

Prakiraan Keganasan Kanker Payudara Menggunakan Metode Klasifikasi k-NN dengan Analisis Atribut Multivariabel

Wahyu Adi Pratama^{1*}, Puput Dwi Mandiri², Bayu Tri Pramono³, Dwi Hartanti⁴

^{1,2,3,4}Teknik Informatika

Universitas Duta Bangsa Surakarta

^{1*}200103113@mhs.udb.ac.id, ²202040332@mhs.udb.ac.id, ³200103088@mhs.udb.ac.id, ⁴dwihartanti@udb.ac.id

Abstrak— Salah satu jenis kanker yang banyak terjadi pada wanita didunia adalah kanker payudara. Pengobatan dan perawatan diperlukan sedini mungkin guna harapan kesembuhan yang lebih baik. Pada jurnal ini, metode k-Nearest Neighbors (k-NN) dengan analisis atribut multivariabel kami usulkan guna memperkirakan keganasan kanker payudara. k-NN merupakan algoritma klasifikasi yang berdasar pada jarak antara data. Pada jurnal ini, metode k-NN yang kami gunakan bermanfaat untuk mengklasifikasikan sampel baru dengan berdasar pada kecocokan dengan tetangga terdekat dari setiap atribut yang diuji. Penggunaan metode k-NN dalam penelitian ini memiliki maksud untuk mengembangkan dan menerapkan metode k-NN dengan analisis atribut multivariabel dalam manfaatnya untuk memperkirakan keganasan kanker. Label akhir dari dataset kanker payudara yang kami gunakan berupa b (benign) atau m (malignant) yang berdasarkan sampel didalam dataset tersebut serta kami juga melakukan evaluasi dari kinerja metode k-NN dalam memperkirakan tingkat keganasan. Kesimpulannya kami mengusulkan pengimplementasian metode k-NN dengan analisis atribut multivariabel dalam memperkirakan keganasan kanker payudara. Metode k-NN memiliki potensi dalam memfasilitasi hasil diagnosis yang lebih akurat dan efektif pada kanker payudara.

Kata kunci : Breast Cancer, k-Nearest Neighbors (kNN), Data Mining.

Abstract— One of the most common types of cancer in women in the world is breast cancer. Treatment and care are needed as early as possible in hopes of a better recovery. In this journal, we propose the k-Nearest Neighbors (k-NN) method with multivariable attribute analysis to estimate breast cancer malignancy. k-NN is a classification algorithm based on the distance between data. In this paper, the k-NN method that we use is useful for classifying new samples based on compatibility with the closest neighbor of each attribute tested. The use of the k-NN method in this study has the intention of developing and implementing the kNN method with attribute analysis. multivariable in its usefulness for estimating cancer malignancy. The final label of the breast cancer dataset that we use is b (benign) or m (malignant) which is based on the sample in the dataset and we also evaluate the performance of the k-NN method in estimating the level of malignancy. In conclusion, we propose the implementation of the k-NN method with multivariable attribute analysis in estimating breast cancer malignancy. The k-NN method has the potential to facilitate a more accurate and effective diagnosis of breast cancer.

Keywords : Breast Cancer, k-Nearest Neighbors (kNN), Multivariable.

I. PENDAHULUAN

1.1 Latar Belakang

Kanker payudara masih menjadi salah satu penyumbang kematian terbesar pada wanita diseluruh dunia setelah kanker paru paru [1]. Dalam hal ini pendeteksian sedini mungkin dari kanker merupakan hal yang memiliki peranan besar dalam kesembuhan pasien [2]. Namun upaya ini bukanlah hal yang mudah. Kanker payudara merupakan kanker yang memiliki kompleksitas yang cukup rumit dan ketidaksamaan didalam dataset atribut. Disebabkan hal tersebut, Penerapan dan pengembangan metode klasifikasi yang mumpuni dan akurat guna memprakirakan keganasan dari kanker payudara menjadi hal yang cukup krusial.

Pada studi terdahulu telah menerapkan beberapa pendekatan dengan klasifikasi untuk memprakirakan sesuatu. Diantara metode yang telah dilakukan termasuk diantaranya Jaringan saraf tiruan (*Artificial Neural Network*), pohon keputusan (*Decision Tree*), dan metode pembelajaran mesin lainnya. Namun, kami dalam penelitian kali ini mengusulkan metode *k-Nearest Neighbors* dengan analisis atribut multivariabel.

k-NN adalah metode pengklasifikasian data baru berbasis mayoritas kesamaan antara tetangga terdekat yang merupakan algoritma supervised learning [3]. Metode klasifikasi sudah banyak digunakan untuk mendukung sistem pengambilan keputusan [4]. Metode k-NN memiliki reputasi yang baik untuk

menyelesaikan berbagai masalah klasifikasi dan dapat digunakan pada pengklasifikasian sampel baru berbasis pada kesamaan dengan tetangga terdekat dalam ruang atribut [5].

1.2 Tujuan Penelitian

Hasil yang diharapkan dari penelitian ini adalah penerapan dan pengembangan metode k-NN dengan analisis atribut multivariabel dalam memprakirakan keganasan kanker payudara. Dengan adanya penelitian ini kami berharap untuk memaksimalkan akurasi prakiraan dan klasifikasi yang lebih baik dalam diagnosis kanker payudara. Dalam penelitian ini kami memiliki dataset yang berisi atribut dari sampel kanker payudara dengan label akhir berupa b (benign) atau m (malignant).

II. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Berikut beberapa tahapan penelitian menggunakan metode k-NN :

1. Menentukan Permasalahan dan Tujuan Penelitian :

Tentukan terlebih dahulu permasalahan apa yang akan di teliti dan tujuan dari permasalahan tersebut yaitu untuk melakukan uji coba dan meneliti apakah metode k-NN dapat diselesaikan dengan analisis atribut mutivariabel.

2. Mencari Dataset

Dataset diambil dari website kaggle.com dimana dataset tersebut berisi atribut dari sampel kanker payudara dengan 2 tingkat keganasan dan diberi label akhir berupa b (benign) atau m (malignant). Benign berarti jinak dan Malignant berarti ganas.

3. Menghitung Jarak

Untuk menemukan titik yang paling dekat dengan titik tertentu, kita perlu mengukur jarak antara titik tersebut dengan titik

lainnya. Kita dapat menggunakan metode yang berbeda seperti Jarak Euclidian, Jarak Hamming[7], Jarak Manhattan[6], dan Jarak Minkowski[8] untuk melakukan perhitungan ini. Dengan menggunakan metode ini, kita dapat mengetahui grup mana yang termasuk dalam poin dan membuat keputusan berdasarkan itu.

4. Mendefinisikan nilai K

Untuk menemukan titik terdekat dengan titik khusus atau nilai K[9], kita harus mengukur seberapa jauh titik tersebut dari titik lainnya. Kita dapat menggunakan berbagai cara untuk mengetahuinya, seperti mengukur dalam garis lurus atau menghitung berapa banyak langkah yang diperlukan. Dengan melakukan ini, kita dapat melihat di grup mana intinya berada dan membuat pilihan berdasarkan itu.

5. Evaluasi

Evaluasi pada penelitian keganasan kanker payudara ini mennggunakan 2 indikator utama yaitu presisi dan recall. Hasil perbandingan yaitu analisis atribut multivariabel. Analisis multivariabel adalah analisis yaang melibatkan 1 variabel bebas dan digunakan ketika berhadapan dengan data yang memiliki setidaknya lebih dari 1 variabel yaitu variabel independen dan dependen.

M	14.68	20.13	94.74	684.5
M	16.13	20.68	108.1	798.8
M	19.81	22.15	130	1260
B	13.54	14.36	87.46	566.3
B	13.08	15.71	85.63	520
B	9.504	12.44	60.34	273.9

Gambar 1. Tabel Data

Adapun penjelasan dari 2 variabel tersebut adalah :

- M (Malignant) berarti ganas.
- B (Benign) berarti jinak.

Pada metode ini kami menggunakan cara Euclidian Distance untuk menghitung jarak k-NN atau nilai rata-rata.

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

Gambar2. Euclidian Distance

Pertama kita menentukan variabel X dan Y terlebih dahulu dan variabel X merupakan nilai absolut pada koordinat titik X, dan nilai Y merupakan nilai absolut pada koordinat titik Y. lalu hitung dengan menggunakan rumus Manhattan diatas untuk menentukan nilai rata-rata.

2.2 Dataset

Ada dua jenis label yang diklasifikasikan dalam dataset kanker payudara yaitu Malignant dan Benign. Dataset berjumlah 204 item terdiri dari X1, X2, Y1, Y2, Z1, Z2. Pada ketiga variabel tersebut dapat dihitung dengan cara Manhattan dalam mencari nilai rata-rata per variabel. Selain menghitung dengan cara ini dapat juga menggunakan rapid miner sebagai alat bantu dalam mengklasifikasikan data. Setelah dilakukan dengan menggunakan excel dan Rapid Miner akurasi kebenaran data mencapai lebih dari 90% bernilai sama atau sesuai.

2.3 RapidMiner

RapidMiner adalah sebuah platform analitik prediktif yang populer digunakan untuk melakukan pemodelan data, analisis prediktif, dan penambahan data. Platform ini menyediakan beragam alat dan fungsionalitas yang memungkinkan pengguna untuk mengimpor data, membersihkan data, menerapkan teknik pemodelan statistik, dan menghasilkan wawasan yang berguna dari data. RapidMiner juga menyediakan antarmuka grafis yang intuitif dan alur kerja berbasis diagram, yang memungkinkan pengguna untuk membangun dan mengelola alur kerja analitik dengan mudah. Dalam alur kerja tersebut, pengguna dapat menggabungkan berbagai langkah seperti impor data, pembersihan data, transformasi variabel, ekstraksi fitur, pemodelan, evaluasi, dan visualisasi..

III. HASIL DAN PEMBAHASAN

Pada penelitian kali ini kami mencari nilai terdekat antara atribut dengan menggunakan euclidean distance yang berdasar pada hubungan antara atribut pada dataset yang berisikan faktor faktor yang menentukan tingkat keganasan kanker payudara. Kami juga mencari tingkat keakurasian data tertinggi dengan mengubah nilai K.

3.1 Sumber Data

Data yang kami gunakan kali ini menggunakan dataset keganasan kanker payudara beserta faktor faktor pendukungnya yang diubah menjadi atribut. Data ini kami buat menjadi dua yang pertama yaitu data training merupakan data yang peruntukan sebagai acuan dalam membangun model klasifikasi [10] sedangkan data testing merupakan sekelompok data yang didalamnya

Tabel 1. Tabel Dataset Keganasan Kanker Payudara

Kolom	Role
ID	Id
Diagnosa	Label
Rata-rata Radius	Atribut
Rata-rata texture	Atribut
Rata-rata perimeter	Atribut
Rata-rata area	Atribut
Rata-rata kelancaran	Atribut
Rata-rata kekompakkan	Atribut

terdapat label atau kelas yang digunakan dalam pengujian keakurasian pola dalam mengklasifikasikan data testing [11].

Data diambil dari website kaggle dimana merupakan dataset faktor faktor penentuan tingkat keganasan kanker payudara yang bersifat public.

Dataset ini terdiri dari enam kolom yang berupa atribut, 1 kolom label dan 1 kolom id.

3.2 Penggunaan Algoritma k-NN

Algoritma k-NN adalah algoritma yang bekerja dengan mencari jarak terdekat antara data training dan data testing. Untuk menghitung jarak tersebut kami menggunakan metode euclidean distance. Dengan menghitung secara manual menggunakan excel dan perhitungan menggunakan

RapidMiner

Tabel 2. Tabel Training Data Lengkap

No	ID	Diagnosis	Rata-rata Radius (X1)	Rata-rata Texture (X2)	Rata-rata Perimeter (Y1)	Rata-rata Area (Y2)	Rata-rata Kelancaran (Z1)	Rata-rata Kekompakan (Z2)
1	842302	M	17,99	10,38	122,8	1001	0,1184	0,2776
2	842517	M	20,57	17,77	132,9	1326	0,08474	0,07864
3	84300903	M	19,69	21,25	130	1203	0,1096	0,1599
4	84348301	M	11,42	20,38	77,58	386,1	0,1425	0,2839
.....								
200	877486	M	19,18	22,49	127,5	1148	0,08523	0,1428
201	877500	M	14,45	20,22	94,49	642,7	0,09872	0,1206
202	877501	B	12,23	19,56	78,54	461	0,09586	0,08087
203	877989	M	17,54	19,32	115,1	951,6	0,08968	0,1198
204	878796	M	23,29	26,67	158,9	1685	0,1141	0,2084

Keterangan :

Label = Diagnosis yang memberi keterangan apakah jenis kanker b (*benign*) atau m (*malignant*)

Tabel 3. Tabel Testing Data Lengkap

No	ID	Diagnosis	Rata-rata Radius (X1)	Rata-rata Texture (X2)	Rata-rata Perimeter (Y1)	Rata-rata Area (Y2)	Rata-rata Kelancaran (Z1)	Rata-rata Kekompakan (Z2)
1	87880	?	13,81	23,75	91,56	597,8	0,1323	0,1768

Data testing adalah sekelompok data yang berisi faktor faktor yang mempengaruhi tingkat keganasan kanker payudara yang akan dikelompokkan menjadi b (*benign*) atau m (*malignant*).

Setelah terdapat data testing dan data training, lalu kami menentukan nilai k-nya. Sebagai contoh kali ini kami menetapkan k=9.

Sebagai berikut langkah langkah menjawab permasalahan diatas, yaitu

1. Nilai k kami tetapkan k=9
2. Menghitung jarak antar atribut menggunakan *euclidean distance* pada data training dan data testing

Kemudian hasil-hasil yang telah ditemukan di urutkan dalam jarak yang terkecil.

Langkah selanjutnya hasil dari pengelompokan “Apakah termasuk tetangga terdekat” yang dikelompokkan menggunakan “Label = Diagnosis” untuk menentukan apakah data tersebut termasuk bagian kelompok b (*benign*) atau m (*malignant*).

3.3 Pengujian data

3.3.1 Hitung Manual

Hasil hitung manual excel didapatkan hasil seperti berikut :

Nilai k kami tetapkan sebesar 9 lalu kami hitung nilai setiap atribut dari data testing dengan setiap atribut dari data training maka didapatkan yang terdekat sebagai berikut :

1. Suku X

Tabel 4. Tabel Hasil Hitung Suku X

RANKING	EUCLIDEAN DISTANCE	DIAGNOSA
5	12,8544 M	
7	31,3229 M	
12	75,7525 M	
4	10,0856 M	
2	7,74 M	
9	54,9704 M	
10	63,0189 M	
13	78,7924 M	
17	122,3416 M	
16	108,8776 M	

6	27,6921 M
19	146,41 M
18	123,7525 M
15	94,0821 M
20	213,9516 M
8	53,5169 M
11	63,1124 M
14	91,5325 M
1	3,3896 M
3	8,4241 B

Hasil akhir euclidean suku X

M = 8

B = 1

2. Suku Y

Tabel 5. Tabel Hasil Hitung Suku Y

RANKING	EUCLIDEAN DISTANCE	DIAGNOSA
14	552,5832426	M
20	877,5361018	M
17	754,5926053	M
3	63,99145568	M
19	848,7259923	M
2	27,14543976	M
15	591,2613297	M
6	128,2304566	M
4	70,10199712	M
1	26,06973149	M
12	348,4762976	M
10	331,7706437	M
16	674,9599692	M
11	333,4735072	M
7	128,9172215	M
8	209,3916734	M
9	234,9021873	M
13	349,8511826	M
18	811,4807453	M
5	116,4792754	B

Hasil akhir euclidean suku Y

M = 8

B = 1

3. Suku Z

Tabel 6. Tabel Hasil Hitung Suku Z

RANKING	EUCLIDEAN DISTANCE	DIAGNOSA
19	0,1294153	M
15	0,090254735	M
4	0,031971393	M
20	0,133923336	M
5	0,043265806	M
1	0,023427334	M
9	0,061150935	M
2	0,025601953	M
6	0,045028102	M
16	0,092120139	M
17	0,10147709	M
7	0,047676514	M
18	0,104844647	M
11	0,074932771	M
13	0,083738283	M
3	0,027775169	M
14	0,088273263	M
8	0,05704244	M
10	0,063050346	M
12	0,080639619	B

Hasil akhir euclidean suku Z

M = 9

B = 0

Setelah ketiga sukunya dihitung jarak euclidean maka didapatkan hasil sebagai berikut : Total m (*malignant*) : 25

Total b (*benign*) : 2

Maka hasil akhir diagnosa pada data testing seperti pada tabel berikut :

Tabel 7. Tabel Hasil Akhir Diagnosa dengan Excel

No	ID	Diagnosa	Rata-rata Radius (X1)	Rata-rata Texture (X2)	Rata-rata Perimeter (Y1)	Rata-rata Area (Y2)	Rata-rata Kelancaran (Z1)	Rata-rata Kekompakan (Z2)
1	87880	M	13,81	23,75	91,56	597,8	0,1323	0,1768

3.3.2 Hitung Otomatis

Hasil hitung otomatis dengan RapidMiner didapatkan hasil seperti berikut :

Nilai k kami tetapkan sebesar 9 lalu kami hitung nilai setiap atribut dari data testing dengan setiap atribut dari data training maka didapatkan yang terdekat sebagai berikut :

Tabel 8. Tabel Hasil Akhir Diagnosa dengan RapidMiner

Row No.	ID	Diagnosa	prediction(Diagnosa)	confidence(M)	confidence(B)
1	87880	M	M	0.603	0.397

Hasil hitung didapati bahwa nilai *confidence* m lebih besar dari b dengan nilai k = 9. Maka hasil diagnosa akhirnya berupa M (*Malignant*).

3.3.2.1 Tingkat Akurasi Menggunakan RapidMiner

Table 9. Tabel Tingkat Akurasi dengan RapidMiner

accuracy: 100.00%			
	true M	true B	class precision
pred. M	1	0	100.00%
pred. B	0	0	0.00%
class recall	100.00%	0.00%	

IV. KESIMPULAN

Berdasarkan perhitungan diatas menggunakan perhitungan manual dengan excel maupun dengan perhitungan otomatis menggunakan rapidminer didapati hasil bahwa diagnosa dari data testing terhadap data training adalah M (*malignant*) dengan nilai akurasi pada rapidminner sebesar 100%. Maka dari itu dapat disimpulkan bahwa hasil prakiraan menggunakan metode k-NN cukup dapat diandalkan sebagai media pendukung pengambil keputusan.

REFERENSI

- [1] N. P. Adimuntja, M. A. Nurdin, and Z. F. Ahmad, "Determinant of Sadari Behavior in the Early Detection Effort of Breast Cancer Among Female Students in the Public Health Faculty of Cenderawasih University," *Jambura J. Heal. Sci. Res.*, vol. 4, no. 2, pp. 574–586, 2022, doi: 10.35971/jjhsr.v4i2.13998.
- [2] E. Marfianti, "Peningkatan Pengetahuan Kanker Payudara dan Ketrampilan Periksa Payudara Sendiri (SADARI) untuk Deteksi Dini Kanker Payudara di Semutan Jatimulyo Dlingo," *J. Abdimas Madani dan Lestari*, vol. 3, no. 1, pp. 25–31, 2021, doi: 10.20885/jamali.vol3.iss1.art4.
- [3] Yunitasari, H. S. Hopipah, and R. Mayasari, "Optimasi Backward Elimination untuk Klasifikasi Kepuasan Pelanggan Menggunakan Algoritme k-nearest neighbor (k-NN) and Naive Bayes," *Technomedia J.*, vol. 6, no. 1, pp. 99–110, 2021, doi: 10.33050/tmj.v6i1.1531.

- [4] D. Random *et al.*, "Komparasi Performansi Algoritma Pengklasifikasi KNN , Bagging," pp. 367–372, 2021.
- [5] I. P. Sari and I. H. Batubara, "Perancangan Sistem Informasi Laporan Keuangan Pada Apotek Menggunakan Algoritma K-NN," *Semin. Nas. Teknol. Edukasi ...*, pp. 689–695, 2021.
- [6] Y. Miftahuddin, S. Umaroh, and F. R. Karim, "Perbandingan Metode Perhitungan Jarak Euclidean, Haversine, Dan Manhattan Dalam Penentuan Posisi Karyawan," *J. Tekno Insentif*, vol. 14, no. 2, pp. 69–77, 2020, doi: 10.36787/jti.v14i2.270.
- [7] Y. N. Gulo, "Penerapan Algoritma Hamming Distance Untuk Pencarian Teks Pada Aplikasi Ensiklopedia Indonesia," *JoGTC J. Glob. Tecnol. Comput.*, vol. 1, no. 2, pp. 50–54, 2022.
- [8] E. Wahyudi and N. I. Pradasari, "Case Based Reasoning untuk Diagnosis Penyakit Jantung Menggunakan Metode Minkowski Distance," *INTECOMS J. Inf. Technol. Comput. Sci.*, vol. 1, no. 1, pp. 119–123, 2018, doi: 10.31539/intecom.v1i1.170.
- [9] T. H. Simanjuntak, W. F. Mahmudy, and Sutrisno Sutrisno, "Implementasi Modified K-Nearest Neighbor Dengan Otomatisasi Nilai K Pada Pengklasifikasian Penyakit Tanaman Kedelai," *Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, No.2, no. 2, pp. 75–79, 2017, [Online]. Available: <http://jptiik.ub.ac.id/index.php/jptiik/article/view/15/21>
- [10] D. Darwis, N. Siskawati, and Z. Abidin, "Penerapan Algoritma Naive Bayes Untuk Analisis Sentimen Review Data Twitter Bmkg Nasional," *J. Tekno Kompak*, vol. 15, no. 1, p. 131, 2021, doi: 10.33365/jtk.v15i1.744.
- [11] W. Musu, A. Ibrahim, and Heriadi, "Pengaruh Komposisi Data Training dan Testing terhadap Akurasi Algoritma C4 . 5," *Pros. Semin. Ilm. Sist. Inf. Dan Teknol. Inf.*, vol. X, no. 1, pp. 186–195, 2021.