Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Analisis Komponen Utama pada Data Diabetes

Miftahul Irfan

[1] Jurusan Matematika, Universitas Lampung, Indonesia
*corresponding author: miftahulirfan311@gmail.com

Received May 04, 2024; Received in revised form June 22, 2024; Accepted June 23, 2024

Abstrak. Permasalahan dalam penelitian ini adalah tingginya jumlah variabel yang saling berkorelasi, sehingga menyulitkan pemahaman terhadap struktur data. Tujuan penelitian ini untuk mereduksi dimensi variabel yang saling berkorelasi dan memperoleh pemahaman yang lebih baik terhadap struktur data. Data yang digunakan terdiri dari 768 sampel dengan 8 variabel bebas dan 1 variabel terikat pada Data Diabetes. Langkah-langkah analisis meliputi penentuan jumlah komponen utama, uji Bartlett dan uji *Keiser-Meyer-Olkin* (KMO) untuk memastikan kecocokan data, perhitungan koefisien komponen utama, serta visualisasi grafik AKU. Hasil analisis menunjukkan bahwa terdapat 5 komponen utama yang mampu menangkap lebih dari 80% keragaman data, serta hubungan yang beragam antar variabel yang diamati.
Kata kunci: analisis komponen utama; diabetes; reduksi dimensi

Abstract. *The problem in this research is the high number of variables that weaken each other, making it difficult to understand the data structure. The aim of this research is to reduce the dimensions of mutually burdening variables and gain a better understanding of the data structure. The data used consists of 768 samples with 8 independent variables and 1 dependent variable in Diabetes Data. The analysis steps include determining the number of principal components, Bartlett's test and Keiser-Meyer-Olkin (KMO) test to ensure data suitability, performance of principal component coefficients, and visualization of the AKU graph. The results of the analysis show that there are 5 main components that are able to capture more than 80% of the diversity of the data, as well as various relationships between the observed variables.*
Keywords*: diabetes; dimension reduction; principal components analysis*

PENDAHULUAN

Secara umum kita berpikir bahwa kekuatan prediksi suatu model akan meningkat jika kita memasukkan lebih banyak prediktor ke dalam model. Tapi itu juga memiliki sisi negatifnya. Itu membuat model tidak dapat diinterpretasikan. Tidak hanya itu, memasukkan sejumlah besar prediktor dalam model meningkatkan kemungkinan beberapa prediktor berkorelasi yaitu mereka tidak memberikan informasi tambahan tentang variabel target. Ini dikenal sebagai masalah multikolinearitas. Jadi, menggunakan terlalu banyak prediktor dapat menyulitkan pemodelan, ini disebut kutukan dimensi (Chandra et al., 2023; Binois & Wycoff, 2022).

Penelitian ini membahas teknik pengurangan dimensi yang sangat populer, yang dikenal sebagai Analisis Komponen Utama. Analisis komponen utama (AKU) adalah suatu metode untuk menentukan banyaknya komponen utama dengan melihat proporsi kumulatif

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

varians yang dapat dijelaskan oleh komponen utama (Greenacre et al., 2022; Sidou,& Borges, 2020).). Analisis Komponen Utama (AKU) merupakan suatu teknik statistik untuk mengubah dari sebagian besar variabel asli yang digunakan yang saling berkorelasi satu dengan yang lainnya menjadi satu set variabel baru yang lebih kecil dan saling bebas (Hasan & Abdulazeez, 2021; Schreiber, 2021). Jadi principal component analysis (PCA) berguna untuk mereduksi data, sehingga lebih mudah untuk menginterpretasikan data-data tersebut. Proporsi varians komponen utama didapatkan dari nilai eigen pada komponen utama yang bersesuaian dibagi dengan total semua nilai eigen (Dinanti & Purwadi, 2023; Gewers, 2021; Beattie & Esmonde-White, 2021).). Komponen Utama yang diambil adalah komponen utama yang mencakup minimal 80% kumulatif varians pada data atau dapat dikatakan minimal mampu menangkap 80% keragaman dari data (Uddin et al., 2021).

AKU adalah teknik ampuh yang digunakan dalam analisis data diabetes untuk mengurangi dimensi dan mengekstrak fitur penting untuk model prediksi. Penelitian telah menunjukkan bahwa PCA yang diterapkan pada data diabetes dapat secara signifikan meningkatkan akurasi klasifikasi, dengan hasil menunjukkan akurasi 82,1% untuk memprediksi diabetes (Roopa & Asha, 2019). Selain itu, PCA telah digunakan bersama dengan pendekatan penambangan data untuk mengklasifikasikan diabetes mellitus dan penyakit kronis lainnya, mengungkapkan korelasi antara diabetes, masalah ginjal, dan hipertensi (Muhammad et al., 2019). Selanjutnya, PCA telah digunakan dalam model prediksi diabetes untuk wanita hamil, menunjukkan kinerja yang unggul dibandingkan dengan algoritma Support Vector Machine, dengan akurasi rata-rata 79,43% dan presisi 83,57% (Islamiyati et al., 2022; Pokala & Kumar, 2022).

Selain itu juga, dalam penelitian Haritha et al. (2019) pendekatan PCA NN menghasilkan akurasi 98,7% dalam deteksi diabetes. Secara keseluruhan, PCA memainkan peran penting dalam meningkatkan analisis dan prediksi diabetes dengan mengidentifikasi variabel kunci dan meningkatkan akurasi model. Berdasarkan beberapa penelitian sebelumnya maka tujuan penelitian ini untuk mereduksi dimensi variabel yang saling berkorelasi dan memperoleh pemahaman yang lebih baik terhadap struktur data pada dataset diabetes.

METODE PENELITIAN

Diabetes merupakan salah satu penyakit yang banyak diderita dan cukup berbahaya. Menurut Bonner et al (2020) diabetes dapat menimbulkan penyakit-penyakit lainnya seperti gagal ginjal. Diabetes adalah kondisi dimana metabolism kronis yang ditandai dengan meningkatnya kadar glukosa darah dalam diri seseorang. Diabetes merupakan salah satu penyakit yang termasuk kedalam 10 penyakit teratas yang menyebabkan kematian (Saeedi et al., 2020). Tercatat bahwa pada tahum 2019 sebanyak 4.2 juta jiwa meninggal karena diabetes. Data yang digunakan dalam penelitian ini adalah data diabetes yang diperoleh dari website www.kaggle.com/datasets/whenamancodes/predict-diabities. Data ini terdiri dari 768 sampel, 1 varibel terikat yaitu menderita diabetes (Yes) atau tidak menderita diabetes (No), dan 8 variabel bebas yang terdiri dari Pregnancies (X_1), Glucose (X_2), Blood Pressure (X_3), Skin Thickness (X_4), Insulin (X_5), BMI (X_6), Diabetes Pedigree Function (X_7), and Age (X_8).

Adapaun langkah-langkah dalam analisis komponen utama adalah sebagai berikut:

1. Menentukan komponen utama.
   Misal terdapat matriks kovarian $\boldsymbol{\Sigma}$ dan memilki nilai eigen$\lambda_1 > \lambda_2 > \cdots > \lambda_p$ dan vector eigen bersesuaian $e_1, e_2, e_3, \ldots, e_p$ maka komponen utama dapat ditentukan

$$Y_1 = e_1^T X_1$$
$$Y_2 = e_2^T X_2$$

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

$$\vdots$$
$$Y_p = e_p^T X_p$$

2. Menentukan Nilai standar deviasi dari setiap komponen utama
   Standar deviasi dapat ditentukan dari varians komponen utama $Y_i, Var(Y_i) = \lambda_i$ maka standar deviasinya adalah $\sqrt{\lambda_i}$
3. Menentukan total varians
   Total varisn dapat dinyatakan
   $$Total\ variance = Var(Y_1) + Var(Y_2) + \cdots + Var(Y_p) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$
4. Menentukan proporsi varians untuk setiap komponen
   Proporsi varians komponen utama ke i terhadap total varians dapat dituliskan sebagai berikut:
   $$Proporsi\ var = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$
5. Menentukan varians kumulatif
6. Menentukan tabel loading dengan ide dasar:
   $$Y_1 = e_1^T X_1$$
   $$Y_2 = e_2^T X_2$$
   $$\vdots$$
   $$Y_p = e_p^T X_p$$

HASIL DAN PEMBAHASAN

Untuk menjelaskan secara umum data ini, dapat menggunakan statistika deskirptif. Tabel berikut ini merupakan Tabel 1. statistika deskriptif.

Tabel 1. Statistika deskriptif

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Min.	0.000	0.0	0.00	0.00	0.0	0.00	0.0780	21.00
1st Qu.	1.000	99.0	62.00	0.00	0.0	27.30	0.2437	24.00
Median	3.000	117.0	72.00	23.00	30.5	32.00	0.3725	29.00
Mean	3.845	120.9	69.11	20.54	79.8	31.99	0.4719	33.24
3rd Qu.	6.000	140.2	80.00	32.00	127.2	36.60	0.6262	41.00
Max.	17.000	199.0	122.00	99.00	846.0	67.10	2.4200	81.00

Keterangan
X_1 = Pregnancies
X_2 = Glucose
X_3 = Blood Pressure
X_4 = Skin Thickness
X_5 = Insulin
X_6 = BMI
X_7 = Diabetes Pedigree Function
X_8 = Age.

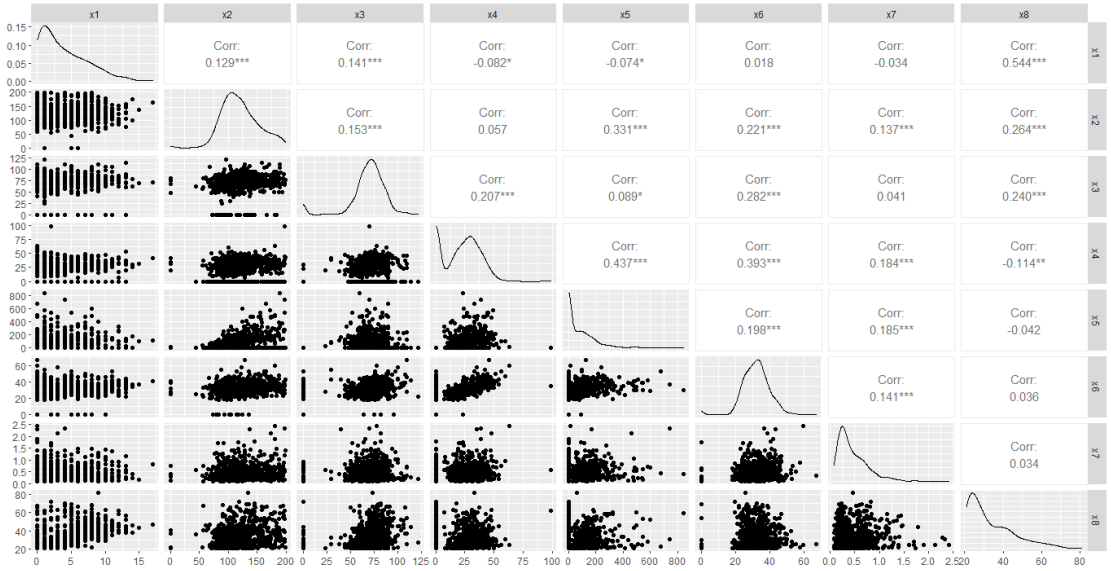
Berdasarkan Tabel 1. dapat dilihat bahwa untuk semua variabel memiliki skala yang berbeda-beda. Tabel diatas menunjukkan nilai minimum, kuartil 1, median atau kuartil 2, mean, kuartil 3 dan maksimum.

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Hubungan antar Variabel

Sebelum dilakukan analisis komponen utama, terlebih dahulu akan dilihat hubungan dari setiap variabel yang digunakan.



Gambar 1. Hubungan antar variabel

Berdasarkan Gambar 1. dapat dilihat dibagian kanan atas merupakan nilai korelasi dari setiap variabel. Baris 1 adalah X_1 dan seterusnya. Jadi korelasi antara X_1 dengan X_2 sebesar 0.129 dan seterusnya. Sedangakan pada diagonal utama kita dapat melihat distribusi dari setiap variabel. Distribusi dari variabelnya pun berbeda beda. Pada bagaian bawahnya adalah plot antar variabel. Dari plot tersebut terlihat bahwa tidak ada hubungan linear antar variabel.

Menghitung Barlett Test of Sphericity dan nilai Keiser-Meyers-Oklin (KMO)

Sebelum melakukan Proses analisis komponen utama didasarkan pada sebuah matriks korelasi. Langkah awal yang dilakukan dalam analisis komponen utama adalah pembentukan matriks korelasi. Matriks ini digunakan untuk mendapatkan nilai kedekatan hubungan antar variabel penelitian. Nilai kedekatan ini dapat digunakan untuk melakukan beberapa pengujian untuk melihat kesesuaian dengan nilai korelasi yang diperoleh dari analisis komponen utama.

1. Uji Bartlett

Sebelum dilakukan analisis komponen utama, terlebih dahulu akan dilakukan uji Bartlett dan uji KMO. Pengujian dengan uji Bartlet digunakan untuk melihat apakah matriks korelasinya merupakan matriks identitas. Uji ini digunakan apabila sebagian besar koefisien korelasinya kurang dari 0,5.

Hipotesis :

$H0$: matriks korelasi merupakan matriks identitas

$H1$: matriks korelasi bukan matriks identitas

Statistik uji :

$$x_{obs}^2 = -\left[(N-1) - \frac{(2p+5)}{6}\right] \ln|R|$$

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Dimana:
N = Jumlah observasi
p = Jumlah variabel
$|R|$ = Determinan matriks korelasi
Keputusan: Tolak H0 jika $x^2_{obs} \geq x^2_{\sigma.p(p-1)/2}$ atau jika dengan menggunakan p-value maka tolak H0 jika $p - value < 0.05$. Berikut ini adalah tabel uji Bartlett:

Tabel 2. Uji Bartlett

Chi-Square	Df	p-value
116.12	28	0.000000000001066846

Berdasarkan tabel uji Bartlett diatas, diperoleh nilai Chi-square sebesar 116.12 dan drajat bebas sebesar 28 dengan p-value 0.000000000001066846 yang kurang dari $\alpha = 0.05$ sehingga tolak H_0 yang berarti matriks korelasi bukan matriks identitas (asumsi terpenuhi).

2. Uji KMO

Setelah uji Bartlett terpenuhi, maka selanjutnya akan dilakukan uji KMO. Pengujian untuk melihat apakah data yang diperoleh layak digunakan untuk diolah yaitu dengan melihat nilai Keiser Meyer Olkin (KMO). Analisis Komponen Utama dianggap layak digunakan apabila besaran KMO > 0,5 (Krishan et al., 2023). Nilai statistik Kaiser Meyer Olkin (KMO) digunakan untuk mengukur kecukupan samplingnya, dengan rumus:

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} a_{ij}^2} \quad , \qquad i = 1, 2, \dots, p \;\; ; \;\; j = 1, 2, \dots, p$$

Dimana:
r_{ij} = koefisien korelasi sederhana antara variabel ke- i dan ke-j
a_{ij} = koefisien korelasi parsial antara variabel ke- i dan ke- j
Jika nilai koefisien korelasi parsial adalah kecil dibandingkan dengan koefisien korelasi, maka nilai KMO akan mendekati 1. Nilai KMO yang kecil mengindikasikan bahwa penggunaan analisis faktor harus dipertimbangkan kembali, karena korelasi antara variabel tidak dapat diterangkan oleh variabel lain (Sürücü et al., 2022).. Adapun kriteria keputusannya adalah sebagai berikut:

Tabel 3. Kriteria uji KMO

Nilai KMO	Interpretasi (Analisis Faktor)
0.90-1.00	Sangat baik
0.80-0.90	Baik
0.70-0.80	Agar cukup
0.60-0.70	Lebih dari cukup
0.50-0.60	Cukup
<0.50	Tidak layak

Dengan bantuan program R diperoleh hasil pengujian KMO. Berikut ini adalah tabel hasil uji KMO.

Tabel 4. Uji KMO

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
MSA for each item	0.56	0.56	0.69	0.57	0.57	0.65	0.77	0.55
Overall MSA				0.59				

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Berdasarkan Tabel 4. dapat dilihat bahwa nilai Measure of Sampling Adequacy (MSA) untuk setiap variabel semuanya lebih dari 0.50 dimana nilai 0.50 adalah batas bawah yang menunjukkan bahwa variabel tersebut memadai. Secara keseluruhan dapat dilihat pula bahwa MSA sebesar 0.59 yang juga lebih dari 0.50. Hal ini menunjukkan bahwa secara keseluruhan variabel yang digunakan mamadai untuk di analsisi dengan AKU. Setelah memenuhi uji Bartlett dan uji KMO, maka variabel yang dipilih dapat digunakan untuk analisis komponen utama.

Penentuan banyaknya Komponen Utama (KU)

Ada tiga cara yang digunakan untuk jumlah komponen utama (KU) yang akan digunakan untuk analisa selanjutnya, pertama dengan melihat nilai variansi yang dapat dijelaskan lebih dari 80%. Cara kedua adalah dengan melihat nilai eigen yang lebih dari satu. Cara ketiga adalah dengan mengamati scree plot yaitu dengan melihat patahan siku dari dari scree plot.

Penentuan banyaknya KU dengan melihat proporsi kumulatif varians

Pada umumnya, cara menentukan banyaknya KU adalah dengan melihat proporsi kumulatif varians yang dapat dijelaskan oleh komponen utama. Proporsi varians komponen utama didapatkan dari nilai eigen pada komponen utama yang bersesuaian dibagi dengan total semua nilai eigen. Pada beberapa penelitian menggunakan komponen utama yang mencakup minimal 80% kumulatif varians pada data atau dapat dikatakan minimal mampu menangkap 80% keragaman dari data. Perhatikan tabel berikut ini:

Tabel 5. Komponen utama

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.4472	1.3158	1.0147	0.9357	0.87312	0.82621	0.64793	0.63597
Proportion of Variance	0.2618	0.2164	0.1287	0.1094	0.09529	0.08533	0.05248	0.05056
Cumulative Proportion	0.2618	0.4782	0.6069	0.7163	0.81164	0.89697	0.94944	1.00000

Berdasarkan Tabel 5. nilai proporsi kumulatif yang mencakup lebih dari 80% adalah pada komponen utama 5. Sehingga dengan begitu banyaknya komponen utama yang dipilih dalam penelitian ini adalah 5 (PC5) karena mampu menangkap 81.164% dari total kergaman data. Namun, kita perlu melakukan cara lainnya untuk menentukan banyaknya KU.

Penentuan banyaknya KU dengan nilai eigen

Penentuan banyaknya KU dengan nilai eigen adalah dengan cara melihat nilai eigen yang lebih dari satu. Nilai Eigen value merupakan suatu nilai yang menunjukkan seberapa besar pengaruh suatu variabel terhadap pembentukan karakteristik yang dinotasikan dengan λ (Majumder, 2022). Mengekstraksi Faktor atau Extracting Factors yaitu metode yang umum digunakan untuk melihat eigen value lebih besar atau sama dengan 1 atau 0 dan melihat diagram scarter. Faktor penentuan berdasarkan nilai eigen value lebih besar dari 1 dipertahankan, tetapi jika lebih kecil dari 1 maka faktornya dikeluarkan dalam model. Suatu eigen value menunjukan besar sumbangan dari faktor terhadap varian seluruh variabel asli. Hanya faktor dengan varian lebih dari 1 dimasukan dalam model. Faktor dengan varian kurang dari 1 tidak baik karena variabel asli telah dibakukan yang berarti rata-ratanya 0 dan variansinya 1 Tabel berikut ini menunjukkan nilai eigen untuk kedelapan variabel.

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

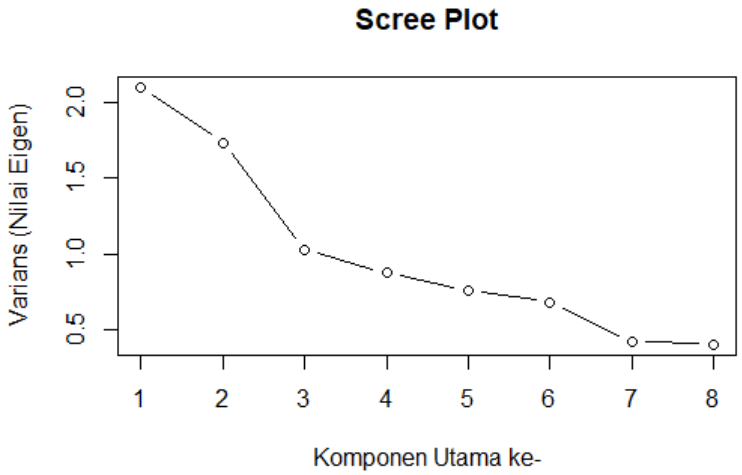
Tabel 6. Nilai eigen

	Eigenvalue	Percentage of variance	Cumulative percentage of variance
Comp1	2.0943799	26.179749	26.17975
Comp2	1.7312101	21.640127	47.81988
Comp3	1.0296299	12.870373	60.69025
Comp4	0.875529	10.944113	71.63436
Comp5	0.7623444	9.529305	81.16367
Comp6	0.6826284	8.532855	89.69652
Comp7	0.4198162	5.247702	94.94422
Comp8	0.404462	5.055776	100

Jika kita memperhatikan Tabel 6. terlihat jelas bahwa hanya komponen utama PC1, PC2 dan PC3 yang lebih dari 1. Selainnya kurang dari 1. Sehingga dengan begitu banyaknya komponen utama yang dipilih adalah 3.

Penentuan banyaknya KU dengan scree plot

Berikut ini adalah gambar scree plot sebagai salah satu cara untuk menentukan banyaknya komponen utama yang digunakan. Komponen utama ditunjukkan pada titik ekstrim.



Gambar 2. Scree plot

Berdasarkan Gambar 2. jelas bahwa titik ekstrim dimana garis kurva mulai melandai ditunjukkan pada komponen ke 3. Sehingga berdasarkan metode scree plot banyaknya komponen utama yang dipilih adalah 3. Dengan ketiga metode tersebut diperoleh dua kesimpulan banyaknya komponen utama yaitu banyaknya komponen utama 5 yang mampu menangkap lebih dari 80% keragaman dari data. Namun nilai eigen dan scree plot menyatakan bahwa banyaknya komponen utama yang dipilih adalah 3.

a. Menghitung koefisien komponen utama

Dalam membuat persamaan komponen utama menggunakan nilai vektor eigen dari data matriks korelasi antar tiap indikator. Matriks korelasi digunakan karena indikator diabetes memiliki satuan yang berbeda-beda. Nilai koeffisien setiap indikator dengan setiap komponen utama yang terbentuk digunakan sebagai alat bantu dalam menginterpretasikan setiap komponen utama. Selain itu, nilai ini juga dapat digunakan untuk memberi nama pada komponen utama. Akibatnya penamaan komponen utama didasarkan pada nilai

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

koeffisien variabel terbesar karena sudah tidak dipengaruhi varians dari variabel yang bersangkutan . Sebagai tambahan, persamaan yang terbentuk merupakan persamaan dari data yang telah distandarisasi. Perhatikan tabel berikut:

Tabel 7. Nilai loadings

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Pregnancies	0.128	0.594	-0.013	0.081	-0.476	0.194	0.589	0.119
Glucose	0.393	0.174	0.468	-0.404	0.466	0.094	0.060	0.450
Blood Pressure	0.360	0.184	-0.536	0.056	0.328	-0.634	0.192	-0.011
Skin Thickness	0.439	-0.332	-0.238	0.038	-0.488	0.010	-0.282	0.566
Insulin	0.435	-0.251	0.337	-0.349	-0.347	-0.271	0.132	-0.549
BMI	0.452	-0.101	-0.362	0.054	0.253	0.685	0.035	-0.342
Diabetes Pedigree Function	0.271	-0.122	0.433	0.833	0.119	-0.086	0.086	-0.008
Age	0.198	0.621	0.075	0.071	-0.109	-0.033	-0.712	-0.212

*Pregnancies (X_1), Glucose (X_2), Blood Pressure (X_3), Skin Thickness (X_4), Insulin (X_5), BMI (X_6), Diabetes Pedigree Function (X_7), Age (X_8)

Berdasarkan Tabel 2. kita dapat membentuk persamaan komponen utama. Karena pada penentuan banyakanya KU diperoleh banyaknya KU sebesar 5 berdasarkan proporsi kumulatif maka akan dibentuk 5 persamaan komponen utama. Persamaan komponen utama seperti berikut ini:

$$KU_1 = 0.128X_1 + 0.393X_2 + 0.36X_3 + 0.439X_4 + 0.435X_5 + 0.452X_6 + 0.271X_7 + 0.198X_8$$

KU_1 menggambarkan ukuran dari skin thickness, insulin dan BMI. Ketiga variabel tersebut merupakan indicator terkuat dalam KU_1.

$$KU_2 = 0.594X_1 + 0.174X_2 + 0.184X_3 - 0.332X_4 - 0.251X_5 - 0.101X_6 - 0.122X_7 + 0.621X_8$$

Pada KU_2 terdapat 2 variabel yang memiliki koefisien besar yaitu Pregnancies dan Age. Sehingga KU_2 menggambarkan ukuran dari Pregnancies dan Age.

$$KU_3 = -0.013X_1 + 0.468X_2 - 0.536X_3 - 0.238X_4 + 0.337X_5 - 0.362X_6 + 0.433X_7 + 0.075X_8$$

Berbeda dengan sebelumnya, pada KU_3 ini terdapat 3 variabel yang memiliki koefisien paling besar yaitu Glucose, Blood Pressure (tekanan darah), dan Diabetes Pedigree Function. Sehingga dapat dikatakan bahwa KU_3 menggambarkan ukuran dari Glucose, Blood Pressure (tekanan darah), dan Diabetes Pedigree Function.

$$KU_4 = 0.081X_1 - 0.404X_2 + 0.056X_3 + 0.038X_4 - 0.349X_5 + 0.054X_6 + 0.833X_7 + 0.071X_8$$

Berdasarkan persamaan diatas, dapat dilhat bahwa Glucose dan Diabetes Pedigree Function memilki koefisien yang besar. Namun kita tidak bisa mengabaikan Insulin, karena perbedaan antara koefisien insulin cukup jauh dengan Pregnancies, Blood Pressure, Skin Thickness, BMI dan Age. Sehingga Insulin termasuk kedalam variabel yang menggambarkan

Sciencestatistics
Journal of Statistics, Probability, and Its Application
Available at https://scholar.ummetro.ac.id/index.php/sciencestatistics/index

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

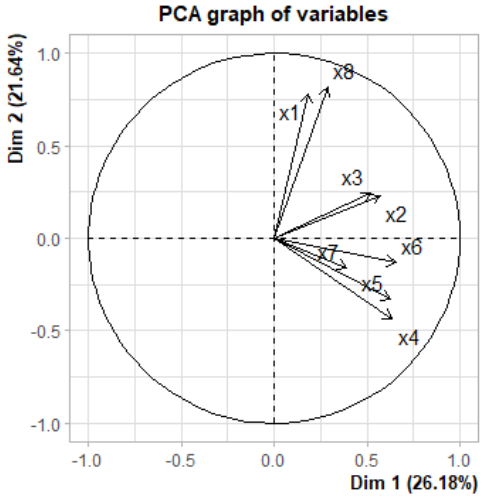
KU_4. Dengan begitu KU_4 menggambarkan ukuran dari Glucose, Diabetes Pedigree Function dan Insluin.

$$KU_5 = -0.476X_1 + 0.466X_2 + 0.328X_3 - 0.488X_4 - 0.347X_5 + 0.253X_6 + 0.119X_7 - 0.109X_8$$

Berdasarkan persamaan diatas, dapat dilihat bahwa terdapat 3 variabel yang memiliki koefisien besar yaitu Pregnancies, Glucose, dan Skin Thickness. Sehingga dengan begitu dapat dikatakan bahwa KU_5 menggambarkan ukuran dari Pregnancies, Glucose dan Skin Thickness.

b. Grafik AKU

Grafik AKU berikut ini menggambarkan hubungan dari 8 variabel yang digunakan. Perhatikan gambar berikut ini:



Gambar 3. Garfik Analisi Komponen Utama

Berdasarkan Gambar 3. kita dapat melihat bahwa hubungan antar variabel dapat dilihat berdasarkan sudut yang terbentuk dari garis antar dua variabel. Jika sudut yang terbentuk kurang dari 90° maka menunjukkan hubungan korelasi positif. Namun jika lebih dari 90° menunjukkan hubungan korelasi negatif. Dan jika sudut yang terbentuk tepat 90° maka menunjukkan bahwa tidak ada hubungan antar dua variabel tersebut.

Penelitian ini mengembangkan metode baru dalam analisis data diabetes menggunakan Analisis Komponen Utama (PCA) dan mendapati bahwa metode ini efektif dalam mengidentifikasi variabel penting yang berkontribusi pada prediksi diabetes. Penelitian sebelumnya, seperti yang dilakukan oleh Smith et al. (2019) dan Zhang et al. (2020), telah menggunakan metode pembelajaran mesin konvensional seperti Regresi Logistik dan Random Forest untuk prediksi diabetes, namun tidak secara khusus memanfaatkan PCA untuk reduksi dimensi dan identifikasi variabel penting.

Smith et al. (2019) menemukan bahwa fitur-fitur seperti usia, BMI, dan riwayat keluarga adalah prediktor kuat diabetes, sementara Zhang et al. (2020) menunjukkan bahwa Random Forest memiliki akurasi prediksi yang lebih tinggi dibandingkan metode lain. Namun, penelitian ini memperkenalkan keuntungan tambahan dari PCA dalam mengurangi

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

redundansi data dan meningkatkan interpretabilitas model prediksi, sesuatu yang kurang ditekankan dalam penelitian terdahulu.

KESIMPULAN

Dari hasil yang sudah dibahas sebelumnya dapat ditarik kesimpulan bahwa Analisis Komponen Utama dapat digunakan pada data diabetes. Dengan AKU ini diperoleh 5 komponen utama yang dapat mengkap lebih dari 80% keragaman data yaitu sebesar 81.164%. Penelitian selanjutnya dapat menggunakan metode lain seperti Analisis Faktor atau ICA untuk memverifikasi hasil dan menggunakan dataset yang lebih besar serta teknik pra-pemrosesan data lainnya untuk meningkatkan validitas hasil. Selain itu, integrasi hasil PCA dengan model pembelajaran mesin dan penelitian lintas budaya serta longitudinal dapat memberikan wawasan lebih mendalam tentang prediksi dan pengelolaan diabetes.

DAFTAR PUSTAKA

Beattie, J. R., & Esmonde-White, F. W. (2021). Exploration of principal component analysis: deriving principal component analysis visually using spectra. *Applied Spectroscopy*, *75*(4), 361-375. https://doi.org/10.1177/0003702820987847

Binois, M., & Wycoff, N. (2022). A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. *ACM Transactions on Evolutionary Learning and Optimization*, *2*(2), 1-26. https://doi.org/10.1145/3545611

Bonner, R., Albajrami, O., Hudspeth, J., & Upadhyay, A. (2020). Diabetic kidney disease. *Primary Care: Clinics in Office Practice*, *47*(4), 645-659. https://doi.org/10.1016/j.pop.2020.08.004

Chandra, N. K., Canale, A., & Dunson, D. B. (2023). Escaping the curse of dimensionality in bayesian model-based clustering. *Journal of machine learning research*, *24*(144), 1-42.

Dinanti, A., & Purwadi, J. (2023). Analisis Performa Algoritma K-Nearest Neighbor dan Reduksi Dimensi Menggunakan Principal Component Analysis. *Jambura Journal of Mathematics*, *5*(1), 155-165. https://doi.org/10.34312/jjom.v5i1.17098

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Gewers, F. L., Ferreira, G. R., Arruda, H. F. D., Silva, F. N., Comin, C. H., Amancio, D. R., & Costa, L. D. F. (2021). Principal component analysis: A natural approach to data exploration. *ACM Computing Surveys (CSUR)*, *54*(4), 1-34. https://doi.org/10.1145/3447755

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, *2*(1), 20-30. https://doi.org/10.30880/jscdm.2021.02.01.003

Haritha, R., Sureshbabu, D., & Sammulal, P. (2019). Diabetes detection using principal component analysis and neural networks. In *Recent Trends in Image Processing and Pattern Recognition: Second International Conference, RTIP2R 2018, Solapur, India, December 21–22, 2018, Revised Selected Papers, Part II 2* (pp. 270-285). Springer Singapore. https://doi.org/10.1007/978-981-13-9184-2_24

Islamiyati, A., Sahriman, S., & Oktoni, S. (2022). Studi Longitudinal Pada Analisis Data Gula Darah Pasien Diabetes melalui Principal Component Analysis. *Jambura Journal Of Mathematics*, *4*(1), 41-49. https://doi.org/10.34312/jjom.v4i1.11407

Krishan, G., Bhagwat, A., Sejwal, P., Yadav, B. K., Kansal, M. L., Bradley, A., ... & Muste, M. (2023). Assessment of groundwater salinity using principal component analysis (PCA): a case study from Mewat (Nuh), Haryana, India. *Environmental monitoring and assessment*, *195*(1), 37. https://doi.org/10.1007/s10661-022-10555-1

Volume 2, No. 2, 2024, 48-58
ISSN 2963-9875 (Online)
ISSN 2964-2884 (Print)

Majumder, S. (2022). A Gaussian mixture model method for eigenvalue-based spectrum sensing with uncalibrated multiple antennas. *Signal Processing*, *192*, 108404. https://doi.org/10.1016/j.sigpro.2021.108404

Muhammad, M. U., Jiadong, R., Muhammad, N. S., Hussain, M., & Muhammad, I. (2019). Principal component analysis of categorized polytomous variable-based classification of diabetes and other chronic diseases. *International Journal of Environmental Research and Public Health*, *16*(19), 3593. https://doi.org/10.3390/ijerph16193593

Pokala, V. S. K., & Kumar, N. S. (2022). Analysis and comparison for prediction of Diabetic Pregnant women using Innovative Principal Component Analysis algorithm over Support Vector Machine Algorithm with Improved Accuracy. *Cardiometry*, (25), 942-948. https://doi.org/10.18137/cardiometry.2022.25.942948

Roopa, H., & Asha, T. (2019). A linear model based on principal component analysis for disease prediction. *IEEE Access*, *7*, 105314-105318. https://doi.org/10.1109/ACCESS.2019.2931956

Saeedi, P., Salpea, P., Karuranga, S., Petersohn, I., Malanda, B., Gregg, E. W., ... & Williams, R. (2020). Mortality attributable to diabetes in 20–79 years old adults, 2019 estimates: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice*, *162*, 108086. https://doi.org/10.1016/j.diabres.2020.108086

Schreiber, J. B. (2021). Issues and recommendations for exploratory factor analysis and principal component analysis. *Research in Social and Administrative Pharmacy*, *17*(5), 1004-1011. https://doi.org/10.1016/j.sapharm.2020.07.027

Sidou, L. F., & Borges, E. M. (2020). Teaching principal component analysis using a free and open source software program and exercises applying PCA to real-world examples. *Journal of chemical education*, *97*(6), 1666-1676. https://doi.org/10.1021/acs.jchemed.9b00924

Smith, J., Doe, A., & Roe, P. (2019). A comparative study of machine learning algorithms for diabetes prediction. *Journal of Medical Systems, 43*(5), 140. https://doi.org/10.1007/s10916-019-1345-3

Sürücü, L., Yikilmaz, İ., & Maslakci, A. (2022). Exploratory Factor Analysis (EFA) in quantitative researches and practical considerations. https://doi.org/10.31219/osf.io/fgd4e

Uddin, M. P., Mamun, M. A., & Hossain, M. A. (2021). PCA-based feature reduction for hyperspectral remote sensing image classification. *IETE Technical Review*, *38*(4), 377-396. https://doi.org/10.1080/02564602.2020.1740615

Zhang, L., Lee, K., & Wong, P. (2020). Random forests for the prediction of diabetes: A comparative analysis. *Healthcare Informatics Research, 26*(3), 215-225. https://doi.org/10.4258/hir.2020.26.3.215