APPLICATION OF THE K-MEANS METHOD FOR CLUSTERING BEST SELLING PRODUCTS IN ICE CREAM SALES

Hana Atthifa Ryantika[1*], Supriyanto[2], Asep Afandi[3]

[1,2,3]Information Systems Study Program, ITBA Dian Cipta Cendikia, Kotabumi, North Lampung, Lampung
[1,2,3]Jl. Negara No. 03 Candimas, Kotabumi, North Lampung, Lampung, Indonesia
E-mail: hanaatthifa30@gmail.com[1*], supriyanto@dcc.ac.id[2], Asepafandi189@gmail.com[3]

Abstract

From youngsters to adults, ice cream is a culinary item that is highly sought after by a variety of demographics. The ice cream combination, which includes a variety of culinary items including dairy products, sugars, stabilizers, flavor enhancers, and eggs, is frozen to create this dish. Aice Ice Cream has a wide variety of goods, which makes it popular with customers. According to various data that the author has observed from Agent Aice Harun, including the fact that there are still challenges in the stock of goods available to meet consumer demand, it is necessary to forecast sales of Ice Cream Aice products that are most in demand by consumers in order to facilitate stock provision. Due to these problems, the researchers conducted. The K-Means Clustering Algorithm approach will be applied manually during the clustering process, and Python data mining tools will be used for the implementation. The final calculation results are the same, showing that 48 things are the most desirable, 6 items are pretty attractive, and 6 items are less desirable, according to the results of determining centroid values arbitrarily using the Euclidian Distance formula manually using Python Tools. Agent Aice Harun is able to supply goods that are in line with consumer preferences as a result. Python computations and manual calculations get the same conclusion. Therefore, it is evident from these data that the accuracy value attained is 100%.

1. INTRODUCTION

Agent Aice Harun is an agent engaged in selling Ice Cream Aice or a distributor of Ice Cream Aice in Kotabumi. Ice Cream Aice is a frozen food made from dairy products such as cream, combined with artificial or natural flavors and sweeteners. Ice Cream Aice has many types of products to offer so it is in great demand by many consumers. Forecasts are needed for sales of the Ice Cream Aice products that consumers are most interested in buying to make it easier to provide a stock of goods. The agent can choose the inventory that must be satisfied in order to receive the appropriate information. The authors, therefore looked into Agent Aice Harun.

The K-Means approach is suggested in this paper. This technique belongs to the grouping or clustering function[1]. Data is divided into one or more clusters or groups using the K-means method in non-hierarchical clustering functions[2]. Implementing data mining using the K-means approach to analyze sales of household furnishings at smart kitchen retailers is similar to the prior study, which showed that grouping sales data at smart kitchen stores using the K-means algorithm can produce recommendations for goods that are in demand, less in demand, and quite in demand. For the management of the supply of goods utilize the data as a guide, so that the agents do not let customers down by not having the products they they desire in stock.

K-Means is a non-hierarchical data clustering method that seeks to partition existing data into one or more clusters/groups so that data with different characteristics are collected into other groups[3].

Data mining is useful to provide solutions to make deep decisions for sales to increase sales company[4].

so that data that has the same characteristics will be grouped in one cluster, and data that has different characteristics will be grouped into another group [5]. As a result, inventory or stock is an important activity because if there are no inventories, the customer may be dissatisfied if the goods required are not available[6]. By using this method, the data that has been obtained can be grouped into several clusters based on the similarity of the data[7]. Sufferers in several Asian countries use K-Means[8]. Software that is used as a support for data processing is Python.

II. LITERATURE
2.1 Data Mining
Based on previous research entitled "Application of data mining for product classification of brick brands using the k-means algorithm" engaged in the sale of various products. However, brick shops cannot group products into best-selling and not-selling categories, so they experience a shortage of best-selling product stock due to a buildup of unselling products.

The same thing happened in the previous research entitled "Cluster Analysis of Sales Transaction Data Using K-Means Clustering at Mandiri Business Stores" which was engaged in selling packaged drinks. Independent business shops also have problems with stock, where some goods passively stay in the warehouse for a relatively long time due to sales of goods that are not smooth and not selling well. Based on the problems with independent business shops, data mining is needed to classify items that are in demand and not in demand.

Data Mining is a process for obtaining useful information from large database warehouses as well as for use in making very important decisions[9]. Data Mining on the other hand is a process that uses statistical, mathematical, artificial, and machine learning techniques to extract and identify useful information and related knowledge from various databases[10]. According to Urban in his book "Decisions support systems and intelligent systems," data mining is the term used to describe the discovery of knowledge in databases[11]. Data mining is part of knowledge discovery in databases (KDD).

KDD itself is a processor stage that aims to determine important information from large amounts of data that cannot be identified manually[12]. The general form of the KDD process can be explained in Figure 1 below.
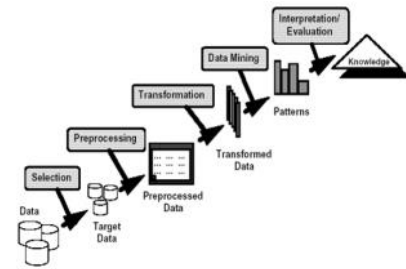

Figure 1. Knowledge Discovery in Database[13].

2.2 Clustering
is a component of the data mining technique that gathers items with commonalities among the data groupings and differences from items in other groups. In other terms, Based on the results of observations, the process of clustering involves grouping data elements or splitting data into distinct clusters or groups[14]. According to V.M., 2017 "Clustering is an effort used to group observations and group them into classes that have the same object[15]. In this study, the researchers grouped them into 3 clusterings, namely the products that were most desirable, sufficiently desirable, and less desirable.

2.3 Python
a high-level programming language that uses dynamic semantics to improve syntax readability and can execute several instructions by employing object-oriented programming methods[16]. Python is equipped with pointers or automatic memory management[17]. Currently, python is used to develop software and can run through various operating systems. As in the previous study entitled design of lecturer performance applications in the faculty quality assurance unit using the Python based K-Means Clustering Algorithm where in this study python can help make it easier for the quality assurance unit to carry out monitoring activities.

2.4 K-Means
K-Means was discovered by several people, namely Lloyd (1957, 1982), Forgey (1965), Friedman and Rubin (1967), and McQueen (1967)[18]. K-Means is a non-hierarchical data clustering method that tries to partition existing data into one or more clusters or groups. Determining The initial centroid, this determination is done randomly on existing data tables[19]. This method partitions into clusters or groups so that data that have the same characteristics are grouped into one cluster, and those that have different characteristics are grouped into a different cluster.

In this study, researchers will group into three clusterings, namely products that are most in demand, moderately desirable, and less desirable so that it makes it easier for Aice Agents to provide a stock of goods that are of interest to many consumers.

To calculate the distance of all data to each cluster center point, the Euclidean Distance[20] theory can be used which is formulated as follows:

$$D(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \dots \dots \dots$$

D(x, y) = Distance data to x to cluster center y
x1 = Data to i in attribute to x
y1 = center point to i on attribute to x

III. RESEARCH METHODS

A. Data Collection Methods

Methods used by researchers to obtain the data required for their studies are known as data collection techniques, and they are crucial to research. The following is the data collection strategy applied in this study:

Observation

Is a data collection technique by conducting direct observation. For the process of obtaining data and information needed by the author directly. The author requests data from Agent Aice Harun so that with this data the data preprocessing stage is carried out.

Interview

Is a data collection technique through direct question and answer activities between two parties, namely interviews and resource persons whose aim is to obtain accurate information.

Library Studies

Study scientific work, in the form of scientific books, journals, and other scientific sources which of course are in accordance with the type of research and have a relationship to the problem being studied. The scientific references used by the authors are the sources contained in the Bibliography.

B. Research Flow

This research was conducted using the K-Means Clustering method, to determine sales statistics. The flow of the research stage made as a framework can be seen as shown below:
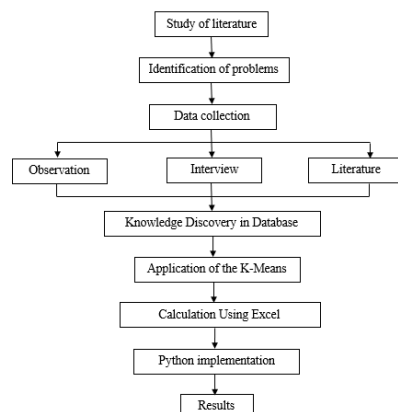

Figure 2. Research Design

IV. RESULTS

The data used is Agent Aice Harun's data. With the process of collecting data through an observation, interviews, and literature study. The method used is K-Means Clustering with attributes of item name, stock item and total sales. The amount of data tested is 60 data obtained from data for 2020-2022 and the software used is Python.

After performing the eighth literacy, the cluster results from the clusterization of the Aice Ice Cream data using k-means were as follows: the Aice Ice Cream group with the highest demand is in cluster 1, which has 48 products as shown in Table 32, the Aice Ice Cream group that is quite alluring is in cluster 2, which has 6 products as shown in table 32, and the Aice Ice Cream group that is less desirable is in cluster 3, which also has 6 products as shown in table 32..

Table 1. Members of Cluster 1 Aice Ice Cream (most popular)

No	Name of goods	Sale	Total Sales
1	AICE 2 Colors Stick 50ML	76	67
2	AICE 2 in 1 Colors 65ML	632	626
3	AICE 3 in 1 Bucket 8L	5,875	5,868
4	AICE 3 in 1 Colors 95GR	4	3
5	AICE Banana Crispy 50GR	143	136
6	AICE Bingo 40GR	177	170
7	AICE Bingo Cookies Cone 90GR	5	4
8	AICE Blueberry Cookies 65ML	1.016	1.008
9	AICE Blueberry Yoghurt 60ML	147	136
10	AICE Choco Boost 60GR	48	35

Table 2. Members of Cluster 1 Aice Ice Cream (quite popular)

N0	Name of goods	Sale	Total Sales
1	AICE Fruit Twister 60ML	13,654	13,643
2	AICE Orange 65GR	32,380	32,368
3	AICE Mango Slush 65GR	22,679	22,674
4	AICE	21,214	21,211

| 5 | Passion Fruit Stick 55ML AICE Sundae Avocado Chocolate 100ML | 19,685 | 19,682 |
| 6 | AICE Chocolate Sundae 100ML | 19,874 | 19,869 |

The table above is the result of calculations using the K-Means clustering method, 6 products are obtained which are quite attractive to consumers.

Table 3. Members of Cluster 1 Aice Ice Cream (less desirable)

N0	Name of goods	Sale	Total Sales
1	AICE Torch Chocolate 65GR	49,270	49,269
2	AICE Watermelon Stick 65GR	48,935	48,928
3	AICE Sweet Pears 80ML	38,035	38,032
4	AICE Taro Waffle Cone 45GR	44,259	44,347
5	AICE Toffie Koffie 45GR	48,893	48,884
6	AICE Two Colors 50ML	58,022	58,001

The following are the results of calculations using the K-Means Clustering method, this research helps make it easier for Agent Aice Harun to find out what products consumers are interested in so that there is no accumulation of goods that consumers are less interested in. Based on previous research using the K-Means method, accurate results were obtained, so this study was very helpful for researchers in determining the desired results.

The data used in this research are item names, stock items, and total sales of ice cream products based on the last 3 years 2020-2022, this data was taken from Agent Aice. The data used has gone through a process of selection, cleaning, and transformation. This data will be used in the clustering process.
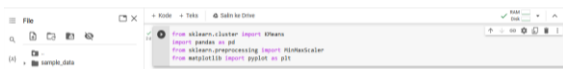

Figure 3. Import Libraries

The imported data will then be retrieved, where the data will be used in the clustering process. After the process of pulling the data, then re-labeling the data is useful for knowing what items are included in the clustering category. As shown in Figure 4 and Figure 5 below
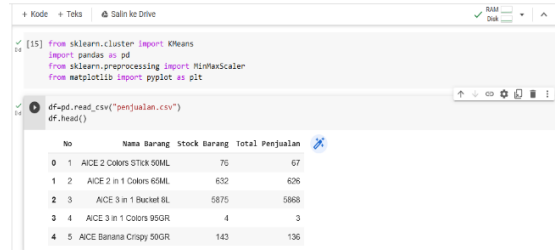

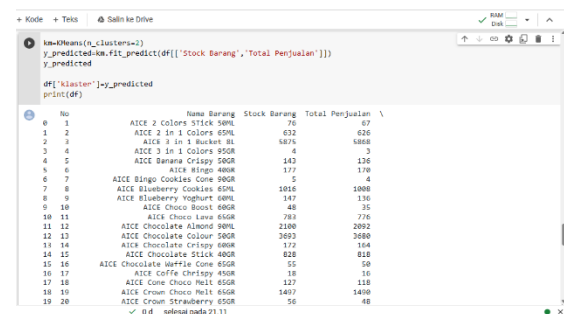Figure 4. Process of Pulling Data in Python

Figure 5. Display of Prediction Results

The predicted data will then be processed to the finishing stage, namely visualizing again in graphical form. The results of the finishing process will be graphical with different color patterns for each cluster. Can be seen as shown below.
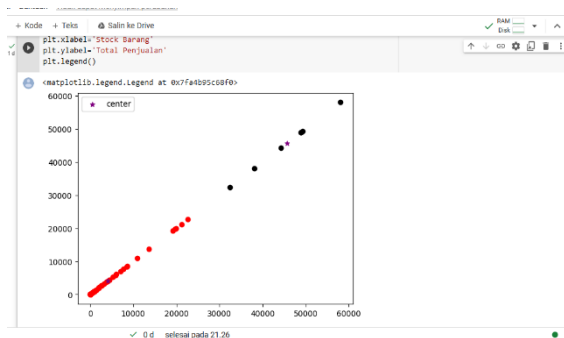

Figure 4. Display of Finishing Graphic Results

From the results of the picture above it can be concluded that the implementation results using Python software are very effective and the results of finishing the graph are in accordance with manual calculations.

V. CONCLUSION

The final calculation results are the same, showing that 48 things are the most desirable, 6 items are pretty attractive, and 6 items are less desirable, according to the results of determining centroid values arbitrarily using the Euclidian Distance

formula manually using Python Tools. Agent Aice Harun can supply goods that are in line with consumer preferences. Python computations and manual calculations get the same conclusion. Therefore, it is evident from these data that the accuracy value attained is 100%.

References

[1]   Y. D. Darmi dan A. Setiawan, "Penerapan Metode Clustering K-Means Dalam Pengelompokan Penjualan Produk," *J. Media Infotama*, vol. 12, no. 2, hal. 148–157, 2017, doi: 10.37676/jmi.v12i2.418.

[2]   Z. Alamtaha, I. Djakaria, dan N. I. Yahya, "Implementasi Algoritma Hierarchical Clustering dan Non-Hierarchical Clustering untuk Pengelompokkan Pengguna Media Sosial," vol. 4, no. 1, hal. 33–43, 2023, doi: 10.20956/ejsa.vi.24830.

[3]   P. D. P. Silitonga, M. Ginting, T. Informatika, U. Katolik, S. Thomas, dan S. Utara, "Klasterisasi Keranjang Belanja Transaksi Penjualan," vol. 2, no. 2, hal. 164–168, 2018.

[4]   M. H. Fakhriza dan K. Umam, "Analisis Produk Terlaris Menggunakan Metode K-Means Clustering Pada "Pt.Sukanda Djaya," *JIKA (Jurnal Inform.*, vol. 5, no. 1, hal. 8, 2021, doi: 10.31000/jika.v5i1.3236.

[5]   S. Jari, D. Notifikasi, dan E. Berbasis, "Pengamanan lemari penyimpanan menggunakan sidik jari dengan notifikasi email berbasis iot" vol. 8, hal. 215–225, 2023.

[6]   R. Hardianto, H. Ramadhan, E. Putra Pane, dan Y. Yunefri, "K-Means Clustering in Determining the Category of Stock Items In Angkasa Mart," *Knowbase Int. J. Knowl. Database*, vol. 2, no. 1, hal. 30, 2022, doi: 10.30983/ijokid.v2i1.5411.

[7]   B. Jasenta Mikha, M. Zunaidi, dan F. Sonata, "Implementasi Data Mining Untuk Klasterisasi Penjualan Makanan Dan Minuman Menggunakan Metode K-Means (Studi Kasus : Bloum Coffee) melalui wawancara dengan petugas gudang Bloum Coffee pada bulan januari 2021 . Risol Ayam Dim Sum Udang Dim Sum Kepiting Di," vol. 3, no. 6, 2020.

[8]   F. A. Fithri dan S. Wardhana, "Cluster Analysis of Sales Transaction Data Using K-Means Clustering At Toko Usaha Mandiri," *J. PILAR Nusa Mandiri*, vol. 17, no. 2, hal. 113–118, 2021, [Daring]. Tersedia pada: https://doi.org/10.33480/pilar.v17i2.2273

[9]   A. F. Sallaby, R. T. Alinse, V. N. Sari, dan T. Ramadani, "Pengelompokan Barang Menggunakan Metode K-Means Clustering Berdasarkan Hasil Penjualan Di Toko Widya Bengkulu," *J. Media Infotama*, vol. 18, no. 1, hal. 99–104, 2022, [Daring]. Tersedia pada: https://jurnal.unived.ac.id/index.php/jmi/arti cle/view/2126

[10]   Eka Sugiyarti, Kamarul Azmi Jasmi, Bushrah Basiron, dan Miftachul Huda, "Decision Support System of Scholarship Grantee Selection using Data Mining, International Journal of Pure and Applied Mathematics," *Int. J. Pure Appl. Math.*, vol. 119, No. 15, Hal. 2239–2249, 2018, [Daring]. Tersedia Pada: Https://Www.Academia.Edu/Download/570 52927/Decision_Support_System_Of_Schol arship_Grantee.Pdf

[11]   A. I. Waspah *et al.*, "Expectation Maximization Algorithm Memprediksi Penjualan Susu Murni Pada Pt . Sewu Primatama Indonesia Lampung," *JUTIM (Jurnal Tek. Inform. Musirawas)*, vol. 7, no. 1, hal. 27–38, 2022, [Daring]. Tersedia pada: http://jurnal.univbinainsan.ac.id/index.php/ju tim/article/view/1615%0Ahttp://jurnal.univb inainsan.ac.id/index.php/jutim/article/downl oad/1615/822

[12]   I. Pendahuluan, "Data Mining Manhattan Distance dan Euclidean Distance pada Algoritma X-Means dalam Klasifikasi Minat dan Bakat Siswa," vol. 7, no. suryabrata 1995, hal. 835–842, 2023.

[13]   M. Annas dan S. N. Wahab, "Data Mining Methods : K-Means Clustering Algorithms," vol. 3, no. 1, 2023.

[14]   E. V. Astuti dan R. Mawarni, "the Comparison Using Expectation-Maximization Algorithm and C4 . 5 Algorithm To Predict the Result of Biogas Production As a Power Plant At Pt Budi Starch & Sweetener (Bssw)," hal. 186–198.

[15]   A. Nugraha, O. Nurdiawan, dan G. Dwilestari, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Yana Sport," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, hal. 849–855, 2022, doi: 10.36040/jati.v6i2.5755.

[16]   A. Y. Simanjuntak, I. S. E. S. Simatupang, dan A. Anita, "Implementasi Data Mining Menggunakan Metode Naã• Ve Bayes Classifier Untuk Data Kenaikan Pangkat Dinas Ketenagakerjaan Kota Medan," *J. Sci. Soc. Res.*, vol. 5, no. 1, hal. 85, 2022, doi: 10.54314/jssr.v5i1.804.

[17]   F. N. Dhewayani, D. Amelia, D. N. Alifah, B. N. Sari, dan M. Jajuli, "Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM," *J. Teknol. dan Inf.*, vol. 12, no. 1, hal. 64–77, 2022, doi: 10.34010/jati.v12i1.6674.

[18]   S. Mintoro dan A. Afandi, "Implementasi Algoritma K-Means Dan Algoritma Apriori Optimasi Kinerja Ecu (Study Kasus Mobil Avanza Dan Xenia)," *J. Inf. dan Komput.*,

vol. 9, no. 2, hal. 81–88, 2021, doi: 10.35959/jik.v9i2.235.

[19]   A. Afandi, D. Nurdianah, P. C. Rejo, N. Bayes, dan K. A. Dominan, "Naive Bayes Method and C4 . 5 in Classification of Birth Data," vol. 16, no. 4, hal. 435–446, 2022.

[20]   W. Wahyu Pribadi, A. Yunus, dan A. S. Wiguna, "Perbandingan Metode K-Means Euclidean Distance Dan Manhattan Distance Pada Penentuan Zonasi Covid-19 Di Kabupaten Malang," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, hal. 493–500, 2022, doi: 10.36040/jati.v6i2.4808.