Journal of Information Systems and Informatics
Vol. 6, No. 2, June 2024 e-ISSN: 2656-4882 p-ISSN: 2656-5935
DOI: 10.51519/journalisi.v6i2.720 Published By DRPM-UBD

Predictive Analytics on Shopee for Optimizing Product Demand Prediction through K-Means Clustering and KNN Algorithm Fusion

Mesi Febima[1], Lena Magdalena[2]

[12]Information Systems Department, Catur Insan Cendekia University, Cirebon, Indonesia
Email: [1]mesi.febima@cic.ac.id, [2]lena.magdalena@cic.ac.id

Abstract

This study focuses on predictive analysis in the context of the Shopee market, aiming to optimize product demand forecasting through the combination of K-Means clustering and KNN algorithms. With the exponential growth of e-commerce platforms like Shopee, accurately predicting product demand is becoming increasingly important for inventory management and marketing strategies. In this research, we propose a novel approach that combines the strengths of K-Means clustering and the KNN algorithm to improve demand prediction accuracy. By leveraging K-Means clustering to group similar products into two clusters, namely "Low Interest" with 64 data points and "High Interest" with 25 data points, we then apply the KNN algorithm to predict demand within each cluster. The KNN algorithm produces two classifications: Low Sales and High Sales. Based on tests using the KNN algorithm with k values of 3, 5, and 7, it was demonstrated that the product "Soraya Bedsheet Cotton Gold Motif Dallas Ask Grey Tua" can be predicted to fall under "High Sales." The sales prediction accuracy rate for Shopee marketplace products is 96%. The implications of these findings indicate that the combination of K-Means and KNN algorithms can improve the accuracy of product demand predictions and optimize inventory and marketing strategies.

Keywords: K-Means, KNN, Product Demand, Sales Prediction, Shopee.

1. INTRODUCTION

A marketplace is an online application or website that bridges various stores in the buying and selling process. Marketplaces enable transactions between Customer to Customer (C2C) or consumer to consumer [1]. Some local marketplaces in Indonesia include Bukalapak, Tokopedia, Lazada, and Shopee [2]. The largest marketplace in Indonesia is Shopee. Shopee is the largest marketplace in Indonesia and a leading online shopping platform in Taiwan and Southeast Asia, tailored to each region. It offers services to customers so they can experience safe, convenient, easy, and fast online shopping through secure logistics and payments [3]. One business entity that uses Shopee as a platform to sell products is Soraya Bedsheet.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Soraya Bedsheet is an industrial company that manufactures and provides bedroom products such as sprays, bed covers, and others. However, on the Shopee marketplace, Soraya Bedsheet encounters several issues. The first issue is the inability to estimate product stock accurately. Secondly, once a product is sold out, Soraya Bedsheet struggles to update its stock. Thirdly, they are unable to view buyer contact numbers. To address these issues, sales prediction is conducted on the Shopee marketplace at Soraya Bedsheet's store. The prediction aims to obtain information and knowledge by using a Data Mining approach.

Data Mining is a process of discovering new patterns from a large dataset using various techniques such as statistics, artificial intelligence, machine learning, and database systems [4]. In the data mining process, algorithms are used, which are divided into several techniques [5]. The combination of two techniques in Data Mining proposed in this research is clustering and classification. This combination of two techniques is called Hybrid Data Mining. Hybrid Data Mining is a process that aids decision-making by combining algorithms and various selection features [6].

This research proposes Hybrid Data Mining with clustering or clustering technique and classification. Clustering is the process of grouping one physical or abstract object that has similarities [7]. The clustering process applies the K-Means algorithm. The K-Means algorithm categorizes entities into K clusters according to their attributes or features, where K is a positive integer [8]. Classification is a part of Data Mining technique to separate or group data and find patterns [9]. The classification process uses the K-Nearest Neighbor algorithm. The K-Nearest Neighbor algorithm, or KNN, is an algorithm for classifying objects based on the distance of new learning data from the K nearest neighbors [4].

The prediction analysis steps on the Shopee marketplace start with clustering the products sold on the Shopee marketplace at Soraya Bedsheet. The clustering process uses the K-Means algorithm to divide the training data into two clusters: the first cluster "High Demand" and the second cluster "Low Demand". The next step is to perform sales classification using the K-Nearest Neighbor (KNN) algorithm. The product data that has been clustered using the K-Means algorithm is used as input data for this step. The resulting classification is sales data that can be used to predict sales on the Shopee marketplace at Soraya Bedsheet.

Research related to the hybrid K-Means and KNN algorithm by [6] focuses on customer loan classification. Research [9] conducts product clustering for sales classification in minimarkets, where clustering functions to group sold products, followed by sales classification. Research [7] classifies Tokopedia tweet content to identify suitable tweet content to attract customer interest using the K-Means Clustering algorithm. Research [10], in predicting the spread of Covid-19 in Indonesia, applies a combination of the K-Means algorithm to determine clusters,

prediction, and mapping with the KNN and ID3 algorithms with 90% accuracy. Research [4] predicts the sales of best-selling products using the KNN algorithm. Research [11] predicts sales of Unilever products by applying the K-Nearest Neighbor algorithm. Furthermore, research [12] applies the KNN algorithm combined with K-Means using the model-based collaboration filtering method used to provide smartphone rating recommendations. Research [13] uses e-commerce data combined with the KNN algorithm to predict poverty levels in an area. Research [14] uses the K-Means and KNN algorithms for diabetes record classification.

The implications of this research are the generation of new knowledge from the application of the Hybrid Algorithm K-Means and K-Nearest Neighbor to provide product sales predictions on the Shopee marketplace for the Soraya Bedsheet store. This can yield more accurate and actionable insights, driving better inventory management and more effective marketing strategies. Furthermore, by predicting product sales on the Shopee marketplace, it can provide Soraya Bedsheet with information about the classification of products into Low Sales and High Sales categories, helping the business remain competitive in a dynamic market by enabling them to anticipate and respond to consumer needs more efficiently.

2. METHODS

This study proposes Hybrid Data Mining with clustering and classification techniques to predict sales on the Shopee marketplace. The clustering technique utilizes the K-Means algorithm to cluster sales products on the Shopee marketplace at Soraya Bedsheet, while the classification technique employs the K-Nearest Neighbor algorithm to provide sales predictions on the Shopee marketplace at Soraya Bedsheet. Figure 1 illustrates the research methodology.

2.1 Data Set

The data used in this study consists of product data from the Shopee marketplace at the Soraya Bedsheet store from September 2023 to March 2024. This data was obtained directly from the Shopee admin at the Soraya Bedsheet store. After completing the data collection stage, the next step is to conduct data preprocessing. This stage includes the process of data cleaning, data selection, and data transformation, as explained as follows.

1) Data Cleaning
This study conducts data cleaning by removing irrelevant data such as eliminating missing values, noisy data, and duplicate data. After data cleaning, the remaining data used amounts to 89 entries. This is because many product names are no

longer available on the Shopee marketplace, and these products are no longer restocked by Soraya Bedsheet.

2) Data Selection
Data selection is performed to choose attributes related to the research and to reduce the complexity of attributes processed in the data transformation stage. This study selects variables to be 4: stock, sold, and price. The results of the data selection process can be seen in Table 1.

Tabel 1. Data Set

No	Product Name (Indonesia)	Stock	Sold	Price
1	Soraya Bedsheet Bantal 3Pillow Premium Standard Size	99	44	160.000
2	Soraya Bedsheet Bantal Cinta Tanpa Sarung	136	14	85.000
3	Soraya Bedsheet Bantal Imut	98	2	80.000
…	……	…	…	…….
87	Soraya Bedsheet Sprei Swan Princess Salem King size	14	1	435.000
88	Soraya Bedsheet Sprei Tencel Motif A Tencel Leaft Blue	6	1	910.000
89	Soraya Bedsheet Sprei Waterproof King Soze	14	1	280.000

Table 1. represents the dataset used in this study to predict sales on the Shopee marketplace, totaling 89 entries.

3) Data Transformation
Before entering the data calculation, it will be transformed by assigning values to each data to facilitate the computation. The value range provided in the study is from 0 to 1.

Table 2. Attribute Weightening

Indicator Variables	Attribute	Value
Stock	≥ 100	1,0
	≥ 50	0,7
	> 50	0,4
Sold	≥ 35	1,0
	≥ 10	0,7
	> 10	0,4
Price	≥ Rp. 500.000	1,0

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Indicator Variables	Attribute	Value
	≥ Rp. 250.000	0,7
	> Rp. 250.000	0,4
Sales Decision	High Sales	0,7 – 1
	Low Sales	< 0,7

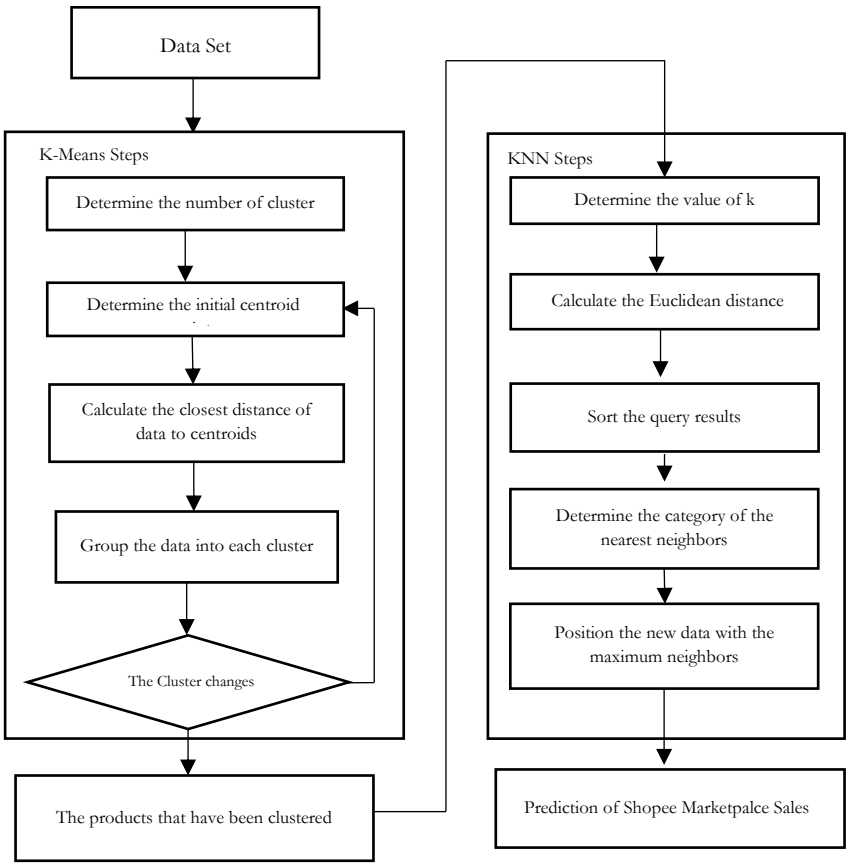

Figure 1. Research Methodology

Figure 1. The research methodology utilized in this study comprises 3 stages: Data Set, K-Means Steps, and KNN Steps. Below are the explanations for each stage.

After determining the weights for each attribute in Table 2, the next step is to transform the data in Table 2 into a standardized format for processing based on Table 2. The transformed data is shown in Table 3.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Table 3. Transformation Data

No	Product Name (Indonesia)	Stock	Sold	Price
1	Soraya Bedsheet Bantal 3Pillow Premium Standard Size	1	1	0,4
2	Soraya Bedsheet Bantal Cinta Tanpa Sarung	1	0,7	0,4
3	Soraya Bedsheet Bantal Imut	1	0,4	0,4
…	……	…	…	…….
87	Soraya Bedsheet Sprei Swan Princess Salem King size	0,4	0,4	0,7
88	Soraya Bedsheet Sprei Tencel Motif A Tencel Leaft Blue	0,4	0,4	1
89	Soraya Bedsheet Sprei Waterproof King Soze	0,4	0,4	0,7

This study applies Hybrid Data Mining, which combines two algorithms consisting of two stages of processing, where the results of the first stage will be used as training data for the second stage. Table 4 is the testing data used, which is transaction data in March 2024 to be classified in the training data.

Table 4. Testing Data

No	Product Name (Indonesia)	Stock	Sold	Price
1	Soraya Bedsheet Sprei Katun Gold Motif Dallas Ask Grey Tua	0,7	1	0,4

2.2 K-Means Algorithm Steps

The first stage involves clustering the products sold on the Shopee marketplace at Soraya Bedsheet store by implementing the K-Means algorithm. The K-Means algorithm categorizes entities into K clusters according to their attributes or features, where K is a positive integer. The K-Means algorithm partitions data with similar characteristics into the same cluster or group, and conversely, data with different characteristics are placed into different clusters or groups [7] . According to [3], the K-Means algorithm has several characteristics, including:

1) K-Means is very fast in the clustering process.
2) K-Means is highly sensitive to the random initialization of centroids.
3) K-Means allows clusters to have no members.
4) K-Means clustering results are unique (always changing; sometimes good, sometimes not good).

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

According to [15], the K-Means algorithm has steps in its process, including:

1. Determine the total number of clusters or groups desired.
2. Determine the initial centroid values randomly.
3. Allocate each data point to the nearest centroid or the average of the data points closest to it. The Euclidean Distance equation is used to measure the distance from data points to the centroid center as shown in equation (1).

$$D(p,q)= \sqrt{(p1-p2)^2+(q1-q2)^2} \quad (1)$$

   Where p and q are data points.
4. Group or merge data to determine the members of each cluster based on distance.
5. Repeat step 2 until the obtained centroids are the same as the previous cluster.

2.3 K-Nearest Neighbors Algorithm Steps

The second step involves performing the sales classification process to predict sales using the K-Nearest Neighbor (KNN) algorithm. The K-Nearest Neighbor algorithm classifies objects based on the distance of new data to the K nearest neighbors. The steps in predicting using the KNN algorithm involve training and testing data based on categories from samples or past data, where the nearest neighbors to the test sample, according to the K training samples, are assigned to the category with the highest probability [16]. The data from the "High Interest" cluster is used in the K-Nearest Neighbor algorithm. The classification in the K-Nearest Neighbor algorithm follows several stages according to [14], including:

1) Determine the value of k.
2) Compute the distance between the target data point and the k neighbors using the Euclidean distance, which can be determined by the following equation (2).

$$D(p,q)= \sqrt{(p1-p2)^2+(q1-q2)^2} \quad (2)$$

   Where p and q are data points.
3) Select and sort the k nearest neighbor groups that have the smallest Euclidean distances.
4) Count the number of data points from each class among the k neighbors identified in the previous step.
5) Position the new data point into the category where the maximum number of neighbors is present. From the classification results, predictions for product sales on the Shopee marketplace at Soraya Bedsheet are made.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3. RESULTS AND DISCUSSION

In this subsection, the research results and discussions are presented by applying the K-Means and K-Nearest Neighbor algorithms using Python tools.

3.1 K-Means Algorithm

The K-Means algorithm resulted in 2 clusters: the first cluster is "Low Interest" and the second cluster is "High Interest". Here are the steps of the K-Means algorithm:

1. Prepare the dataset
   The dataset used in the K-Means algorithm consists of 89 data points with three attributes: stock, sold items, and price, which have been transformed in the Table 3 process.

```
Out[6]: array([[1. , 1. , 0.4],
               [1. , 0.7, 0.4],
               [1. , 0.4, 0.4],
               [1. , 1. , 0.4],
               [0.4, 0.4, 1. ]])
```

Figure 2. K-Means Algorithm Data

   Figure 2 depicts the data used in the K-Means algorithm processing, which has been previously transformed.

2. Determination of cluster count
   The clusters used consist of 2 clusters, namely the "Low Interest" cluster and the "High Interest" cluster. In the K-Means algorithm, the initial centroids are determined. The centroid for C1 is taken from data entry 33, and the centroid for C2 is taken from data entry 63. The initial centroids used can be seen in Table 5.

Table 5. Initial Centroid

Centroid	X	Y	Z
C1	0,4	0,4	1
C2	0,7	0,4	0,7

   Table 5 represents the initial centroids in the K-Means Algorithm calculation with 2 clusters, C1 and C2, with the variable Stock initialized

as (X), the variable Sold as (Y), and the variable Price as (Z). Next, the determination of clusters in the k-means algorithm using Python tools can be seen in Figure 3.

```
#menjalankan K-Means Clustering ke dataset dengan 2 cluster
kmeans = KMeans(n_clusters = 2, init = 'k-means++', random_state = 42)
y_kmeans = kmeans.fit_predict(x)

dataset['f_cluster'] = y_kmeans + 1
dataset.head(89)
```

Figure 3. Cluster Determination

Figure 3. Determination of clusters in the K-Means algorithm using Python, where the number of clusters used is 2 clusters and **kmeans.fit_predict(x)** is used for the computation.

3. Results of K-Means Algorithm

Based on calculations using Python 3.11 software with 89 product data points, 64 data points are classified into cluster 1 "Low Demand" and 25 data points are classified into cluster 2 "High Demand." Table 6. below shows the clustering results for cluster values C1 and C2.

Table 5. Clustering Results for Cluster Values C1 and C2

No	Product Name	C1	C2	C1	C2
1	Soraya Bedsheet Bantal 3Pillow Premium Standard Size	0,89	0,34		C2
2	Soraya Bedsheet Bantal Cinta Tanpa Sarung	0,73	0,13		C2
3	Soraya Bedsheet Bantal Imut	0,67	0,31		C2
4	Soraya Bedsheet Bantal Soft Deluxe Eksklusif 3pillow	0,89	0,34		C2
5	Soraya Bedsheet Bedcover Eco Emily Pink King Size	0,24	0,78	C1	
…	……	…….	……	…….	……
87	Soraya Bedsheet Sprei Swan Princess Salem King size	0,07	0,64	C1	
88	Soraya Bedsheet Sprei Tencel Motif A Tencel Leaft Blue	0,24	0,78	C1	
89	Soraya Bedsheet Sprei Waterproof King Soze	0,07	0,64		

Table 6. Clustering results of products using the K-Means algorithm, where C1 is the "High Demand" cluster and C2 is the "Low Demand" cluster. Subsequently, the K-Nearest Neighbor algorithm is applied using the High Demand cluster as the training data.

3.2 K-Nearest Neighbor Algorithm

Next, the sales classification process is performed to predict sales using the K-Nearest Neighbor algorithm. The K-Nearest Neighbor algorithm will yield 2 decisions: "High Sales" and "Low Sales" using the data obtained from the "Many Fans" cluster. The first step is to determine the number of k values used, namely 3, 5, and 7. The following are the steps of the K-Nearest Neighbor algorithm:

1. Inputiing libraries and dataset training
   The training dataset used is from the "High Demand" cluster. The training dataset is first labeled for classification. High sales are classified with a weight of (0.7 – 1.0) and Low sales with a weight of (<0.7). Table 7. shows the training data used in the K-Nearest Neighbor algorithm.

Table 7. Training Data

No	Product Name	Stock	Sold	Price	Clasiification
1	Soraya Bedsheet Bantal 3Pillow Premium Standard Size	1	1	0,4	High Sales
2	Soraya Bedsheet Bantal Cinta Tanpa Sarung	1	0,7	0,4	High Sales
3	Soraya Bedsheet Bantal Imut	1	0,4	0,4	Low Sales
4	Soraya Bedsheet Bantal Soft Deluxe Eksklusif 3pillow	1	1	0,4	High Sales
5	Soraya Bedsheet Isi Bantal Imut Tanpa Sarung	1	0,4	0,4	Low Sales
6	Soraya Bedsheet Katun Gold Motif Ambrose Colorful	1	0,7	0,7	High Sales
7	Soraya Bedsheet Sarung Bantal Cinta	0,7	0,4	0,7	Low Sales
8	Soraya Bedsheet Sarung Bantal Gendang	1	1	0,4	High Sales
9	Soraya Bedsheet Sarung Bantal Imut Tanpa Isi	1	0,4	0,4	Low Sales
10	Soraya Bedsheet Sarung Bantal Kepala Tambahan	1	1	0,4	High Sales

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

No	Product Name	Stock	Sold	Price	Clasiification
11	Soraya Bedsheet Selimut Bercahaya	0,7	0,4	0,4	Low Sales
12	Soraya Bedsheet Selimut Embos Polos King Size	0,7	0,4	0,4	Low Sales
13	Soraya Bedsheet Selimut Kotak Jepang	1	1	0,4	High Sales
14	Soraya Bedsheet Selimut Salju Single Bermotif	1	1	0,4	High Sales
15	Soraya Bedsheet Selimut Salur Polos	1	1	0,4	High Sales
16	Soraya Bedsheet Set Sarung Kulkas dan Sarung Dispenser	1	1	0,4	High Sales
17	Soraya Bedsheet Soo.Bag Katun Gold	0,7	0,4	0,4	Low Sales
18	Soraya Bedsheet Sprei Antoinette Rimple dan Karet King and Queen Size	1	0,7	0,7	High Sales
19	Soraya Bedsheet Sprei Excellent Katun Gold Motif Pevita	1	0,4	1	High Sales
20	Soraya Bedsheet Sprei Florence Cream Single Size	1	0,4	0,4	Low Sales
21	Soraya Bedsheet Sprei Karet dan Rimple Balmont Abu King and Queen Size	1	0,4	0,7	High Sales
22	Soraya Bedsheet Sprei Katun Gold Motif DR Monogram Biru	0,7	0,7	0,7	High Sales
23	Soraya Bedsheet Sprei Katun Gold Motif Florence Cream	1	0,7	0,7	High Sales
24	Soraya Bedsheet Sprei Katun Gold Motif Florence Cream King Size	1	1	0,7	High Sales
25	Soraya Bedsheet Sprei Rimple dan Karet Bianca King and Queen Size	1	0,4	0,7	High Sales

Table 7. The training dataset consists of 25 data points with 5 variables: Product Name, Stock, Price, Sold, and Classification.

2. Convert data to x_train and y_train
   The data included in x_train consists of attributes such as stock, sold, and price, while the data included in y_train represents the classification.

3. Creating a testing model
   The K-Nearest Neighbor algorithm utilizes testing data, which consists of transaction data for the product named Soraya Bedsheet Cotton Gold Motif Dallas Ask Grey Tua. The testing data has not been classified yet, and testing will be conducted for different values of k. Below are the attribute values for the testing data as shown in Figure 5.

```
#sales prediction
#create testing model
Stock = 0.7
Sold = 1
Price = 0.4
x_new = np.array([Stock, Sold, Price]).reshape(1, -1)
x_new
```

```
array([[0.7, 1. , 0.4]])
```

Figure 4. Testing Data for KNN Algorithm

4. Training Model
   KNN model built using the KneighborsClassifier function from the **sklearn.neighbors** library. The process of classification involves testing 3 values of k, namely k=3, k=5, and k=7. Here are the results of the KNN algorithm with testing of k values in Table 8.

Table 8. KNN Algorithm Results

Value of K	Prediction Using K-Nearest Neighbor
3	High Sales
5	High Sales
7	High Sales

Based on the testing results in Table VI with k values of 3, 5, and 7, it can be concluded that the product "Soraya Bedsheet Katun Gold Motif Dallas Ask Grey Tua" can be predicted to fall under "High Sales." Therefore, Soraya Bedsheet can ensure the availability of this product, especially on the Shopee marketplace, under the name "Soraya Bedsheet Katun Gold Motif Dallas Ask Grey Tua" to increase product stock. This way, customer demand for the "Soraya Bedsheet Katun Gold Motif Dallas Ask Grey Tua" product on the Shopee marketplace will be met, and no customers will miss out on the product due to stock shortages.

3.3 Accuracy Testing

The Hybrid K-Means and K-Nearest Neighbor algorithm underwent accuracy testing to predict sales on the Shopee marketplace at Soraya Bedsheet's store.

Table 9. Accuracy Testing Results

Value of k	Accuracy Testing Results
3	96%
5	96%
7	96%

The accuracy test results in Table 9. show an accuracy of 96% for all three k values. The conclusion from the accuracy test of the Hybrid K-Means and K-Nearest Neighbor algorithm on the Shopee marketplace is 96%. From this accuracy result of 96%, it can be concluded that the Hybrid K-Means and KNN algorithm can improve the accuracy of product demand predictions and optimize inventory and marketing strategies.

4. CONCLUSION

This research presented a method to enhance the precision of product demand forecasts. The study started by applying the K-Means algorithm to cluster sales products from the Soraya Bedsheet store on Shopee, resulting in two distinct clusters: cluster 0 labeled "Low Demand" with 64 data points, and cluster 1 labeled "High Demand" with 25 data points. Following this, the data from the High Demand cluster were utilized as training data for the K-Nearest Neighbor (KNN) algorithm, which produced two sales predictions: Low Sales and High Sales. Testing the KNN algorithm with three different k values (3, 5, and 7) indicated that the product "Soraya Bedsheet Katun Gold Motif Dallas Ask Grey Tua" could be categorized under "High Sales." The combined approach of K-Means and KNN algorithms achieved an accuracy rate of 96%. These findings suggest that integrating K-Means and KNN algorithms can substantially improve the accuracy of product demand predictions, thus optimizing inventory management and marketing strategies.

REFERENCES

[1] S. Roni and C. Crysdian, "Studi Literature Analisis Potensi Pasar Marketplace terhadap Penjualan," *J. Teknol. dan Manaj. Inform.*, vol. 8, no. 2, pp. 134–142, 2022, doi: 10.26905/jtmi.v8i2.9055.

[2] A. Kurniawati and N. Ariyani, "Sales Promotion Strategy on Shopee Marketplace," *Propaganda*, vol. 2, no. 1, pp. 65–79, 2022.

[3] Nurmalasari *et al.*, "Implementation of Clustering Algorithm Method for Customer Segmentation," *J. Comput. Theor. Nanosci.*, vol. 17, no. 2, pp. 1388–1395, 2020, doi: 10.1166/jctn.2020.8815.

[4] S. P. Dewi, N. Nurwati, and E. Rahayu, "Penerapan Data Mining Untuk Prediksi Penjualan Produk Terlaris Menggunakan Metode K-Nearest Neighbor," *Build. Informatics, Technol. Sci.*, vol. 3, no. 4, pp. 639–648, 2022, doi: 10.47065/bits.v3i4.1408.

[5] Nursobah, S. Lailiyah, B. Harpad, and M. Fahmi, "Penerapan Data Mining Untuk Prediksi Perkiraan Hujan dengan Menggunakan Algoritma K-Nearest Neighbor," *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1395–1400, 2022, doi: 10.47065/bits.v4i3.2564.

[6] E. P. W. Mandala, E. Rianti, and S. Defit, "Classification of Customer Loans Using Hybrid Data Mining," *JUITA J. Inform.*, vol. 10, no. 1, p. 45, 2022, doi: 10.30595/juita.v10i1.12521.

[7] J. Rejito, A. Atthariq, and A. S. Abdullah, "Application of text mining employing k-means algorithms for clustering tweets of Tokopedia," *J. Phys. Conf. Ser.*, vol. 1722, no. 1, 2021, doi: 10.1088/1742-6596/1722/1/012019.

[8] O. A. Alghanam, S. N. Al-Khatib, and M. O. Hiari, "Data Mining Model for Predicting Customer Purchase Behavior in e-Commerce Context," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 2, pp. 421–428, 2022, doi: 10.14569/IJACSA.2022.0130249.

[9] Dewi Eka Putri and Eka Praja Wiyata Mandala, "Hybrid Data Mining berdasarkan Klasterisasi Produk untuk Klasifikasi Penjualan," *J. KomtekInfo*, vol. 9, pp. 68–73, 2022, doi: 10.35134/komtekinfo.v9i2.279.

[10] S. Arlis and S. Defit, "Machine Learning Algorithms for Predicting the Spread of Covid–19 in Indonesia," *TEM Journal*, vol. 10, no. 2. pp. 970–974, 2021. doi: 10.18421/TEM102-61.

[11] A. Alfani W.P.R., F. Rozi, and F. Sukmana, "Prediksi Penjualan Produk Unilever Menggunakan Metode K-Nearest Neighbor," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 6, no. 1, pp. 155–160, 2021, doi: 10.29100/jipi.v6i1.1910.

[12] F. A. Prayoga and K. Kusnawi, "Smartphone Recommendation System Using Model-Based Collaborative Filtering Method," *J. Tek. Inform.*, vol. 3, no. 6, pp. 1613–1622, 2022, doi: 10.20884/1.jutif.2022.3.6.413.

[13] T. F. Aulia, D. R. Wijaya, E. Hernawati, and W. Hidayat, "Poverty Level Prediction Based on E-Commerce Data Using K-Nearest Neighbor and Information-Theoretical-Based Feature Selection," pp. 28–33, 2021.

[14] X. Li, J. Zhang, and F. Safara, "Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm," *Neural Process. Lett.*, vol. 55, no. 1, pp. 153–169, 2023, doi: 10.1007/s11063-021-10491-0.

[15] A. H. Nasyuha, Zulham, and I. Rusydi, "Implementation of K-means algorithm in data analysis," *Telkomnika (Telecommunication Comput. Electron.*

Control., vol. 20, no. 2, pp. 307–313, 2022, doi: 10.12928/TELKOMNIKA.v20i2.21986.

[16] Al-Khowarizmi, R. Syah, M. K. M. Nasution, and M. Elveny, "Sensitivity of MAPE using detection rate for big data forecasting crude palm oil on k-nearest neighbor," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 3, pp. 2696–2703, 2021, doi: 10.11591/ijece.v11i3.pp2696-2703.