Journal of Artificial Intelligence and Data Science
Vol. 1, No. 1, 2025, pp. 42-53
DOI:

Customer Loyalty Prediction Using RFM and Product Diversification Features Through Cluster-Derived NPS Proxy Labels

Alina Rusyda Hariadi, Wiyli Yustanti
Program Study of Information System, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia
e-mail address: alinarusyda.22003@mhs.unesa.ac.id (corresponding author)

Unung Istopo Hartanto
Master of Informatics Study Program, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia

Received: 25 September 2025 | Revised: 15 October 2025 | Accepted: 30 October 2025

ABSTRACT

Knowing and forecasting the level of customer loyalty is vital to generate operational efficiency and effective business strategy in the FMCG's distribution industry. This article combines behavioral clustering and supervised machine learning to derive a hybrid analytical model to predict customer loyalty with a Net Promoter Score (NPS) Proxy based on transactional behavior. Based on sales data in PT XYZ (Mojokerto) and real datasets, RFM (Recency, Frequency, Monetary), product diversification were used to feature acquisition of rawdata. Principal Component Analysis (PCA) was employed for dimensionality reduction, and subsequently K-Means, Agglomerative, and DBSCAN clustering techniques were compared by means of internal validity metrics. Agglomerative Clustering was the most successful algorithm with Silhouette Score (0.8919) and Calinski–Harabasz Index (2585.11), as well as low Davies–Bouldin Index (0.5266), which means it produced very compact and distinct clusters. These three clusters were translated into NPS Proxy segments: Detractor, Passive, and Promoter by applying to them different depending on their response behavior. Because of strong class imbalance (98.8% Detractors, 1.06% Passives, 0.12% Promoters), SMOTE was applied to balance the training set before classification. The grid search was applied to tune hyperparameters for KNN, SVM, Gradient Boosting, Logistic Regression and Random Forest machine learning models. In these classifiers, the Random Forest classifier produced highest prediction results for F1-macro = 1.00, Precision = 1.00 and Recall = 1.00 in both training and testing data indicating outstanding generalization with ease in discrimination of overfitting use case. In general, the proposed framework effectively translates transactional data into action-oriented loyalty insights and lead to dependable prediction of customer NPS Proxy labels. Implications of the results Findings underline the relevance of integrating RFM analysis, product-level behavior, clustering validation and machine learning classification in pursuing a scalable customer loyalty management for FMCG distribution environments.

Keywords-RFM; Clustering; NPS Proxy; Clustering; Classification; Customer Loyalty

I. INTRODUCTION

The issue of customer loyalty is one of the most determinant factors of firm performance nowadays (especially in consumer markets where competing based on cost is really hard due to low switching costs, and customers can easily replace your product/brand). It has been most recently evidenced that past research consistently suggests that customer loyalty is directly affected by customer experience, satisfaction and perceived service and relationship quality, all resulting in behavioural responses over time such as trust, repeat purchase and advocacy intention [1-4]. Yum and Kim [5] indicates that the success of business reflects on the customer satisfaction. Happy customers will grow in retention and referrals. In the age of internet commerce, the emergence of massive commercial transaction data has motivated companies to embrace a datadriven methodology to delve into customers in detail for more accurate targeting or retention [6-10].

The Net Promoter Score (NPS) is one of the most commonly used metrics to measure customer loyalty and has been shown empirically to strongly correlate with customer satisfaction, repurchase intention, and business growth [11-16]. Unfortunately, the traditional NPS approach is heavily reliant on customer surveys which can be biased, have low take rates and time sensitivity. Recent studies also show that NPS is largely determined by rather than customers overall long-term behavior and magnified by on their most recent experiences [11]. The result is that companies are limited in their ability to measure loyalty at scale, on an ongoing basis and using objective measurement.

Concurrent to the formulation of loyalty measurement, a popular form of convolution with full frame (Recency, Frequency and Monetary) features accordingly known as RFM is widely regarded for profiling customer value and activity [17, 18]. The model of RFM is applied to improve customer retention and profit in various sectors, like e-commerce, retail, SaaS [19]. In addition to RFM, literatures have pointed out that adding multidimensional product level behaviors (i.e., the variety in product purchasing, combining purchases, and category preferences) is necessary to form more richer customer segments [6-10]. Such enriched behavior characteristics enable organizations to understand more about purchase patterns, channeling engagement and diversity of consumption necessary for creating market-targeted marketing strategies [20, 21]. With the development of artificial intelligence, Machine Learning (ML) has emerged an important technique for predicting customer churn, purchase intention, CLV and loyalty [20–27]. Varieties of ML tools such as ensemble models, hybrid frameworks and balancing solutions have been extensively used to tackle the problems like nonlinear behavior of customers and imbalance in data which are prevalent in practical business scenarios [22, 24, 25]. Although ML has better predictive power, most study prefer to utilize data of NPS which is surveyed data and would limit scalability with objective measuring test across large volume of customers as target label, or just manually defined loyalty classes.

While the literature on RFM segmentation, NPS measurement, and ML-based loyalty prediction is rich, there are still several open questions. First, not many loyalty brand studies make use of behavioral segmentation to directly calculate the loyalty labels, they rather depend on a survey-based NPS measurement. Secondly, few studies integrate RFM with product diversification characteristics to cluster customer behavior. Finally, there is a lack of end-to-end system or model that links the clustering data with derived NPS proxy and machine learning based classification in realistic business practices such as FMCG distribution. Accordingly, there is a need for scalable data-driven algorithms to transform raw transactional data into NPS-like loyalty indicators without using survey as an input. To overcome these weaknesses, this paper suggests a hybrid customer loyalty prediction analytical model using RFM + product usage counts and total spending from PT XYZ (large beverages company in Mojokerto Regency). RFM and product diversified features are utilized to create behavioral groups in the first place. These clusters are then translated to NPS Proxy labels (Promoter, Passive, Detractor) derived from the clusters. Lastly, we train and compare multiple machine learning models to predict loyalty classes using NPS Proxy. Contributions of this research are as follows: (1) we have proposed a scalable alternative to the survey-based NPS using clustering-based proxy labels, (2) we have incorporated multi-dimensional behavioral features for customer segmentation, and (3) we show a machine learning model that can predict the degree of loyalty of customers in FMCG distribution only from their transactional data.

II. METHOD

CRISP-DM (CRoss-Industry Standard Process for Data Mining) is used in this studies. CRISP-DM is de facto standard framework for data mining projects [28]. CRISP-DM presents a well-defined, step-by-step, and domain-independent framework for carrying out data analytics projects which can be helpful in maintaining methodological clarity and reproducibility. The six stages in CRISP-DM is business understanding, data understanding, data preparation, modelling, evaluation and deployment which were complied to be able to provide guidance throughout the entire analytics lifecycle. Although serving as an overview of the methodological framework, this section introduces the details of clustering, NPS proxy calculation, and machine-learning classification in a separate Proposed Method section.

A. Business Understanding

The aim of this research is to establish a model based on real data in order to estimate customer loyalty (inquiry & search for) in the FastMoving Consumer Goods (FMCG) distribution industry. PT XYZ, a Jakarta-based distributor based in Mojokerto Regency, would like to determine the level of customer loyalty by utilizing transactional data rather than NPS, as it is generally done. The business problem revolves around: (1) understanding the nature of customer purchase drivers, (2) creating clusters of customers that share similar behaviors and finally (3) identifying the category of loyalty membership the company's customers belong to. These findings enable managers to make strategic decisions in customer retention, customer management, and targeted marketing.

B. Data Understanding

The data used in this analysis consists of transaction history at the customer level, including purchase dates and quantities of beverage products (AQUA, VIT, and MIZONE), as well as total financial spending for the duration of the study. No direct customer satisfaction or survey-based loyalty annotations were available. The first step was exploratory analysis, to ascertain normality of data distribution and outliers, missing value inspection, and between-group variation in consumption for products.

It was during this step that the preliminary evaluation of features to be used in the analysis as well as a hypothesis about the role of behaviors on customer loyalty prediction were studied.

C. *Data Preparation*

Data preparation process In this section a sequence of preprocessing steps has been performed to convert the raw increment and decrement transactional records into analytically suitable forms. These steps include data preprocessing such as cleaning, detecting/eliminating outliers and missing values handling in case of different kind of attributes. Transactions were summarized by customers, and behavioral data (RFM score; total product consumption; number of different products) at user level were used in the model. District or region-level variables (which were non-numerical) are encoded using the suitable encoding procedure. Quantitative variables were scaled or transformed to make them compatible with later analytic procedures. The processed dataset was the combined features matrix for clustering, NPS Proxy generation and machine learning modeling.

D. *Modelling*

The modelling phase of CRISP-DM is analogous to the wrangling in preparing the dataset for further analysis. In this analysis, the modeling has two key analytical elements:

- Customer segmentation: underpinning behaviour profiling.
- Customer loyalty prediction: based on the understanding of segmentation.

The accurate models (i.e., clustering techniques, NPS Proxy mapping mechanisms and predictive classification algorithms) implemented in our approach are detailed in the proposed method section even though the CRISP-DM modeling phase involves such general tasks as model choice, attribute selection/extraction and data transformation. At this point, inputs to the model and their targets were explicitly defined so as to have a technical pipeline described in the following sections.

E. *Evaluation*

The final step of CRISP-DM makes certain that all models are in line with analysiswise and businesswise goals. Assessment was carried out at two scales for the purpose of this study:

- Clustering internal validity assessment including Silhouette Score, Calinski-Harabasz and Davies–Bouldin index.
- Predictive evaluation for ML models, and fine tuning hyper-parameters using metrics such as accuracy, precision, recall, macro averaged F1-score.

Stratified train–test splits and cross-validation were used to ensure robustness. Class distribution imbalance was also investigated to see if balancing strategies are necessary. The result of the evaluations directed us to pick up the best models for loyalty prediction.

III. PROPOSED METHOD

Figure 1 shows the proposed analytic process flow end-to-end composed of three main components: preprocessing, clustering, and classification. First, the raw transactional data that consist of customer purchase history in various FMCG product categories at PT XYZ is ingested. These descriptive strings are then converted into raw records and feature engineered to define two sets of behavioral indicators: the traditional RFM metrics (Recency, Frequency, Monetary) and the product-level consumption features (AQUA, VIT, MIZONE, and total purchase volume). In the next stage of data preprocessing, we do cleaning, encoding, normalizing and standardizing on selected features to make them consistent and appropriate for modeling tasks. The preprocessed data is input into the behavioral clustering) module, which divides customers according to purchasing behaviors. Besides, a cluster evaluation step is applied to evaluate the quality and truthfulness of generated clusters using built-in internal criteria. Given the behavioral profile of segments, a NPS Proxy Mapping approach is used to categorize all clusters within Promoter, Passive and Detractor classes, generating Loyalty labels from transaction data without involving survey-based NPS answers. The last step, ML classification, considers the NPS Proxy labels as target variable to learn predictive models that predict customer loyalty of unseen customers. Several algorithms are tested with a thorough classification evaluation protocol to identify the best model. The pipeline results in a strong customer loyalty prediction, which allows PT XYZ to make differentiation of profitable customers, observe behavior patterns as well as create targeted retention programs for specific consumers.

In conclusion, Figure 1 highlights the entirety of our proposed holistic pipeline in this work showing how transactional data can be turned into actionable loyalty insights using the fusion of behavioral segmentation, proxy loyalty labeling and supervised ML.

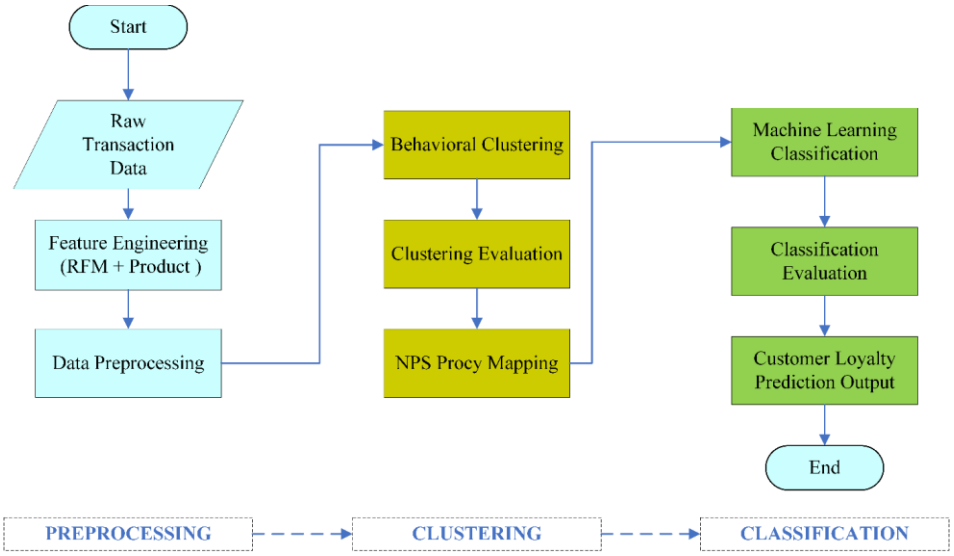

Figure. 1. Proposed Research Framework

A. Feature Engineering

The data were collected within 1 year, April 2024–April 2025, and the number of data obtained was 2450. Let the input transaction data contain the field ID, District, Transaction Date, Product Name, Qty,Price and Amount with reasons shown in Table 1.

TABLE 1. RAW DATA STRUCTURE ON SALES TRANSACTIONS

Attribute	Type	Description
TRANSACTION ID	String	A unique code used to identify each customer or store that transacts.
DISTRICT	String	Geographic location of customers by sales distribution region.
PRODUCT NAME	Numeric	The products customers purchase fall into three main categories: Product A (Aqua), Product B (Vit), and Product C (Mizone).
QUANTITY	Numeric	The number of units of the product purchased in a single transaction.
AMOUNT	Numeric	The total transaction value (in rupiah) paid by the customer for a particular product.
PRICE	Numeric	The unit price of each product is used to calculate the transaction value (amount).
TRANSACTION_DATE	Date	The transaction time when the customer bought products

In addition, a summarization is performed to simplify the analysis of transaction patterns in each customer. The output of this step is transaction data for which it has been rolled up according to customer and includes sub-districts as attributes, along with products. The result of this operation is aggregate information such as the quantity and amount for each product. For instance, for the product A (Aqua), you will be getting recreation of customer transactions for 1 year period between April 2024 and April 2025 along with total purchase units and net sales value. A similar process is followed for the B product (Vit) and Product C (Mizone), making all products follow a uniform way of data summary and are ready to analyze them to the next step. The resulting data structure after transformation is shown in Table 2.

TABLE 2. DATA STRUCTURE FOR CLUSTERING

Attribute	Type	Description
CUSTOMER ID	String	Unique code that represents each customer.
DISTRICT	String	Geographic location of customers by sales distribution region.
RECENCY	Numeric	The difference in days between the customer's last transaction and the analysis reference date.
FREQUENCY	Numeric	The number of unique transactions the customer made during the analysis period.
MONETARY	Numeric	The average value of total customer purchases during the analysis period.
PRODUCT A	Numeric	Total purchase of Aqua (Product A) by customers for one year.
PRODUCT B	Numeric	Total purchases of Vit (Product B) by customers over one year.
PRODUCT C	Numeric	Total purchases of Mizone (Product C) by customers for one year.
TOTAL	Numeric	The amount of payments made by the customer during the analysis period

B. Behavioral Clustering

Figure 2 displays the step by step workflow of behavioral clustering approach adopted in this paper, taking transformed transactional features to best clustering algorithm selection along with NPS Proxy labeling. The process starts with the transformed data set generated after preprocessing and feature engineering. PCA To help visualize the clusters and improve their separability in a reduced feature space, Principal Component Analysis (PCA) is performed to produce a 2-d representation of the data (PCA = 2). While clustering is carried out on the complete multi-dimensional feature space, PCA provides reduced data to aid interpretability of model evaluation. In the cluster phase, three unsupervised methods are individually applied (K-Means, Agglomerative Hierarchical Clustering and DBSCAN) to capture various structural assumptions in data. Since each algorithm can potentially model a different behavior, it is important to compare their accuracy in segmenting customers. To that end, performance of each clustering approach is instead assessed with a set of multiple internal validity indices including the Silhouette Score, Calinski-Harabasz Index and Davies-Bouldin Index. Together, the indices measure cluster cohesion, separation and compactness without reference to external labels.

Finally, the model is selected as a result of the evaluation. More specifically, the framework examines whether each clustering algorithm yields optimal Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index values respectively (on a parameter dependent basis for the latter two) as corresponding metric criterion dictates. The method does not rely on a single indicator for clustering quality, but adopts voting mechanism: the clustering algorithm with three indexes best favoring to it will be selected as the optimal model. Such multi-metric voting approach will further enhance the robustness, as it also reduces the bias associated with a single evaluation measure. When the optimal clustering approach is chosen, models are processed for cluster profiling to analyze behavior profiles within each cluster based on RFM variables, product diversification attributes and purchase frequencies. These profiles are then interpreted by the following NPS Proxy Labeling where each cluster is assigned to Promoter, Passive, Detractor categories based on their purchase behavior and value induced. After clusters are assigned proxy loyalty labels, the procedure ends and behavioral segmentation stage in the framework is finished.

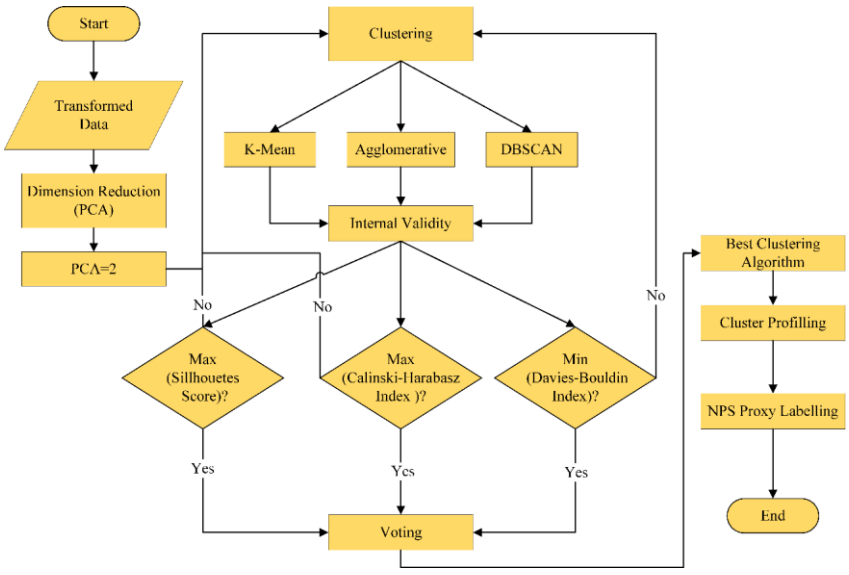

Figure. 2. Proposed Clustering Method

C. Customer Loyalty Prediction

Figure 3 depicts a high-level supervised learning pipeline for scoring customer loyalty from estimated NPS Proxy labels output by the clustering stage. The process starts with transformed data which includes the RFM-Product based features and NPS Proxy labels for Customer Segmentation. Next, we divide the labeled dataset into training and test set with a stratified splitting to maintain class distribution and facilitate fair evaluation. As a result of the unbalance problem faced by NPS Proxy classes, we only apply the SMOTE noise technique on training sub-dataset in order to generate minority-class samples artificially for eliminating learning bias. This method does not introduce data leakage and very well trains the classification models with same proportions in training set, while remaining that testing subset has a representative distribution of population. During modeling stage the five ML approaches are trained concurrently which includes KNN, SVM, Gradient Boosting, RF, and LR. Each algorithm is trained on the balanced training set, and they are tested for prediction performance on the unseen test samples.

The performance evaluation module calculates several classification measures, like accuracy, precision, recall and macro-averaged F1-score, and confusion matrices to measure each model's predictive capacity to correctly predict customer loyalty categories. Subsequently, the protocol objects a clear manual overfitting check, contrasting the model's performance on the learning and testing set. The algorithm that performs very poorly in the training set but shows better performance on the testing set is selected. If overfitting is not observed the pipeline retains the classifier with the highest macro F1-score, which is suitable for a multi-class imbalanced output such as Detractor, Passive and Promoter. The selected model is then used to create customer loyalty prediction for deployment so that PT XYZ can recognize valuable customers, predict risk likelihood of leaving the company and manage promotions with more focus on certain customer group. After all, the Figure 3 shows that a well-controlled classification framework including balanced learning, between model comparison and rigid validation gives good prediction of customer loyalty based on behavior transactional records overall.

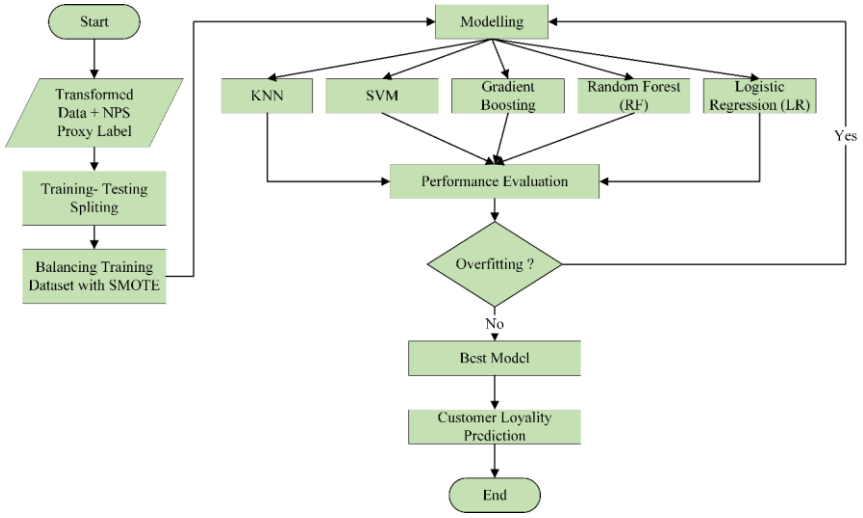

Figure. 3. Classification Workflow

IV. RESULT AND DICUSSION

The findings of our study are discussed in three parts, each corresponding to the design philosophy behind the proposed analytical framework: behavioral clustering, NPS Proxy mapping, and supervised loyalty prediction. Both classes present cardinal data on how transactional information along with RFM and product diversification features can first be shaped into behavior groups and in turn into predictive loyalty labels. The results show not only the performance of clustering and classification model but also practical information from customer purchase behavior in FMCG distribution of PT XYZ. Combining quantitative evaluation with the interpretation of results behaviorally, this section sheds light on how effective, reliable, and managerial plausible is the proposed methodology.

A. Clustering Results and Behavioral Segmentation

In the step of preprocessing before clustering, dimension reduction is performed through PCA to make more effective for clustering and by means of scatter diagram, we can visually see the pattern. We visualise the contribution of variance emanating from engineered transactional features for an optimum number of principal components. Figure 4 presents the results of the Principal Component Analysis (PCA) applied to the engineered transactional features to assess variance contribution and

determine the optimal number of principal components. The left graph displays the scree plot, where the explained variance ratio of each component is plotted. The first three principal components explain approximately 49%, 14% and 12% of the total variation, respectively. For more than the third component, the variance contribution decreases rapidly and there exists a distinct elbow at which point most of the the information is carried by only few principal components. The right graph displays cumulative explained variance, showing how much of the variance builds up as more components are added. The first two components in combination describe about 63% of variance, the first three components sum up to 75%, and first four model components account for as much as 86%. The cumulative variance increases to 94% with 5 components, and becomes nearly 100% at the seventh component. These data indicate that cluster visualizations are best achieved using two principal components in all cases, and four or five components for analyses where a more robust representation of the original feature space is required. In this sense, PCA is employed for visualisation purposes and clustering on full multi-dimension feature space in order to maintain the structure of underlying behavioral patterns.

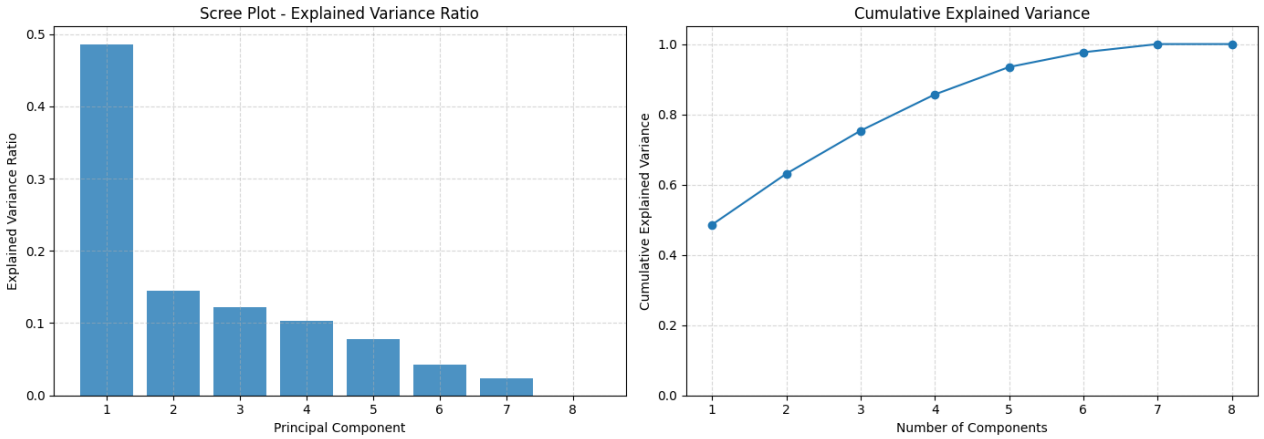

Figure. 4. Reduction Dimension Using PCA

We proceed by clustering using n of 3 and the 2 component feature. This is the famous theory of Net Promoter Score (NPS) which includes 3 types of customers, Promoter, Passive and Detractor. Table 3 shows the internal validity measures achieved by the three cluster algorithms investigated in this work, as well as the number of clusters to which they have been associated: K-Means, Agglomerative Clustering and DBSCAN with k = 3. The performance of the three techniques varies significantly, indicating the complexity and distribution of customer transactional features.

The Agglomerative Clustering achieves the best Silhouette Score (~0.8919) with strong clusteredness among clusters and large distances between them. This is much higher than K-Means' number (0.5432), which means that hierarchical clustering take places in the data better explains the inner structure of the dataset. Furthermore, Agglomerative Clustering obtains a smaller Davies–Bouldin Index (0.5266) than K-Means (0.6147) and DBSCAN (0.6780), indicating that its clusters possess higher compactness and lower overlap in the clustering process. On other hand, DBSCAN obtains the worst performance in all criteria, with a negative Silhouette Score (–0.1727) and an excessively low Calinski–Harabasz Index (47.1694), which means that it was not able to delineate genuinely finding clusters under this setting of parameter values. The potential reason is that the dataset contains no density-based structures for \textsc{DBSCAN} or the distribution of data too uniform for the algorithm to separate clusters. Notice however how K-Means is less effective with not so good Silhouette and Davies–Bouldin measures, proving that it does not achieve enough quality for dissimilar word partitioning. Additionally, the Calinski–Harabasz Index for K-Means (2912.1513) is slightly higher than that of Agglomerative Clustering (2585.1121), however with the cohesion and compactness metrics of Agglomerative Clustering being much higher, on balance, there is strong evidence in favour of Agglomerative Clustering.

By virtue of the aggregate evaluation of all three internal validity measures Agglomerative Clustering is the most appropriate clustering strategy that can be used for behavioral segmentation on this data. This is consistent with the observation that hierarchical clustering, which can be sensitive to small changes in customer purchasing patterns (especially for data sets with nonspherical clusters and variable densities).

TABLE 3. CLUSTERING INTERNAL VALIDITY

Method	N-Cluster	Silhouette Score	Calinski-Harabasz Index	Davies-Bouldin Index
KMeans	3	0.5432	2912.1513	0.6147
Agglomerative	3	0.8919	2585.1121	0.5266
DBSCAN	3	-0.1727	47.1694	0.6780

In order to assess these results, the PCA scatter diagrams of the resulting clusters for the methods are shown in Figure 5 which enables a qualitative comparisons of their separability and pattern formation.

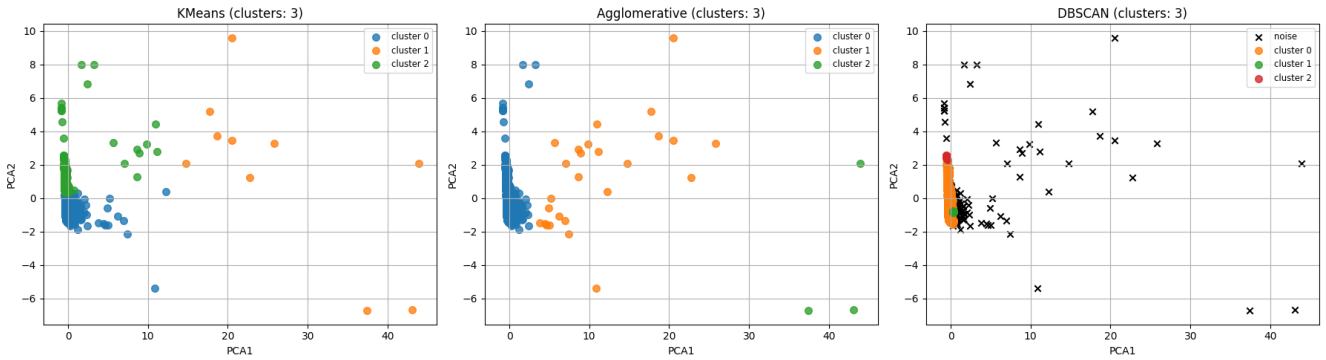

Figure. 5. Scatter Chart for Cluster based on PCA

B. *NPS Proxy Mapping and Customer Loyalty Patterns*

The behavioral profiles of the three customer segments were also investigated by numerical summaries (Table 4) and visualization of normalized features with a radar chart (Figure 6). Together, these representations yield a holistic insight into the purchase intensity and product preference behaviors that set apart each segment of customers, thereby enabling subsequent NPS Proxy assignment.

TABLE 4. CLUSTER CHARACTERISTICS BASED ON MEAN OF FEATURES

Cluster	Recency (Day)	Frequency	Monetary (IDR)	Product A (Aqua- IDR)	Product B (Vit-IDR)	Product C (Mizone - IDR)	Total
0	30.52	50.786	3.08e+05	1.88e+07	6.06e+05	3.37e+05	1.98e+07
1	1.42	369.538	3.03e+07	5.27e+09	2.42e+08	5.58e+08	6.07e+09
2	1.00	2058.333	1.93e+07	3.197e+10	3.94e+08	2.01e+09	3.44e+10

- Cluster 0, located at the base of the octant, exhibits high Recency, low Frequency and low Monetary value representing sporadic and small-scale buying behavior. This pattern is well pronounced in Figure 6 as Cluster 0 recording the highest normalized Recency and near-zero normalized values for the other axes. The low behavioral cost recorded product A (aqua), Product B (Vit) ,Product C (Mizone) confirmed our interpretation that these groups are disloyal and bring limited revenues. This radar plot also shows that the low activity is free from subjective bias, as it creates a small polygon around the origin. These are again highly consistent(?) with the behaviour of Detractors that have poor levels (again however frequent??) engagement and loyalty, as we can not expect them to be different parties.
- Cluster 1 has the middle value of all the three variables with low Recency (which means recent purchase), a reasonable Frequency and higher amount spent than Cluster 0 but less than Cluster 2. In radar chart (Figure 6), Cluster 1 exhibits intermediate ranges on all axes, resulting in a polygonal area coverage larger than Cluster 0 but still significantly smaller than Cluster 2. Product spend exhibits clear involvement across all 3 product categories, notably for Product A and Product C. This customer segment demonstrates regular investment, although not outstanding, as characterised by the Passive NPS profile, representing customers that engage somewhat but are not yet fully invested.
- Cluster 2 is identified as the most valuable and loyal customer segment, consisting of super low Recency but extremely high Frequency consumer purchasing behaviour in addition to exhibiting highest Monetary and aggregate consumption amounts (Table 4). In Figure 6, Cluster 2 presents the highest normalized values in all dimensions but Recency, and exhibits the largest and dominant polygon on radar chart. This signifies great repeat purchasing and volume consumption across the board, and particularly with Product A (Aqua) which drives their spending. The clearly divergence between Cluster 2 and other clusters as a central dominant segment suggest that it plays the traditional role core revenue generator in terms of NPS. (i.e high loyalty with high frequent transaction) risk profile if you were to profile promotor, strong loyality and consisitent behaviour on top-opting for purchasing behavior ultimately.

Taking the numerical information from Table 4 along with visual patterns shown in Figure 6, it can be concluded that the three clusters indeed characterise clearly different behavioural archetypes. Cluster 0 is consistent with Detractors, Cluster 1

with Passives and Cluster 2 with Promoters, validating the NPS Proxy projection whilst demonstrating how RFM and product diversification features help to derive loyalty from transactional data.

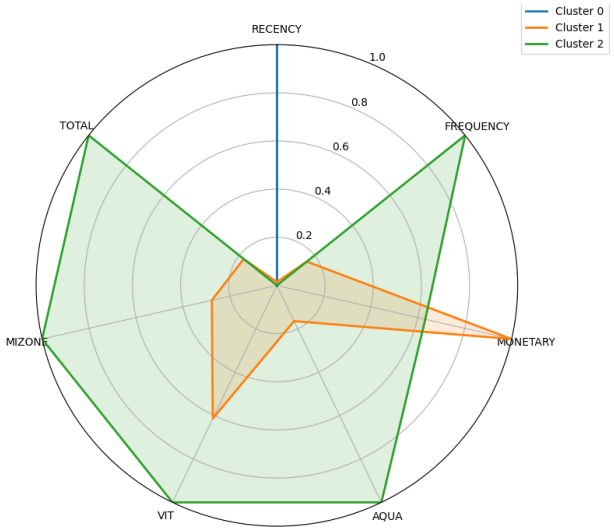

Figure. 6. Radar Chart of Feature Profiles (Normalized) by Cluster

To facilitate the understanding of these findings in a more integrated way, Table 5 reports some basic behavioral attributes for each cluster (RFM dimensions and business interpretation based on consumption rate or order amount as the case may be) alongside with their associated category by NPS Proxy. Such a map can facilitate interpretation and assign mutually comparable weights to different quantitative behavioral measures (such) that the mapping between behavior and loyalty becomes describable in terms of an interpretable, intuitionable basis for decision making and strategic customer management down-stream from the model.

TABLE 5. CLUSTER MAPPING BASED ON NPS

Cluster	RFM Characteristics	Product Behavior	Key Behavioral Profile	Business Interpretation	NPS Proxy
0	Very high Recency (long time since last purchase); very low Frequency and Monetary	Almost no product purchase; Aqua/Vit/Mizone consumption near zero	Inactive customers; negligible purchasing; very low spending	High churn risk; not loyal; may buy only occasionally or have stopped buying	Detractor
1	Low Recency; moderate Frequency and Monetary	Low–medium product consumption; some purchases of Aqua/Vit/Mizone	Active but not intense; stable but moderate spending	Potential to increase loyalty; neutral but not enthusiastic	Passive
2	Lowest Recency; highest Frequency and Monetary	Highest consumption of Aqua, Vit, Mizone; highest total expenditure	Highly active; high and routine spending; multi-product consumers	Very loyal; highest-value tier; strongly engaged	Promoter

C. Classification Model for Customer Loyalty Prediction

Once the NPS Proxy labels have been assigned according to the clustering results, a supervised machine learning was undertaken in order to predict it s customer loyalty category (Detractor, Passive and Promoter). Prior to training, we inspected the distribution of NPS Proxy labels for class balance. As depicted in Figure 7, the dataset is extremely imbalanced: Detractor class dominates the population, and Passive one as well as Promoter one have a very low amount of samples. Such a distribution provides models with a strong learning bias, meaning that models are biased to predict the majority class and are not good at detecting minority loyalty groups. To cope with this problem, we balanced the training dataset with SMOTE (Synthetic Minority Over-sampling Technique) by making sure that all 3 NPS Proxy categories appeared equally in the learning process. This step was introduced after train–test split to prevent any data leakage and ensure the test evaluation remained fair and real-word.

Five machine learning models, including KNN, SVC, Gradient Boosting, Random Forest and Logistic Regression were tested finally after balancing. The performance numbers are listed in Table 6. The observation is from Table 6 that some models (e.g., KNN, SVC, Grandient Boosting) achieve almost perfect training results but display significant decrease of the F1-score (0.64) on testing data. This discrepancy is a demonstration of overfitting, meaning these models are learning the synthetic patterns that SMOTE introduces rather than generalizing well to novel data. In contrast, Random Forest and Logistic Regression achieve great generalization performance, with perfect or near perfect scores on both training/test sets. Random Forest, in

particular, achieves F1-score, precision, and recall of 1.00 on both datasets, highlighting its robustness in handling non-linear relationships and balanced training data. Its optimal hyperparameters, 100 estimators with no limit on tree depth, enable the model to capture complex interactions while maintaining stability across classes.

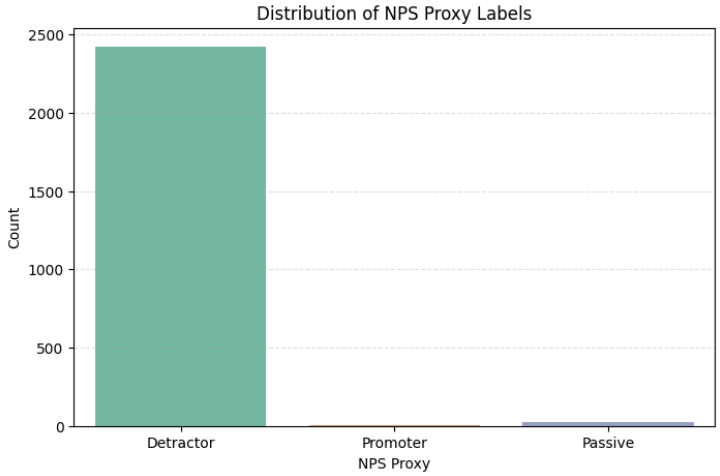

Figure. 7. Cluster Distribution by Count of Members

Due to its better performance, absence of overfitting, and capability for correct classification of the minority loyalty segments (Passive and Promoter), Random Forest would be assumed as the best model to predict customer loyalty. This framework allows PT XYZ the capability to confidently identify high value Promoters, churn risks among Detractors and opportunity segments in Passives.

TABLE 6. CLASSIFICATION PERFORMANCE FOR CUSTOMER LOYALTY PREDICTON

Model	Train			Test			Best Parameter Model
	F1-Score	*Precision*	*Recall*	*F1-Score*	*Precision*	*Recall*	
KNN	1.00	1.00	1.00	0.64	0.61	0.67	{'clf__n_neighbors': 3, 'clf__weights': 'distance'}
SVC	0.98	0.97	1.00	0.64	0.61	0.67	{'clf__C': 1, 'clf__class_weight': None, 'clf__kernel': 'rbf'}
GradientBoosting	1.00	1.00	1.00	0.64	0.61	0.67	{'clf__learning_rate': 0.05, 'clf__n_estimators': 100}
RandomForest	1.00	1.00	1.00	1.00	1.00	1.00	{'clf__class_weight': None, 'clf__max_depth': None, 'clf__n_estimators': 100}
LogisticRegression	0.98	0.97	1.00	1.00	1.00	1.00	{'clf__C': 0.01, 'clf__class_weight': None}

In order to investigate model stability and overfitting, Table 6 contrasts the Train F1-macro with Test F1-macro scores for all models. The plot reveals a similar strong deviation between training and test performance for KNN, SVC and Gradient Boosting, with the former yielding near-perfect training scores of up to around 0.98–1.00, but lower testing F1-macro (around 0.64). This trend further supports the previous finding that these models overfit to the balanced training data and they can hardly generalize when exposed with original class distribution in test. By contrast, Random Forest and Logistic Regression show strong congruence between their training and testing performances. Both models has strong F1-macro scores on test data (Random Forest = 1.00, Logistic Regression = 1.00), which nearly replicates their performance during training. This suggests that these models are not learning only from the synthetic samples produced by SMOTE, but learn decision boundaries that generalize well to new customers.

Figure 8 shows the results corroborate that Random Forest is an appropriate choice as the best classifier. Thus, due to its stable cross-dataset performance, robustness against overfitting and the best predictive capacity for all three loyalty categories (Detractor, Passive, Promoter), is the most suitable candidate for operational use in customer loyalty prediction task.

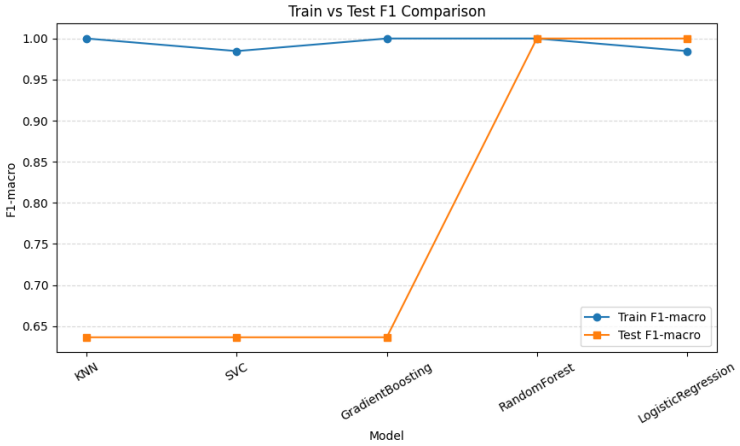

Figure. 8. Line Chart for Predicted Overfitting Diagnosis

V. CONCLUSION

This paper presents an integrated data-driven framework for predicting customer loyalty in the FMCG distribution industry based on transactional data. By integrating RFM factors with product profile, the method is able to comprehensively reflect customer triple behaviors and achieve meaningful segmentation. The cluster analysis phase, based on Silhouette Score, Calinski–Harabasz Index and Davies–Bouldin Index, revealed the Agglomerative Clustering as the most well-performing method allowing to obtain three homogeneous behavioral segments. Mapping, in a second step, of these sections to NPS Proxy labels leads to obvious loyalty archetypes, where Cluster 0 represents Detractors, Cluster 1 corresponds to Passive customers, and Cluster 2 strongly aligns with Promoters. This confirms that transactional patterns alone can serve as a reliable basis for inferring customer loyalty.

The classification phase provided additional confirmation on the predictability of NPS Proxy labels. Although significant class imbalance existed in the initial dataset, we were able to learn using the training set balanced after SMOTE augmentation. Of those algorithms tested, the Random Forest algorithm consistently produced good generalization performance and had higher F1-macro scores on the training as well as testing data compared to other classifiers such as KNN, SVC, Gradient Boosting, Logistic Regression. This stability verifies its capability in representing loyalty behaviors and demonstrates its applicability in practical use.

In summary, the results highlight that combining behavioral clustering, NPS Proxy mapping and machine learning classification provides an effective method to predict high value customers and expected loyalty. The model offers actionable information for distributors, like PT XYZ in Mojokerto, to produce retention strategy, prioritize best customer and better management of all customers. Follow-up work may enhance the approach by adding in temporal features, sentiment information or deep learning models to improve predictive performance.

REFERENCES

[1] C. Aurelia and N. Kusumawati, "The Effect of Online Customer Experience Toward Customer Satisfaction and Customer Loyalty," 2024. doi: 10.2991/978-94-6463-234-7_57.

[2] R. A. Kamaroellah, A. Eliyana, and R. Mubarak, "Service Distribution And Satisfaction Toward Customer Loyalty," *Amwaluna: Jurnal Ekonomi dan Keuangan Syariah*, vol. 5, no. 1, 2021, doi: 10.29313/amwaluna.v5i1.6021.

[3] J. Le Bon, "The Customer Compromise and ComproScore: Toward a New Concept and Metric to Assess Customer Satisfaction, Buying Process, and Loyalty: An Abstract," in *Developments in Marketing Science: Proceedings of the Academy of Marketing Science*, 2019. doi: 10.1007/978-3-030-02568-7_105.

[4] Q. Yang and L. Young-Chan, "What Drives the Digital Customer Experience and Customer Loyalty in Mobile Short-Form Video Shopping? Evidence from Douyin (TikTok)," *Sustainability*, vol. 14, no. 17, 2022, doi: 10.3390/su141710890.

[5] K. Yum, J. Kim, "The Influence of Perceived Value, Customer Satisfaction, and Trust on Loyalty in Entertainment Platforms," *Applied Science,* vol. 14, no. 13, 2024, doi: 10.3390/app14135763.

[6] W. Guo, F. Liu, and X. Zhang, "Research on Insurance Customer Segmentation Model and Marketing Strategy Based on Big Data and Machine Learning," in *ACM International Conference Proceeding Series*, 2021. doi: 10.1145/3469213.3471326.

[7] X. Wang and L. Liu, "Customer segmentation and marketing strategy of commercial banks based on CLV," in *Advances in Intelligent and Soft Computing*, 2012. doi: 10.1007/978-3-642-27334-6_30.

[8] X. Li and Y. S. Lee, "Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm," *Journal of Cases on Information Technology*, vol. 26, no. 1, 2024, doi: 10.4018/JCIT.336916.

[9] H. Do and S. Lee, "Marketing Segmentation Strategy Based on Internal Customers," *The Korean Academic Association of Business Administration*, vol. 37, no. 1, 2018, doi: 10.18032/kaaba.2018.31.7.1307.

[10] F. M. Hilmy, R. A. Nurhaliza, M. Q. Huzyan Octava, and G. Alfian, "Web-based E-Commerce Customer Segmentation System Using RFM and K-Means Model," in *2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2023*, 2023. doi: 10.1109/3ICT60104.2023.10391650.

[11] J. Pechter and A. Kuusik, "NPS from the customer's perspective: The influence of the recent experience," *International Journal of Market Research*, vol. 66, no. 2–3, 2024, doi: 10.1177/14707853231214188.

[12] T. Ho and V. H. Nguyen, "Customer Analytics Using Sentiment Analysis and Net Promoter Score," in *Encyclopedia of Data Science and Machine Learning*, 2022. doi: 10.4018/978-1-7998-9220-5.ch062.

[13] N. Sivabrovornvatana, "Utilizing Net Promoter Score To Assess Customer Satisfaction And Brand Loyalty In The Real Estate Industry Of Thailand," *Journal of Business Leadership and Management*, vol. 1, no. 1, 2023, doi: 10.59762/jblm845920461120231009142231.

[14] L. Eger and M. Mičík, "Customer-oriented communication in retail and Net Promoter Score," *Journal of Retailing and Consumer Services*, vol. 35, 2017, doi: 10.1016/j.jretconser.2016.12.009.

[15] M. Barath, "Net Promoter Score as Measuring Instrument of Customer Brand Loyalty," in *Studies in Systems, Decision and Control*, vol. 421, Springer Science and Business Media Deutschland GmbH, pp. 363–377, 2022. doi: 10.1007/978-3-030-97008-6_16.

[16] B. Hardianto and S. Wijaya, "Analysis of The Impact of Net Promoter Score on Financial Performance With Customer Loyalty As Mediation," *International Journal of Social Service and Research*, vol. 3, no. 6, pp. 1478–1488, Jun. 2023, doi: 10.46799/ijssr.v3i6.401.

[17] L. Abednego, C. E. Nugraheni, and A. Salsabina, "Customer Segmentation: Transformation from Data to Marketing Strategy," *Conference Series*, vol. 4, no. 1, 2023, doi: 10.34306/conferenceseries.v4i1.645.

[18] R. Mandrai, P. Sharma, and B. Borkakaty, "Customer Risk Prediction: A Machine Learning Ensemble Approach," in *International Conference on Electrical, Computer and Energy Technologies, ICECET 2023*, 2023. doi: 10.1109/ICECET58911.2023.10389208.

[19] Y. Suh, "Discovering customer segments through interaction behaviors for home appliance business," *Journal of Big Data*, vol. 12, 2025. doi: 10.1186/s40537-025-01111-y.

[20] J. M. A. M. Ramos and F. A. Silva, "Customer Lifetime Value Prediction: A Machine Learning Approach," 2023. doi: 10.5753/eniac.2023.234262.

[21] I. Z. P. Hamdan, M. Othman, Y. M. M. Hassim, S. Marjudi, and M. M. Yusof, "Customer Loyalty Prediction for Hotel Industry Using Machine Learning Approach," *International Journal on Informatics Visualization*, vol. 7, no. 3, 2023, doi: 10.30630/joiv.7.3.1335.

[22] S. Tavassoli and H. Koosha, "Hybrid ensemble learning approaches to customer churn prediction," *Kybernetes*, vol. 51, no. 3, 2022, doi: 10.1108/K-04-2020-0214.

[23] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, 2022, doi: 10.1007/s00607-021-00908-y.

[24] Y. Beeharry and R. T. Fokone, "Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry," *Concurr Comput*, vol. 34, no. 4, 2022, doi: 10.1002/cpe.6627.

[25] K. Li, C. Xue, Z. Zhao, M. Zhu, X. Cui, S. Xu, and J. Zou, "Deciphering modern customer loyalty: a machine learning approach," 2023. doi: 10.1117/12.3013297.

[26] H. F. Lee and M. Jiang, "A Hybrid Machine Learning Approach for Customer Loyalty Prediction," in *Communications in Computer and Information Science*, 2021. doi: 10.1007/978-981-16-5188-5_16.

[27] I. Z. P. Hamdan and M. Othman, "Predicting Customer Loyalty Using Machine Learning for Hotel Industry," *Journal of Soft Computing and Data Mining*, vol. 3, no. 2, 2022, doi: 10.30880/jscdm.2022.03.02.004.

[28] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model", *Procedia Computer Science,* vo. 181, pp526-534, 2021. doi: 10.1016/j.procs.2021.01.199.