12(1). <https://doi.org/10.1186/s40537-025-01119-4>
- Cheng, K., Zhang, C., Yu, H., Yang, X., Zou, H., & Gao, S. (2019). Grouped SMOTE with Noise Filtering Mechanism for Classifying Imbalanced Data. *IEEE Access*, 7, 170668–170681. <https://doi.org/10.1109/ACCESS.2019.2955086>
- Diwandari, S., Permanasari, A. E., & Hidayah, I. (2024). Comparative Evaluation of Resampling Techniques on Machine Learning for Predicting Customer Purchasing Behavior. *8th International Conference on Information Technology 2024, InCIT 2024*, 103–108. <https://doi.org/10.1109/InCIT63192.2024.10810628>
- Fitriyani, A., Priatna, W., Lestari, T. S., Handayani, D., Munandar, T. B. A., & Amri. (2022). Data Balance Optimization of Fraud Classification for E-Commerce Transaction. *2022 7th International Conference on Informatics and Computing, ICIC 2022*, 1–4. <https://doi.org/10.1109/ICIC56845.2022.10007028>

- Grina, F., Elouedi, Z., & Lefevre, E. (2020). A Preprocessing Approach for Class-Imbalanced Data Using SMOTE and Belief Function Theory. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12490 LNCS, 3–11. https://doi.org/10.1007/978-3-030-62365-4_1
- Hairani, H., Widiyaningtyas, T., & Prasetya, D. D. (2024). Addressing Class Imbalance of Health Data: a Systematic Literature Review on Modified Synthetic Minority Oversampling Technique (SMOTE) Strategies. *International Journal on Informatics Visualization*, 8(3), 1310–1318. <https://doi.org/10.62527/joiv.8.3.2283>
- Husain, G., Nasef, D., Jose, R., Mayer, J., Bekbolatova, M., Devine, T., & Toma, M. (2025). SMOTE vs. SMOTEENN: A Study on the Performance of Resampling Algorithms for Addressing Class Imbalance in Regression Models. *Algorithms*, 18(1), 1–16. <https://doi.org/10.3390/a18010037>
- Ibrahim, N., Ishak, U. M., Ali, N. N. A., & Shaadan, N. (2024). Machine Learning-Based Approaches for Credit Card Debt Prediction. *Malaysian Journal of Computing (Mjoc)*, 9(1), 1722–1733. <https://doi.org/10.24191/mjoc.v9i1.25656>
- Khamsan, M. M., & Maskat, R. (2019). Handling Highly Imbalanced Output Class Label: a Case Study on Fantasy Premier League (Fpl) Virtual Player Price Changes Prediction Using Machine Learning. *Malaysian Journal of Computing*, 4(2), 304–316. <https://www.calculator.net/standard-deviation-calculator.html>
- Rachmatullah, M. I. C. (2022). The Application of Repeated SMOTE for Multi Class Classification on Imbalanced Data. *MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, 22(1), 13–24. <https://doi.org/10.30812/matrik.v22i1.1803>
- Rakib Mollah, M. A., Jahangir Kabir, M. M., Reza, M. S., & Kabir, M. (2024). Adapting Contextual Embedding to Identify Sentiment of E-commerce Consumer Reviews with Addressing Class Imbalance Issues. *2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems: Innovation for Sustainability, ICACCESS 2024*, 1–6. <https://doi.org/10.1109/iCACCESS61735.2024.10499554>
- Selvalakshmi, V., Sree, T. M. U., Saranya, S., Devi, A. U., & Basha, M. S. A. (2025). Enhancing Customer Personality Prediction using Advanced Machine Learning Techniques and Data Balancing Strategies: A Comprehensive Approach to Addressing Imbalanced Datasets in Marketing Analytics. *International Conference on Intelligent Systems and Computational Networks, ICISCN 2025*, 1–7. <https://doi.org/10.1109/ICISCN64258.2025.10934379>
- Shafie, S., Peng Ooi, S., & Khaw, K. W. (2023). Prediction of Employee Promotion Using Hybrid Sampling Method With Machine Learning Architecture. *Malaysian Journal of Computing*, 8(1), 1264–1286. <https://doi.org/10.24191/mjoc.v8i1.18456>
- Zhang, W., & Zeng, N. (2024). An explainable hybrid model of Tourism User Behavior prediction based on SMOTE-XGBoost. *ACM International Conference Proceeding Series*, 406–410. <https://doi.org/10.1145/3696500.3696567>