Jurnal Aplikasi Sains Data
Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Application of K-Prototypes Clustermix Algorithm for Clustering Risk Factors of Diabetes Disease

Martina Hildha[1], Hasih Pratiwi[2] Etik Zukhronah[3]
[1]Statistics Department of Universitas Sebelas Maret
[1]mhildha.mh@student.uns.ac.id

ABSTRACT

Abstract— Diabetes mellitus (DM) is recognized as one of the most rapidly increasing chronic diseases worldwide, posing a significant public health challenge. According to the International Diabetes Federation (IDF), approximately 537 million people were living with diabetes mellitus globally, with projections estimating a rise to 643 million by 2030 and 783 million by 2045. Additionally, the World Health Organization (WHO) reported a 3% increase in mortality rates attributed to diabetes mellitus between 2000 and 2019, underscoring the urgent need for effective risk detection and management strategies. Early identification of risk factors is crucial to mitigating the impact of DM, and clustering analysis offers a promising method for stratifying patients based on risk profiles. This study employs the k-prototypes algorithm, which is particularly suited to clustering datasets with mixed numeric and categorical variables, to analyze DM risk factors. Utilizing data from the 2022 Behavioral Risk Factor Surveillance System (BRFSS) annual survey, the study examines a sample of 2,480 diabetes mellitus patients across the United States. The clustering analysis identified two optimal clusters (k=2) based on a high silhouette score of 0.821, indicating strong cluster cohesion and separation. Cluster 2, consisting of 77 patients, exhibited a higher risk profile for diabetes compared to Cluster 1, which included 2,403 patients. The clusters were characterized by significant differences in average values of key DM risk factors including weight, fruit and vegetable consumption, mental and physical health status, age, alcohol consumption, hypertension, smoking status, physical activity, mobility difficulties, sex, education level, income, and ethnicity. These findings highlight the utility of k-prototypes clustering in identifying high-risk DM subgroups to inform targeted prevention and intervention efforts.

Keywords: diabetes mellitus (DM), mixed data types, clustering, silhoutte score

I. INTRODUCTION

Diabetes mellitus has become one of the chronic diseases with the fastest growing number of patients in the world. According to data released by the International Diabetes Federation (IDF) in the 10th edition of the Atlas publication [1], 537 million people are estimated to have diabetes mellitus, this is predicted to continue to increase to reach 643 million by 2030, and 783 million by 2045. In 2019, diabetes mellitus was the direct cause of 1.5 million deaths, with 48% of total diabetes deaths occurring before the age of 70. Another 460,000 deaths from kidney disease were also attributed to diabetes mellitus, while deaths from elevated blood glucose levels accounted for 20% of total deaths from cardiovascular disease [2].

Diabetes Mellitus is a chronic disease due to the inability of the pancreas to produce enough insulin, as well as manage the use of insulin effectively [3]. Insulin has an important role in managing blood sugar levels in the body. When the body experiences an uncontrolled increase in blood sugar or called hyperglycemia, this can cause various kinds of damage to the body system, especially to vital parts of blood vessels and nerves. According to Ksanti, diabetes has two main types of factors: modifiable factors such as weight, physical activity, hypertension, diet, and dyslipidemia; and non-modifiable factors such as heredity.

* *Corresponding author.*
E-mail address: mhildha.mh@student.uns.ac.id
https://doi.org/10.33005/jasid.v1i1.9

Grouping of risk factors can be done using clustering analysis. Clustering analysis is a part of data mining that functions to group an object based on characteristics that are similar and different from other objects. Clustering is divided into two types, namely hierarchical and non-hierarchical. Hierarchical is done by grouping an object based on a similarity measure, then it will continue on other objects until the cluster will be formed like a tree at the most similar to least similar level. Non-hierarchical clustering is done with initial cluster initiation, and continues without following the hierarchical process [5].

Diabetes risk factors are taken from diabetes patient data including demographic information, medical history, and lifestyle factors. Some of these factors may have different data types, which can be numeric and also categorical so that they can be categorized in mixed data. k-prototypes is a non-hierarchical clustering method that combines k-means (for numeric data) and k-modes (for categorical data) methods, making it suitable for use on datasets containing variables with mixed data types called ClusterMix, both numeric and categorical [6]. In comparison, k-prototypes is a non-hierarchical clustering method that focuses on data clustering applications specifically with mixed attributes (numeric and categorical).

Based on the background explanation above, researchers are interested in discussing mixed data cluster analysis or ClusterMix using the k-prototypes method on DM patient data in the US region. Determination of the k value (number of clusters) is done by measuring the similarity of the distance between clusters with the highest sillhoutte score. The purpose of this study is to obtain optimal clustering results and form patient clusters based on the similarity distance of each variable for the benefit of further handling of DM patients. The clustering results are also expected to be able to help make decisions and actions that will be taken by medical personnel to minimize the risk of DM. Meanwhile, the utilization of machine learning in this study is a learning or reference to the use of the ClusterMix k-prototypes method to handle mixed data.

II. RESEARCH METHODS

A. Data

The data used in this study are secondary data from the Behavioral Risk Factor Surveillance System in 2022 in 50 states in the District of Columbia and three regions in the US. This data is a health survey data regarding diabetes mellitus, which is 2480 data with 15 variables.

This study uses 15 variables that become risk factor variables that cause diabetes mellitus. The research variables used in the analysis are presented in Table I.

TABLE I. RESEARCH VARIABLES

No	Variable	Description	Data Type
1	*Weight*	Body weight (kg)	*Numeric*
2	Fruits	Total fruit consumption per day	*Numeric*
3	Veggies	Number of fruit vegetables per day	*Numeric*
4	MentHlth	Number of days that you feel mentally healthy in a month	*Numeric*
5	PhysHlth	Number of days that you feel physically well within one month	*Numeric*
6	Age	Age at first onset of diabetes (years)	*Numeric*
7	HvyAlcoholConsump	Average number of drinking glasses per day	*Numeric*
8	HighBP	0 : *No* 1 : *Yes*	*Kategoric*
9	*Smoker*	0 : *Never* 1 : *Former* 2 : *Someday* 3 : *Everday*	*Kategoric*
10	*PhysActivity*	0 : *No* 1 : *Yes*	*Kategoric*
11	*DiffWalk*	0 : *No* 1 : *Yes*	*Kategoric*

12	*Sex*	0 : *Female* 1 : *Male*	*Kategoric*
13	*Education*	0 : Tidak bersekolah 1 : SD 2 : SMP 3 : SMA 4 : D3/D4 5 : S1	*Kategoric*
14	*Income*	0 : < $10,000 1 : $10,000 < $15,000 2 : $15,000 < $20,000 3 : $20,000 < $25,000 4 : $25,000 < $35,000 5 : $35,000 < $50,000 6 : $50,000 < $75,000 7 : $75,000 < $100,000 8 : $100,000 < $150,000 9 : $150,000 < $200,000 10 : > $200,000	*Kategoric*
15	*Ethnic*	0 : *Other Race* 1 : *White* 2 : *Black* 3 : *Asian* 4 : *Amarican Indian* 5 : *Hispanic*	*Kategoric*

B. Research Stagees

This research uses the k-prototypes method to perform clustering analysis of diabetes mellitus patients. This clustering method is carried out using Python software. The clustering analysis steps using the k-prototypes method are described in Figure 1.

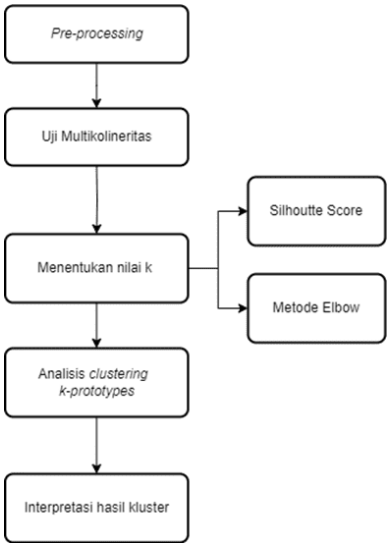

Figure 1. Flowchart

Figure 1 is a water diagram whose steps will be described in more detail as follows:

1. Pre-processing, data exploration which includes describing the data and checking for missing data.
2. Multicollinearity Test.
3. Determine the optimal number of clusters or k by finding the silhouette score and elbow angle in the elbow method.
    - Silhouette Score

      The silhouette score method, introduced by Rousseuw in 1987, is a tool for interpretation and validation of data clustering. The silhouette score has a range from -1 to 1, where values

close to 1 indicate a very appropriate placement of objects in their clusters. Conversely, values close to -1 indicate less appropriate placement of objects [7]. Equation (1) is the formula used to calculate SC:

$$SC = \frac{b_i - a_i}{max\{a_i, a_i\}} \quad (1)$$

a_i: the average distance of the i-th data to all other data in a cluster;

b_i: the average distance of the i-th data to all other data in a cluster;

- Elbhow Method

  The Elbow method is a method of determining the optimal number of clusters can be done through the results of the comparison between the number of clusters that form an angle [8]. If a cluster A and the value in cluster B show a certain angle on the graph in the sense of experiencing a significant decrease, the number of cluster A values is considered the most appropriate. Value comparison is obtained through the calculation of Sum of Square Error (SSE) as in equation (2).

$$SSE = \sum_{k=1}^{k} \sum_{x_i} |x_i - c_k|^2 \quad (2)$$

k : cluster

x_i : distance of i-th object data

c_k : i-th cluster center

4. Clustering DM patient data based on risk factor variables using the k-prototypes algorithm with a mixed distance measure. The distance measure on mixed data types in the k-prototypes method is ideal for data that has two mixed types, namely numeric and categorical. This distance measure is shown in equation (3) [9]. The gamma coefficient (γ) in Equation (3) is calculated based on the average standard deviation (σ) of all numerical variables involved in the study.

$$d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 + \gamma \sum_{s=p+1}^{m} \delta(x_{is}, x_{js}) \quad (3)$$

x_{ik} : i-th observation value on the k-th numerical variable,

x_{jk} : j-th cluster observation value on the kth numeric variable,

p : number of numerical variables

γ : Gamma coefficient

m : number of categorical variables

5. Interpreting the optimal cluster results.

III. RESULTS AND DISCUSSIONS

1. Preprocessing Data

TABLE II. DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

VARIABEL	*MISSING VALUE*	*MIN*	*MED*	*MAX*	*MEAN*	*STD*
WEIGHT	105	31,75	90,26	244,94	92,75	24,3
FRUITS	202	0,00	1,00	100	1,5	6,00
VEGGIES	276	0,00	1,00	203	2,14	7,68
MENTHLTH	38	0,00	0,00	30,0	4,77	9,04
PHYSHLTH	65	0,00	0,00	30,0	7,12	10,8

AGE	181	1,00	50,0	88,0	48,8	14,4
HVYALCOHOLCONSUMP	69	0,00	0,00	720,0	19,56	63,0

Table II. shows the data distribution of all numerical variables and the missing value of each variable, so it needs to be handled before further analysis is carried out. Handling missing value data on numerical variables is done by imputation.

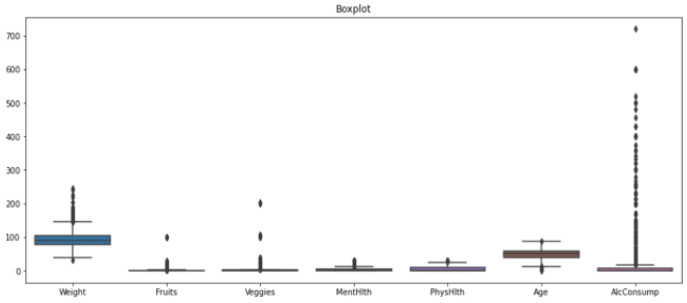

Figure 2. Boxplot of numerical variables

Figure 2 illustrates the distribution of data on numerical variables, where all variables have outlier data so it was decided to use the median value on each variable because the median value is robust to outliers.

TABLE III. DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

Variabel	***Missing Value***	**Keterangan**
HighBP	10	0 = 687 1 = 1783
Smoker	31	0 = 1344 1 = 240 2 = 97 3 = 768
PhysActivity	1	0 = 1141 1 = 1338
DiffWalk	13	0 = 1556 1 = 911
Sex	0	0 = 1244 1 = 1236
Education	3	0 = 8 1 = 107 2 = 165 3 = 707 4 = 720 5 = 770
Income	256	0 = 130 1 = 150 2 = 144 3 = 187 4 = 344 5 = 318 6 = 341 7 = 238 8 = 223 9 = 82 10 = 67
Ethnic	0	0 = 106 1 = 1601 2 = 282 3 = 69 4 = 49 5 = 364

Table III shows the percentage or distribution of data in categorical variables including HighBP, Smoker, PhysActivity, DiffWalk, Sex, Education, Income, and Ethnic variables. In the categorical data used, there are missing values in all categorical variables except the Ethnic and Sex variables. Empty or null data is handled by entering the mode value in the missing value data.

2. Multicollinearity Test

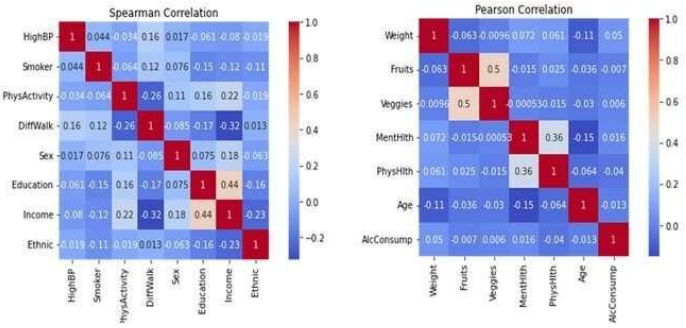

Figure 3. Cross Correlation Heatmap of Numeric Variables, and Categorical Variables

Multicollinearity test is a form of correlation test that exists in all factor variables in the data. If there is multicollinearity, the clusters formed will not be valid. The criterion for multicollinearity is if the relationship coefficient between variables exceeds 0.8. Meanwhile, correlation analysis in the relationship between numerical variables using Pearson correlation and categorical variables using Spearman correlation is used to determine the relationship of each variable. Figure 3 shows the relationship pattern between variables shown in the visual cross corelation heatmap. The redder the color, the stronger the relationship pattern between numerical variables. Figure 3 can be interpreted that between numeric and categorical variables do not have a strong relationship, so the analysis can continue.

3. Determination of Number of Clusters

The determination of the number of clusters (k) is done using the silhouette score and the elbow method. Many clusters with the highest silhouette score will be selected as the optimal cluster value. The following are the results of the Silhoutte Score calculation, namely the number k from 2 to 9 clusters shown in Table IV.

TABLE IV. DESCRIPTIVE STATISTICS OF NUMERICAL VARIABLES

k	*Silhoutte Score*
1	0,8201135
2	0,6439944
3	0,3100345
4	0,3208016
5	0,2311790
6	0,235277
7	0,233861
8	0,248135
9	0,246333

Table IV shows the results of the silhoutte score value of each k (cluster). The best number of clusters can be determined by looking at the highest silhoutte score value. The higher the Silhoutte Score value means that the distance of objects in the cluster is minimal, while the distance of objects with other cluster members is maximum. Based on Table IV, the highest Silhoutte score value is generated at k = 2, which is 0.8201, so that the diabetes mellitus patient data is grouped into 2 clusters.

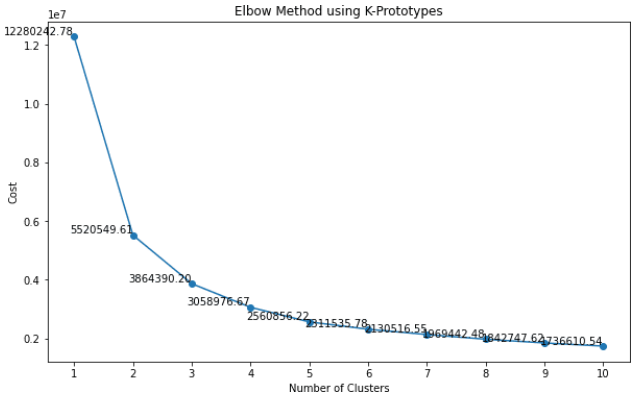

Figure 4. Elbow Method

Figure 4 shows that at k = 1, the SSE value reaches its peak and then begins to decrease significantly at k = 2. When k is at a point after 2, the SSE value decreases constantly. This indicates that right angle formation occurs at k=2.

4. Clustering Analysis with k-prototypes

The application of the k-prototypes algorithm using Python obtained 2 clusters of diabetes mellitus patients with proportions as shown in Figure 5.

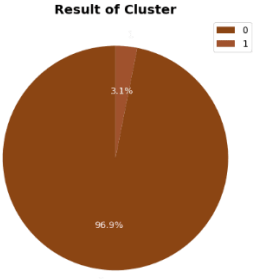

Figure 5. Proportion of Cluster Member Results

Figure 5 shows the results of the clusters formed from the number of k = 2. These results are visualized in the form of a pie chart which can be seen from Figure 4.8. The difference between cluster 1 and cluster 2 is quite significant. The distribution results show that the number of cluster 1 members is 2403 patients or 96.9%, while the number of cluster 2 members is only 77 patients or 3.1%. The Silhoutte Score obtained is 0.82011.

5. Interpretation of Clustering results

The results of clustering analysis using the k-prototypes algorithm are displayed in the form of a comparison table of each cluster on each variable risk factor for diabetes mellitus. The results of the comparison of the characteristics of each cluster can be seen in Table V.

TABLE V. CHARACTERISTICS OF EACH CLUSTERING

Variabel	Cluster 1	Cluster 2	Rata-Rata/ Modus
Weight	92,54	95,96	92,64
Fruits	1,46	1,38	1,46
Veggies	2,01	2,09	2,01
MentHlth	4,68	5,14	4,70
PhysHlth	7,01	4,62	6,93
Age	48,91	48,55	48,9
HvyAlcoholConsump	9,67	310,63	19,02
HighBP	1	1	1
Smoker	0	3	0
PhysActivity	1	1	1
DiffWalk	0	0	0
Sex	0	1	0

Education	5	5	5
Income	5	6	4
Ethnic	1	1	1

Table V shows the salient characteristics of each cluster. Cluster 2 has an average body weight of 95.96, which is greater than the average body weight of all diabetes mellitus patients. While cluster 1, has an average that is almost the same as the average weight of all patients with diabetes mellitus, which is 92.54. The weight indicator is related to BMI where someone with an 'overweight' BMI category is more susceptible to the risk of causing diabetes mellitus.

A diet that involves regular consumption of fruits and vegetables with 2-3 servings per day can reduce the risk of diabetes mellitus [10]. Both clusters have an average of less than 2-3 servings of fruit consumption. Cluster 1 has an average daily fruit consumption of 1.46 servings which is greater than cluster 2 which only has an average of 1.36 servings. Meanwhile, the Veggies variable refers to the amount of daily vegetable consumption of diabetes mellitus patients. Both clusters have a normal average vegetable consumption equal to 2 servings. Cluster 1 has an average daily consumption of vegetables of 2 servings which is less than cluster 2 which only has an average of 2.09 servings.

There is a correlation between high physical and mental health risks for people with diabetes mellitus [11]. This is influenced by good blood sugar management and other factors such as blood pressure and lipids. Cluster 1 has an average number of days of 4.68 days which is less than cluster 2 which only has an average of 5.04 days. Meanwhile, the PhysHlth variable which refers to the number of days in a month where physical health conditions are in good condition shows a significant difference between clusters 1 and 2. Overall diabetes mellitus patients have an average of 6.93 close to 7 days in a period of 30 days (a month). Meanwhile, Cluster 1 has an average number of days of 7.01 or 7 days which is more than cluster 2 which only has an average of 4.62 or close to 5 days.

The Age variable refers to the age when the patient was diagnosed with diabetes mellitus. The results of this study show that the difference between clusters 1 and 2 is not far away and even almost the same. Cluster 1 has an average number of days of 48.91 years which is younger than cluster 2 which only has an average of 48.55 years. Age > 45 years has a risk factor of 1.4 times experiencing abnormal blood sugar levels [12]. This is due to lack of physical activity, increasing body weight, reduced muscle mass and progressive cell shrinkage.

The variable HvyAlcoholConsump which refers to the amount of daily alcohol consumption. The results of this study showed that the differences between clusters 1 and 2 were highly significant. The overall diabetes mellitus patients had an average of 19.02 servings of daily alcohol consumption. Cluster 1 had an average of 9.67 servings of alcohol consumption which was less than cluster 2 which had an average of 310.63 servings. Repeated consumption of large amounts of alcohol is associated with a higher risk for diabetes mellitus [13].

The HighBP variable refers to the history of high blood pressure in patients with diabetes mellitus. Both clusters have a majority of patients who share a history of high blood pressure. A history of elevated systolic and diastolic blood pressure in middle-aged and older adults can increase the risk of diabetes mellitus [14]. PhysActivity variable which refers to physical activity. In this variable, the majority of patients in both cluster 1 and cluster 2 both have physical activity activities. Meanwhile, the DiffWalk variable which refers to complaints of difficulty walking has the majority of patients in both cluster 1 and cluster 2 who do not have complaints of difficulty walking. The Education variable which refers to the last level of education shows that patients in both cluster 1 and cluster 2 are dominated by patients with the last level of education S1 and have the same Ethnic category in the Ethnic variable, namely white.

The Smoker variable or smoking history category shows that all diabetes mellitus patients are dominated by patients who do not have a history of smoking. This also applies to the majority of cluster 1 patients who do not smoke, in contrast to cluster 2, where the majority have a history of active smoking with a high frequency (often). Smoking itself causes resistance to insulin which has an impact on impaired glucose metabolism, thus increasing the risk of diabetes mellitus. The variable Sex or gender shows that overall diabetes mellitus

patients are dominated by women. This also applies to the majority of cluster 1 patients who are predominantly female with an income of $25,000-$35,000, in contrast to cluster 2 which is dominated by males with higher incomes of $35,000-$50,000. According to Damayanti [15], in prevalence women are much more at risk of developing diabetes because physically women have a greater probability of increasing BMI than men.

Income variables show results that are not much different, the majority of cluster 1 patients have an income of $25,000-$35,000, in contrast to cluster 2 which has a higher income of $35,000-$50,000. The higher the income of an individual, the greater the level of lifestyle lived which has an impact on diet, especially the consumption of junk food.

CONCLUSION

The k-prototypes algorithm produced two clusters with a silhouette score of 0.821. Cluster 1 consists of 2403 patients and cluster 2 consists of 77 patients. Cluster 1 shows that the majority of patients are non-smoking females with an income of $25,000-$35,000 and low mean scores on Weight, Veggies, PhysHlth Menthlth, and HvyAlcoholConsump, and high mean scores on PhysHlth, Age, and Fruits variables. Cluster 2 shows that the majority of patients are men who have a smoking habit with high intensity followed by an income of $35,000-$50,000 and have a higher average value than cluster 1 on the variables Weight, Veggies, PhysHlth Menthlth, and HvyAlcoholConsump, and a lower average on the variables PhysHlth, Age, and Fruits. So that patients who are in cluster 2 have a higher risk of diabetes mellitus than patients in cluster 1.

REFERENCES

[1] W. Yudananto, S. S. Remi, and B. Muljarijadi, "Peranan Sektor Pariwisata Terhadap Perekonomian Daerah di Indonesia (Analisis Interregional Input-Output)," Jurnal, vol. 2, no. 4, 2012, Universitas Padjajaran, Bandung.

[2] International Diabetes Federation. *IDF Diabetes Atlas 10th Edition* ; International Diabetes Federation, 2021.

[3] ElSayed, N. A.; Aleppo, G.; Aroda, V. R.; Bannuru, R. R.; Brown, F. M.; Bruemmer, D.; Collins, B. S.; Cusi, K.; Das, S. R.; Gibbons, C. H.; Giurini, J. M.; Hilliard, M. E.; Isaacs, D.; Johnson, E. L.; Kahan, S.; Khunti, K.; Kosiborod, M.; Leon, J.; Lyons, S. K.; Murdock, L. Introduction and Methodology: Standards of Care in Diabetes—2023. *Diabetes Care* **2022**, *46* (Supplement_1), S1–S4.

[4] GBD 2021 Diabetes Collaborators. Global, Regional, and National Burden of Diabetes from 1990 to 2021, with Projections of Prevalence to 2050: A Systematic Analysis for the Global Burden of Disease Study 2021. *The Lancet* **2021**, *402* (10397).

[5] Kshanti, I. A. M. *Pedoman Pemantauan Glukosa Darah Mandiri*, 1st ed.; PB Perkeni: Jakarta, 2019.

[6] Yuan, C.; Yang, H. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J* **2019**, *2* (2), 226–235.

[7] Madhuri, R.; Murty, M. R.; Murthy, J. V. R.; Reddy, P. V. G. D. P.; Satapathy, S. C. Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms. *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India- Vol II* **2014**, *249* (137), 137–144.

[8] Rousseeuw, P. J.; Leroy, A. M. *Robust Regression and Outlier Detection. Wiley Series in Probability and Mathematical Statistics*; John Wiley: New York, 1987.

[9] Bholowalia, P.; Kumar, A. EBK-Means: A Clustering Techiniques Based on Elbow Method and K-Means in WSN. *International Journal of Computer Application (0975-8887)* **2014**, *IX* (105), 17–24.

[10] Huang, Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with

Categorical Values. *Data Mining and Knowledge Discovery* **1998**, *2* (3), 283–304.

[11] Siregar, P. A. Analisis Karakteristik Dan Frekuensi Konsumsi Buah Dan Sayur Pada Penderita Diabetes Dan Non Diabetes. *Darussalam Nutrition Journal, Mei* **2021**, *2021* (1), 61–69.

[12] Kulzer, B. Physical and Psychological Long-Term Consequences of Diabetes Mellitus. *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* **2022**, *66* (4), 503.

[13] Rudi, A.; Kwureh, H. N. Faktor Risiko Yang Mempengaruhi Kadar Gula Darah Puasa pada Pengguna Layanan Laboratorium. *Wawasan Kesehatan* **2017**, *3* (2), 2087–4995.

[14] Wu, X.; Liu, X.; Liao, W.; Kang, N.; Dong, X.; Abdulai, T.; Zhai, Z.; Wang, C.; Wang, X.; Li, Y. Prevalence and Characteristics of Alcohol Consumption and Risk of Type 2 Diabetes Mellitus in Rural China. *BMC Public Health* **2021**, *21* (1). https://doi.org/10.1186/s12889-021-11681-0.

[15] Yang, X.; Chen, J.; Pan, A.; Wu, J. H. Y.; Zhao, F.; Xie, Y.; Wang, Y.; Ye, Y.; Pan, X.-F.; Yang, C.-X. Association between Higher Blood Pressure and Risk of Diabetes Mellitus in Middle-Aged and Elderly Chinese Adults. *Diabetes & Metabolism Journal* **2020**, *44* (3), 436.

[16] Damayanti, S. *Diabetes Melitus Dan Penatalaksanaan Keperawatan*; Nuha Medika; Yogyakarta, 2015.