 ISSN:2302-8149
e-ISSN:2540-9719

Implementasi *Principal Component Analysis* (PCA) dan *Gap Statistic* untuk *Clustering* Kanker Payudara pada Algoritma *K-Means*

Implementation of Principal Component Analysis (PCA) and Gap Statistic for Breast Cancer Clustering in the K-Means Algorithm

[1]Ridha Afifa, [2]Muhammad Itqan Mazdadi*, [3]Triando Hamonangan Saragih, [4]Fatma Indriani, [5]Muliadi
[1,2,3,4,5]Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat
[1,2,3,4,5]Jl. A.Yani Km 36, Banjarbaru, Indonesia
*e-mail: ridhaafifa20@gmail.com

(***received*:** 13 March 2024, ***revised*:** 17 August 2024, ***accepted*:** 11 September 2024)

Abstrak

Kanker payudara menjadi salah satu penyebab kematian paling umum di dunia. Kanker payudara dapat dideteksi menggunakan *data mining*, di mana data diekstraksi menjadi informasi yang berguna. *Clustering* kanker payudara dilakukan untuk membantu pihak medis dalam mengelompokkan karakteristik setiap jenis kanker. Namun, pada data kanker payudara terdapat multikolinieritas data sehingga dapat mempengaruhi hasil *clustering*. Untuk menangani masalah tersebut ditangani menggunakan reduksi dimensi *Principal Component Analysis* (PCA). Metode *Principal Component Analysis* dapat mengatasi masalah multikolinieritas data dan meningkatkan efisiensi komputasi. Selain itu metode *K-Means* juga memiliki kelemahan dalam menentukan jumlah kluster yang optimal, sehingga digunakan metode *Gap Statistic* untuk mencari nilai K optimal yang cocok digunakan pada data kanker payudara. Dalam penelitian ini, dilakukan perbandingan hasil evaluasi dari model *clustering K-Means*, gabungan model *clustering PCA-KMeans* dan gabungan model *clustering PCA-GapStatistic-KMeans*. Dari penelitian ini, didapatkan hasil evaluasi pada model *clustering K-Means* dengan reduksi dimensi PCA dan K optimal *Gap Statistic* lebih baik dibandingkan model *K-Means* tanpa reduksi dimensi. Dengan jumlah kluster yang dihasilkan oleh *Gap Statistic* sebanyak 2 kluster dan hasil evaluasi yang diperoleh sebesar 1.195513.

Kata kunci: kanker payudara, *clustering, k-means, principal component analysis, gap statistic*

Abstract

Breast cancer is one of the most common causes of death worldwide. Data mining can be utilized to detect breast cancer, where information is extracted from data to provide valuable insights. Clustering of breast cancer is conducted to assist medical professionals in grouping the characteristics of each cancer type. However, multicollinearity in breast cancer data can impact clustering results. To address this issue, dimensionality reduction through Principal Component Analysis (PCA) is employed. PCA can effectively handle multicollinearity issues and enhance computational efficiency. Additionally, the K-Means method has limitations in determining the optimal number of clusters. Therefore, the Gap Statistic method is employed to find the optimal K value suitable for breast cancer data. This study compares the evaluation results of the K-Means clustering model, the combined PCA-KMeans clustering model, and the combined PCA-GapStatistic-KMeans clustering model. The findings indicate that the evaluation results for the K-Means model with PCA dimensionality reduction and optimal Gap Statistic K are superior to the K-Means model without dimensionality reduction. The Gap Statistic suggests 2 clusters as the optimal number, with an evaluation result of 1.195513.

Keywords: *breast cancer, clustering, k-means, principal component analysis, gap statistic*

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Kanker payudara merupakan penyakit serius yang sering ditemukan pada wanita[1]. Berdasarkan *Global Cancer Observatory*, di Indonesia, kanker payudara menjadi jenis kanker yang paling umum terjadi, dengan jumlah kasus baru mencapai 65.858 (16,6%) dari total 396.914 kasus baru kanker pada tahun 2020[2]. Dengan meningkatnya jumlah kasus kanker payudara, diperlukan teknik *data mining* untuk membantu pihak medis dalam mengelompokkan diagnosis kondisi pasien kanker payudara berdasarkan variabel-variabel yang mempengaruhi perkembangan kanker payudara[3].

Clustering K-Means dikenal sebagai salah satu algoritma *clustering* yang paling sederhana dan umum karena kemampuannya dalam mengelompokkan data dalam jumlah yang besar dengan waktu komputasi yang relatif cepat dan efisien[4]. Namun, metode ini memiliki beberapa kelemahan, antara lain kesulitan dalam menentukan nilai *cluster* (K) yang tepat dan sensitif terhadap perubahan data [5]. Penetuan jumlah *cluster* yang kurang tepat dapat mempengaruhi informasi *cluster* yang terbentuk dan menghasilkan cluster yang tidak optimal[6]. Untuk menangani penentuan jumlah *cluster* dapat menggunakan *Gap Statistic*. *Gap Statistic* dapat menentukan jumlah klaster lebih konstan dibandingkan pengukuran lainnya[7].

Dalam analisis klaster, setiap variabel diberi nilai yang sama dalam menghitung jarak. Jika terdapat korelasi antara beberapa variabel, hal tersebut dapat mengakibatkan hasil clustering memiliki pembobotan yang tidak seimbang di setiap clusternya[7]. Korelasi yang kuat antara dua atau lebih variabel disebut multikolinieritas[8]. Pengamatan terhadap multikolinearitas merupakan asumsi dalam analisis kluster non hirarki yaitu diharapkan tidak terjadi multikolinearitas atau tidak terdapat korelasi antar variabel[9]. Untuk menangani masalah multikolinearitas pada data, dapat ditangani dengan menggunakan *Principal Component Analysis* (PCA)[10]. Tujuan utama dari PCA adalah mereduksi data yang berdimensi tinggi menjadi lebih sedikit dengan tetap meminimalisasi resiko informasi yang hilang. Metode ini juga memiliki peran mengurangi adanya data *outlier* serta mengatasi asumsi multikolinieritas[11].

Tujuan dari penelitian ini adalah untuk mengetahui hasil pengelompokkan setiap karakteristik kanker payudara menggunakan metode *clustering K-Means*. Selanjutnya, penelitian ini akan dilakukan reduksi dimensi data menggunakan metode *Principal Component Analysis* (PCA) untuk mereduksi dimensi data yang memiliki korelasi yang kuat agar memiliki hasil pembobotan yang seimbang. Setelah itu, dilakukan pencarian jumlah *cluster* terbaik dengan menggunakan metode *Gap Statistic*. Hasil evaluasi dari beberapa pengujian model *clustering* akan dibandingkan untuk mengetahui nilai kinerja terbaik dalam menghasilkan kelompok-kelompok jenis kanker payudara. Diharapkan penelitian ini dapat dijadikan referensi yang berguna bagi pihak medis dalam memahami karakteristik tiap jenis kanker payudara, serta dapat menjadi dasar dalam pengambilan keputusan yang lebih baik dalam menangani pasien kanker payudara dan mengurangi prevalensi penyakit tersebut.

2 Tinjauan Literatur

Penelitian terdahulu [12]menggunakan *Covtype Dataset, Covtype-2 Dataset, Poker Dataset*, dan *Poker-2 Dataset* yang diambil dari *UCI Machine Learning*. Penelitian ini melakukan optimasi menggunakan *Gap Statistic* untuk menentukan jumlah *cluster* (k) yang optimal. Hasil akurasi dari metode *Gap Statistic* mencapai 76,3% dengan mengurangi kompleksitas dari algoritma *K-Means* standar. Hasil optimasi dapat meningkatkan efisiensi dan kecepatan proses pengelompokan.

Penelitian terdahulu [13] menggunakan data penyakit pasien dari Puskesmas Cigugur Tengah. Penelitian ini melakukan perbandingan algoritma *K-Means* dan algoritma *K-Medoids* untuk mengelompokan penyakit pasien berdasarkan penyakit akut dan penyakit tidak akut. Hasil dari penelitian didapatkan *cluster* model 241 data untuk penyakit akut dan 9 data untuk penyakit tidak akut menggunakan algoritma *K-Means*, serta 224 data untuk penyakit akut dan 26 data untuk penyakit tidak akut menggunakan algoritma *K-Medoids*. Algoritma *K-Means* lebih baik daripada *K-Medoids* dengan nilai *Davies Bouldin Index* yang lebih kecil sebesar -0.453, sedangkan algoritma *K-Medoids* sebesar -1.276.

Penelitian terdahulu [14] menggunakan metode *Principal Component Analysis* (PCA) dalam mereduksi data berdimensi tinggi untuk membandingkan hasil pengelompokan data kunjungan wisatawan asing yang diambil dari website BPS menggunakan metode *K-Means* dan gabungan metode *PCA K-Means*. *PCA K-Means* menghasilkan nilai evaluasi yang lebih kecil dibandingkan

 ISSN:2302-8149
e-ISSN:2540-9719

dengan metode *K-Means*, dengan nilai *Davies Bouldin Index* (DBI) sebesar 0,310, sementara *K-Means* hanya mendapatkan hasil nilai DBI sebesar 0,382. Hasil ini menunjukkan bahwa metode *PCA K-Means* mendapatkan nilai evaluasi yang lebih baik, sehingga model *PCA K-Means* digunakan untuk mengelompokan data kunjungan wisman ke Indonesia.

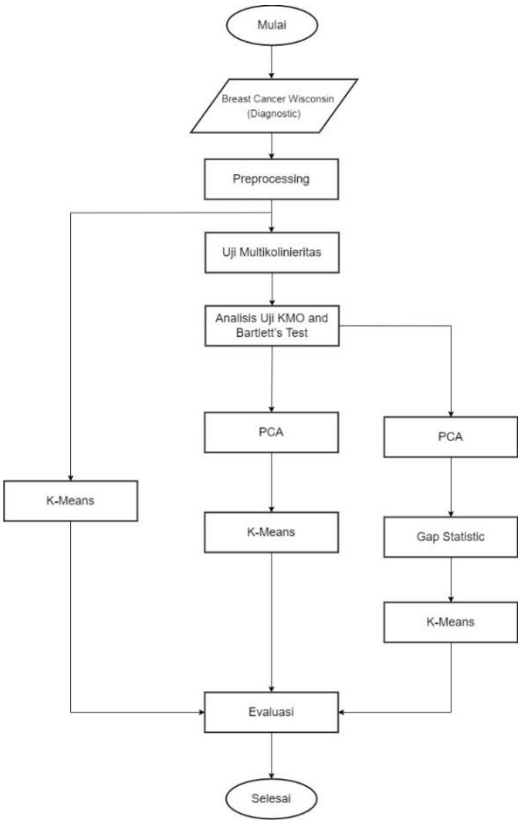
Penelitian terdahulu [15] menggunakan data angka kematian bayi di Kalimantan Barat pada tahun 2018. Penelitian tersebut menggunakan metode *Principal Component Analysis* (PCA) untuk menangani multikolinieritas pada data. Terdapat 2 variabel bebas yang mengalami multikolinieritas. Hasil yang diperoleh dari pembentukan komponen utama menunjukkan nilai sebesar 3,778 dengan *eigenvalue* yang melebihi angka 1. Selanjutnya, variansi kumulatif total mencapai 75,566%. Ini menunjukkan bahwa melalui metode komponen utama, masalah multikolinieritas berhasil diatasi dengan efektif.

Penelitian terdahulu [16] menggunakan *Wisconsin Breast Cancer Dataset*. Penelitian ini melakukan pengelompokkan data kanker payudara menggunakan algoritma *K-Means*. Pada penelitian ini *clustering* dikelompokkan menjadi 2 *cluster* dengan menggunakan metode *Elbow* untuk menentukan nilai K optimal. Hasil akurasi dari prediksi positif menunjukan akurasi sebesar 85% dalam mengklasifikasikan kanker sebagai ganas atau jinak.

Berdasarkan tinjauan literatur yang telah diuraikan, terdapat beberapa metode yang belum diterapkan pada dataset kanker payudara. Dapat diketahui pada penelitian [16] telah melakukan penelitian *clustering K-Means* untuk mengelompokkan data kanker payudara. Sehingga pada penelitian ini berfokus untuk menangani masalah pada data multikolinieritas dengan menggunakan metode *Principal Component Analysis* (PCA) dan kelemahan pada metode *K-Means* dalam menentukan jumlah *cluster* optimal menggunakan metode *Gap Statistic*.

3 Metode Penelitian

Adapun alur penelitian yang dilaksanakan dalam penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Alur penelitian

3.1 Pengumpulan Data

 ISSN:2302-8149
e-ISSN:2540-9719

Dataset yang digunakan yaitu dataset *public* dari *UCI Machine Learning Repository* yaitu *Breast Cancer Wisconsin (Diagnostic)*. Dataset ini memiliki 569 baris data dan 30 atribut. Berikut pada Tabel 1 merupakan sampel dari dataset *Breast Cancer Wisconsin (Diagnostic)*.

Tabel 1. Sampel dataset

radius_ mean	texture_ mean	…	symmetry_worst	fractal_dimension_worst
17.99	10.38	…	0.4601	0.1189
20.57	17.77	….	0.275	0.08902
19.69	21.25	…	0.3613	0.08758
…	…	…	…	…
7.76	24.54	…	0.2871	0.07039

3.2 *Preprocessing* Data

Tahapan *preprocessing* data pada penelitian ini adalah melakukan penanganan *outlier*, standarisasi data, uji multikolinieritas, uji asumsi analisis faktor dan reduksi dimensi.

a. Penanganan *Outlier*

Dataset kanker payudara memiliki data *outlier* diseluruh fitur datasetnya. Untuk mengatasi masalah tersebut ditangani dengan menggunakan metode *winsorizing*. Algoritma *K-Means* sensitif terhadap *outlier*, sehingga dengan menggunakan *winsorizing* dapat mengurangi pengaruh *outlier* sebelum dilakukan *clustering*. Teknik *winsorizing* dilakukan dengan cara mengurangi dampak nilai-nilai ekstrem dengan menggantinya dengan nilai-nilai yang berada pada persentil tertentu, tergantung pada distribusi data[17].

b. Standarisasi Data

Langkah selanjutnya adalah melakukan standarisasi data dengan menggunakan *Z-Score*, karena dataset pada penelitian ini memiliki perbedaan ukuran satuan yang besar pada data antar variabel[18]. Standarisasi *Z-Score* dilakukan dengan cara merubah rentang data agar berada pada standar deviasi 0 dan 1.

c. Uji Multikolinieritas

Setelah itu, akan dilakukan uji multikolinieritas sebelum dilakukan *clustering*. Uji multikolinieritas dilakukan untuk mengidentifikasi apakah ada korelasi tinggi antara variabel-variabel dalam dataset, yang dapat mempengaruhi hasil analisis kluster.

d. Reduksi Dimensi *Principal Component Analysis* (PCA)

Setelah melakukan uji multikolinieritas, jika terdapat korelasi tinggi antar variabel pada data, maka dilakukan reduksi dimensi dengan menggunakan metode *Principal Component Analysis* untuk menangani masalah multikolinieritas pada data. Sebelum dilakukan PCA, terlebih dahulu di lakukan uji asumsi analisis faktor terlebih dahulu dengan melakukan uji *Bartlett*, KMO *(Kaiser-Meyer-Olkin)*, dan MSA *(Measure of Sampling Adequacy)*. Jika uji sudah terpenuhi, selanjutnya melakukan PCA dengan menghitung matriks *covariance* pada atribut, menghitung nilai *eigen value* serta *eigen vector*, dan menghitung nilai *principal component*.

3.3 Pemodelan *Data Mining*

Setelah melalui tahap *preprocessing*, selanjutnya dilakukan proses tahapan menenemukan pola atau pengetahuan dalam data yang berjumlah besar. Metode yang digunakan pada tahap *data mining* yaitu algoritma *K-Means*. Dalam menentukan nilai K pada algoritma *K-Means* dilakukan optimasi dengan menggunakan metode *Gap Statistic*.

3.4 Evaluasi

Evaluasi merupakan tahapan hasil dari kinerja dari teknik *data mining* yang telah dilakukan. Tahap evaluasi pada penelitian ini melakukan perhitungan menggunakan *Davies-Bouldin Index* (DBI). Model yang di evaluasi adalah model *clustering K-Means*, *PCA-KMeans*, *PCA-Gap Statistic-KMeans*. Setelah nilai evaluasi *Davies-Bouldin Index* (DBI) dari ketiga model didapatkan, nilai tersebut dibandingkan untuk mengetahui model dengan nilai evaluasi yang terbaik.

 ISSN:2302-8149
e-ISSN:2540-9719

4 Hasil dan Pembahasan

Hasil penelitian ini akan dibahas secara berurutan sesuai dengan prosedur penelitian.

Tahap Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Breast Cancer Wisconsin (Diagnostic)* dari *UCI Machine Learning Repository*. Dataset kanker payudara ini memiliki jumlah 569 data dengan masing-masing 30 atribut yang menggambarkan karakteristik inti sel dari gambar *Fine Needle Aspirate* (FNA) pada massa payudara. Atribut-atribut dari dataset WDBC dapat dilihat pada Tabel 2.

Tabel 2. Atribut dataset

Nomor	Atribut
1.	radius_mean
2.	texture_mean
3.	perimeter_mean
4.	area_mean
5.	smoothness_mean
6.	compactness_mean
7.	concavity_mean
8.	concave points_mean
9.	symmetry_mean
10.	fractal_dimension_mean
11.	radius_se
12.	texture_se
13.	perimeter_se
14.	area_se
15.	smoothness_se
16.	compactness_se
17.	concavity_se
18.	concave points_se
19.	symmetry_se
20.	fractal_dimension_se
21.	radius_worst
22.	texture_worst
23.	perimeter_worst
24.	area_worst
25.	smoothness_worst
26.	compactness_worst
27.	concavity_worst
28.	concave points_worst
29.	symmetry_worst
30.	fractal_dimension_worst

Tahap *Preprocessing* Data

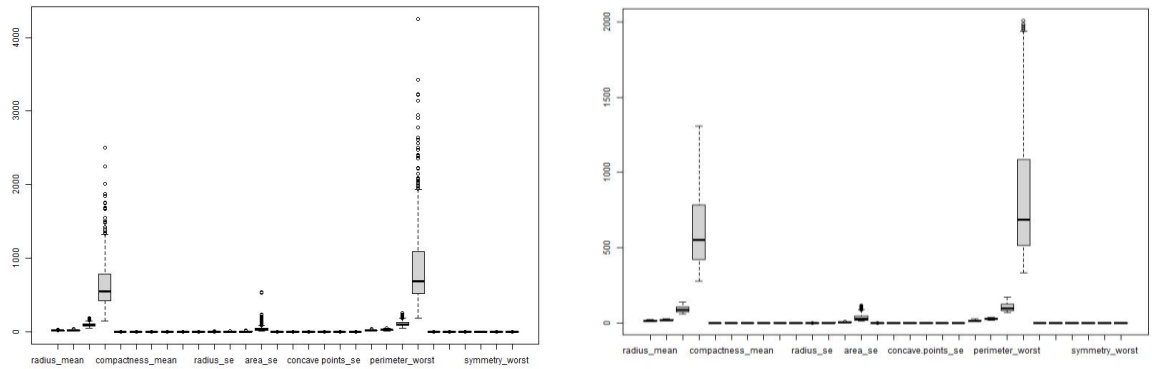
Sebelum memasuki tahap pelatihan model, data perlu melalui tahap *preprocessing* terlebih dahulu agar data yang digunakan bersih dan tidak berdampak buruk pada performa model. Proses *preprocessing* yang dilakukan yaitu penanganan *outlier*, standarisasi data menggunakan *Z-Score*,

 ISSN:2302-8149
e-ISSN:2540-9719

melakukan uji multikolinieritas, melakukan uji asumsi analisis faktor dan reduksi dimensi menggunakan metode *Principal Component Analysis*.

a. Penanganan *Outlier*

Dataset kanker payudara yang digunakan pada penelitian ini memiliki beberapa data yang *outlier* atau data yang menyimpang secara ekstrim dari rata-rata sekumpulan data yang ada. Dalam penelitian ini, penanganan *outlier* dilakukan dengan menggunakan Teknik *Winsorizing*.



Gambar 2. data sebelum dilakukan *winsorizing* (1) dan data setelah *winsorizing* (2)

Pada Gambar 2 (1) diketahui terdapat *outlier* di seluruh fitur dataset, sehingga *outlier* tersebut ditangani menggunakan teknik *winsorizing* dilakukan dengan menggantikan nilai *outlier* menjadi nilai batas ekor atas/bawah distribusi data sebesar 95% (1,5±IQR)[19]. Hasil dari penanganan *outlier* menggunakan *winsorizing* dapat dilihat pada Gambar 2 (2). Berikut pada Tabel 3 merupakan dataset yang telah dilakukan *winsorizing*.

Tabel 3. Data setelah *winsorizing*

Radius_mean	texture_mean	…	symmetry_worst	fractal_dimension_worst
17.9900	13.088	…	0.40616	0.118900
20.5700	17.770	….	0.27500	0.089020
19.6900	21.250	…	0.36130	0.087580
….	…	…	…	…
9.5292	24.540	…	0.28710	0.070390

b. Standarisasi *Z-Score*

Dataset yang sudah dilakukan penanganan outlier dilanjutkan dengan melakukan standarisasi data karena terdapat perbedaan ukuran satuan yang besar antara variabel-variabel yang diteliti. Metode transformasi yang digunakan pada penelitian ini menggunakan *Z-Score*. Berikut pada Tabel 4 merupakan dataset hasil standarisasi *Z-Score.*

Tabel 4. Data hasil standarisasi

radius_ mean	texture_ mean	…	symmetry_worst	fractal_dimension_worst
1.23797067	-1.57287747	…	2.337967483	2.30904628
2.04966164	-0.37158599	….	-0.261493902	0.36948240
1.77280580	0.52130045	…	1.448886683	0.27600944
….	…	…	…	…
-1.423872349	1.365437343	…	-0.021683878	-0.839824002

c. Uji Multikolinieritas

 ISSN:2302-8149
e-ISSN:2540-9719

Langkah berikutnya adalah melakukan pengujian multikolinieritas, yang bertujuan untuk mengetahui tidak adanya korelasi antar atribut atau multikolinearitas yang akan digunakan sebelum melakukan *clustering*. Uji multikolinearitas dilakukan dengan melihat nilai korelasi antar variabel, di mana nilai korelasi tidak melebihi 0.8.

Tabel 5. Nilai korelasi antar variabel

	radius_ mean	texture_ mean	perimeter_ mean	…	fractal_dimension_ worst
radius_mean	1	0.350096858	0.99796361	…	0.04203512
texture_mean	0.350096858	1	0.35733320	…	0.12250684
perimeter_mean	0.997963606	0.357333198	1	…	0.08677954
…	…	…	….	…	…
fractal_dimension_ worst	0.042035124	0.122506837	0.08677954	…	1

Berdasarkan Tabel 5, terlihat bahwa terdapat hubungan korelasi yang kuat antar variabel, di mana beberapa koefisien korelasi melebihi 0,8. Hal ini menunjukkan adanya multikolinearitas pada variabel-variabel penelitian dalam dataset *Breast Cancer Wisconsin*. Oleh karena itu, untuk mengatasi masalah multikolinearitas, metode yang dapat digunakan adalah *Principal Component Analysis* (PCA)[20].

d. Reduksi Dimensi *Principal Component Analysis* (PCA)

Principal Component Analysis diterapkan untuk menangani masalah multikolinieritas data. Metode PCA digunakan untuk mereduksi dimensi dataset sebelum di-*cluster* menggunakan *K-Means*. Sebelum PCA diterapkan, dataset dicek terlebih dahulu dengan melakukan beberapa pengujian asumsi dalam analisis faktor. Hasil uji asumsi analisis faktor yang dilakukan yaitu Uji *Kaiser Meyer Olkin* (KMO), Uji *Bartlett*, dan Uji *Measure Sampling of Adequacy* (MSA) pada Tabel 6 dan Tabel 7.

Tabel 6. Hasil uji KMO dan MSA

Keiser-Meyer-Olkin factor adequacy
Call : KMO (r = cor_bc)
Overall MSA = 0.85
MSA for each item =

radius_mean 0.83	texture_mean 0.67	perimeter_mean 0.90	area_mean 0.85
smoothness_mean 0.82	compactness_mean 0.94	concavity_mean 0.92	concave.points_ mean 0.93
symmetry_mean 0.82	fractal_dimension_ mean 0.87	radius_se 0.83	texture_se 0.50
perimeter_se 0.90	area_se 0.90	smoothness_se 0.65	compactness_se 0.86
concavity_se 0.88	concave.points_se 0.88	symmetry_se 0.57	fractal_dimension_ se 0.84
radius_worst 0.81	texture_worst 0.61	perimeter_worst 0.89	area_worst 0.84
smoothness_worst 0.75	compactness_worst 0.86	concavity_worst 0.88	concave.points_ worst 0.91
symmetry_worst	fractal_dimension_		

 ISSN:2302-8149
e-ISSN:2540-9719

0.70	worst 0.82

Tabel 7. Hasil uji *bartlett*

$chisq	43808.04
$p.value	0
$df	435

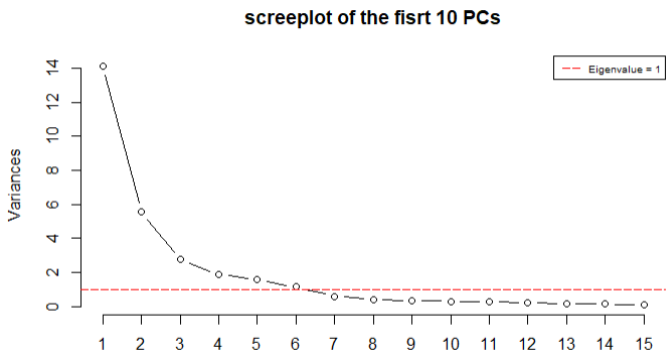
Pada Tabel 6 Uji KMO menunjukkan nilai sebesar 0,85, menunjukkan bahwa data layak untuk dilakukan PCA. Sedangkan hasil Uji MSA menujukkan nilai MSA> 0,5 sehingga dapat disimpulkan variabel dalam penelitian cocok untuk analisis faktor atau PCA. Uji *Bartlett* digunakan untuk mengetahui apakah matriks korelasi hubungan antara variabel adalah matriks identitas, sehingga diketahui korelasi antar variabel. Berdasarkan hasil uji *bartlett* pada Tabel 7 diperoleh nilai p.value= 0,000 < = a 0,05. Hal ini menunjukkan bahwa terdapat korelasi antar variabel sehingga perlu dilakukan PCA.

Setelah dilakukan uji analisis faktor, tahap selanjutnya dilakukan perhitungan kovarian untuk mengukur sejauh mana dua variabel dalam dataset berkorelasi satu sama lain. Semakin tinggi nilai kovarian antara dua variabel, semakin kuat hubungan linier di antara keduanya. Untuk nilai *covariance* antar variabel dapat dilihat pada Tabel 5.

Setelah mendapatkan matriks kovarian, langkah berikutnya adalah menghitung *eigen valu*e dari matriks kovarian. *Eigenvalue* bertujuan untuk mengukur seberapa banyak informasi yang diwakili oleh setiap komponen utama. Semakin besar nilai *eigenvalue*, semakin besar kontribusi komponen tersebut terhadap variabilitas data. Untuk melihat hasil perhitungan *eigenvalue* dapat dilihat pada Tabel 8.

Tabel 8. Hasil perhitungan nilai *eigen value*

	eigenvalue	variance.percent	cumulative.variance.persent
Dim.1	1.412695e+01	4.708984e+01	47.08984
Dim.2	5.541573e+00	1.847191e+01	65.56175
Dim.3	2.750679e+00	9.168930e+00	74.73068
Dim.4	1.881013e+00	6.270042e+00	81.00072
Dim.5	1.590088e+00	5.300293e+00	86.30101
Dim.6	1.149657e+00	3.832191e+00	90.13321
Dim.7	6.072773e-01	2.024258e+00	92.15746
Dim.8	4.341352e-01	1.447117e+00	93.60458
Dim.9	3.500970e-01	1.166990e+00	94.77157
Dim.10	3.129221e-01	1.043074e+00	95.81464



Gambar 3. Plot nilai *eigen value*

 ISSN:2302-8149
e-ISSN:2540-9719

Dari hasil perhitungan nilai eigen, keenam dimensi utama (Dim.1 hingga Dim.6) dengan *eigenvalue* di atas 1 dipertahankan, mencakup sekitar 90.13% variabilitas data, menunjukkan bahwa dimensi tersebut mampu menjelaskan lebih banyak variabilitas data dan memberikan keseimbangan yang baik antara akurasi representasi dan pengurangan dimensi data. Pada Gambar 3, menunjukkan grafik keenam dimensi ini memberikan informasi signifikan dan relevan untuk analisis lanjutan, mempertahankan sebagian besar informasi dari data asli.

Pada tahap selanjutnya dilakukan perhitungan nilai *eigenvector* dari matriks kovarian dan *eigenvalue*. Eigenvektor adalah vektor yang menjelaskan arah dari setiap komponen utama. Eigenvektor ini diurutkan berdasarkan nilai *eigenvalue* dari yang terbesar hingga terkecil. Berikut hasil perhitungan nilai *eigenvector* dapan dilihat pada Tabel 9.

Tabel 9. Hasil perhitungan *eigen vector*

PC1	PC2	PC3	PC4	PC5	…	PC30
-0.2117643	0.24026565	0.005683721	-0.0485079	0.011418336	…	-7.339680e
-0.1105344	0.05554569	-0.17096947	0.57809849	0.020168470	…	1.936934e
-0.2196480	0.22232977	0.009755436	-0.0470236	0.016217141	…	5.321963e
….	…	…	…	…		
-0.1370031	-0.2695277	0.234489825	0.09633447	0.117913495	…	-1.314609

Setelah mendapatkan *eigenvalue* dan *eigenvector* dari matriks kovarian data, langkah berikutnya adalah melakukan perhitungan *Principal Component Analysis* (PCA). Hasil eigen vektor yang telah didapatkan, kemudian di normalisasi dengan cara tiap nilai eigen vektor dibagi dengan panjang vektor. Setelah itu, dilakukan perhitungan jumlah hasil perkalian antara nilai eigen vektor yang dinormalisasi dengan data yang sudah di standarisasi. Sehingga didapatkan hasil perhitungan *principal component score*. Berikut Tabel 10 merupakan hasil data yang diperolah dari perhitungan *principal component*. Setelah melakukan reduksi dimensi menggunakan *Principal Component Analysis* didapatkan dataset baru dengan 6 fitur yang mewakili komponen utama.

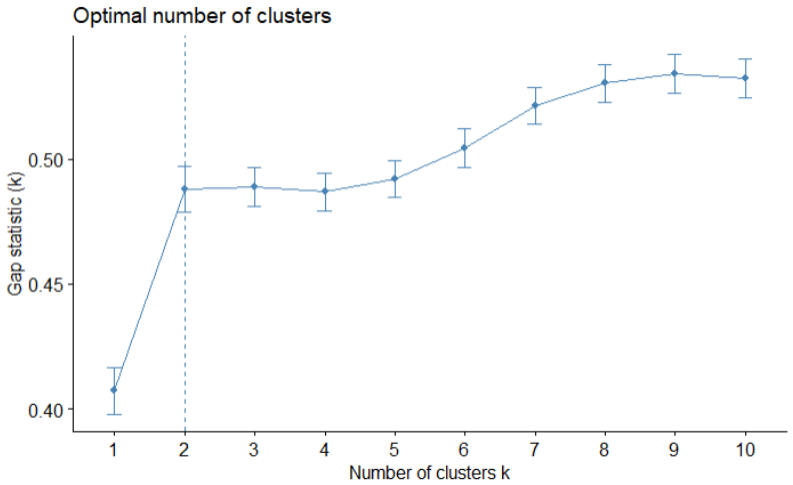
Tabel 10. Data hasil reduksi dimensi PCA

PC1	PC2	PC3	PC4	PC5	PC6
-9.1765708	-1.9017405	1.1762926	-3.3123171	-1.3911460	-1.3711676
-2.8768757	4.3665685	0.5696620	-1.2990994	0.1146076	0.0536220
-6.8039524	1.2291123	0.3638668	-1.1387568	-0.6452750	-0.6932412
…	…	…	…	…	…
4.9551476	0.5384587	-1.9432336	1.9329277	-0.3543618	-1.3116345

Gap Statistic

Hasil grafik pada program R pada Gambar 3 menunjukkan bahwa jumlah kluster optimal menggunakan metode *gap statistic* adalah 2 kluster. Hasil ini diperoleh dari grafik *gap statistic* pada Gambar 3 yang menunjukkan garis vertikal pada nilai 2, menandakan bahwa ini adalah jumlah kluster yang optimal. Berikut grafik nilai gap statistik yang diperoleh pada Gambar 3.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 3. Grafik nilai *gap statistic*

Clustering K-Means

Setelah dataset dilakukan reduksi dimensi menggunakan PCA, selanjutnya akan masuk pada tahap *clustering* menggunakan *K-Means*. Model *clustering* yang dibangun ada 3, yaitu model *clustering K-Means* menggunakan data standarisasi, model *clustering PCA-KMeans*, dan model *clustering PCA-Gap Statistic-KMeans*.

Dalam model *clustering K-Means* dan *clustering PCA-KMeans*, dilakukan uji dengan menggunakan berbagai jumlah kluster (K) seperti 2, 3, 4, 5, 6, 7, dan 8 melalui beberapa percobaan. Sedangkan pada model *Clustering PCA-Gap Statistic-KMeans*, dilakukan uji dengan menggunakan jumlah kluster (K) adalah 2 yang sudah diperoleh dari hasil perhitungan metode *Gap Statistic*.

Berikut Tabel 11, 12, dan 13 merupakan jumlah data yang diperoleh pada tiap cluster terbentuk.

Tabel 11. Jumlah data pada cluster k-means

Jumlah Kluster	Banyak Data							
	C1	C2	C3	C4	C5	C6	C7	C8
2	192	377						
3	110	124	335					
4	83	118	230	138				
5	220	87	83	47	132			
6	196	124	82	57	66	44		
7	96	140	58	44	82	51	98	
8	81	48	56	76	37	113	81	77

Tabel 12. Jumlah Data pada Cluster PCA-K-Means

Jumlah Kluster	Banyak Data							
	C1	C2	C3	C4	C5	C6	C7	C8
2	192	377						
3	124	110	335					
4	85	117	230	137				
5	88	132	51	218	80			
6	67	124	46	57	79	196		
7	58	94	141	80	54	96	46	
8	50	36	84	56	69	81	116	77

Tabel 13. Jumlah Data pada Cluster PCA-Gap Statistic-K-Means

 ISSN:2302-8149
e-ISSN:2540-9719

Jumlah Kluster	Banyak Data	
	C1	C2
2	192	377

Evaluasi *Davies Bouldin Index*

Tahap terakhir melakukan Evaluasi *clustering* menggunakan *Davies Bouldin Index*. Melalui perhitungan DBI, dapat diketahui seberapa baik sebuah cluster terbentuk dengan mempertimbangkan kedekatan antara cluster yang satu dengan yang lain. Semakin rendah nilai DBI, semakin baik hasil klasteringnya. Hasil evaluasi nilai Davies Bouldin Inedex (DBI) disajikan dalam Tabel 14.

Tabel 14. Perbandingan hasil evaluasi DBI

Jumlah Kluster	Hasil Evaluasi DBI		
	K-Means	PCA+K-Means	PCA+Gap Statistic+K-Means
2	1.298284	1.195513	1.195513
3	1.479132	1.341205	
4	1.803953	1.616263	
5	1.776404	1.592884	
6	1.808281	1.594894	
7	1.714904	1.514159	
8	1.690148	1.461446	

Dari hasil evaluasi ketiga pemodelan dapat diketahui pengujian mana yang memiliki nilai kinerja terbaik dalam meng-*cluster* kanker payudara. Berdasarkan Tabel 14 dapat diketahui jumlah *cluster* dan hasil evaluasi dari masing-masing pemodelan algoritma yang digunakan.

Dapat diketahui perbandingan hasil *clustering K-Means* menggunakan data yang telah distandarisasi, di mana terdapat korelasi tinggi antara variabel-variabel yang menyebabkan beberapa variabel memiliki bobot yang lebih besar daripada yang lain. Ketika dua variabel berkorelasi, konsep yang diwakili oleh keduanya akan memiliki pengaruh yang signifikan pada hasil akhir analisis. Sementara itu, hasil *clustering* pada data yang telah ditangani multikolinieritasnya dengan menggunakan reduksi dimensi PCA menunjukkan bahwa tidak ada korelasi yang kuat antar variabelnya. Hal ini menyebabkan cluster yang dihasilkan menjadi lebih stabil ketika perhitungan jaraknya di lakukan, serta dapat memperjelas pola *cluster* dengan mengidentifikasi dimensi yang paling informatif. Melalui evaluasi *cluster* menggunakan *Davies Bouldin Index*, model *PCA-K-Means* memberikan hasil evaluasi terbaik dibandingkan dengan model *K-Means* yang datanya tidak ditangani multikolinieritasnya.

Oleh karena itu, ketika model *PCA-KMeans* digunakan ke tahap selanjutnya, kemudian melakukan pengujian dengan berbagai jumlah *cluster*, pengujian yang dilakukan tidak dapat secara langsung menentukan jumlah cluster yang optimal. Sehingga diterapkan metode *Gap Statistic* untuk mencari nilai K yang optimal. Hasil dari *Gap Statistic* menunjukkan bahwa nilai *cluster* yang optimal adalah 2. Diketahui bahwa nilai kluster dari *Gap Statistic* mendapatkan nilai evaluasi terbaik ketika dibandingkan dengan berbagai nilai *cluster* yang telah diuji sebelumnya. Maka dari itu, *Gap Statistic* dapat mengatasi masalah penentuan nilai k optimal, dibandingkan melakukan pengujian beberapa kali percobaan K kluster.

Berdasarkan Tabel yang sudah disajikan diatas dapat diketahui bahwa model *clustering K-Means* dengan reduksi dimensi PCA dan nilai K optimal dari *Gap Statistic* memperoleh nilai evaluasi terbaik dengan menggunakan jumlah *cluster* adalah 2. Sehingga berdasarkan hasil penelitian ini, penerapan PCA untuk mereduksi dimensi dan *Gap Statistic* untuk menentukan nilai K optimal dari Dataset hasil PCA dapat membantu menangani masalah multikolinieritas data dan penentuan nilai K optimal dari metode *cluster*ing K-Means.

5 Kesimpulan

Berdasarkan penelitian yang telah dilakukan, hasil evaluasi *clustering* kanker payudara menggunakan algoritma *K-Means* mendapatkan nilai DBI sebesar 1.298284 pada kluster 2, 1.479132 pada kluster 3, 1.803953 pada kluster 4, 1.776404 pada kluster 5, 1.808281 pada kluster 6, 1.714904 pada kluster 7, dan 1.690148 pada kluster 8. Selanjutnya hasil evaluasi *clustering* kanker payudara menggunakan algoritma *K-Means* dengan PCA mendapatkan nilai DBI sebesar 1.195513 pada kluster 2, 1.341205 pada kluster 3, 1.616263 pada kluster 4, 1.592884 pada kluster 5, 1.594894 pada kluster 6, 1.514159 pada kluster 7, dan 1.461446 pada kluster 8. Dan hasil evaluasi *clustering* kanker payudara menggunakan algoritma *K-Means* dengan PCA dan *Gap Statistic* mendapatkan nilai DBI sebesar 1.195513 pada kluster 2. Dari penelitian ini dapat ditarik kesimpulan bahwa pemodelan *clustering K-Means* dengan reduksi dimensi PCA dan nilai K optimal dari *Gap Statistic* memberikan hasil evaluasi terbaik, ditunjukkan oleh nilai *Davies Bouldin Index* (DBI) yang paling kecil pada *cluster* 2. Sehingga model ini dapat membantu menangani masalah multikolinieritas data dan mengatasi kelemahan metode *cluster*ing K-Means dalam menentukan nilai K optimal.

Referensi

[1] D. Cahyanti, A. Rahmayani, and S. Ainy Husniar, "Indonesian Journal of Data and Science Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara," *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 39–43, 2020.

[2] Agung Winasis and Ratna Djuwita, "Obesitas dan Kanker Payudara : Literature Review," *Media Publikasi Promosi Kesehatan Indonesia (MPPKI)*, vol. 6, no. 8, pp. 1501–1508, Aug. 2023, doi: 10.56338/mppki.v6i8.3501.

[3] E. Susilowati, A. T. Hapsari, M. Efendi, and P. E. Kresnha, "Diagnosa Penyakit Kanker Payudara Menggunakan Metode K-Means Clustering," *JUST IT: Jurnal Sistem Informasi, Teknologi Informasi dan Komputer*, vol. 10, no. 1, pp. 27–32, 2019, [Online]. Available: https://jurnal.umj.ac.id/index.php/just-it

[4] A. Almayda and S. Saepudin, "Penerapan Data Mining K-Means Clustering Untuk Mengelompokkan Berbagai Jenis Merk Smartphone," in *SISMATIK (Seminar Nasional Sistem Informasi dan Manajemen Informatika)* , 2021.

[5] M. Nishom and M. Y. Fathoni, "Implementasi Pendekatan Rule-Of-Thumb untuk Optimasi Algoritma K-Means Clustering," vol. 03, no. 02, 2018.

[6] S. Ika Murpratiwi, I. Gusti Agung Indrawan, and A. Aranta, "Analisis Pemilihan Cluster Optimal Dalam Segmentasi Pelanggan Toko Reatil," *Jurnal Pendidikan Teknologi dan Kejuruan*, vol. 18, no. 2, 2021.

[7] R. Silvi, "Analisis Cluster dengan Data Outlier Menggunakan Centroid Linkage dan K-Means Clustering untuk Pengelompokkan Indikator HIV/AIDS di Indonesia," *Jurnal Matematika "MANTIK,"* vol. 4, no. 1, pp. 22–31, May 2018, doi: 10.15642/mantik.2018.4.1.22-31.

[8] M. A. Nahdliyah, T. Widiharih, and A. Prahutama, "Metode K-Medoids Clustering Dengan Validasi Silhoutte Index dan C-Index (Studi Kasus Jumlah Kriminalitas Kabupaten/Kota di Jawa Tengah Tahun 2018)," *JURNAL GAUSSIAN*, vol. 8, no. 2, pp. 161–170, 2019, [Online]. Available: http://ejournal3.undip.ac.id/index.php/gaussian

[9] P. N. Safitri, R. Aristawidya, and S. B. Faradilla, "Klasterisasi Faktor-Faktor Kemiskinan Di Provinsi Jawa Barat Menggunakan K-Medoids Clustering," *Journal of Mathematics Education and Science*, vol. 4, no. 2, pp. 75–80, Oct. 2021, doi: 10.32665/james.v4i2.242.

 ISSN:2302-8149
e-ISSN:2540-9719

[10] A. N. Azizah, T. Widiharih, A. R. Hakim, D. Statistika, F. Sains, and D. Matematika, "Kernel K-Means Clustering Untuk Pengelompokan Sungai di Kota Semarang Berdasarkan Faktor Pencemaran Air," *Jurnal Gaussian*, vol. 11, no. 2, pp. 228–236, 2022, [Online]. Available: https://ejournal3.undip.ac.id/index.php/gaussian/

[11] N. Thamrin and A. W. Wijayanto, "Comparison of Soft and Hard Clustering: A Case Study on Welfare Level in Cities on Java Island," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 1, pp. 141–160, Mar. 2021, doi: 10.29244/ijsa.v5i1p141-160.

[12] A. M. El-Mandouh, H. A. Mahmoud, L. A. Abd-Elmegid, and M. H. Haggag, "Optimized K-Means Clustering Model based on Gap Statistic," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 10, no. 1, 2019, [Online]. Available: www.ijacsa.thesai.org

[13] C. A. Sugianto, A. H. Rahayu, and A. Gusman, "Algoritma K-Means Untuk Pengelompokkan Penyakit Pasien Pada Puskesmas Cigugur Tengah," *JOINT (Journal of Information Technology)*, vol. 02, no. 02, pp. 39–44, 2020.

[14] E. Muningsih, N. Hasan, and G. B. Sulistyo, "Bianglala Informatika Penerapan Metode Principle Component Analysis (PCA) untuk Clustering Data Kunjungan Wisatawan Mancanegara ke Indonesia," *Bianglala Informatika*, vol. 8, no. 2, pp. 58–62, 2020, [Online]. Available: www.bps.go.id

[15] Pendi, "Analisis Regresi Dengan Metode Komponen Utama Dalam Mengatasi Masalah Multikolinieritas," *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)*, vol. 10, no. 1, pp. 131–138, 2021.

[16] S. Naveen, N. V. Kashyap, V. P. Kulkarni, A. Sandeep, and M. S. Chakradhar, "Breast Cancer Prediction Using Unsupervised Learning Technique K-Means Clustering Algorithm," in *ViTECoN 2023 - 2nd IEEE International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ViTECoN58111.2023.10157765.

[17] Taufik Hidayat, Mohamad Jajuli, and Susilawati, "Clustering daerah rawan stunting di Jawa Barat menggunakan algoritma K-Means," *INFOTECH : Jurnal Informatika & Teknologi*, vol. 4, no. 2, pp. 137–146, Dec. 2023, doi: 10.37373/infotech.v4i2.642.

[18] R. N. Puspita, "Analisis K-Means Cluster Pada Kabupaten/Kota Di Provinsi Banten Berdasarkan Indikator Indeks Pembangunan Manusia," *Jurnal Ilmiah Pendidikan Matematika, Matematika dan Statistika*, vol. 2, no. 3, 2021, doi: 10.46306/lb.v2i3.

[19] N. Shahadah Qur'ani and A. W. Wijayanto, "Implementasi K-Means dan Hierarchical Clustering Pada PenentuanTingkatan Smart City Tahun 2022 Berdasarkan Motion Index," 2023. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[20] T. Zulyanti, "Perbandingan Pengelompokan Usaha Mikro Kecil Dan Menengah Di Kabupaten Klaten Tahun 2019 Dengan Metode K-Means Dan Clustering Large Application," *Jurnal Statistika Industri dan Komputasi*, vol. 7, no. 1, pp. 46–59, 2022.