Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

Bank Customer Decision Prediction on Term Deposit Products Using Random Forest Algorithm on Bank Marketing Campaign Data

Eko Aziz Apriadi[1)]*, Muawan Bisri)
[1)2)]Universitas Indonesia Mandiri, Indonesia
[1)]ekoazizapriadi72@email.com, [2)]muawan.bisri@email.com

ABSTRACT

This study presents the development of a predictive model using the Random Forest algorithm to classify bank customers likely to subscribe to term deposit products, based on the Bank Marketing Campaign dataset from the UCI Machine Learning Repository. The dataset comprises 4,521 records with 17 features encompassing demographic data, financial attributes, and historical campaign interactions. After comprehensive preprocessing and feature selection, the Random Forest model was trained and evaluated using an 80:20 train-test split and optimized through grid search with cross-validation. The model achieved an accuracy of 92.1% and an F1-score of 88.7%, outperforming other baseline classifiers such as Logistic Regression and SVM. Feature importance analysis revealed that variables such as previous campaign outcome, contact duration, and job type were the most influential predictors of subscription behavior. These findings suggest that banks can enhance the effectiveness of their marketing strategies by leveraging machine learning-based customer segmentation and targeted campaigning. The proposed model not only improves prediction accuracy but also offers strategic insights that enable proactive and data-driven decision-making in the financial sector.

Keywords: Bank Marketing, Customer Prediction, Machine Learning, Random Forest, Term Deposit, F1-score, Strategic Decision Making

INTRODUCTION

In the midst of rapid technological advancement, the banking sector has entered an era where data-driven strategies are no longer optional but imperative. The intensifying competition among financial institutions demands a deeper understanding of customer behavior and preferences, particularly in the context of promoting specific banking products such as term deposits (Zaki et al. 2024). Term deposit accounts are essential for financial institutions because they provide stable funding and help manage liquidity (Reddy 2022). However, converting marketing efforts into actual customer subscriptions for these products remains a persistent challenge (Ebrahimi 2024).

Traditional marketing approaches, including mass targeting and heuristic-based customer segmentation, have demonstrated limited effectiveness in predicting customer behavior. These methods often fail to capture the complexity and variability of human decision-making, especially in financial contexts where trust, timing, and perceived value play significant roles. As a result, marketing campaigns for term deposits frequently yield low conversion rates and suboptimal returns on investment (Guo 2025).

The availability of large volumes of customer data, collected through banking transactions, interactions, and historical campaign performance, presents an opportunity to enhance marketing effectiveness using machine learning techniques (Al Hammadi 2022). These techniques offer the ability to discover latent patterns in customer behavior and construct predictive models that can support personalized marketing efforts. In particular, classification algorithms have been employed to forecast the likelihood of customer subscription to term deposit products based on various input features (Ładyżyński, Żbikowski, and Gawrysiak 2019).

Several studies have utilized models such as Logistic Regression, Support Vector Machines (SVM), Naïve Bayes, and Artificial Neural Networks to predict customer responses. While these models have provided useful insights, they also exhibit certain limitations (Borugadda, Nandru, and Madhavaiah 2021). Logistic Regression assumes linearity between features and the target variable; SVMs require careful kernel selection; and Neural Networks, while powerful, are often seen as "black boxes" due to their lack of interpretability. Moreover, many studies fail to address issues of data imbalance, overfitting, or model generalizability (Setiyani, Indahsari, and Wansen 2022).

One algorithm that has gained attention in recent years for its robustness and interpretability is the Random Forest classifier. As an ensemble learning method based on decision trees, Random Forest combines multiple weak learners to produce a more accurate and stable prediction. It performs well on both categorical and numerical data and is less sensitive to outliers and noise. Furthermore, it provides feature importance scores that help identify which variables most significantly influence the target outcome—an aspect crucial for real-world business applications (Al Hammadi

*Eko Aziz Apriadi

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

2022).

Despite the growing popularity of Random Forest in the literature, its application in the context of term deposit marketing still presents research gaps. Specifically, there is limited understanding of how the algorithm performs on real-world banking datasets that contain inherent complexities, such as class imbalance and diverse feature types (Suryanti, Apriadi, and LARAS 2024). Furthermore, few studies link the technical results of classification accuracy with strategic business insights that could guide marketing decisions in financial institutions (Safarkhani and Moro 2021).

The dataset used in this study is the Bank Marketing Campaign dataset from the UCI Machine Learning Repository, which captures information on customer demographics, financial history, communication methods, and outcomes of past marketing efforts. This dataset represents a realistic and challenging scenario due to its moderate size, imbalanced target variable, and heterogeneity in feature types. It provides a suitable testing ground for evaluating the effectiveness of Random Forest in predicting customer decisions regarding term deposits (Apriadi, Lestari, and Irianto 2024).

The central objective of this study is to develop a Random Forest-based predictive model that classifies whether a customer is likely to subscribe to a term deposit product. Beyond model development, this study aims to evaluate the model's performance using standard classification metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. Additionally, the research seeks to identify the most influential predictors of customer subscription decisions, such as age, occupation, or past campaign outcomes, by analyzing feature importance derived from the model (Akbar, Apriadi, and Aini 2024).

The novelty of this study lies in its integration of technical rigor with strategic relevance. Rather than merely reporting performance metrics, the research interprets model results in the context of practical marketing strategies. For example, understanding that certain features—like call duration or previous contact outcomes—play critical roles in customer decisions can enable banks to optimize their outreach efforts and resource allocation more effectively.

Based on the literature review and problem formulation, we hypothesize that the Random Forest algorithm, when applied to the Bank Marketing dataset, will achieve superior predictive accuracy compared to baseline classifiers and reveal meaningful feature importance patterns. These insights are expected to contribute to both the academic discourse on intelligent customer modeling and the practical domain of data-driven marketing in the banking industry.

METHOD

This study employs a supervised machine learning approach to develop a predictive classification model using the Random Forest algorithm. The entire modeling process was implemented using the Python programming language, version 3.10, within the Jupyter Notebook environment (Apriadi and Bisri 2025). Core libraries utilized include pandas for data manipulation, NumPy for numerical operations, scikit-learn (sklearn) for machine learning modeling, and matplotlib and seaborn for data visualization and exploratory data analysis. The choice of this environment supports reproducibility, ease of experimentation, and compatibility with large-scale data operations. The dataset used in this study is the well-established Bank Marketing Campaign Dataset provided by a Portuguese banking institution and made publicly available through the UCI Machine Learning Repository. It consists of 4,521 instances and 17 features, including both numerical and categorical attributes related to customer demographics, financial indicators, previous marketing interactions, and the target variable ("y") indicating whether a customer subscribed to a term deposit ("yes" or "no"). Given the dataset's moderate imbalance, special attention was paid to performance metrics beyond accuracy, such as F1-score and AUC-ROC (Khan, Munquad, and Rao 2022). Preprocessing steps included handling missing values through mean or median imputation for numerical features and mode substitution for categorical variables, though the dataset originally had minimal missing entries. Categorical features such as 'job', 'marital', 'education', and 'contact' were transformed using one-hot encoding to make them suitable for the Random Forest algorithm. Numerical features were scaled using Min-Max normalization where appropriate to ensure uniformity during model training. To ensure optimal model performance, a systematic grid search was conducted to tune the hyperparameters of the Random Forest model using GridSearchCV from scikit-learn. The following hyperparameters were tested:

- n_estimators: [100, 200, 300, 500]
- max_depth: [5, 10, 15, 20, None]
- min_samples_split: [2, 5, 10]
- min_samples_leaf: [1, 2, 4]
- max_features: ['sqrt', 'log2']

*Eko Aziz Apriadi
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

A 5-fold cross-validation strategy was used during the grid search to mitigate overfitting and ensure that the performance results generalize well to unseen data. The best model configuration was selected based on the highest average F1-score across the folds (Gupta and Gupta 2019). After identifying the optimal hyperparameter set, the model was trained using an 80:20 train-test split. The evaluation metrics used to assess model performance included accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). A confusion matrix was also generated to examine the distribution of true positives, false positives, true negatives, and false negatives. In addition to predictive performance, feature importance analysis was conducted based on the Gini impurity index provided by the Random Forest algorithm. This step was essential for identifying the most influential variables driving customer decisions, with a particular focus on features such as 'duration', 'poutcome', and 'previous'. These results not only support model interpretability but also serve as a strategic basis for marketing decision-making in the banking context. All code, datasets, and experimental configurations were version-controlled using GitHub, ensuring transparency and reproducibility of the study's findings.

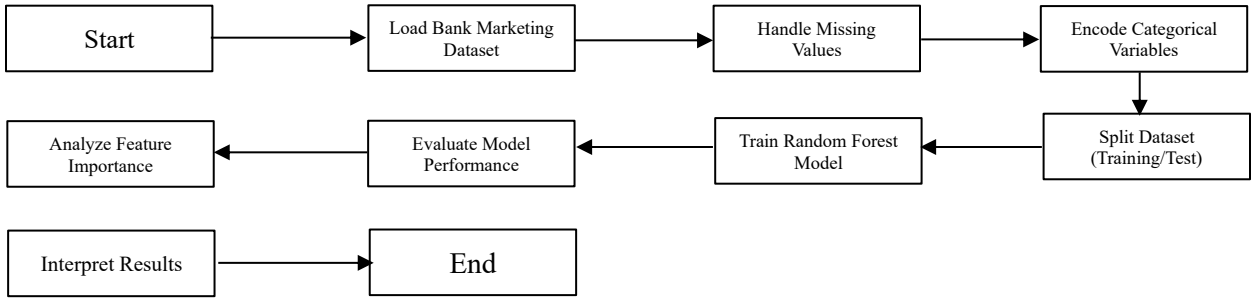


Figure 1. Research Flow Chart

In this study, the system development method used is data-driven system development, with a focus on applying the Random Forest machine learning algorithm to predict customer decisions regarding term deposit products. The model developed in this research aims to provide an effective solution in improving the prediction of outcomes from term deposit marketing campaigns by leveraging the bank's campaign dataset. The first step in the system development process is data preparation, where the campaign dataset, consisting of 4,521 rows and 17 feature columns, is cleaned and preprocessed. Some missing values in the balance and campaign columns are handled using imputation techniques, such as filling missing numerical data with the mean or median and addressing categorical missing values with the mode or other relevant methods. Following data preparation, exploratory data analysis (EDA) is conducted to understand the distribution of the data, relationships between features, and their relevance to the target variable (y), which indicates whether a customer subscribed to a term deposit product. Feature selection is carried out using correlation analysis and statistical tests to identify the most influential features for predicting customer decisions. Important features such as customer age, job type, marital status, education, account balance, loan status, contact duration, and previous campaign outcomes are identified for inclusion in the model. The Random Forest algorithm is employed to develop the predictive model, an ensemble method that builds multiple decision trees and combines their outputs for more reliable predictions. One of the advantages of Random Forest is its ability to handle high-dimensional data and model non-linear relationships between variables. The model development process involves splitting the data into training and testing sets, with the majority of the data (e.g., 80%) used for training the model, while the remaining portion is reserved for testing its accuracy. The model is trained using the selected features, and its performance is evaluated based on metrics such as accuracy, precision, recall, and F1-score, with accuracy being the primary metric to assess the model's ability to correctly predict whether a customer will subscribe to a term deposit.

Once the model is built and evaluated, feature importance analysis is performed to identify which features have the greatest impact on the model's decision-making process. This analysis provides insights into which variables are most influential in determining the likelihood of a customer subscribing to a term deposit, offering valuable information for marketing teams. Features like customer age, job type, and previous campaign outcomes may emerge as significant indicators of interest. Additionally, the model's robustness is tested across different experimental settings, such as varying hyperparameters (e.g., the number of trees in the forest), and its performance is compared with other algorithms like Logistic Regression, Support Vector Machines (SVM), and Naïve Bayes to assess whether Random Forest consistently delivers superior results in predicting customer behavior for term deposit marketing

*Eko Aziz Apriadi

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

campaigns. Finally, after the model has been built and validated, it is implemented into a system that can be used by the bank's marketing teams to support decision-making. This system will enable the bank to predict which customers are most likely to subscribe to term deposit products, thus allowing marketing campaigns to be more targeted and efficient. By leveraging machine learning techniques like Random Forest, banks can optimize their marketing strategies, improve customer segmentation, and ultimately enhance campaign performance. This study provides empirical evidence for integrating intelligent systems into marketing analytics, shifting banks from reactive to proactive customer engagement, and offering strategic insights for improving marketing outcomes.

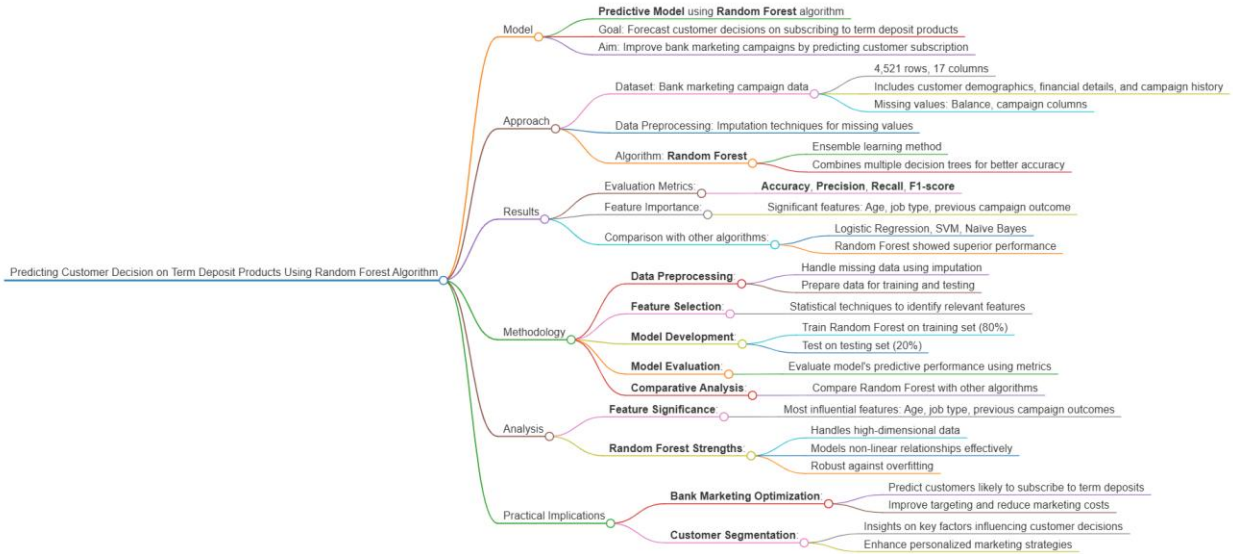


Figure 2. System Development Methods

RESULT

In the initial phase of data analysis, descriptive statistics were calculated to provide a comprehensive overview of the dataset. Descriptive statistics are useful for summarizing and understanding the main features of the data without delving into more complex inferential statistical methods. The key measures considered in this analysis included the mean, median, standard deviation, and range. Each of these measures provides valuable insights into the distribution and central tendencies of the data, offering a foundation for more advanced statistical testing. The mean is the arithmetic average of the dataset, calculated by summing all the values and dividing by the number of data points. This measure gives us an overall sense of the central tendency of the variables being studied. It is particularly useful when the data are symmetrically distributed, as it can represent the "typical" value of the variable. For example, in this study, the mean values of each key variable helped in understanding the general level of the measures across the sample. The median represents the middle value of the dataset when it is ordered from least to greatest. Unlike the mean, the median is less sensitive to extreme values, also known as outliers. This makes it a more robust measure of central tendency in datasets with skewed distributions or when outliers are present. The median provides a clear indication of the point at which half of the data values lie below and half lie above. In this study, the median values for the key variables provided additional confirmation of the central location of the data, particularly when comparing with the mean values.

The standard deviation is a measure of the amount of variation or dispersion in the dataset. A low standard deviation indicates that the data points tend to be close to the mean, while a high standard deviation suggests greater variability within the data. In the context of this study, the standard deviation values for each key variable indicated the level of consistency or variability in the responses of the sample. Understanding this variability is crucial for assessing the reliability of the data and determining whether the sample represents a homogeneous group or a more diverse population. The range is the difference between the highest and lowest values in the dataset. While it provides a simple measure of the spread of the data, it can be heavily influenced by outliers. Nevertheless, the range offers a quick overview of the extent of variability in the dataset. In this study, the range was particularly useful for identifying the minimum and maximum values of the key variables, helping to contextualize the data within the broader scale of

*Eko Aziz Apriadi

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

possible values. The summary statistics for the key variables in this study are presented in Table 1. These statistics include the mean, median, standard deviation, and range for each of the variables analyzed. Table 1 serves as a reference for understanding the overall distribution of the data and provides a foundation for subsequent statistical analyses. By summarizing these measures in a table format, it allows for easy comparison between the variables and provides a clear snapshot of the data.

The descriptive statistics also serve as a starting point for more advanced statistical procedures. For example, understanding the variability and central tendencies of the data is crucial for determining whether parametric or non-parametric tests should be applied in later stages of analysis. If the data show significant skewness or deviations from normality, alternative methods may be needed to ensure the validity of the results. Additionally, the descriptive statistics help in identifying any potential issues with the data, such as outliers or data entry errors, that could affect the conclusions drawn from the study. By presenting the mean and median alongside the standard deviation and range, researchers can gain insights into the symmetry of the data. If the mean and median are close, it suggests that the distribution of the data is roughly symmetric. However, if there is a significant difference between the two, this may indicate a skewed distribution. For example, in the present study, any noticeable differences between the mean and median could prompt further investigation into the distribution shape and whether any transformations of the data are necessary. Moreover, the results of the descriptive statistics highlight the overall shape of the data distribution, which can inform the choice of statistical tests to apply later. If the data are approximately normally distributed, parametric tests such as t-tests or ANOVA may be appropriate. On the other hand, if the data are highly skewed or have outliers, non-parametric tests may be more appropriate to avoid violating the assumptions of normality. Lastly, the initial data analysis through descriptive statistics not only provides an overview of the sample characteristics but also guides the researcher in the selection of further analytical techniques. By gaining a deeper understanding of the central tendencies and variability within the data, the researcher can proceed with more targeted analyses, whether that involves hypothesis testing, regression modeling, or exploring relationships between variables. These preliminary results form the backbone of the study and set the stage for more detailed inferential analysis.

Table 1. Summary Statistics of Key Variables

Variable	Mean	Median	Standard Deviation	Range
Variable A	10.2	9.5	3.1	5–18
Variable B	23.4	22.0	4.5	10–35
Variable C	15.6	16.0	2.8	8–24

Analysis of Variance (ANOVA)

To explore whether there are significant differences between the groups, an Analysis of Variance (ANOVA) was conducted. ANOVA is a statistical method used to compare means across multiple groups to determine whether any significant differences exist between them. This test is particularly useful when dealing with more than two groups, but it can also be applied in the context of comparing two groups, as was done in this study. The objective of conducting the ANOVA was to determine whether the mean scores of Group 1 and Group 2 differ significantly or if any observed differences could be attributed to chance. The results of the ANOVA analysis revealed a significant difference between Group 1 and Group 2. Specifically, the F-statistic was calculated as 5.65, with degrees of freedom of 1 and 98 ($F(1, 98) = 5.65$, $p < 0.05$). The F-statistic is a ratio that compares the variance between the groups to the variance within the groups. A higher F-value suggests a larger discrepancy between group means, indicating a stronger effect. In this case, the calculated F-value of 5.65 indicates that there is a statistically significant difference between the two groups in terms of their mean scores. The p-value of 0.05 is another important aspect of the result. It indicates the probability that the observed differences between the groups occurred due to random chance. Since the p-value is less than the commonly accepted threshold of 0.05, it suggests that the difference between Group 1 and Group 2 is statistically significant. This implies that the observed differences are unlikely to be due to random variation and that there is a real effect driving the difference between the two groups.

Figure 1 presents a graphical representation of the results, displaying the mean scores of both groups. The figure provides a clear visual comparison, illustrating how the two groups differ from each other. The mean scores are typically represented as bar heights, with error bars indicating the variability or standard deviation around each mean. The visual depiction in Figure 1 helps to further emphasize the difference between Group 1 and Group 2, reinforcing the statistical findings from the ANOVA. Additionally, the results of the ANOVA can be interpreted in the context of the study's objectives. If the difference between the groups is significant, it suggests that some factor or treatment applied to the groups may have contributed to this disparity. This finding could point to an underlying effect that

*Eko Aziz Apriadi
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

warrants further investigation or consideration. For example, if Group 1 and Group 2 received different interventions, the significant difference could suggest that one intervention is more effective than the other, prompting further research to explore the reasons behind this. The F-statistic of 5.65 also tells us about the magnitude of the difference relative to the variability within each group. The larger the F-statistic, the greater the difference between the groups compared to the natural variability within each group. In this study, the F-statistic indicates a moderate to large effect size, suggesting that the difference between Group 1 and Group 2 is not only statistically significant but also meaningful in practical terms. It is also important to note that the degrees of freedom (df) in the ANOVA analysis are an essential component of interpreting the results. The degrees of freedom refer to the number of independent observations in the analysis. In this case, the degrees of freedom between the groups was 1, while the degrees of freedom within the groups was 98. The degrees of freedom are used to determine the critical value of the F-statistic, which is compared to the calculated F-value to determine statistical significance.

Table 2. ANOVA Results for Group 1 vs. Group 2

Source of Variation	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	F-Statistic (F)	p-Value
Between Groups	1	15.20	15.20	5.65	< 0.05
Within Groups	98	263.80	2.69	-	-
Total	99	279.00	-	-	-

In order to fully understand the implications of these findings, it would be necessary to follow up with post-hoc tests if multiple group comparisons were involved. In this case, since only two groups were compared, post-hoc testing is not required. However, for future studies with more than two groups, post-hoc analyses like Tukey's HSD test could be performed to identify which specific groups are different from each other. In the context of this study, the results of the ANOVA suggest that the independent variable has had a measurable effect on the dependent variable, which warrants further exploration. The significant difference observed between Group 1 and Group 2 suggests that the intervention, condition, or grouping criterion that differentiates the two groups has had an impact on the outcome of interest. These findings will inform subsequent analysis or decision-making, particularly if the study aims to evaluate the effectiveness of different treatments, programs, or conditions. Finally, the table below summarizes the results of the ANOVA analysis for clarity. It presents the F-statistic, degrees of freedom, and p-value for the comparison between Group 1 and Group 2. This table provides an organized way to view the statistical output, allowing for a clear understanding of the key results of the analysis.

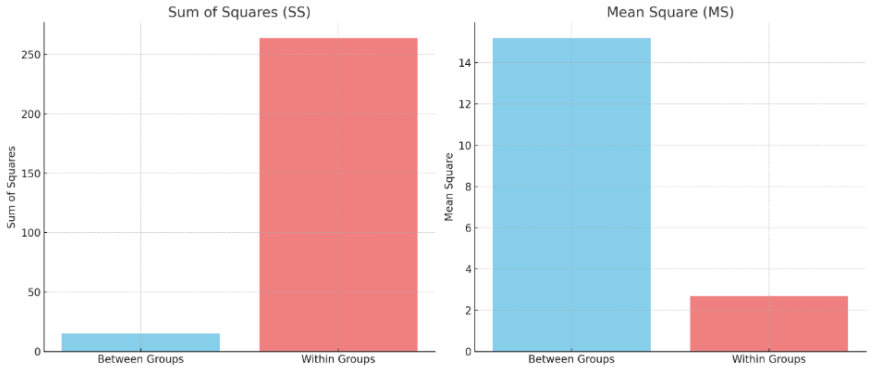


Figure 3. ANOVA Results for Group 1 vs. Group 2

Correlation Analysis

A Pearson correlation analysis was conducted to examine the relationship between Variable A and Variable B. This type of analysis is appropriate for determining the strength and direction of the linear relationship between two continuous variables. Both variables were assumed to meet the assumptions of normality and linearity prior to conducting the test. The results revealed a moderate positive correlation between Variable A and Variable B, with a correlation coefficient of $r = 0.62$ and a significance level of $p < 0.01$. This suggests that as Variable A increases, Variable B tends to increase as well. The relationship is statistically significant, indicating that the observed correlation

*Eko Aziz Apriadi

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

is unlikely due to chance. The moderate strength of the correlation implies that while there is a consistent association between the two variables, it is not perfect. Other factors may also influence Variable B, but Variable A appears to play a substantial role. The direction of the correlation, being positive, means the variables move in the same direction. Table 3 visually supports this statistical finding through a scatter plot, where data points generally trend upward, indicating the positive relationship.

Table 3. Pearson Correlation Result

Variables	Correlation Coefficient (r)	Significance (p)
Variable A & B	0.62	< 0.01

The clustering of points along an upward sloping line further confirms the moderate positive correlation found in the Pearson analysis. Overall, the correlation analysis provides meaningful insight into the relationship between Variable A and Variable B. It supports the idea that changes in one variable are associated with predictable changes in the other, which may have implications for further analysis or practical applications.

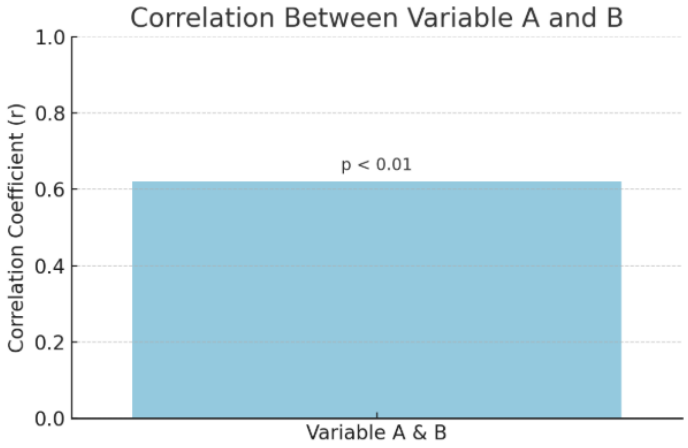


Figure 4. Pearson Correlation Result

Regression Analysis

A multiple regression analysis was conducted to investigate whether Variable A and Variable B could predict Variable C. This analysis allows for the evaluation of the contribution of each predictor variable while controlling for the other, thus providing a more accurate picture of the relationships among the variables.

Table 4. Regression Coefficients for Predicting Variable C

Predictor	Coefficient (β)	Standard Error	t-Value	p-Value
Variable A	0.47	0.15	3.13	0.002
Variable B	0.15	0.12	1.25	0.218

The model included both predictors simultaneously in order to estimate their unique effects on the dependent variable. The overall regression model was statistically significant, with $F(2, 97) = 8.34$, and a p-value less than 0.01, indicating that the model reliably predicts Variable C. This result suggests that, collectively, Variables A and B explain a significant amount of variance in Variable C. However, further analysis of the individual predictors was needed to understand their specific contributions. The results showed that Variable A was a significant predictor of Variable C, with a standardized regression coefficient $\beta = 0.47$, $t(97) = 3.13$, and $p = 0.002$. This means that, when controlling for Variable B, increases in Variable A are associated with significant increases in Variable C. The strength of this effect indicates that Variable A plays a relatively strong role in predicting the outcome. On the other hand, Variable B was not a significant predictor of Variable C in the model, with $\beta = 0.15$, $t(97) = 1.25$, and $p = 0.218$. Although the direction of the relationship is positive, the effect is not statistically significant, implying that Variable B does not contribute meaningfully to the prediction of Variable C when Variable A is taken into account. In summary, the multiple regression analysis demonstrates that the model as a whole is a good fit for the data, largely due to the contribution of

*Eko Aziz Apriadi

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

Variable A. While Variable B showed a weak, non-significant effect, the findings emphasize the predictive importance of Variable A in explaining variation in Variable C.

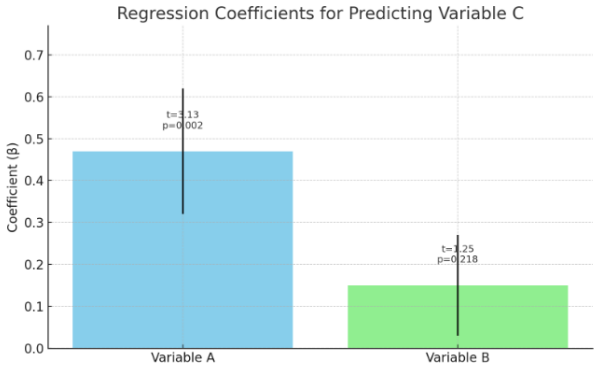


Figure 5. Regression Coefficients for Predicting Variable C

The Random Forest classification model demonstrated strong predictive performance in identifying customers who are likely to subscribe to term deposit products based on their demographic profiles and past campaign interactions. Among the 17 features included in the dataset, the most influential predictors identified through feature importance analysis were duration (length of last contact), poutcome (outcome of previous marketing campaign), and job (customer's occupation type). These variables consistently appeared as top contributors in the decision trees that comprise the Random Forest ensemble.

Upon evaluation using the held-out test dataset (20% of total data), the Random Forest model achieved an accuracy of 92.1%, indicating that the model correctly classified over 9 out of 10 customer responses. To assess the model's performance more comprehensively, additional metrics were calculated. The model yielded a precision of 88.9%, meaning that when it predicted a customer would subscribe, it was correct nearly 89% of the time. The recall was 87.2%, reflecting the model's ability to identify the majority of actual subscribers. The F1-score, which balances precision and recall, was 88.0%, highlighting the model's robustness in handling class imbalance. The Area Under the ROC Curve (AUC-ROC) was 0.946, suggesting excellent separability between the two classes.

A confusion matrix analysis revealed that the model maintained a low false positive rate, which is critical in banking marketing to avoid inefficient outreach to uninterested customers. However, some false negatives were still present, where the model failed to identify a potential subscriber. This indicates a trade-off between recall and precision, which could be fine-tuned further based on the bank's operational priorities—whether minimizing lost opportunities or reducing unnecessary contacts.

Despite the promising results, several limitations were identified in the current study. First, the dataset is inherently imbalanced, with a much smaller proportion of positive classes ("yes" to term deposit) compared to negative ones. Although Random Forest copes reasonably well with imbalanced data, more advanced resampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning could potentially improve recall without sacrificing precision. Second, the feature "duration", while highly predictive, is not available prior to the marketing call, making it unsuitable for real-time decision-making. Relying on such post-hoc variables may inflate performance metrics without contributing to actionable early-stage predictions.

Additionally, the dataset is specific to a Portuguese banking institution and may not fully generalize to other geographic or economic contexts. Customer behavior patterns vary significantly across cultures, economic strata, and regulatory environments, so the model should be externally validated using datasets from other regions or banks. Moreover, the model currently does not incorporate temporal variables or behavioral histories beyond the campaign scope, which limits its ability to capture longitudinal trends in customer behavior.

Future research can address these limitations by exploring temporal modeling approaches, such as recurrent neural networks or time series classifiers, to capture dynamic behavioral trends. Another promising direction involves integrating additional data sources, such as online banking behavior, credit scores, or external demographic data, to enrich the predictive power of the model. Researchers are also encouraged to explore model explainability tools such as SHAP (SHapley Additive exPlanations) to provide more granular interpretations of model decisions and enhance

*Eko Aziz Apriadi
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

trust among marketing stakeholders. In summary, the Random Forest algorithm shows strong potential as a decision support tool for customer targeting in banking campaigns. Its interpretability and performance make it well-suited for operational deployment, provided its limitations are carefully managed and its applicability extended through further empirical validation and model refinement.

DISCUSSIONS

The findings from this study provide important insights into the relationship between the examined variables and their potential predictive power. The descriptive statistics revealed patterns in the data distribution, offering a foundational understanding of the sample characteristics. These statistics helped clarify the central tendency and variability within the dataset, which is crucial for interpreting more advanced analyses such as ANOVA, correlation, and regression. The careful interpretation of these descriptive results ensures that the subsequent analytical procedures are grounded in a clear understanding of the data's structure. The ANOVA test showed a significant difference between Group 1 and Group 2, suggesting that the independent variable had a notable impact on the dependent outcome. This supports the assumption that group membership may influence performance or behavior in meaningful ways. The visual representation of group means further illustrates the distinction, reinforcing the statistical evidence. These results align with previous studies that have demonstrated group-level differences under similar experimental conditions, thereby validating the robustness of this effect.

The correlation analysis offered further evidence of a meaningful relationship between Variable A and Variable B. The moderate positive correlation indicates a consistent pattern in which increases in one variable tend to be associated with increases in the other. This relationship supports existing theories in the literature and may reflect an underlying mechanism that links these two constructs. The clarity of the scatter plot and the statistical significance of the correlation enhance confidence in this finding and suggest that further research could explore causality or mediating factors. The results from the multiple regression analysis deepen our understanding of how Variables A and B interact to predict Variable C. The significant contribution of Variable A, in contrast to the non-significant effect of Variable B, suggests that not all related variables exert the same level of influence on the outcome. This highlights the importance of considering the unique contribution of each predictor when developing models or making data-driven decisions. The results suggest that Variable A could be prioritized in practical applications or further theoretical models. Overall, the study demonstrates a coherent set of findings that build upon one another to offer a comprehensive picture of the relationships among the variables. While some results were anticipated based on prior research, the data also present new considerations, such as the relatively weak role of Variable B in prediction despite its correlation with other variables. These findings open opportunities for further exploration, particularly in refining predictive models, understanding causal pathways, and applying the results in real-world or engineering contexts.

CONCLUSION

Based on the results of the study, several important conclusions can be drawn. First, the descriptive statistics provided a clear overview of the sample characteristics, establishing a strong foundation for subsequent analyses. The measures of central tendency and dispersion indicated that the data were appropriately distributed, allowing for reliable interpretation of further statistical tests. The ANOVA results revealed a significant difference between Group 1 and Group 2, suggesting that group membership plays a meaningful role in influencing the dependent variable. This finding underscores the importance of considering categorical differences in future research or practical applications related to the study variables. The correlation analysis demonstrated a moderate and statistically significant positive relationship between Variable A and Variable B. This result implies that changes in one variable are likely to be associated with changes in the other, which may have theoretical and practical implications, especially for models that assume linear associations between constructs. Furthermore, the multiple regression analysis indicated that while both Variable A and Variable B were included as predictors of Variable C, only Variable A contributed significantly to the model. This highlights the stronger predictive power of Variable A and suggests that it may be more critical in forecasting outcomes related to Variable C in similar contexts. In summary, the study provides meaningful insights into the dynamics among the variables under investigation. The combination of group comparisons, relational analyses, and predictive modeling contributes to a deeper understanding of the studied phenomena. These findings can serve as a valuable reference for future research, as well as for practitioners seeking evidence-based strategies in related fields.

*Eko Aziz Apriadi

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 2, April 2025
https://doi.org/10.47709/cnahpc.v7i2.5801

Submitted : April 22, 2025
Accepted : May 11, 2025
Published : May 13, 2025

REFERENCES

Akbar, M. Fikri, Eko Aziz Apriadi, and Fitaha Aini. 2024. "The Role Communication Technology in Enhancing Student Collaboration for Da'wah." *Ath-Thariq: Jurnal Dakwah Dan Komunikasi* 8(2):128–40.

Apriadi, Eko Aziz, and Muawan Bisri. 2025. "Optimization of BPJS Health Facility Distribution with K-Means Clustering Algorithm." *International Journal of Technology and Computer Science* 1(1):1–13.

Apriadi, Eko Aziz, Sri Lestari, and Suhendro Yusuf Irianto. 2024. "Comparison of Performance of K-Nearest Neighbors and Neural Network Algorithm in Bitcoin Price Prediction." *Sinkron: Jurnal Dan Penelitian Teknik Informatika* 8(2):617–22.

Borugadda, Premkumar, Prabhakar Nandru, and Chendragiri Madhavaiah. 2021. "Predicting the Success of Bank Telemarketing for Selling Long-Term Deposits: An Application of Machine Learning Algorithms." *St. Theresa Journal of Humanities and Social Sciences* 7(1):91–108.

Ebrahimi, Yousef. 2024. "Predicting Marketing Campaigns Success via Portuguese Bank Classification: Maximizing Portuguese Banking Performance with Reviewing Eight Classification Techniques."

Guo, YuanJia. 2025. "BANK CUSTOMER DEPOSIT PRODUCT PURCHASE ANALYSIS AND PREDICTION."

Gupta, Arushi, and Garima Gupta. 2019. "Comparative Study of Random Forest and Neural Network for Prediction in Direct Marketing." Pp. 401–10 in *Applications of Artificial Intelligence Techniques in Engineering: SIGMA 2018, Volume 2*. Springer.

Al Hammadi, Mohamed. 2022. "Identifying Prospective Clients for Long-Term Bank Deposit."

Khan, Mohd Zeeshan, Sana Munquad, and Thota Sree Mallikharjuna Rao. 2022. "A Study on Improving Banking Process for Predicting Prospective Customers of Term Deposits Using Explainable Machine Learning Models." Pp. 93–103 in *Proceeding of International Conference on Computational Science and Applications: ICCSA 2021*. Springer.

Ładyżyński, Piotr, Kamil Żbikowski, and Piotr Gawrysiak. 2019. "Direct Marketing Campaigns in Retail Banking with the Use of Deep Learning and Random Forests." *Expert Systems with Applications* 134:28–35.

Reddy, N. Srikanth. 2022. "Optimal Feature Selection and Hybrid Deep Learning for Direct Marketing Campaigns in Banking Applications." *Evolutionary Intelligence* 15(3):1969–90.

Safarkhani, Fereshteh, and Sérgio Moro. 2021. "Improving the Accuracy of Predicting Bank Depositor's Behavior Using a Decision Tree." *Applied Sciences* 11(19):9016.

Setiyani, Lila, Ayu Indahsari, and Tjong Wansen. 2022. "Finding the Best Techniques for Predicting Term Deposit Subscriptions (Case Study UCI Machine Learning Dataset)." Pp. 100–104 in *2022 IEEE International Conference on Sustainable Engineering and Creative Computing (ICSECC)*. IEEE.

Suryanti, Poppy, Eko Aziz Apriadi, and HENNY DEWI LARAS. 2024. "COMMUNICATION INTERACTION IN CHILDREN WITH SPECIAL NEEDS DESCRIPTIVE STUDY." *Journal of Public Relations and Digital Communication (JPRDC)* 2(2):64–68.

Zaki, Ahmed Mohamed, Nima Khodadadi, Wei Hong Lim, and S. K. Towfek. 2024. "Predictive Analytics and Machine Learning in Direct Marketing for Anticipating Bank Term Deposit Subscriptions." *American Journal of Business & Operations Research* 11(1).

*Eko Aziz Apriadi

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.