SAGA: Journal of Technology and Information Systems
Vol 2, Issue 3, August 2024, Page 295-303
ISSN: 2985-8933 (Media Online)
DOI: 10.58905/SAGA.v2i3.357

Clustering of Generation X and Generation Y Communities in Cybersecurity Using the K-Means Algorithm (Case Study of Depok City, West Java)

Adhitya Eka Wibowo[1], Agung Triayudi[2*]

Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional, Jakarta, Indonesia

Author Email: ekaadit7@gmail.com[1], agungtriayudi@civitas.unas.ac.id[2*]

Abstract. This research aims to explore the differences in understanding and awareness of cybersecurity between Generation X and Generation Y in Depok City, West Java. The K-Means algorithm is used to group communities based on characteristics relevant to cybersecurity. The results of the study show that there are significant differences in understanding and behavior related to cybersecurity between the two generations. Generation X tends to be more cautious in using technology and has a better knowledge of cybersecurity risks, while Generation Y is more proficient in using digital devices and applications but pays less attention to security aspects. Factors that affect the level of cybersecurity awareness in both generation groups include knowledge of cyber threats, education, and demographic factors. The findings of this research can help stakeholders in increasing awareness and knowledge about cybersecurity and developing better solutions to protect users from cyber threats.

Keyword: Clustering, Generation X, Generation Y, Communities, Cybersecurity, K-Means

1 Introduction

Progress is changing people's lives very quickly. One of the most rapid advances is communication and information. The Internet is one example of the development of technology and information, the development of technology and information. Providing convenience such as information communication, transactions, education, and entertainment. Although there are many conveniences offered by the development of the internet, there are also threats lurking on users, such as data theft, hacking, and malware attacks. So, it is hoped that internet users can protect their data from these threats with Cyber Security Cyber security has become a very important issue in the ever-evolving digital era[1]. Cybersecurity is a mechanism to detect computer security vulnerabilities, prevent the threat of cybercrime and recover digital devices that have been affected by cyberattacks.

Based on a publication published by the Central Statistics Agency (BPS), Internet users in Indonesia as of January 2024 are 66.48% or 185 million people out of the total population of Indonesia which amounts to 278.9 million people[2]. Of this figure, it is recorded that it is dominated by community groups aged 25-49 years. A very important generation group in modern society is Generation X, which is the generation whose people were born in the 1965-1980 range. And the Millennial Generation consists of community groups born in the 1981-1996 range. The two generations have different views on the use of technology and the internet. Generation X is transitioning from a non-digital era to a digital era, while Generation Y is growing up with digital technology.

However, Cyber Attacks such as Phishing, Malware, and other hacking attacks do not introduce users to what generation group and what age, so any unwary user can be exposed to cyberattacks. However, there are differences in responding to these cyber attacks, which raises questions about the factors that influence their behavior towards cybersecurity. Whether these factors include knowledge and awareness of cyber threats; previous experience with cybersecurity incidents; cybersecurity education and training, as well as views and perceptions on cyber risks[3].

The purpose of this study is to find the characteristics of the Generation X and Generation Y communities in terms of cybersecurity, using the K-Means Algorithm to group the communities and analyze the factors that affect the actions of the communities of the two generations in terms of cybersecurity. The use of the K-Means Algorithm for community grouping based on relevant characteristics to understand the differences and similarities between Generation X and Generation Y in terms of Cybersecurity. The K-Means algorithm is a good clustering method for finding patterns and groups in big data. It is hoped that by bringing these communities together, patterns of behavior will be found that can provide deeper insights into how each generation faces cybersecurity challenges.

2 Methodology

This study uses a descriptive and quantitative approach. According to Sugiyono (2016) stated that descriptive research is a type of research that aims to identify the relationship between two or more variables. Descriptive research is used to explain how Generation X and Generation Y have cybersecurity awareness. Quantitative methods are used to obtain data about a particular population or sample by conducting data analysis using statistical tools and research tools. The purpose of data collection is to test predetermined hypotheses and group data using the K-Means Algorithm. To ensure that the results of the study do not deviate from the original goal, the author uses this research path as a guideline in carrying out this research.

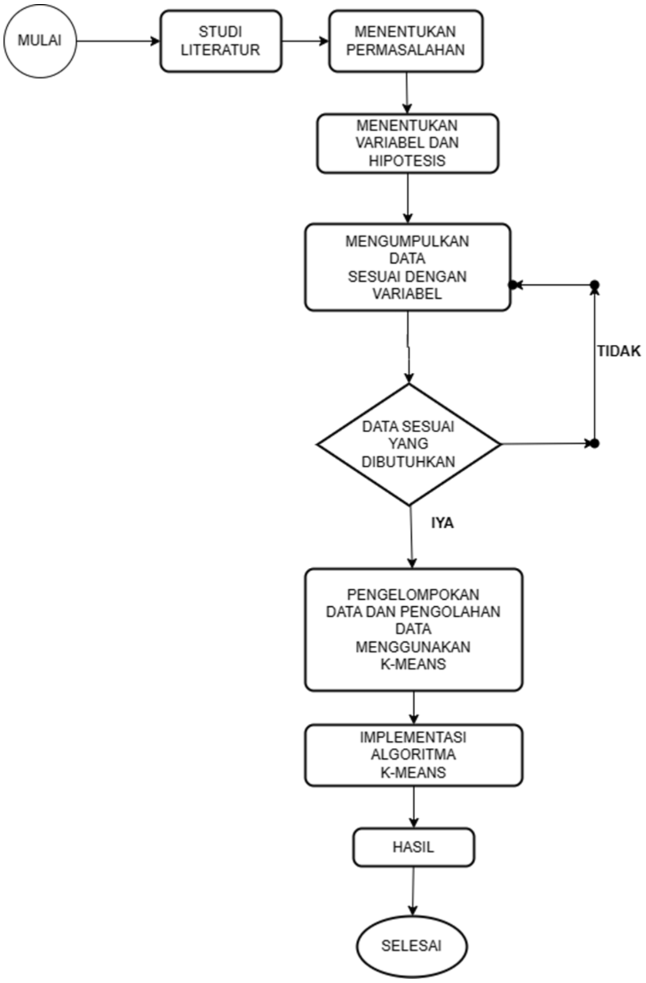
MULAI
STUDI LITERATUR
MENENTUKAN PERMASALAHAN
MENENTUKAN VARIABEL DAN HIPOTESIS
MENGUMPULKAN DATA SESUAI DENGAN VARIABEL
TIDAK
DATA SESUAI YANG DIBUTUHKAN
IYA
PENGELOMPOKAN DATA DAN PENGOLAHAN DATA MENGGUNAKAN K-MEANS
IMPLEMENTASI ALGORITMA K-MEANS
HASIL
SELESAI

Figure 1. Research Flow

2.1 Respondent Datasets/Data Collection

The data taken is the result of a questionnaire that has been distributed to the people of Depok City with the provision that people in the age range of 28 - 59 years. The results of the data that have been shared, are then collected into raw data that will be processed in the next method. The data collected was 200 respondents from several Depok City communities. In data collection, there are several variables that support the process of filling out the Independent Variable questionnaire (Age, Gender, Last Education, Income and Job Sector), while the Dependent Variable (Awareness of Cyber Security)

Table 1. Operational Definition of Independent Variable Demographic Factors

No	Variable	Description
1	Age	28 - 59 Age
2	Gender	Male and Female
3	Last Education	Primary Education (SD), Secondary Education (SMP, SMA), Higher Education (DIII, DIV/S1, S2, S3)
4	Income	Low Income (<5 Million/Month) and High Income (>5 Million/Month)
5	Work	Working or Not Working

Table 2. Definition of Dependent Variable Awareness Factor

Variable	Symbol	Indicators
Cyber Security Knowledge	P-1	I know to ignore messages (SMS, WA) that contain unknown or suspicious links
	P-2	I know to protect my data from cybercrime
	P-3	I know not to give personal information to unknown people (Email, Password, Username)
	P-4	I know Two-Factor Authentication provides double protection in protecting accounts from breaches
	P-5	I know not to share OTP (One-Time Password) codes
	P-6	I know not to access bank accounts (M-banking or I-banking) when using public networks (Wi-Fi)
	P-7	I know how to regularly update apps on my device to improve device security
	P-8	I know regularly update apps on devices to improve device security
	P-9	I know to limit the personal information I share online, such as home address, phone number
	P-10	I know to turn off the internet connection or Wifi when not in use to improve the security of my device

Table 3. Definition of Independent Variable Awareness Factor

Variable	Symbol	Indicators
Cyber Security Awareness	K-1	I am aware to ignore messages (SMS, WA) that contain unknown or suspicious links
	K-2	I am aware of protecting my data from cybercrime
	K-3	I am aware not to give personal information to unknown persons (Email, Password, Username)
	K-4	I am aware that Two-Factor Authentication provides double protection for accounts from breaches
	K-5	I am aware not to share the OTP (One-Time Password) code
	K-6	I am aware not to access my bank account (M-Banking or I-Banking) when using a public network (Wi-Fi)
	K-7	I am aware of regularly updating the app on the device to improve the security of the device
	K-8	I am conscious of using strong, unique passwords
	K-9	I am aware of limiting the personal information I share online, such as home addresses, phone numbers
	K-10	I am conscious of turning off the internet connection or Wifi when not in use to improve the security of my device

2.2 Cleaning Data

Data cleaning is an important step in data mining research. At this stage, the data that has been collected is evaluated to identify and address issues such as missing, duplicate, inconsistent, or irrelevant data[4]. The data cleansing process involves several steps, including the identification and deletion of empty or incomplete data entries, the handling of outliers that may interfere with analysis, and the merging or separation of redundant or inconsistent data entries. In addition, normalization or standardization of data can also be done to ensure consistency in data format and scale. These steps aim to prepare clean and structured data so that data mining analysis can be carried out appropriately and the results produced become more accurate and meaningful. By conducting careful data cleaning, researchers can minimize bias and optimize the quality of the data used in the research process.

2.2 Clustering K-Means

Clustering is one of the important techniques in data mining that is used to group data into similar subsets based on certain characteristics or patterns[5]. The main purpose of clustering is to identify the hidden structures in the data, which can help in further understanding of the groups or categories that exist within them. The K-Means algorithm is one of the most common and simple clustering algorithms used in data analysis. This algorithm works by grouping data points into clusters, where each data point is incorporated into a group that has a centroids closest to it[6]. The main steps in the K-Means Algorithm are as follows:

a. Specify the number of Clusters (K) on the dataset as the centroid value.
b. Calculating the distance between the data and the center point of the Cluster uses the formula from the Euclidean Distance which can be seen in the Euclidean distance theory which is formulated as follows. $De = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2}$ [7]
c. The new cluster center will be defined when all the data has been assigned in the nearby cluster.
d. The process of determining the center point of the cluster and placing the data in the cluster is repeated continuously until the centroid value does not change again.

3 Results and Discussion

3.1 Dataset

The data used came from data collected by the researcher through questionnaires and 200 random samples modeled using K-Means Clustering.

Table 4. Respondent Dataset

NO	USIA	KESADARAN										PENGETAHUAN									
		K_1	K_2	K_3	K_4	K_5	K_6	K_7	K_8	K_9	K_10	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_10
1	45Tahun	4	4	4	3	4	5	5	3	4	5	5	5	5	4	5	5	3	3	4	4
2	33Tahun	4	5	5	4	5	4	5	4	5	4	4	3	4	4	3	5	4	5	4	5
3	58Tahun	2	3	3	2	3	3	2	4	2	3	4	3	3	3	4	3	3	3	4	3
4	58Tahun	2	3	2	3	2	3	2	2	2	2	2	2	2	1	3	1	1	1	3	1
5	47Tahun	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

3.2 Cleaning Data

At the preprocessing stage, namely by normalizing the data first. Data normalization is a technique that aims to map data to a certain scale in the data mining process. This is important because often the data used in the analysis has different scales, so it can cause problems in comparing or combining data. Therefore, a Validity Test and Reliability Test were carried out to evaluate the results **200**of the questionnaire conducted.

Table 5. Results of Validity Test of Consciousness Variables

Variable	Indicators	R Calculate n =	R Table	Resul
Cyber Awareness	K_1	0,650	0,138	Valid
Security	K_2	0,707	0,138	Valid
	K_3	0,718	0,138	Valid
	K_4	0,841	0,138	Valid
	K_5	0,743	0,138	Valid
	K_6	0,815	0,138	Valid
	K_7	0,965	0,138	Valid
	K_8	0,718	0,138	Valid
	K_9	0,617	0,138	Valid
	K_10	0,854	0,138	Valid

Table 6. Results of the Validity Test of Knowledge Variables

Knowledge Cyber Security	P_1	0,721	0,138	Valid
	P_2	0,759	0,138	Valid
	P_3	0,771	0,138	Valid
	P_4	0,919	0,138	Valid
	P_5	0,863	0,138	Valid
	P_6	1,074	0,138	Valid
	P_7	0,984	0,138	Valid
	P_8	0,785	0,138	Valid
	P_9	0,776	0,138	Valid
	P_10	0,890	0,138	Valid

Table 7. Reliability Test Results of Awareness Variables and Knowledge Variables

TEST CRITERIA			
Variable	Reference Value	Cronbach's Alpha Values	Conclusion
Awareness	0.60	0,947	Reliable
Knowledge	0.60	0,946	Reliable

3.3 K-Means Clustering Algorithm

The K-Means Clustering algorithm is one of the most commonly used clustering methods in data analysis. The main goal is to group the data into a number of different categories or clusters. In this study, there are three Clusters, namely: Low Cybersecurity Awareness Level, Medium Cybersecurity Awareness Level, and High Cybersecurity Awareness Level.

Table 8. Initial Cluster Centers

	Clusters		
	1	2	3
Zscore : I am aware of ignoring messages (SMS, WA, Email) that contain unknown or suspicious links	-2,81483	,90521	-1,57482
Zscore: I am aware of protecting my data from cybercrime	-4,01128	,74283	-1,63423
Zscore: I am aware not to give personal information to strangers (Email, Password, Username, OTP)	-3,94173	,77891	-,40125
Zscore: I know to ignore messages (SMS, WA, Email) that contain unknown or suspicious links	-2,76738	,76544	-1,58977
Zscore: I know to protect my data from cybercrime	-3,77464	,81459	-1,48003
Zscore: I know not to give personal information to strangers (Email, Password, Username, OTP)	-3,72420	,83140	-1,44640

It is the first data clustering process before the data is iterated and this data is the process for the formation of three clusters

Table 9. Iteration Process

Iteration	1	2	3
1	3,517	2,790	3,135
2	,597	,051	,117
3	,477	,026	,112
4	,454	,031	,115
5	,410	,000	,072
6	,351	,000	,066
7	,000	,000	,000

Table 9 shows the iteration process in the group grouping from the initial table, which resulted in 7 iterations. In iterations 1 to 6, there are insignificant centroids, and in iteration number 7 there are significant centroids. Thus, all clusters have been formed, and iteration 7 stops with a minimum distance of 9,954.

Table 10. Final Cluster Centers

	Cluster		
	1	2	3
Zscore: I am aware of ignoring messages (SMS, WA, Email) that contain unknown or suspicious links	-2,29816	,51203	-,54465
Zscore: I am aware of protecting my data from cybercrime	-2,22849	,53025	-,59198
Zscore: I am aware not to give personal information to strangers (Email, Password, Username, OTP)	-2,36819	,56782	-,63729
Zscore: I know to ignore messages (SMS, WA, Email) that contain unknown or suspicious links	-2,37484	,59311	-,68392
Zscore: I know to protect my data from cybercrime	-2,43611	,54408	-,57983
Zscore: I know not to give out private information to strangers (Email, Password, Username, OTP)	-2,20567	,57214	-,67546

Table 10 above shows the results of the final clustering process, which forms a cluster with three clusters for each variable. The variables in the End Cluster center table are the result of normalized values. The characteristics for each cluster are obtained from the calculations generated from each variable. Each cluster group has an explanation, as follows:

a. Cluster 1
   The variable of the number of awareness of cybersecurity and the variable of the amount of knowledge were below average.
b. Cluster 2
   The variable of the amount of awareness of cybersecurity is above average and the variable of the amount of knowledge is below the average.
c. Cluster 3
   The variable of the amount of awareness of cybersecurity is above average, and the variable of the amount of knowledge is above average.

Table 11. Final Results of Clustering

Cluster	Frequency	Percent	Valid Percent	Cumulative Percent
1	12	6,0	6,0	6,0
2	123	61,5	61,5	67,5
3	65	32,5	32.5	100,0
Total	200	100,0	100,0	

Based on the table above, the details of the clustering results were obtained from Cluster 1 (C1) as many as 12 respondents, Cluster 2 (C2) as many as 123 respondents and Cluster 3 (C3) as many as 65 respondents.

4 Conclusion

Once the analysis is complete, the researcher will talk about all demographic elements including age, gender, last education, monthly opinions, and occupation.

According to an analysis conducted by the researchers, based on gender differences do not have an impact on their knowledge factors about cybersecurity, the results show that both men and women have the same level of knowledge and awareness of cybersecurity.

In addition, research on generation factors shows that age can affect a person's level of knowledge about cybersecurity, a significant difference in the number of respondents among cluster 1 (C1) groups shows that generation X is more than generation Y. This can happen because the knowledge that everyone has is different due to different ages, so the experience they can get is also different.

Furthermore, an analysis was conducted to evaluate the relationship between educational factors and respondents' knowledge about cybersecurity. Apparently, the latest educational results have a significant impact. The results of the clustering test showed that there was a difference in the perception of cybersecurity between respondents from Primary Education (SD), Secondary (SMP, SMA), and Higher Education (Diploma, Bachelor, Master). Respondents from primary education do not know how to respond to cybersecurity and do not know how to protect their data on digital systems. Respondents from Secondary Education know how to protect their personal data, but they often neglect it in protecting it.

Furthermore, based on the analysis of how income factors affect the understanding and awareness of cybersecurity, income is divided into two categories: Low Income (below Rp 2.5 Million Per Month) and high income (above Rp 2.5 Million Per Month). The Monthly income cluster in C1 is dominated by people with an income below 2.5 Million Rupiah per Month, while the C2 and C3 Clusters are dominated by people with an income above 2.5 Million Rupiah per Month. The monthly income factor can have an impact on the level of public awareness of cybersecurity.

The analysis further looks at a person's employment factors towards cybersecurity knowledge and awareness. The results of the analysis show that occupational factors affect the level of awareness of safety. The people who fall into the C1 group are dominated by people who do not work as many as 8 respondents while the people who work are only half or 4 respondents and for the C2 and C3 groups are dominated by the working people. C2 as many as 122 respondents and C3 62 respondents. Only this shows that working people know more about cybersecurity than people who don't. However, they don't know much about cybersecurity in depth, this is because their work is not only focused on the field of technology.

So that conclusions are made based on the process carried out from the beginning to the end of the research K-Means Algorithm can be used to form a new group based on public awareness of cybersecurity. In the process of forming clustering or grouping, it was successfully carried out where there were 3 (three) clusters obtained from the data. The results of the K-Means algorithm showed that there were 12 respondents who were below average in cyber. The data included in Cluster 2 there were 123 respondents whose level of awareness of cybersecurity was above average but the level of knowledge about cybersecurity was below average. The data included in Cluster 3 contains 65 respondents whose level of awareness and knowledge of cybersecurity is above average. This proves that demographic factors have a significant impact on Public Knowledge and Awareness of Cyber Security.

References

1. Abdullah, M. S., & Ikasari, I. H. (2023). Perkembangan Terbaru Dalam Keamanan Siber, Ancaman Yang Diidentifikasi Dan Upaya Pencegahan. *JRIIN : Jurnal Riset Informatika Dan Inovasi*, *1*(1), 96–98.
2. Statistik, B. P. (2020). *Hasil Sensus Penduduk 2020*. Badan Pusat Statistik. https://demakkab.bps.go.id/news/2021/01/21/67/hasil-sensus-penduduk-2020.html
3. Budi, E., Wira, D., & Infantono, A. (2021). Strategi Penguatan Cyber Security Guna Mewujudkan Keamanan Nasional di Era Society 5.0. *Prosiding Seminar Nasional Sains Teknologi Dan Inovasi Indonesia (SENASTINDO)*, *3*(November), 223–234. https://doi.org/10.54706/senastindo.v3.2021.141
4. Faran, J., & Triayudi, A. (2024). Penerapan Algoritma K-Means Data Mining untuk Clustering Kinerja Karyawan Koperasi. *KLIK: Kajian Ilmiah Informatika Dan Komputer*, *4*(4), 2096–2108. https://doi.org/10.30865/klik.v4i4.1728
5. Jannah, A. R., Arifianto, D., & Kom, M. (2015). Penerapan Metode Clustering dengan Algoritma K-Means untuk Prediksi Kelulusan Mahasiswa Jurusan Teknik Informatika di Universitas Muhammadiyah Jember. *Jurnal Manajemen Sistem Informasi Dan Teknologi*, *1*(1210651237), 1–10.
6. Nurzahputra, A., Muslim, M. A., & Khusniati, M. (2017). Penerapan Algoritma K-Means Untuk Clustering Penilaian Dosen Berdasarkan Indeks Kepuasan Mahasiswa. *Techno.Com*, *16*(1), 17–24. https://doi.org/10.33633/tc.v16i1.1284
7. Tendean, T., & Purba, W. (2020). Analisis Cluster Provinsi Indonesia Berdasarkan Produksi Bahan Pangan Menggunakan Algoritma K-Means. *Jurnal Sains Dan Teknologi*, *1*(2), 5–11.