Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram
https://e-journal.undikma.ac.id/index.php/prismasains/index
e-mail: prismasains.pkpsm@gmail.com

October 2023. Vol. 11, No. 4
p-ISSN: 2338-4530
e-ISSN: 2540-7899
pp. 1002-1017

Application of PCA and K-Means Clustering Methods to Identify Diabetes Mellitus Patient Groups Based on Risk Factors

***Anisa Simanjuntak, Muhammad Siddik Hasibuan**
Computer Science Department, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia

*Corresponding Author e-mail: anisasimanjuntak062001@gmail.com

Received: August 2023; Revised: September 2023; Published: October 2023

Abstract

Diabetes mellitus is a chronic disease characterized by high levels of glucose (sugar) in the blood that is high for a long period of time. Identification is the process of recognizing and determining the characteristics of a particular object or entity. hypertension (high blood pressure), smoking and lack of physical activity can affect the condition of diabetes mellitus patients. Therefore, an approach is needed that can identify groups of diabetic patients based on their risk factors, so that appropriate management and treatment can be carried out. The purpose of this study is to apply PCA method by reducing data dimension to identify the linear combination of the most contributing risk factors in diabetes mellitus patient data and apply K-Means Clustering to cluster into groups based on similar risk factors. The methods to be used are Principal Component Analysis (PCA) and K-Means Clustering. type of quantitative research, this research can be categorized as analytic research, variables are risk factors for diabetes mellitus disease. The results of research using the PCA (principal component analysis) method obtained 9 main components (PC) 86.9275%. correlation between attributes and principal components, then a matrix component is formed with a loading value that the greater the value, the stronger the correlation with the principal component formed with a cut off point of loading value> 0.4 regardless of positive and negative. By using the K-Means Clustering method, The clustering results obtained are divided into 3 groups of diabetes patients based on existing risk factors. Centroid C1 represents a group of diabetes mellitus patients whose condition is at a mild level, while Centroid C2 represents a group of diabetes mellitus patients who are at a moderate level, and Centroid C3 represents a group of patients with severe or dangerous diabetes mellitus.

Keywords: Diabetes Mellitus, Principal Component Analysis, K-Means Clustering

How to Cite: Simanjuntak, A., & Hasibuan, M. (2023). Application of PCA and K-Means Clustering Methods to Identify Diabetes Mellitus Patient Groups Based on Risk Factors. *Prisma Sains : Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram, 11*(4), 1002-1017. doi:https://doi.org/10.33394/j-ps.v11i4.9263

https://doi.org/10.33394/j-ps.v11i4.9263

INTRODUCTION

Diabetes mellitus is a chronic disease characterized by high blood glucose (sugar) levels over a long period of time. There are three types of diabetes mellitus, namely Type I Diabetes Mellitus, Type II Diabetes Mellitus, and Gestational Diabetes Mellitus (Nuraisyah, 2018). According to data from the Ministry of Health, in 2018 there were around 10.3 million people living with diabetes in Indonesia (Kemenkes RI, 2018). According to data from the International Diabetes Federation, in 2021 there will be around 537 million adults suffering from diabetes worldwide. The prevalence of diabetes in Indonesia also continues to increase from year to year, and in 2021 it is estimated that there will be around 10.7 million people suffering from diabetes in Indonesia (IDF, 2021). According to the World Health Organization or WHO, the number of people with diabetes continues to increase, there have been 42 million or 71% in the world. Diabetes is the fourth leading cause of death in the world with 1.6 million people each year, followed by cancer (9.0 million), cardiovascular (17.9 million), and

respiratory (3.9 million) diseases. These four disease groups account for more than 80% of premature deaths (WHO, 2018).

Identification is the process of recognizing and determining the characteristics of a particular object or entity. In a healthcare context, identification is often used to identify individuals who are at risk of developing a certain disease or medical condition based on certain risk factors. Effective and efficient diabetes management requires identification based on several risk factors found in the patient's condition. Various risk factors such as age, gender, family history, obesity (overweight), hypertension (high blood pressure), smoking and lack of physical activity can affect the condition of diabetes mellitus patients. Therefore, an approach is needed that can identify groups of diabetes patients based on their risk factors, so that appropriate management and treatment can be carried out (Prasatya et al., 2020).

In today's digital era, information and communication technology can be utilized to help group diabetes patients based on their risk factors. The methods to be used are Principal Component Analysis (PCA) and K-Means Clustering. PCA is a multivariate analysis technique used to extract information from high-dimensional and complex data sets. The goal is to reduce the dimension of the data by selecting a number of factors or components that are most important, and can explain most of the data variation. Meanwhile, K-Means Clustering is an unsupervised learning method used to group objects or data in a set into several different groups or clusters, based on the similarity between the objects or data. The goal is to find patterns or structures in data where there is no previous label or class (Ilu et al., 2022).

The PCA (Principal Component Analysis) method is used to identify key risk factors or biomakers, analyze the principal components, and classify diabetes mellitus patients into groups based on specific risk factors. PCA involves standardization techniques, covariance or correlation calculations, matrix decomposition, and dimensionality reduction in variables in the dataset or research attributes. In this study, Sarulla Health Center was used as a research site or case study. public health center Sarulla is one of the health centers in North Tapanuli Regency serving health checks, referrals, health letters etc. This public health centerserves various public health center programs such as health checks (check ups), making health certificates, outpatient care, removing stitches, changing dressings, sewing wounds, pulling teeth, checking tension, pregnant tests, examining children, testing blood type, uric acid, cholesterol, blood sugar (diabetes) and others. After I conducted research research, there were 297 people with diabetes mellitus in 2020 - 2023. Based on the problem identification, this research focuses on the application of PCA and K-means clustering methods used to identify diabetes patients based on risk factors.

The K-Means Clustering method is used to group diabetic patients based on the main risk factors found in the PCA analysis. K-Means Clustering uses several stages including initial centroid initialization, grouping diabetic patient data into the closest cluster to the initial centroid calculated using Euclidean distance, centroid update, iteration, and cluster evaluation (Jamal et al., 2018).

The following previous research is related and relevant to the research entitled "Application of K-Means and C4.5 Methods for Predicting Diabetics" which concludes that data mining can provide solutions to provide alternative HbA1c examination service decisions for diabetics who will control or consult a doctor and can also be used to determine patterns or HbA1c prediction rules which require the use of training data that is large, varied, and there are unique elements in the sample data (Prasatya et al., 2020).

Followed by a further study entitled "Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction" which concluded that the number of features for breast cancer classification from the original WBC dataset can be reduced by feature extraction, i.e. transforming the original data by principal component decomposition (eigenvectors) and also by K-means clustering technique. The matrix measurement results show the dimensionality reduction by K-means clustering is almost as good as PCA and at least two clusters have fewer features (Jamal et al., 2018).

The study entitled "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques" which concluded that in designing an efficient model for diabetes prediction using the PCA method for dimensionality reduction, k-means for clustering, and logistic regression for classification. logistic regression models performed at a better level in predicting diabetes, compared to the results obtained when other algorithms were used in our study and other studies (Zhu et al., 2019).

The novelty of this research with previous research is that previous research ((Prasatya et al., 2020) and (Zhu et al., 2019)) is the same as predicting diabetics. The first previous research used K-means and C4. 5 method and the 2rd previous research used the pca- and K-means methods by improving the logistic regression model. In the second study (Jamal et al., 2018) using PCA and K-means Clustering methods to reduce dimensions and classify breast cancer while my research uses PCA and K-means Clustering Methods to identify groups or grouping (Clustering) of diabetes mellitus patients based on existing risk factors.

The results of this analysis are expected to provide insight into the characteristics of patients with diabetes mellitus and assist physicians at sarulla health center in developing better treatment plans based on relevant risk factors. By categorizing patients into groups based on similar risk variable values, doctors can develop effective and efficient treatment strategies for each group of diabetes mellitus patients.

Based on the method approach used in this research, the objectives of this research can be explained as follows: Applying PCA and K-Means Clustering Methods to identify diabetes mellitus patient groups based on risk factors, Applying PCA methods by reducing data dimensions to identify linear combinations of the most contributing risk factors in diabetes mellitus patient data and applying K-Means Clustering to group into groups based on similar risk factors, The approach in the study is expected to be able to determine the risk factors that most affect the identified diabetes patient groups.

METHOD

This research uses quantitative methods, which is research that uses numerical data to be analyzed and the results can be measured using clustering evaluation metrics such as SSE (Sum of Squared Errors), Silhouette Coefficient, and others (Nasution & Hasibuan, 2020). In this case, PCA and K-Means Clustering methods were used to generate different groups of patients based on their risk factors, so that the results could be quantitatively measured and interpreted. This study was conducted by collecting risk factor variables for diabetes mellitus patient groups, collecting data on respondent samples, namely data on visiting diabetes mellitus patients collected through anthropometric measurements.

Type diabetes mellitus is the most common disease and has many sufferers in Sarulla Health Center. So that Diabetes Mellitus is the most likely to be identified. Identification of diabetes mellitus patient groups is carried out based on two risk factors found, namely: Risk factors that cannot be changed and risk factors that can be changed. The following Research Variables from the above risk factors, among others.

Table 1. Data Table of PCA and K-Means Clustering Research Needs

Data Table of PCA and K-Means Clustering Research Needs						
NO	**Age (Years)**	**Gender**	**Family History**	**Physical Activity**	**HbA1c**	**Unhealthy Diet**
1	Kel-n age	L/P	Yes/No	Fair/Low	GDH kel-n	Yes/No
2	Age of kel-n	L/P	Yes/No	Moderate/Low	GDH kel-n	Yes/No
3	Age of nth grade	L/P	Yes/No	Moderate/Low	GDH kel-n	Yes/No
4	Age n	L/P	Yes/No	Moderate/Low	GDH kel-n	Yes/No
NO	**Systolic Blood Pressure**	**Diastolic Blood Pressure**	**Smoking Status**	**Blood Sugar Level (mg/dL)**	**Diet**	**Obesity Status**
1	TDS kel-n	TDD kel-n	Yes/No	GD kel-n	Healthy/Not	(kg/m2)
2	TDS kel-n	TDD kel-n	Yes/No	GD kel-n	Healthy/Not	BW/LP kel-n
3	TDS kel-n	TDD kel-n	Yes/No	GD kel-n	Healthy/No	BW/LP kel-n
4	TDS kel-n	TDD kel-n	Yes/No	GD kel-n	Healthy/Not	BW/TB/LP kel-n

Table 1 is a description of the data required in the study in the approach with PCA and K-Means Clustering to identify risk factors contributing to diabetes mellitus and form groups of patients at risk of diabetes based on relevant research attributes.

PCA reduction was performed to reduce the dimensionality of the data by projecting the data from a higher attribute space to a lower attribute space. To determine the risk factors that most influence the occurrence of diabetes mellitus or have high accuracy in identifying patient groups, PCA analysis can be performed and observe the contribution of each attribute in explaining data variability. Accuracy is a measure of the extent to which the prediction or classification performed by a model or method can match correctly with the actual value or actual category. In this case, risk factors that have high accuracy will provide prediction results that are closer to the actual situation, namely correctly identifying patients who are at high risk of developing diabetes mellitus. In the context of PCA and K-means clustering analysis, after applying PCA and selecting the most significant risk factors, the next step is to combine the results with clustering methods to form patient groups. With high accuracy, the groups will effectively separate patients at high risk of diabetes mellitus from the rest of the group.

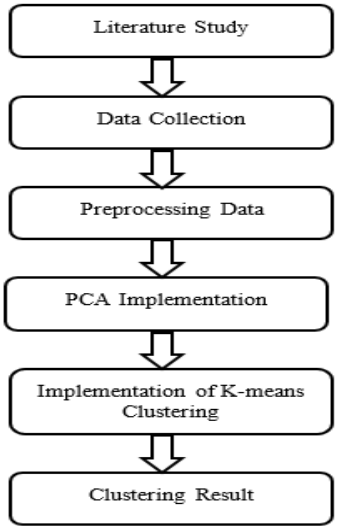

Figure 1. Research conceptual framework

RESULTS AND DISCUSSION

Data Analysis

At the data analysis stage that will be carried out, namely: selecting attributes that will be used in data mining. Of course, not all attributes are included in the dataset used in the data mining process because only those that act as identification references are selected. The data used in this study are medical records of diabetes mellitus patients from 2021, 2022, and 2023. The data was collected from the Sarulla Health Center and consisted of 297 diabetes mellitus patients. Then the data is selected to determine which risk factors for diabetes mellitus are most suitable for use as research attributes. The attributes used in this study can be seen in Table 2.

Table 2. Data Analysis Results

Patient Name	Age (Years)	Gender	Family History	Physical Activity	HbA1c	Unhealthy Diet	Hipertensi		Smoking Status	Blood Sugar Levels	Diet	Obesity Status (BMI)
							Systolic Blood Pressure	Diastolic Blood Pressure				
Rolina Pasaribu	35	P	Available	Enough	6,2	No	170	60	No	280	No	20,6
Simon Sianturi	48	L	No	Enough	6,5	No	160	90	Yes	330	No	26
Winda Pasaribu	50	P	No	Less	5,8	No	150	80	No	310	No	22,9
Rosalinda Sitompul	60	P	No	Less	7,8	No	140	90	No	240	No	27,3
Natanael Simanungkalit	61	L	Available	Enough	6,4	No	140	80	No	366	No	21,4
Julindah Purba	56	P	Available	Less	7,5	Yes	150	70	No	226	No	28,5

Patient Name	Age (Years)	Gender	Family History	Physical Activity	HbA1c	Unhealthy Diet	Hipertensi		Smoking Status	Blood Sugar Levels	Diet	Obesity Status (BMI)
							Systolic Blood Pressure	Diastolic Blood Pressure				
Wardi Siahaan	66	L	No	Enough	6,8	No	160	80	Yes	254	No	25,6
Febri Yanti Gultom	62	P	Available	Less	7,0	Yes	160	70	No	282	No	22,2
Frengki Julio Nababan	53	L	No	Enough	7,5	No	140	70	Yes	328	No	23,1
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
Aspu Sianturi	58	L	Available	Enough	6,0	No	165	80	No	220	No	26,6

Preprocessing Data

Before the data mining process can be carried out, a data preprocessing or data cleaning process needs to be carried out. The purpose of this process is to ensure the quality of the data selected at the data selection stage. Before the data preprocessing stage is carried out, a scan of the dataset will be carried out using the RapidMiner tool.

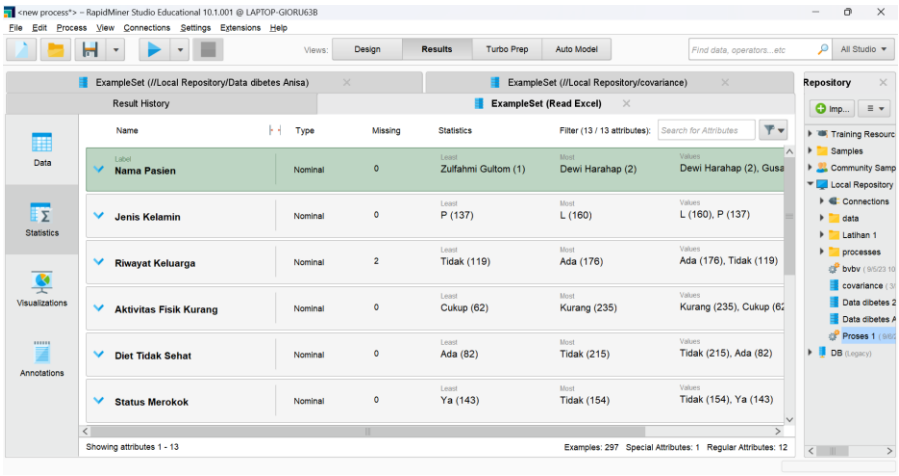

Figure 2. Dataset Scan Result

Figure 2 shows that there are no missing values or noisy (inconsistent data in the diabetes mellitus patient medical record dataset. For the existence of data duplication, it cannot be known through the results of data scanning, therefore at the data preprocessing stage, data duplication cleaning is carried out to remove the same research data. The process and results of data preprocessing can be seen in Figure 3. and Table 3. in the data transformation stage.

Data Transformation

At this stage, the data will be transformed so that it is suitable for use in the data mining process. Such as the attributes of gender, family history, physical activity, unhealthy diet, smoking status, and patterns that are still nominal data types. So it must be converted into numeric data types so that the PCA and K-means Clustering algorithms can be run.

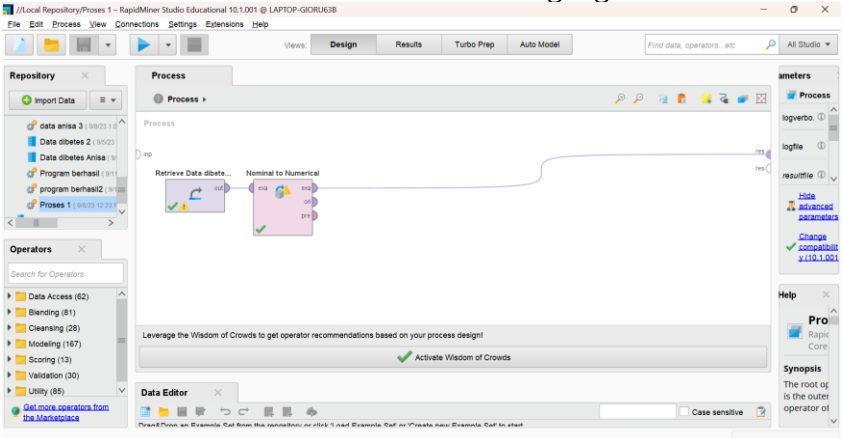

Figure 3. Data Transformation Process

Based on Figure 3, the data transformation stage is carried out using the nominal to numerical operator to convert attributes with nominal data types into numeric. Then, Table 3 shows the results of the data transformation that has been done before. In table 3 it can be seen that gender consisting of male (L) = 0, and female (P) = 1. Family History consisting of the word there = 1, no = 0 and other nominal type variables.

Table 3. Data Transformation Results

Patient Name	Gender	Family Care	Physical Activity	Healthy Diet	Smoking Status	Diet	Age	Systolic Blood Pressure	Diastolic Blood Pressure	HbA1c	Blood Sugar Levels	Obesity Status
Rolina Pasaribu	1	1	0	0	0	1	35	170	60	6,2	280	20,6
Simon M Sianturi	0	0	0	0	1	1	48	160	90	6,5	330	26
Winda Pasaribu	1	0	1	0	0	1	50	150	80	5,8	310	22,9
Rosalinda Sitompul	1	0	1	0	0	1	60	140	90	7,8	240	27,3
Natanael Simanungkalit	0	1	0	0	0	1	61	140	80	6,4	366	21,4
Julindah Purba	1	1	1	1	0	1	56	150	70	7,5	226	28,5
Wardi Siahaan	0	0	0	0	1	1	66	160	80	6,8	254	25,6
Febri nti Gultom	1	1	1	1	0	1	62	160	70	7,0	282	22,2
Frengki Julio Nababan	0	0	0	0	1	1	53	140	70	7,5	328	23,1
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
Rusmini Sitanggang	1	1	1	0	0	1	72	134	90	7,0	328	23,1

PCA Implementation (*Principal Component Analysis)*

The implementation of the PCA algorithm in this study uses Rapidminer tools. The operators used are the read excel operator for inputting the preprocessing dataset, the PCA operator for dimensionality reduction.

1. Table of mean and standard deviation of each variable in the data.

Table 4. *Mean* and *Standart deviation*

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
Gender	297	0	297	0,000	1,000	0,451	0,498
Family History	297	0	297	0,000	1,000	0,596	0,492
Physical Activity	297	0	297	0,000	1,000	0,791	0,407
Healthy Diet	297	0	297	0,000	1,000	0,273	0,446
Smoking Status	297	0	297	0,000	1,000	0,488	0,501
Diet	297	0	297	0,000	1,000	0,946	0,226
Age	297	0	297	25,000	94,000	56,987	9,949
Systolic Blood Pressure	297	0	297	120,000	206,000	160,276	15,729
Diastolic Blood Pressure	297	0	297	50,000	90,000	75,465	9,154
HbA1c	297	0	297	4,700	9,000	6,542	0,794
Blood Sugar Level	297	0	297	200,000	585,000	310,071	73,498
Obesity Status	297	0	297	0,918	34,200	26,237	3,185

2. Standardize data using Rapidminer tools by using the Normalize operator

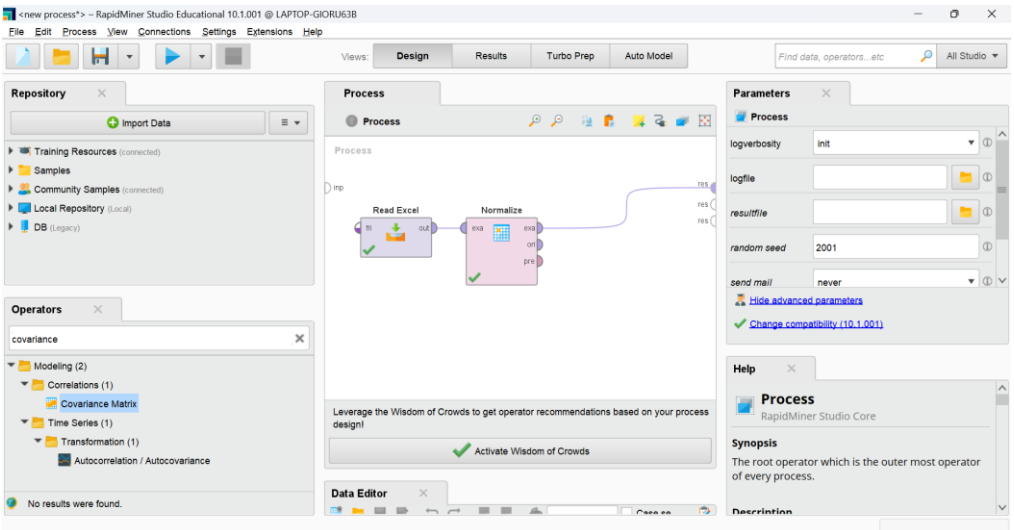

Figure 4. Data Standardization Process

Based on Figure 4.3, the data standardization stage is carried out using the normalize operator. Standardization is done by subtracting the average of each variable and dividing it by

the standard deviation of each variable in the data so that the standardization results are obtained in Figure 4 below.

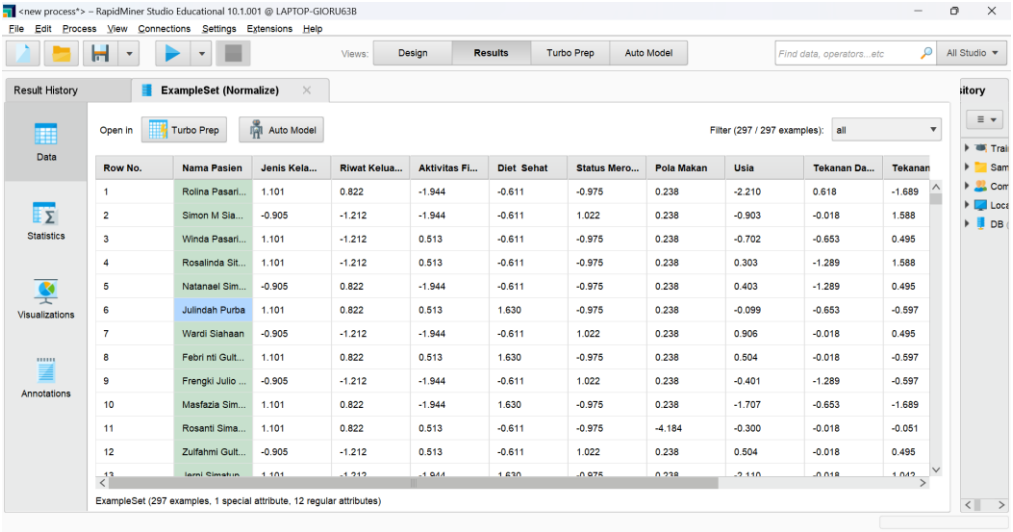

Figure 5. Data Standardization Results

3. After the data is standardized, we calculate the covariance matrix (a matrix containing the covariance between pairs of variables) of the standardized dataset using the Covariance Matrix operator.

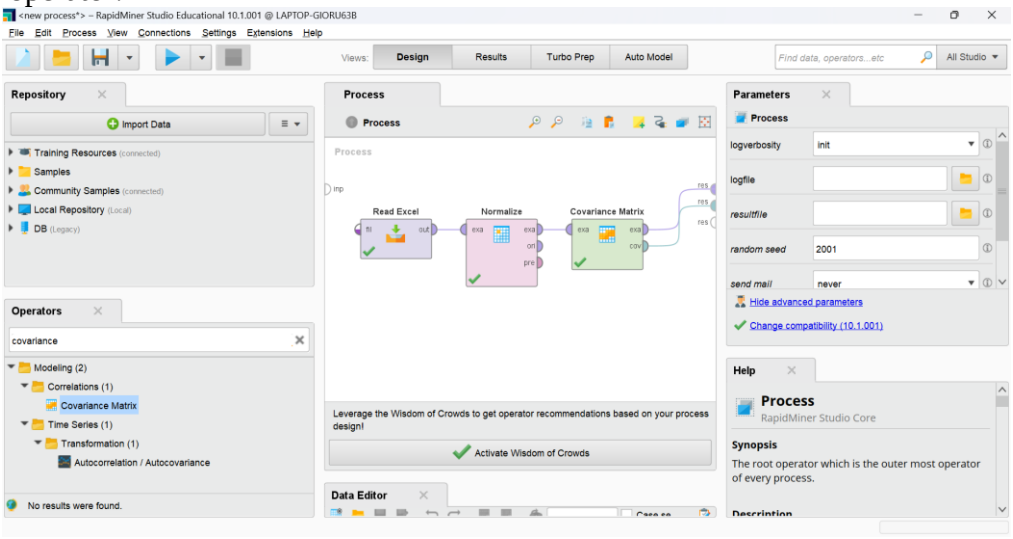

Figure 6: Process of Calculating the Covariance Matrix

4. Table of Correlation Matrix and Covariance Matrix of the standardized dataset

Table 5. Correlation and Covariance Matrix

Patient Name	Gender	Family Care	Physical Activity	Healthy Diet	Smoking Status	Diet	Age	Systolic Blood Pressure	Diastolic Blood Pressure	HbA1c	Blood Sugar Levels	Obesity Status
Gender	**1**	**-0,012**	**-0,117**	**0,113**	**-0,777**	**-0,053**	**-0,071**	**-0,021**	**0,055**	**-0,008**	**-0,127**	**0,079**
Family History	**-0,012**	**1**	**0,033**	**-0,004**	**-0,033**	**0,016**	**-0,027**	0,000	**-0,025**	**0,028**	**0,039**	**-0,029**
Physical Activity	**-0,117**	**0,033**	**1**	**-0,076**	**0,170**	**0,061**	**0,054**	**0,100**	**0,099**	**-0,014**	**0,054**	**0,144**
Healthy Diet	**0,113**	**-0,004**	**-0,076**	**1**	**-0,144**	**0,012**	**-0,059**	**0,133**	**0,038**	**-0,083**	**-0,080**	**0,076**
Smoking Status	**-0,777**	**-0,033**	**0,170**	**-0,144**	**1**	**0,084**	**0,106**	**0,045**	**0,008**	**0,016**	**0,155**	**-0,020**
Diet	**-0,053**	**0,016**	**0,061**	**0,012**	**0,084**	**1**	**0,012**	**-0,047**	**-0,035**	**0,050**	**0,123**	**0,133**
Age	**-0,071**	**-0,027**	**0,054**	**-0,059**	**0,106**	**0,012**	**1**	**-0,005**	**0,005**	**0,032**	**0,074**	**-0,005**
Systolic Blood Pressure	**-0,021**	0,000	**0,100**	**0,133**	**0,045**	**-0,047**	**-0,005**	**1**	**-0,108**	**-0,095**	**-0,049**	**0,099**
Diastolic Blood Pressure	**0,055**	**-0,025**	**0,099**	**0,038**	**0,008**	**-0,035**	**0,005**	**-0,108**	**1**	**0,030**	**-0,044**	**0,124**
HbA1c	**-0,008**	**0,028**	**-0,014**	**-0,083**	**0,016**	**0,050**	**0,032**	**-0,095**	**0,030**	**1**	-0,002	**0,178**
Blood Sugar Level	**-0,127**	**0,039**	**0,054**	**-0,080**	**0,155**	**0,123**	**0,074**	**-0,049**	**-0,044**	-0,002	**1**	**-0,098**
Obesity Status	**0,079**	**-0,029**	**0,144**	**0,076**	**-0,020**	**0,133**	**-0,005**	**0,099**	**0,124**	**0,178**	**-0,098**	**1**

5. Table of eigenvalues of the covariance matrix. Eigenvalue represents how much variation can be explained by each eigenvector. In determining the number of principal components (PC) there are several approaches, one of which is to look at the cumulative proportion of variance that can be explained by the principal components. Table 5 below indicates that 9

main components (PC) should be chosen, because the nine main components have been able to capture 86.9275% of the data diversity. This can also be seen through the scree plot in Figure 6.

Table 6. Eigenvalue Result

Value	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Eigenvalue	**1,973**	**1,350**	**1,243**	**1,102**	**1,020**	**1,010**	**0,994**	**0,918**	**0,822**	0,702	0,653	0,213
Variability (%)	16,439	11,251	10,356	9,186	8,497	8,421	8,280	7,652	6,847	5,852	5,442	1,779
Cumulative %	16,439	27,690	38,045	47,231	55,728	64,149	72,428	80,080	**86,927**	92,779	98,221	100,000

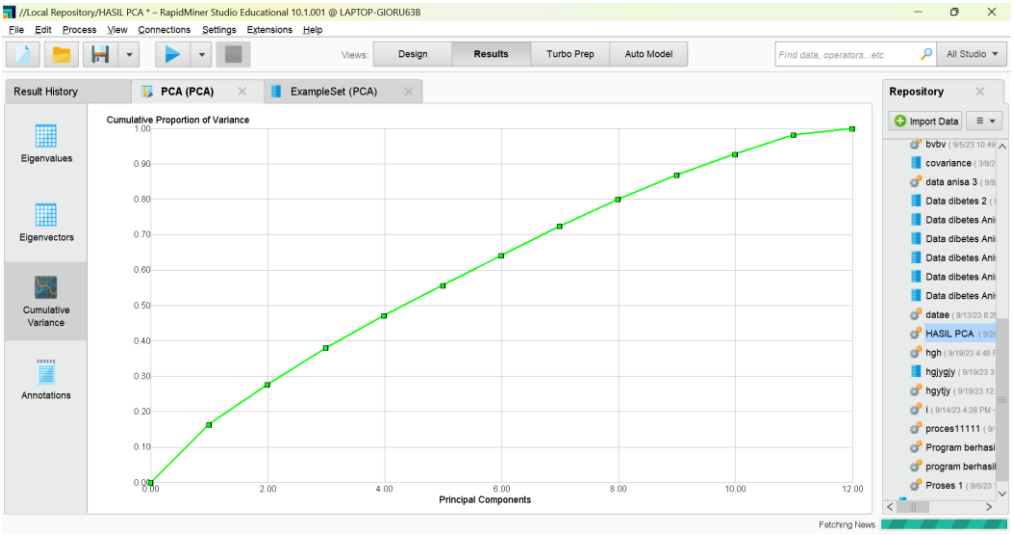

Figure 7. *Scree Plot*

6. Table of eigenvectors of the covariance matrix. The eigenvectors are sorted based on the highest eigenvalue which will be the main component of the PCA.

Table 7. *Eigenvectors values*

Variable	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Gender	-0,620	0,067	-0,129	0,111	0,193	-0,155	0,116	-0,139	0,047	-0,023	-0,059	0,694
Family History	0,005	-0,028	-0,068	0,398	0,399	0,673	0,138	0,390	-0,178	-0,119	0,074	0,040
Physical Activity	0,224	0,395	0,156	-0,018	0,518	0,034	0,122	-0,378	-0,034	0,573	-0,107	-0,026
Healthy Diet	-0,220	0,118	0,425	0,142	-0,102	-0,121	-0,369	0,581	0,135	0,468	0,020	0,036
Smoking Status	0,638	0,016	0,118	-0,132	-0,149	0,082	-0,109	0,096	-0,019	-0,040	0,045	0,712
Diet	0,129	0,273	-0,163	0,583	-0,125	-0,246	-0,349	-0,102	-0,438	-0,133	-0,351	-0,021
Age	0,162	0,040	-0,122	-0,069	0,209	-0,562	0,531	0,492	-0,262	-0,017	0,017	-0,025
Systolic Blood Pressure	0,012	0,128	0,670	0,167	-0,023	-0,010	0,336	-0,055	0,255	-0,414	-0,391	0,031
Diastolic Blood Pressure	-0,038	0,360	-0,154	-0,489	0,383	0,015	-0,409	0,233	0,122	-0,363	-0,295	-0,053
HbA1c	0,030	0,342	-0,443	0,031	-0,420	0,220	0,312	0,166	0,369	0,235	-0,379	0,004
Blood Sugar Level	0,248	-0,138	-0,217	0,417	0,302	-0,276	-0,139	-0,007	0,685	-0,088	0,183	-0,029
Obesity Status	-0,054	0,683	0,050	0,061	-0,169	0,002	0,068	-0,039	0,017	-0,226	0,660	-0,038

7. To determine the attributes included in the 9 principal components, the rotation is described in Table 4.7. In table 4.7, the correlation between attributes and principal components is explained, then a matrix component is formed with a loading value that the greater the value, the stronger the correlation to the principal component formed with a cut off point of loading value> 0.4 without looking at positive and negative.

Table 8. *Loading Factor values*

Variabel	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
Gender	-0,321	0,077	-0,144	0,116	0,195	-0,155	**-0,871**	-0,133	0,043	-0,019	-0,048	0,321
Family History	0,007	-0,033	-0,076	0,418	0,403	**0,676**	0,138	0,374	-0,161	-0,100	0,059	0,018
Physical Activity	0,314	0,459	0,174	**0,523**	0,019	0,034	0,121	-0,362	-0,031	0,481	-0,087	-0,012
Healthy Diet	-0,309	0,137	0,474	0,149	-0,103	-0,121	-0,368	**0,557**	0,122	0,392	0,016	0,017
Smoking Status	**0,896**	0,019	0,132	-0,139	-0,151	0,082	-0,109	0,092	-0,018	-0,034	0,036	0,329
Diet	0,181	0,317	-0,182	**0,612**	-0,126	-0,247	-0,348	-0,098	-0,397	-0,112	-0,284	-0,010
Age	0,228	0,046	-0,136	-0,072	0,211	**-0,565**	0,529	0,471	-0,237	-0,014	0,014	-0,012
Systolic Blood Pressure	0,017	0,148	**0,747**	0,175	-0,023	-0,010	0,335	-0,052	0,231	-0,347	-0,316	-0,014
Diastolic Blood Pressure	-0,054	0,418	**-0,572**	-0,514	0,387	0,015	-0,408	0,224	0,111	-0,304	-0,238	-0,025
HbA1c	0,042	0,397	-0,424	0,032	**-0,494**	0,221	0,311	0,159	0,334	0,197	-0,306	0,002
Blood Sugar Level	0,349	-0,160	-0,242	0,438	0,305	-0,277	-0,139	-0,007	**0,621**	-0,074	0,148	-0,013
Obesity Status	-0,076	**0,794**	0,056	0,064	-0,171	0,002	0,067	-0,038	0,015	-0,190	0,534	-0,018

8. Loading values in bold indicate the loading value> 0.4 which means there is a correlation between the attributes with the principal components formed, then the position of each component is obtained to form 9 principal components as in table 4.8 below.

Table 9. Principal Component Result

No	Variable	*Principal Component* (PC)	*Loading Value*	*Variance Cumulatif %*	*Eigenvalue*
1	Smoking Status	PC 1	0,896	16 %	1,973
2	Obesity Status	PC 2	0,794	28 %	1,350
3	Systolic Blood Pressure and Diastolic Blood Pressure	PC 3	0,747 -0,572	38 %	1,243
4	Phisical activety and diet	PC 4	0,523 0,612	47 %	1,102
5	HbA1c	PC 5	-0,494	56 %	1,020
6	Family History and Age	PC 6	0,676 -0,565	64 %	1,010
7	Gender	PC 7	-0,871	72 %	0,999
8	Unhealthy Diet	PC 8	0,557	80 %	0,918
9	Blood sugar level	PC 9	0,621	86 %	0,822

9. Then data reduction is carried out using the PCA operator so as to produce the principal component value found in Figure 8 below.

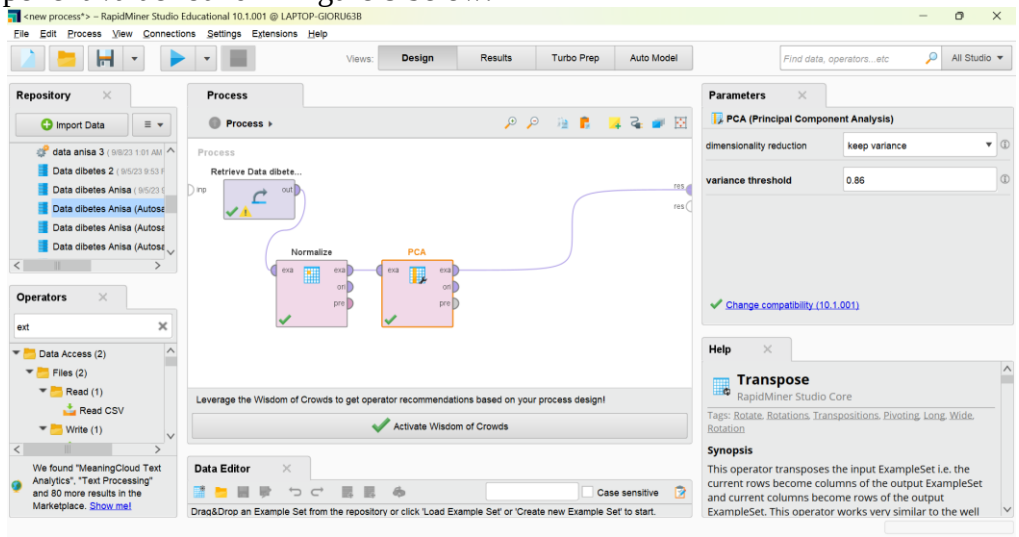

Figure 8. PCA Reduction Process

Based on Figure 4.6, the data reduction process is carried out using the PCA operator so as to produce the principal component value contained in Table 10 below.

Table 10. PCA Reduction Result Value

pc_1	pc_2	pc_3	pc_4	pc_5	pc_6	pc_7	pc_8	pc_9
1,87615	-2,65751	-0,22171	1,45482	1,05466	1,47761	0,21793	-1,08488	-0,13569
-0,79967	-0,35668	0,46650	-1,24204	1,24811	-0,19862	1,67184	0,02529	0,62242
1,14477	-0,68081	0,45899	-0,58579	-0,80057	-0,83124	0,88647	-1,60148	-0,20304
1,26636	1,60724	2,01766	-1,53099	0,13496	-0,55491	0,00128	-0,44618	-0,19479
0,02939	-2,04823	1,80703	0,17745	-0,18996	0,10375	0,66473	1,11637	-0,23529
1,80102	1,24699	0,16404	0,73212	0,45123	0,70080	-0,28083	0,83144	-0,51755
-0,89644	-0,49244	0,46895	-1,25899	1,73831	-0,86352	0,01168	0,73506	-0,55543
1,41842	-0,32017	-0,20267	0,97444	-0,48891	0,00302	-0,37809	1,06106	-0,25622
-1,02919	-1,47321	1,64045	-0,44747	2,33358	-0,21792	0,60291	0,07781	0,33008
.....	.....	.....	.....	.....	.....	.....	.....	.....
1,33951	0,85362	0,23974	-1,67130	-0,70194	-1,14364	-0,10704	-0,58164	-1,12577

10. Visualization results of PCA (Principal Component Analysis) reduction on each variable in the data.

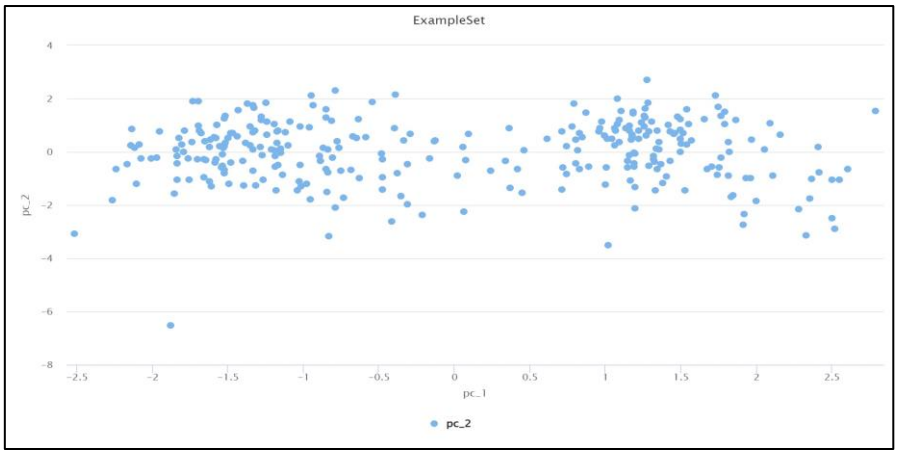

Figure 9. Visualization of PCA Reduction

Implementation of K-Means Clustering

In the implementation stage of the k-means algorithm, the data processed is the medical record data of patients with diabetes mellitus from the reduction of Principal Component Analysis in 2021, 2022, and 2023 which are listed in table 11 below.

Tabel 11. Nilai *Principal Component* (PC)

Patient Name	pc_1 (Smoking Status)	pc_2 (Obesity Status)	pc_3 (TDS and TDD)	pc_4 (Physical activity and Diet)	pc_5 (HbA1c)	pc_6 (Family History and Age)	pc_7 (Gender)	pc_8 (Unhealthy Diet)	pc_9 (Blood Sugar Level)
Rolina Pasaribu	1,87615	-2,65751	-0,22171	1,45482	1,05466	1,47761	0,21793	-1,08488	-0,13569
Simon M Sianturi	-0,79967	-0,35668	0,46650	-1,24204	1,24811	-0,19862	1,67184	0,02529	0,62242
Winda Pasaribu	1,14477	-0,68081	0,45899	-0,58579	-0,80057	-0,83124	0,88647	-1,60148	-0,20304
Rosalinda Sitompul	1,26636	1,60724	2,01766	-1,53099	0,13496	-0,55491	0,00128	-0,44618	-0,19479
Natanael Simanungkalit	0,02939	-2,04823	1,80703	0,17745	-0,18996	0,10375	0,66473	1,11637	-0,23529
Julindah Purba	1,80102	1,24699	0,16404	0,73212	0,45123	0,70080	-0,28083	0,83144	-0,51755
Wardi Siahaan	-0,89644	-0,49244	0,46895	-1,25899	1,73831	-0,86352	0,01168	0,73506	-0,55543
Febri Yanti Gultom	1,41842	-0,32017	-0,20267	0,97444	-0,48891	0,00302	-0,37809	1,06106	-0,25622
Frengki Julio Nababan	-1,02919	-1,47321	1,64045	-0,44747	2,33358	-0,21792	0,60291	0,07781	0,33008
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
Rusmini Sitanggang	0,72637	0,13800	2,43856	-0,47868	-1,94445	0,41444	-0,22192	0,83774	-0,54277

1. Determining the value of the number of clusters (k), the number of clusters is 3, namely cluster (C1) light, cluster (C2) medium, and cluster (C3) heavy.
2. Initialize the centroid randomly or randomly found in table 12 which is selected for the cluster of diabetes mellitus patients in order 1, 5, 7.

Table 12. Initial Centroid

Cluster	Patient Name	Smoking Status	Obesity Status	TDS and TDT	Physical Activity and Diet	HbA1c	Family History and Age	Gender	Unhealthy Diet	Blood Sugar Levels
C1	Wardi Siahaan	-0,89644	-0,49244	0,46895	-1,25899	1,73831	-0,86352	0,01168	0,73506	-0,55543
C2	Natanael Simanungkalit	0,02939	-2,04823	-2,04823	0,17745	-0,18996	0,10375	0,66473	1,11637	-0,23529
C3	Rolina Pasaribu	1,87615	-2,65751	-0,22171	1,45482	1,05466	1,47761	0,21793	-1,08488	-0,13569

3. Calculate the distance of each data point to the centroid, using the euclidean distance theory. After the calculation is done, the results are obtained like this:

Table 13. Calculation of Euclidean distance iteration 1 of each data

N0	Patient Name	Distance to Cluster			Result
		C1	C2	C3	
1	Rolina Pasaribu	5,409985398	4,24509866	**0**	C3
2	Simon M Sianturi	**2,314672834**	3,529943094	5,193150734	C1
3	Winda Pasaribu	4,177454503	**3,768441476**	4,308222464	C2
4	Rosalinda Sitompul	**3,968662969**	4,6128542	6,158341457	C1
5	Natanael Simanungkalit	3,529822515	**0**	4,24509866	C2
6	Julindah Purba	**4,30806255**	4,338980729	4,57781213	C1
7	Wardi Siahaan	**0**	3,171371981	5,449385402	C1
8	Febri Yanti Gultom	4,109335702	**3,283309788**	3,92882994	C2
9	Frengki Julio Nababan	**2,31242304**	3,121431165	4,812215405	C1
.....	.....	.....	.....	.....	.....
297	Rusmini Sitanggang	4,620485983	**3,224869833**	4,14653783	C2

4. Recalculate the position of the centroids:
   Calculate the mean or average of the data points in each cluster to get the new centroids position.

- The new C1 centroid:

1. Pc_1 (Smoking Status) = (-0,79967 + 1,26636 +1,80102 +…..+ 0,38009) /159 = -0,488890377
2. Pc_2 (Obesity Status) = (-0,35668 + 1,60724 + 1,24699 +……+ -0,56352) / 159 = 0,385462579
3. Pc_3 (TDS and TDD) = (0,46650 + 2,01766 + 0,16404 + ……+ 2,21836) / 159 = -0,199052201
4. Pc_4 (Physical Activity and Diet) = (-1,24204 + -1,53099 + 0,73212 + ……+ -1,01376) / 159 = -0,337979874
5. Pc_5 (HbA1c) = (1,24811 + 0,13496 + 0,45123 +……+ 1,90128) / 159 = 0,302815409
6. Pc_6 (Family History) = (-0,19862 + -0,55491 + 0,70080 + ……+ -0,18488) / 159 = -0,128177296
7. Pc_7 (Gender) = (1,67184 + 0,00128 + -0,28083 + ……+ -0,10823) / 159 = -0,129666855
8. Pc_8 (Unhealthy Diet) = (0,02529 + -0,44618 + 0,83144 +……+ -0,78520) / 159 = -0,154876792
9. Pc_9 (Blood Sugar Level) = (0,62242 + -0,19479 + 0,51755 +……+ 0,18136) / 159 = 0,076806792

- New C2 centroid:

1. Pc_1 (Smoking Status) = (1,14477 + 0,02939 + 1,41842 + …….+ 2,58897) / 95 = 0,324511579
2. Pc_2 (Obesity Status) = (-0,68081 + -2,04823 + -0,32017 +…….+ -0,51195) / 95 = -0,510261789
3. Pc_3 (TDS and TDD) = (0,45899 + 1,80703 + -0,20267 +…….+ 0,55968) / 95 = 0,602987
4. Pc_4 (Physical Activity and Diet) = (-0,58579 + 0,17745 + 0,97444 +……+ 0,41151) / 95 = 0,256916316
5. Pc_5 (HbA1c) = (-0,80057 + -0,18996 + -0,48891 +…….+ 0,78582) / 95 = -0,582886737
6. Pc_6 (Family History) = (-0,83124 + 0,10375 + 0,00302 +…….+ -0,03282) / 95 = 0,038364947
7. Pc_7 (Gender) = (0,88647 + 0,66473 + -0,37809 +……..+ -0,64617) / 95 = -0,032484211
8. Pc_8 (Unhealthy Diet) = (-1,60148 + 1,11637 + 1,06106 +…….+ 1,26667) / 95 = 0,065942842
9. Pc_9 (Blood Sugar Levels) = (-0,20304 + -0,23529 + -0,25622 +……..+ 2,42748) / 95 = 0,336346947

- New C3 centroid:

1. Pc_1 (Smoking Status) = (1,87615 + 2,47266 + 2,37703 +…….+ 1,17590) / 43 = 1,164993023
2. Pc_2 (Obesity Status) = (-2,65751 + -2,53193 + -1,25882 +……+ -2,14064) / 43 = -0,279714651
3. Pc_3 (TDS and TDD) = (-0,22171 + -0,51398 + -0,95995 + ……..+ 0,40320) / 43 = -0,59863000
4. Pc_4 (Physical Activity and Diet) = (1,45482 + 1,24578 + 0,62956 +…….+ 1,00456) / 43 = 0,870385349
5. Pc_5 (HbA1c) = (1,05466 + 1,24097 + 0,50873 +…….+ 1,41085) / 43 = 0,26686186
6. Pc_6 (Family History) = (1,47761 + 1,06288 + 0,0024 +…….+ -0,63922) / 43 = 0,128362791
7. Pc_7 (Gender) = (0,21793 + 1,19328 + 0,53246 +……..+ 0,97086) / 43 = -0,242687674
8. Pc_8 (Unhealthy Diet) = (-1,08488 + 0,49653 + 1,45172 +……..+ -1,46885) / 43 = -0,085473488

9. Pc_9 (Blood Sugar Levels) = (-0,13569 + -0,82969 + -1,42242 +…….+ -1,11379) /43 = -0,108915116

5. Repeat steps 2 and 3 until the centroid position does not change and because there is no data that moves clusters and the 1st and 2nd clusters have the same result, the process of calculating the new centroid is stopped and ends at iteration 2. We can see in table 4.13 below, that the clustered data is divided into three groups of diabetes patients based on existing risk factors. Centroid C1 represents a group of diabetes mellitus patients whose condition is at a mild level, while Centroid C2 represents a group of diabetes mellitus patients who are at a moderate level, and Centroid C3 represents a group of patients with severe or dangerous diabetes mellitus (Kurniawan et al., 2022).

Table 14. Cluster Result

No	Cluster Result	
	Cluster	**Number of Patients**
1	C1 (Mild)	131
2	C2 (Medium)	150
3	C3 (Heavy)	16

Clustering Results With Rapidminer

1. In the results stage, the k-means algorithm implementation uses Rapidminer tools. The operators used are the read excel operator to input the preprocessed dataset, the PCA operator for dimension reduction, the K-means Clustering operator to run the k-means algorithm.

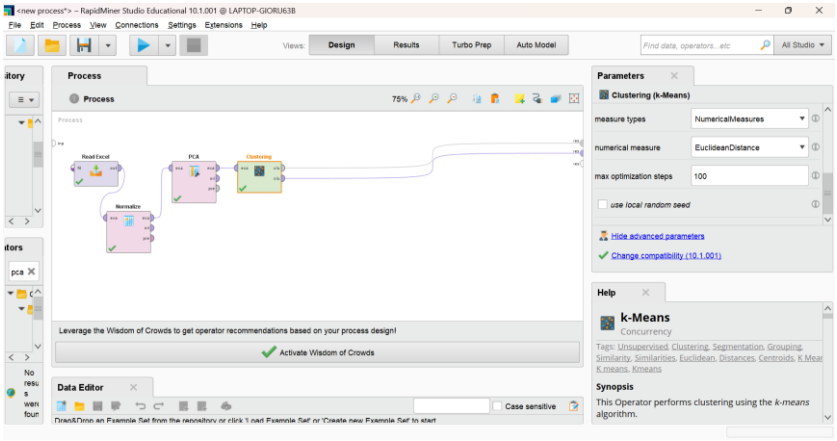

Figure 10. K-means Clustering Process

2. The Clustering Model Result display of the processed data is shown in Figure 10 and Figure 11 below.

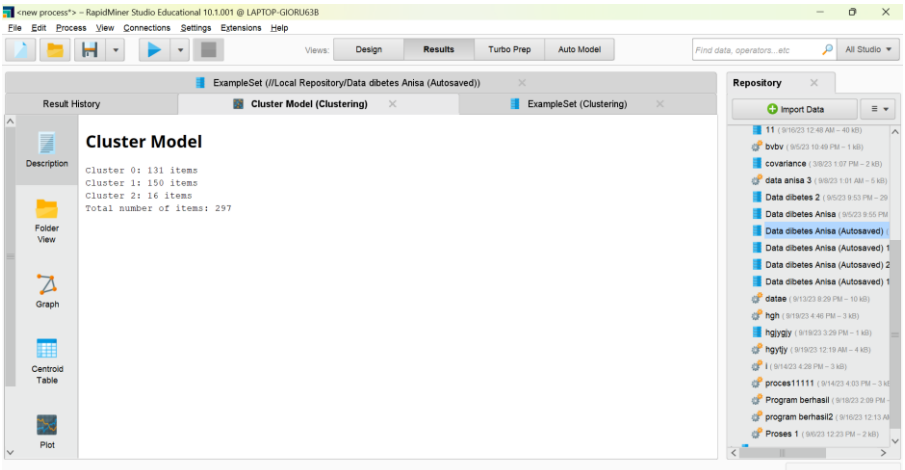

Figure 11. *Cluster Model*

3. The display of Clustering 1 results that have been processed is in Figure 12 below.

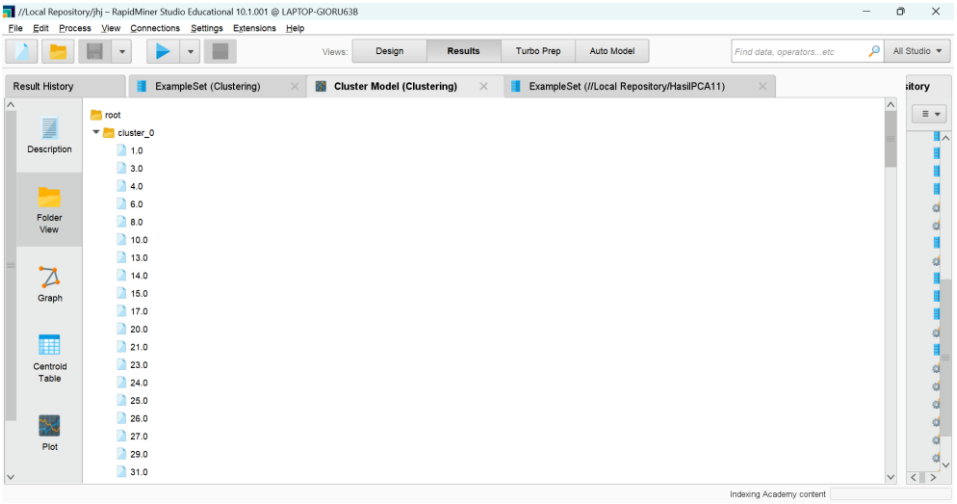

Figure 12. Clustering Result 1

4. The display of Clustering 2 results that have been processed is in Figure 13 below.

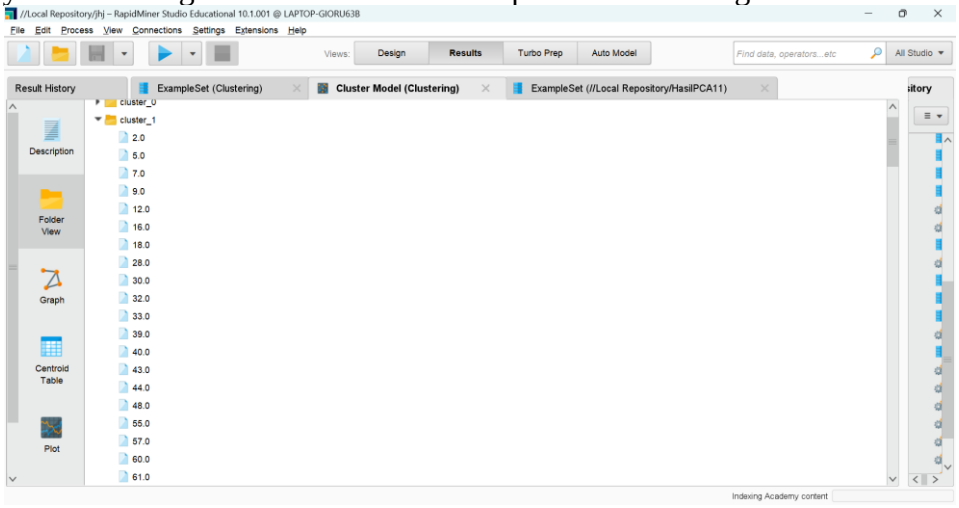

Figure 13. Clustering Result 2

5. The display of the processed Clustering 3 results is shown in Figure 14 below.

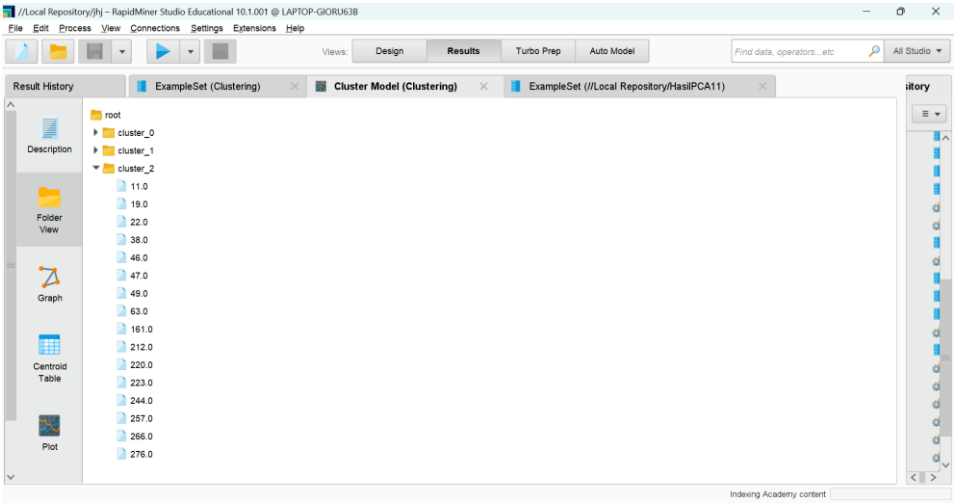

Figure 14. Clustering Result 3

6. Clustering visualization results for each variable in the data are in figure 15 below.

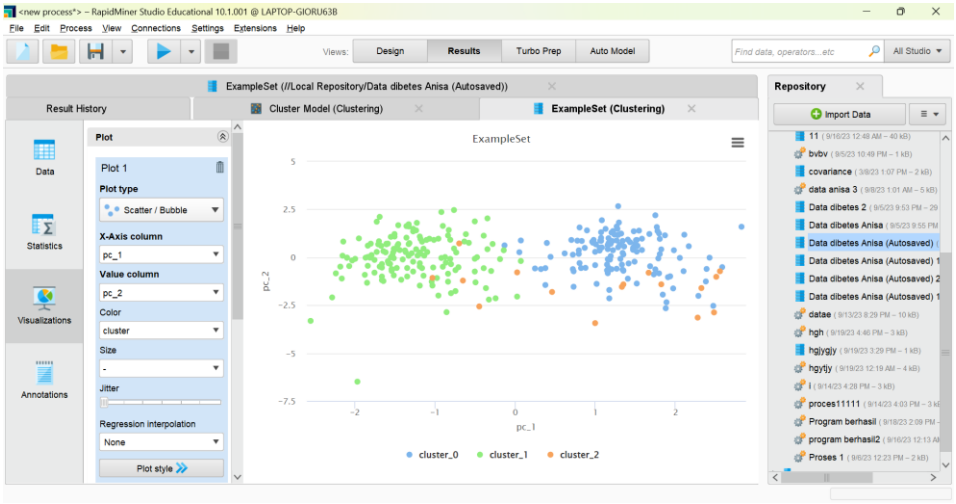

Figure 15. Visualization of Clustering Results

Discussion

Cluster 1 (Mild) consists of 131 diabetes mellitus patients whose male gender is 98 and female is 33 and aged 35-60 years. Most have no family history of diabetes, most smoke, low obesity levels, average high blood pressure (Hypertension) between 130/75 -160/90. HbA1c ranged from 5.1 - 6.5. With blood sugar levels between 207-330.

Cluster 2 (Moderate) consists of 150 patients with diabetes mellitus, 58 of whom are male and 92 of whom are female and aged 50 - 72 years. Most have a family history of diabetes, many have a smoking status, a moderate level of obesity, an average high blood pressure (hypertension) between 134/90 - 180/65. HbA1c ranged from 5.8 - 7.0. With blood sugar levels between 230 - 490.

Cluster 3 (Severe) consisted of 16 diabetes mellitus patients, 6 males and 10 females, aged 52 - 94 years. Most did not smoke, high obesity levels, high average blood pressure between 7.0-8.5. With blood sugar levels between 280 - 585.

Clustering of diabetes mellitus patients was done on the grounds that there is a large variation in the characteristics of diabetes mellitus patients, and these groups can be used to direct the treatment and management of the disease according to its severity. Cluster 1 tends to have milder diabetes, while Cluster 3 has more severe diabetes. In addition, there are several possible contributing risk factors, such as family history, smoking, obesity, and high blood pressure, which can affect the severity of the disease.

This discussion is in line with the results of research by Yulianti et al., 2022 which concluded that the final grouping based on PCA and k-means clustering shows that the use of male family planning in Indonesia is not good. This is because in cluster 2, most of the married men use family planning poorly and the percentage in almost all provinces reaches 90%, so that the male family planning promotion strategy is given to the two clusters. Likewise, research by No et al., 2023 concluded that clustering of diabetic patient data at the Mojokerto Health Center using the K-Means algorithm method to understand the patterns and characteristics of diabetic patients. And also strengthened by the research of Hediyati & Suartana, 2021 which concluded that the designed system has successfully used PCA to reduce the features of the dataset obtained from the agricultural production data of Bojonegoro Regency in 2017-2020 which originally consisted of 12 columns and 430 rows, into datasets with 1 PC, 2 PC, and 3 PC. The system is able to perform clustering using the K-Means algorithm with a different number of clusters on two different types of datasets, namely the original dataset and the dataset that has undergone the dimension reduction process.

CONCLUSION

Based on the results of research on the application of the PCA method and k-means clustering to identify groups of diabetes mellitus patients based on risk factors at the Sarulla Health Center, the authors draw the following conclusions: By using the PCA (principal component analysis) method in helping to reduce 12 attributes or research variables to facilitate the clustering process. So that the results obtained are 9 main components (PC). The nine principal components have been able to capture 86.9275% of the data diversity. To determine the attributes included in the 9 principal components, Table 8 explains the correlation between attributes and principal components, then a matrix component is formed with a loading value that the greater the value, the stronger the correlation to the principal component formed with a cut off point of loading value> 0.4 without looking at positive and negative. Using the K-Means Clustering method is used in clustering patients with diabetes mellitus as a result of PCA reduction in 2021, 2022, and 2023. The clustering results obtained are divided into 3 groups of diabetes patients based on existing risk factors. Centroid C1 represents a group of diabetes mellitus patients whose condition is at a mild level, while Centroid C2 represents a group of diabetes mellitus patients who are at a moderate level, and Centroid C3 represents a group of patients with severe or dangerous diabetes mellitus.

RECOMMENDATION

For further research of this kind, it is expected to conduct research with different methods or algorithms, which are certainly in accordance with case studies that occur in the field so that they can compare the results with a combination of PCA and K-means Clustering methods. And the program can be developed again to get better results, accurate and in accordance with the development of technological science.

ACKNOWLEDGMENT

This research is the result of Anisa Simanjuntak's final project research, Computer Science Study Program, Faculty of Science and Technology, State Islamic University of North Sumatra (UINSU). The researcher would like to thank those who have contributed to this research.

REFERENCES

Abdillah, A. A., & Prianto, B. (2019). Pembelajaran Mesin Menggunakan Principal Component Analysis dan Support Vector Machines untuk Mendeteksi Diabetes. *Jurnal Matematika Dan Sains*, *24*(1), 10–14. https://doi.org/10.5614/jms.2019.24.1.2

Agustanti, D., & Purbianto, P. (2022). Pengaruh Konsumsi Air Alkali Terhadap Kadar Glukosa Darah Pada Pasien Diabetes Mellitus. *Jurnal Ilmiah Keperawatan Sai Betik*, *16*(2), 149. https://doi.org/10.26630/jkep.v16i2.3099

Azizah, U. N., Wurjanto, M. A., Kusariana, N., & Susanto, H. S. (2022). Hubungan Kualitas Tidur dengan Kontrol Glikemik pada Penderita Diabetes Melitus : Systematic Review. *Jurnal Epidemiologi Kesehatan Komunitas*, *7*(1), 411–422. https://doi.org/10.14710/jekk.v7i1.13310

Bastian, A. (2018). Penerapan Algoritma K-Means Clustering Analysis Pada Penyakit Menular Manusia (Studi Kasus Kabupaten Majalengka). *Jurnal Sistem Informasi*, *14*(1), 28–34. https://doi.org/10.21609/jsi.v14i1.566

Hayqal, H. H. Q., Oni Soesanto, & Yuana Sukmawaty. (2022). K-Means Clustering dan Principal Component Analysis (PCA) Dalam Radial Basis Function Neural Network (RBFNN) Untuk Klasifikasi Data Multivariat. *Journal of Mathematics Theory and Application*, *4*(1), 1–7. https://doi.org/10.31605/jomta.v4i1.1757

Hediyati, D., & Suartana, I. M. (2021). Penerapan Principal Component Analysis (PCA) Untuk

Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro. *Journal of Information Engineering and Educational Technology*, *5*(2). https://doi.org/10.26740/jieet.v5n2.p49-54

IDF. (2021). *International Diabetes Federation*. Diabetes Research and Clinical Practice. https://doi.org/10.1016/j.diabres.2013.10.013

Ilu, S. Y., Rajesh, P., & Mohammed, H. (2022). Prediction of COVID-19 using long short-term memory by integrating principal component analysis and clustering techniques. *Informatics in Medicine Unlocked*, *31*(June), 100990. https://doi.org/10.1016/j.imu.2022.100990

Jamal, A., Handayani, A., Septiandri, A. A., Ripmiatin, E., & Effendi, Y. (2018). Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction. *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, *9*(3), 192. https://doi.org/10.24843/lkjiti.2018.v09.i03.p08

Kemenkes RI. (2018). *Penyakit Diabetes Melitus*. https://p2ptm.kemkes.go.id/informasi-p2ptm/penyakit-diabetes-melitus

Kesuma Dinata, R., & Hasdyna, N. (2020). *Machine Learning.pdf* (M. S. DR. Fajriana, S.Si. (ed.); Pertama). Unimal Press.

Kurniawan, R. A., Hasibuan, M. S., Piramida, P., & Ramadhan, R. S. (2022). Penerapan Algoritma K-Means Untuk Clustering Tempat Makan Di Batubara. Journal of Computer Science and Informatics Engineering (CoSIE), 01(1), 10–18. https://doi.org/10.55537/cosie.v1i1.27

Nasution, M. Z., & Hasibuan, M. S. (2020). Pendekatan Initial Centroid Search Untuk Meningkatkan Efisiensi Iterasi Klustering K-Means. *Techno.Com*, *19*(4), 341–352. https://doi.org/10.33633/tc.v19i4.3875.

No, V., Hal, J., Elang, A., Setyadji, S., Wibowo, A. P., Ngurah, I. G., Matthew, A., Pratama, R. B., Masyhuda, T. A., Sinaga, A. A., Purwanti, E., & Werdiningsih, I. (2023). *Analisis Klaster Data Pasien Diabetes untuk Identifikasi Pola dan Karakteristik Pasien*. *5*(3), 172–182.

Nuraisyah, F. (2018). Faktor Risiko Diabetes Mellitus Tipe 2. *Jurnal Kebidanan Dan Keperawatan Aisyiyah*, *13*(2), 120–127. https://doi.org/10.31101/jkk.395

Prasatya, A., Siregar, R. R. A., & Arianto, R. (2020). Penerapan Metode K-Means Dan C4.5 Untuk Prediksi Penderita Diabetes. *Petir*, *13*(1), 86–100. https://doi.org/10.33322/petir.v13i1.925

Purbolaksono, M. D., Irvan Tantowi, M., Imam Hidayat, A., & Adiwijaya, A. (2021). Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(2), 393–399. https://doi.org/10.29207/resti.v5i2.3008

Riskesdas. (2018). Laporan Provinsi Sumatera Utara Riskesdas 2018. In *Badan Penelitian dan Pengembangan Kesehatan*.

Simeftiany Indrilemta Lomo, Endang Darmawan, & Sugiyarto. (2023). Cluster analysis of type II Diabetes Mellitus Patients with the Fuzzy C-means method. *Annals of Mathematical Modeling*, *3*(1), 24–31. https://doi.org/10.33292/amm.v3i1.28

WHO. (2018). *Noncommunicable diseases*. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

Yulianti, T. R., Siregar, K. N., Prabawa, A., & Fadhilah, N. (2022). Identifikasi Atribut dengan Principal Component Analysis dan K-Means Clustering Sebagai Dasar Penyusunan Strategi Promosi KB Pria di Indonesia. *Jurnal Biostatistik, Kependudukan, Dan Informatika Kesehatan*, *2*(2), 79. https://doi.org/10.51181/bikfokes.v2i2.5868

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*(January), 100179. https://doi.org/10.1016/j.imu.2019.100179