FEATURE SELECTION COMPARATIVE PERFORMANCE FOR UNSUPERVISED LEARNING ON CATEGORICAL DATASET

Rachmad Fitriyanto[1*]; Mohamad Ardi[2]

Information System Department[1]
Informatics Engineering Department[2]
STMIK PPKIA Tarakanita Rahmawati, Tarakan, Indonesia[1,2]
https://ppkia.ac.id[1,2]
fitriyanto7477@gmail.com[1*], ardi@ppkia.ac.id[2]
(*) Corresponding Author

Abstract*— In the era of big data, Knowledge Discovery in Databases (KDD) is vital for extracting insights from extensive datasets. This study investigates feature selection for clustering categorical data in an unsupervised learning context. Given that an insufficient number of features can impede the extraction of meaningful patterns, we evaluate two techniques—Chi-Square and Mutual Information—to refine a dataset derived from questionnaires on college library visitor characteristics. The original dataset, containing 24 items, was preprocessed and partitioned into five subsets: one via Chi-Square and four via Mutual Information using different dependency thresholds (a low-mid-high scheme and dynamic quartile thresholds: Q1toMax, Q2toMax, and Q3toMax). K-Means clustering was applied across nine variations of K (ranging from 2 to 10), with clustering performance assessed using the silhouette score and Davies-Bouldin Index (DBI). Results reveal that while the Mutual Information approach with a Q3toMax threshold achieves an optimal silhouette score at K=7, it retains only 4 features—insufficient for comprehensive analysis based on domain requirements. Conversely, the Chi-Square method retains 18 features and yields the best DBI at K=9, better capturing the intrinsic characteristics of the data. These findings underscore the importance of aligning feature selection techniques with both clustering quality and domain knowledge, and highlight the need for further research on optimal dependency threshold determination in Mutual Information.*

Keywords*: Chi-Square Test, Dynamic Dependency Threshold, Feature Selection, Mutual Information, Unsupervised Learning*

Intisari— *Di era big data, Knowledge Discovery in Databases (KDD) memiliki peranan penting dalam mengekstraksi informasi dari dataset yang besar. Penelitian ini mengkaji performa teknik seleksi fitur untuk klasterisasi data kategorikal. Penelitian ini mengevaluasi dua teknik seleksi fiture, chi-Square dan Mutual Information, untuk menghasilkan dataset yang dapat diproses menghasilak karakteristik pengunjung perpustakaan perguruan tinggi. Dataset asli, yang terdiri dari 24 item, dipra-proses dan dibagi menjadi lima subset: satu subset melalui Chi-Square dan empat subset melalui Mutual Information dengan menggunakan empat macam dependency threshold yaitu Low-Mid-High, dan 3 dari dynamic dependency threshold (Q1toMax, Q2toMax, dan Q3toMax. Hasil seleksi fitur dievaluasi menggunakan K-Means variasi nilai K mulai K=2 hingga K=10. Hasil klasterisasi dievaluasi kembali menggunakan silhouette score dan Davies-Bouldin Index (DBI). Hasil penelitian menunjukkan bahwa meskipun pendekatan Mutual Information dengan ambang Q3toMax mencapai skor silhouette optimal pada K=7, metode tersebut hanya mempertahankan 4 fitur, jumlah yang tidak mencukupi untuk ekstraksi informasi karakteristik pengunjung perpustakaan. Sebaliknya, metode Chi-Square mempertahankan 18 fitur dan menghasilkan DBI terbaik pada K=9, sehingga lebih mampu menangkap karakteristik intrinsik data. Hal ini menunjukkan diperlukannya integrasi teknik seleksi fitur dengan domain knowledge untuk menentukan ukuran dataset yang optimal.*

Kata Kunci*: Chi-Square Test, Dynamic Dependency Threshold, Mutual Information, Seleksi Fitur, Unsupervised Learning.*

INTRODUCTION

In the era of big data, Knowledge Discovery in Databases (KDD) plays a critical role in extracting valuable patterns and insights from large datasets

KDD offered function for the process of uncovering significant patterns, trends, and insights from vast datasets which is a pivotal aspect of modern data analysis. Feature selection, a critical step within KDD, plays an essential role in refining the dataset by eliminating irrelevant or redundant features, thereby enhancing the efficiency and effectiveness of data mining algorithms (Sosa-Cabrera et al., 2024; Tadesse et al., 2022; Tsamardinos et al., 2022). Although extensively studied within the context of supervised learning, the importance of feature selection in unsupervised learning, particularly for categorical dataset clusterization, warrants further exploration (Bhadra et al., 2022; Büyükkeçeci & Okur, 2023; Hopf & Reifenrath, 2021; Pudjihartono et al., 2022; Sosa-Cabrera et al., 2024; Yang et al., 2022).

Unsupervised learning algorithms, such as clustering, aim to discern inherent structures within data without predefined labels or targets. When dealing with categorical datasets, the challenge of feature selection becomes even more pronounced due to the discrete nature of the attributes and the absence of clear performance metrics. Inappropriate or excessive features can lead to suboptimal clustering results, obscuring meaningful patterns and potentially leading to erroneous interpretations.

However, feature selection in unsupervised learning presents several unique challenges. Firstly, the absence of labeled data makes it difficult to evaluate the relevance and quality of features directly. Unlike supervised learning, where feature importance can be assessed based on target variables, unsupervised learning requires alternative approaches to measure the impact of features on the clustering outcome. Secondly, categorical data adds another layer of complexity due to the nominal nature of the variables, often necessitating specialized techniques for feature selection and distance measurement (Fitriyanto & Syafiqoh, 2024). Lastly, the high dimensionality of many real-world datasets can exacerbate the issue of the curse of dimensionality, making it essential to identify the most informative subset of features to ensure robust and meaningful clustering results (Peng et al., n.d.; Ting et al., 2021; Yan et al., 2021).

In addressing these challenges, the use of Chi-square test and Mutual Information technique for feature selection is particularly pertinent. The Chi-square test is a statistical measure used to evaluate the independence between categorical variables. By assessing the degree of association between features, the Chi-square test helps in identifying those features that significantly contribute to the clustering structure(Párraga-Valle et al., 2020; Tang, 2024). This method is especially useful in handling categorical data, providing a robust mechanism to filter out irrelevant attributes.

On the other hand, Mutual Information measures the amount of information shared between two variables, indicating the degree of dependency between them (Covert et al., 2023; Liu & Motani, 2022). In the context of feature selection for clustering, Mutual Information can be leveraged to evaluate the relationship between features and the clustering outcome, even in the absence of labelled data. By quantifying the shared information, this technique aids in selecting features that enhance the clustering process, leading to more accurate and interpretable clusters. One notable research gap in the existing literature pertains to the dependency of clustering outcomes on the mutual information threshold used for feature selection.

While mutual information has proven effective in evaluating the dependency between features and clustering outcomes, the selection of an appropriate threshold remains a significant challenge (Prasetiyowati et al., 2021). Previous research has not clearly established guidelines or best practices for determining the optimal mutual information threshold, leading to inconsistent results and potential bias in feature selection processes. This lack of clarity hinders the reproducibility and generalizability of studies utilizing mutual information for feature selection (Rohadi, 2023).

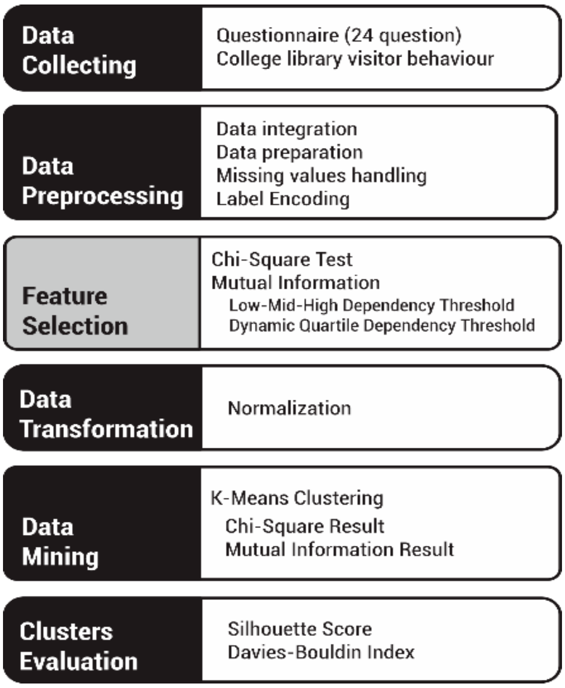
Despite the growing importance of feature selection in unsupervised learning, particularly in categorical dataset clustering, several key challenges remain unresolved. Unlike supervised learning, where feature relevance can be evaluated based on predefined class labels, unsupervised feature selection lacks a direct performance metric, making it difficult to determine the most informative attributes. As a result, many clustering models suffer from noisy, redundant, or irrelevant features, leading to suboptimal cluster formations and reduced interpretability.

While statistical and information-theoretic methods such as Chi-Square tests and Mutual Information have been explored for feature selection, their application to categorical clustering remains limited and underdeveloped. Many existing approaches are either designed for numerical datasets or fail to scale effectively with high-dimensional categorical data. Additionally, there is no consensus on the best feature selection strategy for categorical clustering, making it challenging to establish a standardized framework for improving clustering performance in KDD.

Therefore, the fundamental research problem addressed in this study is How can effective feature selection methods improve categorical data clustering in unsupervised learning, and what are the best techniques for selecting relevant features in the absence of labelled data? This research seeks to bridge the gap by evaluating and improving feature selection techniques for categorical clustering, ensuring that the most relevant features are retained while maintaining the integrity of discovered patterns in KDD applications.

MATERIALS AND METHODS

This study adopts a quantitative research design to investigate the impact of feature selection techniques on categorical data clustering in an unsupervised learning setting. The research follows an experimental approach, where different feature selection methods, particularly the Chi-Square test and Mutual Information, are applied to categorical datasets to evaluate their effectiveness in improving clustering performance. Figure 1 gave an illustration this main research stages.



Source: (Research Result, 2025)
Figure 1. Research Stages

The research stages depict from figure 1 adopt KDD framework from our previous study (Fitriyanto & Syafiqoh, 2024), involves data collecting, data preprocessing, feature selection, clustering analysis, and performance evaluation based on internal validation metrics. The datasets used in this research collected from questionnaire contained 24 questions about college library visitors characteristics as shown in Table 1.

Table 1. Questionnaire Items

Code	Question
Q01	Respondent department
Q02	Respondent sex
Q03	Respondent age
Q04	Active Semester
Q06	Previous High school category
Q07	Have you registered or ever applied for membership at the university library?
Q08	Have you ever visited the university library?
Q09	Have you ever accessed the university library's website?
Q10	In which semester did you first receive information about the university library?
Q11	Where did you obtain information about the university library?
Q12	On average, how many times do you visit the university library per month?
Q13	What is your purpose for visiting the university library?
Q14	Did your previous high school have a library?
Q15	On average, how many times did you visit your school library per month when you were in 12th grade?
Q16	What obstacles prevent you from visiting the university library frequently or at all?
Q17	What type of media do you read most often?
Q18	When was the last time you purchased reading materials such as books, magazines, or newspapers?
Q19	How much time do you spend reading per day?
Q20	Have you ever completed reading an entire book?
Q21	What book genre do you prefer the most?
Q22	Have you ever visited a bookstore (either online or in person)?
Q23	What is your primary source of academic references?
Q24	Do you have close friends who frequently read (physical or online media, excluding social media)?

Source: (Research Result, 2025)

The Twenty three questions are close-ended question with categorical multiple choice, and one question is open-ended question about the age of respondent. . The questionnaire consists of twenty-three close-ended questions with categorical multiple-choice responses and one open-ended question regarding the respondent's age. The target respondents are active college students in their 2nd semester and students who have completed their final project in the 7th or 8th semester. A purposive sampling technique was employed to ensure that the selected respondents had relevant experiences with library usage and academic reading habits. The questionnaire was distributed online via Google Forms, and a total of 140 responses were collected. The sample size was deemed sufficient for exploratory analysis within the given population

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 22, No. 1 March 2025 | DOI https://doi.org/10.33480/techno.v22i1.6512

The second research stages, contain of data preparation and data preprocessing, involves data tabulation, missing-value handling and data normalization using min-max normalization. The third stage are implementing feature selection techiques, chi-square test and mutual information. On this stage we studied and implement both feature selection techniques. On the mutual information implemetation, we proposed two techniques of dependency threshold for feature selection. The first technique use 3 range of threshold of mutual information score, low, mid and high. The low threshold for score between 0 to 0.19, the middle threshold for score between 0.2 to 1 and the high threshold for score more than 1. The second technique proposed are dynamic quartile threshold. We adopt quartile concept (Q1,Q2,Q3) to generate parts on each features based on quartiles values.

The result from third stage are 5 datasets comprises one dataset from Chi-Square feature selection result, one dataset from mutual information with low-mid-hight threshold (MI-LMH) and three datasets from dynamic quartile threshold (MI-Q1toMax, MI-Q2toMax, MI-Q3toMax).

In the research fourth stage, we clusterized the five datasets using K-Means Clustering with 9 variations of K values, start from K=2 until K=10. The clustering process conducted with rapidminer tool which used also for calculating Davies-Bouldin Index (DBI) for each K values. Other evaluation metrics used in this study is silhouette score (SSc), calculated use jupyter notebook. Based on DBI and SSc, we determined the best clustered data and the suitable feature selection tehniques according the clustering results.

RESULTS AND DISCUSSION

The results of the Chi-square test for feature selection on the categorical dataset with alpha 0.05 are summarized in Table 2.

Table 2. Chi-Square Test Results

Feature	X^2	P Value	Results
Q01	302.798	0.0006209	Significant
Q02	151.9762647	0.0100886	Significant
Q03	1296.56419	0.0000003	Significant
Q04	655.639422	0.0012589	Significant
Q05	654.5967758	0.4860681	Not Significant
Q06	1394.618606	0.0000000	Significant
Q07	278.1559001	0.0000000	Significant
Q08	284.377798	0.0000000	Significant
Q09	237.0809342	0.0000000	Significant
Q10	1374.411161	0.0000000	Significant
Q11	832.5192173	0.0272942	Significant
Q12	524.1340496	0.7800251	Not Significant
Q13	1340.649878	0.0000000	Significant
Q14	334.6602267	0.0000000	Significant
Q15	776.999	0.0000000	Significant
Q16	1187.0925	0.0000000	Significant
Q17	219.7617237	0.0000000	Significant
Q18	591.4363317	0.0000000	Significant
Q19	446.516836	0.0000000	Significant
Q20	479.205809	0.0000000	Significant
Q21	1343.409944	0.0000000	Significant
Q22	246.5116204	0.0000000	Significant
Q23	331.8886777	0.5531018	Not Significant
Q24	85.80934089	0.9773909	Not Significant

Source : (Research Result, 2025)

From Table 1, it can be observed Features Q05, Q12, Q23 and Q24 are not significant to others and will be removed from dataset.

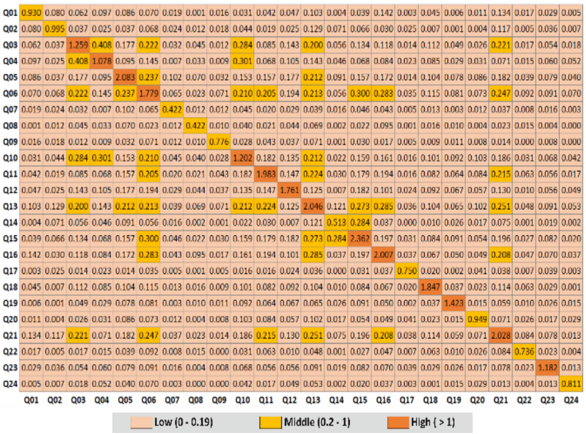
The mutual information calculation result, visualized as matrix in figure 2. Based on mutual information score, we set 2 dependency threshold categories. The first category is set the threshold into 3 range scores, low(0 – 0.19), middle (0.2 – 1.0) and high(>1.0). this first category applied for all mutual information score with excluding the score between feature to it self. Features with low mutual information scores removed from dataset, while features with score between middle threshold to high, retained in dataset.

Fitur	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24
Q01	0.930	0.080	0.062	0.097	0.086	0.070	0.019	0.001	0.016	0.031	0.042	0.047	0.103	0.004	0.039	0.142	0.003	0.045	0.006	0.011	0.134	0.017	0.029	0.005
Q02	0.080	0.995	0.037	0.025	0.037	0.068	0.024	0.012	0.018	0.044	0.019	0.025	0.129	0.071	0.066	0.030	0.025	0.007	0.001	0.004	0.117	0.005	0.036	0.007
Q03	0.062	0.037	1.259	0.408	0.177	0.222	0.032	0.045	0.012	0.284	0.085	0.143	0.200	0.056	0.134	0.118	0.014	0.112	0.049	0.026	0.221	0.017	0.054	0.018
Q04	0.097	0.025	0.408	1.078	0.095	0.145	0.007	0.033	0.009	0.301	0.068	0.105	0.143	0.046	0.068	0.084	0.023	0.085	0.029	0.031	0.071	0.015	0.060	0.052
Q05	0.086	0.037	0.177	0.095	2.083	0.237	0.102	0.070	0.032	0.153	0.157	0.177	0.212	0.091	0.157	0.172	0.014	0.104	0.078	0.086	0.182	0.039	0.079	0.040
Q06	0.070	0.068	0.222	0.145	0.237	1.779	0.065	0.023	0.071	0.210	0.205	0.194	0.213	0.056	0.300	0.283	0.035	0.115	0.081	0.073	0.247	0.092	0.091	0.070
Q07	0.019	0.024	0.032	0.007	0.102	0.065	0.422	0.012	0.012	0.045	0.020	0.029	0.039	0.016	0.046	0.043	0.005	0.013	0.003	0.012	0.037	0.008	0.016	0.003
Q08	0.001	0.012	0.045	0.033	0.070	0.023	0.012	0.422	0.010	0.040	0.021	0.044	0.069	0.002	0.022	0.095	0.001	0.016	0.010	0.004	0.023	0.015	0.004	0.000
Q09	0.016	0.018	0.012	0.009	0.032	0.071	0.012	0.010	0.776	0.028	0.043	0.037	0.071	0.001	0.030	0.017	0.005	0.009	0.011	0.008	0.014	0.000	0.008	0.000
Q10	0.031	0.044	0.284	0.301	0.153	0.210	0.045	0.040	0.028	1.202	0.182	0.135	0.212	0.022	0.159	0.161	0.016	0.101	0.092	0.103	0.186	0.031	0.068	0.042
Q11	0.042	0.019	0.085	0.068	0.157	0.205	0.020	0.021	0.043	0.182	1.983	0.147	0.224	0.030	0.179	0.194	0.016	0.082	0.064	0.084	0.215	0.063	0.056	0.017
Q12	0.047	0.025	0.143	0.105	0.177	0.194	0.029	0.044	0.037	0.135	0.147	1.761	0.125	0.007	0.182	0.101	0.024	0.092	0.067	0.057	0.130	0.010	0.056	0.049
Q13	0.103	0.129	0.200	0.143	0.212	0.213	0.039	0.069	0.071	0.212	0.224	0.125	2.046	0.121	0.273	0.285	0.036	0.104	0.065	0.102	0.251	0.048	0.091	0.053
Q14	0.004	0.071	0.056	0.046	0.091	0.056	0.016	0.002	0.001	0.022	0.030	0.007	0.121	0.513	0.284	0.037	0.000	0.010	0.026	0.017	0.075	0.001	0.019	0.002
Q15	0.039	0.066	0.134	0.068	0.157	0.300	0.046	0.022	0.030	0.159	0.179	0.182	0.273	0.284	2.362	0.197	0.031	0.084	0.091	0.054	0.196	0.027	0.082	0.020
Q16	0.142	0.030	0.118	0.084	0.172	0.283	0.043	0.095	0.017	0.161	0.194	0.101	0.285	0.037	0.197	2.007	0.037	0.067	0.050	0.049	0.208	0.047	0.070	0.037
Q17	0.003	0.025	0.014	0.023	0.014	0.035	0.005	0.001	0.005	0.016	0.016	0.024	0.036	0.000	0.031	0.037	0.750	0.020	0.002	0.041	0.038	0.007	0.039	0.003
Q18	0.045	0.007	0.112	0.085	0.104	0.115	0.013	0.016	0.009	0.101	0.082	0.092	0.104	0.010	0.084	0.067	0.020	1.847	0.037	0.023	0.114	0.063	0.029	0.001
Q19	0.006	0.001	0.049	0.029	0.078	0.081	0.003	0.010	0.011	0.092	0.064	0.067	0.065	0.026	0.091	0.050	0.002	0.037	1.423	0.015	0.059	0.010	0.026	0.015
Q20	0.011	0.004	0.026	0.031	0.086	0.073	0.012	0.004	0.008	0.103	0.084	0.057	0.102	0.017	0.054	0.049	0.041	0.023	0.015	0.949	0.071	0.026	0.017	0.029
Q21	0.134	0.117	0.221	0.071	0.182	0.247	0.037	0.023	0.014	0.186	0.215	0.130	0.251	0.075	0.196	0.208	0.038	0.114	0.059	0.071	2.028	0.084	0.078	0.013
Q22	0.017	0.005	0.017	0.015	0.039	0.092	0.008	0.015	0.000	0.031	0.063	0.010	0.048	0.001	0.027	0.047	0.007	0.063	0.010	0.026	0.084	0.736	0.023	0.004
Q23	0.029	0.036	0.054	0.060	0.079	0.091	0.016	0.004	0.008	0.068	0.056	0.056	0.091	0.019	0.082	0.070	0.039	0.029	0.026	0.017	0.078	0.023	1.182	0.013
Q24	0.005	0.007	0.018	0.052	0.040	0.070	0.003	0.000	0.000	0.042	0.017	0.049	0.053	0.002	0.020	0.037	0.003	0.001	0.015	0.029	0.013	0.004	0.013	0.811

Source: (Research Result, 2025)

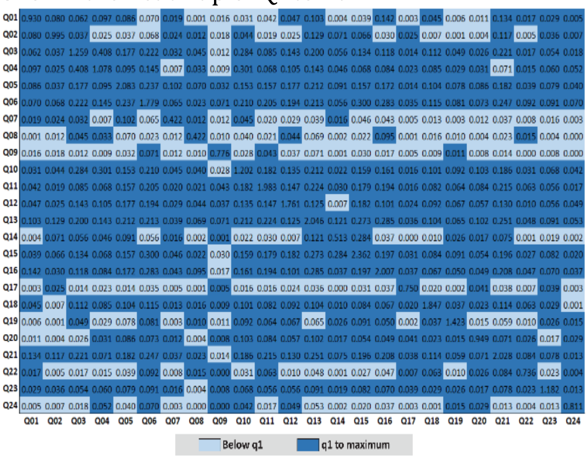
Figure 2. Chi-Square Mutual Information Matrix

Figure 3 shown heatmap of chi-square mutual information score that has been categorized into three dependency threshold. Except mutual information low score, 24 mutual information scores on middle and high categories are values between feature to itself, made this score did not included to selection process. From this result, it is concluded that only mutual information scores in middle categories retained in dataset, there are Q03, Q04, Q05, Q06, Q10, Q11, Q13, Q14, Q15, Q16, and Q21.



Source: (Research Result, 2025)
Figure 3. Chi-Square MI Score Heatmap

The second categories of mutual information dependency threshold proposed in this research is dynamic quartile threshold. This category, generate 1st, 2nd and 3rd quartile which are different on each features mutual information scores. Based on these quartile values, we developed 3 rules for feature selection. First, remove scores below Q1 or retain scores between Q1 to maximum score (Q1toMax). Second, retain scores between Q2 to maximum score (Q2toMax) and the third is retain score between Q3 to maximum score (Q3toMax). Figure 4 shown the heatmap of Q1toMax.



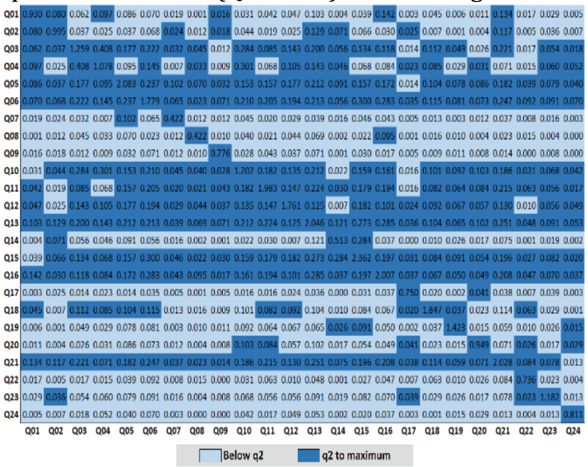
Source:(Research Result, 2025)
Figure 4. Q1toMax Heatmap

The selection of features to be retained or removed from the dataset is carried out by calculating the percentage of mutual information values appearing in each row of the feature matrix. Table 3 contains these percentage values.

Table 3. Percentage Appearing in Q1toMax

Feature	Persen	Feature	Persen
Q01	56.5%	Q13	100.0%
Q02	52.2%	Q14	47.8%
Q03	95.7%	Q15	100.0%
Q04	87.0%	Q16	100.0%
Q05	100.0%	Q17	17.4%
Q06	100.0%	Q18	91.3%
Q07	30.4%	Q19	52.2%
Q08	21.7%	Q20	78.3%
Q09	13.0%	Q21	95.7%
Q10	100.0%	Q22	47.8%
Q11	100.0%	Q23	100.0%
Q12	95.7%	Q24	30.4%

Features with a percentage of less than 50% are removed from the dataset, while features with a minimum percentage of 50% are retained. The results show that at the Q1toMax threshold, 18 features are retained: Q01, Q02, Q03, Q04, Q05, Q06, Q10, Q11, Q12, Q13, Q15, Q16, Q18, Q19, Q20, Q21, and Q23. Second threshold result of dynamic quartile threshold (Q2toMax) shown in figure 5.



Source:(Research Result, 2025)
Figure 5. Q2toMax Heatmap

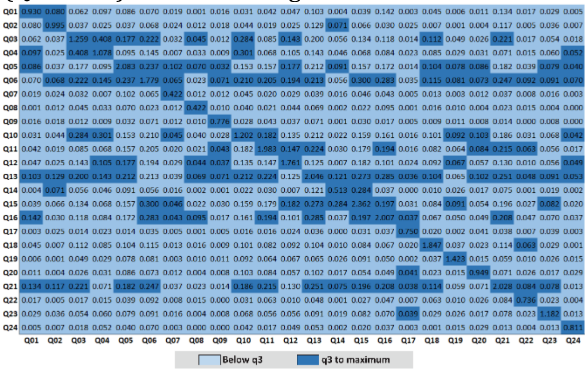
"The percentage of mutual information values appearing in each row of the feature matrix with the Q2toMax threshold is shown in Table 4.

Table 4. Percentage Appearing in Q2toMax

Feature	Persen	Feature	Persen
Q01	21.74%	Q13	100.00%
Q02	30.43%	Q14	8.70%
Q03	86.96%	Q15	95.65%
Q04	56.52%	Q16	100.00%
Q05	95.65%	Q17	4.35%
Q06	100.00%	Q18	39.13%
Q07	4.35%	Q19	17.39%
Q08	4.35%	Q20	21.74%
Q09	0.00%	Q21	95.65%
Q10	95.65%	Q22	0.00%
Q11	95.65%	Q23	13.04%
Q12	95.65%	Q24	0.00%

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 22, No. 1 March 2025 | DOI https://doi.org/10.33480/techno.v22i1.6512

The features retained at the Q2toMax threshold total 13, namely Q03, Q04, Q05, Q06, Q10, Q11, Q12, Q13, Q15, Q16, and Q21. The third threshold result of the dynamic quartile threshold (Q3toMax) is shown in Figure 6.



Source:(Research Result, 2025)
Figure 6. Q3toMax Heatmap

The features retained at the Q3toMax threshold total 4, namely Q05, Q06, Q13, and Q21. A comparison of feature selection results from the five methods used is illustrated in Figure 7.

CS	LMH	Q1 to Max	Q2 to Max	Q3 to Max
Q01		Q01		
Q02		Q02		
Q03	Q03	Q03	Q03	
Q04	Q04	Q04	Q04	
	Q05	Q05	Q05	Q05
Q06	Q06	Q06	Q06	Q06
Q07				
Q08				
Q09				
Q10	Q10	Q10	Q10	
Q11	Q11	Q11	Q11	
	Q12	Q12	Q12	
Q13	Q13	Q13	Q13	Q13
Q14	Q14			
Q15	Q15	Q15	Q15	
Q16	Q16	Q16	Q16	
Q17				
Q18		Q18		
Q19		Q19		
Q20		Q20		
Q21	Q21	Q21	Q21	Q21
Q22				
		Q23		

Source: (Research Result, 2025)
Figure 7. Feature Selection Result Datasets

All five datasets illustrated on figure xx processed on the research fourth stages by clusterized use K-Means Clustering with Rapidminer Tool. The clusterization conducted with 9 K values variation, start from K=2 until K=10. Each clusterization with each K values, evaluated use 2 metrics evaluation, Silhouette Score calculated with Jupyter Notebook and Davies-Bouldin Index with Rapidminer. Table 5 until 9 shown the values of both metrics from five datasets clusterization.

Table 5. Silhouette Score and DBI on CS Dataset

K	Silhouette Score	DBI
2	0.07134489870369704	0.095
3	0.06566427264922067	0.066
4	0.05992200202267188	0.068
5	0.05496875702666082	0.071
6	0.05674252452552738	0.075
7	0.05289824087045086	0.066
8	0.06585368181872374	0.064
9	0.04251910758984117	0.062
10	0.04630911912931734	0.065

Source: (Research Result, 2025)

Table 6. Silhouette Score and DBI on LMH Dataset

K	Silhouette Score	DBI
2	0.1199201821596599	0.153
3	0.1305380236580989	0.119
4	0.1159028748245709	0.128
5	0.1270351224612208	0.123
6	0.1340168096417899	0.119
7	0.1241950906630058	0.125
8	0.1207292542551403	0.119
9	0.1438223485636899	0.103
10	0.1273698319283217	0.129

Source: (Research Result, 2025)

Table 7. Q1toMax's Silhouette Score and DBI

K	Silhouette Score	DBI
2	0.0669501845352940	0.107
3	0.0746678268495167	0.083
4	0.0696196436243296	0.090
5	0.0658663768598787	0.081
6	0.0743782566214256	0.084
7	0.0684799310879453	0.077
8	0.0707201729810179	0.078
9	0.0502150236353305	0.089
10	0.0649687150633831	0.086

Source: (Research Result, 2025)

Table 8. Q2toMax's Silhouette Score and DBI

K	Silhouette Score	DBI
2	0.1121824306087677	0.155
3	0.1232611865907965	0.120
4	0.1130570488278988	0.129
5	0.1215713590794960	0.125
6	0.1323938934216536	0.135
7	0.1141578990978350	0.126
8	0.1264024425364859	0.122
9	0.0851750139990721	0.131
10	0.0988452539010624	0.133

Source: (Research Result, 2025)

Table 9. Q3toMax's Silhouette Score and DBI

K	Silhouette Score	DBI
2	0.2509677008324908	0.271
3	0.2825964811885622	0.234
4	0.2897665313520913	0.203
5	0.3080882342690044	0.202
6	0.22375592644913886	0.207
7	0.31027517725843834	0.241
8	0.25001176906284456	0.214
9	0.20168154097554916	0.242
10	0.2343218821130369	0.211

Source: (Research Result, 2025)

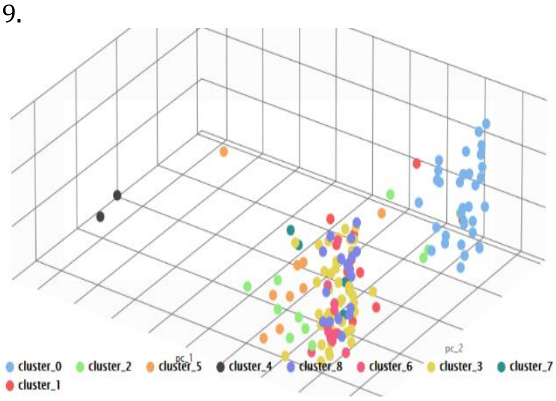
All clusterization's metric evaluation compared each other to determined the optimal K value as shown as in tabel 10.

Table 10. Silhouette Score and DBI Comparison

Dataset	Silhouette Score	K		DBI
Chi-Square	0.071	2	9	0.062
LMH	0.144	9	9	0.103
Q1toMax	0.075	3	7	0.077
Q2toMax	0.132	6	3	0.120
Q3toMax	0.310	7	5	0.202

Source: (Research Result, 2025)

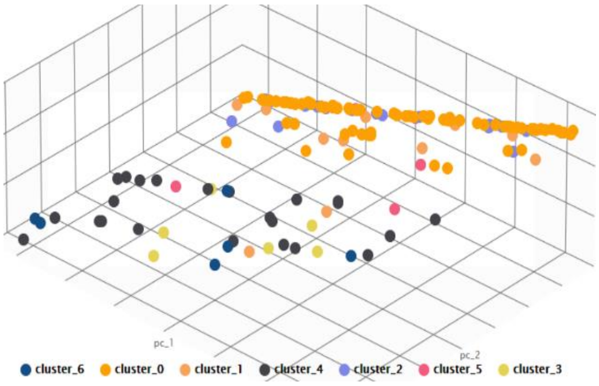
The optimal K value based on the silhouette score is determined by the highest silhouette score, while the optimal K value based on the DBI is chosen from the lowest DBI value. The comparison results in Table 9 show that the optimal K values differ between the silhouette score and DBI. Based on the silhouette score, the best K is K=7 for the dataset obtained from feature selection using mutual information with the dynamic dependency threshold Q3toMax. In contrast, based on the DBI value, the best K is K=9 for the dataset obtained from feature selection using the Chi-Square method. A visual comparison of K=9 from Chi-Square feature selection and K=7 from Mutual Information Q3toMax feature selection is shown in Figures 8 and 9.



Source: (Research Result, 2025)
Figure 8. Chi-Square Clustered Result Scatter Plot

The visualization of clustered data from feature selection using the Chi-Square method shows that there is one cluster (cluster_4) on the left side of the image that has the potential to be an outlier or singleton due to having only two members. In contrast, no such occurrence is observed in the clustering results for the dataset obtained from feature selection using mutual information with the Q3toMax threshold, as shown in Figure 9.

The clustering results shown in Figure 9 do not indicate any outlier or singleton clusters. However, the feature selection using mutual information with the Q3toMax threshold retains only 4 out of 24 features. From a domain knowledge perspective, considering the purpose of questionnaire design during the data collection stage, having only 4 features is insufficient to describe the characteristics of campus library visitors.



Source: (Research Result, 2025)
Figure 9. Q3toMax Clustered Result Scatter Plot

Therefore, based on the feature selection and clustering analysis conducted in this study, there are two decisions was made, first is to use the Chi-Square feature selection results. This approach retains a sufficient number of features (18) to achieve the data collection objectives. Although one cluster has the potential to be a singleton, it may also represent a unique and easily identifiable characteristic. Second, use the second-best silhouette score to select optimal K. From table 9, the second best silhouette score belong to K=9 from mutual information score with low-mid-high dependency threshold, contains of 12 data features.

The findings from selected dataset based on mutual information score provide several important implications. From a practical perspective, the results highlight the need for university libraries to enhance their outreach efforts, particularly among early-semester students who may have limited awareness of library services. Libraries could implement targeted orientation programs or digital engagement strategies to encourage student participation. Theoretically, the research contributes to the understanding of student reading behaviors and library usage, supporting previous studies that emphasize the role of academic resources in student success. Methodologically, this study demonstrates the effectiveness of purposive sampling in capturing diverse perspectives across different academic levels. Future research could expand the sample size or explore qualitative methods to gain deeper insights into students' motivations and barriers related to library use.

CONCLUSION

In this study, the selection of feature selection techniques for clustering categorical datasets is determined based on the quality of the resulting clusters and the domain or business knowledge underlying the data collection process. A limited number of features may hinder data users or business analysts from effectively extracting meaningful insights from the formed clusters. Furthermore, understanding the advantages and drawbacks of outliers within clustered data can serve as an additional consideration when selecting an appropriate feature selection technique for specific cases.

This research demonstrates the application of feature selection in an unsupervised learning context, expanding its traditional use beyond supervised learning. However, several aspects warrant further investigation. The discrepancy in optimal K values between the silhouette score and the Davies-Bouldin Index (DBI) requires deeper exploration to provide greater certainty for unsupervised learning practitioners in determining the appropriate number of clusters. Additionally, the determination of dependency thresholds in mutual information remains an open research challenge, necessitating further studies across different dataset variations.

ACKNOWLEDGEMENT

The authors express appreciation and sincere gratitude to The Indonesian Ministry of Higher Education, Science, and Technology for their generous grant by contract numbers 113/E5/PG.02.00PL/2024 through Junior Lecture Research program scheme, which has played a crucial role in supporting this research project. Their belief in the significance of this research and their commitment to advancing knowledge in this field have been instrumental in its realization

REFERENCE

Bhadra, T., Mallik, S., Hasan, N., & Zhao, Z. (2022). Comparison of five supervised feature selection algorithms leading to top features and gene signatures from multi-omics data in cancer. *BMC Bioinformatics*, *23*(S3), 153. https://doi.org/10.1186/s12859-022-04678-y

Büyükkeçeci, M., & Okur, M. C. (2023). A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning. *Gazi University Journal of Science*, *36*(4), 1506–1520. https://doi.org/10.35378/gujs.993763

Covert, I., Qiu, W., Lu, M., Kim, N., White, N., & Lee, S.-I. (2023). *Learning to Maximize Mutual Information for Dynamic Feature Selection* (arXiv:2301.00557). arXiv. https://doi.org/10.48550/arXiv.2301.00557

Fitriyanto, R., & Syafiqoh, U. (2024). Multilevel Modal Value Analysis for Interpreting Categorical K-Medoids Clusters Data. *Jurnal Techno Nusa Mandiri*, *21*(2), 134–143. https://doi.org/10.33480/techno.v21i2.5796

Hopf, K., & Reifenrath, S. (2021). *Filter Methods for Feature Selection in Supervised Machine Learning Applications—Review and Benchmark* (arXiv:2111.12140). arXiv. https://doi.org/10.48550/arXiv.2111.12140

Liu, S., & Motani, M. (2022). *Improving Mutual Information based Feature Selection by Boosting Unique Relevance* (arXiv:2212.06143). arXiv. https://doi.org/10.48550/arXiv.2212.06143

Párraga-Valle, J., García-Bermúdez, R., Rojas, F., Torres-Morán, C., & Simón-Cuevas, A. (2020). Evaluating Mutual Information and Chi-Square Metrics in Text Features Selection Process: A Study Case Applied to the Text Classification in PubMed. In I. Rojas, O. Valenzuela, F. Rojas, L. J. Herrera, & F. Ortuño (Eds.), *Bioinformatics and Biomedical Engineering* (Vol. 12108, pp. 636–646). Springer International Publishing. https://doi.org/10.1007/978-3-030-45385-5_57

Peng, D., Gui, Z., & Wu, H. (n.d.). *Interpreting the Curse of Dimensionality from Distance Concentration and Manifold Effect.*

Prasetiyowati, M. I., Maulidevi, N. U., & Surendro, K. (2021). Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. *Journal of Big Data*, *8*(1), 84. https://doi.org/10.1186/s40537-021-00472-4

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. *Frontiers in Bioinformatics*, *2*, 927312. https://doi.org/10.3389/fbinf.2022.927312

Rohadi, P. B. (2023). *Optimasi Metode Naïve Bayes Menggunakan Seleksi Fitur Mutual Information Untuk Klasifikasi Teks Ujaran Kebencian*. Universitas Pembangunan Nasional "Veteran."

Sosa-Cabrera, G., Gómez-Guerrero, S., García-Torres, M., & Schaerer, C. E. (2024). Feature Selection: A perspective on inter-attribute cooperation. *International Journal of Data Science and Analytics*, *17*(2), 139–151. https://doi.org/10.1007/s41060-023-00439-z

Tadesse, G. A., Ogallo, W., Cintas, C., & Speakman, S. (2022). *Model-free feature selection to facilitate automatic discovery of divergent subgroups in tabular data* (arXiv:2203.04386). arXiv. https://doi.org/10.48550/arXiv.2203.04386

Tang, C. (2024). Review on Application of Chi-square Statistic in Text Classification in Recent Five Years. *Applied and Computational Engineering*, *97*(1), 115–118. https://doi.org/10.54254/2755-2721/97/20241397

Ting, K. M., Washio, T., Zhu, Y., & Xu, Y. (2021). *Breaking the curse of dimensionality with Isolation Kernel* (arXiv:2109.14198). arXiv. https://doi.org/10.48550/arXiv.2109.14198

Tsamardinos, I., Charonyktakis, P., Papoutsoglou, G., Borboudakis, G., Lakiotaki, K., Zenklusen, J. C., Juhl, H., Chatzaki, E., & Lagani, V. (2022). Just Add Data: Automated predictive modeling for knowledge discovery and feature selection. *Npj Precision Oncology*, *6*(1), 38. https://doi.org/10.1038/s41698-022-00274-8

Yan, X., Sarkar, M., Gebru, B., Nazmi, S., & Homaifar, A. (2021). *A Supervised Feature Selection Method For Mixed-Type Data using Density-based Feature Clustering* (arXiv:2111.08169). arXiv. https://doi.org/10.48550/arXiv.2111.08169

Yang, Y., Wang, W., Fu, H., & Kuo, C.-C. J. (2022). *On Supervised Feature Selection from High Dimensional Feature Spaces* (arXiv:2203.11924). arXiv. https://doi.org/10.48550/arXiv.2203.11924