Vol. 7, No. 4, December 2025

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Taxpayer Classification Using K-Means Clustering to Support CRM Strategy Development: Case Study of Prabumulih City Samsat

Bimmo Fathin Tammam[1], Ali Ibrahim[2*], Dwi Rosa Indah[3], Ahmad Fali Oklilas[4], Yadi Utama[5]

[1,3,5] Information System, Faculty of Computer Science, Sriwijaya University, Indonesia

[2]Master of Computer Science, Faculty of Computer Science, Sriwijaya University, Indonesia

[4]Computer System, Faculty of Computer Science, Sriwijaya University, Indonesia

Email: bimmofathintammam@gmail.com[1], aliibrahim@unsri.ac.id[2*], dwirosaindah@unsri.ac.id[3], ahmadfalioklilas@unsri.ac.id[4] , yadiutama@unsri.ac.id[5]

correspondence*

Received: October 15, 2025

Revised: Nov 28, 2025

Accepted: Dec 2, 2025

Published: Dec 10, 2025

Corresponding Author:

Author Name*:

Ali Ibrahim

Email*:

aliibrahim@unsri.ac.id

DOI:

10.63158/journalisi.v7i4.1365

Abstract. Effective management of taxpayer data is crucial for enhancing compliance and optimizing regional revenue. This study addresses the limited use of data-driven taxpayer segmentation in local Samsat institutions by applying K-Means Clustering to support targeted Customer Relationship Management (CRM) strategies. A dataset of 3,999 motor vehicle taxpayer records from September 2025 was processed through feature selection, scaling, and clustering. The analysis identified three distinct taxpayer groups based on payment timeliness, compliance consistency, and vehicle age. Cluster validity was confirmed using the Davies-Bouldin Index, yielding a value of -41.327 for k = 3, supported by ANOVA for statistical significance. The findings highlight how clustering can reveal taxpayer behavior patterns, guiding personalized services and compliance programs. This study's novelty lies in integrating clustering outcomes with practical CRM strategies for public agencies, offering a data-driven approach to improve taxpayer engagement and regional revenue. However, the study is limited by its focus on a single-period dataset and vehicle-related attributes.

Keywords: K-Means Clustering, Taxpayer Segmentation, Customer Relationship Management, Davies–Bouldin Index, Public Service Analytics

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

In the digital era, managing taxpayer data efficiently has become a significant challenge for the public sector [1]. The sheer volume and complexity of administrative records often overwhelm traditional methods of management, highlighting the need for more sophisticated analytical tools to improve policy-making and service delivery [2]. With the rise of digital technologies, government agencies now have access to vast amounts of data that, if properly analysed, can lead to improved decision-making processes. One of the most effective strategies in this context is data segmentation, which allows for targeted public service strategies by grouping taxpayers based on similar characteristics [3]. This approach not only optimizes resource allocation but also helps tailor services to specific taxpayer needs.

A powerful analytical tool used in segmentation is K-Means Clustering, a method that groups data points based on similarity by minimizing the distance between data points and their respective cluster centroids [4], [5]. In vehicle tax management, K-Means Clustering can be particularly useful for analysing patterns in payment behaviour, delinquency levels, and compliance rates [6], [7], [8]. By leveraging tools such as RapidMiner and the Davies–Bouldin Index for evaluating cluster validity, agencies can gain insights into taxpayer behaviours that are critical for optimizing tax collection processes [9]. Numerous studies have demonstrated the versatility and effectiveness of this method across various domains. For instance, Nur et al. [10] applied K-Means to segment property taxpayers, while Khan et al. [11] worked on improving the interpretability of clustering results. Ridwan et al. [12] and Gopalakrishnan [13] employed this technique in the e-commerce and banking sectors, showcasing its broad applicability in diverse industries.

In addition to clustering, the integration of Customer Relationship Management (CRM) strategies has proven valuable in enhancing taxpayer engagement. CRM frameworks enable the personalized communication of services, which can significantly improve taxpayer compliance by targeting specific groups with tailored messages [15], [16]. For example, understanding taxpayers' behaviour through segmentation can help design outreach programs that increase awareness and provide tailored solutions to specific compliance challenges. Factors such as digital adoption, penalties, and public awareness

play a critical role in shaping taxpayers' willingness to comply with regulations [17], [18], [19], [20]. By combining segmentation with CRM strategies, public agencies can create more effective and targeted interventions, thus improving both compliance rates and taxpayer satisfaction.

Despite these advancements, few studies have integrated K-Means segmentation with CRM strategy formulation, especially in the context of regional Samsat institutions, where vehicle tax collection is a critical aspect of local revenue generation. This gap is particularly evident in Prabumulih City, where an analysis of 3,999 taxpayer records from September 2025 reveals that 1,000 taxpayers (approximately 25%) consistently delay payments. The current system at the Prabumulih Samsat relies on a uniform service approach and manual processes, which prevents effective differentiation between compliant and non-compliant taxpayers. This inefficiency leads to suboptimal resource allocation and missed opportunities for engagement and improved compliance.

Therefore, this study aims to address this gap by utilizing K-Means clustering to segment vehicle taxpayers in Prabumulih City. By identifying distinct taxpayer groups based on their payment behaviours, the study will develop tailored CRM strategies for each segment. These strategies will focus on enhancing compliance through personalized communication and targeted interventions. Additionally, the study will provide actionable recommendations to optimize revenue management at the Prabumulih City Samsat. Ultimately, this research will contribute to a more data-driven approach to taxpayer engagement, offering valuable insights that could be applied to improve the efficiency of regional tax collection processes across the country.

2. METHODS

This study follows a systematic research process to ensure validity, reliability, and scientific accountability, as illustrated in Figure 1.1. The process begins with data selection from the SAMSAT database, followed by preprocessing to ensure data quality. K-Means clustering is applied to segment taxpayers based on payment behavior, then evaluated using Davies-Bouldin Index, statistical tests, scatter plots, and centroid analysis. Finally, CRM strategies are formulated based on the clustering results, ensuring integration between problem definition, methodology, and analysis [21].

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

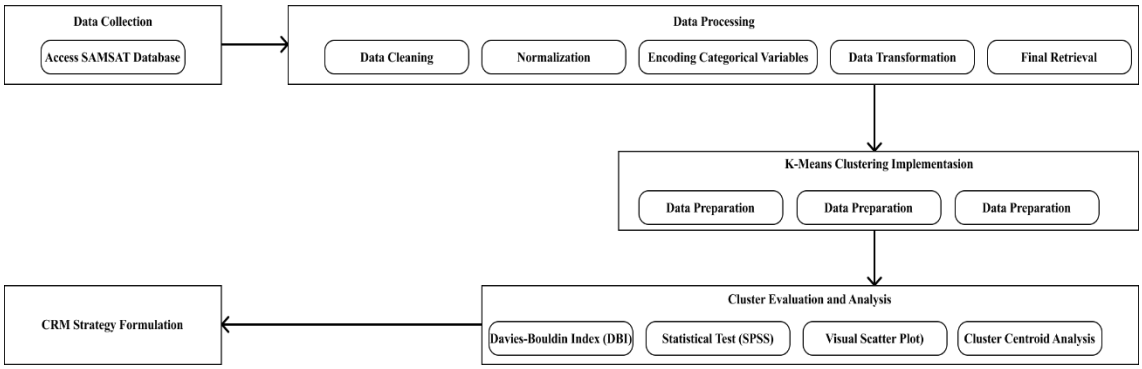

Figure 1. Flowchart of Research Stages

2.1 Data Collection

Data were collected from the SAMSAT Kota Prabumulih database for September 2025, representing a one-month cross-sectional snapshot of motor vehicle tax transactions. The dataset includes taxpayer information, vehicle attributes, and motor vehicle tax payment records. An initial verification process was conducted to ensure completeness, consistency, and accuracy by checking for missing entries, validating payment records, and confirming the correctness of taxpayer and vehicle data. The initial dataset contained 5,000 taxpayer records, which was reduced to 4,000 valid entries after systematic verification and removal of incomplete or inconsistent data. This verification ensures that the dataset is reliable and suitable for the preprocessing and clustering stages. The data collection and validation procedures follow administrative research standards to maintain validity, reliability, and representativeness [22], [23].

2.2 Data Preprocessing

Data preprocessing was carried out using RapidMiner Studio to prepare the September 2025 dataset for K-Means clustering. This stage is essential because clustering model performance depends heavily on data structure and quality [24]. The preprocessing workflow, illustrated in Figure 2, began with data cleaning to remove duplicate entries, correct inconsistent values, and discard non-contributive attributes. Table 1 summarizes the attributes removed during preprocessing and their justification.

Table 1. Removed Attributes and Justification

Attribute	Data Type	Reason for Removal
Owner Address	Text	High cardinality, no analytical value for payment behavior
Chassis Number	Alphanumeric	Unique identifier, no clustering relevance
Engine Number	Alphanumeric	Unique identifier, no clustering relevance
Owner Name	Text	Privacy-sensitive, not related to behavior

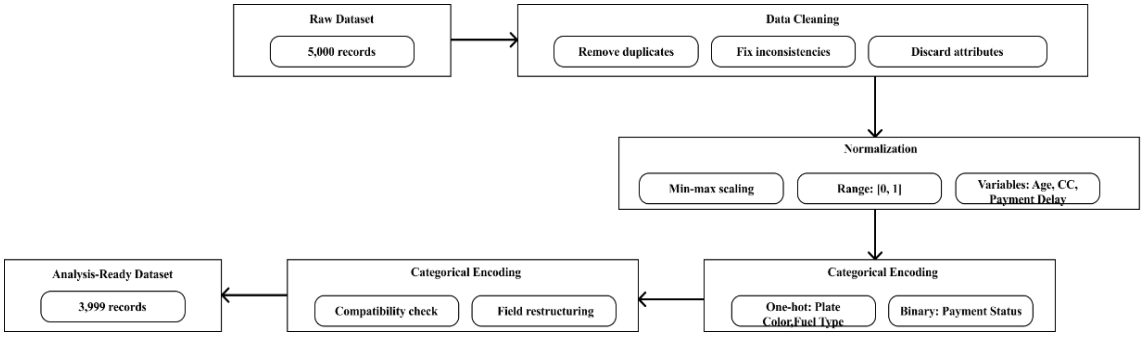

Figure 2. Preprocessing Pipeline Workflow

The dataset was then standardized through min-max normalization, scaling key numerical attributes (vehicle age, engine capacity, payment delay) to a 0–1 range [25]. This scaling is critical for K-Means because the algorithm uses Euclidean distance calculations [26]. Variables with larger ranges (e.g., engine capacity: 100-2,500 cc) would dominate cluster formation compared to smaller-range variables (e.g., payment status: 0-1). Normalization ensures all attributes contribute equally to distance measurements and prevents bias toward high-magnitude features [27].

Categorical attributes were converted into numerical form through encoding techniques [28]. Plate color and fuel type, both nominal variables with multiple categories, were transformed using one-hot encoding to preserve non-ordinal relationships [29]. For example, plate color {White, Black, Red} was encoded into three binary columns (TNKB_White, TNKB_Black, TNKB_Red). Payment status, being dichotomous (On Time/Late), was converted using binary encoding (0/1) for efficiency [30]. Additional data transformation steps restructured fields to ensure compatibility with RapidMiner's K-

Means operator. After preprocessing, the dataset contained 3,999 records ready for clustering. Table 2 presents the complete preprocessing workflow.

Table 2. Preprocessing Steps in RapidMiner

Step	Description	Purpose	Output
Data Cleaning	Remove duplicates, fix inconsistencies, discard irrelevant attributes (Table 2)	Ensure data integrity and reduce noise	Cleaned dataset (3,999 records)
Normalization	Scale numerical attributes to 0–1 using min-max normalization	Prevent dominance in distance calculation	Normalized dataset
Encoding	One-hot encoding (plate color, fuel type); binary encoding (payment status)	Convert categorical to numeric form	Numeric dataset
Transformation	Restructure fields for RapidMiner compatibility	Ensure algorithm compatibility	Transformed dataset
Final Retrieval	Import prepared dataset	Ready for K-Means clustering	Analysis-ready dataset (3,999 records, 8 attributes)

2.3 K-Means Clustering Implementation

K-Means clustering is used in this study to segment taxpayers based on payment behaviors and related attributes. The resulting clusters (high, medium, and low compliance) support the development of CRM strategies tailored to taxpayer characteristics. K-Means partitions data into k clusters through an iterative process involving centroid initialization, distance calculation, data assignment, and centroid updates. The algorithm continues until the centroids stabilize, indicating convergence. Before clustering, the optimal number of clusters was determined using the Within-Cluster Sum of Squares (WCSS). WCSS decreases as k increases, but the reduction becomes minimal after a certain point. The Elbow Method is used to identify this point.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 3. WCSS Results for Different k Values

Number of Clusters (k)	WCSS Value	Decrease (Δ)
2	9186.924	–
3	9182.091	4.833
4	9180.681	1.410

The sharp decline occurs between k = 2 and k = 3, while the reduction becomes smaller from k = 3 to k = 4. Therefore, k = 3 was selected as the optimal number of clusters.

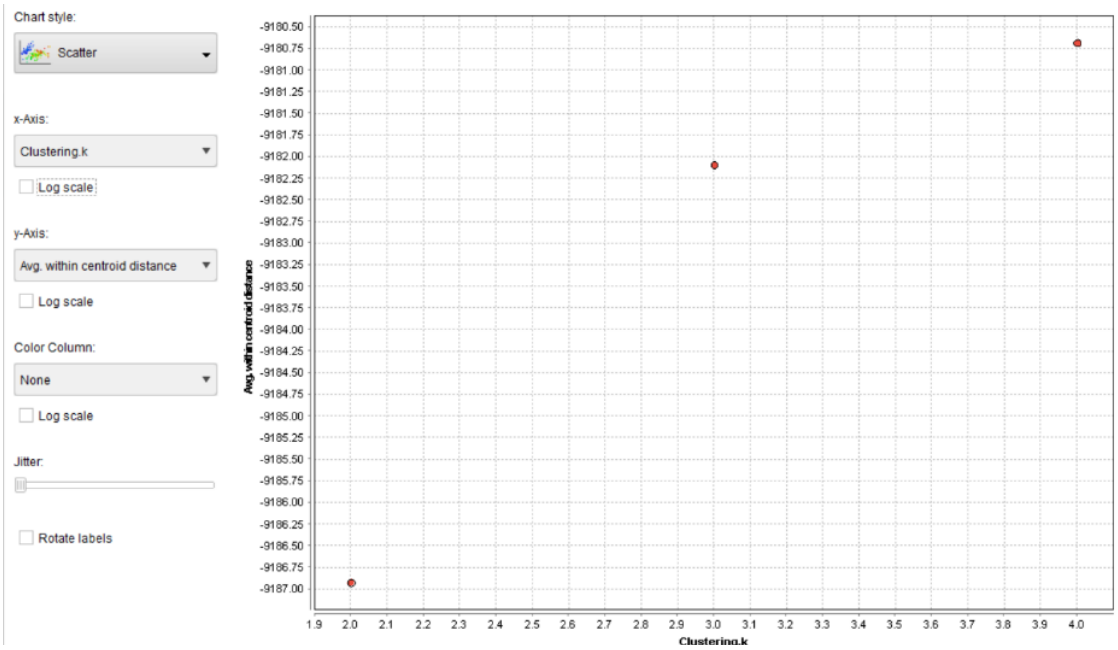

Figure 3 WCSS Plot for Determining the Optimal Number of Clusters

Figure 3 shows the relationship between the number of clusters and their WCSS values. The elbow appears at k = 3, confirming the optimal cluster configuration for this study. After determining the optimal value of k, the clustering process was carried out following the workflow shown in Figure 4.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

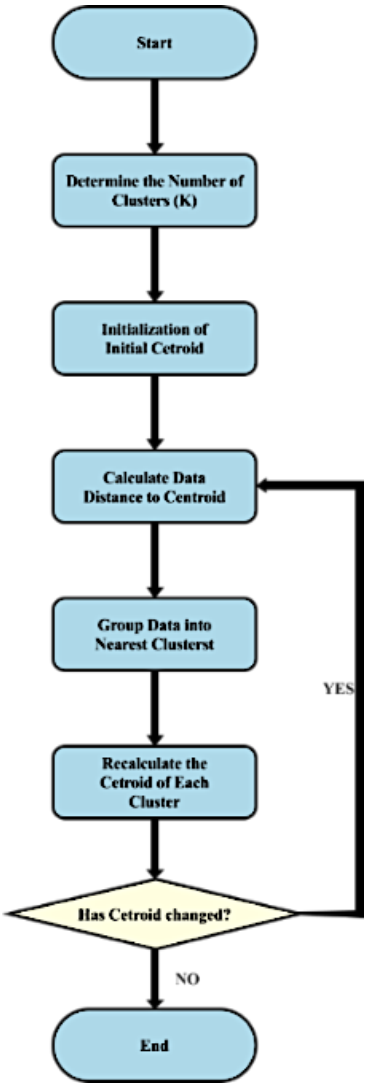

Figure 4. Flowchart of K-Means Clustering Algorithm

Based on Figure 4, the workflow of the K-Means algorithm can be described as follows:

1) Determine the desired number of clusters (k).
2) Initialize the initial centroid values randomly from the available dataset.
3) The centroid for each cluster is computed using Equation 1.

$$v_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} \quad (1)$$

Where

v_{ij}= centroid of the i-th cluster for the j-th variable

N_i= total data points in the i-th cluster

i, k= cluster indices

j= variable index

x_{kj}= value of the k-th data point in the j-th variable within the cluster

4) Calculate the distance between each data point and the centroids using Euclidean Distance (Equation 2):

$$D_e = \sqrt{(x_i - \bar{s}_i)^2 + (y_i - \bar{t}_i)^2} \quad (2)$$

Where

D_e= Euclidean Distance

i= number of objects

(x, y)= coordinates of the object

$(\bar{s}, \bar{t})$= coordinates of the cluster centroid

5) Assign data points to the cluster with the nearest centroid based on the minimum distance calculated using Equation (2).

6) Recalculate the centroid positions based on the mean values within each cluster using Equation (1).

7) Repeat steps 3 to 5 until centroids no longer change significantly (convergence), indicating that the clustering structure has stabilized.

Through these steps, taxpayer data can be grouped objectively, making the clustering results easier to analyze and apply in the development of CRM strategies at Samsat Kota Prabumulih.

2.4 Cluster Evaluation and Analysis

Cluster quality was evaluated using the Davies–Bouldin Index (DBI) to determine the optimal number of clusters, with lower DBI values indicating more distinct and well-separated clusters [32]. The DBI measures the ratio of within-cluster scatter to between-cluster separation, providing a quantitative assessment of clustering quality where values closer to zero signify superior cluster definition and minimal overlap between groups. In this study, negative DBI values were obtained due to the use of normalized data in the range of 0 to 1, which can result in negative distance calculations when cluster centroids are positioned relative to the scaled feature space. Negative DBI values do not indicate poor clustering quality; rather, they suggest exceptionally well-separated clusters where between-cluster distances are substantially larger than within-cluster scatter, confirming that the algorithm successfully identified distinct taxpayer segments with minimal overlap.

Statistical validation using ANOVA in IBM SPSS was performed on key variables including vehicle age, engine capacity, and payment status, where a p-value less than 0.05 indicates significant differences among clusters [33]. The ANOVA test examines whether the means of these variables differ significantly across the identified clusters, thereby confirming that the clustering algorithm has successfully partitioned taxpayers into groups with genuinely distinct characteristics rather than arbitrary divisions. Additionally, Chi-Square tests were applied to categorical variables such as payment status to validate the association between cluster membership and compliance behavior. This multi-method validation approach combining the DBI for internal cluster quality assessment, ANOVA for numerical variables, and Chi-Square tests for categorical variables ensures the clustering results are robust, reliable, and statistically valid, providing a solid foundation for subsequent CRM strategy formulation.

2.5 CRM Strategy Formulation

Clusters were systematically utilized to develop and implement tailored Customer Relationship Management (CRM) strategies that align with the specific behavioral patterns and compliance characteristics of each taxpayer segment. High-compliance taxpayers, who consistently demonstrate responsible tax behavior and timely payment patterns, may receive appreciation through reward programs, incentive schemes, or priority service benefits to maintain and reinforce their positive compliance behavior. Conversely, low-compliance clusters, characterized by irregular payment patterns or delayed submissions, receive targeted educational campaigns, personalized guidance materials, or automated reminder systems designed to improve their understanding of tax obligations and encourage better compliance practices. This differentiated approach not only addresses the diverse needs of various taxpayer segments but also strengthens long-term taxpayer engagement and fosters a more positive relationship between tax authorities and citizens based on the unique characteristics and requirements identified within each cluster.

3. RESULTS AND DISCUSSION

This chapter presents results from the data processing and analytical procedures, describing findings systematically and discussing their connection to research objectives and theoretical foundations. It provides a comprehensive overview of taxpayer patterns

and characteristics identified through clustering, along with implications for CRM strategy formulation at Samsat Kota Prabumulih.

3.1 Data Collection

The dataset used in this study consists of secondary data obtained through an official data-sharing agreement with Samsat Kota Prabumulih. The data were extracted directly from the internal Samsat information system and contain complete records of registered motor vehicles within the Prabumulih area, including vehicle characteristics, owner information, and annual tax payment status. The initial dataset contained 5,000 taxpayer records. However, several issues such as incomplete entries, inconsistent formatting, invalid values, and duplicate records were identified during the initial inspection. After a systematic manual cleaning process to remove these errors, 4,000 valid entries remained and were used for further analysis. This ensures that the dataset accurately represents the taxpayer population and is suitable for clustering. The dataset includes key attributes relevant to the identification of taxpayer behavior and compliance patterns. These variables cover demographic information, vehicle specifications, and tax payment indicators. A summary of the attributes used in this study is presented in Table 4.

Table 4. Motor Vehicle Taxpayer Data Attributes

No	Attribute	Data Type	Description
1	Vehicle Number	Nominal	Vehicle registration number
2	Owner Name	Nominal	Name of the vehicle owner
3	Owner Address	Nominal	Residential address of the vehicle owner
4	Year	Numeric	Vehicle year of manufacture
5	TNKB	Categorical	Vehicle plate color
6	Fuel Type	Categorical	Type of fuel (Gasoline/Diesel)
7	CC	Numeric	Engine capacity (cc)
8	Samsat	Categorical	Samsat office where the vehicle is registered
9	Tax Due Date	Date	Due date for annual tax payment
10	Payment Status	Categorical	Tax payment status (Overdue/On Time)

A preview of the cleaned dataset before import into RapidMiner Studio version 9.10.008 is shown in Figure 5. The preview confirms the removal of duplicated and invalid entries, with all variables structured consistently and ready for preprocessing and clustering analysis.

A	B	C	D	E	F	G	H	I	J	K
NO	NOMOR POLISI	NAMA PEMILIK	ALAMAT PEMILIK	TAHUN	TNKB	BBM	CC	SAMSAT	TGL PKB	Status
1	BG1306CS	SURYA ANANTA	JL. PADAT KARYA RT/RW 010/001 KEL. GUNUNG IBUL KEC. PRABUMULIH	2000	PUTIH	SOLAR	2499	PRABUMULIH	01/09/2025	Menunggak
2	BG1328CS	AHMAD HAQIQI RINO, SPT	JL JAMBAT AKAR NO 43 RT/RW 030/012 KEL MANGGA BESAR KEC	2023	PUTIH	BENSIN	1987	PRABUMULIH	01/09/2025	Menunggak
3	BG1437CO	HADI BASTIAN	JL. TANJUNG LAUT NO. 069 RT/RW 003/007 KEL. GUNUNG IBUL KEC. PRABUMULIH TIMUR KOTA PRABUMULIH	2016	HITAM	BENSIN	1496	PRABUMULIH	01/09/2025	Menunggak
4	BG1501CO	MIRNAWATI BOMANTARA	JL.JENDRAL SUDIRMAN RT/RW 003/002 KEL. PATIH GALUNG KEC.	2016	HITAM	BENSIN	1497	PRABUMULIH	01/09/2025	Menunggak
5	BG1797CP	TRIDES NOPRIADI	KEPODANG INDAH BLOK D5/09 RT/RW 002/003 KEL. PATIH GALUNG KEC. PRABUMULIH BARAT KOTA PRABUMULIH	2001	HITAM	SOLAR	2499	PRABUMULIH	01/09/2025	Menunggak
6	BG8675CI	SYAMSUDDIN	JLN FLORES GANG MALAKA NO 44 RT/RW 002/001 KEL. GUNUNG IBUL BARAT KEC. PRABUMULIH TIMUR KOTA PRABUMULIH	2015	PUTIH	BENSIN	1495	PRABUMULIH	01/09/2025	Menunggak
7	BG2268CY	JULIAN SAPUTRA	JL KP LEGOK RT. 002 RW. 001 KEL.SUKARAJA KEC. PRABUMULIH SELATAN PRABUMULIH	2021	HITAM	BENSIN	155	PRABUMULIH	01/09/2025	Menunggak
8	BG2892C	SUPARTI	JL. AIR MENDIDIH RT/RW 003/003 KEL. SUKARAJA KEC. PRABUMULIH SELATAN KOTA PRABUMULIH	2023	PUTIH	BENSIN	156	PRABUMULIH	01/09/2025	Menunggak
9	BG2898C	SITI MARIYAM	JL.MUARA TIGA RT/RW 002/002 KEL. ANAK PETAI KEC. PRABUMULIH UTARA KOTA PRABUMULIH	2023	PUTIH	BENSIN	124	PRABUMULIH	01/09/2025	Menunggak
10	BG3809CA	HANTER SINAGA	JL. SUMATRA GG. RAMBANG LUBAI RT/RW 008/004 KEL. GUNUNG IBUL KEC. PRABUMULIH TIMUR KOTA PRABUMULIH	2022	PUTIH	BENSIN	109	PRABUMULIH	01/09/2025	Menunggak
11	BG3816CA	ENDANG YUNITA	JL.PELITA RT/RW 003/003 KEL. GUNUNG IBUL BARAT KEC. PRABUMULIH TIMUR KOTA PRABUMULIH	2022	PUTIH	BENSIN	109	PRABUMULIH	01/09/2025	Menunggak
12	BG3818CA	SUYATMI	JL. TAMAN MURNI RT/RW 001/001 KEL. GUNUNG IBUL BARAT KEC. PRABUMULIH TIMUR KOTA PRABUMULIH	2022	PUTIH	BENSIN	109	PRABUMULIH	01/09/2025	Menunggak
13	BG3841CA	AHMAD TANZILI	JL.VOLLY NO 229/02 RT/RW 002/001 KEL. PRABU JAYA KEC. PRABUMULIH TIMUR KOTA PRABUMULIH	2022	PUTIH	BENSIN	156	PRABUMULIH	01/09/2025	Menunggak
14	BG3847CA	MUHAMMAD RAHMAT	JL. SEMINUNG PERUM GRIYA CIPTA UTAMA BLOK B-3 RT/RW 002/005 KEL MUARA DUA KEC. PRABUMULIH TIMUR KOTA PRABUMULIH	2022	PUTIH	BENSIN	124	PRABUMULIH	01/09/2025	Menunggak
15	BG3857CA	NAZIFA NASPALA DEWI	JL.TROMOL RT/RW 001/001 KEL. SUKARAJA KEC. PRABUMULIH SELATAN KOTA PRABUMULIH	2022	PUTIH	BENSIN	109	PRABUMULIH	01/09/2025	Menunggak
16	BG5235CN	NAIMAH	JL.KERINCI BLOK A NO.1 M.DUA KOTA PRABUMULIH PRABUMULIH	2010	HITAM	BENSIN	110	PRABUMULIH	01/09/2025	Menunggak
17	BG5764CZ	KECAMATAN CAMBAI	JL.JEND.SUDIRMAN NO.10 KEL.CAMBAIKEC.CAMBAI KOTA PRABUMULIH	2015	MERAH	BENSIN	113	PRABUMULIH	01/09/2025	Menunggak
18	BG5983CC	ROSADI	JL.KOP TOYA NO.942 RT.02 KOTA PRABUMULIH	2004	HITAM	BENSIN	100	PRABUMULIH	01/09/2025	Menunggak
19	BG6022CZ	PEMERINTAH KOTA PRABUMULIH	JL. JENDRAL SUDIRMAN KM 12 SINDUR CAMBAI KEL.SINDUR KEC. CAMBAI KOTA PRABUMULIH	2022	MERAH	BENSIN	113	PRABUMULIH	01/09/2025	Menunggak
20	BG6023CZ	PEMERINTAH KOTA PRABUMULIH	JL. JENDRAL SUDIRMAN KM 12 SINDUR CAMBAI KEL.SINDUR KEC. CAMBAI KOTA PRABUMULIH	2022	MERAH	BENSIN	113	PRABUMULIH	01/09/2025	Menunggak
21	BG6024CZ	PEMERINTAH KOTA PRABUMULIH	JL. JENDRAL SUDIRMAN KM 12 SINDUR CAMBAI KEL.SINDUR KEC. CAMBAI KOTA PRABUMULIH	2022	MERAH	BENSIN	113	PRABUMULIH	01/09/2025	Menunggak
22	BG6025CZ	PEMERINTAH KOTA PRABUMULIH	JL. JENDRAL SUDIRMAN KM 12 SINDUR CAMBAI KEL.SINDUR KEC. CAMBAI KOTA PRABUMULIH	2022	MERAH	BENSIN	113	PRABUMULIH	01/09/2025	Menunggak
23	BG6028CZ	PEMERINTAH KOTA PRABUMULIH	JL. JENDRAL SUDIRMAN KM 12 SINDUR CAMBAI KEL.SINDUR KEC. CAMBAI KOTA PRABUMULIH	2022	MERAH	BENSIN	113	PRABUMULIH	01/09/2025	Menunggak

Figure 5. Sample of Motor Vehicle Taxpayer Dataset (Cleaned) in Excel

3.2 Data Preprocessing

The study utilized a comprehensive dataset of 5,000 motor vehicle taxpayer records obtained from the internal database of Samsat Kota Prabumulih. This initial dataset contained detailed information about vehicle owners and their respective vehicles, including registration numbers, owner identities, vehicle specifications, payment histories, and compliance status. However, the raw data required substantial cleaning and transformation before it could be used for analytical purposes. Manual data cleaning procedures were first conducted to systematically identify and remove obvious errors, inconsistencies, and duplicate entries. This initial cleaning phase reduced the dataset to 4,000 valid entries, which were then imported into RapidMiner Studio version 9.10.008 for automated preprocessing operations.

During the preprocessing stage in RapidMiner, several critical data transformation procedures were applied to ensure the dataset met the technical requirements for K-Means clustering. Duplicate entries were further verified and removed, while missing

values were addressed using appropriate imputation methods such as mean substitution for numerical fields or mode substitution for categorical fields.

Numerical attributes, including vehicle age, engine capacity (CC), and annual tax payment amounts, were normalized to a 0–1 scale using min-max normalization. This step was essential to prevent variables with larger numeric ranges from dominating distance calculations in K-Means. Categorical attributes, including vehicle plate color, fuel type, and tax payment status, were encoded into numerical representations using dummy coding or ordinal mapping, allowing the algorithm to calculate Euclidean distances correctly. After preprocessing, the final dataset consisted of 3,999 high-quality, structured entries. Each entry contained complete information across all relevant attributes, with all variables properly scaled, encoded, and formatted for K-Means analysis.

Table 5. Preprocessing Steps, Purpose, and Output in RapidMiner

Preprocessing Stage	Action Taken	Dataset Size After Step
Manual Cleaning	Removed obvious duplicates and errors	4,000
Retrieve Data	Imported into RapidMiner	3,999
Data Cleaning	Removed additional duplicates and corrected missing/inconsistent values	3,999
Normalization	Numerical attributes scaled to 0–1	3,999
Encoding Categorical Variables	Converted TNKB, Fuel Type, Payment Status to numeric	3,999
Data Transformation	Aggregated/restructured columns for K-Means input	3,999

These preprocessing procedures ensured that the dataset was clean, consistent, and structured, providing a reliable basis for the subsequent K-Means clustering analysis to identify patterns in taxpayer behavior and compliance levels. By systematically addressing data quality issues, standardizing variable scales, and converting all attributes into numerical formats, the preprocessing phase eliminates potential sources of bias and error that could distort clustering results. This comprehensive data preparation enhances

the accuracy and validity of the clustering model and ensures that the identified taxpayer segments reflect genuine behavioral patterns.

	NO	NOMOR POL...	NAMA PEMILIK	ALAMAT PE...	TAHUN	TNKB	BBM	CC	SAMSAT	TGL PKB	Status
1	1	BG1306CS	SURYA ANANTA	JL. PADAT KA...	2000	PUTIH	SOLAR	2499	PRABUMULIH	Sep 1, 2025	Menunggak
2	2	BG1328CS	AHMAD HAQIQI RINO, SPT	JL JAMBAT A...	2023	PUTIH	BENSIN	1987	PRABUMULIH	Sep 1, 2025	Menunggak
3	3	BG1437CO	HADI BASTIAN	JL. TANJUNG...	2016	HITAM	BENSIN	1496	PRABUMULIH	Sep 1, 2025	Menunggak
4	4	BG1501CO	MIRNAWATI BOMANTARA	JL.JENDRAL ...	2016	HITAM	BENSIN	1497	PRABUMULIH	Sep 1, 2025	Menunggak
5	5	BG1797CP	TRIDES NOPRIADI	KEPODANG I...	2001	HITAM	SOLAR	2499	PRABUMULIH	Sep 1, 2025	Menunggak
6	6	BG8675CI	SYAMSUDDIN	JLN FLORES...	2015	PUTIH	BENSIN	1495	PRABUMULIH	Sep 1, 2025	Menunggak
7	7	BG2268CY	JULIAN SAPUTRA	JL KP LEGO...	2021	HITAM	BENSIN	155	PRABUMULIH	Sep 1, 2025	Menunggak
8	8	BG2892C	SUPARTI	JL. AIR MEN...	2023	PUTIH	BENSIN	156	PRABUMULIH	Sep 1, 2025	Menunggak
9	9	BG2898C	SITI MARIYAM	JL.MUARA TI...	2023	PUTIH	BENSIN	124	PRABUMULIH	Sep 1, 2025	Menunggak
10	10	BG3809CA	HANTER SINAGA	JL. SUMATRA...	2022	PUTIH	BENSIN	109	PRABUMULIH	Sep 1, 2025	Menunggak
11	11	BG3816CA	ENDANG YUNITA	JL.PELITA RT...	2022	PUTIH	BENSIN	109	PRABUMULIH	Sep 1, 2025	Menunggak
12	12	BG3818CA	SUYATMI	JL. TAMAN M...	2022	PUTIH	BENSIN	109	PRABUMULIH	Sep 1, 2025	Menunggak
13	13	BG3841CA	AHMAD TANZILI	JL.VOLLY NO...	2022	PUTIH	BENSIN	156	PRABUMULIH	Sep 1, 2025	Menunggak
14	14	BG3847CA	MUHAMMAD RAHMAT	JL. SEMINUN...	2022	PUTIH	BENSIN	124	PRABUMULIH	Sep 1, 2025	Menunggak
15	15	BG3857CA	NAZIFA NASPALA DEWI	JL.TROMOL ...	2022	PUTIH	BENSIN	109	PRABUMULIH	Sep 1, 2025	Menunggak
16	16	BG5235CN	NAIMAH	JL.KERINCI ...	2010	HITAM	BENSIN	110	PRABUMULIH	Sep 1, 2025	Menunggak
17	17	BG5764CZ	KECAMATAN CAMBAI	JL.JEND.SU...	2015	MERAH	BENSIN	113	PRABUMULIH	Sep 1, 2025	Menunggak
18	18	BG5983CC	ROSADI	JL.KOP.TOYA...	2004	HITAM	BENSIN	100	PRABUMULIH	Sep 1, 2025	Menunggak

ExampleSet (3,999 examples, 0 special attributes, 11 regular attributes)

Figure 6. Display of Preprocessed Motor Vehicle Taxpayer Dataset in RapidMiner

3.3 K-Means Clustering Implementation

This stage applies K-Means clustering to group motor vehicle taxpayers based on payment behavior and relevant attributes, including Vehicle Year, Engine Capacity (CC), Fuel Type, and Tax Payment Status. The primary goal is to identify underlying taxpayer patterns and behavioral similarities that can inform the design of effective, targeted Customer Relationship Management (CRM) strategies at Samsat Kota Prabumulih. By segmenting taxpayers into homogeneous groups, the analysis enables the development of differentiated service approaches tailored to the specific characteristics of each cluster.

The K-Means algorithm was implemented in RapidMiner Studio version 9.10.008 using the preprocessed dataset. The clustering workflow includes data import, application of the K-Means operator with specified parameters, automatic cluster assignment based on minimum distance to cluster centroids, visualization of results, and evaluation using internal performance metrics such as the Davies–Bouldin Index (DBI) and Within-Cluster Sum of Squares (WCSS). Multiple cluster configurations were tested to determine the optimal number of clusters. Figure 5 illustrates the complete clustering process workflow as configured in RapidMiner, showing the sequence of operators and data flow from

input to final output. Clusters were formed by iteratively assigning taxpayers to the nearest centroid and recalculating centroids until convergence. The centroid of each cluster is calculated using the formula as shown in Equation 1. where v_{ij}represents the centroid of the i-th cluster for the j-th variable, $N_i$is the total number of data points in the cluster, and $x_{kj}$is the value of the k-th data point in the j-th variable. These steps are repeated iteratively until the centroid positions stabilize, indicating convergence. Distances from each taxpayer to the cluster centroids were measured using the Euclidean Distance as shown in Equation 2. where D_e represents the distance between a data point and the centroid, x denotes the coordinates of the data point, and c denotes the coordinates of the cluster centroid. This mathematical formulation provides the foundation for the K-Means algorithm to measure similarity and assign each taxpayer record to the most appropriate cluster based on proximity.

The optimal number of clusters was determined by evaluating multiple cluster solutions (k = 2, 3, 4) using DBI and WCSS. The configuration with k = 3 was selected, as it provided the lowest DBI value and a balance between model simplicity and meaningful differentiation of taxpayer behavior. The resulting clusters revealed distinct taxpayer segments:

a) High-compliance cluster: newer vehicles with timely payments
b) Low-compliance cluster: older vehicles with delayed payments
c) Intermediate cluster: mixed characteristics in vehicle age and payment patterns

These segments inform CRM strategies such as reward programs for compliant taxpayers and targeted reminders or educational campaigns for low-compliance groups. Differentiation among clusters provides tax authorities with a clear framework for resource allocation and intervention design, enabling more efficient and effective taxpayer engagement. Cluster visualization using RapidMiner's Cluster Model Visualizer confirmed clear separation between clusters, validating that the algorithm captured behavioral differences effectively. The visual representation provides an intuitive and accessible way to communicate findings to stakeholders, supporting informed decision-making regarding CRM strategy implementation.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

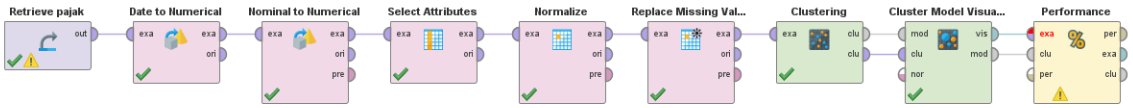

Figure 7 Clustering Process Workflow in RapidMiner

3.4 Cluster Evaluation and Analysis

This section presents a comprehensive evaluation and analysis of the K-Means clustering results, aiming to translate numerical outputs into meaningful insights regarding taxpayer profiles, behavioral patterns, and practical implications for Customer Relationship Management (CRM) strategies. The clustering quality was first assessed using the Davies–Bouldin Index (DBI), which measures cluster separation and compactness, with smaller values indicating more well-defined and homogeneous clusters. The evaluation of different cluster configurations showed that k = 3 produced the lowest DBI value of -41.327, indicating the highest cluster separation and internal consistency among the taxpayers. Therefore, this configuration was selected as the optimal model for further analysis.

Table 6. DBI Evaluation Results

Number of Clusters (k)	DBI Value	Quality
2	-31.011	Good
3	-41.327	Very Good
4	-34.244	Fair

To provide context for interpreting the clusters, descriptive statistics were calculated for key variables. The dataset had a mean vehicle year of 2015 with a standard deviation of 5 years, an engine capacity averaging 1,500 CC with a standard deviation of 450 CC, and an average on-time payment rate of 75%. These statistics enable the translation of scaled centroid values back to actual data, allowing for a more intuitive understanding of cluster characteristics. The robustness of the clustering results was verified through statistical tests. ANOVA was applied to numerical variables, such as Vehicle Year and Engine Capacity, revealing significant differences among clusters.

Table 7. ANOVA Results for Numerical Variables

Variable	Between Groups Sum of Squares	df	Mean Square	F	Sig.
Vehicle Year	11803.797	2	5901.899	102.906	0.000
Engine Capacity (CC)	33663976.033	2	16831988.016	15.646	0.000

while Chi-Square tests on the categorical variable of payment status also indicated a statistically significant association with cluster membership.

Table 8. Chi-Square Test for Payment Status

Test	Value	df	Significance (2-sided)
Pearson Chi-Square	1795.438	2	0.000
Likelihood Ratio	2185.849	2	0.000

These results confirm that the algorithm successfully segmented taxpayers into meaningful groups that differ significantly in both numerical and categorical attributes. The cluster centroids, originally scaled between 0 and 1 for K-Means computation, were converted to actual values to facilitate interpretation. Cluster 0 consisted of compliant taxpayers with average-year vehicles and smaller engine capacities, reflecting an on-time payment rate of 77%, while Cluster 1 represented potential late payers with older vehicles and medium to large engines, sharing the same on-time payment rate of 77% but differing in vehicle characteristics. Cluster 2 contained non-compliant taxpayers, primarily late payers with newer vehicles, with an on-time payment rate of only 23% .

Table 9. Cluster Centroid Values

Cluster	Vehicle Year	Engine Capacity (CC)	Payment Status On Time	Payment Status Late	Interpretation
0	0.008	-0.362	0.577	-0.577	Compliant taxpayers with average-year vehicles and small engines

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Cluster	Vehicle Year	Engine Capacity (CC)	Payment Status On Time	Payment Status Late	Interpretation
1	-0.673	0.451	0.577	-0.577	Potential late payers with older vehicles and medium-large engines
2	0.654	0.005	-0.577	1.422	Non-compliant taxpayers, mostly late payers with newer vehicles

These centroids effectively summarize the average attributes of each cluster, serving as prototypes that capture the typical profile of taxpayers in each segment. Visualization using the RapidMiner Cluster Model Visualizer further supported the analysis. The scatter plot illustrated clear separation among clusters, with distinct regions in the multidimensional space representing differences in Vehicle Year and Payment Status, and boxplots highlighted the variations in these attributes across the three clusters. These visualizations provided intuitive and accessible interpretations of clustering results, reinforcing the statistical findings and enabling the identification of key behavioral differences between groups.

Figure 8. Scatter Plot of Clustering Results

The clustering analysis also informs practical CRM strategy development. Compliant taxpayers in Cluster 0 can be targeted with reward programs or loyalty incentives to maintain their timely payment behavior. Cluster 1 taxpayers, identified as potential late payers, may benefit from proactive monitoring and reminder communications to reduce payment delays. Non-compliant taxpayers in Cluster 2 require more direct intervention, such as active collection efforts and personalized notifications. By aligning cluster insights with CRM actions, the Samsat can allocate resources efficiently, design targeted interventions, and implement personalized communication strategies, ultimately improving overall tax compliance rates and collection efficiency. This integrated evaluation demonstrates that the K-Means algorithm effectively captures meaningful behavioral patterns in taxpayer data, and the analysis provides a data-driven foundation for decision-making, policy formulation, and the implementation of differentiated CRM strategies tailored to the specific characteristics and compliance profiles of each taxpayer segment.

3.5 CRM Strategy Formulation

The K-Means clustering method effectively divided motor vehicle taxpayers at Samsat UPTD I Prabumulih into three separate categories, facilitating the creation of focused Customer Relationship Management (CRM) approaches tailored to compliance behaviors, payment trends, and vehicle attributes. This categorization establishes a data-informed basis for developing customized service strategies that respond to the distinct requirements and obstacles faced by each taxpayer category.

Cluster 0 (Compliant Taxpayers) consists of individuals who possess relatively new vehicles, generally under five years of age, and consistently fulfill their tax obligations punctually. These individuals exemplify optimal compliance standards and necessitate approaches centered on maintaining engagement and building lasting loyalty. CRM techniques for this segment involve tailored communications delivered via diverse platforms including electronic mail, text messaging, and smartphone applications to sustain interaction and recognize their exemplary conduct. Reward initiatives can be established, providing advantages such as reduced fees for subsequent renewals, preferential treatment at Samsat facilities, faster transaction processing, or special access to online platforms. Furthermore, regular acknowledgment messages and appreciation initiatives strengthen their positive practices and enhance their dedication

to continued compliance. Through nurturing robust connections with this segment, Samsat can guarantee consistent revenue generation and develop advocates who may positively influence their social circles.

Cluster 1 (Inconsistent Taxpayers) encompasses individuals who possess vehicles with moderate engine specifications and display variable payment habits, occasionally meeting deadlines but sometimes failing to pay on time or postponing payments. These individuals need CRM approaches emphasizing awareness-building, habit adjustment, and preventive involvement. Educational programs should be created to enhance understanding regarding the significance of prompt tax fulfillment, the regulatory ramifications of delinquency, and the advantages of regular compliance with tax requirements. Automated notification systems should be deployed at calculated timeframes prior to payment deadlines via text messages, electronic correspondence, instant messaging platforms, or mobile application alerts. Performance-linked rewards can be established, including modest fee reductions or penalty exemptions for individuals who demonstrate improved payment regularity within a specified timeframe. Moreover, customized subsequent communications following missed obligations can assist in reconnecting these individuals and directing them toward compliance. The aim is to progressively guide this segment toward more reliable payment practices through encouraging measures and ongoing assistance.

Cluster 2 (Non-Compliant Taxpayers) includes individuals who possess aging vehicles, frequently exceeding ten years in age, and routinely postpone or completely disregard their tax payment responsibilities. This segment represents the most significant challenge regarding compliance and revenue generation. CRM approaches for this category must be more rigorous and intervention-oriented. Personal outreach via telephone contacts, residential visits, or face-to-face meetings at Samsat locations can assist in determining the underlying reasons for non-compliance, including economic limitations, insufficient knowledge, procedural challenges, or additional obstacles. Preventive alerts should be distributed regularly and via numerous communication channels to guarantee message receipt and attention. Educational initiatives designed for this segment should highlight the legal ramifications of payment failure, encompassing possible sanctions, vehicle registration cancellation, and limitations on vehicle operation. Adaptable payment arrangements, including phased payment options

or conditional penalty abatements, can be provided to alleviate financial pressure and promote incremental compliance. Cooperative methods involving local leaders, residential organizations, or social influence may prove beneficial in accessing and convincing this population. The ultimate goal is to restore this segment to compliance through an integrated approach of assistance, awareness, regulation enforcement, and incentive-based pathways toward regularization.

Table 8 delivers an extensive overview of the suggested CRM approaches aligned with each category, emphasizing the principal attributes of each taxpayer segment together with focused intervention methods. This table demonstrates how knowledge obtained from K-Means clustering can be successfully converted into concrete, implementable service enhancements that improve taxpayer contentment, cultivate enduring dedication, elevate compliance percentages, and ultimately advance optimized regional revenue acquisition. The varied approaches embody a contemporary, evidence-based methodology to public administration, illustrating the importance of combining machine learning methodologies with customer engagement frameworks in governmental service contexts.

By applying CRM strategies aligned with the clustering results, Samsat can enhance public service effectiveness, reduce tax arrears, and optimize regional revenue collection. This analysis demonstrates that data mining techniques not only support detailed data analysis but also provide a robust empirical basis for strategic decision-making in public service management.

Table 10. Recommended CRM Strategies for Each Cluster

Cluster	Taxpayer Characteristics	Recommended CRM Strategies
0	Compliant taxpayers, generally own new vehicles, consistently pay taxes on time	• Loyalty programs: digital certificates, priority access, exclusive app features • Personalized communication: birthday greetings, maintenance tips, tax reminders • Referral rewards: points redeemable for discounts or merchandise

Cluster	Taxpayer Characteristics	Recommended CRM Strategies
		• Premium services: bundled tax and insurance packages
1	Taxpayers with inconsistent payment patterns, typically own medium vehicles with moderate engine capacity	• Automated reminders via app and SMS before due dates • Interactive payment dashboard to monitor status and payment trends • Behavior-based incentives: reward points for consecutive on-time payments • Educational campaigns through call center and digital channels • Predictive analysis to identify and prevent potential defaults
2	Non-compliant taxpayers, usually own older vehicles, frequently delay tax payments	• Direct engagement supported by payment history data • Coordination with local authorities for monitoring and education • Educational campaigns: videos/webinars on consequences and benefits • Flexible payment plans and penalty restructuring • Proactive notifications with personalized urgency messaging • Targeted behavioral interventions to maximize compliance

3.6 Discussion

The K-Means clustering analysis at Samsat UPTD I Prabumulih successfully identified three meaningful taxpayer clusters based on vehicle age, engine capacity, and payment behavior. Statistical validation using the Davies–Bouldin Index (DBI = -41.327 for k = 3), ANOVA ($p < 0.05$), and Chi-Square tests ($p < 0.05$) confirmed that the clusters were well-separated and reliable. The negative DBI value indicates exceptionally strong separation between groups. Cluster 0 consists of compliant taxpayers with a 77% on-time payment

rate, Cluster 1 includes inconsistent taxpayers with similar payment rates but older vehicles, while Cluster 2 represents non-compliant taxpayers with only 23% timely payments, posing the greatest revenue challenge.

These results align with previous studies using K-Means for tax segmentation, such as Nur et al. [10], Asti et al. [6], and Wahyuni and Sriani [7]. However, this research advances existing work by directly connecting clustering outputs to practical CRM strategies—an aspect rarely addressed in regional Samsat studies. Unlike studies that focus solely on algorithm performance, this research integrates multiple validation metrics, strengthening the credibility of the segmentation results.

Translating centroid values into actionable CRM strategies highlights the practical contribution of this study. Cluster 0 benefits from loyalty and appreciation programs, Cluster 1 requires predictive reminders and educational interventions, while Cluster 2 needs intensive outreach and flexible payment arrangements. This tiered approach enables more efficient resource allocation and supports evidence-based public service management.

The analysis reveals several important behavioral insights. The similar payment rates of Clusters 0 and 1 despite different vehicle profiles suggest that compliance is driven by factors beyond vehicle characteristics alone. Meanwhile, the presence of newer vehicles in Cluster 2 indicates that non-compliance may be influenced by awareness gaps, administrative barriers, or behavioral factors rather than financial limitations. These findings underscore the need for adaptive and behavior-focused taxpayer engagement strategies.

This study is not without limitations. The dataset covers only one period (September 2025), limiting long-term behavioral analysis. The attributes used are primarily vehicle-related, excluding socio-economic or demographic factors. The study also relies solely on K-Means, which assumes spherical clusters, and the proposed CRM strategies have not yet been tested in real implementation.

Future research should incorporate multi-year data, richer behavioral variables, and comparisons with alternative clustering methods such as DBSCAN or hierarchical

clustering. Evaluating the real-world impact of the CRM strategies proposed here would provide valuable insights. Broader application across multiple Samsat offices would also enable cross-regional comparison and enhance generalizability.

Overall, this study demonstrates the value of integrating K-Means clustering with CRM strategy development for public service institutions. The approach offers a data-driven framework for understanding taxpayer behavior, improving compliance, and optimizing revenue management. The novelty lies in bridging machine learning insights with practical policy applications, presenting a replicable model for other government agencies seeking to modernize administrative decision-making.

4. CONCLUSION

This study successfully achieved its objectives by applying K-Means Clustering to manage motor vehicle taxpayer data in Prabumulih City and formulating targeted Customer Relationship Management (CRM) strategies. The research demonstrated how data mining techniques can transform raw administrative data into actionable intelligence, supporting technology-based improvements in public service management. The clustering results effectively grouped taxpayers based on payment behavior and vehicle characteristics, producing three meaningful segments: compliant, at-risk, and non-compliant taxpayers. These clusters provide clear behavioral patterns that enhance the understanding of taxpayer profiles and fulfill the first specific objective of identifying structured payment tendencies within the taxpayer population. Building on this segmentation, tailored CRM strategies were developed for each cluster. Loyalty and appreciation programs were proposed for compliant taxpayers, predictive and personalized reminders for at-risk taxpayers, and direct engagement through educational campaigns for non-compliant taxpayers. These translated strategies fulfill the second objective by offering differentiated approaches that can strengthen taxpayer engagement and improve overall compliance.

The research also contributes practical recommendations that support broader institutional goals, addressing the third objective. The integration of clustering insights with CRM strategies provides Samsat Prabumulih with a framework for optimizing local revenue collection, improving service quality through targeted resource allocation, and

enhancing institutional relationships with taxpayers. This demonstrates the value of evidence-based decision-making enabled by data mining and CRM integration. However, this study has several limitations. The dataset was limited to administrative and vehicle-related attributes, without incorporating behavioral, socio-economic, or geographic variables that may further influence compliance. Additionally, clustering results were derived from a single period of data, which may not fully capture seasonal variations or long-term behavioral trends. Future research should expand on these limitations by integrating additional variables, exploring temporal or longitudinal clustering to observe behavioral changes over time, and comparing different clustering algorithms such as DBSCAN or hierarchical clustering. Further studies may also evaluate the real-world implementation of CRM strategies developed in this research to measure their effectiveness in improving compliance and service quality.

ACKNOWLEDGMENT

The authors express sincere gratitude to Samsat UPTD I Prabumulih for providing access to the taxpayer data. Special thanks to colleagues and academic mentors for their valuable guidance. This study was supported by Sriwijaya University's resources and facilities. The authors acknowledge all parties who contributed to this research.

REFERENCES

[1] Samrah, "Implementasi Customer Relation Management (CRM) dalam memberikan pelayanan pada PT. Ranum Jaya Abadi Kabupaten Sidrap," Skripsi S.Sos, Prog. Studi Manajemen Dakwah, Fak. Ushuluddin Adab dan Dakwah, IAIN Parepare, Parepare, Indonesia, 2024.

[2] Suharmanto, W. S. Utami, N. Pratiwi, dan M. Faisal, "Penerapan data mining menggunakan algoritma K-Means untuk clustering perokok usia lebih dari 15 tahun," *Bulletin of Information Technology (BIT)*, vol. 4, no. 4, hal. 501-507, Des. 2023, doi: 10.47065/bit.v3i1.1067.

[3] M. R. P. Pratama, M. I. Fieldi, M. S. Albani, M. A. Fachrozi, F. R. Aderiyana, K. D. Tania, dan A. Meiriza, "Perbandingan algoritma K-Means, K-Medoid, dan DBSCAN untuk clustering kualitas hidup Indonesia dalam perspektif Knowledge Management dan Data Discovery," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 4, hal. 5903-

5910, Agu. 2025.

[4] L. M. Rumaropen dan F. Mahananto, "Evaluation to implementation of Customer Relationship Management in scope of social media to increase interest customer," *JIMU (Jurnal Ilmu Manajemen Universitas Bhayangkara Jakarta Raya)*, vol. 7, no. 1, hal. 37-47, Apr. 2025, doi: 10.31599/jmu.v5i1.

[5] K. A. Yanuar dan H. Firmansyah, "Penerapan algoritma K-Means untuk clustering pemodelan pengetahuan pengguna menggunakan RapidMiner," *Jurnal Komputer dan Informatika*, vol. 3, no. 1, hal. 26-32, Mar. 2025.

[6] D. Asti, M. S. Hasibuan, dan P. A. Siregar, "Penerapan algoritma K-Means untuk mengetahui tingkat kepatuhan wajib pajak kendaraan bermotor pada UPT Samsat Medan Selatan," *Journal of Computer Science and Informatics Engineering (CoSIE)*, vol. 2, no. 4, hal. 190-198, Okt. 2023, doi: 10.31599/cosie.v2i4.

[7] S. Wahyuni dan Sriani, "Penerapan algoritma K-Means untuk pengelompokan kepatuhan wajib pajak bumi dan bangunan di Kota Medan," *CESS (Journal of Computer Engineering, System and Science)*, vol. 10, no. 1, hal. 325-334, Jan. 2025.

[8] E. Wiradiansya, L. Elfianty, dan J. Fredricka, "Klasterisasi data kendaraan bermotor berdasarkan tunggakan pajak pada Kantor Samsat Kabupaten Bengkulu Selatan menggunakan metode K-Means clustering," *JUKI: Jurnal Komputer dan Informatika*, vol. 6, no. 2, hal. 164-173, Nov. 2024.

[9] A. M. Nur, H. Bahtiar, dan M. A. Jannah, "Implementasi Algoritma K-Means Clustering Dalam Mengelompokkan Kepatuhan Wajib Pajak Bumi dan Bangunan Dengan Optimasi Elbow," *Infotek: Jurnal Informatika dan Teknologi*, vol. 8, no. 1, pp. 181-192, Januari 2025. DOI: 10.29408/jit.v8i1.27975.

[10] C. Wulandari, Y. Ansori, dan K. F. H.H., "CRISP-DM Method On Indonesian Micro Industries (UMKM) Using K-Means Clustering Algorithm," *MATICS: Jurnal Ilmu Komputer dan Teknologi Informasi*, vol. 14, no. 2, pp. 35-40, September 2022.

[11] R. H. Khan, D. F. Dofadar, dan Md. G. R. Alam, "Explainable Customer Segmentation Using K-means Clustering," in *Proc. IEEE UEMCON*, 2021. DOI: 10.1109/UEMCON53757.2021.9666609.

[12] A. Ridwan, S. Setiadi, dan R. Maulana, "Optimization of Product Placement on E-commerce Platforms with K-Means Clustering to Improve User Experience," *International Journal Software Engineering and Computer Science (IJSECS)*, vol. 4, no. 1, pp. 133-147, April 2024. DOI: 10.35870/ijsecs.v4i1.2328.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[13] K. Gopalakrishnan, "Customer Segmentation Using K-Means Clustering for Targeted Marketing in Banking," *International Journal of Artificial Intelligence & Machine Learning (IJAIML)*, vol. 3, no. 2, pp. 89-94, Juli-Desember 2024. DOI: 10.5281/zenodo.13627403.

[14] F. R. Sucahyo, I. H. Santi, M. F. Rahmat, dan D. Fahrizal, "Comparing K-Means and K-medoids algorithms for clustering hamlet regions by tax liabilities in tax determination documents," *International Journal of Science and Technology Research Archive*, vol. 8, no. 1, pp. 69-78, Maret 2025. DOI: 10.53771/ijstra.2025.8.1.0023.

[15] S. B. Syahputro, T. Chairunnisya, F. Apriyanti, J. Akbar, dan H. Marpaung, "Penerapan Customer Relationship Management (CRM) Upaya untuk Meningkatkan Loyalitas Pelanggan," *Jurnal Ekonomi Manajemen Dan Bisnis*, vol. 1, no. 2, pp. 147-151, November 2023.

[16] S. Yadav, "Creating Customer Value: A Comprehensive Analysis of Contemporary CRM Strategies," *International Journal of Research Publication and Reviews*, vol. 5, no. 8, pp. 3525-3529, Agustus 2024.

[17] S. Febriansyah dan S. Wahyuni, "Pengaruh Kebijakan e-Samsat, Tax Compliance Cost, Kualitas Pelayanan, dan Sanksi Perpajakan Terhadap Kepatuhan Wajib Pajak Dinas Samsat Kabupaten Pidie," *Jurnal Akuntansi dan Keuangan (JAK)*, vol. 11, no. 2, pp. 101-110, 2023. DOI: 10.29103/jak.v11i2.8729.

[18] N. Rizkiani, "The Effect of Taxpayer Awareness, Quality of Service, and Tax Penalties on Taxpayer Compliance at Samsat Bersama Office in the East Jakarta," *International Journal of Multidisciplinary Research and Literature (IJOMRAL)*, vol. 1, no. 2, pp. 127-137, Maret 2022.

[19] S. Prayitna dan B. Witono, "Pengaruh Sistem Samsat Drive Thru, Kesadaran Wajib Pajak, Sanksi Pajak, Pengetahuan Perpajakan Dan Akuntabilitas Pelayanan Publik Terhadap Kepatuhan Wajib Pajak Dalam Membayar Pajak Kendaraan Bermotor (Studi Pada Wajib pajak SAMSAT Kota Surakarta)," *IKRAITH-EKONOMIKA*, vol. 5, no. 1, pp. 134-141, Maret 2022.

[20] A. E. P. Benggu dan T. W. Damayanti, "Pengaruh penerimaan penggunaan e-samsat terhadap kepatuhan perpajakan di Kota Kupang dengan technology acceptance model (TAM)," *Entrepreneurship Bisnis Manajemen Akuntansi (E-BISMA)*, vol. 5, no. 2, pp. 239-256, 2022. DOI: 10.37631/ebisma.v5i2.1107.

[21] A. Susilawati, A. S. M. Al Obaidi, A. Abduh, F. S. Irwansyah, dan A. B. D. Nandiyanto, "How to do research methodology: From Literature Review, Bibliometric, Step-by-step Research Stages, to Practical Examples in Science and Engineering Education," *Indonesian Journal of Science & Technology*, vol. 10, no. 1, pp. 1-40, April 2025. DOI: 10.17509/ijost.v10i1.78637.

[22] H. Lakhlij, "Rethinking the Data Gathering Techniques of Qualitative Methods for Social Sciences Research," 2024.

[23] G. Daruhadi dan P. Sopiati, "Research Data Collection," *International Journal of Social Service and Research (IJSSR)*, vol. 4, no. 7, pp. 1-18, Juli 2024.

[24] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2022.

[25] P. Tan, M. Steinbach, A. Karpatne, and V. Kumar, Introduction to Data Mining, 2nd ed. Harlow, UK: Pearson, 2022.

[26] C. C. Aggarwal, Data Mining: The Textbook, 2nd ed. Cham, Switzerland: Springer, 2022, doi: 10.1007/978-3-319-14142-8.

[27] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2023, doi: 10.1007/978-0-387-84858-7.

[28] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning with Applications in Python. Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-38747-0.

[29] S. Raschka and V. Mirjalili, Machine Learning with PyTorch and Scikit-Learn: Develop Machine Learning and Deep Learning Models with Python. Birmingham, UK: Packt Publishing, 2022.

[30] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.

[31] Q.-V. Doan, T. Amagasa, T.-H. Pham, T. Sato, F. Chen, and H. Kusaka, "Structural k-means (S k-means) and clustering uncertainty evaluation framework (CUEF) for mining climate data," *Geosci. Model Dev.*, vol. 16, no. 8, pp. 2215–2233, Apr. 2023, doi: 10.5194/gmd-16-2215-2023.

[32] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: A comprehensive review," *Heliyon*, vol. 11, hlm. e41953, 2025. DOI: 10.1016/j.heliyon.2025.e41953.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

[33] Y. Sheng, R. Bond, R. Jaiswal, J. Dinsmore, and J. Doyle, "Augmenting K-Means Clustering With Qualitative Data to Discover the Engagement Patterns of Older Adults With Multimorbidity When Using Digital Health Technologies: Proof-of-Concept Trial," *J. Med. Internet Res.*, vol. 26, hlm. e46287, Mar. 2024. DOI: 10.2196/46287.