Fidelity : Jurnal Teknik Elektro
Vol. 3, No. 3, September 2021, page 63-67

p-ISSN: 2686-3650
e-ISSN: 2686-3642

Teknik Data Mining untuk Prediksi Kanker Payudara yang Efisien

Irfan Solikin
R&D
PT. Bhumi Jati Power
Jakarta, Indonesia
irfan.solikin@pt-bjp.co.id

***Abstrak*—Salah satu keganasan paling umum pada wanita, kanker payudara, juga merupakan salah satu penyebab utama kematian. Menurut Organisasi Kesehatan Dunia, kanker payudara sekarang menjadi keganasan paling umum di antara wanita di seluruh dunia. Untuk menyelamatkan nyawa, identifikasi dini kanker payudara sangat penting. Akurasi klasifikasi dari database Wisconsin Breast Cancer (WBC) digunakan untuk membandingkan berbagai pengklasifikasi Data Mining dalam penelitian ini. Bertujuan untuk akurasi prediksi yang tinggi, pekerjaan ini bermaksud untuk mengembangkan model klasifikasi akurat untuk prediksi kanker payudara yang sepenuhnya menggunakan informasi berharga yang ditemukan dalam data klinis. Berdasarkan data WBC, kami telah menjalankan uji coba. Ini dibagi menjadi dua set: set latihan 499 pasien dan set tes dunia nyata 200. Menggunakan perangkat lunak Weka, eksperimen ini menganalisis enam strategi kategorisasi dan menemukan bahwa Support Vector Machine (SVM) lebih akurat dalam memprediksi masa depan daripada teknik lain yang diuji. Keakuratan beberapa teknologi deteksi kanker payudara sedang diselidiki dan dibandingkan. SVM lebih cocok untuk menangani kesulitan klasifikasi seperti prediksi kanker payudara. Jadi kami menyarankan untuk menerapkan temuan ini pada masalah klasifikasi lainnya juga.**

Kata Kunci—kanker payudara; klasifikasi; Pohon keputusan, Naïve Bayes, MLP, Regresi Logistik SVM, KNN dan weka

I. Pengantar

Data mining, juga dikenal sebagai penemuan pengetahuan dalam database didefinisikan sebagai "ekstraksi informasi implisit, sebelumnya tidak diketahui, dan berpotensi berguna dari data" [8] [16]. Ini mencakup serangkaian proses yang dilakukan secara otomatis, yang tugasnya adalah menemukan dan mengekstrak fitur tersembunyi (seperti: berbagai pola, keteraturan, dan anomali) dari kumpulan data besar. Klasifikasi adalah salah satu masalah yang paling banyak dipelajari dalam pembelajaran mesin dan data mining [8]. Memprediksi hasil penyakit adalah salah satu tugas yang paling menarik dan menantang untuk mengembangkan aplikasi data mining.

Algoritma klasifikasi bertujuan untuk membangun model yang akurat dari kumpulan data pelatihan yang diketahui dan kemudian menggunakan model ini untuk mengkategorikan data baru. Klasifikasi data kanker dapat membantu dalam meramalkan hasil penyakit atau menemukan genetika tumor pada pasien kanker payudara. Pada wanita, kanker payudara adalah salah satu bentuk penyakit yang paling umum. Semua kanker lainnya termasuk, kanker payudara adalah penyebab kematian paling umum bagi wanita. Kanker adalah penyakit yang mengubah sifat-sifat sel-sel tubuh dan mendorong pembentukan sel-sel yang menyimpang. Tumor terbentuk ketika kumpulan sel kanker tumbuh besar secara tidak normal. Telah terjadi peningkatan kasus kanker payudara di seluruh dunia. Wanita mengkhawatirkannya sebagai masalah kesehatan yang signifikan [1]. Ketika datang untuk mencegah kematian terkait kanker payudara, deteksi dini penyakit ini sangat penting. Namun, terapi dini memerlukan deteksi dini kanker payudara. Pendekatan yang akurat dan dapat diandalkan diperlukan untuk membedakan antara tumor payudara jinak dan ganas untuk membuat diagnosis dini. Mendeteksi kanker payudara secara otomatis adalah masalah medis yang parah. Dengan demikian, membangun pendekatan diagnosis yang praktis dan akurat sangat penting. Terutama dalam diagnosis medis, teknologi pembelajaran mesin menjadi semakin populer. Diagnosis medis adalah salah satu aspek yang paling menantang dari penggunaan medis.

Klasifikasi data kanker dapat membantu dalam meramalkan hasil penyakit atau menemukan genetika tumor pada pasien kanker payudara. Dalam ilmu kedokteran, salah satu masalah yang paling menantang adalah menentukan penyakit pasien berdasarkan beberapa tes. Akibatnya, diagnostik medis semakin mengandalkan sistem pengklasifikasi. Berikut ini adalah struktur organisasi makalah: Bagian 2 mencakup pekerjaan terkait. Bagian 3 berfokus pada metode klasifikasi dengan sangat rinci. Hasil dan evaluasi teknik kategorisasi dan hasil akhir disajikan pada Bagian 4. Bagian 5 akan memberikan kesimpulan.

II. Penelitian Terkait

Data SIER digunakan oleh Bellachia et al. [2] untuk memeriksa tiga algoritma prediksi deteksi kanker payudara. Mereka menemukan bahwa algoritma C4.5 melakukan yang terbaik, dengan tingkat akurasi 86,7%. Data SIER telah diproses sebelumnya oleh Delen et al. [6] untuk menghapus data yang berlebihan dan hilang. Analisis mereka terhadap tiga model prediksi menemukan bahwa pohon keputusan (C5) adalah yang paling akurat, dengan tingkat akurasi 93,6 persen pada sampel ketidaksepakatan data SIER. Endo dkk. [3] menggunakan teknik pembelajaran mesin standar untuk memperkirakan tingkat kelangsungan hidup wanita dengan kanker payudara. Persentase yang tinggi dari kasus yang menguntungkan digunakan dalam penelitian ini, berdasarkan data program SIER (18,5 persen). Ketika datang ke presisi, jaringan saraf tiruan dan model pohon

keputusan J48 terikat untuk tempat pertama, sementara regresi logistik memiliki spesifisitas tertinggi. Dalam hal menggabungkan Bagging dan Boosting, Kotsiantis et al. [13] menggunakan pelajar dasar yang berbeda seperti C4.5, Nave Bayes, OneR dan Decision Stump untuk melakukannya. Kumpulan data benchmark repositori pembelajaran mesin UCI telah digunakan untuk menguji algoritme ini.

III. Teknik Mengklasifikasikan

Tujuan utama dari penelitian data mining dan pembelajaran mesin adalah untuk mengembangkan pengklasifikasi yang sangat akurat dan cepat untuk database besar. Salah satu aspek terpenting dari data mining adalah pembuatan sistem klasifikasi yang efektif. Beberapa metode klasifikasi yang paling populer termasuk Pohon Keputusan (pendekatan Naive Bayesian), Jaringan Syaraf Tiruan (Regresi Logistik), SVM (KNN), dan Regresi Logistik (SVM).

A. Decision Tree (J48)

Analisis data dan pembuatan model pohon keputusan adalah dua teknik standar untuk memprediksi hasil dalam penambangan data[10]. Pohon klasifikasi, misalnya, dapat digunakan untuk memprediksi nilai kategoris menggunakan ramalan. Dengan satu cabang dan sub-pohon untuk setiap hasil yang mungkin, pohon keputusan adalah pengklasifikasi yang muncul sebagai struktur pohon. Node adalah node daun yang menunjukkan nilai atribut target atau kelas dari contoh yang diuji, atau node keputusan, yang menentukan beberapa pengujian yang akan dilakukan pada nilai atribut tunggal. Dengan memulai dari akar pohon dan melanjutkan ke simpul daun, Anda dapat menggunakan pohon keputusan untuk menentukan klasifikasi sebuah contoh.

B. Neural Networks

Kompleks, fungsi non-linier dapat dimodelkan menggunakan jaringan saraf. Banyak unit yang saling berhubungan membentuk struktur atau jaringan (neuron buatan). Dimungkinkan untuk menerapkan perhitungan atau fungsi lokal menggunakan unit-unit ini yang terdiri dari properti input/output. Misalnya, jika jumlah input yang tertimbang melebihi ambang tertentu, metode ini menghasilkan output. Neuron lain dalam jaringan dapat menggunakan hasilnya sebagai input, terlepas dari ukurannya. Prosedur ini berulang sampai produk akhir tercapai.

C. Naive Bayes (NB)

Menghasilkan model prediksi statistik dengan cepat dapat dilakukan dengan menggunakan pendekatan Naive Bayes. Teorema Bayesian adalah dasar dari NB. Korelasi kelas-atribut dianalisis menggunakan teknik klasifikasi ini untuk menentukan probabilitas bersyarat untuk hubungan antara nilai-nilai atribut. Data pelatihan kelas kursus digunakan untuk menghitung kemungkinan setiap jenis. P(C=c) disebut sebagai "probabilitas sebelumnya". Algoritme juga mempertimbangkan apakah x menerima c selain probabilitas sebelumnya, karena karakteristik diasumsikan independen. Kemungkinan setiap fitur dikalikan dengan probabilitas ini untuk sampai pada hasil akhir. Distribusi frekuensi set pelatihan dapat digunakan untuk memperkirakan probabilitas.

D. Logistic Regression (LR)

Pemodelan data biner dengan LR dianggap sebagai praktik standar dalam statistik [16]. Alih-alih menggunakan regresi linier, yang menetapkan model linier untuk setiap kelas dan memprediksi kasus yang tidak terlihat berdasarkan sebagian besar model, alternatif yang lebih baik adalah menggunakan regresi multikelas. Sebuah model dibangun selama proses prediksi alih-alih menyatakan perkiraan yang tepat tentang kemungkinan suatu peristiwa akan terjadi. Masalah dua kelas, misalnya, menetapkan kasus ke kelas yang ditentukan sebagai "1" untuk YA dan "0" untuk YA dan "TIDAK" jika probabilitasnya lebih dari 50 persen.

E. Support Vector Machine (SVM)

Analisis subklasifikasi dan regresi dapat dilakukan dengan menggunakan metode pembelajaran terawasi SVM. Ruang multidimensi dapat dibagi menjadi dua kelompok menggunakan SVM, sebuah teknik yang mencoba menemukan pemisah linier (hyper-plane) antara titik-titik data dari setiap kelas. Ketika datang untuk menggunakan SVM, ide-ide teori pembelajaran statistik diterapkan. SVM didasarkan pada gagasan untuk menemukan hyperplane sebagai segmentasi dari dua kelas untuk mengurangi kesalahan klasifikasi. SVM menggunakan vektor pendukung (tupel pelatihan) dan margin untuk menentukan hyperplane (vektor pendukung). Sebuah SVM dapat dilatih menggunakan algoritma SMO, metode yang sederhana dan cepat.

F. K-Nearest Neighbor (KNN)

Kesamaan tersebut digunakan untuk mengklasifikasikan instance pada K-Nearest Neighbor (KNN) [8]. Suatu hal diatur ketika mayoritas tetangganya menyetujuinya. K adalah bilangan bulat positif setiap saat. Ada koleksi barang yang telah ditentukan dari mana tetangga dipilih. Atribut numerik N-dimensi mewakili sampel pelatihan. Satu bagian mewakili ruang dimensi-N. Ruang pola N-dimensi digunakan untuk menyimpan semua sampel pelatihan. Ketika diberikan model yang tidak diketahui, k-nearest neighbor classifier mengeksplorasi ruang pola untuk k sampel pelatihan terdekat. Konsep Euclidean tentang "kedekatan" digunakan untuk menggambarkan kedekatan dua titik. Model baru diberikan klasifikasi paling umum di antara k tetangga terdekatnya. Ketika k=1, sampel pelatihan yang paling sebanding dengan model anonim dalam ruang pola disediakan. IBK adalah pengklasifikasi WEKA untuk pengklasifikasi ini.

IV. Hasil dan Diskusi

Sebanyak 699 catatan pasien termasuk dalam dataset. Validasi crossover 10 kali lipat digunakan untuk memvalidasi hasil prediksi perbandingan enam pendekatan data mining umum, dengan 241 atau 34,5 persen menderita kanker payudara. Sisanya 458, atau 65,5 persen, tidak. Validasi crossover K-fold biasanya digunakan untuk mengurangi kesalahan sampel acak ketika membandingkan akurasi model prediksi yang berbeda. Ada k lipatan dalam dataset, masing-masing dengan jumlah contoh yang sama. Pelatihan dan pengujian diulang k kali, dengan satu lipatan dipilih untuk pengujian lebih lanjut dan sisanya untuk pelatihan lanjutan. Sepuluh lipatan data digunakan dalam penyelidikan ini, dengan satu lipatan digunakan untuk pengujian dan sembilan lainnya digunakan sebagai pelatihan. Sepuluh variabel dalam Tabel 1 merangkum

temuan diagnostik untuk setiap subjek dalam kumpulan data seperti yang disebutkan di atas. Salah satu dari sepuluh faktor tersebut adalah variabel respon yang menunjukkan apakah seorang pasien menderita kanker payudara atau tidak (yaitu, ganas atau jinak). Data pelatihan diambil secara acak dari seluruh dataset dan langsung dimasukkan ke dalam metode penambangan yang diusulkan.

Tabel 1. memberikan informasi atribut.

No.	Atribut	Domain
1	Ketebalan rumpun	1-10
2	Keseragaman ukuran sel	1-10
3	Keseragaman bentuk sel	1-10
4	Adhesi marginal	1-10
5	Ukuran sel epitel tunggal	1-10
6	Bare nuclei	1-10
7	Kromatin hambar	1-10
8	nukleolus normal	1-10
9	Mitosis	1-10
	Class	2 for benign, 4 for malignant

A. Metode Evaluasi

Ketiga metode data mining ini diuji menggunakan toolbox Weka [14]. Ada beberapa alat di Weka, dan semuanya bekerja sama untuk membantu Anda mengklasifikasikan, memprediksi, mengelompokkan, mengaitkan, dan memvisualisasikan data Anda. Kami menggunakan WEKA versi 3.6.9 untuk menilai kinerja dan efektivitas enam model prediksi kanker payudara berbeda yang dikembangkan menggunakan berbagai metode dan teknik lain. Program WEKA memberi para peneliti dan pengembang kerangka kerja yang terdefinisi dengan baik untuk mengembangkan dan mengevaluasi model. Tingkat kesalahan dan waktu komputasi digunakan untuk mengevaluasi kinerja pengklasifikasi. Sensitivitas dan Spesifisitas digunakan untuk memprediksi keakuratan klasifikasi. Setiap waktu komputasi pengklasifikasi dipertimbangkan. Spesifisitas, sensitivitas, dan akurasi keseluruhan adalah metrik yang digunakan untuk evaluasi.

Untuk menghitung sensitivitas atau tingkat positif sebenarnya (TPR), bagi jumlah positif palsu (TP + FN) dengan jumlah total positif palsu (TN + FP). True positive (TP) adalah jumlah sampel positif yang berhasil diprediksi. Negatif palsu (FN) adalah jumlah model positif yang salah ditunjukkan. True negatif (TN) adalah jumlah sampel negatif yang diprediksi dengan benar.

Seringkali, nilai-nilai ini ditampilkan dalam matriks kebingungan, seperti yang terlihat pada Tabel 2. Akhirnya, frekuensi prediksi yang benar dan salah ditampilkan dalam matriks klasifikasi. Ini membandingkan nilai sebenarnya dari kumpulan data uji dengan nilai yang diproyeksikan dari model yang dilatih.

Tabel 2. Matriks Kebingungan. Diprediksi

Sebenarnya	Diprediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

B. Hasil. Tabel 3 menunjukkan matriks konfusi untuk masing-masing teknik Klasifikasi, dan Tabel 4 menunjukkan parameter kinerja (akurasi, sensitivitas, spesifisitas, tingkat kesalahan, dan waktu) yang dihasilkan dari matriks.

Tabel 3. Matriks Data yang Membingungkan dari Pelatihan dan Pengujian

Algoritma	Data pelatihan (499)		
	Hasil yang diinginkan	Hasil Keluaran	
		Benign	Malignant
J48	Benign	308	13
	Malignant	9	169
Naive bayes	Benign	310	11
	Malignant	5	173
MLP	Benign	307	14
	Malignant	12	166
Logistic	Benign	314	7
	Malignant	9	169
SVM(SMO)	Benign	315	6
	Malignant	6	172
KNN(IBK)	Benign	314	7
	Malignant	17	161

Algoritma	Data Pengujian (200)		
	Hasil yang diinginkan	Hasil Keluaran	
		Benign	Malignant
J48	Benign	129	8
	Malignant	8	55
Naive bayes	Benign	128	9
	Malignant	2	61
MLP	Benign	130	7
	Malignant	10	53
Logistic	Benign	131	6

	Malignant	9	54
SVM(SMO)	Benign	131	6
	Malignant	5	58
KNN(IBK)	Benign	130	7
	Malignant	5	58

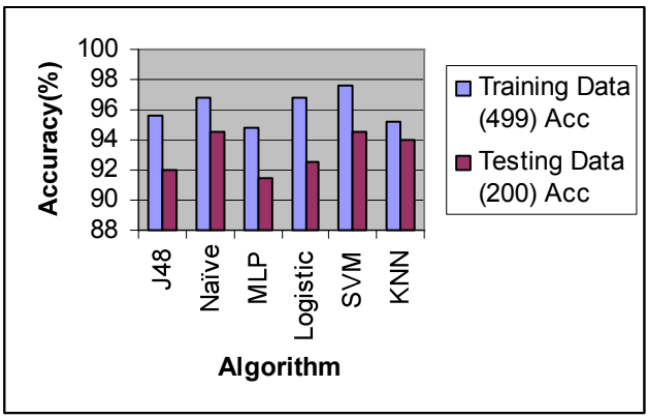
Tabel 4. Kinerja Data Pelatihan dan Pengujian

Algoritma	Data pelatihan (499)				
	Acc	Senst	Spec	Err	Time
J48	95.59	0.96	0.949	4.41	0.09
Naive Bayes	96.79	0.966	0.972	3.21	0.05
MLP	94.78	0.956	0.933	5.22	4.03
Logistic	96.79	0.978	0.949	3.21	0.23
SVM(SMO)	97.59	0.981	0.966	2.41	0.73
KNN(IBK)	95.19	0.978	0.904	4.81	0

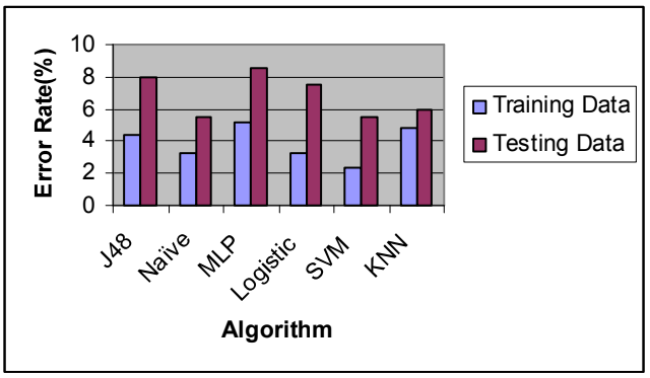
Algoritma	Data Pengujian (200)				
	Acc	Senst	Spec	Err	Time
J48	92	0.942	0.873	8	0.08
Naive Bayes	94.5	0.934	0.968	5.5	0.05
MLP	91.5	0.949	0.841	8.5	1.94
Logistic	92.5	0.956	0.857	7.5	0.23
SVM(SMO)	94.5	0.956	0.921	5.5	0.69
KNN(IBK)	94	0.949	0.921	6	0

Acc - Accuracy, Senst - Sensitivity, Spec - Specificity, Err - Error Rate

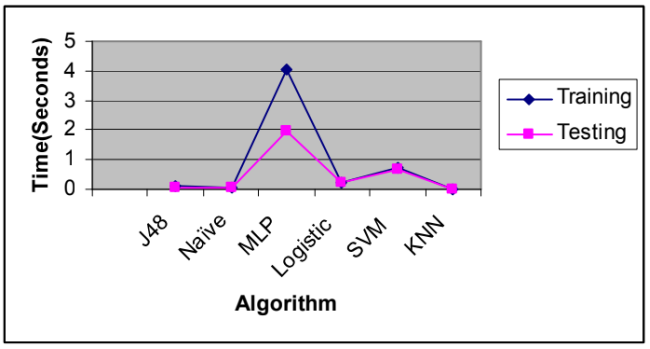
Gambar 1 dan 2 menunjukkan bahwa model Klasifikasi dengan akurasi terbaik (97,59 persen) dan tingkat kesalahan terendah (2,41 persen), seperti yang ditunjukkan pada tabel di atas, adalah SVM.



Gambar 1. Akurasi metode Klasifikasi



Gambar 2. Tingkat kesalahan metode Klasifikasi



Gambar 3: Waktu Eksekusi metode Klasifikasi

V. KESIMPULAN

Keakuratan beberapa strategi klasifikasi dinilai dalam penelitian ini menggunakan algoritma pengklasifikasi tertentu. Salah satu tantangan terbesar bagi peneliti data mining dan machine learning adalah mengembangkan pengklasifikasi yang akurat dan efisien untuk aplikasi medis. SVM mengungguli pengklasifikasi lain dalam hal kinerja. Dengan demikian, SVM menunjukkan hasil dalam data pasien untuk penyakit Kanker Payudara. Oleh karena itu, klasifikasi berbasis penyakit kanker payudara harus menggunakan pengklasifikasi SVM karena lebih akurat, tingkat kesalahan lebih rendah, dan berkinerja lebih baik.

REFERENSI

[1] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

[2] A.Bellachia and E.Guvan,"Predicting breast cancer survivability using data mining techniques", Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining, 2006.

[3] A. Endo, T. Shibata and H. Tanaka (2008), Comparison of seven algorithms to predict breast cancer survival, Biomedical Soft Computingand Human Sciences, vol.13, pp.11-16.

[4] Breast Cancer Wisconsin Data [online]. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data.

[5] Brenner, H., Long-term survival rates of cancer patients achieved by the end of the 20th century: a period analysis. Lancet. 360:1131–1135, 2002.

[6] D. Delen, G. Walker and A. Kadam (2005), Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine, vol.34, pp.113-127.

[7] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005.

[8] J. Han and M. Kamber, Data Mining—Concepts and Technique (The Morgan Kaufmann Series in Data Management Systems), 2nd ed. San Mateo, CA: Morgan Kaufmann, 2006.

[9] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA:Morgan Kaufmann; 1993.

[10] Mitchell, T. M., Machine Learning, McGraw-Hill Science/Engineering/Math, 1997

[11] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[12] Razavi, A. R., Gill, H., Ahlfeldt, H., and Shahsavar, N., Predicting metastasis in breast cancer: comparing a decision tree with domain experts. J. Med. Syst. 31:263–273, 2007.

[13] S.B.Kotsiantis and P.E.Pintelas,"Combining Bagging and Boosting", International Journal of Information and Mathematical Sciences, 1:4 2005.

[14] Vapnik, V. N., The nature of statistical learning theory. Springer, Berlin, 1995.

[15] Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

[16] Witten H.I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques, Second edition, Morgan Kaufmann Publishers, 2005.