Terbit *online* pada laman web jurnal : https://ojisnu.nuponorogo.or.id/index.php/ijitech/

Indonesian Journal Of Information Technology

| ISSN (Print) 0000-0000 | ISSN (Online) 3047-5511 |

Klasifikasi Kanker Payudara Menggunakan Data Mining Tool (WEKA)

Krisna Nuresa Qodri *
Program Studi Teknologi Informasi, Universitas Muhammadiyah Klaten, Indonesia
* *krisna@umkla.ac.id*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 20 April 2025
Revisi Akhir: 27 Mei 2025
Diterbitkan *Online*: 26 Juni 2025

KATA KUNCI

kanker; data mining; klasifikasi; naïve bayes; c4.5; support vector machine.

KORESPONDENSI

Telepon: +6285176962114
E-mail: krisna@umkla.ac.id

ABSTRACT

Cancer is one of the non-communicable diseases that is currently still a serious public health problem in the world. Cancer can be divided into two types, which is benign cancer and malignant cancer. Data mining has begun to be applied in various fields, one of which is in the field of health data. By exploring information or knowledge in data, it allows health facilities to improve care for cancer patients, especially breast cancer. This study focuses on classifying the type of cancer suffered by patients. The algorithms used include naïve bayes, C4.5, and support vector machine. The results of testing the three algorithms found that the support vector machine algorithm obtained the highest accuracy, which was 96.9957%.

1. PENDAHULUAN

Kanker merupakan salah satu penyakit tidak menular atau penyakit non infeksi yang hingga saat ini masih menjadi masalah kesehatan masyarakat serius di dunia. WHO melaporkan bahwa setiap tahunnya terdapat 6,25 juta penderita kanker, 2/3 diantaranya terjadi di negara yang sedang berkembang [1]. Apabila tidak dikendalikan, diperkirakan 26 juta orang akan menderita kanker dan 17 juta orang meninggal karena kanker pada tahun 2030. Kejadian tersebut akan terjadi lebih pesat di negara miskin dan berkembang [2]. Kanker dapat menyerang semua kelompok umur, strata sosial ekonomi dan strata pendidikan dari strata pendidikan rendah hingga tinggi [3].

Kanker payudara merupakan salah satu jenis kanker yang banyak diderita oleh kaum wanita waluapun dapat juga ditemukan pada kaum pria. Sebanyak 16% kematian akibat kanker pada wanita dewasa disebabkan oleh kanker payudara [4]. Pada seluruh dunia terdapat lebih dari 1,1 juta kasus baru per tahun pada perempuan yang didiagnosis kanker payudara dan 410.000 perempuan meninggal akibat penyakit tersebut [5].

Tingkat kelangsungan hidup penderita kanker payudara sangat dipengaruhi pada tangkat keganasan kanker pada saat didiagnosis [6]. Oleh sebab itu, diagnosis dini diperlukan untuk menyediakan perawatan yang tepat untuk pasien dan untuk mengurangi tingkat morbiditas dan mortalitas. Diagnosa yang tepat diperlukan untuk berbagai jenis kanker penting bagi dokter untuk membantu mereka mengidentifikasi dan memilih pengobatan yang tepat [7].

Pada umumnya, kanker payudara diobati dengan operasi, yang mungkin diikuti oleh radiasi, kemoterapi, dan terapi hormon [8]. Pada awal pengobatan pada penderita kanker payudara, penyakit tersebut dapat kambuh setiap saat. Namun, sebagian besar kasus kekambuhan cenderung terjadi dalam 5 tahun pertama setelah perawatan [9]. Kelenjar getah bening, tulang, paru-paru, hati, dan otak adalah beberapa bagian umum kekambuhan di luar wilayah payudara [10].

Penggalian data atau data mining merupakan proses untuk mengidentifikasi dan mengekstrasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Proses data mining menggunakan statistik, matematika, kecerdasan buatan dan machine learning [11].

Salah satu pendekatan yang dapat dilakukan adalah dengan menggunakan *support vector machine* (svm), C4.5, naïve bayes untuk melakukan klasifikasi dalam menentukan tingkatan kanker antara kanker jinak atau ganas. Penelitian ini mengidentifikasi faktor-faktor yang terkait untuk menentukan jenis kanker. Dalam penelitian ini beberapa akurasi dari beberapa algoritma juga dibandingkan. Tujuan utama dari penelitian ini adalah untuk memprediksi jenis kanker payudara berdasarkan jenisnya.

2. TINJAUAN PUSTAKA

Banyak penelitian mengenai data mining telah dilakukan. Berikut rangkuman mengenai beberapa penelitian mengenai kanker payudara menggunakan teknik data mining. Jacob et al. [12] telah membandingkan beberapa bebrapa algoritma klasifikasi dengan menggunakan dataset Wisconsin Breast Cancer diagnosis. Hasil dari penelitian mereka menunjukkan bahwa algoritma klasifikasi *Random Tree* and C4.5 menghasilkan akurasi 100%. Namun mereka menggunakan atribut 'Waktu' (*Time to recur/ Disease-free Survival*) bersama dengan parameter lain untuk memprediksi hasil kekambuhan atau tidak kambuhnya kanker payudara di antara pasien. Dalam makalah ini, atribut 'Waktu' tidak dapat diandalkan untuk memprediksi tingkatan penyakit.

Delen et al. [13] menggunakan data SEER (antara periode 1973-2000 dengan sebanyak 202,932 data) data kanker payudara untuk memprediksi tingkat harapan hidup pasien dengan menggunakan metode 10 *fold cross validation*. Hasilnya menunjukkan bahwa algoritma decision tree (C5) adalah algoritma terbaik dengan akurasi 93.6%, algoritma artificial neural network (ANN) juga menghasilkan akurasi yang baik dengan akurasi sebesar 91.2%. Model *logistic regression* tidak cukup baik dengan hanya mendapatkan akurasi sebesar 89.2%.

Chih-Lin Chi et al. [14] menggunakan model ANN untuk dua dataset Breast Cancer Prognosis. Mereka memperkirakan kemungkinan rekurensi kanker payudara dan mengelompokkan pasien dengan prognosis yang baik (> 5 tahun) dan buruk (<5 tahun).

Falk et al. [15] menggunakan model Gaussian Mixture Regression (GMR) pada dataset WPBC telah menyimpulkan bahwa kinerja GMR lebih baik daripada kinerja dari *Classification and Regression Trees* (CART) dalam memprediksi kekambuhan kanker payudara pada pasien. Pendharkar et al. [16] menggunakan beberapa algoritma data mining untuk menemukan pola dalam kanker payudara. Penelitian yang mereka lakukan menunjukkan bahwa data mining dapat digunakan dalam menemukan pola yang sama dalam kasus kanker payudara, yang dapat sangat membantu dalam deteksi dini dan pencegahan penyakit ini.

2.1. Naïve Bayes

Naive Bayes adalah algoritma klasifikasi untuk klasifikasi biner (dua kelas) dan multi-kelas. Teori Bayes dirumuskan (1):

$$P(h|d) = \frac{(P(d|h)*P(h)}{P(d)} \quad (1)$$

Dimana:

- P(h|d) adalah kemungkinan hipotesis h diberikan data d. Ini disebut probabilitas posterior.
- P(d|h) adalah probabilitas data d mengingat bahwa hipotesis h benar.
- P(h) adalah kemungkinan hipotesis h menjadi benar (terlepas dari datanya). Ini disebut probabilitas sebelumnya dari h.
- P(d) adalah kemungkinan data (terlepas dari hipotesis).

Setelah menghitung probabilitas posterior, pilih hipotesis dengan probabilitas tertinggi. Ini adalah kemungkinan hipotesis maksimum (MAP). Ini bisa ditulis sebagai (2):

$$MAP\ (h) = \max \frac{(P(d|h)*P(h)}{P(d)} \quad (2)$$

P (d) adalah istilah normalisasi yang memungkinkan kita untuk menghitung probabilitas. Jika ada jumlah genap di setiap kelas dalam data pelatihan, maka probabilitas masing-masing kelas (misalnya P (h)) akan sama. Sekali lagi, ini akan menjadi istilah konstan dalam persamaan dan menjatuhkannya sehingga berakhir dengan (3):

$$MAP\ (h) = (\max P(d|h)) \quad (3)$$

2.2. C4.5

Algoritma ini adalah penerus ID3 yang dikembangkan oleh Quinlan Ross pada tahun 1993. Algrotima tersebut juga dikenal sebagai algoritma J48. Algoritma tersebut juga berdasarkan algoritma Hunt. Algoritma tersebut secara serial diimplementasikan seperti ID3. Dengan menggunakan algoritma ini, pemangkasan dapat dilakukan yang menggantikan *node* internal dengan simpul daun sehingga mengurangi tingkat kesalahan seperti ID3. C4.5 menangani baik atribut kategoris dan kontinu untuk membangun *decision tree*. Untuk menangani atribut kontinyu, C4.5 membagi nilai atribut menjadi dua partisi berdasarkan ambang yang dipilih sehingga semua nilai di atas ambang sebagai satu anak dan sisanya sebagai anak lain. Algoritma tersebut juga dapat menangani nilai-nilai atribut yang hilang. C4.5 menggunakan metode *gain ratio* untuk mengevaluasi atribut pemecahan yaitu untuk membangun decision tree [17]. Ini menghilangkan bias dari perolehan informasi ketika ada banyak nilai hasil dari suatu atribut.

C4.5 digunakan dalam masalah klasifikasi dan itu adalah algoritma yang paling banyak digunakan untuk membangun DT. Algoritma ini dapat digunakan untuk membangun pohon keputusan yang lebih kecil atau lebih besar dan lebih akurat dan algoritma ini cukup efisien waktu. Algoritma ini digunakan untuk menangani atribut berkelanjutan, misaslnya suhu. C4.5 meningkatkan efisiensi komputasi.

2.3. Support Vector Machine

Support vector machine [SVM] adalah algoritma pelatihan untuk aturan klasifikasi dari kumpulan data yang melatih penggolong; kemudian digunakan untuk memprediksi kelas sampel baru. didasarkan pada konsep bidang keputusan yang menentukan batas keputusan dan titik untuk membentuk batas keputusan antara kelas yang disebut ancaman vektor dukungan sebagai parameter. SVM didasarkan pada algoritma pembelajaran mesin, yang diciptakan oleh vapnik pada tahun 1960-an. dan struktur prinsip minimalisasi risiko untuk mencegah *over fitting*. Ada 2 implementasi kunci dari teknik SVM: pemrograman matematis dan fungsi kernel [18]. SVM menemukan hyperplane optimal antara titik data dari kelas yang berbeda dalam ruang dimensi yang tinggi.

3. METODOLOGI

3.1. Data

Dataset yang akan digunakan dalam penelitian ini adalah datset Wisconsin breast cancer. Dataset tersebut memiliki atribut sebanyak 10, antara lain: *Clump Thickness, Cell Size Uniformity, Cell Shape Uniformity, Marginal Adhesion, Single Epi Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses integer*, dan *Class* { benign, malignant}. Dataset tersebut terdiri dari 699 data.

3.2. Prediction Model

Data mining adalah proses mengekstraksi pola dan pengetahuan yang menarik dari data. Pada penelitian ini berfokus pada penggunaan model klasifikasi untuk memprediksi jenis kanker payudara. Penjelasan singkat tentang algoritma ini dan penerapan spesifiknya untuk penelitian ini adalah sebagai berikut.

3.2.1. Algoritma klasifikasi

Teknik-teknik machine learning dapat digunakan untuk mengklasifikasikan objek-objek yang berbeda berdasarkan kumpulan data pelatihan yang nilai hasilnya diketahui. Pada penelitian ini menggunakan tiga algoritma klasifikasi, antara lain SVM, Naive Bayes and C5.0. pada SVM (Support Vector Machines), data pertama kali dikonversi ke satu set poin dan kemudian diklasifikasikan ke dalam kelas yang dapat dipisahkan secara linier. Model Naive Bayes bekerja dengan memperkirakan probabilitas kumpulan data yang dapat dimasukkan ke kelas menggunakan aturan Bayes. algoritma C5.0 adalah decision tree yang secara rekursif memisahkan pengamatan di cabang untuk membangun sebuah pohon dengan tujuan meningkatkan akurasi prediksi. Ini adalah versi perbaikan dari algoritma C4.5 dan ID3 [19]. Ini juga menawarkan metode penguat yang kuat untuk meningkatkan akurasi algoritma klasifikasi ini [20].

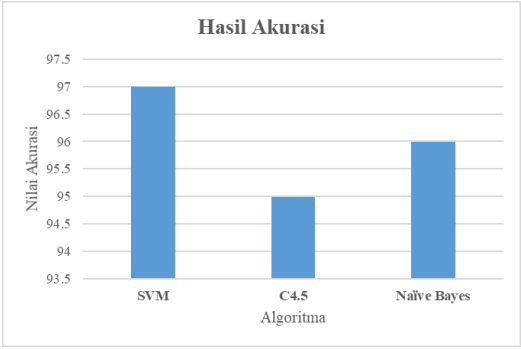
3.3. WEKA

WEKA adalah sistem data mining yang dikembangkan oleh University of Waikato di Selandia Baru yang mengimplementasikan algoritma data mining. WEKA adalah kumpulan algoritma *machine learning* untuk tugas-tugas penambangan data. Algoritma yang akan digunakan akan langsung diterapkan ke dataset yang akan digunakan.

4. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan mode pengujian *cross-validation*. *Cross-validation* adalah teknik untuk menilai atau memvalidasi keakuratan model yang dibangun pada dataset. Mode tes ini mengevaluasi algoritma menggunakan folds value. Penelitian ini menggunakan validasi K-Fold, yang membagi dataset ke sejumlah partisi K secara acak untuk diuji. Nilai dari folds yang digunakan adalah 10.

Hasil dari pengujian menggunakan metode naïve bayes, C4.5 dan *Support Vector Machine* dapat dilihat pada gambar 1. Correctly Classified Instance (CCI) adalah persentase data yang diklasifikasi dengan benar. Berbeda dengan CCI, *Incorrect Classified* adalah nilai persentase data yang diklasifikasikan dalam kelas yang salah. *Precision* adalah nilai pasti dari informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. *Recall* adalah tingkat keberhasilan sistem dalam menemukan kembali suatu informasi. *F-Measure* adalah nilai harmonik atau nilai rata-rata dari presisi dan nilai *recall*. Untuk hasil yang lebih lengkap dapat dilihat pada tabel 1.



Gambar 1. Hasil Klasifikasi

Gambar 1 adalah merupakan hasil klasifikasi dengan menggunakan tiga algoritma yaitu algoritma support vector machine, C4.5 dan naïve bayes. Hasil yang diperoleh dengan menggunakan svm menunjukkan bahwa sebanyak 446 data diklasifikasikan sebagai kanker jinak dan 232 diklasifikasikan sebagai kanker ganas. Pada algoritma naïve bayes sebanyak 436 diklasifikasikan sebagai kanker jinak dan 235 diklasifikasikan sebagai kanker ganas. Algoritma C4.5 memberikan hasil, sebanyak 438 diklasifikasikan sebagai kanker jinak dan 226 diklasifikasikan sebagai kanker ganas.

Tabel 1. Hasil Performa Algoritma

		Classifier		
		SVM	C4.5	Naïve Bayes
Hasil	CCI	96.9957 %	94.9928 %	95.9943 %
	ICI	3.0043 %	5.0072 %	4.0057 %
	Precision	0.970	0.950	0.962
	Recall	0.970	0.950	0.960
	F-Measure	0.970	0.950	0.960
	ROC Area	0.968	0.971	0.986

Hasil dari tabel 1 memperlihatkan bahwa algoritma SVM memperoleh hasil *correctly classified instances* (CCI) sebesar 96.9957 %, *incorrectly classified instances* (ICI) sebesar 3.0043 %, *precision* sebesar 0.970, *recall* sebesar 0.970, *f-measure* sebesar 0.970 dan *receiver operating characteristic curve* (ROC) *area* sebesar 0.968. Algoritma C.45 memperoleh hasil *correctly classified instances* (CCI) sebesar 94.9928 %, *incorrectly classified instances* (ICI) sebesar 5.0072 %, *precision* sebesar 0.950, *recall* sebesar 0.950, *f-measure* sebesar 0.950 dan *receiver operating characteristic curve* (ROC) *area* sebesar 0.971. Algoritma naïve bayes memperoleh hasil *correctly classified instances* (CCI) sebesar 95.9943 %, *incorrectly classified instances* (ICI) sebesar 4.0057 %, *precision* sebesar 0.962, *recall* sebesar 0.960, *f-measure* sebesar 0.950 dan *receiver operating characteristic curve* (ROC) *area* sebesar 0.986.

Hasil dari ketiga algoritma yang digunakan untuk melakukan pengujian diketahui bahwa SVM adalah algoritma yang memiliki akurasi tertinggi dibandingkan dengan dua

algoritma yang lainnya. SVM mengklasifikasikan dengan benar 446 data dan 232 data di kelas yang salah.

5. KESIMPULAN DAN SARAN

Pada penelitian ini, algoritma klasifikasi Naive Bayes, C4.5, dan *support vector machine* digunakan untuk mengklasifikasi jenis dari pernyakit kanker payudara. Hasil dari pengujian menunjukkan bahwa algoritma naïve bayes memperoleh akurasi sebesar 95.9943 %, algoritma C4.5 memperoleh akurasi sebesar 94.9928 %, dan algoritma *support vector machine* memperoleh akurasi sebesar 96.9957 %. Hasil dari pengujian yang dilakukan menggunakan ketiga algoritma tersebut diketahui bahwa algoritma *support vector machine* memperoleh akurasi tertinggi dengan akurasi sebanyak 96.9957 %.

DAFTAR PUSTAKA

[1] M. Nadjib Bustan, Epidemiologi Penyakit Tidak Menular. Rineka Cipta, 2007.

[2] T. F. Prastiwi, "Kualitas Hidup Penderita Kanker," J. Psychol. Univ. Negeri Semarang, vol. 1, no. 1, pp. 21–27, 2013.

[3] Kemenkes, "Penderita Kanker Diperkirakan Menjadi Penyebab Utama Beban Ekonomi Terus Meningkat," 2012.

[4] W. H. Organization, "World Health Statistics 2008.," 2008.

[5] R. Ratnawati, H. Al Rasyid, and Y. D. Maharani, "Mekanisme Koping Perempuan Survivor Kanker Payudara Dalam Mempertahankan Kualitas Hidup (Kualitatif Tentang Kualitas Hidup Perempuan Survivor Kanker Payudara Of Kota Malang)," 2013.

[6] V. Beral et al., "Screening for Breast Cancer in England: Past and Future," NHSBSP Publ., vol. 61, pp. 30–32, 2006.

[7] C. Chin-Hsing, L. Ann-Shu, L. Jiann-Der, and W. H. Yang, "3D image reconstruction of bladder by nonlinear interpolation," Math. Comput. Model., vol. 22, no. 8, pp. 61–72, 1995.

[8] R. Lin and P.Tripuraneni, "Radiation therapy in early-stage invasive breast cancer," Indian J Surg Oncol, vol. 2, pp. 101–111, 2011.

[9] T.Saphner, D.C.Tormey, and R.Gray, "Annual hazard rates of recurrence for breast cancer after primary therapy," J Clin Oncol, vol. 14, pp. 2738–2746, 1996.

[10] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E.Ward, and D. Forman, "Global cancer statistics," CA Cancer J Clin, vol. 61, pp. 69–90, 2011.

[11] E. Turban, R. Sharda, D. Delen, and D. King, Introduction to business intelligence. Business Intelligence: A Managerial Approach. 2011.

[12] S. G. J. Ramani and R. Geetha, "Efficient Classifier for Classification of Prognosis Breast Cancer Data Through Data Mining Techniques," in Proceedings of the World Congress on Engineering and Computer Science 2012, 2012.

[13] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artif. Intell. Med., vol. 34, pp. 113–127, 2004.

[14] C. L. Chi, W. N. Street, and W. H. Wolberg, "Application of artificial neural network-based survival analysis on two breast cancer datasets," Am. Med. Informatics Assoc. Annu. Symp., pp. 130–134, 2007.

[15] T. H. Falk, H. Shatkay, and W.-Y. Chan, "Breast cancer prognosis via gaussian mixture regression," Can. Conf. Electr. Comput. Eng., 2006.

[16] P. C. Pendharkar, J. A. Rodger, G. J. Yaverbaum, N. Herman, and M. Benner, "Association statistical mathematical and neural approaches for mining breast cancer patterns," Exper. Syst. with Appl., vol. 17, pp. 223–232, 1999.

[17] J.R. QUINLAN, "Induction of Decision Trees," Res. Dev. Expert Syst. XV, no. Chapter 2, pp. 15–26, 1999.

[18] J. G., W. D., H. T., and T. R., "Support Vector Machines," An Introd. to Stat. Learn., pp. 337–372, 2013.

[19] J. R. Quinlan, "Simplifying Decision Trees," Int. J. Man Mach. Stud., vol. 27, pp. 221–234, 1987.

[20] A. Max, S. Weston, and M. Culp, "Package ' C50 ,'" 2018.