

Application Of K-Means And Random Forest Algorithms For Trader Segmentation And Activity Status Prediction At GKINVEST

Ramadhan Alvito Alfian^{1*}, Safrizal²

^{1,2}Sistem Informasi Universitas Pembangunan Jaya, Banten, Indonesia

* Corresponding author:

Email: alvitoxd121@gmail.com

Abstract.

Trader activities at trading firms exhibit heterogeneous characteristics, necessitating an analytical approach to identify segmentation patterns and trader activity status. This study aims to analyze GKInvest trader segmentation using the K-Means algorithm and to build a Random Forest model to predict trader activity status based on historical activity characteristics. The dataset comprises trader data with key variables including 30-day transaction volume, 30-day login count, 30-day transaction count, days since last activity (recency), account tenure, number of active days, and the number of instruments traded. K-Means was employed as an unsupervised learning method to establish segments, while Random Forest was used independently as a supervised learning method to classify active versus inactive status. The K-Means analysis yielded three segments: Passive Traders (Churn Risk) comprising 1,135 traders (56.7%), Highly Active Traders (VIP) comprising 437 traders (21.9%), and Regular Active Traders comprising 428 traders (21.4%). The Random Forest model achieved an accuracy of 83%; performance for the "Active" class reached a precision of 84%, recall of 93%, and F1-score of 88% (with a weighted average F1-score of 82%). Feature importance analysis revealed that 30-day transaction volume made the largest relative contribution (~0.218), followed by 30-day login count (~0.184) and 30-day transaction count (~0.183). The findings demonstrate that K-Means and Random Forest can independently provide complementary perspectives for understanding the activity characteristics and status of GKInvest traders, serving as a foundation for developing data-driven trader activity monitoring systems.

Keywords: K-Means, Random Forest, trader segmentation, trader activity, classification and machine learning.

I. INTRODUCTION

Digital transformation in the financial sector generates vast amounts of user activity data with increasingly diverse characteristics. In financial trading services, trader activity data reflects not only executed transactions but also user engagement levels captured through metrics such as transaction volume and frequency, login counts, the number of instruments traded, time since the last activity, and account tenure. This diversity of characteristics offers companies the opportunity to leverage data as a foundation for analyzing user behavior. Advances in machine learning enable the processing of activity data using unsupervised learning to discover user group patterns and supervised learning to classify users based on historical data characteristics. In the financial context, machine learning approaches have been employed to identify user segments based on transaction patterns and to predict user conditions or behaviors. Research by Perdana et al. demonstrates that patterns regarding transaction volume, timing, and type can be utilized specifically via the K-Means algorithm to segment users into groups with distinct behavioral characteristics [1].

This situation is relevant for trading firms like GKInvest, which have traders exhibiting diverse levels and patterns of activity. Variations in transaction volume and frequency, login intensity, the number of instruments traded, the time elapsed since the last activity, and account tenure can reveal distinct trader engagement characteristics. Traders with high, moderate, or low activity levels display service usage patterns that cannot always be explained by aggregate indicators alone. Therefore, an analysis capable of segmenting traders based on similarities in activity characteristics is required so that the company can gain a more structured overview of its trader base. Furthermore, the company requires a predictive approach to identify trader activity status based on activity patterns recorded in the data [2], [3].

In addition to segmentation, identifying a trader's activity status can be achieved through a classification approach using supervised learning algorithms. One such method is Random Forest, a classification algorithm based on an ensemble of decision trees that makes decisions by combining data

characteristics. In the banking sector, Random Forest has been applied to classify customer conditions based on historical data. Azmi and Voutama, utilized Random Forest and Decision Trees to predict bank customer churn, using data from 10,000 customers and evaluating the models via a confusion matrix [4]. Cong, also applied Random Forest to predict customer churn in the banking sector, employing SMOTE and ADASYN to address class imbalance and evaluating the model using accuracy, recall, precision, specificity, F1-score, and ROC-AUC [5]. These findings demonstrate that Random Forest is suitable for classification based on historical activity characteristics, even though the context of customer churn differs from the trader activity status examined in this study [4], [5].

In this study, a trader's activity status is defined as a category of activity condition based on specific operational criteria, rather than as a measure of customer churn. This distinction is significant because a trader's activity level can fluctuate without necessarily indicating that they have ceased using the company's services. Previous research has demonstrated the use of K-Means to identify user groups based on transaction characteristics and Random Forest for classification based on historical data. However, there remains scope for further development regarding the separate application of these two approaches to analyze trader segmentation and activity status within the context of a trading company. This gap provides the rationale for employing K-Means to identify trader groups based on activity characteristics and using Random Forest independently to predict trader activity status.

This study utilizes K-Means and Random Forest as two analytical approaches that serve distinct yet complementary functions. K-Means is employed to segment traders based on similarities in activity characteristics, while Random Forest is used independently to build a predictive model for activity status based on trader activity variables. Consequently, the K-Means clustering results are not utilized as features or inputs for the Random Forest model. Both methods are applied to generate two analytical perspectives on the same dataset: a segmentation perspective and a predictive perspective. The study's contribution lies in applying these approaches within the context of the trading company GKInvest, thereby enabling the analysis of trader group characteristics and activity status prediction capabilities within a unified research framework. The analysis results can subsequently be visualized via a web-based dashboard to present segmentation and prediction information in a more structured manner.

Based on the foregoing, this study aims to identify trader segments at GKInvest based on activity characteristics using the K-Means algorithm, analyze the characteristics of each resulting segment, and build a Random Forest model to predict trader activity status. The prediction model's performance is evaluated using accuracy, precision, recall, F1-score, and a confusion matrix, while the segmentation results are analyzed based on the characteristics of each cluster and appropriate evaluation metrics. The study's findings are expected to provide a data-driven overview of the heterogeneity in trader activity and the model's predictive capability regarding activity status, thereby serving as analytical information for the company to monitor and manage trader activities at GKInvest.

II. METHODS

This study employs a quantitative approach utilizing machine learning-based analytical methods to analyze the activity characteristics of traders at GKInvest. The research dataset comprises data from 2,000 traders, characterized by seven variables representing trader behavior and activity intensity: 30-day transaction volume, 30-day login count, 30-day transaction count, days since last activity, account age (in days), number of active days within a 30-day period, and the number of traded instruments. These seven variables serve to represent trader activity characteristics across dimensions such as transaction intensity, access frequency, activity continuity, and the diversity of traded instruments.

The analysis utilizes two algorithms serving distinct functions: K-Means, an unsupervised learning method for trader segmentation, and Random Forest, a supervised learning method for predicting trader activity status. This approach enables the analysis of trader characteristics from two perspectives: grouping based on activity pattern similarities and classification based on predefined activity statuses. Applying K-Means complemented by evaluations using the Elbow Method and Silhouette Score helps determine the appropriate number of clusters during the clustering process [6].

The initial phase of the research involves data preprocessing to ensure the dataset is suitable for analysis. This stage encompasses checks for data completeness, duplicates, missing values, and data types, as well as the selection of relevant variables and data scaling adjustments. The seven trader activity variables utilized in the study are 30-day transaction volume, 30-day login count, 30-day transaction count, days since last activity, account age (in days), number of active days within a 30-day period, and the number of traded instruments. In the K-Means analysis, all numerical variables were standardized using StandardScaler. Standardization was performed to ensure that differences in value ranges between variables did not cause variables with larger scales to dominate distance calculations during cluster formation. Consequently, each variable could contribute more proportionally to the trader grouping process.

In the Random Forest analysis, the seven variables served as predictor variables (features), while trader activity status categorized as Active or Inactive—served as the target variable. Unlike K-Means, which relies on distance metrics to form groups, Random Forest utilizes patterns of relationships between predictor variables and target labels to perform classification. Stages such as preprocessing, data splitting, modeling, and evaluation are crucial components in developing machine learning-based classification models [7].

Trader segmentation was conducted using the K-Means algorithm to group traders based on similarities in their activity characteristics. K-Means operates by forming groups based on the proximity of objects to a group center (centroid), thereby placing traders with relatively similar activity characteristics into the same cluster. The seven trader activity variables used as the basis for cluster formation were transaction volume, number of logins, number of transactions, recency, account age, number of active days, and number of traded instruments.

The determination of the number of clusters was carried out by considering both the Elbow Method and the Silhouette Score. The Elbow Method was used to observe changes in inertia values—or the within-cluster sum of squares—across various alternative cluster counts. Meanwhile, the Silhouette Score was employed to evaluate internal density and inter-cluster separation. Using both approaches simultaneously provided a more informative basis for evaluating the number of groups to be used in the segmentation [6], [8].

Based on these evaluation results, this study adopted a three-cluster configuration for trader segmentation. Following the clustering process, each group was analyzed based on its number of members and the mean values of the seven activity variables. The characteristics of each cluster were interpreted after the grouping results were obtained; consequently, labels such as "active trader," "moderate trader," and "passive trader" were assigned based on the emerging data patterns rather than being predetermined prior to the clustering process.

Prediction of trader activity status was performed using the Random Forest algorithm, independently of the K-Means segmentation results. Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to generate a final classification decision. In this study, the dataset was split into 80% training data and 20% testing data using a stratified split technique based on the distribution of trader activity status.

The Random Forest model was constructed using 300 decision trees with a random state of 42. The seven trader activity variables served as predictor variables, while trader activity status served as the target variable. These predictor variables included 30-day transaction volume, 30-day login count, 30-day transaction count, days since last activity, account age (in days), number of active days within the last 30 days, and the number of instruments traded. The model was trained on the training data and subsequently used to generate predictions on the testing data. In addition to generating class predictions, Random Forest is used to identify the relative importance (feature importance) of each variable in the classification process. Feature importance indicates the relative contribution of each variable to the decisions made by the model. These results are then used to determine which activity variable plays the most significant role in predicting a trader's activity status [9].

Based on the Random Forest model results, the feature importance ranking of the seven trader activity variables shows that 30-day transaction volume has the highest relative importance, with a value of 0.218. The subsequent variables are the number of logins in 30 days (0.184), the number of transactions in 30 days

(0.182), days since last activity (0.133), account age in days (0.131), number of active days in a 30-day period (0.111), and the number of instruments traded (0.040). These values serve as the basis for identifying the most dominant variables in predicting trader activity status.

The Random Forest model was evaluated using a confusion matrix and four classification metrics: accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correct predictions across the entire test dataset. Precision measures the model's accuracy in predicting traders belonging to the "Active" class. Meanwhile, recall measures the model's ability to identify traders who actually belong to the "Active" class.

Furthermore, the F1-score is used to represent the balance between precision and recall. A confusion matrix is employed to identify the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) generated by the model. Utilizing the confusion matrix alongside accuracy, precision, recall, and the F1-score enables a more comprehensive evaluation; this approach not only indicates the rate of correct predictions but also provides insights into classification errors for each specific class [7], [9].

The results from K-Means and Random Forest are subsequently analyzed together during the discussion phase to provide a comprehensive overview of the activity characteristics and activity status of GKInvest traders. Although both methods are employed within the same research framework, the K-Means clustering results are not used as input variables for the Random Forest model. K-Means serves as a segmentation method based on similarities in activity characteristics, whereas Random Forest functions as a predictive method for trader activity status.

The K-Means segmentation results are presented through the distribution of member counts and the characteristic profiles of each cluster, based on seven trader activity variables. Meanwhile, the Random Forest results are presented via feature importance rankings, the confusion matrix, and metrics for accuracy, precision, recall, and the F1-score. The feature importance ranking is used to identify the variables that contribute most significantly to predicting trader activity status. All analysis results can be visualized using tables, graphs, and web-based dashboards to present analytical information regarding trader characteristics, segmentation, and activity status predictions.

III. RESULT AND DISCUSSION

Result

This study analyzes data from 2,000 GKInvest traders based on seven activity variables: 30-day transaction volume, 30-day login frequency, 30-day transaction count, days since last activity, account age (in days), number of active days within a 30-day period, and the number of instruments traded. These seven variables serve to represent trading intensity, platform engagement, and the temporal characteristics of trader activity. The analysis employs K-Means to identify segments based on similarities in activity characteristics and Random Forest to predict trader activity status. The use of clustering techniques to create user profiles based on transaction patterns has also been applied in research within the financial and trading market sectors. For instance, Tan et al. utilized K-Means to identify trader behavioral profiles in the Malaysian derivatives market, demonstrating that trading activity characteristics can be used to group traders with distinct behavioral patterns. [10].

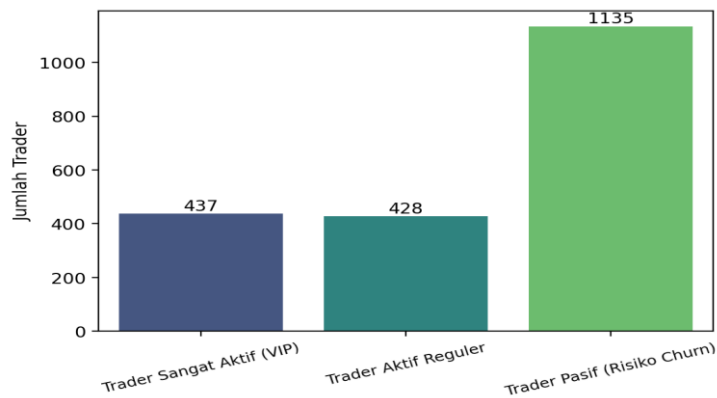
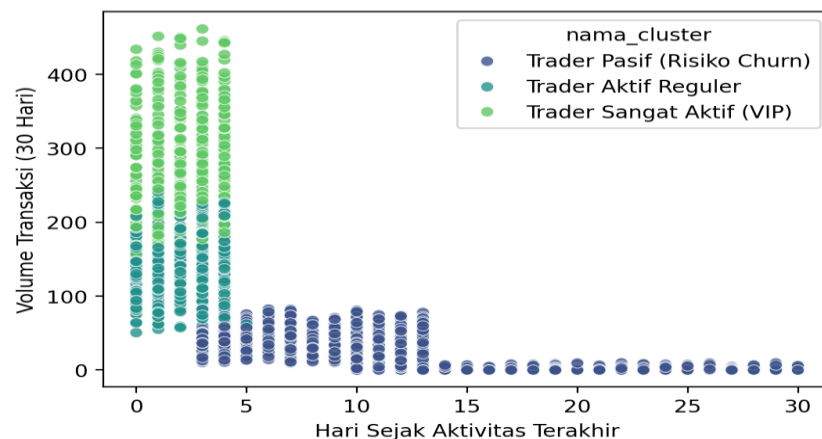
Trader Segmentation Using K-Means

Based on the grouping results using the K-Means algorithm, data from 2,000 traders were successfully mapped into three clusters, with the following distribution and interpretation: [11], [12].

- Passive Traders (Churn Risk): 1,135 traders (56.7%)
- Highly Active Traders (VIP): 437 traders (21.9%)
- Regular Active Traders: 428 traders (21.4%)

Table 1. Distribution and Key Characteristics of Trader Clusters

Cluster	Amount Trader	Proportion	Average Transaction Volume	Recency Characteristics	Interpretation
C1	1.135	56,7%	349,7	Volume transaksi rendah, hari sejak aktivitas terakhir panjang (recency tinggi)	Trader Pasif (Risiko Churn)
C2	437	21,9%	4.344,90	Volume transaksi sangat tinggi, recency sangat singkat	Trader Sangat Aktif (VIP)
C3	428	21,4%	1.233,80	Volume transaksi sedang/rutin, recency singkat	Trader Aktif Reguler
Total	2.000	100%	—	—	3 segmen

**Fig. 1.** Distribution of GKInvest Trader Clusters**Fig. 2.** Relationship between Transaction Volume and Recency

The distribution indicates that the majority of traders fall into the Passive Trader (Churn Risk) segment, comprising 1,135 traders (56.7%), while the Highly Active Trader (VIP) group accounts for 437 traders (21.9%) and the Regular Active Trader group comprises 428 traders (21.4%). Differences in transaction volume and recency serve as the primary basis for distinguishing the characteristics of these three groups [13], [14].

Prediction of Activity Status Using Random Forest

The subsequent analysis employs Random Forest to predict the activity status of traders. The model was constructed using the same seven activity variables used in the K-Means analysis, but with classification targets defined as "Active" and "Inactive" statuses. The dataset was split into 80% training data and 20% testing data, resulting in 1,600 records used for training and 400 for testing. The Random Forest model was built using 300 decision trees with a random state of 42.

The use of Random Forest for classifying customer or user status is widely applied in research based on activity data. For instance, Azmi and Voutama utilized Random Forest and a confusion matrix to predict customer churn using bank client data [4]. Other studies on churn prediction also demonstrate that Random Forest can be used to classify customers based on activity and financial characteristics [4][15].

Table 2. Random Forest Model Evaluation Results

Metrik / class	Precision	Recall	F1 Score	Support
Active	0,84 (84,00%)	0,93 (93,00%)	0,88 (88,00%)	267
Inactive	0,82 (82,00%)	0,63 (63,00%)	0,71 (71,00%)	133
Accuracy	—	—	0,83 (83,00%)	400
Macro Avg	0,83 (83,00%)	0,78 (78,00%)	0,80 (80,00%)	400
Weighted Avg	0,83 (83,00%)	0,83 (83,00%)	0,82 (82,00%)	400

Evaluation results indicate that the Random Forest model achieved an accuracy of 83.00% (332 out of 400 test data points correctly classified). For the "Active" class, the model achieved a precision of 84.00%, a recall of 93.00%, and an F1 score of 88.00% [4].

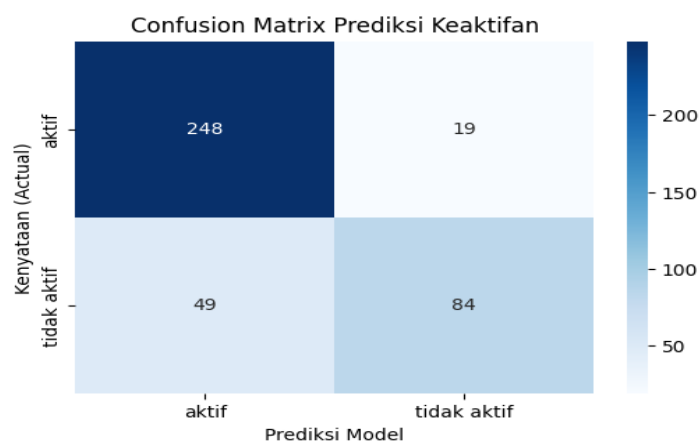


Fig. 3. Confusion Matrix Random Forest Model

The confusion matrix evaluation records the following details:

- True Positive (TP): 248 active traders correctly predicted as active.
- False Negative (FN): 19 active traders incorrectly predicted as inactive.
- False Positive (FP): 49 inactive traders incorrectly predicted as active.
- True Negative (TN): 84 inactive traders correctly predicted as inactive.

The total test dataset consists of 400 entries (248 + 19 + 49 + 84). Correct predictions account for 332 entries (83.00%), while classification errors account for 68 entries (17.00%).

Feature Importance in Predicting Activity Status

Random Forest was subsequently used to identify the relative contribution of the seven variables to the classification decision through feature importance. This analysis provides information on the variables that contribute the most to the model's decision-making process.

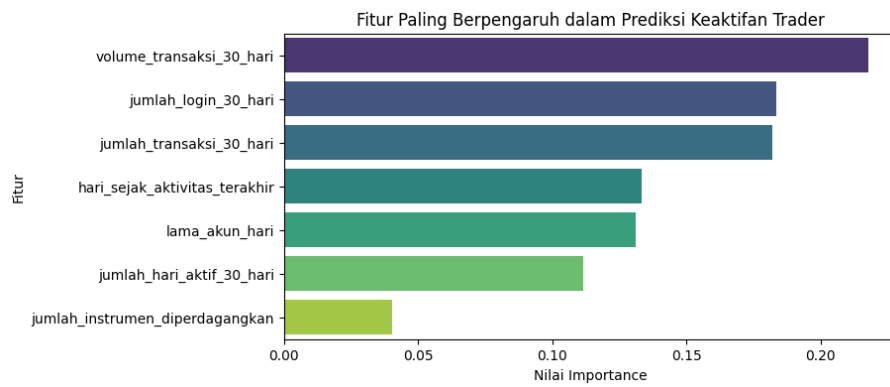


Fig. 4. Feature Importance Random Forest

The feature importance results indicate that 30-day transaction volume makes the largest relative contribution (~ 0.218), followed by the number of logins in the last 30 days (~ 0.184), the number of transactions in the last 30 days (~ 0.182), days since last activity (~ 0.133), account age in days (~ 0.131), number of active days in the last 30 days (~ 0.111), and the number of traded instruments (~ 0.040).

Table 3. Feature Importance Random Forest

Ranking	Variable	Importance
1	30-day transaction volume	0,218
2	Number of logins (30 days)	0,184
3	Number of transactions (30 days)	0,182
4	Days since last activity	0,133
5	Account age (days)	0,131
6	Number of active days (30 days)	0,111
7	Number of instruments traded	0,04

The feature importance results shown in Figure 3 and Table 5 indicate that 30-day transaction volume is the primary indicator, exerting the highest level of influence (~ 0.218). Subsequently, user interaction intensity represented by the number of logins over 30 days (~ 0.184) and the number of transactions over 30 days (~ 0.183) emerges as the next crucial factor in distinguishing between active and inactive traders. This suggests that the more frequently traders log in and engage in high-volume transactions, the greater the probability of them being classified as active traders [2], [15].

Discussion

The research results indicate that GKInvest trader activity follows a heterogeneous pattern. The K-Means algorithm successfully categorized traders into three main groups: Passive Traders (Churn Risk)—comprising 1,135 traders (56.7%); Highly Active Traders (VIP)—comprising 437 traders (21.9%); and Regular Active Traders—comprising 428 traders (21.4%). This segmentation demonstrates that trader activity characteristics are not uniform but are instead divided based on varying levels of intensity and engagement. These differences are particularly evident in transaction volume and recency characteristics; thus, employing a clustering method provides a more structured overview of trader activity behavior than simply looking at the total number of traders. The K-Means approach has also been utilized in user

segmentation studies based on transaction patterns to identify user groups with distinct behavioral characteristics [2], [3]. These findings align with Tan et al., who demonstrated that trading activity characteristics can be used to form trader groups with distinct behavioral patterns [10].

The Passive Trader (Churn Risk) group represents the largest segment, consisting of 1,135 traders, or 56.7% of the total dataset. The dominance of this group indicates that more than half of the traders in the dataset exhibit relatively low activity levels. Based on cluster characteristics, this group is defined by low transaction volumes and higher recency values. This condition reflects a relatively longer time gap since the traders' last activity compared to the more active trader groups. Consequently, this segment requires greater attention regarding the management of trader activity. The use of recency variables and transaction activity characteristics to identify user groups at risk of relationship issues with the company has also been observed in customer segmentation studies within the financial sector [3].

The significant proportion of "Passive Traders" indicates that trader management strategies should not focus solely on high-activity traders. The 56.7% of traders falling into this category represent a user segment at risk of declining platform engagement. This mapping enables GKInvest management to implement more targeted monitoring. Retention or re-engagement programs can be directed at traders exhibiting a combination of low transaction activity and high recency. Consequently, clustering results serve as a basis for identifying groups requiring special attention, rather than applying a uniform approach to all traders. A segmentation approach based on transaction characteristics also supports the formulation of tailored strategies for user groups with distinct behaviors [2], [3].

However, the "Churn Risk" associated with the Passive Trader group should be understood as an indicator based on activity characteristics, rather than definitive proof that churn has actually occurred. Traders with low activity and high recency have not necessarily ceased using the service entirely. Therefore, segmentation results are best used to identify groups at risk of declining activity, thereby signaling a need for further monitoring or follow-up strategies. This distinction is crucial to avoid overinterpreting clustering results as actual churn events. Research on customer churn indicates that identifying churn propensity is a predictive task that must be distinguished from the state of actual churn [4], [15].

The second group consists of "Highly Active Traders" (VIP), comprising 437 traders or 21.9% of the total dataset. This group is characterized by very high transaction volumes and very low recency. These characteristics indicate that traders in this group possess high trading intensity and a strong level of engagement with trading activities. Although smaller in number than the Passive Trader group, the VIP group is significant because it represents a segment of traders with activity levels far higher than those of other groups. Segmentation based on trading activity intensity is also employed by Tan et al. to distinguish trader profiles in derivatives markets according to trading activity characteristics [10].

From a management perspective, the "Highly Active Traders" group can be viewed as a segment whose engagement levels need to be maintained. The management strategy for this group differs from that of the "Passive Traders." While the passive group requires an approach aimed at revitalizing activity, the VIP group requires strategies that sustain consistent activity and the relationship with the company. Thus, K-Means segmentation provides a foundation for companies to implement distinct approaches based on the specific characteristics of each trader group. Utilizing segmentation results to apply tailored treatments or strategies to user groups aligns with research by Pratama and Putri, which demonstrates that user profiles based on transaction patterns can support more targeted strategies [2].

Meanwhile, the "Regular Active Traders" group comprises 428 traders, representing 21.4% of the total dataset. This group exhibits an average transaction volume approximately 1,233.80 falling between that of Passive Traders and Highly Active Traders, alongside a relatively short recency period. These conditions indicate that while the group engages in fairly routine trading activity, it has not yet reached the activity levels seen in the VIP group. Regular Active Traders represent a crucial group to retain, given their solid level of platform engagement. The pattern of group formation based on varying transaction intensities demonstrates that K-Means can effectively identify user profiles with distinct activity characteristics [2], [10].

The differing characteristics of these three groups highlight transaction volume and recency as critical factors in distinguishing trader activity levels. Passive Traders show an average transaction volume of approximately 349.7 and relatively high recency, whereas Highly Active Traders exhibit an average volume of around 4,344.90 and very low recency. Regular Active Traders occupy the middle ground between these two groups, with an average transaction volume of approximately 1,233.80. The pattern indicates that trader segmentation is determined not only by the volume of transaction activity but also by the recency of that activity. The use of recency alongside frequency or transaction volume characteristics as a basis for segmentation is also observed in RFM and K-Means approaches within the financial sector.

These segmentation results imply that trader activity is best understood through a combination of characteristics rather than a single indicator. Transaction volume provides insight into the intensity of trading activity, while recency offers information on how recently the activity occurred. When these two characteristics are analyzed concurrently, the distinctions between passive, regularly active, and highly active traders become more apparent. Consequently, the application of K-Means in this study provides a data-driven approach to understanding the heterogeneity of GKInvest traders. These findings align with research utilizing K-Means for user profiling based on transaction patterns, which demonstrates that combining transaction characteristics can yield distinct user groups with varying behaviors.

In terms of prediction, the Random Forest model demonstrated solid performance with an accuracy of 83.00%. Out of 400 test samples, 332 were correctly classified, while 68 were misclassified. This accuracy figure indicates that the model is capable of predicting trader activity status with a high degree of precision based on the activity variables employed. These results reinforce the suitability of Random Forest as a supervised learning approach for identifying trader activity status based on historical activity patterns. The application of Random Forest for classifying customer conditions or behaviors in the financial sector further demonstrates its utility, with performance evaluated through metrics such as accuracy, precision, recall, and F1-score.

The model's ability to recognize active traders is remarkably high, evidenced by a Recall score of 93.00% for the "Active" class, with 248 out of 267 actual active traders successfully detected. This Recall score indicates that the model can identify the vast majority of traders who are truly active. In the context of trader management, this capability is significant as it provides companies with an analytical tool to identify traders exhibiting active behavior patterns based on historical data. The use of recall as a key indicator for evaluating a model's ability to detect specific classes is also observed in Random Forest-based churn prediction research [5].

In addition to Recall, the model achieved a Precision of 84.00% and an F1-score of 88.00% for the "Active" class. A Precision of 84.00% indicates that the majority of traders predicted as "Active" were indeed active traders. Meanwhile, the F1-score of 88.00% reflects a relatively good balance between the model's ability to identify active traders and the accuracy of its predictions. Consequently, the model's performance is evidenced not only by its 83.00% accuracy but also by the combination of Precision, Recall, and F1-score metrics for the "Active" class. Using multiple evaluation metrics simultaneously is important for assessing the performance of classification models, as accuracy alone does not always provide a complete picture of performance across individual classes [4], [5].

Although the model demonstrates high capability in identifying the "Active" class, there is a performance disparity regarding the "Inactive" class. The recall for the "Inactive" class stands at 63.00%, indicating that the model's ability to identify traders who are actually inactive is lower than its ability to identify active traders. The confusion matrix reveals 49 False Positives traders who are actually Inactive but predicted as Active and 19 False Negatives traders who are actually Active but predicted as Inactive. These results indicate that the model still exhibits classification errors that must be considered when using it as a decision-support tool. Evaluations utilizing confusion matrices and various classification metrics are standard practice in churn prediction research to assess a model's class-discrimination capabilities in greater detail.

This difference in classification performance suggests that the activity patterns of Active traders are relatively easier for the model to recognize than those of Inactive traders. This may be because active traders

tend to generate clearer activity patterns through transactions, logins, and trading volumes. Conversely, inactive traders may exhibit more diverse characteristics, making the boundary between Active and Inactive classes difficult to distinguish perfectly based on the available variables. Consequently, the 83.00% accuracy rate should be interpreted alongside other metrics to gain a comprehensive understanding of the model's performance. Multi-metric evaluation is particularly important when the distribution or characteristics of the classes are not fully balanced.

Feature importance analysis reveals that transaction intensity and access frequency are the most dominant indicators in determining a trader's activity status. The 30-day transaction volume variable makes the largest relative contribution (approximately 0.218), followed by the number of logins over 30 days (0.184) and the number of transactions over 30 days (0.182). Meanwhile, days since last activity had a value of approximately 0.133, account age days was 0.131, active days in 30 days was 0.111, and number of traded instruments was 0.040. Utilizing feature importance within the Random Forest model helps identify the most informative variables for the classification process.

The dominance of 30-day transaction volume indicates that the magnitude of trading activity is the variable providing the most information to the model for distinguishing between trader activity statuses. This result aligns with the K-Means findings, as transaction volume is also a key characteristic distinguishing Passive Traders, Regular Active Traders, and Highly Active Traders. Thus, there is consistency between the segmentation and prediction results, confirming that trading intensity is a crucial characteristic in characterizing GKInvest trader activity. The findings regarding the importance of transaction patterns in shaping user profiles are consistent with research by Pratama and Putri, while studies by Tan et al. demonstrate that trading activity characteristics can serve as a basis for constructing trader profiles [2], [10].

Furthermore, the "number of logins in 30 days" variable ranking second demonstrates that the frequency of platform access plays a crucial role in distinguishing a trader's activity status. This indicates that trader activity is reflected not only in the transactions performed but also in the intensity of interaction with the platform. Traders who access the platform more frequently generate richer activity patterns, thereby providing the model with stronger information for classification purposes. This interpretation supports the view that user profiles can be constructed from a combination of transaction patterns and user interactions with digital services, rather than relying on a single activity indicator.

The "number of transactions in 30 days" variable, with a feature importance value of approximately 0.182, further underscores the significance of transaction frequency. While transaction volume indicates the magnitude of trading activity, the number of transactions reveals how often such activity occurs. The combination of these two variables suggests that a trader's activity status is more closely linked to the intensity and frequency of activity than to the mere occurrence of transactions. These results align with user behavior analysis approaches that utilize transaction characteristics to differentiate between user groups [3].

These findings demonstrate that indicators based on transaction activity and platform interaction are the most dominant variables within the model. In Random Forest research, identifying informative features is critical, as features containing high information content contribute significantly to classification performance. Prasetyowati et al. demonstrated that feature selection and data characteristic handling can enhance Random Forest performance. Similarly, Cong utilized feature importance within a Random Forest model to identify variables influencing customer churn prediction. Thus, the findings of this study reinforce the notion that historical activity data provides robust information for distinguishing trader activity statuses.

When the results from K-Means and Random Forest are viewed together, the two algorithms offer complementary perspectives. K-Means provides insights into trader segmentation based on similarities in activity characteristics, whereas Random Forest offers information regarding the predictive capability concerning trader activity status. In other words, K-Means addresses the question of how traders can be grouped based on their behavior, whereas Random Forest helps identify activity status based on patterns in activity variables. These two methods are employed independently; thus, the K-Means clustering results do not serve as input variables for the Random Forest model. Since using K-Means for segmentation and a classification model for prediction represents two distinct analytical functions, keeping them separate allows for a clearer interpretation of the segmentation and prediction results.

Combining these two results offers practical benefits for GKInvest. The "Passive Trader" group can be prioritized in monitoring strategies and retention programs, given their large share of the total and their relatively low activity levels. Strategies for "Regular Active Traders" can focus on maintaining consistent activity, while "Highly Active" or "VIP" traders can be prioritized for strategies aimed at sustaining high engagement levels. Meanwhile, the Random Forest model serves as a tool to identify activity status based on recorded activity patterns. Such segmentation and prediction approaches support more targeted user management strategies based on behavioral characteristics identified in the data [2], [3], [15]. Thus, the research findings not only provide insight into trader characteristics but also offer an analytical foundation for the company to formulate more targeted trader management strategies.

Overall, the results demonstrate that the heterogeneity of GKInvest trader activity can be identified using the K-Means approach, and activity status can be predicted using Random Forest. K-Means identified three main groups, with "Passive Traders" comprising the largest segment at 56.7%, while the Random Forest model achieved an accuracy of 83.00% and a recall of 93.00% for the "Active" class. Feature importance analysis further revealed that transaction volume, login frequency, and the number of transactions were the most dominant indicators in the prediction. These findings indicate that machine learning-based approaches can provide more structured information regarding trader characteristics and activities, thereby supporting trader monitoring and management processes at GKInvest.

IV. CONCLUSION

Based on the analysis, the application of K-Means and Random Forest provides comprehensive insights into the segmentation and activity prediction of GKInvest traders:

1. K-Means Segmentation: Resulted in three trader groups: Passive Traders (Churn Risk) comprising 1,135 traders (56.7%), Highly Active Traders (VIP) comprising 437 traders (21.9%), and Regular Active Traders comprising 428 traders (21.4%).
2. Random Forest Prediction: Achieved an accuracy of 83.00% on 400 test samples. For the Active class, Precision was 84.00%, Recall was 93.00%, and the F1 Score was 88.00%, with a total of 248 True Positives, 84 True Negatives, 49 False Positives, and 19 False Negatives.
3. Feature Importance: The most influential indicators for predicting trader activity, in descending order, were 30-day transaction volume (~0.218), 30-day login count (~0.184), 30-day transaction count (~0.182), days since last activity (~0.133), account age in days (~0.131), number of active days in the last 30 days (~0.111), and number of instruments traded (~0.040).

REFERENCES

- [1] S. A. Perdana, S. F. Florentin, and A. Santoso, "Analisis segmentasi pelanggan menggunakan K-Means clustering studi kasus aplikasi Alfagift," *Sebatik*, vol. 26, no. 2, pp. 446–457, 2022, doi: <https://doi.org/10.46984/sebatik.v26i2.1991>.
- [2] S. F. Pratama and N. A. Putri, "User Profiling Based on Financial Transaction Patterns: A Clustering Approach for User Segmentation," *Int. J. Appl. Inf. Manag.*, vol. 4, no. 4, pp. 217–228, 2024, doi: <https://doi.org/10.47738/ijaim.v4i4.92>.
- [3] H. Rizkyanto and L. G. Ford, "Customer segmentation of personal credit using recency, frequency, monetary (RFM) and K-means on financial industry," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 4, 2023, doi: <https://doi.org/10.14569/IJACSA.2023.0140417>.
- [4] A. F. Azmi and A. Voutama, "Prediksi churn nasabah bank menggunakan klasifikasi random forest dan decision tree dengan evaluasi confusion matrix," *KOMPUTA J. Ilm. Komput. dan Inform.*, vol. 13, no. 1, pp. 111–119, 2024, doi: <https://doi.org/10.34010/komputa.v13i1.12639>.
- [5] T. T. Cong, "Banking customer churn prediction using random forest based on smote and adasyn approach," *Tap chí Phát Triển và Hội Nhập*, no. 78, pp. 86–91, 2024, doi: <https://doi.org/10.61602/jdi.2024.78.11>.
- [6] S. Juanita and R. D. Cahyono, "K-means clustering with comparison of elbow and silhouette methods for medicines clustering based on user reviews," *J. Tek. Inform.*, vol. 5, no. 1, pp. 283–289, 2024, doi: <https://doi.org/10.52436/1.jutif.2024.5.1.1349>.

- [7] M. I. Prasetyowati, N. U. Maulidevi, and K. Surendro, "The accuracy of Random Forest performance can be improved by conducting a feature selection with a balancing strategy," *PeerJ Comput. Sci.*, vol. 8, p. e1041, 2022, doi: <https://doi.org/10.7717/peerj-cs.1041>.
- [8] M. Guntara and N. Lutfi, "Optimasi cacah klaster pada klasterisasi dengan algoritma kmeans menggunakan silhouette coefficient dan elbow method," *J. Teknol. Inf.*, vol. 2, no. 1, pp. 43–52, 2023, doi: <https://doi.org/10.26798/juti.v2i1.944>.
- [9] Y. Dong and Z. Zhang, "Deep forest modeling: An interpretable deep learning method for mineral prospectivity mapping," *J. Geophys. Res. Mach. Learn. Comput.*, vol. 1, no. 4, p. e2024JH000311, 2024, doi: <https://doi.org/10.1029/2024JH000311>.
- [10] E. H. L. Tan, Y. Hamed, H. Daud, M. A. F. Abdul Wahab, A. A. A. Azhar, and S. Y. Tan, "Profiling investor behavior in the Malaysian derivatives market using K-means clustering," *Front. Artif. Intell.*, vol. 8, p. 1640776, 2025, doi: <https://doi.org/10.3389/frai.2025.1640776>.
- [11] F. M. Ilyas and S. S. Priscila, "An optimized clustering quality analysis in k-means cluster using silhouette scores," *Explain. AI Appl. Hum. Behav. Anal.*, pp. 49–63, 2024, doi: [10.4018/979-8-3693-1355-8.ch004](https://doi.org/10.4018/979-8-3693-1355-8.ch004).
- [12] M. Yusran, R. Fadillah, and R. Shawn, "Segmentasi Pelanggan E-Commerce Berbasis Perilaku Belanja Menggunakan Algoritma K-Means Clustering dan Evaluasi Silhouette Score," *J. Comput. Informatics Res.*, vol. 5, no. 3, pp. 482–488, 2026, doi: <https://doi.org/10.47065/comforch.v5i3.3079>.
- [13] A. Haryoko, A. Putra, A. Arifia, and A. Nurlifa, "Penerapan Algoritma K-Means dengan Pendekatan RFM untuk Analisis Loyalitas Pelanggan pada Bisnis Online," *Curtina*, vol. 6, no. 1, pp. 9–17, 2025, doi: <https://doi.org/10.55719/curtina.v6i1.1745>.
- [14] R. Bhuria et al., "Ensemble-based customer churn prediction in banking: a voting classifier approach for improved client retention using demographic and behavioral data," *Discov. Sustain.*, vol. 6, no. 1, p. 28, 2025, doi: <https://doi.org/10.1007/s43621-025-00807-8>.
- [15] R. A. de Lima Lemos, T. C. Silva, and B. M. Tabak, "Propension to customer churn in a financial institution: A machine learning approach," *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11751–11768, 2022, doi: <https://doi.org/10.1007/s00521-022-07067-x>.