Contemporary Accounting Case Studies

Volume 3, Nomor 1, March 2024 Article 5

IMPLEMENTATION OF DATA ANALYSIS USING K-MEANS CLUSTERING ALGORITHM FOR THE EXPANSION OF TELECOMMUNICATIONS SERVICES - CASE STUDY ON TELECOMMUNICATIONS COMPANIES

Dendy Tryanda
Master of Accounting Program, Faculty of Economics and Business, Universitas Indonesia
dendy.tryanda@ui.ac.id

Isnaeni Achdiat
Master of Accounting Program, Faculty of Economics and Business, Universitas Indonesia
isnaeni.achdiat@gmail.com

Contemporary Accounting Case Studies,
March 2024, Vol. 3, No. 1, pg. 62-79

DATA ANALYSIS USING K-MEANS CLUSTERING ALGORITHM FOR THE EXPANSION OF TELECOMMUNICATIONS SERVICES - CASE STUDY ON TELECOMMUNICATIONS COMPANIES

Dendy Tryanda [*]***, Isnaeni Achdiat***
Master of Accounting Program, Faculty of Economics and Business, Universitas Indonesia

ABSTRACT

Internet fixed broadband products or internet products using cables is a products that are rarely used by Indonesian people, even though in the Covid-19 era with a work from home system, people need a good quality internet. This study aims to help PT ABC get new customers by cross-selling products to customers of its subsidiary PT XYZ who also use internet fixed broadband products, but not products from PT ABC by using the machine learning and the k-means clustering algorithm with the KNIME tool for the k-means clustering process and the R Programming tool for the process of finding the optimal number of clusters. The result of this algorithm finds that there are four types of PT XYZ customer clusters with the main characteristics from the revenue side. The results of data correlation, showing that cluster 2 and cluster 3 are potential clusters with 2123 customers, also areas that have few customers are area 1 and area 4, then the estimated minimum revenue that will be generated is IDR 8,937,830,000.

Keywords: *Cross-Selling, K-Means Clustering, Machine Learning*

[*] Corresponding Author's Email: dendy.tryanda@ui.ac.id

1. INTRODUCTION

According to a report from the Indonesian Central Bureau of Statistics, internet users in the Indonesian population in the 2016-2021 period have increased both in urban and rural areas (Badan Pusat Statistika, 2021). In fact, it can be seen from Figure 1, until 2026 internet users in Indonesia is projected to increase.

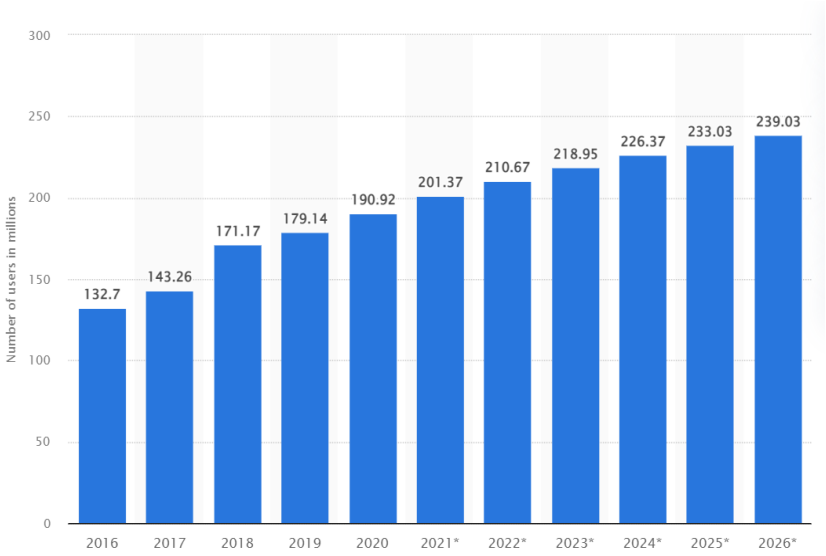


Figure 1. Internet User Forecast in Indonesia

This increase is due to the success that has been done by the government and the public who work together for the process of developing telecommunications network infrastructure in Indonesia. The availability of internet access facilities can be done from home using a wired telephone network or from outside the home such as offices, cafes and other places. Retrieved from the press release on the Kominfo website No. 366/HM/KOMINFO/10/2021 on October 12, 2021, it was stated that internet users in Indonesia as of January 2021 reached 202.6 million people (Biro Humas Kementerian Kominfo, 2021).

The COVID-19 pandemic that has hit Indonesia since early 2020 has had a lot of impact on the Indonesian economy. Many things have been done by the government to prevent the spread and transmission of COVID-19, ranging from the PSBB (Large-Scale Social Restrictions) policy to PPKM (Enforcement of Restrictions on Community Activities). The policy implemented by this government makes several sectors have to carry out activities at home such as teaching and learning activities, seminars, religious activities and even restaurants are asked not to serve the dine-in process during this policy to avoid high mobility and meeting activities, even for certain sector workers it is recommended to work online or Work From Home (WFH) (PP RI No. 21, 2020).

Telecommunications companies are also competing to add capacity and improve their network infrastructure at various points to support the policies set by the government. There are two alternative types of internet connections that can be used during the pandemic, namely fixed broadband internet (internet using wired networks)

and mobile broadband (internet without wired networks). These two alternatives are very different in terms of the benefits obtained, such as if using fixed broadband internet, the quality of the resulting network tends to be stable and can be used by many people. However, on the other hand, it cannot be denied that mobile broadband users are very numerous in Indonesia. Based on data from The World Bank, Indonesians who subscribe to fixed broadband internet in 2020 only amounted to 11,722,218 subscribers or around 4% of the total population of Indonesia in 2020 (The World Bank, 2022) this maybe due to the low ODP (Optical Distribution Point) in Indonesia.

Currently in Indonesia, there are seven brands of fixed broadband internet products, one of which is the product of PT ABC which is a telecommunications network and service provider company. In addition to its fixed broadband internet products, PT ABC also has a subsidiary, PT XYZ with 65% share ownership, which is a mobile broadband service provider in Indonesia. PT XYZ itself currently has more than 170 million subscribers throughout Indonesia. To maximize company profits, PT ABC also sells its fixed broadband internet products to customers of PT XYZ. However, not all customers of PT XYZ use products from PT ABC since there are several other brands of fixed broadband internet products in Indonesia that customers may choose.

Therefore, it is necessary for PT XYZ to find a way to identify its customers who use fixed broadband internet products other than PT ABC's products. To acquisition these customers so that PT ABC can get new customer segmentation, PT ABC can apply machine learning with the k-means clustering algorithm using the KNIME tool for the k-means clustering process and the R Programming tool for the process of finding the optimal number of clusters. This acquisition process is necessary because in addition to PT XYZ being a subsidiary of PT ABC, PT ABC's fixed broadband internet products have better internet quality and have a wide network scope and selection of various internet packages.

The k-means clustering algorithm is not a perfect data management because each method has its own advantages and disadvantages, but if the customer segmentation and cross-selling process is to be carried out, the application of the k-means clustering algorithm is the right choice because it can recognize customers better based on their characteristics so that can provide an appropriate approach (Dqlab, 2022). In addition, the k-means clustering method is easy to implement and does not require a long time and is flexible, k-means clustering is also efficient for large data. Basically, clustering is a process that produces a partition for a particular data set that matches the criteria of the clustering without prior knowledge of the data set (Ahmed and Islam, 2020; Dias and Fernandes, 2020). The k-means clustering algorithm is an algorithm that divides a data set into a predefined set of k clusters and no overlap occurs, so each data point has its own group (Alibuhtto and Mahat, 2020). The results of this research are expected to be utilized by companies to get new fixed broadband internet customers and increase company profits.

2. LITERATURE REVIEW

2.1. MACHINE LEARNING

Machine learning was first introduced by Arthur Samuel in 1959 when he worked at IBM, the term was introduced to describe the delivery pattern of "learning" in Artificial Intelligence (AI) (Gogas and Papadimitriou, 2021). Machine learning is a part of artificial intelligence (AI) and computer science. Machine learning uses software applications to predict data more accurately and uses algorithms to predict values. Machine learning also creates and provides data and information in the form of observations. Machine learning itself is divided into four categories, namely supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

Basically, machine learning will provide identification related to the function where an accurate result will be produced beyond expectations (Ho, et al., 2021). Machine learning consists of three components (UC Berkeley, 2022), namely:

1. Decision process: machine learning will make a prediction or classification based on the data we input.
2. Error function: serves to evaluate the prediction or classification produced and see the accuracy of a model.
3. Optimization process: is the process of updating until the final decision of a model is obtained.

Machine learning is needed for several reasons, namely for data scale where companies are currently faced with a lot of very large data variations, therefore machine learning is needed so that processing the data can be more efficient and the resulting data is accurate. Then, another reason is that with machine learning, companies can find unexpected findings from the data analyzed by machine learning.

2.2. UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning for unlabeled data, when it has input variables but without corresponding output variables. Unsupervised learning uses algorithms that are useful for identifying, analyzing, scanning unlabeled data and looking for relationships from the data, these algorithms will find patterns or groupings of data without the need for help from users. This shows that unsupervised learning is an ideal solution for us to do exploratory data analysis, create cross-selling strategies, and determine customer segmentation.

The purpose of using unsupervised learning is get to know know more about the data how the distribution between each data. Unsupervised learning itself is divided into two (Manne, 2021), namely:

1. Clustering: clustering is a grouping of data based on certain patterns or behavior.
2. Association: association is a way to find rules that explain the data.

2.3. CLUSTERING

Clustering is the most widely studied method in unsupervised learning classification (Bindhu and Ranganathan, 2021) because this method reveals a property of data through

a learning of unlabeled data points. This method uses the technique of separating data points into certain clusters so that later data points in this cluster will have similar properties to each other compared to data points outside the separated cluster (Shrikant, et al., 2022).

This clustering method is a common technique for statistical data analysis and is usually applied to non-positional features especially in unsupervised learning classification (Pera, María Sol, et al., 2021). In its rules, this clustering method is divided into four categories (Shi, Congming, et al., 2021; Starczewski, Artur, et al., 2021), namely:

1. Partition-based clustering e.g. K-Means method and K-Medoids method.
2. Hierarchy-based clustering e.g. Divisive method and Agglomerative method
3. Density-based clustering e.g. DBSCAN method
4. Grid-based clustering e.g. Clique method

Basically, this clustering method is the most efficient method used and promising in the current era because the method is easy to implement (Farhan, Naureen, et al., 2021).

In its application, there are several things that are required and must be considered in using the clustering method, namely:

1. Scalability of large amounts of data and data with high dimensions
2. Ability to analyze various forms of data
3. There are clusters in unexpected forms
4. Ability to handle outliers and noise
5. Interpretation and usability

2.4. K-MEANS CLUSTERING

K-means clustering is a widely used method for clustering in unsupervised learning algorithms (Patil, Saurabh, et al., 2021). This method is widely used by various applications because it is simple to implement and effective and this method is suitable for large data sets with high feature dimensions and this algorithm is very efficient (Awad, Fouad H., and Murtadha M. Hamad, 2022) (Wu, Chunqiong, et al., 2021). In this k-means clustering method, an observation will be made which will later form a classification of various groups, each of which is similar to each other and each group/ cluster is represented by a center called a centroid which is obtained from the average value of group/ cluster observations.

In determining the optimal number of groups / clusters in using the k-means clustering method, it can be done with several approaches as follows:

1. Elbow Method
2. Average Silhouette Method
3. Gap Statistical Method

To implement the k-means clustering method, several steps need to be taken as follows (Cohn, et al., 2021) (Hutagalung, et al., 2021):

a. Determine the number of clusters (k) in the data;
b. Determine the centroid value of k;
c. Associate each centroid that is closest and calculate the distance;

d. Update the position of each centroid to be the average position of the closest and associated points;
e. Repeat steps a and b until the centroid is optimized.

2.5. DATA QUALITY

In maximizing the quality of data, it is necessary to apply relevant controls and practices. A data can be said to be of poor quality if it is manifested in missing data, duplicate data, has highly correlated variables, outliers and has a large number of variables it also related to accounting information for auditor. If this poor quality data is still run, it will cause a problem to build and apply a new model (Gudivada, et al., 2017). Data quality analysis consists of four functional categories (Evans, 2020) (Gudivada, et al., 2017), namely:

1. Descriptive analysis is an analysis in which the data we have will be categorized, consolidated and classified and then converted into information that aims to understand and analyze the performance of a business.
2. Diagnostic analysis is an analysis that will help us determine the causes of missing, incomplete, duplicate and inconsistent data factors that cause poor data quality.
3. Predictive analysis is an analysis that can predict and detect a pattern or relationship in data by referring to historical data.
4. Prescriptive analysis is an analysis that helps us to identify the best alternative to minimize errors and maximize the purpose of using data.

2.6. CROSS-SELLING

Cross-selling is an old term that often encounter in everyday life, especially for those who work in marketing. Basically, cross-selling is an activity that involves selling multiple products of various types and brands to the same customer or arguably selling an additional product or service to customers who have already purchased a product or service from the company (Özdemir and Bayrakli, 2022). The purpose of cross-selling is so that companies can make other sales in addition to the products or services that customers are likely to buy, besides that cross-selling reduces the company's costs rather than having to acquire new customers (Özdemir and Bayrakli, 2022).

2.7. PREVIOUS RESEARCH

Research using the k-means clustering algorithm on telecommunication service customers has been carried out, among others, can be seen from Table 1 below:

Table 1. List of Previous Research

No	Title	Author	Year	Result
1	*Target Selection of Customers for Communication Service*	Md Iqbal	2020	The use of k-means clustering has successfully created computational

No	Title	Author	Year	Result
	Providers using K-Means Clustering and Back propagation Artificial Neural Network			models for telecommunication systems using data mining and soft computing techniques. Data mining techniques were used to group telecom subscribers into various sectors based on similarities in network usage. Historical call usage data of various customers is used to categorize them into five clusters referring to five different telecom service schemes using K-means clustering algorithm.
2	*Data Mining for Marketing in Telecommunication Industry*	Rokhmatul Insani, Hira Laksmiwati Soemitro	2016	K-Means Clustering is more effective for related cases because the results can identify profitable customers and useful for companies to provide other offers to their customers.

3. RESEARCH METHODS

This research aims to group potential customers of PT XYZ fixed broadband internet users who are potential to be offered products owned by PT ABC with 43,873 data. The data is divided into 4 areas, namely Sumatra, Jabodetabek Jabar, Java Bali and Pamasuka (Papua, Maluku, Sulawesi, Kalimantan). The initial data from this research will be processed first in order to produce quality data. Then, after that proceed with processing the data into clustering using the k-means algorithm as shown in Figure 2. In processing data into clustering, it will be assisted by research tools, namely KNIME (Konstanz Information Miner) software version 4.6.3 and R Programming version R-4.2.2. The results of this clustering are then stored in the database for PT ABC to use in starting the cross-selling process for its products.

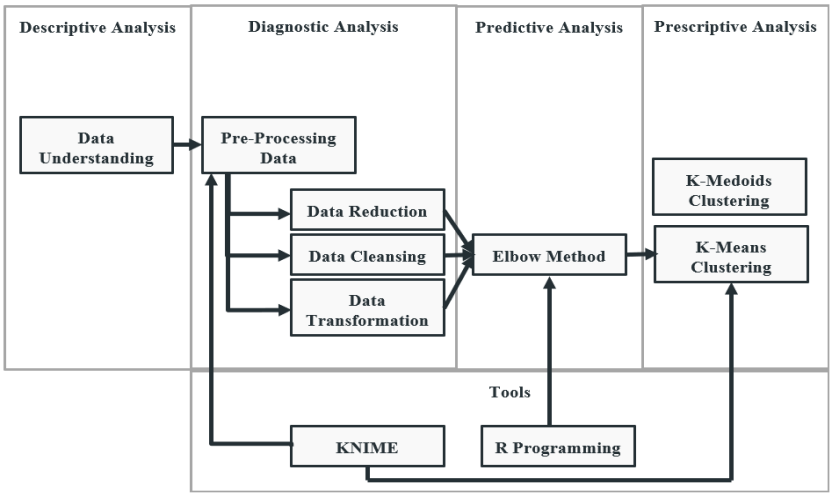


Figure 2. Research Framework

Data Quality Analysis Process

In the application of data analysis to produce quality data, the analysis process starts from descriptive analysis, diagnostic analysis, predictive analysis to prescriptive analysis because the research start from looking at data quality to finally using the data for certain purposes for the company.

Data Understanding

This stage is a process for us to understand the characteristics of the data to be analyzed and helps us determine the data attributes that can be used in the data analysis process.

Data Reduction

In this process, researchers reduce or reduce data attributes that are not needed in the research process. Usually the reduced data is data that has a relationship with each other so that it does not need to be used and can be represented by other data. Then, irrelevant data in research that has no relationship and is related to the research being conducted.

Data Cleansing

In this process, researchers will clean invalid data, then data that has duplicates and blanks and negative data. This process is very important in improving the efficiency of the results of a data analysis because it helps to improve data quality.

Data Transformation

In this process, because researchers use the k-means clustering algorithm, the data processed must be in the form of numbers, so this process is needed to change attribute data that was previously not a number into a numeric value so that the data can be managed.

K-Means Clustering

The initial stage of implementing the k-means clustering algorithm is to determine k centroids that refer to the desired number of clusters as in Figure 3. In the end, each cluster will have a centroid point which is the mean of the cluster. Then, in order to

determine the number of clusters to be optimal, it is necessary to calculate using the curve of the elbow method.

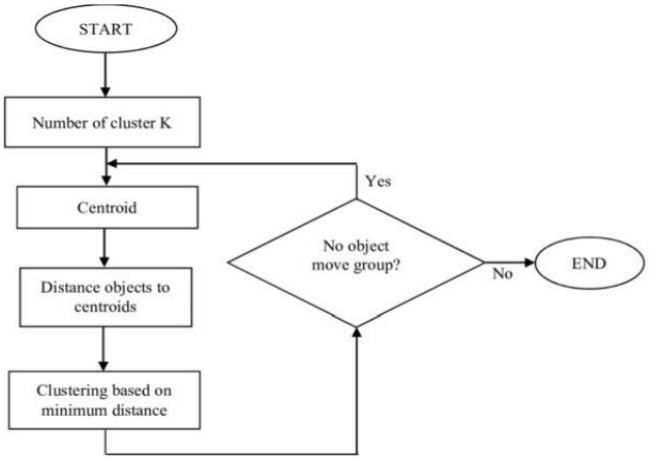


Figure 3. K-Means Clustering Process

R Programming and KNIME

The KNIME program is used in the first stage with several stages of using the nodes that will be used as follows:

1. Using the CSV Reader node to read the dataset
2. Using the Column Filter node for the data reduction process
3. Using the Missing Value node to look for missing data and the Numeric Outlier node to see if there is data that has characteristics that are significantly different from the existing data, these two nodes are part of the data cleansing process.

After that, the R program will be used to determine the optimal number of clusters for the research conducted. The results of the R program were re-entered into the KNIME program to be processed in the k-Means Clustering node.

4. ORGANIZATION PROFILE

The object of this research is PT ABC and PT XYZ. PT ABC is a public company. PT ABC is the largest telecommunications company in Indonesia with 11 subsidiaries with direct ownership and actively operating, 25 subsidiaries with indirect ownership and 9 affiliated entities. PT XYZ itself is one of the 11 subsidiaries with direct ownership and actively operating owned by PT ABC. To create and increase value for customers, PT ABC organizes its business based on customer segments or Customer Facing Units (CFU). Therefore, PT ABC categorizes its product portfolio into five segments, namely Consumer, Mobile, Enterprise, Wholesale and International Business, and other segments.

PT XYZ is a subsidiary of PT ABC with 65% share ownership and a telecommunications company from Singapore with 35% share ownership. PT XYZ is a company engaged in telecommunications with the widest network coverage in Indonesia which currently operates more than 251,116 BTS serving around 176 million

customers spread across Indonesia from remote, outermost islands to even the country's border zones with a total of 5535 employees.

5. RESULT AND DISCUSSION

The first step in starting research is to collect and identify and standardize the data format. In this study, the data used is data on PT XYZ customers who use fixed broadband internet products in 2021 which consists of 43,873 databases and the data used is data in the form of .csv files.

Data Understanding

In appendix 1, the attributes of PT XYZ customer data consisted of 9 character data types and 9 numeric data types.

Data Reduction

In this data reduction process, researchers reduced some attributes that were not needed. The data was reduced from 18 attributes to 13 attributes. Researchers did not include the dominant_recharge_channel, first_rank_category and first_rank_apps attributes because they were not relevant to the research to be carried out because they contained data on the use of applications that were often used by customers. Researchers also do not use the region_lacci and cluster_lacci attributes because they are adequately represented by the district and area_lacci attributes which cover a wide range.

Data Cleansing

In this data cleansing process, researchers cleaned the data that was invalid and had duplicates and blank or negative values. This process resulted in a reduction in data rows, which previously had 43,873 data reduced to 43,819 data.

Data Transformation

This process is the process of converting attribute data that was previously not a number into a numeric value, but for this data there is no need for this data transformation process because the existing data is appropriate, namely only numeric data.

Elbow Method

The elbow method is used to determine the best k cluster value owned by the researcher's data so that it can be continued for the k-means clustering process. If seen from Figure 4, using R Programming it is found that the best k cluster values are in the four clusters.

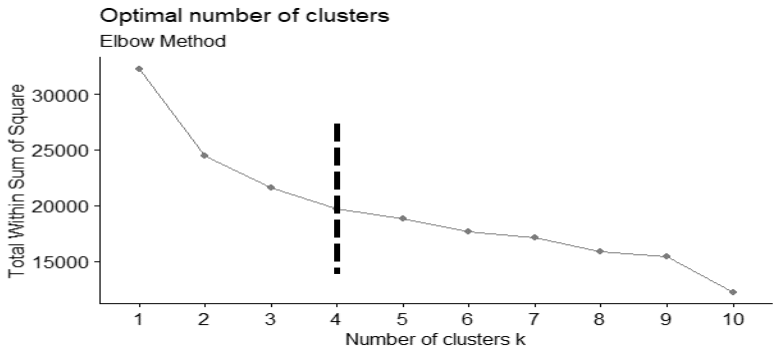
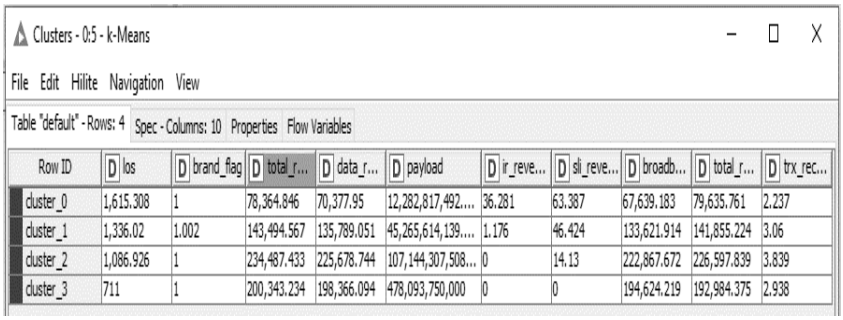


Figure 4. Number of Clusters

K-Means Clustering

After the data is processed in the k-Means node using KNIME, the output in the form of a table of attribute results from each cluster as shown in Figure 5.



Clusters - 0:5 - k-Means

File Edit Hilite Navigation View

Table "default" - Rows: 4 Spec - Columns: 10 Properties Flow Variables

Row ID	los	brand_flag	total_r...	data_r...	payload	ir_reve...	sli_reve...	broadb...	total_r...	trx_rec...
cluster_0	1,615.308	1	78,364.846	70,377.95	12,282,817,492....	36.281	63.387	67,639.183	79,635.761	2.237
cluster_1	1,336.02	1.002	143,494.567	135,789.051	45,265,614,139....	1.176	46.424	133,621.914	141,855.224	3.06
cluster_2	1,086.926	1	234,487.433	225,678.744	107,144,307,508...	0	14.13	222,867.672	226,597.839	3.839
cluster_3	711	1	200,343.234	198,366.094	478,093,750,000	0	0	194,624.219	192,984.375	2.938

Figure 5. Output K-Means Clustering

Correlation of Clustering Data Results

Before categorizing customers based on their characteristics, looking the relationship between each data is a must for determine the main characteristics that use for this research. The correlation results show that the revenue variable side has a correlation with each other between the data where the resulting correlation value is> 0.7 which means the correlation between the variables is very strong.

Customer Category Based on Cluster

After mapping customers based on the amount of revenue as shown in Table 2, it is found that cluster 0 and cluster 1 can be categorized as non-prospect users because they generate low revenue. Meanwhile, cluster 2 and cluster 3 are customers with the prospect user category because they generate large revenue.

Table 2. Customer Segment Characteristics

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**
los (months)	**1615**	1336	1086	**711**

total_revenue (IDR Bn)	78.4	143.5	**234.5**	200.3
data_revenue (IDR, 000)	70.4	135.8	**225.7**	198.4
payload (GB)	**12**	45	107	**478**
ir_revenue (IDR)	**36.2**	1.2	0	0
sli_revenue (IDR)	63.4	**46.4**	14.1	0
broadband_package_ revenue (IDR, 000)	67.6	133.6	**222.9**	194.6
total_recharge (IDR, 000)	79.6	141.8	**226.4**	193
trx_recharge	2.2	3.06	**3.8**	2.9
Kategori	Non – Potential Users	Non – Potential Users	Potential Users	Potential Users

Potential Cluster Analysis Based on Regional Area

If seen from Table 3 which is data on the number of potential customers using fixed broadband internet products that can be offered by PT ABC products, it can be seen that based on the four areas, it can be seen that Area 1 (Sumatra) and Area 4 (Pamasuka - Papua, Maluku, Sulawesi, Kalimantan) have the fewest customers using fixed broadband internet products, there are many factors that determine this, one of which is inadequate networks and facilities for these areas.

Table 3. Number of Potential Customers Based on Area

	Cluster 2	**Cluster 3**
AREA 1	314	7
AREA 2	893	25
AREA 3	613	24
AREA 4	239	8

Potential Cluster Revenue Analysis

If calculating the estimated revenue that the company will get, an example of a customer buying a regular bundling package only for the internet with a subscription fee of Rp 330,000 / month and an installation fee of Rp 250,000, then the estimated revenue that can be obtained by the company in one year with a total of 2123 customers that can be offered is around Rp 8,937,830,000 excluding other costs that must be incurred by the company.

6. CONCLUSION AND RECOMMENDATION

The application of k-means clustering is sufficient to describe the characteristics of customers who can be offered fixed broadband internet products from PT ABC. From the results of this study, it was found that the number of customer groupings of fixed broadband internet users who could be offered PT ABC products were spread in 4 clusters. Then, the correlation results show that the main characteristics of customers can be seen in terms of revenue generated. After the analysis, it is found that cluster 2 and cluster 3 are potential customers to be offered PT ABC's fixed broadband internet products because they generate large revenue with a total of 2123 customers with an estimated minimum revenue of around Rp 8,937,830,000.

Therefore, the recommendation that can be given for the short term is that companies can carry out massive marketing campaigns such as year-end promos, beautiful date promos, and other discounts. Companies can also do bundling promos, for example, if you buy a product from PT XYZ, you will get points that can be used to buy fixed broadband internet products owned by PT ABC. To carry out this marketing campaign, the company can use the SMS channel which will be broadcast to customers at certain times.

For long-term recommendations, the company must check and expand ODP (Optical Distribution Point) in all areas, especially areas with very low subscribers so that they can convince customers that fixed broadband internet products from PT ABC can be used in the area and there are no obstacles, but also the company must have completed the regulatory process in the area before expanding ODP so that everything is done with clear rules.

REFERENCES

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

Alibuhtto, M. C., & Mahat, N. I. (2020). Distance based k-means clustering algorithm for determining number of clusters for high dimensional data. Decision Science Letters, 9(1), 51–58. https://doi.org/10.5267/j.dsl.2019.8.002

Awad, Fouad H., and Murtadha M. Hamad. "Improved k-Means Clustering Algorithm for Big Data Based on Distributed SmartphoneNeural Engine Processor." *Electronics* 11.6 (2022): 883.

Bindhu, V., and G. Ranganathan. "Hyperspectral Image Processing in Internet of Things model using Clustering Algorithm." *Journal of ISMAC* 3.02 (2021): 163-175.

Cohn, Ryan, and Elizabeth Holm. "Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data." *Integrating Materials and Manufacturing Innovation* 10.2 (2021): 231-244.

Dias, L. A., Ferreira, J. C., & Fernandes, M. A. (2020). Parallel implementation of K-means algorithm on FPGA. *IEEE Access*, *8*, 41071-41084.

Farhan, Naureen, et al. "Clustering Approaches for Efficient Radio Resource Management in Heterogeneous Networks." (2021).

Gogas, P., Papadimitriou, T. Machine Learning in Economics and Finance. *Comput Econ* 57, 1–4 (2021). https://doi.org/10.1007/s10614-021-10094-w

Gudivada, Venkat, Amy Apon, and Junhua Ding. "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations." *International Journal on Advances in Software* 10.1 (2017): 1-20.

Ho, Winky KO, Bo-Sin Tang, and Siu Wai Wong. "Predicting property prices with machine learning algorithms." *Journal of Property Research* 38.1 (2021): 48-70.

Hutagalung, Juniar, et al. "Covid-19 cases and deaths in southeast Asia clustering using k-means algorithm." *Journal of Physics: Conference Series*. Vol. 1783. No. 1. IOP Publishing, 2021.

Insani, Rokhmatul, and Hira Laksmiwati Soemitro. "Data mining for marketing in telecommunication industry." 2016 IEEE Region 10 Symposium (TENSYMP). IEEE, 2016.

Iqbal, Md. "Target Selection of Customers for Communication Service Providers using K-Means Clustering and Back propagation Artificial Neural Network."

Manne, Ravi. "Machine learning techniques in drug discovery and development." *International Journal of Applied Research* 7.4 (2021): 21-28.

ÖZDEMİR, Yunus Emre, and Selim Bayrakli. "A Case Study on Building a Cross-Selling Model through Machine Learning in the Insurance Industry." *Avrupa Bilim ve Teknoloji Dergisi* 35 (2022): 364-372.

Patil, Saurabh, et al. "Study of Customer Segmentation Using K-Means Clustering and RFM Modelling." *Journal of Engineering Sciences* 12.6 (2021): 556-559.

Pera, María Sol, et al. "pyUPMASK: an improved unsupervised clustering algorithm." *arXiv preprint arXiv:2101.01660* (2021).

Peraturan Pemerintah Republik Indonesia Nomor 21 tahun 2020 tentang Pembatasan Sosial Berskala Besar Dalam Rangka Percepatan Penanganan Corona Virus Disease 2019 (COVID-19).

Shi, Congming, et al. "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm." *EURASIP Journal on Wireless Communications and Networking* 2021.1 (2021): 1-16.

Shrikant, Khyaati, et al. "A Comparative Study of Clustering Algorithm." *Intelligent Computing and Networking*. Springer, Singapore, 2022. 219-235.

Starczewski, Artur, et al. "A novel grid-based clustering algorithm." *Journal of Artificial Intelligence and Soft Computing Research* 11 (2021).

Website Berkeley Education - https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/

Website BPS Statistik Telekomunikasi Indonesia 2021 - https://www.bps.go.id/publication/2022/09/07/bcc820e694c537ed3ec131b9/statistik-telekomunikasi-indonesia-2021.html

Website Kominfo - https://www.kominfo.go.id/content/detail/37471/siaran-pers-no-366hmkominfo102021-tentang-tren-pendanaan-startup-meningkat-menkominfo-berharap-indonesia-tambah-satu-decacorn/0/siaran_pers#:~:text=%E2%80%9CJumlah%20pengguna%20internet%20yang%20mencapai,Covid%2D19%2C%E2%80%9D%20jelasnya

Webiste World Bank - https://data.worldbank.org/indicator/IT.NET.BBND?locations=ID

Wu, Chunqiong, et al. "k-Means Clustering Algorithm and Its Simulation Based on Distributed Computing Platform." *Complexity* 2021 (2021).

APPENDIX

Appendix 1. Described Picture in Introduction

No	Attribute Name	Description	Data Type
1	apps	Fixed broadband internet product used	Character
2	los	Length of subscription time	Numeric
3	brand_flag	Customer type (prepaid and postpaid)	Character
4	total_revenue	Total overall revenue	Numeric
5	data_revenue	Total data revenue	Numeric
6	payload	Data payload	Numeric
7	ir_revenue	Total internet roaming revenue	Numeric
8	sli_revenue	Total SLI revenue	Numeric
9	broadband_package_revenue	Total broadband package revenue	Numeric
10	total_recharge	Total recharge	Numeric
11	trx_recharge	Total recharge transaction	Numeric
12	dominan_recharge_channel	Channel / Apps used to recharge	Character
13	first_rank_category	Frequently accessed application categories	Character
14	first_rank_apps	Applications that are often accessed	Character
15	kabupaten	Customer's district location	Character
16	region_lacci	Region Location Area Code Cell ID	Character
17	area_lacci	Area Location Area Code Cell ID	Character
18	cluster_lacci	Cluster Location Area Code Cell ID	Character

Appendix 2. k-Means Result Data of Fixed Broadband Internet Product Users

	los	total_ revenue	data_ revenue	payload	ir_ revenue	sli_ revenue	broadband_ package_ revenue	total_ recharge	trx_ recharge
los	**1**	0,072	0,036	-0,082	-0,003	0,007	0,03	0,072	-0,07
total_ revenue	0,072	**1**	**1**	0,441	0,033	0,103	**1**	**1**	0,26
data_ revenue	0,036	**1**	**1**	0,511	-2,762	0,033	**1**	**0,8**	0,223
payload	-0,082	0,441	0,511	**1**	-0,006	-0,002	0,521	0,37	0,141
ir_revenue	-0,003	0,033	-2,762	-0,006	**1**	-9,031	-5,42	0,03	0,001
sli_revenue	0,007	0,103	0,033	-0,002	-9,031	**1**	0,02	0,106	0,013
broadband _package _revenue	0,03	**1**	**1**	0,521	-5,42	0,02	**1**	**0,8**	0,21
total_ recharge	0,072	**1**	**0,8**	0,37	0,03	0,11	**0,8**	**1**	0,32
trx_ recharge	-0,07	0,261	0,223	0,141	0,001	0,013	0,21	0,32	**1**

Appendix 3. k-Means Result Data of Fixed Broadband Internet Product Users

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Grand Total**
BH	**7085**	**3061**	**490**	**8**	**10644**
AREA 1	999	541	68	-	1608
AREA 2	2625	1114	207	5	**3951**
AREA 3	2708	1053	170	3	3934
AREA 4	753	353	45	-	1151
CB	**2028**	**767**	**207**	**34**	**3036**
AREA 1	269	154	31	6	460
AREA 2	766	268	73	11	1118

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Grand Total
AREA 3	854	279	77	12	**1222**
AREA 4	139	66	26	5	236
FM	**13538**	**6128**	**880**	**10**	**20556**
AREA 1	1450	1096	144	1	2691
AREA 2	7116	2653	393	3	**10165**
AREA 3	4083	1771	242	5	6101
AREA 4	889	608	101	1	1599
MP	**1538**	**504**	**89**	**3**	**2134**
AREA 1	159	80	8	-	247
AREA 2	716	204	46	2	**968**
AREA 3	544	158	23	1	726
AREA 4	119	62	12	-	193
RI	**3955**	**2016**	**343**	**5**	**6319**
AREA 1	668	376	58	-	1102
AREA 2	1723	753	161	3	**2640**
AREA 3	1185	573	83	2	1843
AREA 4	379	314	41	-	734
XH	**797**	**279**	**50**	**4**	**1130**
AREA 1	57	36	5	-	98
AREA 2	348	107	13	1	**469**
AREA 3	216	62	18	1	297
AREA 4	176	74	14	2	266
Grand Total	**28941**	**12755**	**2059**	**64**	**43819**