ISSN 2723-6471

Analysis of Product Recommendation Models at Each Fixed Broadband Sales Location Using K-Means, DBSCAN, Hierarchical Clustering, SVM, RF, and ANN

Nurvita Trianasari[1,*], Thifan Anjar Permadi[2]

[1,2] *School of Economics and Business, Telkom University, Bandung, Indonesia*

(Received: February 11, 2024; Revised: March 20, 2024; Accepted: April 21, 2024; Available online: May 31, 2024)

Abstract

The telecommunications industry proliferates in the digitalization era, especially Fixed Broadband services. Fast and stable internet access is essential, especially at sales locations with appropriate products. This research aims to develop an optimal product recommendation model for each sales location, using machine learning with a mixed method approach, with a combination method of clustering and classification, where the clustering method is used for the geographic segmentation stage. Then, the results of each cluster from the geographic segmentation are used as input for the classification method, which is a stage called sales forecasting. Next, the performance analysis measured the accuracy level of each combination of models. The best model combines clustering and classification models, which, on average, across all clusters, gives the best accuracy value. The data used in this research is GIS-based POI data and sales history data, which is internal data from a telecommunications company in Indonesia. From the tests carried out in this research, the best model combination is the K-Means and the Random Forest models, with an accuracy value of 82.08%. Meanwhile, the lowest performance resulted from a combination of the K-Means and ANN models with an accuracy value of 79.50%. With an average combination model performance above 80%, this research shows that using mixed methods with clustering and classification can provide valuable insights in subsequent research, especially in the context of the telecommunications industry, especially in fixed broadband services.

Keywords: Fixed Broadband, Product Recommendation, Geographic Segmentation, Sales Forecasting, Mixed-Method

1. Introduction

The telecommunications industry, especially Fixed Broadband services, has been one of the sectors that has experienced rapid growth in recent years. With digitalization sweeping the world, fast and stable internet access has become a basic need for individuals and businesses. Fixed broadband services, which currently use fiber optic technology, are usually called Fiber to the x (FTTx) depending on the target market, such as Fiber to the Home (FTTH) which targets the residential and household segments, Fiber to the Building (FTTB) which targeting the apartment and other building segments, and Fiber to the Enterprise (FTTE) which focuses on the enterprise segment. Fixed broadband services have become the leading solution to meet the need for reliable internet connectivity. In Indonesia, the government has set a vision to provide equal internet access throughout the country. The need for reliable internet connectivity aligns with the national development vision to improve community welfare and advance the digital economy. Equitable internet access to remote and rural areas is one of the priorities in efforts to realize this vision. To achieve this goal, penetration of fixed broadband services, especially in FTTH, is very important.

The high demand for fixed broadband services, accelerated by the COVID-19 pandemic, is one of the triggers for work from home which can be used as a flexible work option in exceptional cases [1], has encouraged telecommunications companies to develop more effective marketing strategies to meet this demand. Increasing competition in the telecommunications industry, especially in fixed broadband services, encourages companies to look for effective marketing strategies and maintain high sales performance to achieve sales and financial targets. In winning the competition and as technology develops, digital marketing is a strategy that is quite effective in increasing sales performance [2], with the adaptation of digital technology [3], one of which is a strategy based on big data analytics

***Corresponding author: Thifan Anjar Permadi (thifananjar@student.telkomuniversity.ac.id)**
DOI: https://doi.org/10.47738/jads.v5i2.210

which has emerged as a driving factor for change in the business world with its use in increasing accuracy in decision making and improving the performance of sustainable Industry 4.0 applications [4].

Concerning efforts to improve sales performance, many previous studies have used big data analytics as research, such as [5], which revealed a positive perception between big data analytics and sales performance. One application of big data analytics in a digital marketing strategy is sales forecasting, predicting future sales based on historical data [6]. Research related to sales forecasting, which aims to provide product recommendations, has been carried out in the e-commerce industry [7] and retail store industry [8], and has positive impact to improve customer choice [9] and sales performance [10]. However, research on product recommendations focusing on fixed broadband services still needs to grow. Research related to fixed broadband services is usually more oriented to the technology used in the service rather than to business and marketing aspects.

Therefore, this research aimed to fill this gap by providing new insights into the research that focuses on fixed broadband services, which are services in the telecommunications industry that develop an appropriate perspective model to provide practical product recommendations for each sales location using internal data from a leading telecommunication company in Indonesia which contained two primary datasets: GIS-Based POI and Sales History based on machine learning with combining clustering and classification. Clustering is used for geographic segmentation with models, namely K-Means, DBSCAN, and Hierarchical Clustering, and classification is used for sales forecasting based on product recommendation using models, namely SVM, RF, and ANN. This research also aims to measure the performance of the models developed at each sales location and determine the most optimal and appropriate combination of models to provide product recommendations at each sales location for fixed broadband services. Thus, the main question of this research is how to select, measure, and evaluate the most suitable model for providing product recommendations at each fixed broadband sales location using GIS-based POI data and sales history data from a leading telecommunication company in Indonesia.

The paper follows this structure: Section 1 presents a brief background, research gap, and research question. Section 2 reviews the literature on geographic segmentation as part of geomarketing, sales forecasting, and a mixed-method analysis approach. Section 3 explains the research method. Section 4 provides results and discussion. The final section presents the conclusion, implications, and future research.

2. Literature Review

Identifying the components that influence this is necessary to determine the appropriate method and model for providing product recommendations at each sales location for fixed broadband services. The main components in this research are geographic segmentation and sales forecasting, which refers to the primary factors influencing sales success in fixed broadband services: sales location factors and the type of product offered to customers. Geographic segmentation is used to identify sales locations based on their characteristics, while sales forecasting provides predictions about the most suitable products to be marketed in each location.

2.1. Geomarketing and Geographic Segmentation

Geography knowledge plays an essential role in every organization and field of study. The use of geography in marketing is known as geomarketing [11]. Some researchers define it as a specific application of spatial-based marketing [12]. In contrast, other researchers provide a more detailed explanation, where geomarketing is a series of methods that manipulate geographic-based data, which focuses more on analysis rather than strategy making or decision making [13] or a marketing approach that assumes that individuals living in the same area have similar socio-demographic, economic, and cultural characteristics [14]. Geomarketing can also be referred to as a marketing instrument that has the potential to help decision-makers address essential issues in market analysis [15]. At its core, geomarketing combines geographic information with technology in various marketing elements, utilizing geographic data in the marketing research process.

In today's global market, geomarketing has become very important because taking advantage of opportunities, such as finding untapped market gaps and being one step ahead of competitors, is essential. Even though internet technology has changed how geography is viewed in marketing, attention to each sales location remains crucial in line with market growth. Consumers can still be identified based on their geographic location, which explains why some internet-based

companies experience problems delivering to their customers [13]. Geomarketing in business requires GIS. GIS is an abbreviation of Geographical Information System, which is a computer platform created to manage information about geographic areas [16] in the form of a system that is a combination of software, data, and hardware to collect, manage, and display spatial information to help solve development and management problems complex one [17].

GIS can map information related to various sales locations, including demographic data, infrastructure, and market share, based on geographical aspects. Some previous studies have adopted GIS to understand markets and develop more effective marketing strategies. For example, GIS is a decision-support tool for tourism planning and marketing [18]. From this system, users can easily visualize and analyze the spatial distribution of visitors along with their demographic information. This function can help develop or strengthen a business plan or marketing strategy. Another example is using GIS as a decision tool to analyze the spatial distribution of the soft drinks industry in Kenya [19]. This study develops a multiple regression model to predict sales by considering sales figures from certain distribution outlets and demographic and socio-economic characteristics.

In geomarketing, the data needs to be categorized into groups before analyzing location data. This process is known as market segmentation. Market segmentation can be done in several ways, such as geographic, demographic, psychographic, and behavioral. Geographic segmentation is essential and may be considered the first step towards international marketing, followed by demographic and psychographic segmentation. The geo-cluster approach combines demographic and geographic data to create more accurate or specific profiles [20]. In geographic segmentation, data is grouped or categorized based on geographic criteria.

Along with the development of digital technology, the clustering method approach, part of machine learning, is commonly used to carry out segmentation, including geographic segmentation. By utilizing this technique, more detailed and accurate market segments can be identified, which in turn can increase understanding of consumer preferences and assist in developing more thoughtful marketing strategies.

2.2. Sales Forecasting

Previous research related to sales forecasting has been carried out in various industrial sectors such as automotive [21], food products [22], fashion [23], electronics [24], clothing [25], and others. However, previous research on sales forecasting in the telecommunications industry, especially fixed broadband services, is challenging to obtain. However, techniques and research methods from other industries can be adapted for this research. In food products, research [22] uses artificial neural network models and evolutionary computing in its research related to sales forecasting. In the e-commerce industry, research [7] uses principal component analysis (PCA) and K-means clustering to build a product recommendation model based on sales forecasting.

Based on a literature review of previous research, the methods used in research related to forecasting methods, including sales forecasting, can be generally classified into two main categories: AI-based and non-AI-based models. AI-based models, such as artificial neural networks, machine learning, and support vector machines (SVM), utilize advanced algorithms to model and predict data patterns by processing large and complex amounts of data. On the other hand, non-AI-based models such as linear regression, time series analysis, and ARIMA models rely on more traditional statistical approaches to analyze historical trends and make predictions. Non-AI-based models are more straightforward to implement and are often quite effective for data with more stable and predictable patterns but are not suitable for use in large datasets and non-linear data. Meanwhile, AI-based models can analyze complex data without understanding the relationship between input and output variables from the start, so they are very suitable for implementing large datasets. However, they are often criticized for needing a theoretical basis and more understandable interpretations. However, they are still popular because they can improve prediction accuracy [26].

Considering that this research involves a non-linear dataset, and the main objective is to make accurate predictions in sales forecasting, choosing an AI-based model that uses classification techniques is a strategic implementation choice. This technique was chosen because of its ability to process and analyze complex data to produce prediction results with reasonably good accuracy. Apart from that, AI in business creates excellent opportunities in the marketing field because as it develops, artificial intelligence makes identification easier [27].

2.3. Mixed-Method Approaches

Most previous research on sales forecasting uses all existing data to create prediction models without considering how well the training dataset matches the testing dataset. Using all data can reduce prediction accuracy because the training dataset may contain too much information that could be more useful for the testing dataset, leading to errors during model training [28]. To overcome this and obtain good prediction results without spending much computing time, some recent studies propose using clustering algorithms to divide the prediction data into groups with similar characteristics before building a prediction model. For example, research [29] combines the clustering method with K-Means Clustering and Support Vector Regression. It provides insight that a hybrid sales forecasting scheme (using a combination of clustering and classification methods) is effective for use in research related to sales forecasting. Based on this, this research will adopt a mixed-method approach with clustering and classification methods using AI-based models.

This research uses K-Means Clustering, DBScan, and Hierarchical Clustering for the clustering method. K-Means Clustering has the advantage of ease of implementation and scalability, making it possible to group data into rounded and separate groups. Meanwhile, DBScan can overcome outliers and makes it possible to find irregular or irregularly shaped groups. On the other hand, Hierarchical Clustering has the advantage of building a hierarchical structure in the data, which makes it possible to understand the relationships between groups in a multilevel way. These three models provide different perspectives regarding the form of clusters formed so that they can provide variance analysis in determining the best model.

In the classification method, this research uses several models that have their respective advantages, namely SVM, RF, and ANN. SVM is known for its ability to handle high-dimensional datasets and can handle both linear and non-linear data. RF has the advantage of dealing with overfitting problems and can provide stable accuracy estimates by combining the results of several decision trees. Meanwhile, ANN is known for its ability to handle high-complexity problems and can learn from poorly structured data to provide more accurate and robust classification results by utilizing the advantages of each of these models.

3. Method

Figure 1 depicts the research framework of this study, which consists of four main steps: Data Collection, Data Preparation, Model Development, and Model Evaluation. Each stage is a part integral to the research process, which aims to investigate and analyze the phenomenon under study comprehensively.

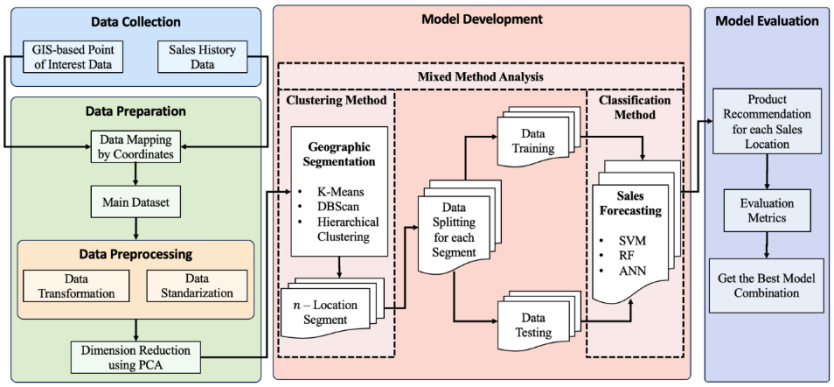


Figure 1. The research framework

3.1. Data Collection

This research utilizes two main types of data: GIS-based Point of Interest (POI) data and sales history data over three months. GIS-based POI data are directly downloaded from an internal server of a leading telecommunications company in Indonesia. POI refers to grids with an area of 250 m2, which include various information, such as the availability of telecommunications networks, population size, number of households, etc. This data was taken on November 1, 2023, and consists of 280,292 rows, each representing one POI grid. In addition, historical sales data was obtained from the

 ISSN 2723-6471

same company and directly downloaded from the internal sales performance dashboard. This sales data was collected for three consecutive months, November 2023 to January 2024, containing 94,914 rows representing one sales transaction in each row. To avoid potential biases and the generalizability of the findings, the geographic scope of the data used in this research is limited to encompassing 15 provinces and 197 districts/cities in Indonesia from four big island, such as Kalimantan (Borneo Island), Sulawesi (Celebes Island), Maluku (Moluccas Island), and Papua, with details shown in table 1.

GIS-based POI data is dynamic data that provides information on geographic conditions at a particular time. It will change over time because it is influenced by sales activity and churn. For example, GIS-based POI data taken on November 1, 2023, will, of course, have different values from data taken on December 1, 2023, because, during that month, there was sales activity or churn, which affected the value of port availability, demographic conditions, and market share. Therefore, GIS-based POI data was only taken once in this study. Next, sales history data is collected for three months, representing sales results from November 2023 to January 2024, so that sales results can be observed for three months at each location based on geographical conditions on November 1, 2023.

Table 1. Limitation of geographic area in the datasets

Island	Province	Cities/Districts	Sub Districts	Grid-POIs
Kalimantan (Borneo Island)	Central Kalimantan	14	130	17,365
	North Kalimantan	5	43	5,194
	South Kalimantan	13	148	42,394
	West Kalimantan	24	272	63,171
Maluku (Moluccas Island)	Maluku	11	82	11,609
	North Maluku	10	110	7,731
Papua	Papua	26	124	10,999
	West Papua	13	125	7,123
Sulawesi (Celebes Island)	Central Sulawesi	13	162	18,095
	Gorontalo	6	74	7,243
	North Sulawesi	15	166	14,346
	South Sulawesi	24	289	48,501
	Southesast Sulawesi	17	202	20,130
	West Sulawesi	6	65	6,390

3.2. Data Preparation

The initial process in data preparation is mapping the two types of data used, namely GIS-based POI data and sales history data, based on geographic coordinates, namely latitude and longitude. This mapping process aims to identify the dominant product types sold in each POI grid based on recorded sales history. Hence, each POI grid will receive an additional attribute called "most_package," which will be assigned a value of "1" if the high-value product dominates sales in that grid and a value of "0" if the low-value product dominates. This new attribute will become a predicted variable, while other attributes in the GIS-based POI data will become predictor variables used in the subsequent analysis process. The data resulting from this mapping then becomes the primary dataset. Table 2 is a breakdown of the variables or data attributes from the primary dataset.

Table 2. Dataset Attributes

Data Source	Attributes			
	Attribute Category	***Attribute Name***	***Function***	***Description***
	Demography	Population	Predictors	Number of Population within a grid

Data Source	Attributes			
	Attribute Category	***Attribute Name***	***Function***	***Description***
GIS-Based POI Data		Household	Predictors	Number of Households within a grid
		Potential Household	Predictors	The number of households can potentially be subscribed within the grid
		Potential Household Layer	Predictors	The level of households is potentially subscribing within the grid. (High, Mid, Low)
	Market Share	Market Share	Predictors	Market share levels within a grid
		Market Share Layer	Predictors	Market share characteristics (high, mid, low)
		GTM Profile	Predictors	Go to Market Recommendation (regular, aggressive, no competitors)
	Network	Ports Available	Predictors	Network profiles
		Ports Used	Predictors	
		Reserved Ports	Predictors	
		Broken Port	Predictors	
		Total Ports	Predictors	
		ODP Red	Predictors	
		ODP Yellow	Predictors	
		ODP Green	Predictors	
		ODP Black	Predictors	
		ODP Orange	Predictors	
		Total ODP	Predictors	
Sales History Data	Products	Most Packages	Predicted	Most packages are sold in a grid based on sales history. ("1" for High Value Package and "0" for Low Value Package)

After the primary dataset is formed, the next step is to carry out data preprocessing, which includes a series of essential steps to prepare the data. One of the main processes in data preprocessing is Exploratory Data Analysis (EDA), which aims to examine and understand the characteristics of existing data. This process includes identifying and handling missing values, cleaning the data from outliers that may be disturbing, selecting the most relevant features for analysis, and normalizing the data distribution to better match the assumptions of the model. This process may also include steps such as data transformation and standardization, which are necessary to ensure data consistency and accuracy. Appropriate data preprocessing can increase the potential for the dataset used to be ready for use in creating accurate and reliable models in subsequent analysis.

Due to the large number of attributes or variables in the dataset, the first step before entering the dataset into the model development stage, especially in the context of the clustering method, is to carry out dimension reduction using the Principal Component Analysis (PCA) method. The main goal of PCA is to reduce the dimensionality of attributes to two main variables that represent most of the variation in the dataset, allowing a more straightforward but informative representation of the available data. By using only these two variables, the complexity of the model can be reduced without losing essential information in the original dataset. This step allows for a more efficient and effective analysis process in further development of the clustering model. The algorithm's complexity can be reduced by applying PCA with dimension reduction carried out on a large scale.

3.3. Model Development

To explore and analyze data with a comprehensive approach, this research applies a mixed method that combines clustering and classification techniques. In general, the stages in model development begin with the implementation of the geographic segmentation stage using the clustering method using a dataset that has been processed through the dimension reduction stage using PCA. This step is carried out to understand the spatial and geographic patterns

underlying the data. The clustering methods used, namely K-Means, DBScan, and Hierarchical Clustering, are used to form groups based on interconnected geographical characteristics. These models are selected because they have different prespective methods for the clustering processes. K-means, DBSCAN, and hierarchical clustering are popular clustering algorithms with their own strengths and weaknesses. K-means is simple and scalable but requires the number of clusters to be specified manually and can be sensitive to initial values and the curse of dimensionality. DBSCAN can handle clusters of arbitrary shapes and densities but is sensitive to the choice of Eps and MinPts parameters and has high computational cost for large data sets. Hierarchical clustering does not require the number of clusters to be specified and can handle clusters of varying densities, but it has a lower efficiency and is not as effective for large data sets.

The next step is applying the classification method to each cluster to forecast sales. However, before the classification process begins, the dataset must go through the data splitting stage, where the data is divided into two parts, namely training data to train the model and testing data to test the model's performance. In the classification method, this research uses the SVM, RF, and ANN models. SVM separates data into two classes by finding the optimal hyperplane that maximizes the distance between the classes. RF combines the results of multiple decision trees to minimize overfitting and improve prediction accuracy.

Meanwhile, ANN is an artificial neural network model inspired by the structure and function of human neural networks, capable of handling high-complexity problems and learning from data that needs to be better structured. The aim of using these three models is to optimize sales forecasting predictions, with each model providing contributions based on the strengths and weaknesses of its function. The results of the development model will later be evaluated to determine the best model combination.

This research focuses on utilizing six models, comprising three clustering and three classification models. The selection of this specific number of models is driven by computational time considerations rather than the unsuitability of other models for the research purpose. The research aims to balance analytical complexity and computational efficiency by opting for these six models. Focussing on these models allows for thorough and meticulous analysis without sacrificing excessive computing time. Thus, emphasizing these six models is expected to yield accurate and meaningful research outcomes.

3.3.1. K-Means Clustering

A very well-known standard clustering method for clustering is K-means [30]. K-means organizes objects from the same set into several exclusive groups; for example, given a dataset D consisting of n objects and k, the number of clusters to be formed. The K-means algorithm organizes these objects into k partitions (k n), where each part represents a cluster. This method uses a centroid-based technique shown in formula (1).

$$d(y,x) = \sum_{i=1}^{D} |x^i - y^i| \quad (1)$$

With x^i is the coordinate in the th dimension i. However, K-means algorithm is very sensitive to the initial location of the cluster center, where one of the processes of object mining is partitioning an existing object into one or more clusters whose characteristics are similarly grouped in the same cluster [31].

3.3.2. Density-Based Spatial Clustering of Applications with Noise

The Density-based spatial clustering of applications with noise (DBSCAN) algorithm is a clustering model generally used in data mining and machine learning. This approach considers a group as a region containing dense or dense objects, separated by areas of low density (which represent noise). This method creates reasonably high-density clusters and identifies groups of any kind in a spatial database that includes noise. DBSCAN defines a group as the most extensive set of density-connected points. All objects not included in these groups will be considered as noise.

3.3.3. Hierarchical Clustering

Hierarchical Clustering is the recursive division of a dataset into increasingly smaller groups. The input is a weighted graph whose edge weights represent pairwise similarities or dissimilarities between data points [32]. Hierarchical Clustering is represented by a rooted tree where each leaf represents a data point, and each internal node represents a

group containing its descendant leaves. By grouping data structures into hierarchies, the Hierarchical Clustering approach can work [30].

3.3.4. Support Vector Machine

SVM is a machine learning algorithm that maps data in a high-dimensional feature space through a nonlinear mapping function. SVM classifies the training data vector ($\vec{x}_i$)into two segments ($\vec{y}_i$)represented in the formula (2).

$$G = (\overrightarrow{x_i} \in R^n; y_i = -1 \text{ or } 1; i = 1,2,..,N \quad (2)$$

According to the SVM algorithm, the first step is determining the point closest to the line separating the two classes. This point is known as the support vector. Once the points are identified, the distance between the line and the support vector is calculated, and this distance is referred to as the margin. This algorithm aims to maximize the margin, thereby producing an optimal line or hyperplane.

3.3.5. Random Forest

Random Forest or random decision forest is a classification method in ensemble learning. Random Forest uses several decision trees to determine classification. First, the training subset is randomly selected from the training dataset. Second, trees are generated randomly and trained using the training subset. The parent node divides the subset into two sub nodes, and the information impurity due to this division is formulated in formula (3).

$$\Delta g(N) = g(N) - P_L g(N_L) - P_R g(N_R) \quad (3)$$

Where g(N)is the Gini impurity measure at node N, $P_L$is the population proportion of the left child node N_L and P_R is the proportion of the right child node N_R. Next, each tree predicts the training dataset, and the prediction results produced by all the trees are averaged to obtain the final output of sales result predictions. The final output of the Random Forest is shown in formula (4).

$$\hat{y} = \frac{1}{N_{trees}} \sum_{i=1}^{N_{trees}} y_i \quad (4)$$

Where $y$is the final output, N_{trees} is the number of trees and $y_i$is the result of each tree.

3.3.6. Artificial Neural Networks

Prediction performance using the ANN method is proper when dealing with non-linear systems with high dimensions [33]. This research uses a feed-forward neural network method, which consists of one or more input layers, one or more hidden layers, and one output layer, where each neuron in one layer conveys information to all neurons in the next layer. In this research, the ANN model consists of one input layer with the number of neurons adjusted to the number of data attributes or predictor variables, and one output layer representing the dependent variable, namely product recommendations (Y). The output of hidden neurons (V_L) and product recommendations (Y) are formulated in formula (5).

$$V_1 = \sum_{i=1}^{3} h(w_{1i}x_i + b_1) \quad (5)$$

$$V_L = \sum_{i=1}^{3} h(w_{Li}x_i + L) \quad (6)$$

$$Y = \sum_{i=1}^{L} h(w_j x_j + \beta) \quad (7)$$

Where w_{Li} is the input weight, x_i is the input neuron, b_L is the hidden layer threshold, w_L is the output weight, V_L is the output of the hidden neuron, β is the output layer threshold, h(x) is the activation function and Y is the output neuron (product recommendation)

3.4. Model Evaluation

A thorough evaluation of model performance is necessary to ensure optimal results. This research uses a mixed approach that combines clustering and classification models to improve analysis. The process starts with clustering to divide geographic data into groups. The results then serve as the primary input in the classification model for sales forecasting. This model aims to provide specific product recommendations for each fixed broadband service sales location so that marketing strategies can be more targeted.

It is necessary to search for the best model combination by thoroughly evaluating the performance of all models involved to ensure a high level of precision and accuracy in product recommendations. This evaluation considers the best accuracy of each model combination in each data group. The accuracy results of each cluster will be mapped based on the clustering and classification models so that analysis can be easily carried out.

The final step is calculating the average accuracy of all clusters in each clustering model and provides deeper insight into how the model performs across different geographic data sets. From the results of this analysis, a combination of clustering and classification models that provide the highest accuracy will be identified to determine the most effective model for providing product recommendations at each fixed broadband service sales location.

4. Result and Discussion

All analysis and testing in this research used the Jupyter Notebook application based on the Python 3.0 programming language. So, the research process uses modules available in the Python library itself. Data analysis and visualization programs can help achieve a more profound understanding; one uses an application based on the Python programming language, with English commands and easy-to-follow syntax, offering a free and open-source alternative to traditional techniques and applications [34].

The results of this research are a series of steps obtained from the results at each stage, so an analysis of the results at each stage is required. The first stage of this research began by mapping GIS-based POI data consisting of 280,292 lines and sales history data with a total of 94,914 lines. The mapping results found that 14.33% of the entire POI grid contained sales, which then became the primary dataset with 40,155 rows, while the other part was deleted to maintain the research focus on the POI grid with sales transactions. Next, this mapping process is the exploratory data analysis stage, which aims to understand the characteristics of the data in more depth. At the EDA stage, a data cleaning process is carried out to remove rows that have null values, followed by distribution normalization and scale standardization to reduce outliers. This process produces a cleaner dataset ready for use in the following analysis stage, as shown in figure 2.

From the EDA stage, 22.14% of the data, or 8,889 POI rows, were clean for use in the next stage, namely principal component analysis to reduce the dimensions of the data while retaining significant information. Data with dimensionally reduced will then be used in clustering and classification models but must go through data standardization to obtain a uniform scale. Thus, this stage is the first step in understanding data patterns and preparing them for further analysis. These steps are critical in maintaining the quality and accuracy of the analysis that will be carried out in subsequent stages of this research.

4.1. Analysis of Clustering Results

The initial analysis carried out focuses on examining various clustering methods. This step is crucial as it ensures the creation of effective geographic location clusters, which can subsequently be utilized optimally in classification methods. To achieve the best results, parameter tuning is conducted for each clustering model. Following this, a thorough analysis is performed to identify the most optimal parameters that yield the most effective clustering outcomes. By fine-tuning these parameters, the clustering process can be enhanced, ensuring that the resulting clusters are highly precise and can be used to improve the performance of subsequent classification tasks.

4.1.1. K-Means Clustering

The k-Means Clustering model uses the Silhouette and Elbow Method to determine the optimal number of clusters. With the Silhouette approach, the highest coefficient indicates the most optimal number of clusters, whereas, with the Elbow Method approach, the line at the point closest to an elbow shape or the most significant fracture shape shows the most optimal number of clusters Figure 2 shows the graph resulting from elbow method and silhouette method to find the optimal number of clusters for k-Means clustering.

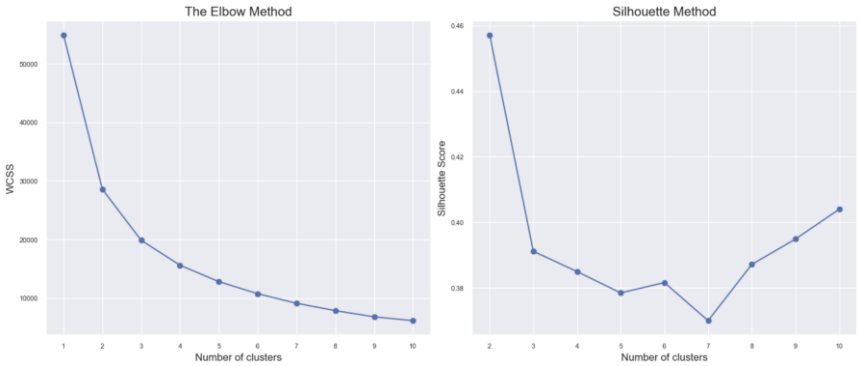


Figure 1. Elbow method and Silhouette Method results for K-Means Clustering

Based on the Silhouette test results, the most optimal number of clusters is 2, with a coefficient value of 0.457170715. Meanwhile, test results using the Elbow Method show that the line shows a fracture or is closest to an elbow at the number of clusters = 2. Figure 2 shows a graph of the Silhouette and Elbow Method results. Therefore, based on these two techniques, the optimal number of clusters produced through K-Means Clustering is 2 clusters, with the data in cluster 1 being 6,051 grids, while cluster 2 is 2,838 grids. As seen in figure 3, the cluster shape from K-Means has boundaries that tend to be straight; this is because K-Means assumes that the cluster has a well-defined geometric shape.

4.1.2. DBSCAN

Analysis of the clustering method using the DBSCAN model was carried out using tools based on the Python programming language, so the implementation was carried out using modules that were available in the library. The library used for DBSCAN clustering analysis in the Python programming language is the Sckit-learn library with the "dbscan" module. Using this module, the mandatory parameters needed are the "epsilon value" and "min_samples" parameters. The research was carried out by looking for the best combination of values between the parameters "epsilon value" and "min_samples" to obtain cluster results with appropriate density and minimal outliers.

Based on the analysis results, the optimal cluster shape with minimal outliers was obtained using the parameter values epsilon = 0.2000967021 and min_samples = 9. The results of applying these parameters produced three clusters. Figure 3 shows a visualization of the cluster results. If observed, the clusters produced from DBSCAN are divided based on density, where cluster 1, with the densest density, produces 4,973 data, and cluster 2, with medium density, produces 3,162 data. In comparison, cluster 3, with the lowest density, produces 754 data spread across two other clusters. However, several points should still be cluster 1 or 2 members but instead become cluster 3, so the cluster results can be partially perfect and still have outliers.

Figure 2. Visualization of results of clustering models

4.1.3. Hierarchical Clustering

In the analysis using Hierarchical Clustering or what can also be called Agglomerative Clustering, the optimal number of clusters can be determined using a dendrogram, a graphical visualization of the relationship between data points in one dataset. A dendrogram shows how data points are grouped based on their distance or similarity. The perspective obtained from the dendrogram can be used to select the optimal number of clusters by looking at where there are cuts in the dendrogram that provide results that best suit the data analysis needs by selecting a cut-off level in the dendrogram that produces the desired clusters.

Based on the dendrogram results, the optimal number of clusters is 2 clusters, with the amount of data in cluster 1 of 5,812 data and cluster 2 of 3,077. Figure 4 shows the dendrogram resulting from hierarchical clustering and figure 3 shows the results of the clusters formed. Based on figure 3, the cluster results with hierarchical clustering show the most appropriate cluster division visually.

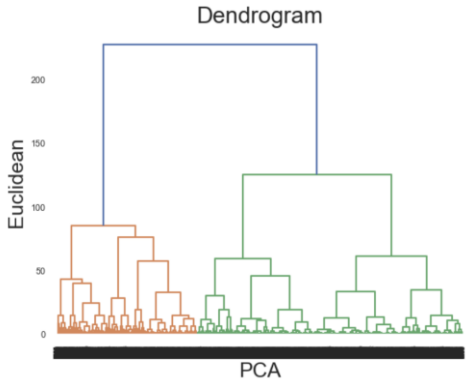


Figure 3. Dendrogram in Hierarchical Clustering

4.2. Analysis of Classification Results

At this stage, each cluster formed through the clustering process will become input data for the classification model, namely the SVM, RF, and ANN models. Then, an analysis was carried out on the accuracy of each classification model on the data in each cluster, totaling 7 clusters, with details, Kmeans = 2 Clusters, DBSCAN = 3 Clusters, and Hierarchical Clustering = 2 Clusters. This process analyzes how well each classification model can predict and classify data in each cluster. The best model combination can be determined at the evaluation stage based on the accuracy values obtained.

Analysis and testing at this classification stage use the "scikit-learn" module in the Python library, including the parameters required for each model formed. To provide maximum prediction results, research at this stage focuses on finding the best parameters through testing parameters within a specific range, called parameter tuning. In general, the test graphic results of this parameter tuning process are shown in figure 5. In contrast, the specific test results will be explained in the discussion of analysis and testing for each classification model.

4.2.1. Support Vector Machine

Based on tips and trick on the library documentation for SVM model [35], this research performs tuning on the C (Complexity) value parameter, with the RBF kernel type, to get good accuracy results in the SVM model; in SVM models, value C is a parameter that controls the trade-off between the penalty for misclassification on the training data and the maximum margin. Higher values indicate that the model will tend to account for each training data point more, which can lead to a better fit to the training data but also increase the risk of overfitting. Conversely, lower C-Values emphasize larger maximum margins, which can result in a more general model but less precision on the training data. In practice, C-Values are often adjusted through a parameter-tuning process to balance model accuracy and generalization.

The SVM model was tested by running model training and varying the C value with a range between 0.1 and 4.0 with an interval of 0.1. The best value C in testing the SVM model is determined from the accuracy value obtained, which

is tested on each trained cluster. Table 3 shows the various best C-Values from each cluster, and a graph between the variance of the C-Values and the accuracy results is shown in figure 5.

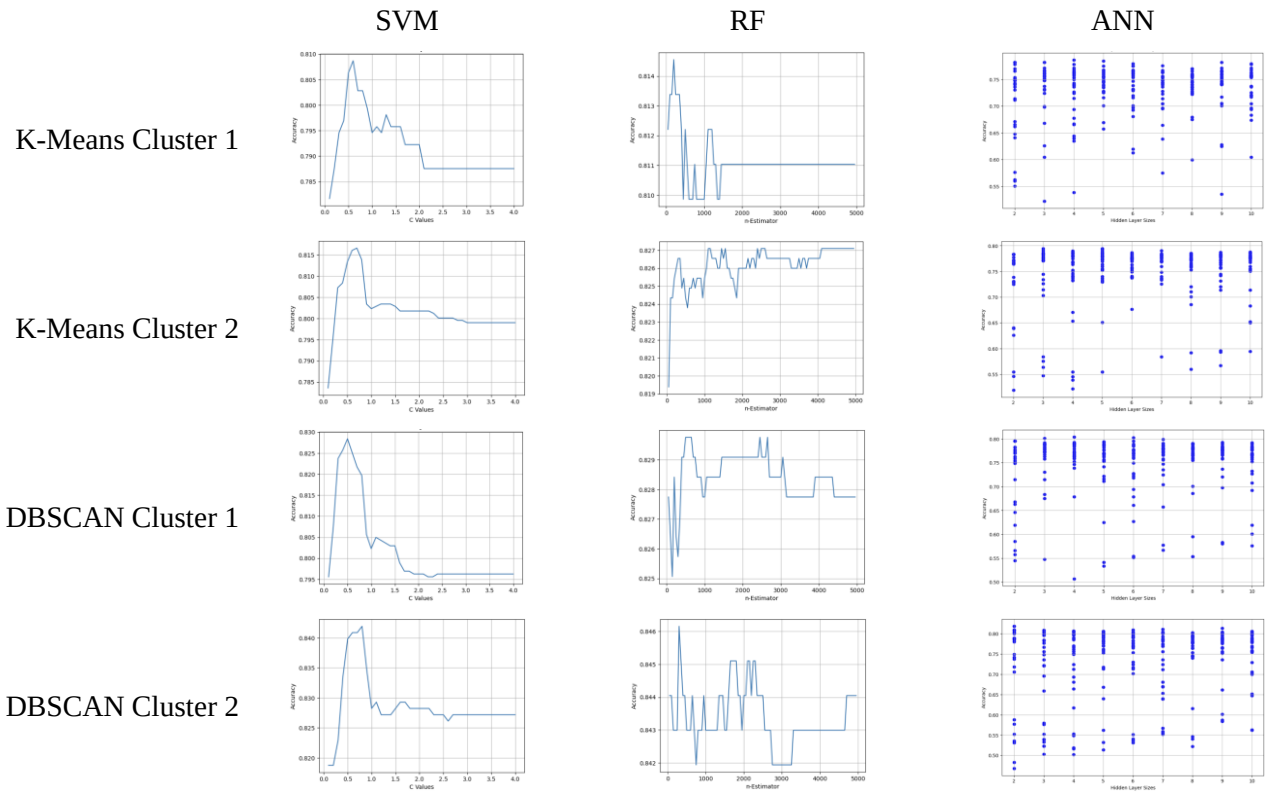
4.2.2. Random Forest

Based on the scikit-learn library documentation on the Random Forest Classifier model [36], the required parameter is the n_estimators parameter. The n_estimators parameter in the Random Forest model is the number of decision trees to build in the ensemble. The more trees that are built, the more complex the model becomes and the more potent its ability to handle data diversity or capture complex patterns. However, adding trees can also increase the training time and memory required. Typically, larger values of n_estimators produce more powerful models, but there is a point where the increase in model performance becomes significantly reduced. On the other hand, a n_estimator value that is too small can cause the model to be less powerful and tend to overfit the training data.

Therefore, in the Random Forest model, testing is done by tuning the n_estimators parameters and analyzing the best parameter values based on the accuracy obtained. The n_estimator parameter tested is in the range of values from 50 to 5000 with an interval of 50. The best value n_estimators in testing this RF model is determined from the accuracy value obtained, which is tested on each trained cluster. Table 3 shows the various n_estimators best values from each cluster, and a graph between the value variances n_estimators and the accuracy results are shown in figure 5.

4.2.3. Artificial Neural Networks

In this research, testing the ANN model is the most complicated test compared to the two previous models. Based on sciencekit-learn documentation, Training with an ANN model using Multi-layer Perceptron classifier [37], three essential parameters need to be tested, namely hidden_layer_sizes, which determines the architecture of the ANN in terms of the number and size of hidden layers; learning_rate _init, which is the initial learning rate that regulates how much the model weights are updated during each iteration learning, and max_iter, which determines the number of iterations (epochs) to be performed during model training. Combining the three parameters plus many test clusters requires much time to run this model. Therefore, the ANN model used in this research is a feed-forward neural network type, which requires faster training time than backpropagation.



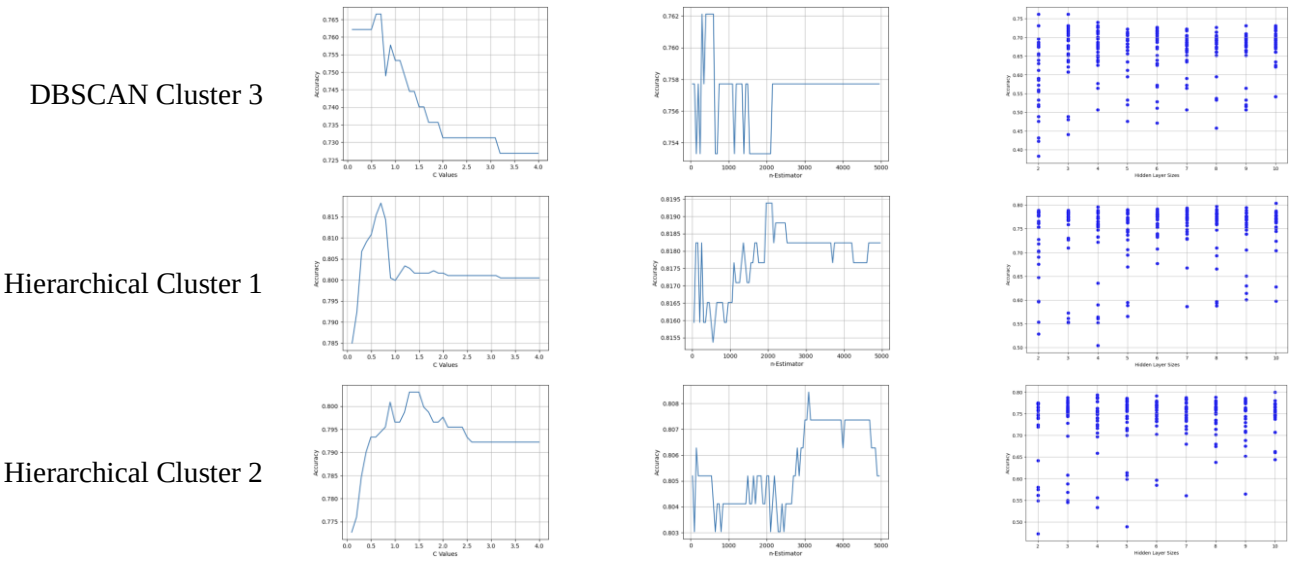


Figure 4. Graph of tuning parameters vs accuracy in the classification model

In testing the parameters used in this ANN model, the hidden_layer_sizes value was tested in the range of values 2 to 10, then the learning_rate _ init settings were tested with varying values of 0.001, 0.01, 0.1, and 0.5, and max_iter was tested with varying values of 100, 500, 1000, 5000, 10000, 50000, 100000. The best combination of values for the three parameters in testing the ANN model is determined from the accuracy values obtained, which are tested on each trained cluster. Table 3 shows the various n_estimatorbest values from each cluster, and a graph between the value variances n_estimatorthe accuracy results are shown in figure 5.

Table 3. Classification model test results

Clustering Models	Clusters	SVM (C-Value = 0.1-0.4, steps = 0.1)		Random Forest (n_estimator = 50-5000, steps = 50)		Artificial Neural Networks (hidden_layer_sizes = 2-10, learning_rate = 0.001-0.5, max_iter = 100-100000)			
		Best C-Value	*Accuracy per Cluster (%)*	*Best n-estimator*	*Accuracy per Cluster (%)*	*Best hidden_layer_size*	*Best learning_rate_init*	*Best max_iter*	*Accuracy per Cluster (%)*
K-Means Clustering	Cluster 1	0.6	80.87	200	81.45	4	0.5	500	78.64
	Cluster 2	0.7	81.66	1100	82.71	3	0.5	5000	79.46
DBSCAN	Cluster 1	0.5	82.84	500	82.97	4	0.1	1000	80.43
	Cluster 2	0.8	84.19	300	84.61	2	0.5	500	81.87
	Cluster 3	0.6	76.65	200	75.77	2	0.5	500	76.21
Hierarchical Clustering	Cluster 1	0.7	81.82	2000	81.93	10	0.5	500	80.39
	Cluster 2	1.4	80.30	3100	80.84	10	0.5	500	79.98

4.3. Evaluation

At the evaluation stage, the analysis process is carried out to check the performance of the various models used, namely the combination of clustering and classification models. This analysis not only pays attention to overall performance but also focuses on the performance of each cluster formed from the clustering model, which is 7 clusters in this case. The results of this analysis are then presented in table 4, which maps the relationship between clustering and classification models and displays the accuracy results of each combination.

 ISSN 2723-6471

Table 4. Performance evaluation on model combinations

Clustering Models	Clusters	SVM		RF		ANN	
		Accuracy per Cluster (%)	***Avg. Accuracy (%)***	***Accuracy per Cluster (%)***	***Avg. Accuracy (%)***	***Accuracy per Cluster (%)***	***Avg. Accuracy (%)***
K-Means Clustering	Cluster 1	80.87	81.26	81.45	**82.08**	78.64	*79.05*
	Cluster 2	81.66		82.71		79.46	
DBSCAN	Cluster 1	82.84	81.22	82.97	81.17	80.43	79.50
	Cluster 2	84.19		84.61		81.87	
	Cluster 3	76.65		75.77		76.21	
Hierarchical Clustering	Cluster 1	81.82	81.06	81.93	81.38	80.39	80.18
	Cluster 2	80.30		80.84		79.98	

The mapping results show that the combination of K-Means with Random Forest stands out as the best, with an accuracy of 82.08%. This number is calculated from the average accuracy value for each existing cluster. However, on the contrary, the combination of K-Means with ANN shows the lowest performance with an accuracy of 79.50%. These results illustrate that some model combinations can provide more consistent and satisfactory results than others.

This evaluation shows that the ANN model tends to provide less than optimal results because its accuracy tends to be below 80%. However, cluster 2 in the DBSCAN model stands out as an exception, with reasonable accuracy performance when combined with the rest of the classification models. With impressive performance values above 80%, around 80% of product recommendations offered on each POI grid are expected to be on target, positively impacting overall sales.

5. Conclusion

Based on this research, the method applied involves a mixture of clustering methods and three classification models, namely K-Means, DBSCAN, and Hierarchical Clustering, as well as Support Vector Machine, Random Forest, and Artificial Neural Network. Clustering is used for geographic segmentation. In the sales forecasting stage, classification is applied to provide product recommendations at each location called POI, a grid with an area of 250 m2. The test and analysis results show that model performance is measured through accuracy values after the training process at the classification stage. The evaluation was based on the combination of models and each cluster produced, including 7 clusters formed, two from K-Means, three from DBSCAN, and two from Hierarchical Clustering. Combining K-Means and Random Forest is the most effective model, with an average accuracy of 82.08%. K-Means with ANN show the lowest performance, with an average accuracy of 79.50%.

However, it should be noted that the performance of the ANN model still needs to be improved by exploring enhancing its performance through a more comprehensive parameter setting, including adjusting parameters for the MPLClassifier in scikit-learn, such as the number of layers, neurons per layer, activation functions, learning rate, and regularization techniques. These adjustments can significantly impact the model's performance and generalization ability, offering opportunities to enhance its effectiveness across various applications. Nonetheless, it is crucial to recognize that such parameter tuning may demand higher computational resources. Thus, researchers must carefully consider the computing device choice and the dataset's dimensions and features utilized.

To advance future research, researchers can expand upon the study by integrating additional models to enhance the understanding of optimal model combinations. Incorporating measurements of computational complexity and practical feasibility is vital when comparing different model combinations. This holistic approach ensures that accuracy is not the sole criterion for evaluating research insights. By considering factors such as computational efficiency and practical applicability, researchers can comprehensively analyze the effectiveness and suitability of various model combinations in real-world contexts. Moreover, future studies could explore alternative avenues such as market trends, customer

preferences, and other relevant factors better to understand the dynamics influencing decision-making processes beyond product recommendations. Additionally, segmentation can encompass various aspects such as product, pricing, and others, not limited to geographical segmentation. Therefore, these model combinations can be applied widely across different segments, offering versatile insights into decision-making processes.

With an average performance above 80%, this research shows that using mixed methods between clustering and classification can provide valuable insights in subsequent research, especially in the context of the telecommunications industry, especially in fixed broadband services. The implications of these findings for the telecommunications industry in Indonesia are expected to be significant. By leveraging mixed methods like clustering and classification, telecommunication companies can enhance their marketing strategies by providing more targeted product recommendations based on geographic segmentation, leading to increased sales performance and customer satisfaction.

Moreover, these findings align with broader industry trends and challenges. In an increasingly competitive telecommunications landscape, companies seek innovative ways to improve their marketing effectiveness and meet the evolving needs of their customers. By adopting advanced analytics techniques like clustering and classification, companies can better understand customer behavior and preferences, allowing them to tailor their offerings more effectively. Additionally, by embracing data-driven approaches, companies can stay ahead of industry trends and proactively address challenges such as market saturation and changing consumer demands. Therefore, the insights provided by this research have the potential to inform strategic decision-making and drive positive outcomes for the telecommunications industry in Indonesia.

6. Declarations

6.1. Author Contributions

Conceptualization: N.T. and W.A.A.; Methodology: W.A.A.; Software: N.T.; Validation: N.T., W.A.A.; Formal Analysis: N.T., W.A.A.; Investigation: N.T.; Resources: W.A.A.; Data Curation: W.A.A.; Writing Original Draft Preparation: N.T. and W.A.A.; Writing Review and Editing: W.A.A. and N.T.; Visualization: N.T.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

Due to the nature of the research, and due to ethical and legal reasons, the supporting data is not available.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R. Farooq and A. Sultana, “The potential impact of the COVID-19 pandemic on work from home and employee productivity,” *Measuring Business Excellence*, vol. 26, no. 3, pp. 1-12, 2022, doi: 10.1108/MBE-12-2020-0173.

[2] Tabiat, “The Impact of Digital Marketing on Sales Performance: The Case of Lebanese Pharmaceutical Companies,” *European Journal of Business and Management Research*, vol. 7, no. 4, pp. 1-9, 2022, doi: 10.24018/ejbmr.2022.7.4.1600.

[3] K. Plangger, D. Grewal, K. de Ruyter, and C. Tucker, “The future of digital technologies in marketing: A conceptual framework and an overview,” *Journal of the Academy of Marketing Science*, vol. 50, no. 6. Springer, pp. 1125–1134, Nov.

01, 2022. doi: 10.1007/s11747-022-00906-2.

[4] S. Gupta, T. Justy, S. Kamboj, A. Kumar, and E. Kristoffersen, "Big data and firm marketing performance: Findings from knowledge-based view," *Technol Forecast Soc Change*, vol. 171, no. 1, p. 120986, Oct. 2021, doi: 10.1016/j.techfore.2021.120986.

[5] M. Shahbaz, C. Gao, L. Zhai, F. Shahzad, A. Luqman, and R. Zahid, "Impact of big data analytics on sales performance in pharmaceutical organizations: The role of customer relationship management capabilities," *PLoS One*, vol. 16, no. 4 April 2021, pp. 1-7, 2021, doi: 10.1371/journal.pone.0250229.

[6] Z. L. Sun, T. M. Choi, K. F. Au, and Y. Yu, "Sales forecasting using extreme learning machine with applications in fashion retailing," *Decis Support Syst*, vol. 46, no. 1, pp. 411–419, Dec. 2008, doi: 10.1016/j.dss.2008.07.009.

[7] S. Bandyopadhyay, S. S. Thakur, and J. K. Mandal, "Product recommendation for e-commerce business by applying principal component analysis (PCA) and K-means clustering: benefit for the society," *Innov Syst Softw Eng*, vol. 17, no. 1, pp. 1-7, 2021, doi: 10.1007/s11334-020-00372-5.

[8] C. Udokwu, P. Brandtner, F. Darbanian, and T. Falatouri, "Improving Sales Prediction for Point-of-Sale Retail Using Machine Learning and Clustering," *Machine Learning and Data Analytics for Solving Business Problems*, Springer, vol. 1, no. 1, pp. 55-7, 2022. doi: 10.1007/978-3-031-18483-3_4.

[9] S. Senecal and J. Nantel, "The influence of online product recommendations on consumers' online choices," *Journal of Retailing*, vol. 80, no. 2, pp. 1-9, 2004, doi: 10.1016/j.jretai.2004.04.001.

[10] B. Pathak, R. Garfinkel, R. Gopal, R. Venkatesan, and F. Yin, "Empirical analysis of the impact of recommender systems on sales," *Journal of Management Information Systems*, vol. 27, no. 2, pp. 1-8, 2010, doi: 10.2753/MIS0742-1222270205.

[11] V. Ramadani, D. Zendeli, S. Gerguri-Rashiti, and L. P. Dana, "Impact of geomarketing and location determinants on business development and decision making," *Competitiveness Review*, vol. 28, no. 1, pp. 9-16, 2018, doi: 10.1108/CR-12-2016-0081.

[12] P. Latour and J. L. Floc'h, Géomarketing: principes, méthodes et applications. Ed. d'Organisation, 2001. [Online]. Available: https://books.google.co.id/books?id=m9HGAQAACAAJ

[13] G. Cliquet and J. Baray, Location-Based Marketing: Geomarketing and Geolocation. Wiley, 2020. [Online]. Available: https://books.google.co.id/books?id=nFbbDwAAQBAJ

[14] J. M. Lehu, L'encyclopédie du marketing. in Les Références. Eyrolles, 2012. [Online]. Available: https://books.google.co.id/books?id=EKtYRuPKT5sC

[15] C. Chasco Yrigoyen, "El Geomarketing y la Distribución Comercial," Investigación y Marketing, vol. 1, 2003.

[16] P. A. Burrough, R. A. McDonnell, and C. D. Lloyd, Principles of Geographical Information Systems. OUP Oxford, 2015. [Online]. Available: https://books.google.co.id/books?id=kvoJCAAAQBAJ

[17] G. Cliquet, Geomarketing: Methods and Strategies in Spatial Marketing. in ISTE. Wiley, 2013. [Online]. Available: https://books.google.co.id/books?id=Hg29bQB2EjwC

[18] S. K. Supak, H. A. Devine, G. L. Brothers, S. Rozier Rich, and W. Shen, "An Open Source Web-Mapping System for Tourism Planning and Marketing," *Journal of Travel and Tourism Marketing*, vol. 31, no. 7, pp. 1-13, 2014, doi: 10.1080/10548408.2014.890153.

[19] S. M. Musyoka, S. M. Mutyauvyu, J. B. K. Kiema, F. N. Karanja, and D. N. Siriba, "Market segmentation using geographic information systems (GIS): A case study of the soft drink industry in Kenya," *Marketing Intelligence and Planning*, vol. 25, no. 6, pp. 23-30, 2007, doi: 10.1108/02634500710819987.

[20] P. Kotler and K. L. Keller, MarkKotler, P., & Keller, K. L. (2022). Marketing Management. Global Edition (Vol. 16/E), vol. 16/E, no. 4. 2022.

[21] A. Sa-Ngasoongsong, S. T. S. Bukkapatnam, J. Kim, P. S. Iyer, and R. P. Suresh, "Multi-step sales forecasting in automotive industry based on structural relationship identification," *in International Journal of Production Economics*, vol. 1, no. 1, pp. 1-12, 2012. doi: 10.1016/j.ijpe.2012.07.009.

[22] P. Doganis, A. Alexandridis, P. Patrinos, and H. Sarimveis, "Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing," *J Food Eng*, vol. 75, no. 2, pp. 34-41, 2006, doi: 10.1016/j.jfoodeng.2005.03.056.

[23] T. M. Choi, C. L. Hui, N. Liu, S. F. Ng, and Y. Yu, "Fast fashion sales forecasting with limited data and time," *Decis Support Syst*, vol. 59, no. 1, pp. 1-8, 2014, doi: 10.1016/j.dss.2013.10.008.

[24] E. Hadavandi, H. Shavandi, and A. Ghanbari, "An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: Case study of printed circuit board," *Expert Syst Appl*, vol. 38, no. 8, pp. 1-7, 2011, doi: 10.1016/j.eswa.2011.01.132.

[25] S. Thomassey, "Sales forecasts in clothing industry: The key success factor of the supply chain management," *Int J Prod Econ*, vol. 128, no. 2, pp. 13-21, 2010, doi: 10.1016/j.ijpe.2010.07.018.

[26] H. Song, R. T. R. Qiu, and J. Park, "A review of research on tourism demand forecasting: Launching the Annals of Tourism Research Curated Collection on tourism demand forecasting," *Ann Tour Res*, vol. 75, no. 1, pp. 23-30, 2019, doi: 10.1016/j.annals.2018.12.001.

[27] M. I. P. Bagaskara and N. Trianasari, "The Effect Of Artificial Intelligence On Smart Customer Experience With Moderation Of Technology Readiness (Case Study On Go Food Application)," *JHSS (JOURNAL OF HUMANITIES AND SOCIAL STUDIES)*, vol. 7, no. 3., pp. 1066–1069, 2023.

[28] I. F. Chen and C. J. Lu, "Sales forecasting by combining clustering and machine-learning techniques for computer retailing," *Neural Comput Appl*, vol. 28, no. 9, pp. 23-31, 2017, doi: 10.1007/s00521-016-2215-x.

[29] C. J. Lu and C. C. Chang, "A hybrid sales forecasting scheme by combining independent component analysis with k-means clustering and support vector regression," *Scientific World Journal*, vol. 2014, no. 1, pp. 26-34, 2014, doi: 10.1155/2014/624017.

[30] A. K. Tyagi, A. Abraham, F. K. Hussain, A. Kaklauskas, and R. J. Kannan, *Machine Learning, Blockchain Technologies and Big Data Analytics for IoTs: Methods, technologies and applications*. Institution of Engineering and Technology, vol. 16, no. 1, pp. 24-30, 2022. doi: 10.1049/pbse016e.

[31] M. N. Hutasoit, R. Y. Fa'rifah, and R. Andreswari, "Application of Data Mining For Clustering Car Sales Using The K-Means Clustering Algorithm," *International Journal of Information System & Technology Akreditasi*, vol. 7, no. 2, pp. 118-124, 2023. doi: 10.30645/ijistech.v7i2.307

[32] V. Cohen-Addad, V. Kanade, F. Mallmann-Trenn, and C. Mathieu, "Hierarchical clustering: Objective functions and algorithms," *Journal of the ACM*, vol. 66, no. 4, pp. 34-40, 2019, doi: 10.1145/3321386

[33] A. Alamsyah and T. B. A. Nugroho, "Predictive modelling for startup and investor relationship based on crowdfunding platform data," *in Journal of Physics: Conference Series*, vol. 971, no. 1, pp. 13-20, 2018, p. 012002.

[34] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory data analysis using python," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 36-42, 2019, doi: 10.35940/ijitee.L3591.1081219.

[35] "1.4. Support Vector Machines," scikit-learn.org, https://scikit-learn.org/stable/modules/svm.html#id13 (accessed Mei 18, 2024).

[36] "sklearn.ensemble.RandomForestClassifier," scikit-learn.org, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed Mei 18, 2024)

[37] "sklearn.neural_network.MLPClassifier," scikit-learn.org, https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html (accessed Mei 18, 2024)