JURNAL FOURIER | Oktober 2023, Vol. 12, No. 2, 79-85 ISSN 2252-763X
DOI: 10.14421/fourier.2023.122.79-85 E-ISSN 2541-5239

Klasifikasi Penyakit Kanker Payudara menggunakan Metode K-Nearest Neighbors (KNN)

Micha Annata Shinami[1], Saiful Bahri[2]

[12]Program Studi Matematika, Fakultas Sains dan Matematika, Universitas Islam Negeri Sunan Ampel Surabaya, Jl. Dr. Ir. H. Soekarno No. 682, Gn. Anyar, Surabaya, Indonesia

Korespondensi; Micha Annata Shinami, Email: michaannatash@gmail.com

Abstrak

Kanker payudara adalah jenis tumor ganas yang tumbuh di sel-sel payudara. Kanker payudara dibagi menjadi dua tingkatan yaitu Tingkat jinak dan ganas, pada tingkatan jinak benjolan yang ada di payudara masih bisa mengecil dan hilang, tetapi pada tingkatan ganas sudah susah untuk diobati bahkan bisa menyebabkan kematian. Penelitian ini menggunakan metode Klasifikasi K-Nearest Neighbors (KNN) dengan menghasilkan nilai k = 8 dengan akurasi sebesar 77% dan terdapat 62 sampel data yang termasuk dalam kelas positif dan prediksinya benar. Berdasarkan Confusion Matrix, pengklasifikasian kanker payudara ini mendapatkan nilai akurasi sebesar 77%, nilai presisi sebesar 76% dan nilai recall sebesar 71%.

Kata Kunci: Kanker Payudara, Klasifikasi, *K-Nearest Neighbors, Confusion Matrix.*

Abstract

Breast cancer is a type of malignant tumor that grows in breast cells. Breast cancer is divided into two levels, namely benign and malignant levels. At the benign level, lumps in the breast can still shrink and disappear, but at the malignant level it is difficult to treat and can even cause death. This research uses the K-Nearest Neighbors (KNN) classification method which produces a value of k = 8 with an accuracy of 77% and there are 62 data samples which are included in the positive class and the predictions are correct. Based on the Confusion Matrix, this breast cancer classification has an accuracy value of 77%, a precision value of 76% and a recall value of 71%.

Keywords: Breast Cancer, K-Nearest Neighbors, Classification, Confusion Matrix.

Pendahuluan

Kanker payudara adalah termasuk salah satu penyakit yang sering dialami oleh perempuan. Kanker ini dapat tumbuh bila terdapat pertumbuhan sel yang tidak normal pada payudara, sel tersebut membelah lebih cepat dibandingkan sel normal dan kemudian sel tersebut membentuk benjolan.[1] Pada kanker payudara stadium lanjut, sel-sel abnormal ini menyebar melalui kelenjar getah bening dan kemudian ke organ lain di tubuh. Kanker payudara terbagi menjadi dua tingkatan, yaitu kanker payudara jinak dan ganas yang tidak menyebar atau merusak jaringan di sekitarnya, dan benjolan dapat mengecil dan hilang dengan sendirinya. Kanker payudara tingkat ganas ditandai dengan benjolan yang ada disekitar payudara jika ditekan maka akan terasa nyeri yang berlebihan dan mudah menyebar dan merusak jaringan dan organ lain yang ada di dekatnya[2].

Kanker payudara adalah penyakit kanker nomor satu dan juga penyebab utamanya yaitu kematian akibat kanker di seluruh dunia setiap tahunnya. Menurut WHO (2020), prevalensi kanker payudara sebanyak 2.261.419 kasus, dengan kasus kanker terbanyak yang menyerang perempuan. Gejala kanker

payudara termasuk benjolan di payudara, keluarnya cairan berdarah dari puting, dan perubahan bentuk atau tekstur puting atau payudara[3].

Deteksi dini mengenai kanker payudara sangat penting dilakukan, yaitu melalui berbagai pemeriksaan seperti biopsi payudara, termografi payudara, mamografi, duktografi, dan USG payudara (USG). Termografi payudara adalah prosedur diagnostik berdasarkan tingkat kimia payudara dan aktivitas pembuluh darah untuk mendeteksi sel kanker payudara pada tahap awal. Termografi payudara sangat sensitif dalam menggambarkan perubahan suhu dan pembuluh darah yang mengindikasikan sel-sel abnormal pada payudara, namun jika terdapat tumor, termografi payudara tidak dapat menunjukkan lokasi tumor, jadi sebaiknya dilakukan pada waktu yang bersamaan. waktu dengan mamografi, agar hasil pemeriksaannya lengkap. Mamograf*i* adalah metode pemeriksaan payudara dengan menggunakan sinar x yang memiliki kadar rendah dan secara umumnya lebih diperbolehkan pada perempuan yang telah berumus diatas empat puluh tahun[4].

Penelitian sebelumnya yaitu mengenai klasifikasi penyakit kulit dengan menggunakan algoritma *K-Nearest Neighbors* (KNN) dan diperoleh nilai akurasi yang sebesar 65%[5], juga penelitian yang dilakukan oleh Permana Putra, Akim M H Pardede, dan Siswan Syahputra mengenai analisis metode *K-Nearest* Neighbors (KNN) dalam klasifikasi data iris bunga[6]. Dalam penelitian ini algoritma yang digunakan adalah algoritma klasifikasi *K-Nearest Neighbors* (KNN).

Landasan Teori

Berikut ini beberapa penjelasan yang berkaitan dengan kanker payudara dan metode yang diperlukan untuk melakukan klasifikasi pada penyakit kanker payudara menggunakan metode K-Nearest Neighbors (KNN)

1. Kanker Payudara

Kanker payudara adalah penyakit yang sering dialami oleh kaum perempuan dengan tidak memandang usia yang bisa menyerang anak kecil, dewasa dan usia tua. Penyakit kanker payudara harus dengan segera diobati, jika sudah terkena kanker payudara yang tingkat ganas maka kemungkinan besar akan berakibat buruk bagi penderitamya yang bisa menyebabkan kematian sehingga pemeriksaan dan pengobatan harus dilakukan sejak dini karena penting sekali bagi keselamatan penderita kanker payudara[7].

2. Data Mining

Data mining adalah proses dalam menemukan pola yang menarik dari sejumlah data besar yang tersimpan. Data mining ini berhubungan dengan analisa data dan penggunaan perangkat lunak untuk mencari pola dan persamaan dalam data dengan mengekstrasi pola yang sebelumnya tidak jelas[8].

3. Klasifikasi

Klasifikasi adalah suatu metode dalam penentuan anggota di suatu kelas yang telah ditentukan sebelumnya. Anggota pada kelas tersebut didasarkan pada persamaan karakter dari data yang ada. Pada saat pengklasifikasian, dataset dibagi menjadi dua yaitu data training dan data testing yang Dimana keduanya berperan dalam menentukan hasil akurasi yang tepat dari penerapan metode klasifikasi yang digunakan[9].

4. Normalisasi Data

Normalisasi data adalah tahap dimana data dikategorikan ke dalam skala tertentu. Dalam proses normalisasi data ini bisa menggunakan nilai *Z-Score* yang terdiri anatara angka positif dan negatif tidak terbatas[10]. Berikut adalah rumusnya[11]:

$$Z = \frac{X_i - \mu}{\sigma}$$

Keterangan:
X_i : nilai data yang ke-i
μ : nilai rata-rata
σ : varians

5. Algoritma K-Nearest Neighbor

Algoritma K-Nearest Neighbour (KNN) adalah metode klasifikasi untuk menentukan target baru berdasarkan jarak terdekat antara data training dengan data testing. Algoritma KNN ini banyak digunakan dalam klasifikasi karena mudah diimplementasikan dan memberikan akurasi yang baik. Algoritma KNN bekerja dengan menghitung jarak Euclidean tiap sampel yang berbeda dari kelas-kelas yang sudah ditentukan sebelumnya dan kemudian memilih tetangga terdekat dari setiap kategori lalu akan diberikan sampel kategori berdasarkan k yang terdekat[12]. Berikut merupakan rumus untuk menentukan nilai k dan rumus untuk menghitung jarak Euclidean:

a. Rumus menentukan nilai k[13]:

$$k = \sqrt{N}$$

b. Rumus untuk menghitung jarak Euclidean:

$$D(x_1, x_2) = \sqrt{\sum_{i=1}^{n}(x_1 - x_2)^2}$$

6. Confusion Matrix

Confusion matrix adalah metode yang digunakan untuk menggambarkan hasil suatu model klasifikasi dengan cara membandingkan hasil prediksi model dengan nilai yang sebenarnya dari data uji. Dalam penelitian ini, Confusion matrix digunakan untuk mengevaluasi kinerja dari metode klasifikasi. Confusion Matrix dapat membantu memahami kinerja dari model KNN secara detail dan dapat memperoleh hasil klasifikasi dengan benar[14]. Berikut adalah table Confusion Matrix

Tabel 1. Confusion Matrix

Classification	Classification	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Keterangan:
a. *True Positive* (TP): data yang diklasifikasi dengan benar sebagai nilai positif.
b. *True Negative* (TN): data yang diklasifikasi dengan benar sebagai nilai negatif.
c. *False Positive* (FP): data yang diklasifikasi salah dengan nilai positif
d. *False Negative* (FN): data yang diklasifikasi salah dengan nilai negatif

Kemudian untuk mengevaluasi dan mengukur kinerja hasil dari klasifikasi dilakukan dengan cara menghitung *accuracy*, *precision* dan *recall*. Rumusnya yaitu sebagai berikut[15]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + TN}$$

Bahan dan Metode

Metode yang digunakan dalam penelitian ini adalah metode Klasifikasi *K-Neares Neighbors* (KNN). Penelitian ini mengambil data dari Kaggle https://www.kaggle.com/code/laurencens/knn-breast-cancer-kanker-payudara. Data yang diperoleh tersebut akan dianalisis menggunakan Klasifikasi *K-Neares Neighbors* (KNN) dengan variabel yang digunakan adalah sebagai berikut:

Tabel 2. Data Kanker Payudara

Pid	Age	Meno	Size	Grade	Nodes	Pgr	Er	Hormon	Rfstime	Status
132	49	0	18	2	2	0	0	0	1838	0
1575	55	1	20	3	16	0	0	0	403	1
1140	56	1	40	3	3	0	0	0	1603	0
…	…	…	…	…	…	…	…	…	…	…
586	51	0	30	3	2	1152	38	1	1760	0
1273	64	1	26	2	2	1356	1144	1	1152	0
1525	57	1	35	3	1	1490	209	1	1342	0
736	44	0	21	2	3	1600	70	0	629	0
894	80	1	7	2	7	2380	972	1	758	0

Riset ini dilakukan dengan langkah-langkah sebagai berikut. Langkah pertama yang harus dilakukan yaitu input data yang ada di *Microsoft Excel*. Langkah kedua yaitu melakukan memproses data yang sudah diinput. Langkah ketiga yaitu menormalisasikan data. Langkah keempat yaitu membagi data menjadi 2 bagian yaitu data *training* dan data *testing*. Langkah kelima yaitu menerapkan Algoritma KNN. Langkah keenam yaitu menentukan nilai k, dan langkah ke tujuh adalah membuat *Confusion Matrix* agar mempermudah dalam menentukan nilai akurasi.

Hasil dan Pembahasan

Penelitian ini diawali dengan mengolah data kanker payudara yang jumlahnya sebanyak 686 data. Dalam data ini terdapat 10 variabel independent dan 1 variabel dependent. Berikuta adalah sampel datanya:

Tabel 3. Data Kanker Payudara

Pid	Age	Meno	Size	Grade	Nodes	Pgr	Er	Hormon	Rfstime	Status
132	49	0	18	2	2	0	0	0	1838	0
1575	55	1	20	3	16	0	0	0	403	1
1140	56	1	40	3	3	0	0	0	1603	0
…	…	…	…	…	…	…	…	…	…	…
1525	57	1	35	3	1	1490	209	1	1342	0
736	44	0	21	2	3	1600	70	0	629	0
894	80	1	7	2	7	2380	972	1	758	0

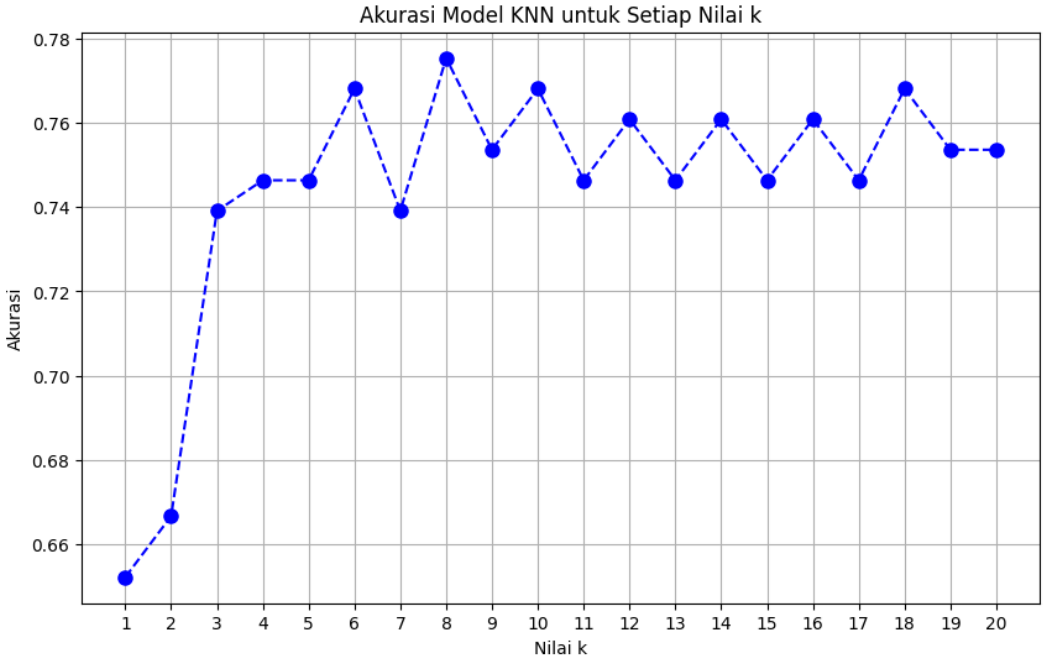
Kemudian dilakukannya normalisasi data menggunakan Min – Max dengan rentang nilai -1 sampai dengan 1 agar rentang jarak antar data tidak terlalu jauh.

Tabel 4. Hasil Normalisasi Data

Pid	Age	Meno	Size	Grade	Nodes	Pgr	Er	Hormon	Rfstime
0.07	0.47	0.0	0.12	0.5	0.02	0.00	0.00	0.0	0.69
0.86	0.57	1.0	0.14	1.0	0.30	0.00	0.00	0.0	0.14

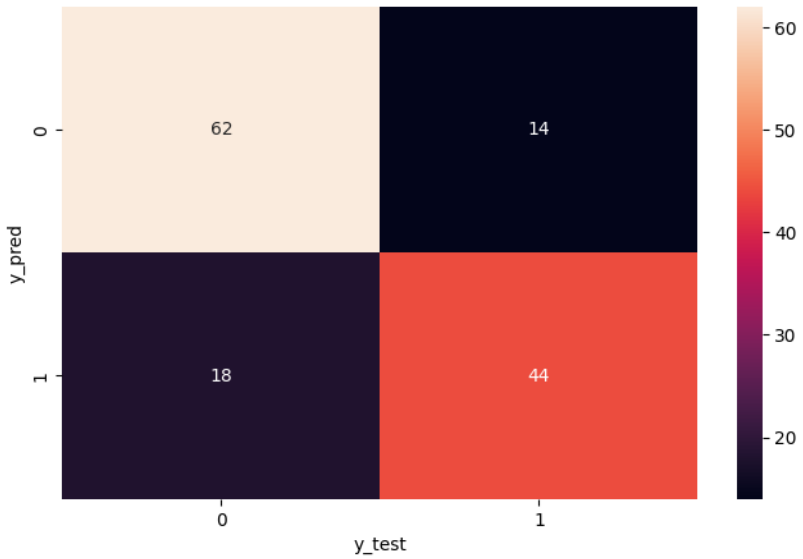
0.63	0.59	1.0	0.31	1.0	0.04	0.00	0.00	0.0	0.60
…	…	…	…	…	…	…	…	…	…
0.83	0.61	1.0	0.27	1.0	0.00	0.62	0.18	1.0	0.50
0.40	0.38	0.0	0.15	0.5	0.04	0.67	0.06	0.0	0.23
0.49	1.00	1.0	0.03	0.5	0.12	1.00	0.84	1.0	0.28

Setelah data di normalisasikan, lanjut ke tahap selanjutnya yaitu membagi data menjadi 2 bagian, yaitu data training dan data testing. Rasio perbandingannya yaitu 80:20 atau 80% data training dan 20% data testing. Lalu mencari nilai k tetangga terdekat yaitu antara k = 4, k = 6 dan k = 8.



Gambar 1. Akurasi Model KNN untuk setiap nilai k = 1 sampai k = 20

Berdasarkan Gambar 1. yang merupakan grafik akurasi diatas dapat dilihat bahwa akurasi k = 4 nilainya 77% tertinggi daripad yang lainnya. Selanjutnya gambar *Confusion Matrix* digunakan untuk mengetahui baik atau buruk sebuah pengklasifikasian yang dilakukan.



Gambar 2. *Confusion Matrix* Hasil Klasifikasi Kanker Payudara

Tabel 5. Hasil Akurasi, Presisi, dan Recall

	precision	*recall*	*fl-score*	*Support*
0	0.78	0.82	0.79	76
1	0.76	0.71	0.73	62
accuracy			0.77	13B
macro avg	0.77	0.76	0.76	13B
Wighted avg	0.77	0.77	0.77	13B

Berdasarka Gambar 2 yang menunjukka evaluasi model Confusion Matrix dan Tabel 5 diatas, didapatkan nilai akurasi sebesar 77%, nilai presisi sebesar 76% dan nilai recall sebesar 71%. Kemudia dapat disimpulkan pula bahwa terdapat 62 sampel data yang termasuk dalam kelas positif dan prediksinya benar.

Kesimpulan

Dalam melakukan klasifikasi penyakit kanker payudara yang diterapkan menggunakan Algoritma *K-Nearest Neighbor (KNN)* dengan menggunakan 10 variabel independent dan 1 variabel dependent dengan jumlah data sebanyak 686 data. Dibutuhkan normalisasi data agar jaraknya tidak terlalu jauh lalu membagi data menjadi data training dan data testing dengan 80% data training dan 20% data testing. Dan didapatkan nilai k yang akurasinya tinggi yaitu k = 8 dengan 77% serta didaparkan nilai akurasi sebesar 77%, nilai presisi sebesar 76% dan nilai recall sebesar 71%.

Referensi

[1] I. N. Atthalla, A. Jovandy, and H. Habibie, "Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K Nearest Neighbor," *Pros. Annu. Res. Semin.*, vol. 4, no. 1, pp. 148–151, 2018.

[2] A. Rizka, M. K. Akbar, and N. A. Putri, "Carcinoma Mammae Sinistra T4bN2M1 Metastasis Pleura," *AVERROUS J. Kedokt. dan Kesehat. Malikussaleh*, vol. 8, no. 1, p. 23, 2022, doi: 10.29103/averrous.v8i1.7006.

[3] I. Yulianti, H. Setyawan, and D. Sutiningsih, "Faktor-Faktor Risiko Kanker Payudara," *J. Kesehat. Masy.*, vol. 4, no. 4, pp. 401–409, 2016.

[4] H. Dewi, "Analisis risiko kanker payudara berdasar riwayat pemakaian kontrasepsi hormonal dan usia," *J. Berk. Epidemiol.*, vol. 3, no. 1, pp. 12–23, 2015.

[5] S. R. Listyanto, "Implementasi K-Nearest Neighbor Untuk Mengenali Pola Citra Dalam Mendeteksi Penyakit Kulit," *UDiNus Repos.*, pp. 1–7, 2015.

[6] M. I. P. Putra, D. T. Murdiansyah, and A. Aditsania, "Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Penyakit Kanker Payudara Tugas Akhir diajukan untuk memenuhi salah satu syarat memperoleh gelar sarjana dari Program Studi S1 Ilmu Komputasi M Ikhsan Perdana Putra Program Studi," *E-proceeding Eng.*, vol. 6, no. 1, pp. 2431–2441, 2019.

[7] A. Herawati, S. Rijal, A. St Fahira Arsal, R. Purnamasari, D. Amelia Abdi, and S. Wahid, "Karakteristik Kanker Payudara," *Fakumi Med. J. J. Mhs. Kedokt.*, vol. 2, no. 5, pp. 359–367, 2022.

[8] *Data Mining*.

[9] P. Putra, A. M. H. Pardede, and S. Syahputra, "Analisis Metode K-Nearest Neighbour (Knn) Dalam Klasifikasi Data Iris Bunga," *J. Tek. Inform. Kaputama*, vol. 6, no. 1, pp. 297–305, 2022.

[10] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.

[11] M. A. Imron, "Peningkatan Akurasi Algoritma k-Nearest neighbor menggunakan Normalisasi Z-Score dan particle

Swarm Optimization untuk Prediksi Customer Churn," 2020.

[12] A. Yandi Saputra and Y. Primadasa, "Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbour Implementation of Classification Method to Predict Student Graduation Using K-Nearest Neighbor Algorithm," *Techno.Com*, vol. 17, no. 4, p. 9, 2018.

[13] F. L. D. Cahyanti, W. Gata, and F. Sarasati, "Implementasi Algoritma Naïve Bayes dan K-Nearest Neighbor Dalam Menentukan Tingkat Keberhasilan Immunotherapy Untuk Pengobatan Penyakit Kanker Kulit," *J. Ilm. Univ. Batanghari Jambi*, vol. 21, no. 1, p. 259, 2021, doi: 10.33087/jiubj.v21i1.1189.

[14] J. Tp. A. K.-M. D. N. B. U. M. P. P. T. R. S. B. T. K. P. P. P. (PERSERO)eknik Informatika *et al.*, "Perbandingan Algoritma K-Means Dan Naive Bayes Untuk Memprediksi Prioritas Pembayaran Tagihan Rumah Sakit Berdasarkan Tingkat Kepentingan Pada PT.Pertamina (PERSERO)," vol. 13, no. 2, pp. 1–8, 2021.

[15] M. Sholeh, D. Andayati, and R. Y. Rachmawati, "Data Mining Model Klasifikasi Menggunakan Algoritma K-Nearest Neighbor Dengan Normalisasi Untuk Prediksi Penyakit Diabetes," *TeIKa*, vol. 12, no. 02, pp. 77–87, 2022, doi: 10.36342/teika.v12i02.2911.