Perbanas International Conference on Economics, Business, Management, Accounting and IT (MASTER) 2026
"Integrity-Driven Innovation: Reimagining Business, Finance, and Digital Transformation for Sustainable Impact"

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Random Forest-Based Prediction of Customer Responses to Marketing Campaigns

Sri Wahyuni Aprianti[1,a], Yoga Armando[2,b], Elliana Gautama[3,c]

[1,2,3] Data Science Study Program, Faculty of Information Technology, Perbanas Institute

[a)]Corresponding author: sri.wahyuni06@perbanas.id,

[b)]yoga.armando08@perbanas.id,

[c)]elliana@perbanas.id

Abstract - *A company's ability to predict customer responses to marketing campaigns is one of the key factors in improving budget efficiency and increasing sales conversions. However, conventional approaches based on intuition or simple rules often result in inaccurate targeting, leading to inefficient use of marketing resources. This study proposes the application of the Random Forest algorithm to predict whether customers will respond to marketing campaigns based on customer profile data and purchasing behavior. The dataset used in this research was obtained from Kaggle (Customer Personality Analysis), which contains demographic information, purchase history, and records of previous campaign interactions. This study is limited to the analysis of customer responses based on demographic characteristics, purchasing behavior, and previous marketing campaign interactions available in the dataset. In addition, the classification process only focuses on two classes, namely customers who respond positively to marketing campaigns (1) and customers who do not respond to marketing campaigns (0). The research process includes data cleansing, feature engineering, data splitting, Random Forest model training, and evaluation using accuracy, precision, recall, and F1-score metrics. The results show that the model is capable of performing binary classification between responsive and non-responsive customers with good predictive performance.*

Keywords: Customer Response Prediction, Random Forest, Machine Learning, Marketing Campaigns

INTRODUCTION

In the era of data-driven marketing, the ability to accurately predict customer behavior has become a valuable competitive advantage(Mittal & Kumar, 2024). A fundamental problem faced by marketers is the inability of conventional systems to process and interpret multidimensional customer data, including demographic factors, purchasing behavior, product preferences, and previous campaign interactions, simultaneously and accurately(James et al., 2024). Therefore, a method is needed that is able to integrate these variables into a reliable and interpretive predictive model.

Machine learning, especially ensemble methods like Random Forest, has proven effective in handling high-volume data with diverse feature dimensions (Sarker, 2021). The study uses a Customer Personality Analytics dataset that includes demographic profiles, transaction history, and a track record of customer engagement with previous campaigns (Kumar et al., nd, 2022). This study applied binary classification to predict which customers responded and

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

did not respond to a marketing campaign. The model performance was evaluated using accuracy, precision, recall, and F1-score metrics.

Although machine learning techniques have been widely applied in customer response prediction, existing studies mainly focus on improving predictive performance while providing limited discussion on how demographic characteristics, purchasing behavior, and previous campaign interactions collectively influence customer responses to marketing campaigns (Campaigns, 2024; Singh, 2024). Therefore, this study contributes by implementing the Random Forest algorithm to predict customer responses using the Customer Personality Analysis dataset while identifying the most influential variables affecting customer engagement.

LITERATURE RIVIEW

Customer Behavior and Segmentation

Understanding customer behavior is the foundation of any effective marketing strategy. Customer personality analysis allows companies to segment consumers based on demographic, psychographic, and purchasing behavior attributes, so they can design more relevant and appropriate offerings (Brito et al., 2021).

Machine Learning and Marketing

Machine learning has changed the way companies understand and interact with their customers. With the ability to analyze patterns from large-scale historical data, machine learning algorithms are able to generate predictions that are much more accurate than conventional statistical models (Sarker, 2021).

Random Forest Algorithm

Random Forest is an ensemble learning algorithm that builds a large number of decision trees in parallel, then combines the predicted results of each tree to produce a more accurate and stable final decision(Brito et al., 2021). The Random Forest algorithm works through several stages, including bootstrap sampling, random feature selection, decision tree construction, majority voting, and model evaluation.

Data Preprocessing and Feature Engineering

Data quality is a major determining factor in the performance of machine learning models. Comprehensive preprocessing processes, including missing value handling, normalization of numerical features, and coding of categorical variables, have been shown to significantly improve the accuracy and stability of predictive models (Anant et al., 2025).

Evaluation of the Clarification Model

Model evaluations on unbalanced datasets, where the positive class (the responding customer) is much smaller than the negative class, cannot rely solely on accuracy metrics. F1's precision, recall, and score metrics provide a more comprehensive picture of the model's ability to correctly identify minority classes (Sujon et al., 2025).

Previous Research

(Mittal & Kumar, 2024) It shows that machine learning-based models are able to predict campaign responses with a level of accuracy that consistently exceeds traditional statistical approaches. In line with these findings, (Rocha et al., 2026) It also confirms that the integration of customer behavior data into predictive models significantly strengthens the personalization capabilities of marketing campaigns.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

Resaerch Gap

Previous studies have demonstrated the effectiveness of machine learning algorithms in predicting customer responses to marketing campaigns (Campaigns, 2024). Many studies primarily focus on predictive accuracy while providing limited discussion regarding the contribution of customer demographic and purchasing behavior variables to prediction outcomes. The interpretability of influential features affecting customer responses is often underexplored. Therefore, this study seeks to address these gaps by implementing the Random Forest algorithm and examining the importance of customer-related features in predicting marketing campaign responses.

Literature Synthesis

Based on the literature review that has been conducted, there is a strong consensus among researchers that machine learning approaches, especially ensemble methods such as Random Forest, offer superior solutions over conventional models in predicting customer responses (Ji et al., 2022). The success of the model relies heavily on the quality of pre-processing data, rigor in feature engineering, and the selection of appropriate evaluation metrics (Sarker, 2021).

Conceptual Framework

The conceptual framework of this study describes the flow of the process of predicting customer responses to marketing campaigns using the Random Forest algorithm-based machine learning approach. Overall, the framework consists of three main layers: (1) a data input layer that includes demographic attributes, purchase behavior, and campaign interaction history; (2) a processing layer that includes the stages of preprocessing, feature engineering, model training, and prediction; and (3) an output layer in the form of a binary classification of customer responses that are evaluated using F1 accuracy, precision, recall, and score metrics.

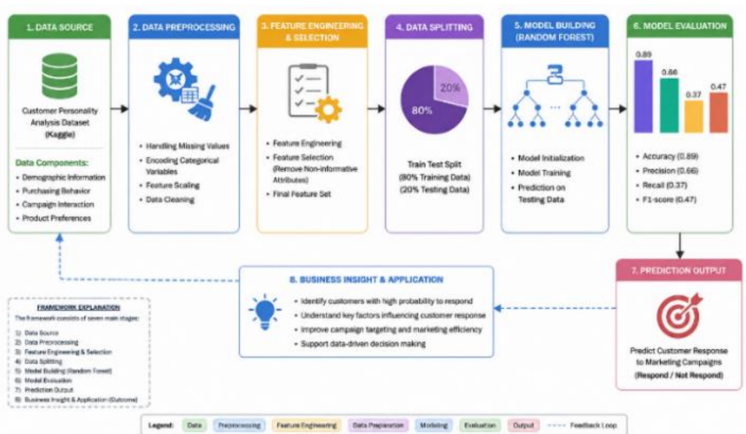

FIGURE 1. Conceptual Framework

As shown in figure 1, this conceptual framework maps the logical relationship between raw data inputs, transformation stages, algorithm applications, and model output evaluations. Each component in this framework is designed to ensure that the prediction process runs systematically, transparently, and can be replicated by other researchers in the future.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

RESEARCH METHODOLOGY

Research Design

The dataset consists of 2240 customer data with 29 variables that include demographic information, purchasing behavior, revenue, transaction history, and customer responses to marketing campaigns. The target variable used in this study is Response, which is classified into two classes, namely customers who respond to campaigns (1) and customers who do not respond to campaigns (0). (Sarker, 2021).

Data Source

This study uses the public dataset Customer Personality Analysis from Kaggle which represents customer data in the context of retail and marketing companies. The target field used is Responses, where a value of 1 indicates a customer who received the campaign and a value of 0 indicates a person who did not respond.

Dataset Overview and Implementation Environment

The following table presents a sample of the Customer Personality Analysis dataset used in this study before the preprocessing and modeling stages were conducted.

TABLE 1. Customer Dataset Overview

ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines
5524	1957	Graduation	Single	58138	0	0	04/09/2012	58	635
2174	1954	Graduation	Single	46344	1	1	08/03/2014	38	11
4141	1965	Graduation	Together	71613	0	0	21/08/2013	26	426
6182	1984	Graduation	Together	26646	1	0	10/02/2014	26	11
5324	1981	PhD	Married	58293	1	0	19/01/2014	94	173

MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	NumWebPurchases
88	546	172	88	88	3	8
1	6	2	1	6	2	1
49	127	111	21	42	1	8
4	20	10	3	5	2	2
43	118	46	27	15	5	5

NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5
10	4	7	0	0	0
1	2	5	0	0	0
2	10	4	0	0	0
0	4	6	0	0	0
3	6	5	0	0	0

AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
0	0	0	3	11	1
0	0	0	3	11	0
0	0	0	3	11	0
0	0	0	3	11	0
0	0	0	3	11	0

The implementation environment and dataset structure support efficient data analysis, preprocessing, and machine learning model development throughout the research process.

Dataset Input and Exploration

The dataset used in this study is a Dataset of Customer Personality Analysis obtained from Kaggle. The dataset is imported using Panda and explored using functions such as head(), info(), and describe() to understand the structure and characteristics of the data.

TABLE 2. Initial Dataset Exploration Process

Process	Purpose
head()	Display the first row of a dataset
info()	View dataset structure and data types
isnull().sum()	Check for missing values

The exploration process identified demographic information, purchasing behavior, and campaign response data within the dataset, confirming that the data structure and attributes were suitable for preprocessing and machine learning implementation.

Data Prepocessing

Data pre-processing is done to improve the quality of the data set before the machine learning modeling stage. Preprocessing activities include checking for missing values, removing incomplete records, encoding categorical variables, and standardizing numerical features.

TABLE 3. Pre-Data Processing Activities

Activities	Method
Missing Value Check	isnull().sum()
Removal of Missing Values	dropna()
Categorical Coding	LabelEncoder()
Feature Scaling	StandardScaler()

The preprocessing stage transforms the raw customer dataset into a structured format suitable for machine learning analysis through missing value handling, categorical encoding, and data standardization, improving dataset consistency and supporting effective model training.

Feature Engineering and Feature Selection

The feature engineering process focused on preparing customer demographic, purchasing behavior, and campaign interaction variables for machine learning analysis. No explicit feature selection technique was applied, as Random Forest is capable of evaluating feature importance internally during model training.

Data Splitting

The dataset was divided into training and testing data using the train_test_split function from Scikit-learn with an 80:20 ratio. A random_state value of 42 was applied to ensure consistent and reproducible results during the model training and evaluation process. To assess stability and reduce dependence on a single split, model selection used stratified k-fold cross-validation (k = 5) on the training set, reporting the mean and standard deviation of F1 across folds, the held-out test set was used only for final evaluation.

Feature Scaling

Feature normalization was performed using the StandardScaler from Scikit-learn to standardize numerical features into a similar scale. This process helps improve model performance by preventing features with larger value ranges from dominating the learning process.

Implementation of the Random Forest Model

The Random Forest model is implemented using the RandomForestClassifier class from the Scikit-learn library for binary classification of customer responses. The model uses 100 decision trees (n_estimators=100) to improve prediction accuracy and reduce overfitting during the classification process.

Evaluation Matrix

The evaluation of the model's performance is carried out using four main metrics: accuracy, precision, recall, and F1 score. Precision, recall, and F1-score were used to handle potential class imbalance problems in the dataset (Sujon et al., 2025). Confusion matrix visualization is also used to analyze the distribution of misclassifications in more depth, while feature importance graphs are used to identify the most influential variables in model predictions.

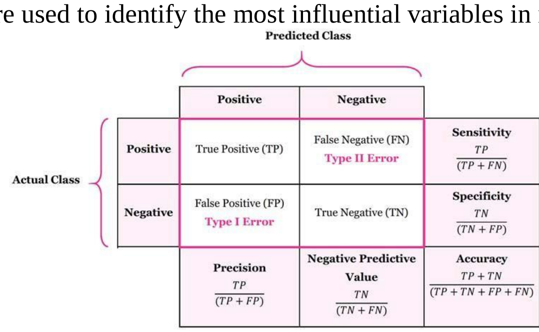

FIGURE 2. Confusion Matrix Formula

As shown in figure 2, confusion matrix visualization is also used to analyze the distribution of misclassifications in more depth through True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) classifications(Powers, 2011). In addition, feature importance graphs are used to identify the most influential variables in model predictions. The distribution of customer responses indicated a class imbalance between responsive and non-responsive customers. Therefore, model evaluation was not limited to accuracy but also included precision, recall, and F1-score metrics(Campaigns, 2024).

Perbanas International Conference on Economics, Business, Management, Accounting and IT (MASTER) 2026
"Integrity-Driven Innovation: Reimagining Business, Finance, and Digital Transformation for Sustainable Impact"

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

RESULTS AND DISCUSSION

Correlation Heatmap

Correlation analysis was conducted to identify the relationships between variables in the dataset before the modeling process. The heatmap visualization helps illustrate the strength and direction of correlations among demographic attributes, purchasing behavior, and customer campaign response variables.

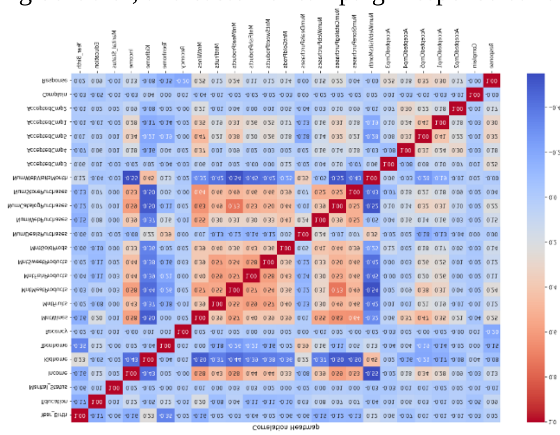

FIGURE 3. Correlation Heatmap Between Variables

As shown in figure 3, several purchasing behavior variables demonstrate moderate to strong positive correlations with one another, indicating that customer spending patterns are interconnected. In contrast, some demographic variables show weaker relationships, suggesting that not all customer characteristics have the same influence on marketing campaign responses.

Customer Response Distribution

Before the modeling stage, the distribution of customer responses is analyzed to understand the balance between customers who accept and reject marketing campaigns.

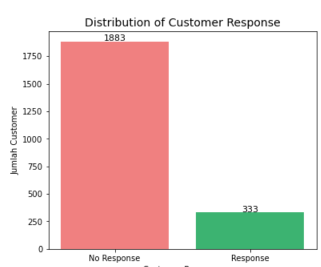

FIGURE 4. Customer Response Distribution

As shown in figure 4, the distribution of customer response classes in the dataset. The visualization shows that unresponsive customers dominate the dataset, while positively responsive customers represent a smaller proportion. This imbalance can affect the predictive performance of machine learning models, especially in identifying minority classes.

Random Forest Model

The Random Forest algorithm is implemented as the primary classification model to predict customer responses to marketing campaigns using training data that has passed the preprocessing stage. The Random Forest model generates multiple decision trees to classify customer responses based on specific feature conditions. The following visualization represents one of the decision trees generated during the training process.

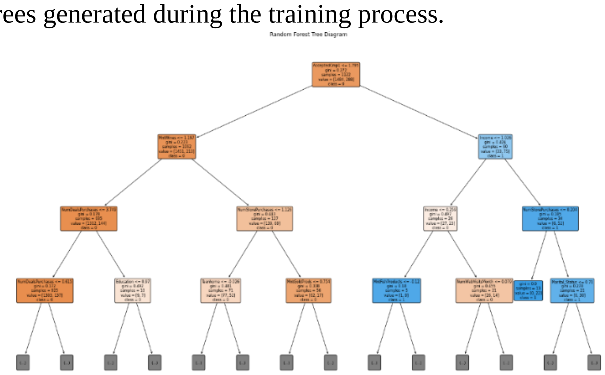

FIGURE 5. Random Forest Tree

As shown in figure 5, the decision tree divides customer data into several branches based on variables such as income, product purchases, and campaign acceptance history. Each branch represents a classification rule used by the Random Forest model to predict whether customers are likely to respond to marketing campaigns.

Model Evaluation Results

The model evaluation was carried out using several classification metrics, including accuracy, precision, recall, and F1 score, to measure the effectiveness of the Random Forest model in predicting customer retention. The results of the evaluation are presented in the following visualization.

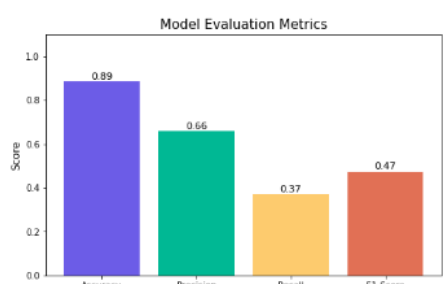

FIGURE 6. Model Evaluation Results

As shown in figure 6, the Random Forest model achieved strong classification performance with an accuracy value of 0.89. F1's precision, memory, and score values of 0.66, 0.37, and 0.47 suggest that the model is relatively effective in predicting customer responses, although some positive responses have still not been successfully detected.

Confusion Matrix

To further analyze the prediction performance, a confusion matrix is generated using a test dataset. The confusion matrix compares the actual customer response with the predictive classification generated by the Random Forest model.

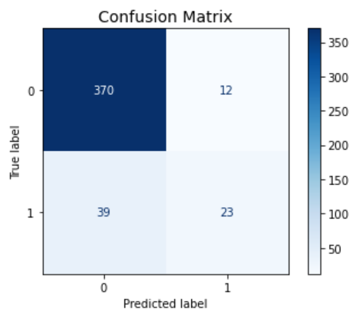

FIGURE 7. Confusion Matrix

As shown in figure 7, the confusion matrix results from the test dataset. This model successfully classifies most of the unresponsive customers correctly while still identifying the portion of customers who responded positively to the marketing campaign. Although some positive responses remain undetected, the model shows a good ability to minimize false positive predictions.

Discussion

The Random Forest model achieved an accuracy of 0.89, precision of 0.66, recall of 0.37, and an F1-score of 0.47, indicating good overall performance in predicting customer responses to marketing campaigns. Variables related to income, recency, and spending behavior were identified as important contributors to the prediction process. The relatively low recall may be influenced by class imbalance in the dataset.

The results demonstrate that Random Forest is capable of capturing customer behavior patterns and generating reliable predictions for marketing applications. Nevertheless, further improvements such as class balancing techniques and hyperparameter optimization may help increase the model's ability to detect responsive customers.

From a practical perspective, the proposed model can support marketing decision-making by identifying customer segments with a higher probability of responding to promotional campaigns. This approach may improve campaign targeting efficiency and marketing resource allocation.

CONCLUSION

This study demonstrates that the Random Forest algorithm can effectively predict customer responses to marketing campaigns using demographic characteristics, purchasing behavior, and previous campaign interaction data. The model achieved an accuracy of 0.89, precision of 0.66, recall of 0.37, and an F1-score of 0.47, indicating good overall classification performance. In addition, income, recency, and spending behavior were identified as important variables influencing customer response predictions.

By identifying customers who are more likely to respond to marketing campaigns, organizations can allocate marketing resources more efficiently and enhance campaign performance. Companies can utilize customer demographic characteristics, purchasing behavior, and campaign interaction history to identify potential customers more accurately and develop more personalized marketing strategies.

Proceedings of the Perbanas International Seminar on Economics, Business, Management, Accounting and IT
[June 2026] [ISSN 3047-0951]

REFERENCES

Anant, P., Shweta, G., Ghandat, S., & Ghorpade, S. (2025). Fraud Detection Using Machine Learning. 11(05), 184–190.

Brito, A., Feio, A., Andriolo, B., Penha, E., Santos, J., Nobre, M., Santis, M., Arruda, R., Feio, R., & Ribeiro, S. (2021). Aplicabilidade do uso de tecnologia móvel para registros em saúde durante a assistência pré-natal. Revista Eletrônica Acervo Saúde, 13, e6758. https://doi.org/10.25248/reas.e6758.2021

Campaigns, M. (2024). Predictive Modeling of Customer Response to Marketing Campaigns.

James, B., Denis, K., & Peter, D. (2024). Implementing Artificial Intelligence in Traditional B2B Marketing Practices : An Activity Theory Perspective. Information Systems Frontiers, 26(3), 1025–1039. https://doi.org/10.1007/s10796-022-10294-1

Ji, Y., Sun, A., Zhang, J., & Li, C. (2022). A Critical Study on Data Leakage in Recommender System Offline Evaluation. ACM Transactions on Information Systems, 41. https://doi.org/10.1145/3569930

Kumar, V., Reinartz, W., & Edition, T. (n.d.). Customer Relationship Management.

Mittal, S., & Kumar, V. (2024). Analyzing the role of technological capabilities and digital marketing on the performance of e-commerce-based SMEs. Indian Journal of Marketing, 54(12), 45–60.

Rocha, S., Fernando, J., Melo, B., Ramos, D., Carvalho, R., Silva, R., Pains, A., Cristina, G., Chung, S., Oliveira, L. L. De, Schmidt, D., Fortes-silva, R., & Copatti, C. E. (2026). Natural Dietary Supplementation with Elionurus muticus Essential Oil Enhances Growth Performance and Modulates Physiological Responses to Transport Stress in Nile Tilapia. 1–23.

Sarker, I. H. (2021). Machine Learning : Algorithms , Real - World Applications and Research Directions. SN Computer Science, 2(3), 1–21. https://doi.org/10.1007/s42979-021-00592-x

Singh, M. (2024). Machine Learning in Marketing Analytics. 13(4).

Sujon, K. M., Hassan, R., Choi, K., & Samad, A. (2025). empirical evidence from advanced statistics , ML , and XAI for evaluating business predictive models.