ISSN ONLINE. 2988-1900

Klasifikasi Kanker Payudara Menggunakan Metode K-Nearest Neighbor

Naomi Angela Meyana[1], Yessica Tarida Sheevana Sitorus[2], Achmad Zanbar Soleh[3]

Universitas Padjadjaran[1]
yessica20001@mail.unpad.ac.id[1]

Abstract

Abstrak. *Kanker payudara merupakan salah satu penyakit yang paling banyak menyebabkan kematian. Tumor pada kanker payudara dibagi menjadi dua, yaitu benign atau biasa disebut jinak dan malignant atau biasa disebut juga ganas. Klasifikasi tersebut diperlukan untuk menentukkan tindakan medis apa yang diperlukan dalam mengatasi tumornya. Klasifikasi ini menggunakan algoritma K Nearest Neighbour (KNN). Data yang digunakan bersifat open source, yang dapat diakses di Kaggle. Hasil penelitian ini menunjukkan bahwa metode KNN dapat digunakan untuk menentukan jenis tumor pada kanker payudara, apakah ganas atau jinak dengan tingkat akurasi 98%.*

Kata kunci: kanker payudara, klasifikasi, k-nearest neighbour

I. PENDAHULUAN

World Health Organization (WHO) menyatakan bahwa kanker payudara terdapat di urutan kelima dalam kanker yang paling banyak menyebabkan kematian. Kanker payudara adalah sekelompok sel tidak normal pada payudara yang terus tumbuh berupa ganda. Akhirnya, sel-sel ini menjadi bentuk benjolan di payudara. Jika benjolan tersebut tidak dibuang atau dikontrol, sel-sel kanker bisa menyebar pada bagian tubuh lainnya. Menurut survey yang dilakukan WHO, sekitar 8%-9% wanita mengidap kanker payudara. Oleh karena itu, kanker payudara merupakan salah satu penyakit mematikan di dunia, termasuk Indonesia.

Tumor pada kanker payudara secara umum dibagi menjadi dua, yaitu *benign* atau biasa disebut jinak dan *malignant* atau biasa disebut juga ganas. Perbedaan paling mendasar dari kedua tipe tumor payudara adalah perkembangannya. Tumor yang ganas berkembang menjadi sel kanker, sementara tumor yang sifatnya jinak tidak bersifat kanker dan bisa sembuh dengan sendirinya.

Salah satu metode dalam pengolahan data adalah klasifikasi. Klasifikasi merupakan cara pengelompokkan benda berdasarkan ciri – ciri yang dimiliki oleh objek klasifikasi. Dalam prosesnya, klasifikasi dapat dilakukan dengan banyak cara baik secara manual maupun dengan bantuan teknologi. Klasifikasi dapat diartikan sebagai pengelompokan fitur ke dalam kelas yang sesuai[9].

Pengklasifikasian kanker payudara diperlukan untuk menyediakan model klasifikasi yang dapat digunakan untuk melakukan prediksi maupun pengambilan keputusan medis lainnya. Hal ini tentunya akan berguna pada bidang kesehatan, tetapi penelitian ini juga dapat berguna dalam bidang asuransi khususnya dalam bidang aktuaria.

Tabel Morbiditas Indonesia merupakan kebutuhan dari industri asuransi sebagai acuan bagi para aktuaris mengembangkan produk dan menetapkan premi khususnya produk asuransi jiwa dan kesehatan yang memberi perlindungan terhadap penyakit kritis. Manfaat tabel ini adalah agar perusahan dapat menentukkan harga premi yang kompetitif dan wajar sekaligus memudahkan aktuaris dalam penentuan harga premi dan pengembangan produk. Penelitian ini dapat berguna dalam bidang aktuaria terutama dalam pembuatan Tabel Morbiditas Indonesia.

Pada penelitian ini, klasifikasi kanker payudara akan dilakukan dengan metode K-Nearest Neighbor (KNN). Metode KNN digunakan karena memiliki kelebihan, yaitu merupakan lazy learning

ISSN ONLINE. 2988-1900

algorithm, karena tidak membutuhkan banyak data dalam pengaplikasiannya. Metode KNN juga mudah diimplementasikan dan hasil akhirnya mudah untuk diterjemahkan.

Metode KNN melakukan klasifikasi terhadap objek berdasarkan jarak terdekat terhadap tetangganya. Nilai jarak inilah yang digunakan sebagai nilai kedekatan/kemiripan antara data uji dengan data latih. Sehingga algoritma KNN cocok untuk diterapkan dalam mengklasifikasi kanker payudara.

Penelitian ini sebelumnya sudah dibahas pada artikel dengan judul "Diagnosis Kanker Payudara Menggunakan Machine Learning Dengan Algoritma K-Nearest Neighbor" untuk menentukkan apakah tumor pada kanker payudara adalah jinak atau ganas. Penelitian tersebut membuat aplikasi untuk menginputkan nilai dalam penentuan diagnosis. Namun, dalam penelitian tersebut tidak dilakukan evaluasi model, dimana evaluasi model penting untuk melihat keakuratan model tersebut.

Berdasarkan latar belakang masalah yang telah dipaparkan sebelumnya, maka muncul suatu permasalahan yaitu perlunya pengklasifikasian kanker payudara, apakah tumor yang ada dikategorikan jinak atau ganas. Maksud dari penelitian ini adalah untuk mengaplikasikan metode K-Nearest Neighbor dalam klasifikasi kanker payudara. Adapun tujuan dari penelitian ini:

1. Melakukan klasifikasi pada kanker payudara untuk menentukkan apakah tumor pada kankernya ganas atau jinak.
2. Mengevaluasi performa algoritma *K-Nearest Neighbour* pada klasifikasi kanker payudara.

Manfaat yang diperoleh berdasarkan penelitian ini:

1. Bagi penulis, penelitian ini dapat menambah pengetahuan dalam penerapan algoritma *K-Nearest Neighbour* pada klasifikasi kanker payudara.
2. Bagi pembaca, penelitian ini dapat menambah pengetahuan tentang algoritma *K-Nearest Neighbour*.
3. Bagi perusahaan asuransi, penelitian ini dapat melakukan analisis terhadap kondisi pasien dan membandingkannya dengan hasil yang dikeluarkan rumah sakit.

II. METODE PENELITIAN

2.1 *Data Preparation*

Penelitian ini menggunakan dataset Kanker Payudara Wisconsin yang dapat diakses pada Kaggle. Dataset Kanker Payudara Wisconsin terdiri dari 569 sampel dengan 32 atribut.

Tabel 1. Deskripsi Fitur-Fitur pada Data

No	Atribut	Deskripsi
1	radius_mean	Rata-rata jarak dari pusat ke titik-titik pada keliling sel inti dari benjolan di payudara
2	texture_mean	Standar deviasi nilai *grayscale* pada sel inti dari benjolan di payudara
3	perimeter_mean	Ukuran rata-rata sel inti dari benjolan di payudara
4	area_mean	Ukuran rata-rata dari sel inti dari benjolan di payudara
5	smoothness_mean	Rata-rata variasi lokal dalam panjang radius sel inti dari benjolan di payudara
6	compactness_mean	Rata-rata untuk keliling^2 / luas sel inti dari benjolan di payudara dikurangi 1
7	concavity_mean	Rata-rata kecekungan sel inti dari benjolan

ISSN ONLINE. 2988-1900

		di payudara
8	concave points_mean	Rata-rata untuk jumlah bagian cekung sel inti dari benjolan di payudara
9	symmetry_mean	Rata-rata untuk simetrisitas dari kontur sel inti dari benjolan di payudara
⋮	⋮	⋮
26	compactness_worst	Nilai rata-rata terbesar untuk keliling^2 / luas sel inti dari benjolan di payudara dikurangi 1
27	concavity_worst	Nilai rata-rata terbesar untuk tingkat keparahan bagian cekung pada sel inti dari benjolan di payudara
28	concave points_worst	Nilai rata-rata terbesar untuk jumlah bagian cekung pada sel inti dari benjolan di payudara
29	symmetry_worst	Nilai rata-rata terbesar untuk simetrisitas pada sel inti dari benjolan di payudara
30	fractal_dimension_worst	Nilai rata-rata terbesar untuk "perkiraan *coastline*" sel inti dari benjolan di payudara dikurangi 1
31	ID *Number*	Nomor pasien
32	Diagnosis	diagnosis dari pasien (B=*benign*, M=*malignant*)

2.2 *Data Preprocessing*

a. Pemisahan data

Pemisahan data ini dilakukan untuk membagi data menjadi dua bagian yaitu data latih dan data uji. Data latih digunakan untuk mengenali karakteristik pasien yang terkena kanker payudara maupun yang tidak terkena kanker payudara. Data uji dalam uji coba berguna terhadap model klasifikasi yang dihasilkan, serta menilai performa dari model klasifikasi dengan membandingkan hasil klasifikasi model terhadap tiap data dalam data testing dengan label sebenarnya.

b. Normalisasi data

Pada penelitian ini, metode yang digunakan untuk normalisasi data adalah normalisasi min-maks. Metode ini melakukan transformasi linear terdapat data asli dengan rumus :

$$X_n = \frac{X_0 - X_{min}}{X_{maks} - X_{min}} \quad (1)$$

Keterangan:

$X_n = nilai\ baru\ untuk\ variabel\ x$

$X_0 = nilai\ asli\ variabel\ x$

$X_{min} = nilai\ minimum\ dari\ variabel\ x$

ISSN ONLINE. 2988-1900

$X_{maks} = nilai\ maksimum\ dari\ variabel\ x$

Keuntungan dari metode ini adalah keseimbangan nilai perbandingan antardata saat sebelum dan sesudah proses normalisasi, serta tidak ada bias yang dihasilkan oleh metode ini. Kekurangan dari metode ini adalah ketika ada data baru, metode ini memungkinkan terjebak "*out of the bound*" error [6].

2.3 *K-Nearest Neighbor* (KNN)

Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasikan sebelumnya [7]. jarak terpendek dari data uji ke data latih untuk menentukan jarak tetangga K terdekat. Setelah mendapatkan KNN, kemudian diambil mayoritas dari KNN untuk dijadikan prediksi dari sampel uji yang diambil berdasarkan dekat atau jauhnya tetangga yang disebut jarak Euclidean .

Cara kerja algoritma KNN :

a) Menentukan parameter K, dengan $K \equiv$ jumlah tetangga terdekat.

b) Menghitung jarak antara data baru dengan semua data latih, dengan:

$$Euclidean = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (2)$$

Keterangan :

$p_i = data\ uji$

$q_i = data\ latih$

$n = \ jumlah\ variabel$

c) Mengurutkan data tersebut dan menetapkan tetangga terdekat berdasarkan jarak minimum ke K.

d) Memeriksa kelas dari jarak terdekat.

e) Menggunakan mayoritas sederhana dari kelas tetangga terdekat sebagai kelas data yang akan dievaluasi

2.4 Evaluasi Model

Selanjutnya akan dilakukan evaluasi model untuk melihat nilai akurasi sebuah model dengan melakukan *Confusion Matrix*. *Confusion Matrix* adalah metode yang digunakan untuk mengukur kinerja suatu model klasifikasi. Metode ini akan menampilkan dan membandingkan nilai aktual atau nilai sebenarnya dengan nilai hasil prediksi model yang dapat digunakan untuk mendapatkan matrik evaluasi seperti *Accuracy*, *Precision*, *Recall* dan *F1-Score*. *Confusion Matrix* dapat digambarkan sebagai berikut:

Actual Values			
	0	TP	FP
	1	FN	TN
		0	1
		Predicted Values	

Gambar 1. *Confusion Matrix*

Keterangan:

True Positive (TP) : Jumlah data yang bernilai Positif dan diprediksi benar Positif

False Positive (FP) : Jumlah data yang bernilai Negatif tetapi diprediksi Positif

True Negative (TN) : Jumlah data yang bernilai Negatif dan diprediksi benar Negatif

False Negative (FN) : Jumlah data yang bernilai Positif tetapi diprediksi Negatif

Accuracy adalah tingkat kedekatan antara nilai prediksi dengan nilai sebenarnya dengan rumus sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

Precision adalah rasio dari item relevan yang dipilih terhadap jumlah item yang terpilih dengan rumus sebagai berikut:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Recall adalah rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia dengan rumus sebagai berikut:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

F-Measure adalah harmonic mean antara nilai Precision dan Recall dengan rumus sebagai berikut:

$$F - Measure = \frac{2(Precision.Recall)}{Precision + Recall} \quad (6)$$

III. HASIL DAN PEMBAHASAN

Berikut adalah data yang digunakan dalam penerapan metode klasifikasi *K-Nearest Neighbor*. Data merupakan dataset dari pasien penderita kanker payudara yang terdiri dari 569 data dan 32 atribut.

Tabel 2. Dataset *Breast Cancer Disease*

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	...	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
842302	M	0,77708333	10.38	122.08.00	1001	...	4,62222222	4,94375	1,84305556	3,195138889	0,825694444
842517	M	20.57	0,76180556	132.09.00	1326	...	1,29583333	1,67777778	186	275	0.08902
84300903	M	0,83958333	21.25	130	1203	...	2,94791667	3,12777778	243	2,509027778	0.08758
84348301	M	11.42	20.38	77.58.00	386.01.00	...	6,01597222	4,77013889	1,78819444	4,609722222	173
84358402	M	20.29	14.34	135.01.00	1297	...	205	00.04	1,12847222	1,641666667	0.07678
843786	M	12.45	15.07	82.57.00	477.01.00	...	3,64513889	3,71875	1,20902778	2,767361111	0,863888889
844359	M	18.25	0,85972222	119.06.00	1040	...	1,78888889	2,62777778	1,34166667	2,127083333	0.08368
84458202	M	0,59097222	0,89097222	90.02.00	577.09.00	...	2,55694444	1,85972222	1,08055556	2,219444444	0,799305556
844981	M	13	0,93194444	87.05.00	519.08.00	...	3,75069444	539	206	3,040277778	0,744444444
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
925292	B	14.05	27.15.00	91.38.00	600.04.00	...	1,57222222	0,92083333	0,72777778	225	0.08321
925311	B	11.02	29.37.00	70.67	386	...	0.05494	0	0	1,0875	0.05905
925622	M	15.22	30.62	103.04.00	716.09.00	...	5,49791667	01.17	1,63611111	2,839583333	0,978472222
926125	M	0,89722222	25.09.00	143	1347	...	2,90694444	4,58263889	1,76527778	2,034027778	0.09873
926424	M	21.56	22.39	142	1479	...	1,46736111	2,85208333	1,53888889	206	0.07115
926682	M	20.13	28.25.00	131.02.00	1261	...	1,33472222	2,23263889	1,13055556	1,786111111	0.06637
926954	M	16.06	28.08.00	108.03.00	858.01.00	...	2,14861111	2,36319444	0,98472222	1,540277778	0,543055556
927241	M	20.06	29.33.00	140.01.00	1265	...	6,02847222	6,51875	265	2,838194444	124
92751	B	0,34444444	24.54.00	47.92	181	...	0.06444	0	0	1,99375	0.07039

ISSN ONLINE. 2988-1900

3.1 Data Preprocessing

Data yang telah disiapkan melalui tahap *preprocessing* sebelum bisa digunakan. Pada tahap ini dilakukan normalisasi data menggunakan metode *Min-Max Normalization* pada persamaan (1). Normalisasi data ini dilakukan agar semua nilai variabel berada pada interval 0-1 sehingga mencegah bias dari nilai variabel yang jauh lebih besar daripada nilai pada variabel lainnya.

Selain normalisasi, dilakukan pemisahan data. Pemisahan data dilakukan dengan perbandingan 80:20 untuk data latih dan data uji. Data pengamatan berjumlah 569, maka data pengamatan pertama sampai ke-455 akan berperan sebagai data latih dan data ke-456 sampai ke-569 akan berperan sebagai data uji.

3.2 Implementasi dan Evaluasi Model KNN

Setelah dilakukan normalisasi data dan pemisahan data, dataset penyakit kanker payudara dapat digunakan dalam implementasi model KNN. Untuk pengukuran jarak akan digunakan persamaan (2).

Akan dilakukan ilustrasi perhitungan menggunakan data pengamatan ke-456 sebagai data uji.

a) Menentukan nilai K

Dalam pemilihan nilai K tidak terdapat aturan resmi, namun cukup lazim dilakukan dengan cara mengambil nilai akar kuadrat dari jumlah data.

$$\sqrt{569} = 23.85372 \simeq 24$$

Akan digunakan nilai K=24 untuk klasifikasi KNN ini. Artinya 24 titik data dengan jarak terdekat dengan data uji akan masuk dalam *decision barrier* untuk menentukan kelas dari data uji.

b) Menghitung *Eucledian Distance*

$$Euclidean = \sqrt{\sum_{i=1}^{30}(p_i - q_i)^2}$$

$$= \sqrt{(0.521037 - 0.52103744)^2 + (0.7105174 - 0.02265810)^2 + \ldots + (0.142398 - 0.41886396)^2}$$

$$= 2{,}168343$$

Perhitungan jarak ini dilakukan terhadap ke-455 data latih untuk setiap 1 set data uji yang akan diklasifikasi. Dari ke-455 jarak yang sudah dihitung, akan diambil 24 titik data yang memiliki jarak paling dekat dengan titik data uji.

Tabel 3. Jarak 24 Titik Data Terdekat

rank	*euclidean distance*	*diagnosis*	rank	*euclidean distance*	*diagnosis*	rank	*euclidean distance*	*diagnosis*
1	0,184928	B	9	0,280478	B	17	0,315887	B
2	0,218133	B	10	0,280802	B	18	0,322522	B
3	0,220689	B	11	0,289078	B	19	0,323587	B
4	0,226177	B	12	0,290138	B	20	0,32592	B
5	0,227504	B	13	0,290269	B	21	0,327797	B
6	0,239431	B	14	0,293001	B	22	0,331711	B
7	0,245598	B	15	0,293891	B	23	0,332646	B
8	0,252642	B	16	0,298745	B	24	0,335428	B

c) Memberikan Kelas Kepada Data Uji

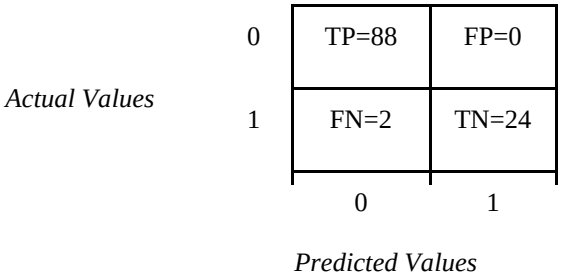
Dari ke-24 data latih yang memiliki jarak paling dekat dengan titik data uji, keseluruhannya merupakan kelas B atau *Benign*. Sehingga untuk data pengamatan ke-456 diklasifikasikan sebagai *Benign*.

ISSN ONLINE. 2988-1900

d) Membuat *Confusion Matrix*

Setelah klasifikasi 114 data uji selesai, untuk perhitungan menggunakan bantuan perangkat lunak seperti RStudio, akan dihasilkan *Confusion Matrix* yang berfungsi untuk mengevaluasi proses pengklasifikasian. Ukuran dari matriks ini akan mengikuti jumlah kelas yang dimiliki masing-masing dataset, untuk dataset Breast Cancer Disease dimiliki 2 kelas, sehingga *Confusion Matrix* berukuran 2x2.

Dihasilkan matriks yang tertera pada Gambar 2, 0 Menggambarkan *Benign* dan 1 menggambarkan *Malignant*. Nilai *True Positive* sejumlah 88 menunjukan bahwa dari 114 data uji, 88 diantaranya merupakan kanker payudara *Benign* dan tepat diprediksi sebagai *Benign*. Nilai *False Positive* sejumlah 0 menunjukan bahwa dari 114 data uji, 0 diantaranya merupakan kanker payudara *Benign* namun kelirru diprediksi sebagai *Malignant*. Nilai *False Negative* sejumlah 2 menunjukan bahwa dari 114 data uji, 2 diantaranya merupakan kanker payudara *Malignant* namun keliru diprediksi sebagai *Benign*. Nilai *True Negative* sejumlah 24 menunjukan bahwa dari 114 data uji, 24 diantaranya merupakan kanker payudara *Malignant* dan tepat diprediksi sebagai *Malignant*.



Gambar 2. *Confusion Matrix Breast Cancer Disease*

Dari *confusion matrix* yang dihasilkan, dapat dilakukan evaluasi kinerja klasifikasi *K Nearest Neighbor* pada dataset Breast Cancer Disease menggunakan empat indikator yang diantaranya adalah *Accuracy*, *Precision*, *Recall*, *F1-Score*, keempat indikator ini dihitung menggunakan persamaan (3), (4), (5), (6) dan memperoleh hasil sebagai berikut

Accuracy : 0.9824561/1

Precision : 1/1

Recall : 0.9777778/1

F1-Score : 0.988764/1

IV. KESIMPULAN

Telah dilakukan klasifikasi terhadap kanker payudara menggunakan dataset Breast Cancer Disease menggunakan Algoritma *K Nearest Neighbor*, dan evaluasi kinerja algoritma yang sangat baik dengan *Accuracy* 0.9824561/1, *Precision* : 1/1, *Recall* : 0.9777778/1, dan *F1-Score* : 0.988764/1. Kedepannya, dapat dilakukan klasifikasi terhadap data di lapangan medis maupun asuransi menggunakan Algoritma ini.

ISSN ONLINE. 2988-1900

DAFTAR PUSTAKA

[1] Admojo, F.T., 2020. Klasifikasi Aroma Alkohol Menggunakan Metode KNN. *Indonesian Journal of Data and Science*, *1*(2), pp.34-38.

[2] Atthalla, I.N., Jovandy, A. and Habibie, H., 2018. Klasifikasi Penyakit Kanker Payudara Menggunakan Metode K-Nearest Neighbor. *Pros. Annu. Res. Semin*, *4*(1), pp.978-979.

[3] Chazar, C., Nursyamsi, I. and Herwanto, P., 2021, October. Diagnosis Kanker Payudara Menggunakan Machine Learning Dengan Algoritma K-Nearest Neighbor. In *Prosiding Seminar Nasional Inovasi dan Adopsi Teknologi (INOTEK)* (Vol. 1, No. 1, pp. 182-191).

[4] Fear, E.C., Meaney, P.M. and Stuchly, M.A., 2003. Microwaves for breast cancer detection?. *IEEE potentials*, *22*(1), pp.12-18.

[5] Nasution, D.A., Khotimah, H.H. and Chamidah, N., 2019. Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN. *CESS (Journal of Computer Engineering, System and Science)*, *4*(1), pp.78-82.

[6] Putra, Sitiatava Rizema. 2015. *Buku Lengkap Kanker Payudara*. Laksana.

[7] Ramadhan, N.G. and Adhinata, F.D., 2021. Teknik SMOTE dan Gini Score dalam Klasifikasi Kanker Payudara. *RADIAL: Jurnal Peradaban Sains, Rekayasa dan Teknologi*, *9*(2), pp.125-134.

[8] Saniy, R.N., Sitorus, Y.T.S., Meyana, N.A., Najwa, S., Pravitasari, A.A. and Indrayatna, F., 2022. Penerapan Algoritma K-Nearest Neighbor pada Klasifikasi Penyakit Jantung. *E-Journal BIAStatistics| Departemen Statistika FMIPA Universitas Padjadjaran*, *2022*(1), pp.222-229.

[9] Wibawa, A.P., Guntur, M., Purnama, A., Akbar, M.F. and Dwiyanto, F.A., 2018. Metode-metode Klasifikasi. In *Prosiding Seminar Ilmu Komputer Dan Teknologi Informasi* (Vol. 3, No. 1).