RESEARCH ARTICLE • OPEN ACCESS

Implementation of K-Prototypes with Feature Selection in Clustering Cervical Cancer Patients based on Risk Factors

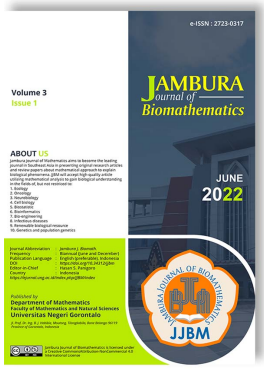
Wanda Puspita Hati, Devvi Sarwinda, and Bevina Desjwiandra Handari

Volume 6, Issue 3, Pages 234–240, September 2025

Received 7 February 2025, Revised 19 March 2025, Accepted 19 August 2025, Published Online 12 September 2025

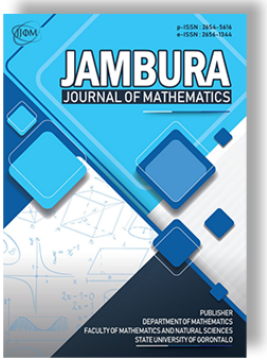
To Cite this Article : W. P. Hati, D. Sarwinda, and B. D. Handari,"Implementation of K-Prototypes with Feature Selection in Clustering Cervical Cancer Patients based on Risk Factors", *Jambura J. Biomath*, vol. 6, no. 3, pp. 234–240, 2025, *https://doi.org/10.37905/jjbm.v6i3.30552*

JOURNAL INFO • JAMBURA JOURNAL OF BIOMATHEMATICS



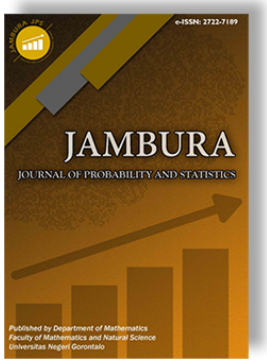
Homepage	:	http://ejurnal.ung.ac.id/index.php/JJBM/index
Journal Abbreviation	:	Jambura J. Biomath.
Frequency	:	Quarterly (March, June, September and December)
Publication Language	:	English
DOI	:	https://doi.org/10.37905/jjbm
Online ISSN	:	2723-0317
Editor-in-Chief	:	Hasan S. Panigoro
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jjbm/oai
Google Scholar ID	:	XzYgeKQAAAAJ
Email	:	editorial.jjbm@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

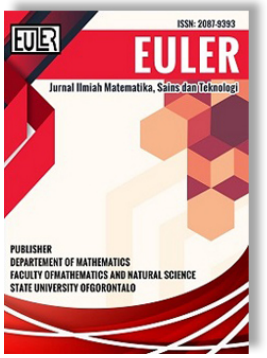


Jambura Journal of Mathematics

Jambura Journal of Mathematics Education



Jambura Journal of Probability and Statistics



EULER : Jurnal Ilmiah Matematika, Sains, dan Teknologi

Jambura Journal of Biomathematics, Volume 6, Issue 3, Pages 234–240, September 2025 https://doi.org/10.37905/jjbm.v6i3.30552

Research Article

Check for updates

Implementation of K-Prototypes with Feature Selection in Clustering Cervical Cancer Patients based on Risk Factors

Wanda Puspita Hati[1], **Devvi Sarwinda**[1], **and Bevina Desjwiandra Handari**[1,*]

[1]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok 16424, Indonesia*

ARTICLE HISTORY
Received 7 February 2025
Revised 19 March 2025
Accepted 19 August 2025
Published 12 September 2025

KEYWORDS
K-Prototypes
Variance threshold
Correlation coefficient
Cervical cancer risk factor
Clustering

ABSTRACT. *Cancer is a leading cause of death worldwide, resulting in nearly 10 million deaths or almost one-sixth of all deaths in 2020. Effective primary prevention measures can prevent at least 40% of cancer cases. Cancer mortality rates are higher in developing countries than in developed countries, reflecting disparities in addressing risk factors, detection success, and available treatments. Women in developing countries most frequently suffer from cervical cancer. It is crucial for communities, especially women, to have knowledge about the risk factors for cervical cancer. One potential solution to this issue is the role of machine learning in analyzing cervical cancer patient data. This study uses the K-Prototypes clustering algorithm, which can cluster mixed data, both numerical and categorical. Cervical cancer risk factor data were used in this research. Feature selection was performed to improve the performance of the K-Prototypes algorithm, using feature selection methods Variance Threshold and Correlation Coefficient. The best performance of the K-Prototypes algorithm was obtained using the Correlation Coefficient, as reviewed based on a Silhouette Coefficient of 0.6, a Davies-Bouldin Index of 0.6, and a Calinski-Harabasz Index of 1.080. Interpretation of the clusters formed revealed major differences in the characteristics of risk factors between two clusters, namely age, menopause, and health conditions such as leukorrhea, bleeding, lower abdominal pain, and loss of appetite. Meanwhile, factors related to previous history, reproductive health, and nutritional issues did not show significant differences. The K-Prototypes algorithm is expected to be a solution in identifying groups based on cervical cancer risk factors to assist medical professionals in decision-making and subsequent actions, as well as to provide knowledge to the public.*

1. Introduction

According to the World Health Organization (WHO), cancer was responsible for nearly 10 million deaths, or about one-sixth of all deaths in 2020. One in five people globally will develop cancer during their lifetime. Preventing cancer is one of the biggest public health challenges of the 21st century. The mortality rate from cancer is higher in developing countries compared to developed countries. This discrepancy indicates variations in risk factor management, detection effectiveness, and treatment availability [1]. The modeling of cancer dynamics can be found in [2].

Cancer is a non-communicable disease characterized by the abnormal growth of malignant tissues or cells in the body. Cancer cells proliferate rapidly, invade nearby areas, and spread to other parts of the body. Cervical cancer affects the cervical area of the uterus (cervical neck), a region in the female reproductive organ that serves as the entrance to the uterus, located between the uterus and the vaginal canal (vagina) or the lower uterus [3]. One of the main triggers of cervical cancer is infection with the Human Papillomavirus (HPV), which is commonly spread through sexual contact with individual [4].

In developing countries, cervical cancer is the most prevalent among women [5]. According to the Indonesian Ministry of Health's estimates in 2018 [6], the incidence of cervical cancer in women, as new cases, ranged from 90 to 100 per 100,000 people, with about 40,000 new cases occurring each year. Furthermore, in 2022, WHO reported that cervical cancer ranked as the fourth most common cancer among women worldwide, with an estimated 660,000 new cases and 350,000 deaths [1]. Society, especially women, needs knowledge about cervical cancer risk factors; however, many lack this knowledge due to limited access to healthcare services and a lack of understanding about the disease [7]. The impact of cervical cancer cases in Indonesia poses a significant challenge for patients and their families, healthcare workers, and potentially places a substantial financial burden on the government. Therefore, increasing efforts in prevention and early detection are crucial for all parties involved.

Machine learning plays an essential role in efficiently and accurately identifying data related to cervical cancer risk factors. One approach to problem-solving using machine learning is clustering. P. Gupta et al. did detection and subsequent prevention using data mining techniques on patient information to predict the occurrence of cervical cancer [8]. In addition, previous research have applied clustering algorithm for patient's data from a disease, such as Clustering Cervical Cancer based on Comparison between Euclidean and Manhattan using K-Means Method [9], Data Clustering Mining Applying the K-Means Algorithm, Cer-

*Corresponding Author.

Email : *bevina@sci.ui.ac.id* (B. D. Handari)
Homepage : http://ejurnal.ung.ac.id/index.php/JJBM/index / E-ISSN : 2723-0317

Table 1. Cervical cancer stage

Stage	Information
I	The carcinoma is strictly confined to the cervix
IA	Invasive carcinoma that can be diagnosed only by microscopy, with maximum depth of invasion $\leq 5\ mm$
IA1	Measured stromal invasion $\leq 3\ mm$ in depth
IA2	Stromal invasion measuring $\geq 3\ mm$ and $< 5\ mm$ in depth
IB	Invasive carcinoma with stromal invasion $\geq 5\ mm$, confined to the cervix uteri
IB1	Invasive carcinoma with stromal invasion $\geq 5\ mm$ and $< 2\ cm$ in greatest dimension
IB2	Invasive carcinoma measuring $\geq 2\ cm$ and $< 4\ cm$ in greatest dimension
IB3	Invasive carcinoma $\geq 4\ cm$ in greatest dimension
II	Tumor extends beyond the uterus but not to the pelvic wall
IIA	Without parametrial invasion
IIA1	Invasive carcinoma $\leq 4\ cm$ in greatest dimension
IIA2	Invasive carcinoma $> 4\ cm$ in greatest dimension
IIB	With parametrial invasion
III	Tumor involves the pelvic wall and/or the lower third of the vagina, and/or causes hydronephrosis or non-functioning kidney
IIIA	Tumor involves the lower third of the vagina without extension to the pelvic wall
IIIB	Tumor extends to the pelvic wall and/or causes hydronephrosis or non-functioning kidney
IIIC	Tumor involves pelvic and/or para-aortic lymph nodes, regardless of tumor size or extent
IIIC1	Pelvic lymph node metastasis
IIIC2	Para-aortic lymph node metastasis
IV	Extended beyond the true pelvis or has involved (biopsy proven) the mucosa of the bladder or rectum
IVA	Spread of the growth to adjacent pelvic organs
IVB	Spread to distant organs

vical Cancer Behaviour Risk [10], Clustering of Risk Factors for Coronary Heart Disease Using The K-Prototypes Algorithm [11], Innovative Incremental K-Prototypes Based Feature Selection for Medicine and Healthcare Applications [12], Cluster Analysis of Diabetes Patients' Data for Identify Pattern and Characteristics Patients [13], Comparison of Various clustering techniques for diagnosis of breast cancer using DBSCAN and Hierarchical Clustering [14], Pre Cervical Cancer Detection on Visual Inspection of Acetic Acid (VIA) Test Image Using K-Means Clustering Method [15], and Optimization of K-Means Attribute Selection Using Correlation Matrix in Patient Disease Clustering [16].

This study uses data on the risk factors of cervical cancer patients at X hospital in DKI Jakarta from 2021 to 2023 with 1166 observations and 36 features. The K-Prototypes algorithm was chosen due to its ability to handle issues related to mixed data types, both numerical and categorical. Feature selection is conducted to improve the algorithm's efficiency. Two feature selection methods are used: the Variance Threshold and the Correlation Coefficient. The Variance Threshold can eliminate features with low variability which do not provide much information for clustering, and the Correlation Coefficient method reduced feature pairs with high correlation by removing one of the correlated features, as they are considered to provide redundant information for clustering. Both feature selection methods were chosen because they are effective for both numerical and categorical features, and also do not require labeled data. The performance of the K-Prototypes algorithm will be evaluated by reviewing the Silhouette Coefficient, Davies-Bouldin Index, Calinski-Harabasz Index, and Running Time.

2. Research Methodology

2.1. Stage and Risk Factors of Cervical Cancer

According to the Indonesian Ministry of Health's Guidelines for the Management of Cervical Cancer, issued in 2018, the initial step to reduce cervical cancer incidence is to identify risk factors. Risk factors are any factors that increase the likelihood of developing a disease [17]. The risk factors that may increase the likelihood of developing cervical cancer, include early age at first sexual intercourse, multiple sexual partners, smoking, high parity, long-term use of oral contraceptive pills, low economic status, co-infection with sexually transmitted infections, unprotected sexual intercourse, immune system disorders, HPV genotypes 16 and 18, co-infection with HIV, male circumcision, and poor diet [18].

The stage of cervical cancer is determined clinically by considering the size of the tumor and the extend of its spread. A high stage indicates that the cancer has spread extensively. A low stage indicates that the cancer has not spread extensively [19, 20]. The cervical cancer stage can be found in Table 1.

2.2. Feature Selection

Feature selection is a feature learning process where the goal is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results [21]. In this research, feature selection using Variance Threshold and Correlation Coefficient was used. Variance Threshold determines the variance threshold and eliminating feature with a variance below the threshold which can be calculated by eq. (1):

$$Var(x_j) = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n}, \tag{1}$$

where $x_j,\ \ j = 1,\ 2,\ \ldots,\ m$ are individual values from all features, $\bar{x}_j,\ \ j = 1,\ 2,\ \ldots,\ m$ are mean values from all features, and n are total observation.

On the other hand, Correlation Coefficient is the statistical measure that indicates the strength of the linear relationship between two features. Two of the Correlation Coefficient methods are Pearson and Kendall. Pearson Correlation (r) measures

Table 2. K-Prototypes workflow

Step	Description
1	Determine the number of clusters (k) to be formed. In this research, we used Elbow method.
2	Select the centroids (c) from the dataset randomly.
3	Calculate the distance for each data point based on the distance measure in eq. (4). Assign each observation to the cluster closest to it.
4	Re-evaluate the similarity of observations to the current centroid. If any observation is found to be closer to another cluster than its current cluster, reassign the observation to the closer cluster and update the cluster memberships.
5	Repeat steps 2, 3, and 4 until no observations change clusters or convergence is achieved. Convergence criteria indicate that there are no changes in cluster memberships during the subsequent iteration.

Table 3. SC Value Interpretation

SC value	interpretation
$0.7 < SC \leq 1.0$	There is a "very strong bond" between observations and the formed cluster.
$0.5 < SC \leq 0.7$	There is a "fairly strong bond" between observations and the formed cluster.
$0.25 < SC \leq 0.5$	There is a "weak relationship" between observations and the formed cluster.
$SC \leq 0.25$	There is "no bond" between observations and the formed cluster.

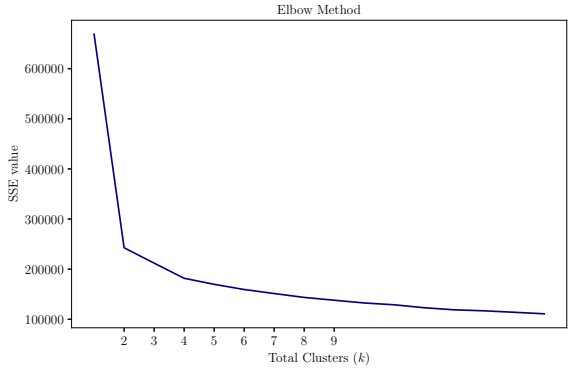


Figure 1. Elbow method.

the strength and direction of the linear relationship between two continuous variables that are normally distributed which can be calculated by eq. (2):

$$r_{xy} = \frac{n\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left[n\sum x_i^2 - \left(\sum x_i\right)^2\right]\left[n\sum y_i^2 - \left(\sum y_i\right)^2\right]}}, \tag{2}$$

where n are total observations, x_i is i-th values of feature x, y_i is i-th values of feature y, and $x_i y_i$ is i-th values of feature x, y.

Kendall Correlation (τ) measures association based on the difference between the probability of concordance and discordance between two observed features, X and Y, which can be calculated by eq. (3):

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}, \tag{3}$$

where C is the number of concordant pairs, D is the number of discordant pairs, and n is the number of observation.

2.3. *K-Prototypes Algorithm*

Clustering is the process of dividing a large set of data points into smaller clusters so that the data points in each cluster have unique characteristics. The K-Prototypes algorithm, proposed by Huang in 1998 [22], is an extension of the K-Means and K-Modes algorithms. K-Prototypes groups a mixed dataset containing both numerical and categorical features into k different clusters [12]. It is well-known as a non-hierarchical clustering algorithm that handles mixed data due to its clear, measurable, and convergent capabilities. It introduces a new representation of the cluster center (centroid) and provides a new definition of the similarity measure between observations and the centroid which can be calculated by eq. (4):

$$d(x_i^m, z_i^m) = \sum_{l=1}^{p_r}(x_{ij}^r - z_{lj}^r)^2 + \gamma_l \sum_{j=l+1}^{p_t} \delta(x_{ij}^t, z_{lj}^t), \tag{4}$$

where

$d(x_i^m, z_i^m)$: distance of i-th mixed observation to l-th cluster centroid,
$\sum_{l=1}^{p_r}(x_{ij}^r - z_{lj}^r)^2$: Euclidean Distance used in K-Means,
$\sum_{j=l+1}^{p_r}\delta(x_{ij}^t, z_{lj}^t)$: Simple Matching used in K-Modes,
γ_l : standard deviation of numeric feature in each l-th cluster,
p_r : total numeric features,
p_t : total categorical features.

The workflow of the K-Prototypes algorithm is detailed in Table 2.

Based on step 1, the Elbow method is used to determine the optimal number of clusters by observing a point where the number of clusters forms an elbow by calculating the Sum of Squared Error (SSE) for each number of clusters as in eq. (5). If the

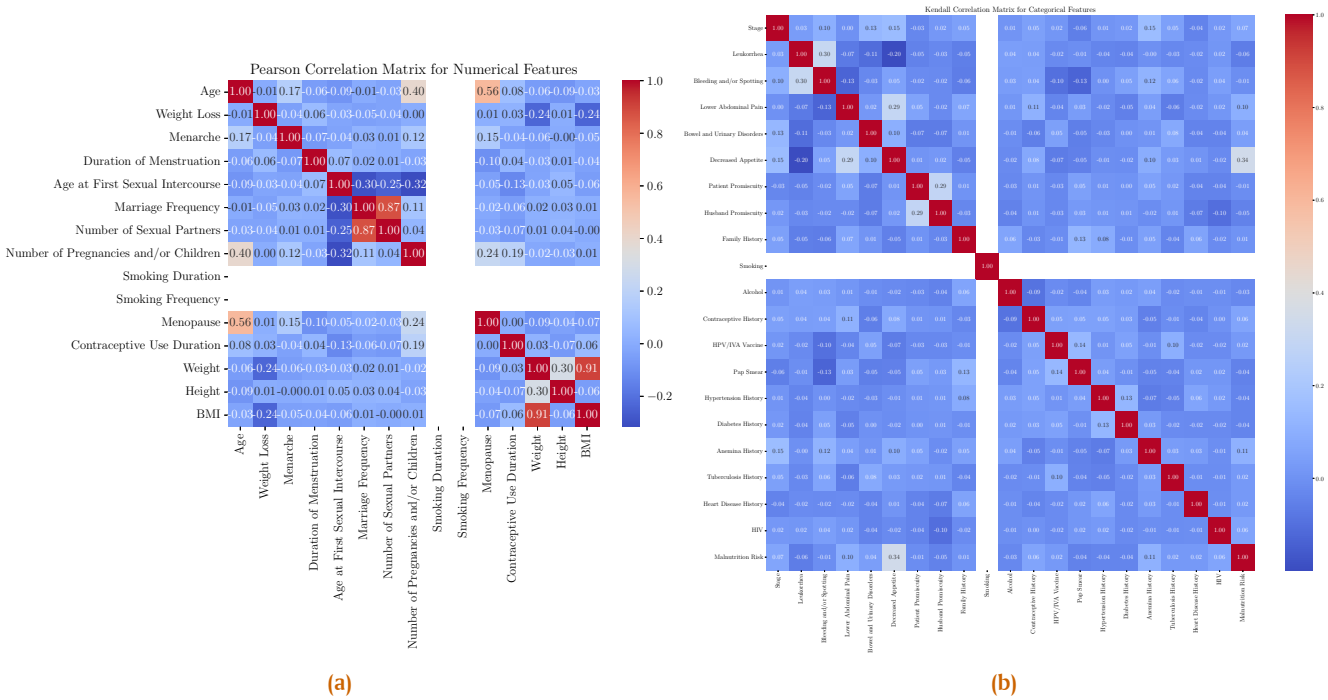


Figure 2. (a) Pearson correlation. (b) Kendall correlation.

Table 4. Performance evaluation of K-Prototypes

Performance Index	Without Feature Selection	Variance Threshold	Correlation Coefficient
Silhouette Coefficient	0.543	0.544	0.604
Davies-Bouldin Index	0.723	0.721	0.600
Calinski-Harabasz Index	1235.650	1240.673	1808.677
Running Time (s)	16.907	3.433	4.048

plot or the values of SSE show a significant decrease between the first and second clusters, or if there is a noticeable change in the slope of the SSE curve, then that number of clusters is considered optimal [23].

$$SSE = \sum_{k=1}^{k} \sum_{i=1}^{n} \|x_{i,k} - u_k\|^2, \tag{5}$$

where k is the total number of clusters, n is the total number of observations, $x_{i,k}$ is the feature value from i-th observation from k-th cluster, and u_k is the centroid from k-th cluster.

2.4. *Performance Evaluation of Clustering*

To evaluate the performance of clustering algorithms, we used the Silhouette Coefficient, Davies-Bouldin Index, and Calinski-Harabasz Index. Silhouette Coefficient (SC) is the quality of clustering can be assessed by measuring the distance between each observation within a cluster to points in other clusters, evaluating how well each observation fits its assigned cluster [11]. Table 3 shows the interpretation of the cluster formed based on SC value [24].

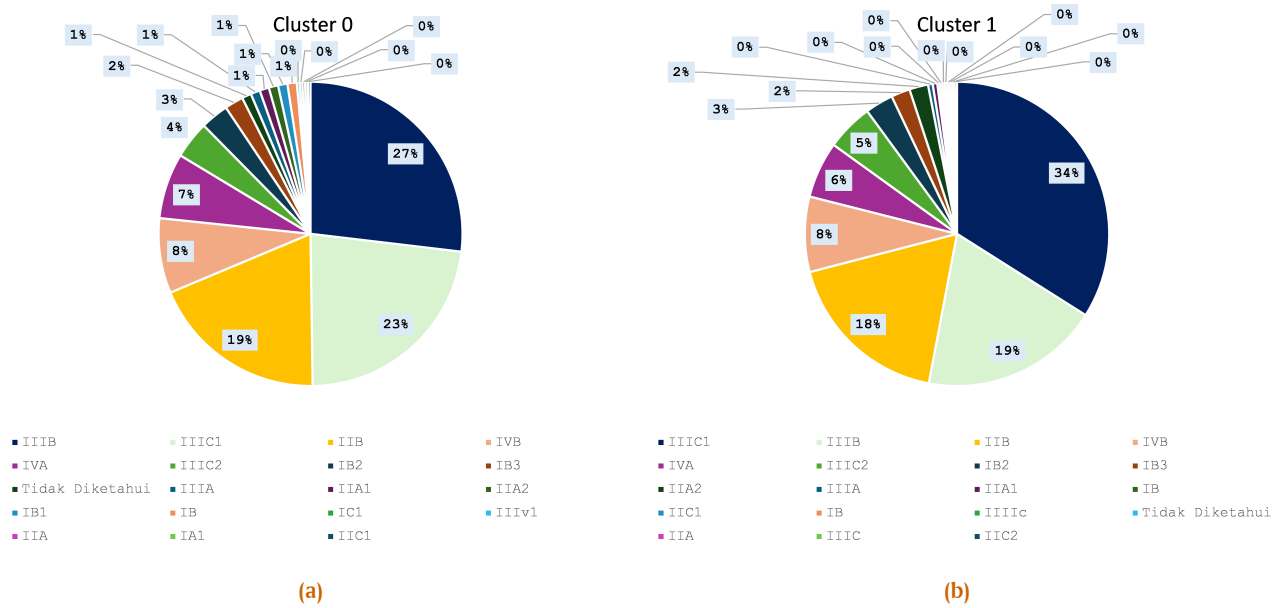
Davies-Bouldin Index (DBI) is measures the ratio of the average spread within clusters to the average spread between clusters [25]. A lower DBI value indicates that clusters are well-defined. On the other hand, Calinski-Harabasz Index (CHI) is a ratio between the between-cluster dispersion and within-cluster dispersion [12]. A higher CHI value indicates that the clustering model is well-defined.

3. Result And Discussion

In this study, risk factor data from cervical cancer patients consisting of 1166 observations with 36 features, including 15 numerical features (e.g. Weight Loss, Height, BMI) and 21 categorical features were used. Risk factors are analyzed by the K-Prototypes algorithm to produce several clusters, each with unique characteristics, and distinct characteristics between clusters. The effectiveness of the K-Prototypes algorithm is expected to demonstrate that it can serve as an alternative solution in applying machine learning to identify groups of cervical cancer risk factors, aiding in the development of cervical cancer screening techniques for healthcare professionals and enhancing public knowledge. Implementation of K-Prototypes in clustering cervical cancer patients based on risk factors is conducted with three simulations: 1) K-Prototypes Implementation without Feature Selection, 2) K-Prototypes Implementation with Variance Threshold, and 3) K-Prototypes Implementation with Correlation Coefficient.

3.1. *Data Preprocessing*

LabelEncoder from the scikit-learn (sklearn) library in Python is used for the Stage, Patient Promiscuity, and Husband's Promiscuity features. Meanwhile, other categorical features have



(a) (b)

Figure 3. Patients distribution based on cancer stage in each cluster.

already been converted to numerical form such as 0 and 1. Missing values were identified, three features with the highest percentage of missing values: Menopause (years) (35%), Duration of Menstruation (days) (31%), Number of Sexual Partners (15%). Missing values for numerical features were imputed using the mean, categorical features imputed by mode, and Menopause ans Age at First Sexual Intercourse imputed by median. Outliers were identified using Interquartile Range (IQR) method. Two features with the highest outlier percentage are Smoking Duration (years) and Smoking Frequency (cigarettes/day) with percentages of 20.92%. Observations that detected an outlier based on IQR method were eliminated.

3.2. *K-Prototypes Implementation with Variance Threshold*

The first simulation uses variance threshold as the feature selection method. The variance from each feature was calculated. For numerical features, the threshold was set by the 25th quantile or first quantile (Q_1) of all variance values, which is 0.425. As a result, 4 numerical features were deleted. For categorical features, threshold set by the 50th percentile or median of all variance values, which is 0.186. As a result, 11 categorical features were deleted. Based on the variance threshold, 36 features became 21 features.

Elbow method were used to determine the optimal number of clusters, as shown in Figure 1, where the elbow formed on $k = 2$. Furthermore, K-Prototypes algorithm implemented with $k = 2$ resulted in 342 patients (47.7%) in cluster 0 and 380 patients (52.6%) in cluster 1.

3.3. *K-Prototypes Implementation with Correlation Coefficient*

The Pearson Correlation matrix in Figure 2a. shows that Marriage Frequency-Number of Sexual Partners has high correlation 0.87 and Weight (kg)-BMI $\left(\frac{kg}{m^2}\right)$ has high correlation 0.91. Marriage Frequency and Weight (kg) were chosen for deletion. Smoking Duration (years) and Smoking Frequency (cigarettes/day) do not show correlation values because all values are 0 or all patients do not smoke. Since these values are constant, both features were removed as they do not provide information.

The Kendall Correlation matrix in Figure 2b was used for categorical features and there are no categorical feature pairs with high correlation. Smoking does not show correlation values because all values are 0 or all patients do not smoke. Since these values are constant, smoking was removed because it provides no information for the clustering process in the next stage.

Based on Correlation Coefficient, 15 numerical features became 11, and 21 categorical features became 20. Hence, the initial total of 36 features became 31. Elbow Method was also used similar to previous simulation, resulting in a total optimum cluster k of 2. K-Prototypes algorithm with k=2 resulted in 342 patients (47.7%) in cluster 0 and 380 patients (52.6%) in cluster 1.

3.4. *Performance Evaluation of K-Prototypes*

The performance of clustering is considered better when the Silhouette Coefficient value approaches 1, clustering performance improves as the Davies-Bouldin Index value approaches 0, and clustering performance is better when the Calinski-Harabasz Index value is higher. Therefore, based on those three indices, the simulation of the K-Prototypes algorithm with feature selection using the Correlation Coefficient performs better compared to the other two simulations.

Clustering performance is better when the running time is faster. In Table 4, the running time includes the time taken for the feature selection process. Running time of the K-Prototypes algorithm using the Variance Threshold and the Correlation Coefficient is much better than the method without feature selection. The simulation of the K-Prototypes algorithm with feature selection using the Correlation Coefficient performs better compared to the other two simulations. The clustering performance between the simulation of the K-Prototypes algorithm without feature selection and with feature selection using the Variance

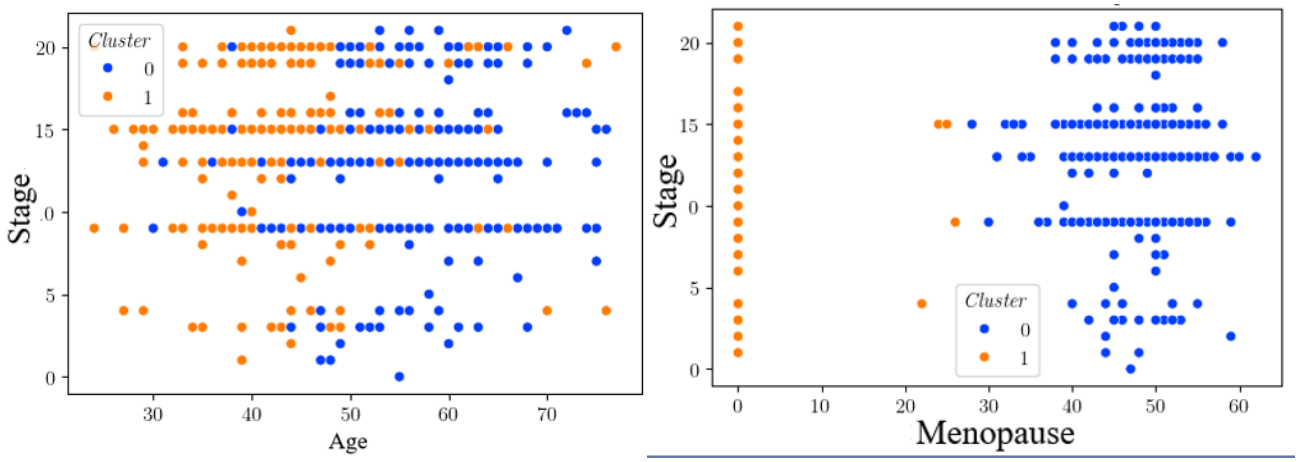


Figure 4. (a) Patient distribution based on age and stage. (b) Patient distribution based on menopause and stage.

Threshold is not very significant, as the Variance Threshold removes many features but does not improve the performance of the K-Prototypes algorithm.

Based on the details above, it can be concluded that the K-Prototypes algorithm for clustering cervical cancer patients based on risk factors performs better with feature selection using the Correlation Coefficient.

3.5. Analysis of Cluster Characteristics

Analysis of the characteristics of each cluster's cervical cancer risk factors as a result of implementing the K-Prototypes algorithm, where risk factors have distinct characteristics between cluster 0 and cluster 1 as shown in Figure 3.

As shown in Figure 3b, cluster 1 is dominated by patients with cervical cancer stage IIIC1 for 34%. In other words, 66% of cluster 1 consists of patients with cervical cancer stages other than stage IIIC1 (patients with pelvic lymph node metastasis). On Figure 3a, cluster 0 is dominated by patients with cervical cancer stage IIIB for 27%. Both clusters have relatively few patients in the early stages (stage IA1-IIA2), with a significant increase observed at stage IIB, indicating that early detection may be less common.

As shown in Figure 4a, patients who are older (average age of 55 years in cluster 0) are diagnosed with lower stages of cancer (IIIB). Patients who are younger (average age of 45 years in cluster 1) are diagnosed with more advanced stages of cancer (IIIC1). Hence, older age does not necessarily correlate with a more advanced cancer stage. As shown on Figure 4b, patients who are pre-menopausal (cluster 1) are diagnosed with more advanced stages of cancer (IIIC1). Patients who have experienced menopause at the age of 48 (cluster 0) are diagnosed with lower stages of cancer (IIIB). Hence, menopause likely affects the diagnosis of cancer stages.

Based on clinical symptoms including bleeding and/or spotting and leukorrhea, these symptoms are more likely in cluster 1 where most patients are at stage IIIC1. On the other hand, based on clinical symptoms including lower abdominal pain and decreased appetite, these symptoms are more likely in cluster 0 where most patients are at stage IIIB. The main differences between the two clusters lie in age, menopausal status, and several health conditions such as leukorrhea, bleeding and/or spotting, lower abdominal pain, and decreased appetite. Risk factors related to past medical history, reproductive health, and nutritional issues may not be significant differentiating factors between the two clusters.

4. Conclusion

Implementation of K-Prototypes in clustering cervical cancer patients based on risk factors aims to produce several clusters, each with unique characteristics, and distinct characteristics between clusters. The best-performing algorithm for K-Prototypes involved feature selection using Correlation Coefficient. The initial number of 36 features became 31 features after feature selection. The Elbow method used for determined the optimal number of clusters which is two clusters. Hence, the implementation of K-Prototypes with number clusters are two resulted patients in cluster 0 are 47.4% (342 patients) and patients in cluster 1 are 52.6% (380 patients). Key performance indices for the best K-Prototypes algorithm from the simulations are: 1) Silhouette Coefficient: 0.6; 2) Davies Bouldin Index: 0.6; 3) Calinski-Harabasz Index: 1.806; and 4) Running Time: 4.048 s. It was found that the Variance Threshold feature selection method did not significantly improve performance.

The implementation of K-Prototypes for clustering cervical cancer patients based on risk factors using Correlation Coefficient feature selection yielded good performance. Primary differences between the two clusters are age, menopause, leukorrhea, bleeding and/or spotting, lower abdominal pain, and decreased appetite. However, the simulations did not fully differentiate the characteristics of risk factors between clusters, e.g. for risk factors such as reproductive health and nutritional issues.

Author Contributions. **Hati, W. P.**: software, data curation, formal analysis, writing original draft preparation. **Sarwinda, D.**: software, methodology, validation, writing review, supervision. **Handari, B. D.**: Conceptualization, methodology, validation, writing review and editing, project administration, funding acquisition.

Acknowledgement. The authors gratefully acknowledge the support of

the editors, reviewers, and colleagues in both the research process and the preparation of this manuscript.

Funding. This research was funded by Universitas Indonesia, Hibah Publikasi Terindeks Internasional (PUTI) Q2 Fiscal year 2025-2026, No.:PKS-366/UN2.RST/HKP.05.00/2025.

Conflict of interest. The authors declare no competing interests.

Data availability. Not applicable.

References

[1] WHO, "Cervical cancer," https://www.who.int/news-room/fact-sheets/detail/cervical-cancer, 2024, Accesed on 7 February.

[2] N. Fitriyati, S. A. Faizah, and T. E. Sutanto, "Prediction of the change rate of tumor cells, healthy host cells, and effector immune cells in a three-dimensional cancer model using extended kalman filter," *Jambura Journal of Biomathematics (JJBM)*, vol. 5, no. 1, pp. 27–37, 2024. DOI:10.37905/jjbm.v5i1.24672

[3] Misgiyanto and D. Susilawati, "Hubungan antara dukungan keluarga dengan tingkat kecemasan penderita kanker serviks paliatif," *Jurnal Keperawatan*, vol. 5, no. 1, pp. 1–15, 2014. DOI:10.22219/jk.v5i1.1855

[4] I. Rasjidi, "Epidemiologi kanker serviks," *Indonesian Journal of Cancer*, vol. 3, no. 3, 2009. DOI:10.33371/ijoc.v3i3.123

[5] F. Hardiyanti, J. Harlan, and E. Hermawati, "The association between knowledge and preventive behavior of cervical cancer among woman employees in the companies in jakarta," *Indonesian Journal of Cancer*, vol. 14, no. 1, pp. 8–15, 2020. DOI:10.33371/ijoc.v14i1.666

[6] M. K. R. INDONESIA, "Pedoman nasional pelayanan kedokteran tata laksana kanker serviks," *Ministry of Health of the Republic of Indonesia*, 2018.

[7] D. Y. C. Sogukkuyu and O. Ata, "Diagnosing cervical cancer using machine learning methods," in *HORA 2022 - 4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*, pp. 1–3, 2022. DOI:10.1109/HORA55278.2022.9800033

[8] P. Gupta, I. Jindal, and A. Goyal, "Early detection and prevention of cervical cancer," in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pp. 1–4, 2019. DOI:10.1109/I2CT45611.2019.9033800

[9] S. Widodo, H. Brawijaya, and S. Samudi, "Clustering kanker serviks berdasarkan perbandingan euclidean dan manhattan menggunakan metode k-means," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 687, 2021. DOI:10.30865/mib.v5i2.2947

[10] R. M. F. Lubis *et al.*, "Data clustering mining applying the k-means algorithm, cervical cancer behavior risk," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 7, no. 2, p. 819, 2023. DOI:10.30865/mib.v7i2.6088

[11] E. S. Setianingsih *et al.*, "Clustering of risk factors for coronary heart disease using the k-prototypes algorithm," in *International Conference on Electrical, Computer, Communications and Mechatronics Engineering, ICECCME 2023*, 2023. DOI:10.1109/ICECCME57830.2023.10252558.

[12] S. Gorrab, F. B. Rejab, and K. Nouira, "Innovative incremental k-prototypes based feature selection for medicine and healthcare applications," *Smart Innovation, Systems and Technologies*, pp. 282–291, 2023. DOI:10.1007/978-981-99-3311-2_25

[13] A. E. Satriatama *et al.*, "Analisis klaster data pasien diabetes untuk identifikasi pola dan karakteristik pasien," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 5, no. 3, pp. 172–182, 2023. DOI:10.47233/jteksis.v5i3.828

[14] R. D. H. Devi and P. Deepika, "Performance comparison of various clustering techniques for diagnosis of breast cancer," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1–5, 2015. DOI:10.1109/ICCIC.2015.7435711

[15] R. Ariyani *et al.*, "Pre cervical cancer detection on visual inspection of acetic acid (via) test image using k-means clustering method," in *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, pp. 131–135, 2020. DOI:10.1109/ICIMCIS51567.2020.9354317

[16] A. Bengnga, R. Ishak, "Optimalisasi seleksi atribut k-means menggunakan correlation matrix pada clustering penyakit pasien optimization of k-means attribute selection using correlation matrix in patient disease clustering," *Jambura Journal of Electrical and Electronics Engineering*, vol. 7, no. 2, pp. 141–148, 2025. DOI:10.37905/jjeee.v7i2.28010

[17] I. J. Ratul *et al.*, "Early risk prediction of cervical cancer: A machine learning approach," in *19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2022*, pp. 1–4, 2022. DOI:10.1109/ECTI-CON54298.2022.9795429

[18] O. M. Omone and M. Kozlovszky, "The associations between hpv-infections associated risk factors and cervical cancer associated risk factors using chi-square method," in *INES 2022 - 26th IEEE International Conference on Intelligent Engineering Systems 2022, Proceedings*, pp. 225–230, 2022. DOI:10.1109/INES56734.2022.9922618

[19] P. A. Cohen *et al.*, "Cervical cancer," *The Lancet*, vol. 393, no. 10167, pp. 169–182, 2019. DOI:10.1016/S0140-6736(18)32470-X

[20] M. Saleh *et al.*, "Cervical cancer: 2018 revised international federation of gynecology and obstetrics staging system and the role of imaging," *American Journal of Roentgenology*, vol. 214, no. 5, pp. 1182–1195, 2020. DOI:10.2214/AJR.19.21819

[21] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[22] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 12, pp. 283–304, 1998. DOI:10.1023/A:1009769707641

[23] Z. R. Fadilah and A. W. Wijayanto, "Perbandingan metode klasterisasi data bertipe campuran: One-hot-encoding, gower distance, dan k-prototype berdasarkan akurasi (studi kasus: Chronic kidney disease dataset)," *Journal of Applied Informatics and Computing*, vol. 7, no. 1, pp. 63–73, 2023. DOI:10.30871/jaic.v7i1.5857

[24] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. DOI:10.1016/0377-0427(87)90125-7

[25] S. Gorrab, F. B. Rejab, and K. Nouira, "Real-time k-prototypes for incremental attribute learning using feature selection," *Machine Learning and Data Analytics for Solving Business Problems*, pp. 165–187, 2022. DOI:10.1007/978-3-031-18483-3_9