Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Classification of Spending Segmentation in Mobile Game Applications Using Random Forest and Decision Tree Algorithms

Dewa Restu Putra Wicaksana[1], Rangga Anom[2*], Syahrina Musyarafah[3], Giatika Chrisnawati[4]

[1,2,3,4] *Program Studi Infirmatika, Fakultas Teknik dan Informatika, Universitas Bina Sarana Informatika,Indonesia*
17230936@bsi.ac.id [1], 17230546@bsi.ac.id [2*], 17230130@bsi.ac.id [3], giatika.gcw@bsi.ac.id [4]

Abstract

This research aims to classify spending segmentation in mobile game users using Random Forest and Decision Tree algorithms. The dataset consists of demographic attributes, gameplay behavior, session frequency, and historical spending records. Several preprocessing steps uwere applied, including missing value handling, label encoding, one-hot encoding, and feature scaling. The data were divided into an 80:20 training-testing ratio, and hyperparameter tuning was performed using GridSearchCV. The results indicate that Random Forest achieved higher accuracy compared to Decision Tree, demonstrating better generalization for multiclass segmentation (Low, Medium, High spenders). This study shows the potential of machine learning in predicting user spending behavior to support data-driven monetization strategies in mobile game applications.

Keywords: *Classification, Spending Segment, Random Forest, Decision Tree, Mobile Game, Machine Learning.*

1. Introduction

Mobile games have become one of the fastest-growing sectors in the digital entertainment industry, generating significant revenue primarily through in-app purchases. The rapid growth of mobile gaming is supported by the widespread use of smartphones and the increasing popularity of free-to-play business models. In this model, user spending behavior varies widely, ranging from players who rarely spend money to those who consistently make high-value purchases [1]. [2] Understanding these differences is crucial for game developers, as spending behavior directly affects revenue sustainability, user retention, and monetization strategies. Therefore, identifying and classifying user spending patterns has become an important research topic in mobile game analytics.

Spending segmentation is a commonly used approach to categorize users based on their transaction behavior, such as Low, Medium, and High spenders [3] [1]. This segmentation allows developers to design personalized promotions, optimize in-game offers, and improve player engagement. Research by Lee [4] highlights that game design elements such as reward systems, progression mechanics, and in-app purchase features strongly influence player spending behavior. Highly engaged players tend to exhibit different purchasing patterns compared to casual players, making segmentation an essential step before applying predictive modeling techniques. By grouping users according to their spending levels, developers can focus their marketing and development efforts more effectively [5].

Machine learning techniques provide powerful tools to analyze large-scale and complex user data in mobile game applications [2]. Classification algorithms such as Decision Tree and Random Forest are widely used due to their ability to handle mixed data types and produce interpretable results. Previous studies have demonstrated the effectiveness of these algorithms in various classification tasks. [6] showed that both Decision Tree and Random Forest perform well in classification problems, although their performance may vary depending on data characteristics. Additionally, [7] demonstrated that Random Forest, combined with data balancing techniques such as SMOTE, can improve classification performance in datasets with class imbalance, which is a common issue in spending segmentation problems.

Based on these considerations, this study focuses on classifying mobile game users into spending segments using Random Forest and Decision Tree algorithms. The dataset used includes demographic attributes, gameplay behavior, session activity, and historical spending records [4]. This research aims to compare the performance of both algorithms in terms of accuracy and generalization ability. By evaluating

and analyzing the results, this study seeks to provide insights into the suitability of Random Forest and Decision Tree models for spending segmentation in mobile game applications, as well as to support data-driven decision-making for effective monetization strategies .

2. Literature

2.1. Previous Research

Several results of literature reviews conducted by various research journals related to the research Classification of Spending Segmentation in Mobile Game Applications Using Random Forest and Decision Tree Algorithms:

Paper [1] by Atmaja (2025) discusses the use of the Random Forest algorithm to segment fashion product sales, addressing imbalanced data. To address this issue, the SMOTE method and parameter optimization were applied to improve model performance. The results showed that Random Forest was able to provide more stable and accurate classification performance after data balancing. This approach is relevant to the current research, as segmenting mobile game user spending also has imbalanced data between low- and high-spending users [7].

Paper [2] by Jehad & Yousif (2020) explains how the Random Forest and Decision Tree algorithms use a machine learning approach to classify fake news. The dataset was preprocessed and feature extracted using the TF-IDF method before classification. The results show that both algorithms perform well, with Decision Tree producing slightly higher accuracy than Random Forest on the dataset used. This shows the effectiveness of random forest and decision tree as the main algorithms in the classification process of game user spending segmentation [6].

Paper [3] by Lee (2025) explores how mobile game design, particularly in-app purchase features, can influence player spending behavior. The study explains that elements such as reward systems, game progression, and virtual item offerings play a crucial role in encouraging players to make in-app purchases. Highly engaged players tend to have different spending patterns than casual players, leading to user segmentation based on their spending behavior. This shows the importance of user segmentation before carrying out classification modeling such as Random Forest and Decision Tree [4].

Paper [4] by Sari et al. (2023) examined the use of the LightGBM and Random Forest algorithms in the process of classifying potential customers based on demographic and behavioral data. Both algorithms are decision tree-based methods capable of handling data with a large number of features and complex relationships between variables. The results showed that Random Forest has advantages in terms of model stability and interpretability, while LightGBM excels in processing speed and computational efficiency. These findings support the use of Random Forest as an effective classification method [8].

2.2. Spending Segmentation

Spending segmentation is the process of grouping users or customers based on their spending patterns and levels within a service or application [7]. This segmentation aims to understand differences in user transaction behavior, allowing management to make more informed decisions, particularly regarding marketing and monetization strategies. In the context of mobile gaming apps, spending segmentation typically divides users into several categories, such as Low Spenders, Medium Spenders, and High Spenders. Low-spending users generally play casually and rarely make in-app purchases [1]. Conversely, high-spending users tend to be more active, play frequently, and are more interested in paid features or premium items. Meanwhile, medium-spending users fall somewhere in between these two groups.

The spending segmentation process involves analyzing user data, such as playing frequency, session duration, number of transactions, and total spending [4]. This data is then processed using data analysis techniques or machine learning to automatically and accurately segment users. Implementing spending segmentation enables game developers to identify high-value users, increase player retention, and support data-driven decision-making in the development and management of mobile game applications.

2.3. Classification

Classification is a machine learning method that aims to group data into specific classes or categories based on patterns learned from previous data. The classification process utilizes attributes or features to predict class labels, such as grouping users into low, medium, or high spenders. This method is widely used because it can facilitate automated, data-driven decision-making [6].

2.4. Decision Tree

A Decision Tree is a classification algorithm that works by creating a decision tree structure, where each node represents a decision-making process based on a specific attribute value. This model is easy to understand and interpret, but it has the disadvantage of being prone to overfitting if the data used is complex and the tree depth is not restricted [6].

2.5. Random Forest

Random Forest is a development of Decision Tree that uses an ensemble approach, combining multiple decision trees to produce more accurate and stable predictions. By combining the results from multiple trees, Random Forest can reduce overfitting and improve model generalization, making it frequently used in classification problems involving complex data [7].

3. Research Methods

This section describes the methodology used in this study, starting from dataset preparation to model evaluation. The dataset used is ”Mobile Game InApp Purchase” in .csv format containing user demographic information, gameplay behavior, and spending-related attributes for spending segmentation analysis. Data preprocessing includes data cleaning, handling missing values, encoding categorical features, normalization, and splitting the dataset into training and testing sets using an 80:20 ratio. Classification models are developed using Decision Tree and Random Forest algorithms. The models are evaluated using accuracy, precision, recall, F1-score, and confusion matrix.

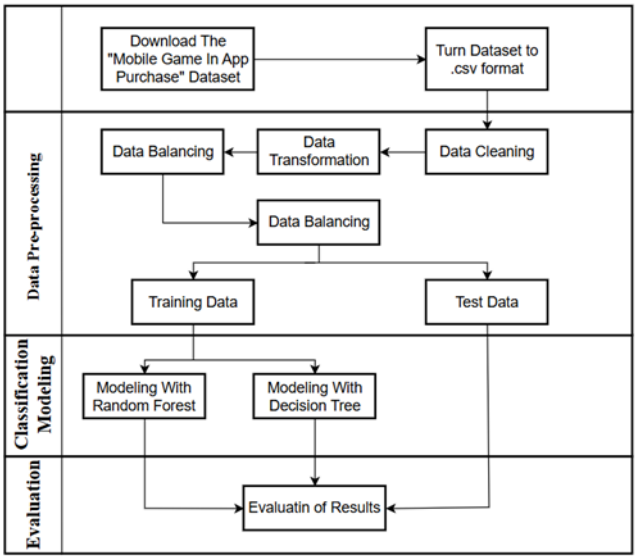

Fig. 1: Research Flow Diagram

3.1. Dataset Description

The dataset used in this study consists of several attributes related to the behavior and characteristics of mobile game players. These attributes include demographic information, such as age and country, gameplay behavior variables, including playtime hours and session frequency, as well as purchase history indicators, such as average session length and spending behavior.

The target variable in this study is SpendingSegment, which classifies users into three categories: Low, Medium, and High spenders.

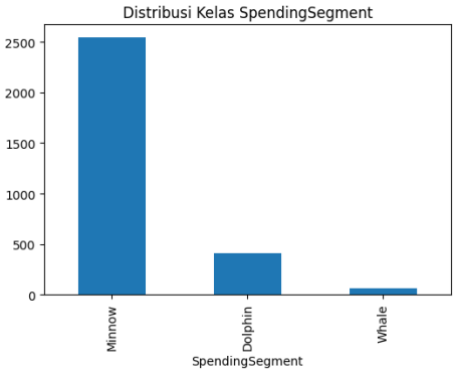

Fig. 2: Distribution of SpendingSegment Classes

Figure 2 shows that the Minnow class dominates the dataset, followed by Dolphin and Whale. This class imbalance may affect model performance and motivates the use of techniques such as stratified sampling or SMOTE.

3.2. Preprocessing

Data preprocessing steps include:

a. Removing irrelevant or leakage-prone columns (UserID, LastPurchaseDate, InAppPurchaseAmount).
b. Handling missing values using median for numerical attributes and mode for categorical attributes.
c. Applying Label Encoding to convert the target classes into numerical values.
d. Applying One-Hot Encoding to convert categorical features into binary representations.
e. Scaling numerical features using StandardScaler to standardize feature ranges.
f. Splitting the dataset into training and testing sets using an 80:20 ratio with stratified sampling to preserve class distribution.
g. Handling class imbalance using SMOTE (Synthetic Minority Oversampling Technique) when necessary.

These steps ensure the dataset is clean, consistent, and ready for machine learning model training.

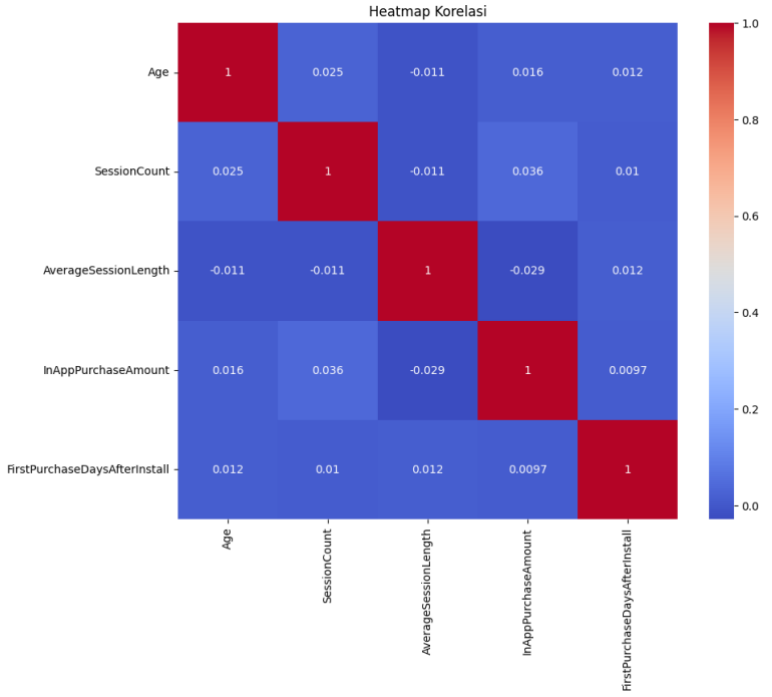

Fig .3: Correlation heatmap among numerical features

Figure 3 illustrates the correlation level among numerical attributes. Overall, the correlation values are relatively low (<0.05), indicating that the dataset does not contain multicollinearity problems.

3.3. Modeling

GTwo machine learning algorithms were used:

a. Random Forest

An ensemble learning algorithm that constructs multiple decision trees and aggregates their results. Hyperparameters tuned using GridSearchCV include:

a. Number of estimators
b. Maximum depth
c. Minimum samples split

b. Decision Tree

A tree-structured model that splits data based on entropy or gini criteria.
Hyperparameters tuned include:

a. Criterion

b. Maximum depth
c. Minimum samples split

Both models were evaluated using:

a. Accuracy
b. Precision
c. Recall
d. F1-Score
e. Confusion Matrix

Additionally, 5-Fold Cross Validation was applied to ensure reliability and robustness.

4. Results and Discussion

The Random Forest model achieved a higher accuracy compared to the Decision Tree model.

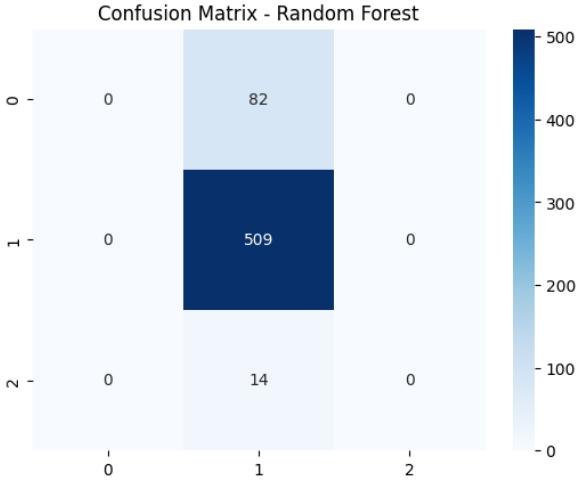

Fig .4: Confusion Matrix of Random Forest

As shown in Figure 4, the Random Forest model correctly classified most instances of the Minnow and Dolphin classes, although misclassification still occurred in minority classes due to data imbalance.

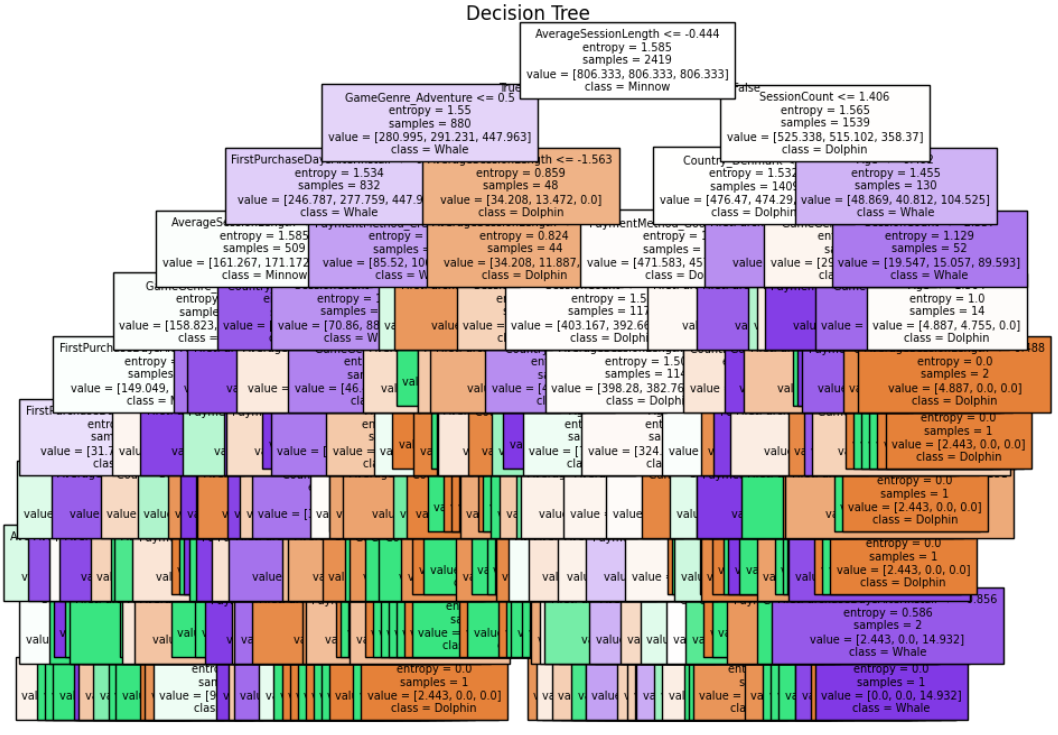

Fig. 5: Visualization of Decision Tree

Figure 5 presents the complete structure of the Decision Tree model. The tree contains many branches, reflecting a high level of model complexity that may lead to overfitting.

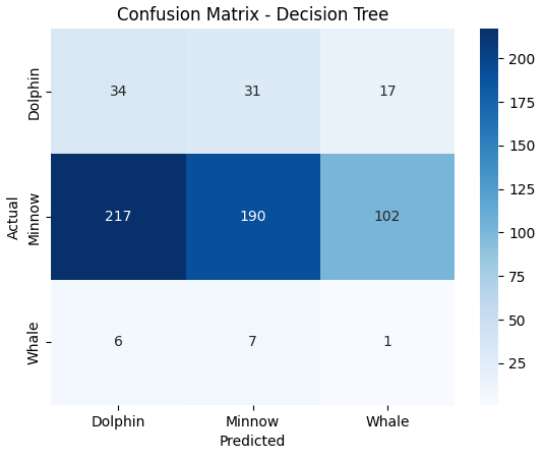

Fig .6: Confusion Matrix of Decision Tree

Figure 6 shows that the Decision Tree model performs less accurately compared to Random Forest, with a higher number of misclassifications across all classes, particularly between Minnow and Dolphin.

Based on the evaluation metrics:

a. Random Forest demonstrated better generalization and stability across all classes.
b. Decision Tree, while simpler and more interpretable, showed lower performance.
c. Feature importance analysis from Random Forest indicated that gameplay duration, session frequency, and average session length were the most influential variables in predicting spending segmentation.

Confusion Matrix results show that Random Forest classified more instances correctly across Low, Medium, and High classes.
The superior performance of Random Forest can be attributed to its ensemble learning mechanism, which reduces variance and prevents overfitting.
Decision Tree, although interpretable and easy to visualize, tends to overfit on training data and shows lower performance on unseen data.
These results indicate that Random Forest is more suitable for multiclass segmentation tasks in mobile game environments. The findings can support game developers in identifying high-value players and improving targeted marketing strategies.

5. Conclusion

This study demonstrates that the Random Forest algorithm provides better accuracy and generalization for spending segmentation in mobile game applications compared to the Decision Tree algorithm. Machine learning proves to be effective in analyzing player behavior and identifying spending patterns based on gameplay and demographic attributes.The insights obtained from this research can help game developers optimize monetization strategies, retain high-value users, and personalize in-game experiences.

Future work may include testing advanced algorithms such as Gradient Boosting, XGBoost, LightGBM, and experimenting with larger, real-world datasets for improved performance.

Acknowledgement

The author would like to express sincere gratitude to all parties who have supported the completion of this research. Special thanks to A.I and Machine Learning lecturer for their valuable insights, constructive feedback and continuous encouragement during the development of this study. Furthermore, appreciationis extended to family and friends for their unwavering moral support and motivation.This research would not have been possible without the inspiration and collaboration of those who have shared their time and knowledge. All contributions are deeply appreciated.

References

[1] S. M. Tham and G. P. Perreault, “A Whale of a Tale: Gaming Disorder and Spending and Their Associations With Ad Watching in Role-Playing and Loot-Box Gaming.” [Online]. Available: http://igi.camh.net/doi/pdf/xxxx
[2] F. J. Zagreb, “Machine Learning Based Models for Prediction of Player Behavior and Purchase Decisions in Digital Games,” 2018.
[3] M. Imani, “Evaluating Classification and Sampling Methods for Customer Churn Prediction under Varying Imbalance Levels.”
[4] G. K. S. Lee, “In-App Purchase Behaviour and Game Design in Mobile Gaming: A Review of Monetisation and Player Experience,” *International Journal of Mobile Applications and Technologies*, vol. 1, no. 1, May 2025, doi: 10.51137/wrp.ijmat.2025.glit.45789.

[5] J. Wu, "Distributed Intelligent Optimization of E-commerce User Purchase Data Mining using Spark Framework," *Informatica (Slovenia)*, vol. 48, no. 20, pp. 29–40, 2024, doi: 10.31449/inf.v48i20.6779.

[6] R. Jehad and S. A. Yousif, "Fake news classification using random forest and decision tree (J48)," *Al-Nahrain Journal of Science*, vol. 23, no. 4, pp. 49–55, Dec. 2020, doi: 10.22401/ANJS.23.4.09.

[7] G. T. Atmaja, "Enhancing Fashion Product Sales Segmentation Using Random Forest with SMOTE and Hyperparameter Optimization," *Jurnal Ilmu Komputer dan Informatika*, vol. 5, no. 1, pp. 45–56, Aug. 2025, doi: 10.54082/jiki.280.

[8] L. Sari, A. Romadloni, R. Lityaningrum, and H. D. Hastuti, "Implementation of LightGBM and Random Forest in Potential Customer Classification," *TIERS Information Technology Journal*, vol. 4, no. 1, pp. 43–55, Jun. 2023, doi: 10.38043/tiers.v4i1.4355.