



**IMPLEMENTASI *PRINCIPAL COMPONENT ANALYSIS* (PCA) DAN
GAP STATISTIC UNTUK *CLUSTERING* KANKER PAYUDARA PADA
ALGORITMA *K-MEANS***

Skripsi

**Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Sarjana Strata-1 Ilmu Komputer**

Oleh

RIDHA AFIFA

NIM 1911016320015

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
JANUARI 2024**



IMPLEMENTASI *PRINCIPAL COMPONENT ANALYSIS* (PCA) DAN *GAP STATISTIC* UNTUK *CLUSTERING* KANKER PAYUDARA PADA ALGORITMA *K-MEANS*

Skripsi

**Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Sarjana Strata-1 Ilmu Komputer**

Oleh

RIDHA AFIFA

NIM 1911016320015

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
JANUARI 2024**

SKRIPSI

IMPLEMENTASI *PRINCIPAL COMPONENT ANALYSIS* (PCA) DAN *GAP STATISTIC* UNTUK *CLUSTERING* KANKER PAYUDARA PADA ALGORITMA *K-MEANS*

Oleh:
RIDHA AFIFA
NIM 1911016320015

Telah dipertahankan di depan Dosen Penguji pada Tanggal 4 Januari 2024
Susunan Dosen Penguji:

Pembimbing I



M. Itqan Mazdadi, S.Kom., M.Kom.
NIP. 199006122019031013

Dosen Penguji I



Fatma Indriani, S.T., M.I.T., Ph.D.
NIP. 198404202008122004

Pembimbing II



Triando H. Saragih, S.Kom., M.Kom.
NIP. 199308242019031012

Dosen Penguji II



Muliadi, S.Kom., M.Cs.
NIP. 197804222010121002



Banjarbaru, 4 Januari 2024
Koordinator Program Studi Ilmu Komputer

Iwan Budiman, S.T., M.Kom.
NIP. 19770325 2008121001

PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diberikan orang lain, kecuali yang secara tertulis diacu di dalam naskah ini dan disebutkan dalam Daftar Pustaka

Banjarbaru, 4 Januari 2024



RIDHA AFIFA
NIM. 1911016320015

ABSTRAK

IMPLEMENTASI *PRINCIPAL COMPONENT ANALYSIS* (PCA) DAN *GAP STATISTIC* UNTUK *CLUSTERING* KANKER PAYUDARA PADA ALGORITMA *K-MEANS*

(Oleh: Ridha Afifa; Pembimbing: Muhammad Itqan Mazdadi, S.Kom., M.Kom. dan Triando Hamonangan Saragih, S.Kom., M.Kom.; 2024; 79 halaman)

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

IMPLEMENTATION OF PRINCIPAL COMPONENT ANALYSIS (PCA) AND GAP STATISTIC FOR BREAST CANCER CLUSTERING IN THE K-MEANS ALGORITHM (By: Ridha Afifa; Supervisor: Muhammad Itqan Mazdadi, S.Kom., M.Kom. and Triando Hamonangan Saragih, S.Kom., M.Kom.; 2024; 79 pages)

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

PRAKATA

Assalamualaikum Warahmatullahi Wabarakatuh.

Puji syukur penulis panjatkan ke hadirat Allah SWT karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul “Implementasi *Principal Component Analysis (PCA)* Dan *Gap Statistic* Untuk *Clustering* Kanker Payudara Pada Algoritma *K-Means*” untuk memenuhi syarat dalam menyelesaikan pendidikan program S-1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat. Tak lupa pula penulis panjatkan shalawat dan salam ke hadirat Rasulullah Muhammad SAW beserta para sahabat, keluarga, dan pengikut beliau hingga yaumul qiyamah.

Pada lembar ini penulis ingin menyampaikan ucapan terima kasih kepada pihak-pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan skripsi ini, adapun yang dimaksud adalah sebagai berikut :

1. Keluarga terutama orang tua yang selalu memberikan bantuan, semangat, doa dan dukungan dalam proses penyelesaian skripsi ini.
2. Bapak Muhammad Itqan Mazdadi S.Kom., M.Kom. selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran penyelesaian skripsi ini.
3. Bapak Triando Hamonangan Saragih, S.Kom., M.Kom. selaku dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
4. Seluruh Dosen dan staf Program Studi Ilmu Komputer FMIPA ULM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.
5. Teman-teman keluarga Ilmu Komputer angkatan 2019, terima kasih untuk canda, tawa, perjuangan yang sudah dilewati bersama, untuk semua kenangan manis yang telah terukir selama ini. Senang bisa menjadi salah satu bagian dari kehidupan kalian.
6. Ucapan terimakasih kepada Rabiatus Nisa, Zaina Fadia, Hartati, Nurmolika, dan Maimunah yang telah kebersamai melewati lika liku perkuliahan serta memberikan dukungan dalam proses pengerjaan skripsi.

7. Semua pihak yang tidak dapat disebutkan satu persatu yang telah turut membantu dalam penyelesaian skripsi ini.

Penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan, maka saran dan kritik dari semua pihak sangat diharapkan demi penyempurnaan pada penelitian selanjutnya. Diharapkan skripsi ini dapat bermanfaat bagi semua pihak dan dapat menambah pengetahuan kita semua.

Banjarbaru, 4 Januari 2024



Penulis

DAFTAR ISI

SKRIPSI.....	i
PERNYATAAN.....	ii
ABSTRAK	iii
<i>ABSTRACT</i>	iv
PRAKATA	v
DAFTAR ISI.....	vii
DAFTAR TABEL.....	ix
DAFTAR GAMBAR	xii
DAFTAR LAMPIRAN.....	xiii
BAB I	1
PENDAHULUAN	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	3
1.3 Batasan Masalah	3
1.4 Tujuan Penelitian	4
1.5 Manfaat Penelitian.....	4
BAB II.....	5
TINJAUAN PUSTAKA	5
2.1 Kajian Terdahulu	5
2.2 Keaslian Penelitian	6
2.3 Kanker Payudara.....	8
2.4 <i>Data Mining</i>	9
2.5 Multikolinieritas	11
2.6 <i>Outlier</i>	11
2.7 Standarisasi Data	12
2.8 <i>Principal Component Analysis (PCA)</i>	12
2.9 <i>Gap Statistic</i>	17
2.10 <i>Clustering</i>	18
2.11 <i>K-Means</i>	19

2.12 <i>Davies Boudin Index</i>	20
BAB III	22
METODE PENELITIAN.....	22
3.1 Alat Penelitian	22
3.2 Bahan Penelitian	22
3.3 Variabel Penelitian.....	22
3.4 Prosedur Penelitian	23
BAB IV	26
HASIL DAN PEMBAHASAN.....	26
4.1 Hasil.....	26
4.1.1 Pengumpulan Data	26
4.1.2 <i>Preprocessing Data</i>	28
4.1.2.1 Penanganan <i>Outlier</i>	28
4.1.2.2 Standarisasi <i>Z-Score</i>	30
4.1.2.3 Principal Component Analysis.....	31
4.1.2.3.1 Uji Bartlett, KMO, MSA	31
4.1.3.2 Perhitungan <i>Covariance</i>	33
4.1.4 <i>Gap Statistic</i>	36
4.1.5 <i>Clustering K-Means</i>	39
4.1.5.1 Model <i>Clustering K-Means</i>	39
4.1.5.2 Model <i>Clustering PCA-K-Means</i>	51
4.1.5.3 Model <i>Clustering PCA-GapStatistic-K-Means</i>	64
4.1.6 <i>Davies Bouldin Index</i>	68
4.2 Pembahasan	69
BAB V.....	75
PENUTUP.....	75
5.1 Kesimpulan.....	75
5.2 Saran	75
DAFTAR PUSTAKA	76
LAMPIRAN	
RIWAYAT HIDUP	

DAFTAR TABEL

Tabel	Halaman
Tabel 1. Keaslian Penelitian.....	6
Tabel 2. Rancangan Penelitian.....	8
Tabel 3. Kriteria Keputusan	15
Tabel 4. Sampel dataset.....	22
Tabel 5. Deskripsi atribut dataset.....	26
Tabel 6. Contoh Data Hasil <i>Winsorizing</i>	30
Tabel 7. Contoh Data Hasil Standarisasi <i>Z-Score</i>	30
Tabel 8. Nilai Korelasi antar Variabel	31
Tabel 9. Hasil Uji KMO dan MSA	32
Tabel 10. Hasil Uji <i>Bartlett</i>	32
Tabel 11. Hasil Perhitungan <i>Covariance</i>	33
Tabel 12. Hasil Perhitungan <i>Eigen Value</i>	33
Tabel 13. Hasil Perhitungan Eigen Vektor	35
Tabel 14. Hasil Perhitungan Nilai <i>Principal Component</i>	35
Tabel 15. Contoh Data Hasil PCA	36
Tabel 16. Hasil Logaritma <i>Within-Cluster Sum of Squares</i>	36
Tabel 17. Hasil Rata-rata Logaritma <i>Within-Cluster Sum of Squares</i>	37
Tabel 18. Hasil Perhitungan <i>Gap Statistic</i>	38
Tabel 19. Contoh Penentuan Centroid Awal K=2	39
Tabel 20. Contoh Perhitungan <i>Euclidean Distance</i>	40
Tabel 21. Penetapan jarak <i>cluster</i> ke titik pada data iterasi 0	40
Tabel 22. Contoh perhitungan <i>centroid</i> baru	40
Tabel 23. Hasil centroid terakhir pada iterasi 1	41
Tabel 24. Hasil centroid terakhir K=2	41
Tabel 25. Hasil <i>cluster</i> terbentuk pada K=2	42
Tabel 26. Hasil <i>centroid</i> terakhir K=3	42
Tabel 27. Hasil <i>cluster</i> terbentuk pada K=3	43
Tabel 28. Hasil <i>centroid</i> akhir K=4	44

Tabel 29. Hasil <i>cluster</i> terbentuk pada K=4	44
Tabel 30. Hasil <i>centroid</i> akhir K=5	45
Tabel 31. Hasil <i>cluster</i> terbentuk pada K=5	45
Tabel 32. Hasil <i>centroid</i> akhir K=6	46
Tabel 33. Hasil <i>cluster</i> terbentuk pada K=6	47
Tabel 34. Hasil <i>centroid</i> akhir K=7	48
Tabel 35. Hasil <i>cluster</i> terbentuk pada K=7	48
Tabel 36. Hasil <i>centroid</i> akhir K=8	49
Tabel 37. Hasil <i>cluster</i> terbentuk pada K=8	49
Tabel 38. Jumlah data yang di- <i>cluster</i>	50
Tabel 39. Contoh Penentuan <i>Centroid</i> Awal K=2	51
Tabel 40. Contoh Perhitungan <i>Euclidean Distance</i>	52
Tabel 41. Penetapan jarak <i>cluster</i> ke titik pada data iterasi 0	52
Tabel 42. Contoh perhitungan <i>centroid</i> baru	52
Tabel 43. Hasil <i>centroid</i> terakhir pada iterasi 1	53
Tabel 44. Hasil <i>centroid</i> terakhir K=2	53
Tabel 45. Hasil <i>cluster</i> terbentuk pada K=2	54
Tabel 46. Hasil <i>centroid</i> terakhir K=3	54
Tabel 47. Hasil <i>cluster</i> terbentuk pada K=3	55
Tabel 48. Hasil <i>centroid</i> terakhir K=4	56
Tabel 49. Hasil <i>cluster</i> terbentuk pada K=4	56
Tabel 50. Hasil <i>centroid</i> terakhir K=5	57
Tabel 51. Hasil <i>cluster</i> terbentuk pada K=5	58
Tabel 52. Hasil <i>centroid</i> terakhir K=6	59
Tabel 53. Hasil <i>cluster</i> terbentuk pada K=6	59
Tabel 54. Hasil <i>centroid</i> terakhir K=7	60
Tabel 55. Hasil <i>cluster</i> terbentuk pada K=7	61
Tabel 56. Hasil <i>centroid</i> terakhir K=8	62
Tabel 57. Hasil <i>cluster</i> terbentuk pada K=8	63
Tabel 58. Jumlah data yang di- <i>cluster</i>	63
Tabel 59. Contoh Penentuan <i>Centroid</i> Awal K=2	64

Tabel 60. Contoh Perhitungan <i>Euclidean Distance</i>	65
Tabel 61. Penetapan jarak <i>cluster</i> ke titik pada data iterasi 0	65
Tabel 62. Contoh perhitungan <i>centroid</i> baru	66
Tabel 63. Hasil <i>centroid</i> terakhir pada iterasi 1	66
Tabel 64. Hasil <i>centroid</i> terakhir K=2	66
Tabel 65. Hasil <i>cluster</i> terbentuk pada K=2	67
Tabel 66. Jumlah data yang di- <i>cluster</i>	67
Tabel 67. Hasil DBI K-Means	68
Tabel 68. Hasil DBI PCA + K-Means	68
Tabel 69. Hasil DBI PCA + Gap Statistic + K-Means.....	68
Tabel 70. Perbandingan Hasil Evaluasi DBI	71
Tabel 71. Perbandingan Hasil Cluster Terbentuk	72
Tabel 72. Hasil Kemiripan <i>Clustering PCA-KMeans</i> dengan Data WDBC.....	73

DAFTAR GAMBAR

Gambar	Halaman
Gambar 1. Tahap-tahap <i>data mining</i>	9
Gambar 2. Alur Penelitian.....	23
Gambar 3. <i>Outlier</i> pada data	29
Gambar 4. Data setelah dilakukan penanganan <i>outlier</i>	29
Gambar 5. Plot dari Nilai Eigen Value	34
Gambar 6. Grafik Nilai <i>Gap Statistic</i>	38
Gambar 7. Hasil visualisasi K=2.....	42
Gambar 8. Hasil visualisasi K=3.....	43
Gambar 9. Hasil visualisasi K=4.....	45
Gambar 10. Hasil visualisasi K=5.....	46
Gambar 11. Hasil visualisasi K=6.....	47
Gambar 12. Hasil visualisasi K=7.....	49
Gambar 13. Hasil visualisasi K=8.....	50
Gambar 14. Hasil visualisasi K=2.....	54
Gambar 15. Hasil visualisasi K=3.....	56
Gambar 16. Hasil visualisasi K=4.....	57
Gambar 17. Hasil visualisasi K=5.....	59
Gambar 18. Hasil visualisasi K=6.....	60
Gambar 19. Hasil visualisasi K=7.....	62
Gambar 20. Hasil visualisasi K=8.....	63
Gambar 21. Hasil visualisasi K=2.....	67
Gambar 22. Grafik Perbandingan Hasil Evaluasi DBI	71

DAFTAR LAMPIRAN

Lampiran 1. Source Code Import Library	81
Lampiran 2. Source Code Import Dataset.....	81
Lampiran 3. Source Code Memeriksa Outlier Dengan Boxplot	81
Lampiran 4. Source Code Penanganan Outlier Menggunakan Winsorizing	81
Lampiran 5. Source Code Standarisasi Data dengan Min-Max Scaler	81
Lampiran 6. Source Code Mengecek Multikolinieritas	81
Lampiran 7. Souce Code Uji Asumsi Analisis Faktor (Uji KMO dan MSA).....	82
Lampiran 8. Souce Code Uji Asumsi Analisis Faktor (Uji Bartlett)	82
Lampiran 9. Source Code Reduksi Dimensi Principal Component Analysis (PCA) 82	
Lampiran 10. Source Code Mencari Nilai K Optimal menggunakan Gap Statistic ..	83
Lampiran 11. Source Code Model Clustering K-Means.....	83
Lampiran 12. Source Code Model Clustering PCA + K-Means.....	83