JAMBURA JOURNAL OF MATHEMATICS
Jambura J. Math. Vol. 5, No. 1, pp. 155–165, February 2023
Journal Homepage: http://ejurnal.ung.ac.id/index.php/jjom
DOI: https://doi.org/10.34312/jjom.v5i1.17098

Analisis Performa *Algoritma K-Nearest Neighbor* dan Reduksi Dimensi Menggunakan *Principal Component Analysis*

Aldila Dinanti[1], Joko Purwadi[1,*]

[1]*Program Studi Matematika, Universitas Ahmad Dahlan, Yogyakarta 55191, Indonesia*
[*]*Corresponding author. Email: joko@math.uad.ac.id*

ABSTRAK

Paper ini membahas tentang performa *Algoritma K-Nearest Neighbor* dengan reduksi dimensi menggunakan *Principal Component Analysis (PCA)* pada kasus klasifikasi penyakit diabetes. Banyaknya variabel dan data yang besar pada penyakit diabetes relative memerlukan waktu komputasi yang lama, sehingga dibutuhkan reduksi dimensi untuk mempercepar proses komputasi. Metode reduksi dimensi yang digunakan pada penelitian ini adalah PCA. Setelah dilakukan reduksi dimensi, dilanjutkan dengan klasifikasi menggunakan *Algoritma K-Nearest Neighbor*. Hasil penelitian pada studi kasus penyakit diabetes menunjukkan bahwa reduksi dimensi menggunakan *PCA* menghasilkan 3 komponen utama dari 8 variabel pada data asli yaitu PC1, PC2, dan PC3. Hasil klasifikasi menggunakan *Algoritma K-Nearest Neighbor* menunjukkan bahwa pada pengambilan 3 parameter ketetanggaan (K) terdekatnya yaitu untuk *K*=3, *K*=5 dan *K*=7. Pada kasus *K*=3, diperoleh hasil akurasi sebesar 67,53%, pada pengambilan *K*=5 diperoleh akurasi 72,72 %, dan untuk *K*=7 diperoleh hasil akurasi sebesar 77,92%. Dengan demikian, disimpulkan bahwa performa akurasi terbaik untuk klasifikasi penyakit diabetes dicapai pada *K*=7 dengan akurasi sebesar 77,92%.

Kata Kunci:
Principal Component Analysis; *K-Nearest Neighbor*; Reduksi Dimensi; Penyakit Diabetes

ABSTRACT

This paper discusses the performance of the K-Nearest Neighbor Algorithm with dimension reduction using Principal Component Analysis (PCA) in the case of diabetes disease classification. A large number of variables and data on the diabetes dataset requires a relatively long computation time, so dimensional reduction is needed to speed up the computational process. The dimension reduction method used in this study is PCA. After dimension reduction is done, it is continued with classification using the K-Nearest Neighbor Algorithm. The results on diabetes case studies show that dimension reduction using PCA produces 3 main components of the 8 variables in the original data, namely PC1, PC2, and PC3. Then classification result using K-Nearest Neighbor shows that by choosing 3 closest neighbor parameters (K), for K = 3, K = 5, and K = 7. The result for K = 3 has an accuracy of 67,53%, for K = 5 had an accuracy is 72,72%, and for K=7 had an accuracy of 77,92%. Thus, it was concluded that the best accuracy performance for the classification of diabetes was achieved at K=7 with an accuracy of 77.92%.

e-ISSN: 2656-1344
Received: 22 November 2022 | Accepted: 19 January 2023 | Online: 1 February 2023

Keywords:
Priciple Component Analysis; K-Nearest Neighbor; Dimensional Reduction; Diabetes Desease

Format Sitasi:

A. Dinanti and J. Purwadi, "Analisis Performa *Algoritma K-Nearest Neighbor* dan Reduksi Dimensi Menggunakan *Principal Component Analysis*", *Jambura J. Math.*, vol. 5, No. 1, pp. 155–165, 2023, doi: https://doi.org/10.34312/jjom.v5i1.17098

1. Pendahuluan

Tingginya kematian akibat penyakit diabetes, dan kebanyakan penderita tidak mengetahui bahwa mereka sedang mengidap diabetes, menjadikan diabetes sebagai salah satu penyakit yang berbahaya. Penyebab utama diabetes disebabkan karena kadar glukosa yang meningkat dalam darah dan biasa disebut sebagai penyakit *silent killer* [1]. Adapun beberapa faktor resiko dari penyakit ini adalah usia, kehamilan atau riwayat melahirkan, berat badan, hipertensi, ketebalan kulit, insulin, indeks massa tubuh, riwayat keluarga, dan umur [2].

Besarnya ukuran data dan banyaknya variabel dalam data penyakit dabetes, menjadi permasalahan dalam proses komputasi. Proses komputasi akan membutuhkan waktu yang lebih lama dan hasil menjadi kurang maksimal. Untuk mengatasi hal tersebut dibutuhkan reduksi dimensi terhadap data dengan tanpa menghilangkan informasi-informasi penting yang terdapat dalam dataset [3–5].

Teknik reduksi dimensi yang sering digunakan dalam beberapa penelitian antara lain *Principle Component Analysis* (PCA), *Factor Analysis* (FA) dan *Discriminant Analysis* (DA). Ketiganya merupakan beberapa teknik analisis multivariat, dimana perbedaan keduanya terletak pada apakah kelas respon diperhatikan atau tidak [6]. Teknik FA dan DA tetap memperhatikan variabel target atau respon, sedangan PCA tidak memperhatikan kelas respon [7, 8]. Metode PCA telah digunakan dalam beberapa penelitian untuk reduksi dimensi antara lain penelitian yang dilakukan oleh Magdalena pada tahun 2021 dalam menentukan faktor – faktor yang mempengaruhi lama studi [9]. Dalam penelitian yang lain PCA digunakan untuk mereduksi faktor – faktor yang mempengaruhi dalam hal peramalan dan kualitas produk [10–13]. Pada penelitian ini digunakan metode PCA untuk mereduksi dimensi pada kasus penyakit diabetes, yang diduga variabel – variabelnya memiliki kesamaan karakteristiknya, selain itu kelas – kelas pada variabel responnya tidak diperhatikan sehingga PCA layak utuk digunakan.

Setelah diperoleh komponen utama hasil reduksi dimensi variabel, selanjutnya dilakukan proses klasifikasi. Proses klasifikasi bertujuan untuk menemukan fungsi yang menjelaskan kelas data yang labelnya belum dikatehui dengan memperkirakan kelas tersebut. Salah satu algoritma klasifikasi yang sering digunakan yaitu Algoritma *K-Nearest Neighbor* [14]. Algoritma *K-Nearest Neighbor* dikenalkan pada tahun 1967 oleh Cover dan Hart [15], sering disebut sebagai lazy *learner*. *K-Nearest Neighbor* merupakan suatu metode klasifikasi terhadap data objek yang didasarkan pada data pembelajaran yang jaraknya dekat dengan perhitungan menggunakan jarak *euclidean* [16]. Tujuannya yaitu untuk mengklasifikasi data baru berdasarkan atribut dan data training dengan klasifikasi berdasarkan mayoritas kategorinya, beberapa penelitian menggunakan Algoritma telah dilakukan dalam beberapa bidang kesehatan, *e-commerce*, transportasi dan fisika [17–21]. Berdasarkan penelitian yang telah dilakukan algoritma *K-Nearest Neighbor* menghasilkan tingkat akurasi yang baik dalam berbagai kasus tersebut. Pada

penelitian ini dikombinasikan antara metode reduksi dimensi menggunakan PCA dengan algoritma *K-Nearest Neighbor* dan selanjutnya diuji bagaimana performa dari kombinasi keduanya pada kasus penyakit diabetes.

Bagaimana hasil klasifikasi yang terbentuk dan bagamana tingkat akurasi menjadi menarik untuk diteliti. Bagaimana mendapatkan hasil akurasi terbaik pada proses pengklasifikasian data besar dengan menyederhanakan data, mereduksi dimensi dan waktu komputasi, serta pengelompokan variabel yang memiliki kesamaan dari segi jarak. Bagaimana performa Algortima *K-Nearest Neighbor* dengan reduksi dimensi PCA yang diterapkan pada kasus penyakit diabetes menjadi fokus penelitian ini.

2. Metode

Pada penelitian ini metode yang digunakan pada proses reduksi dimensi yaitu menggunakan metode *Principal Component Analysis* dan pada klasifikasi menggunakan metode *K-Nearest Neiighbor*. Dimana kedua metode tersebut akan diterapkan pada data penyakit diabetes dengan jumlah data 798 dengan 8 variabel independen dan 1 variabel dependen. Data tersebut merupakan data sekunder yang diperoleh dari *Kaggle*.

Adapun langkah-langkah dari penelitian ini yaitu normalisasi data dengan menggunakan *Z-Score*. Selanjutnya melakukan reduksi dimensi dengan membentuk matriks kovarian, menghitung nilai eigen dan vektor eigen, dan membentuk komponen utama dari hasil vektor eigen yang diperoleh. Kemudian dari hasil reduksi dimensi akan dilakukan klasifikasi dengan mencari parameter K, menentukan jarak, dan mencari kelas mayoritas yang kemudian hasilnya akan dievaluasi dan dilihat bagaimana penerapan metode ini pada data penyakit diabetes.

Adapun langkah-langkah pada penelitian ini adalah sebagai berikut :

1. Normalisasi merupakan salah satu tahapan preprocessing yang biasa digunakan untuk perentangan data tinggi yang mempengaruhi hasil dari pengolahan data. data menggunakan *Z-Score* dengan persamaan berikut :

$$Z_{i,j} = \frac{x_{i,j} - \overline{X}}{\sigma},\ i = 1,2,3,\ldots,n, j = 1,2,3,\ldots,m. \tag{1}$$

2. Reduksi dimensi menggunakan *Principal Component Analysis* dengan langkah sebagai berikut :

   (a) Menghitung matriks kovarian
   Matriks kovarian digunakan pada saat mencari nilai eigen dan vektor eigen. Untuk menentukan matriks kovarian digunakan persamaan (2):

$$Cov\left(x_j, x_k\right) = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_{i,j} - \overline{X}_j\right)\left(x_{i,k} - \overline{X}_k\right), j \text{ dan } k = 1,2,3,\ldots,n \tag{2}$$

   Dari persamaan (2) dilakukan perhitungan untuk setiap variabel dan hasil dari masing – masing data disusun dalam matriks sehingga diperoleh matriks kovarian pada persamaan (3), yaitu

$$S = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & s_{n,n} \end{bmatrix}_{n \times n}. \tag{3}$$

(b) Menghitung nilai eigen dan vektor eigen
Untuk mendapatkan nilai eigen (λ) digunakan persamaan (4), yaitu dengan mengambil determinan dari matriks persamaan $S - \lambda I = 0$, yaitu

$$|S - \lambda I| = 0. \tag{4}$$

Setelah diperoeh nilai eigen (λ), selanjutnya dicari vektor eigen menggunakan persamaan (5), yaitu

$$(S - \lambda I)\, a = 0. \tag{5}$$

(c) Menentukan komponen utama
Penentuan komponen utama diperoleh dari kombinasi linear antara matriks vektor eigen dan matriks data sehingga diperoleh persamaan berikut:

$$PC_1 = \sum_{j=1}^{n} a_{j,1} x_j = a_{1,1}x_1 + a_{2,1}x_2 + \cdots + a_{n,1}x_n$$

$$PC_2 = \sum_{j=1}^{n} a_{j,2} x_j = a_{1,2}x_1 + a_{2,2}x_2 + \cdots + a_{n,2}x_n$$

$$\vdots$$

$$PC_n = \sum_{j=1}^{n} a_{j,n} x_j = a_{1,n}x_1 + a_{2,n}x_2 + \cdots + a_{n,n}x_n$$

.
Selanjutnya dapat dibentuk matriks sebagai berikut:

$$\begin{bmatrix} PC_1 \\ PC_2 \\ \vdots \\ PC_n \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{2,1} & \cdots & a_{n,1} \\ a_{1,2} & a_{2,2} & \cdots & a_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1,n} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix}_{n \times n} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n \times 1}$$

sehingga diperoleh persamaan (6), yaitu

$$PC_{n \times 1} = a^T X^T_{n \times 1} \tag{6}$$

dengan
a^T = Transpose matriks vektor eigen,
PC = Matriks variabel baru,
X^T = Transpose matriks data.
Setelah diperoleh komponen utama selanjutnya dilakukan klasifikasi dengan membagi data menjadi 2, masing-masing untuk data *training* dan data *testing* dengan perbandingan 9:1.

3. Klasifikasi menggunakan *K-Nearest Neighbor*
   Tahap ketiga adalah melakukan klasifikasi menggunakan algoritma *K-Nearest Neighbor,* untuk menemukan fungsi yang menjelaskan kelas data dan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Adapun tahapan pada *K-Nearest Neighbor* adalah sebagai berikut:
   (a) Menentukan parameter K atau banyaknya ketetanggaan terdekatnya.
   (b) Menghitung jarak antara data *training* dan data *testing* menggunakan persamaan *euclidean* sebagai berikut:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, i = 1,2,3,\ldots,n \tag{7}$$

   dengan
   x_i = data *training* pada data ke-i,
   y_i = data *testing* pada data ke-i,
   $d(x,y)$ = jarak *euclidean*.
   (c) Mengurutkan jarak yang diperoleh secara ascending atau dari yang terkecil ke yang terbesar.
   (d) Menentukan kelas yang bersesuaian.
   (e) Kelas yang mayoritas yang akan menentukan atau membentuk kelas pada data baru.
4. *Confussion Matriks*
   Untuk menguji performa algoritma *K-Nearest Neghbor* digunakan *confussion matriks. Confussion matriks* digunakan untuk mengetahui informasi prediksi antara data aktual dengan data prediksi hasil dari klasifikasi. Adapun *confussion matriks* dapat dilihat pada Tabel 1.

Tabel 1. Tabel *confussion matriks*

Kelas Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Keterangan
TP = Banyaknya data positif yang terklasifikasi dengan benar (*True Positif*),
FP = Banyaknya data positif yang terklasifikasi dengan salah (*False Positif*),
FN = Banyaknya data negatif yang terklasifikasi dengan salah (*False Negatif*),
TN = Banyaknya data negatif yang terklasifikasi dengan benar (*True Negatif*).

Perhitungan nilai akurasi merupakan rasio dari semua data dan terklasifikasi benar yang dihitung menggunakan persamaan (8), yaitu

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}. \tag{8}$$

Untuk rasio dari hasil yang terklasifikasi benar (positif) dan data prediksi positif, maka digunakan persamaan presisi (*precission*) yang ditampilkan pada persamaan (9), yaitu

$$Presisi = \frac{TP}{TP + FP}. \tag{9}$$

Nilai sensitivitas (*recall*) yang merupakan rasio dari semua data yang terklasifikasi positif dan diprediksi benar (positif) dihitung digunakan persamaan (9), yaitu

$$Recall = \frac{TP}{TP + FN}, \tag{10}$$

dan untuk *F1-Score* nya digunakan persamaan (11), yaitu

$$F1 - Score = 2\left(\frac{presisi * recall}{presisi + recall}\right). \tag{11}$$

Adapun Proporsi Total Variansi (PTV) dihitung dengan menggunakan persamaan (12), yaitu

$$\text{PTV} = \frac{\lambda_i}{\sum_{i-1}^{8} \lambda_i} \times 100\%, \text{ dengan } i = 1, 2, \ldots, 8. \tag{12}$$

3. Hasil dan Pembahasan

Pada penelitian ini digunakan data penyakit Diabetes dengan jumlah 798 data dengan 8 variabel independen dan 1 variabel dependen. Variabel-variabel yang terdapat pada data penyakit diabetes tersebut yaitu terdiri dari jumlah kehamilan (x_1), konsentrasi glukosa(x_2), tekanan darah (x_3), jenis ketebalan kulit(x_4), jenis insulin (x_5), indeks massa tubuh(x_6), jenis fungsi silsilah diabetes (x_7), dan umur (x_8). Selanjutnya dilakukan reduksi dimensi menggunakan PCA dan Algoritma *K-Nearest Neigbor*.

3.1. Principal Component Analysis (PCA)

PCA digunakan untuk mengurangi variabel yang saling berkorelasi menjadi variabel baru yang bebas tanpa menghilangkan informasi-informasi penting pada data. Korelasi antar variabel dilakukan dengan menentukan matriks varians kovarians (S). Adapun langkah – langkah metode PCA adalah sebagai berikut :

1. Langkah pertama adalah melakukan normalisasi data, menggunakan persamaaan (1) sehingga diperoleh matriks data yang telah dinormalisasi sebagai berikut:

$$Z = \begin{bmatrix} 0.639 & 0.848 & \cdots & 1.426 \\ -0.844 & -1.123 & \cdots & -0.190 \\ \vdots & \vdots & \ddots & \vdots \\ -0.844 & -0.873 & \cdots & -0.087 \end{bmatrix}_{798\times 8}.$$

2. Langkah kedua adalah mencari matriks kovarian menggunakan persamaan (2) dan disajikan sesuai dengan persamaan (3), untuk menguji korelasi antar variabel. Hasil perhitungan diperoleh matriks S sebagai berikut :

$$S = \begin{bmatrix} 1.0013 & 0.1296 & \cdots & 0.5450 \\ 0.1296 & 1.0013 & \cdots & 0.1296 \\ \vdots & \vdots & \ddots & \vdots \\ 0.5450 & 0.2638 & \cdots & 1.0013 \end{bmatrix}_{8\times 8}.$$

3. Langkah ketiga yaitu mencari nilai eigen dan vektor eigen untuk pengambilan komponen utama. Variabel baru yang terbentuk diambil dari nilai eigen ≥ 1 dimana nilai eigen tersebut menunjukkan besarnya sumbangan dari komponen terhadap variansi seluruh variabel asli. Dalam menentukan jumlah komponen utama yang akan dipilih yaitu dengan mencari nilai proporsi kumulatif variansi dan nilai eigen terbesar. Nilai eigen diperoleh dengan merujuk pada persamaan (4) dan proporsi total variansi diperoleh menggunakan persamaan (12). Hasil perhitungan secara lengkap disajikan pada Tabel 2.

Tabel 2. Nilai eigen, proporsi total variansi, dan proporsi kumulatif

Komponen	Nilai Eigen (λ)	Proporsi Total Variansi (%)	Kumulatif(%)
1	2,09711	26,18	26,18
2	1,73346	21,64	47,82
3	1,03097	12,87	60,69
4	0,87667	10,94	71,63
5	0,76333	9,53	81,16
6	0,68351	8,53	89,70
7	0,42036	5,52	94,94
8	0,40498	5,06	100

Pada Tabel 2 didapatkan 3 komponen utama dimana pengambilan komponen utama tersebut didasarkan pada nilai eigen $\lambda \geq 1$. Komponen utama pertama yang didapatkan yaitu dengan nilai eigen 2,09711 dengan variansi 26,18%. Komponen utama kedua yaitu dengan nilai eigen 1,73346 dengan variansi 21,64%. Berikutnya komponen utama yang ketiga yaitu nilai eigen 1,03097 dengan variansi 12,87%. Ketiga komponen tersebut dapat menjelaskan total variansi kumulatif sebesar 60,69%. Selanjutnya dicari nilai vektor eigen berdasarkan nilai eigen yang diperoleh sebelumnya. Hasil yang diperoleh mengindikasikan bahwa terdapat 3 komponen utama yang selajutnya elemen penyusun ketiga komponen utama tersebut dipilih berdasarkan nilai *loading* terbesar. Nilai *loading* merupakan nilai identifikasi korelasi dari faktor yang terbentuk dengan variabel. Semakin dekat hubungan faktor dengan variabel maka semakin besar juga nilai *loading*nya. Nilai *loading* dari ketiga komponen tersebut disajikan pada Tabel 3.

Tabel 3. Nilai *loading*

Variabel	PC_1	PC_2	PC_3
x_1	-0,128432	-0,593786	**0,013087**
x_2	-0,393083	**-0,174029**	-0,467923
x_3	-0,360003	-0,183892	**0,535494**
x_4	-0,439824	**0,331965**	0,237674
x_5	-0,435026	**0,250781**	-0,336709
x_6	-0,451941	0,100960	**0,361865**
x_7	-0,270611	**0,122069**	-0,433189
x_8	-0,198027	-0,620598	**-0,075248**

Tabel 3 menjelaskan tentang hubungan antara variabel asli dengan variabel baru yang dibentuk dengan *Principal Component Analysis.* Tabel 3 menunjukkan bahwa nilai *loading* terbesar pada masing – masing variabel x ditunjukkan oleh nilai yang dicetak tebal, yaitu x_2, x_4, x_5, dan x_7 pada PC_2, x_1, x_3, x_6, dan x_8 pada PC_3, sementara tidak terdapat variabel yang memenuhi pada PC_1. Variabel-vaariabel

tersebut yang selanjutnya dikelompokkan kedalam komponen utama PC yang baru. Adapun penjelasan hubungan antar variabel tersebut dapat dilihat pada Tabel 4.

Tabel 4. Hasil PCA berdasarkan nilai *loading*

PC	*Loading*	Variansi yang dijelaskan
PC_2	x_2 x_4 x_5 x_7	21,64%
PC_3	x_1 x_3 x_6 x_8	12,87%

Berdasarkan nilai *loading* pada Tabel 3 dan rekapan hasil PCA nilai *loading* pada Tabel 4, ditunjukkan variabel-variabel yang mewakili setiap komponen utama (PC). Komponen utama pertama atau PC_1 tidak diwakili oleh variabel apapun, sementara untuk komponen utama kedua atau PC_2 diwakili oleh variabel konsentrasi glukosa (x_2), jenis ketebalan kulit (x_4), jenis insulin (x_5), dan silsilahh diabetes (x_7). Adapun untuk komponen utama ketiga atau PC_3 diwakili oleh variabel kehamilan (x_1), variabel tekanan darah (x_3), indeks massa tubuh (x_6), dan variabel umur atau (x_8). Selanjutnya ditentukan nilai dari komponen utama yang terbentuk dari kombinasi linear berikut:

$$\begin{aligned}
PC_1 &= \sum_{i=1}^{8} a_{j1}^T X_j^T = a_{1,1}X_1 + a_{2,1}X_2 + a_{3,1}X_3 + \cdots + a_{8,1}X_8 \\
&= -0,12843X_1 - 0,39308X_2 - 0,36000X_3 + \cdots - 0,19802X_8 \\
PC_2 &= \sum_{i=1}^{8} a_{j2}^T X_j^T = a_{1,2}X_1 + a_{2,2}X_2 + a_{3,2}X_3 + \cdots + a_{8,2}X_8 \\
&= -0,59378X_1 - 0,17402X_2 - 0,18389X_3 + \cdots - 0,62058X_8 \\
PC_3 &= \sum_{i=1}^{8} a_{j3}^T X_j^T = a_{1,3}X_1 + a_{2,3}X_2 + a_{3,3}X_3 + \cdots + a_{8,3}X_8 \\
&= 0,01307X_1 - 0,467923X_2 - 0,535494X_3 + \cdots - 0,075248X_8
\end{aligned}$$

yang menghasilkan nilai komponen utama pada Tabel 5.

Tabel 5. Nilai komponen utama PC

PC_1	PC_2	PC_3
-0,068503	-1,234895	-0,095930
1,121683	0,733852	0,712938
⋮	⋮	⋮
1,060324	-0,837062	-0,425030
0,839892	1,151755	1,009178

Tabel 5 merupakan hasil perolehan data yang telah direduksi dimensi dengan menggunakan *Principal Component Analysis* dimana dari data yang awalnya

mempunyai 8 variabel setelah direduksi berubah menjadi data dengan 3 variabel baru. Data tersebut selanjutnya digunakan untuk melakukan klasifiksi dengan menggunakan *K-Nearest Neighbor*.

3.2. Algoritma K-Nearest Neighbor

Proses klasifikasi dengan *K-Nearest Neighbor* memiliki input data sebesar 798 data yang selanjutnya dibagi menjadi data *training* dan data *testing* dengan perbandingan 90% data *training* dan 10% data *testing*. Pemilihan nilai K pada penelitian ini dibatasi untuk nilai K = 3, K = 5, dan K = 7, untuk menguji bagaimana performa algoritma *K-Nearest Neighbor*. Hasil prediksi kelas disajikan pada Tabel 6.

Tabel 6. Hasil prediksi kelas sesuai dengan K

K=3	*K=5*	*K=7*
0	0	0
1	1	1
0	0	0
⋮	⋮	⋮
0	0	0

Pada tahapan ini dilakukan evaluasi hasil dari ketepatan klasifikasi terhadap data penyakit diabetes dengan menggunakan *confussion matriks.* Hasil evaluasi disajikan pada Tabel 7.

Tabel 7. *Confussion* matriks pada *K=3, K=5* dan *K=7*

Kelas Aktual	Prediksi *K=3*		Prediksi *K=5*		Prediksi *K=7*	
	Positif	Negatif	Positif	Negatif	Positif	Negatif
Positif	11	11	13	9	14	11
Negatif	14	41	12	43	6	46

Setelah itu dihitung nilai akurasi, presisi, *Recall* dan *F1 Score* menggunakan persamaan (8-11). Performa dari hasil klasifikasi menggunakan *K-Nearest Neighhbor* dengan bantuan *Jupyter Notebook* disajikan pada Tabel 8.

Tabel 8. Hasil akursi, presisi, *recall* dan *F1 score KNN*

	K=3	*K=5*	*K=7*
Akurasi	67,53%	72,72%	77,92%
Presisi	50%	59%	70%
Recall	44%	52%	56%
F1-Score	47%	55%	62%

Berdasarkan nilai pada Tabel 7 dan Tabel 8 dapat disimpulkan bahwa hasil uji performansi menggunakan 3 kelas klasifikasi, setelah dilakukan reduksi dimensi menggunakan PCA menujukkan bahwa hasil akurasi, presisi, *Recall* dan *F1 Score* yang tertinggi terdapat pada kelas klasifikasi *K=7*, yaitu sebesar 77,92%.

4. Kesimpulan

Hasil dari penerapan reduksi dimensi dengan PCA mampu menghasilkan 3 komponen utama dari variabel baru dari 8 variabel asli, komponen utama yang terbentuk adalah

PC1, PC2 dan PC3. PC_1tidak terdapat variabel yang mewakili, PC_2 meliputi variabel konsentrasi glukosa (x_2), jenis ketebalan kulit (x_4), jenis insulin (x_5) dan silsilah diabetes (x_7). PC_3 terdapat variabel kehamilan (x_1), variabel tekanan darah (x_3), indeks massa tubuh (x_6), dan variabel umur atau (x_8). Hasil performa klasifikasi menggunakan Algoritma *K-Nearest Neighbor* dengan 3 kelas klasifikasi menghasilkan nilai akurasi tertinggi pada parameter kelas klasifikasi K=7. Pada kasus data penyakit diabetes, data positif yang terklasifikasi dengan benar sebesar 14 data, data negatif yang terklasifikasi dengan benar sebesar 46 data, data positif yang terklasifikasi dengan salah sebesar 11 data, dan data negatif yang terklasifikasi dengan salah sebesar 6 data. Hasil yang diperoleh menunjukkan bahwa klasifikasi menggunakan Algoritma *K-Nearest Neighbor* dengan data yang sudah direduksi dimensi menggunakan PCA menghasilkan performa yang baik.

Referensi

[1] M. A. Mujib and A. Jatissidi, "Perancangan Buku Ilustrasi Mengenai Panduan Penyakit Deabetes," *Pantarei*, vol. 5, no. 2, pp. 1–8, 2021.

[2] Kemenkes, *Buku Pintar Kader POSBINDU*. Jakarta: Direktorat Jenderal Pencegahan dan Pengendalian Penyakit, 2019.

[3] R. Firliana, R. Wulanningrum, and W. Sasongko, "Implementasi Principal Component Analysis (PCA) Untuk Pengenalan Wajah Manusia," *Nusantara of Engineering*, vol. 2, no. 1, pp. 65–69, 2015.

[4] E. Joko and S. Suparman, "Reduksi Dimensi untuk Meningkatkan Performa Metode Fuzzy Klastering pada Big Data," *Science, Technology, Engineering, Economics, Education, and Mathematics*, vol. 1, no. 1, pp. 27–36, 2020.

[5] J. F. Hair, "Multivariate Data Analysis: An Overview," in *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 904–907, doi: 10.1007/978-3-642-04898-2_395.

[6] W. K. Härdle and L. Simar, *Applied Multivariate Statistical Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, doi: 10.1007/978-3-662-45171-7.

[7] I. T. Jolliffe, *Principal Component Analysis*, ser. Springer Series in Statistics. New York: Springer-Verlag, 2002, doi: 10.1007/b98835.

[8] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. London: Pearson Education Inc., 2002.

[9] M. Wangge, "Penerapan Metode Principal Component Analysis (PCA) Terhadap Faktor-faktor yang Mempengaruhi Lamanya Penyelesaian Skripsi Mahasiswa Program Studi Pendidikan Matematika FKIP UNDANA," *Jurnal Cendekia : Jurnal Pendidikan Matematika*, vol. 5, no. 2, pp. 974–988, apr 2021, doi: 10.31004/cendekia.v5i2.465.

[10] T. Rahmawati, "Aplikasi Principal Component Analysis (Pca) Untuk Mereduksi Faktor-Faktor Yang Berpengaruh Dalam Peramalan Konsumsi Listrik," *Teknomatika*, vol. 7, no. 1, pp. 31–41, 2014.

[11] T. Saepurohman and B. E. Putro, "Analisis Principal Component Analysis (PCA) Untuk Mereduksi Faktor-Faktor yang Mempengaruhi Kualitas Kulit Kikil Sapi," in *Seminar dan Konferensi Nasional IDEC 2019*, Surakarta, 2019.

[12] F. Fitrianingsih and S. Sugiyarto, "Implementasi Analisa Komponen Utama untuk Mereduksi Variabel yang Mempengaruhi Perbaikan pada Fungsi Ginjal Tikus," *Jurnal Ilmiah Matematika*, vol. 6, no. 2, pp. 62–68, oct 2019, doi: 10.26555/konvergensi.v6i2.19549.

[13] M. S. Noya van Delsen, A. Z. Wattimena, and S. Saputri, "Penggunaan Metode Analisis Komponen Utama Untuk Mereduksi Faktor-Faktor Inflasi di kota Ambon," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 11, no. 2, pp. 109–118, dec 2017, doi: 10.30598/barekengvol11iss2pp109-118.

[14] Yu-Wei and D. Chiu, *Machine Learning with R Cookbook*. Birmingham: Packt Publishing, 2015.

[15] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, jan 1967, doi: 10.1109/TIT.1967.1053964.
[16] A. Setiawan, "Perbandingan Penggunaan Jarak Manhattan, Jarak Euclid, dan Jarak Minkowski dalam Klasifikasi Menggunakan Metode KNN pada Data Iris," *Jurnal Sains dan Edukasi Sains*, vol. 5, no. 1, pp. 28–37, may 2022, doi: 10.24246/juses.v5i1p28-37.
[17] M. R. Alghifari and A. P. Wibowo, "Penerapan Metode K-Nearest Neighbor Untuk Klasifikasi Kinerja Satpam Berbasis Web," *Jurnal Teknologi dan Manajemen Informatika*, vol. 5, no. 1, pp. 1–10, jun 2019, doi: 10.26905/jtmi.v5i1.3074.
[18] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 29–33, jul 2020, doi: 10.33096/ijodas.v1i2.11.
[19] I. A. Angreni, S. A. Adisasmita, M. I. Ramli, and S. Hamid, "Pengaruh Nilai K Pada Metode K-Nearest Neighbor (KNN) Terhadap Tingkat Akurasi Identifikasi Kerusakan Jalan," *Rekayasa Sipil*, vol. 7, no. 2, pp. 63–70, jan 2019, doi: 10.22441/jrs.2018.v07.i2.01.
[20] M. S. Fajri, N. Septian, and E. Sanjaya, "Evaluasi Implementasi Algoritma Machine Learning K-Nearest Neighbors (kNN) pada Data Spektroskopi Gamma Resolusi Rendah," *Al-Fiziya: Journal of Materials Science, Geophysics, Instrumentation and Theoretical Physics*, vol. 3, no. 1, pp. 9–14, aug 2020, doi: 10.15408/fiziya.v3i1.16180.
[21] D. Sebastian, "Implementasi Algoritma K-Nearest Neighbor untuk Melakukan Klasifikasi Produk dari beberapa E-marketplace," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 5, no. 1, pp. 51–61, may 2019, doi: 10.28932/jutisi.v5i1.1581.