Optimization of the Number of Clusters in Grouping Characteristics of Hajj Pilgrim Candidates in Tebing Tinggi City Using the K-Means Algorithm and Elbow Method

Mhd. Septian, Muhammad Irfan Sarif

Abstract

Every year, thousands of prospective Hajj pilgrims from Tebing Tinggi City register and participate in manasik (Hajj rituals) guidance. The diversity of their characteristics, including age, gender, education, occupation, and frequency of guidance, demands different service strategies. This study aims to group prospective Hajj pilgrims based on demographic characteristics using the K-Means algorithm with cluster number optimization via the Elbow method. The data used is the registration data of prospective Hajj pilgrims for 2023 from KBIHU in Tebing Tinggi City, totaling **189 records**. The variables used include age, gender, education, and occupation. The results show that the optimal number of clusters is **K=3** with an inertia value of 152.47 at K=3. Cluster 1 (Elderly) is dominated by pilgrims aged over 50 years with secondary education and occupations as housewives/retirees. Cluster 2 (Productive Adults) consists of pilgrims aged 30-50 years with higher education and occupations as civil servants/private employees. Cluster 3 (Youth) comprises pilgrims under 30 years old with student status. This study provides recommendations for different guidance strategies for each cluster: an intensive approach with visual/audio media for Cluster 1, flexible blended learning for Cluster 2, and independent digital-based guidance for Cluster 3.

Keywords: *K-Means, Elbow Method, Clustering, Hajj Pilgrims, Optimization, Data Mining, Tebing Tinggi*

Mhd. Septian[1]
[1]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: muhammadtian34@gmail.com[1]

Muhammad Irfan Sarif[2]
[2]Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: irfanberbagi@gmail.com[2]
2nd International Conference on Islamic Community Studies (ICICS)
Theme: History of Malay Civilisation and Islamic Human Capacity and Halal Hub in the Globalization Era
https://proceeding.pancabudi.ac.id/index.php/ICIE/index

Introduction

1.1 Background

Hajj is the fifth pillar of Islam and an obligation for every financially and physically able Muslim. Every year, thousands of prospective Hajj pilgrims from Tebing Tinggi City register and prepare themselves through manasik guidance organized by the Hajj and Umrah Guidance Group (KBIHU) and the Ministry of Religious Affairs (Kemenag). Based on data obtained, the number of Hajj registrants in Tebing Tinggi City in 2023 reached **189 prospective pilgrims** with diverse characteristics. The diversity of characteristics of prospective Hajj pilgrims, which includes age, gender, education level, and occupation, demands different service strategies so that guidance effectiveness can be achieved (Shella & Eviyanti, 2025). Unfortunately, currently the demographic data of prospective Hajj pilgrims stored at KBIHU has not been optimized into strategic information. This data is only stored as administrative archives and has not been utilized to support decision-making. The development of information technology, particularly data mining, opens opportunities to process large amounts of data into new knowledge. Clustering is a data mining technique that groups data based on similarity of characteristics (Wakhidah, 2010). The K-Means algorithm is a popular choice due to its simplicity and efficiency (Kodinariya & Makwana, 2013). This study aims to optimize the number of clusters in grouping the characteristics of prospective Hajj pilgrims in Tebing Tinggi City using the K-Means algorithm and Elbow method, and to formulate recommendations for manasik guidance strategies based on the resulting cluster profiles.

1.2 Problem Formulation

1. What are the demographic characteristics of prospective Hajj pilgrims in Tebing Tinggi City in 2023?
2. What is the optimal number of clusters using the Elbow method on the Hajj pilgrim data?
3. What are the profiles of each resulting cluster?
4. What are the recommended manasik guidance strategies based on the cluster profiles?

1.3 Research Objectives

1. To identify the demographic characteristics of prospective Hajj pilgrims in Tebing Tinggi City in 2023.
2. To determine the optimal number of clusters using the Elbow method.
3. To analyze the profile of each cluster.
4. To formulate recommendations for manasik guidance strategies.

Literature Review

2.1 Data Mining and Clustering

Data mining is the process of extracting valuable information from large datasets to support decision-making (Mardi, 2017). Clustering is a data mining technique that groups data into several clusters based on similarity of characteristics, where objects within one cluster have high similarity while objects between clusters have low similarity (Wakhidah, 2010).

2.2 K-Means Algorithm

The K-Means algorithm aims to partition data into K clusters by minimizing the sum of squared distances between data points and cluster centers (centroids), known as inertia or within-cluster sum of squares (WCSS) (Kodinariya & Makwana, 2013). The objective function of K-Means:

$$J = \Sigma_{j=1}^{k} \Sigma_{i=1}^{n} \|x_i^{(j)} - c_j\|^2$$

2.3 Elbow Method

The Elbow method is used to determine the optimal number of clusters by plotting the value of K against inertia. The "elbow" point where the decrease in inertia begins to slow down is considered the optimal K (Purnima & Arvind, 2014).

Research Methodology

3.1 Data Source

The study uses registration data of prospective Hajj pilgrims for 2023 from KBIHU in Tebing Tinggi City. The data consists of 189 records with variables: name, queue number, gender, occupation, education, place of birth, date of birth, address, village, sub-district.

3.2 Research Variables

Table 1. Research Variables

No	Variabel	Tipe Data	Keterangan
1	Usia	Numerik	Dihitung dari tahun lahir (2024 - tahun lahir)
2	Jenis Kelamin	Kategorik	L = Laki-laki, P = Perempuan
3	Pendidikan	Kategorik	SD, SLTP, SLTA, D1/D2/D3, S1, S2, S3
4	Pekerjaan	Kategorik	Ibu Rumah Tangga, PNS, BUMN/BUMD, TNI/Polri, Swasta, Dagang, Pelajar/Mahasiswa, Lainnya, Pensiunan

3.3 Research Procedure

Research stages:
1. Data Collection: Collecting registration data of prospective Hajj pilgrims
2. Data Pre-processing:
   - Calculating age from birth year
   - Encoding categorical variables
   - Data normalization
3. Determining Optimal K: Running K-Means for K=1 to K=10
4. K-Means Implementation: Performing clustering with optimal K
5. Cluster Profile Analysis: Interpreting characteristics of each cluster

Results

4.1 Data Characteristics

Based on data from 189 prospective Hajj pilgrims in Tebing Tinggi City in 2023, descriptive statistics are as follows:

Table 2. Descriptive Statistics of Hajj Pilgrim Candidate Data, Tebing Tinggi City (2023)

Variabel	N	Mean	Std Dev	Min	Max
Usia (tahun)	189	42,7	11,3	13	75
Jenis Kelamin (L=1, P=0)	189	0,39	-	0	1
Pendidikan (skala 1-7)	189	4,2	1,3	1	7
Usia per Kelompok					
- Usia < 30 tahun	31	22,5	4,2	13	29
- Usia 30-50 tahun	112	41,8	6,1	30	50
- Usia > 50 tahun	46	58,3	6,8	51	75

Demographic Analysis:
- Age: The average age of pilgrims is 42.7 years, ranging from 13 years (youngest) to 75 years (oldest). The majority are in the productive age group of 30-50 years (112 people or 59.3%).

- Gender: The composition of pilgrims is dominated by women, 115 people (60.8%), while men are 74 people (39.2%).
- Education: Education level is dominated by Bachelor's graduates (64 people), Senior High School (53 people), and Master's graduates (25 people).

4.2 Distribution of Occupation and Education

Table 3. Distribution of Occupations of Hajj Pilgrim Candidates

Pekerjaan	Jumlah	Persentase
Pegawai Negeri Sipil	42	22,2%
Swasta	38	20,1%
Ibu Rumah Tangga	35	18,5%
BUMN / BUMD	19	10,1%
Pelajar / Mahasiswa	14	7,4%
Dagang	11	5,8%
Lainnya	11	5,8%
TNI / Polri	9	4,8%
Pensiunan	6	3,2%
Total	**189**	**100%**

Table 4. Distribution of Education of Hajj Pilgrim Candidates

Pendidikan	Jumlah	Persentase
SD	2	1,1%
SLTP	9	4,8%
SLTA	53	28,0%
D1/D2/D3	11	5,8%
S1	64	33,9%
S2	25	13,2%
Tidak tercatat	25	13,2%
Total	**189**	**100%**

4.3 Determining Optimal Number of Clusters with Elbow Method

The K-Means algorithm was run for K=1 to K=10 on normalized data. The inertia calculation results:

Table 5. Inertia Values for Each K Value

K (Jumlah Cluster)	Inertia (WCSS)	Penurunan Inertia
1	356,28	-
2	231,56	124,72
3	**152,47**	**79,09**
4	118,34	34,13
5	97,62	20,72
6	84,31	13,31
7	74,58	9,73
8	67,42	7,16
9	61,89	5,53
10	57,34	4,55

Based on the Elbow graph, the elbow point occurs at K=3, where the decrease in inertia starts to slow down significantly (from a decrease of 79.09 at K=3 to 34.13 at K=4). Thus, the optimal number of clusters is 3 clusters.

4.4 Clustering Results with K=3

Implementation of the K-Means algorithm with K=3 produced the following grouping:

Table 6. Distribution of Members per Cluster

Cluster	Jumlah Anggota	Persentase
Cluster 1	68	36,0%
Cluster 2	72	38,1%
Cluster 3	49	25,9%
Total	**189**	**100%**

4.5 Cluster Profile Analysis

Table 7. Average Characteristics per Cluster

Variabel	Cluster 1	Cluster 2	Cluster 3
Usia (tahun)	54,3	41,2	26,7
Jenis Kelamin (% Laki-laki)	32,4%	44,4%	40,8%
Pendidikan (skala 1-7)	3,2 (SLTA)	5,1 (S1)	3,8 (SLTA/D3)
Pekerjaan Dominan	IRT, Pensiunan	PNS, Swasta	Pelajar/Mhs, Swasta

Table 8. Complete Profile of Each Cluster

Aspek	Cluster 1	Cluster 2	Cluster 3
Rentang Usia	> 50 tahun	30-50 tahun	< 30 tahun
Rata-rata Usia	54,3 tahun	41,2 tahun	26,7 tahun
Komposisi Gender	67,6% Perempuan	55,6% Perempuan	59,2% Perempuan
	32,4% Laki-laki	44,4% Laki-laki	40,8% Laki-laki
Pendidikan	SLTA (44%)	S1 (58%)	SLTA (41%)
	S1 (28%)	S2 (22%)	S1 (29%)
Pekerjaan Utama	IRT (41%)	PNS (36%)	Pelajar/Mhs (29%)
	Pensiunan (15%)	Swasta (28%)	Swasta (24%)
	Swasta (12%)	BUMN (15%)	IRT (18%)
Jumlah Anggota	68 orang	72 orang	49 orang
Persentase	36,0%	38,1%	25,9%
Label Cluster	**Lansia Produktif**	**Profesional Dewasa**	**Muda Potensial**

4.6 Interpretation of Cluster Characteristics

Cluster 1: Productive Elderly (68 people, 36%)
- Dominated by pilgrims aged over 50 years with an average of 54.3 years
- Majority female (67.6%) with a background of Senior High School education
- Main occupations: Housewife (41%) and retiree (15%)
- Characteristics: Have mature life experience, but may require a more patient guidance approach and easily understood media

Cluster 2: Professional Adult (72 people, 38.1%)
- The largest group with productive age 30-50 years (average 41.2 years)
- High education level (58% Bachelor, 22% Master)
- Occupations: Civil Servant (36%), Private (28%), State-Owned Enterprise (15%)
- Characteristics: Have work commitments, require flexible and time-efficient guidance

Cluster 3: Promising Youth (49 people, 25.9%)
- Youngest pilgrims with an average age of 26.7 years (range 13-29 years)
- Dominated by students (29%) and young private employees (24%)
- Education level equivalent to Senior High School to Bachelor
- Characteristics: Familiar with technology, responsive to digital media, high enthusiasm

4.7 Recommendations for Manasik Guidance Strategies

Table 9. Guidance Strategy Recommendations by Cluster

Aspek	Cluster 1 (Lansia Produktif)	Cluster 2 (Profesional Dewasa)	Cluster 3 (Muda Potensial)
Pendekatan Bimbingan	Intensif, tatap muka, sabar dan berulang	Fleksibel, blended learning (tatap muka + online)	Mandiri, berbasis teknologi, interaktif
Media Pembelajaran	- Modul cetak ukuran besar - Video demonstrasi lambat - Audio rekaman - Poster visual	- E-learning (LMS) - Video pendek tutorial - Webinar weekend - Modul PDF	- Aplikasi mobile - Konten TikTok/Reels - VR simulasi manasik - Quiz interaktif
Jadwal Bimbingan	Pagi hari (08.00-10.00) 2x seminggu Reguler	Sore hari (16.00-18.00) Sabtu/Minggu Fleksibel	Self-paced On-demand Akses 24/7
Fokus Materi	- Praktik dasar manasik - Kesehatan jamaah lansia - Doa dan dzikir	- Manajemen waktu ibadah - Efisiensi pelaksanaan - Persiapan fisik	- Pemahaman mendalam - Tsaqofah haji - Simulasi digital
Metode Evaluasi	- Tes lisan - Praktik langsung - Pendampingan personal	- Kuis online - Tugas mandiri - Diskusi kelompok	- Studi kasus - Simulasi digital - Peer evaluation
Pendamping/Kelompok	Kelompok kecil (5-7 orang) Pendamping senior	Kelompok sedang (10-15 orang) Asisten pembimbing	Forum diskusi online Mentor sebaya

4.8 Guidance Program Design

Program for Cluster 1: "Elderly-Friendly Manasik"
- Duration: 8 sessions (2 months)
- Each session: 90 minutes
- Method: Lecture + practice + Q&A
- Media: Large-sized pocket book, slow tutorial videos, murottal audio
- Evaluation: Manasik simulation at the end of the program

Program for Cluster 2: "Weekend Professional Manasik"
- Duration: 6 sessions (3 months)
- Each session: 120 minutes (Saturday/Sunday)
- Method: Blended (40% online, 60% face-to-face)
- Media: LMS, digital modules, tutorial videos
- Evaluation: Online quizzes and limited simulation

Program for Cluster 3: "Digital Millennial Manasik"
- Duration: 4 weeks (self-paced)
- Platform: Mobile app + webinars
- Method: Fully online with discussion forums
- Media: 3D animated videos, VR simulation, infographics, podcasts
- Evaluation: Gamification (points and certificates)

Conclusion

Based on the research results on data from 189 prospective Hajj pilgrims in Tebing Tinggi City in 2023, it can be concluded:

1. Demographic characteristics of prospective Hajj pilgrims in Tebing Tinggi City:
a) Average age 42.7 years, dominated by productive age 30-50 years (59.3%)
b) Gender composition: 60.8% female, 39.2% male
c) Education: dominated by Bachelor (33.9%) and Senior High School (28.0%)
d) Occupation: Civil Servant (22.2%), Private (20.1%), Housewife (18.5%)

2. The optimal number of clusters using the Elbow method is K=3, indicated by the elbow point on the graph with a decrease in inertia from 152.47 (K=3) to 118.34 (K=4).

3. Profiles of the three clusters formed:
a) Cluster 1 (Productive Elderly - 36.0%): Age >50 years, Senior High education, occupation Housewife/Retiree
b) Cluster 2 (Professional Adult - 38.1%): Age 30-50 years, Bachelor/Master education, occupation Civil Servant/Private
c) Cluster 3 (Promising Youth - 25.9%): Age <30 years, Senior High/Bachelor education, occupation Student/Private

4. Guidance strategy recommendations:
a) Cluster 1: Intensive approach with visual/audio media ("Elderly-Friendly Manasik" program)
b) Cluster 2: Flexible blended learning ("Weekend Professional Manasik" program)
c) Cluster 3: Independent digital-based guidance ("Digital Millennial Manasik" program)

Suggestions

For further research:
1. Add variables such as guidance frequency and distance of residence to KBIHU
2. Use other clustering algorithms such as K-Medoids or DBSCAN
3. Expand data coverage to several years for trend analysis

For KBIHU and Kemenag:
1. Develop an integrated database system for recording pilgrim data
2. Implement segmentation in manasik guidance services
3. Develop various guidance media according to cluster characteristics
4. Conduct periodic evaluation of guidance program effectiveness

References

[1] Herdiana, I., et al. (2025). A More Precise Elbow Method for Optimum K-means Clustering. arXiv preprint, arXiv:2502.00851.
[2] Hidayat, T. (2022). Klasterisasi data jamaah umroh menggunakan metode K-Means Clustering (Studi Kasus di AET Travel Indonesia). Masters thesis, Universitas Putra Indonesia YPTK.
[3] Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal of Advanced Research in Computer Science and Management Studies, 1(6), 2321-7782.
[4] Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. Jurnal Edik Informatika, 2(2), 213-219.

[5] Muningsih, E., & Yogyakarta, A. B. S. I. (2017). Optimasi jumlah cluster k-means dengan metode elbow untuk pemetaan pelanggan. Prosiding Seminar Nasional ELINVO, September, 105--114.
[6] Patro, S. G. K., & sahu, K. K. (2015). Normalization: A Preprocessing Stage. Iarjset, 20--22.
[7] Purnima, B., & Arvind, K. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. International Journal of Computer Applications, 105(9), 17--24.
[8] Shella, D. M., & Eviyanti, A. (2025). Implementation of the K-Means Algorithm for Clustering the KBIH Jabal Nur Hajj Pilgrim Area. UMSIDA Preprints Server.
[9] Wakhidah, N. (2010). Clustering Menggunakan K-Means Algorithm. Jurnal Transformatika, 8(1), 33.
[10] Winaryanti, H. S., Hadi, H. P., & Rachmawanto, E. H. (2024). Customer Segmentation Using K-Means Clustering with RFM Method (Case Study: PT. Dewangga Travindo Semarang). Journal of Applied Intelligent System, 9(1).
[11] Yusuf, A. (2025). Computational Implementation of Linear Reference Elbow (LRE) Method for Optimal Cluster Determination in k-Means Algorithm. Journal of Physics: Conference Series, 3139, 012038.