KOMPARASI ALGORITMA DATA MINING UNTUK KLASIFIKASI PENYAKIT KANKER PAYUDARA

M. Faizal Kurniawan [(1)] Ivandari [(2)]
STMIK Widya Pratama Pekalongan
Jl. Patriot 25 Pekalongan Telp (0285) 427816
[(1)] Email: faizal@stmik-wp.ac.id
[(2)] Email: ivandarialkaromi@gmail.com

ABSTRAK

Kanker merupakan salah satu penyakit mematikan. Pada tahun 2012 International Agency for Research of Cancer (IARC) mencatat kasus penyakit kanker sebanyak 14.067.894 jiwa dan lebih dari 8,2 juta jiwa meninggal dunia akibat penyakit kanker. Sedangkan dalam 5 tahun terakhir tercatat penderita kanker payudara merupakan yang terbanyak yaitu 19,2% dari keseluruhan kasus. Pencatatan terhadap penyakit kanker banyak dilakukan guna mengantisipasi dan menganalisa pasien sejak dini agar dapat dilakukan pencegahan. Salah satu yang dilakukan adalah dengan menggunakan teknik klasifikasi data mining. Dengan melakukan klasifikasi data mining data lampau yang sebelumnya telah dikumpulkan dapat dijadikan sebuah pengetahuan baru. Beberapa teknik klasifikasi data mining terbukti baik dan menghasilkan akurasi yang tinggi. Dalam penelitian ini akan dilakukan komparasi algoritma K-Nearest Neighbour, Naive Bayes dan Decission Tree C4.5 untuk klasifikasi penyakit kanker payudara. Penelitian ini membuktikan bahwa dari ketiga model algoritma tersebut Naive Bayes memiliki tingkat akurasi terbaik yaitu 95,85%. Sedangkan algoritma KNN memperoleh tingkat akurasi sebesar 94,70% dan Decission Tree C4.5 memperoleh tingkat akurasi sebesar 94,70%..

***Kata Kunci** : Klasifikasi Kanker Payudara, Naive Bayes, KNN, Decission Tree C4.5*

1. Pendahuluan

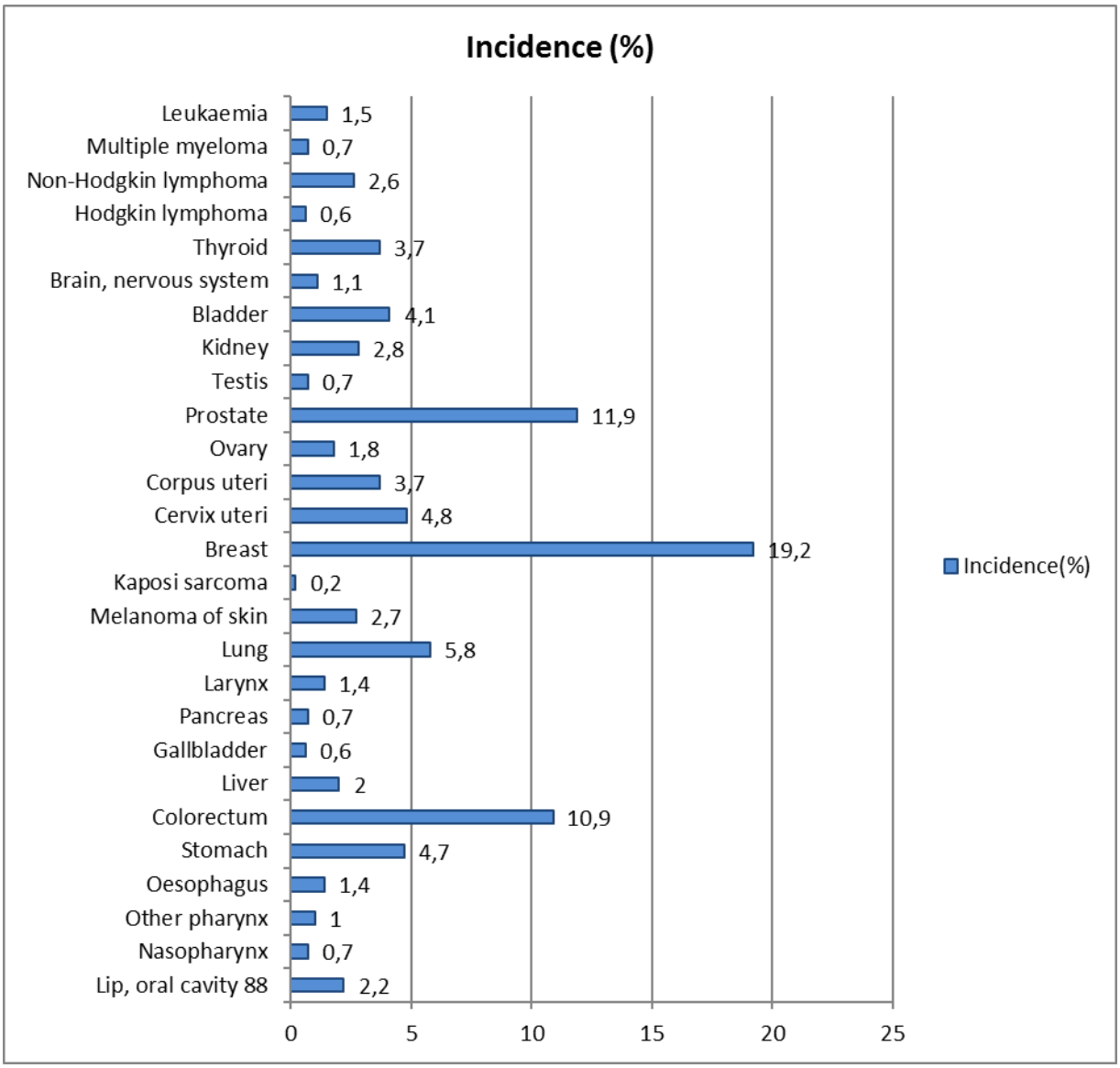
1.1 Latar Belakang

Kanker merupakan salah satu penyakit mematikan. Pada tahun 2012 International Agency for Research of Cancer (IARC) mencatat kasus penyakit kanker sebanyak 14.067.894 jiwa dan lebih dari 8,2 juta jiwa meninggal dunia akibat penyakit kanker. Sedangkan dalam 5 tahun terakhir tercatat penderita kanker payudara merupakan yang terbanyak yaitu 19,2% dari keseluruhan kasus. Gambar 1.1 merupakan grafik presentase penderita jenis kanker dalam 5 tahun terakhir, sedangkan tabel 1.1 merupakan data penderita kanker berdasarkan data yang dihimpun oleh GLOBOCAN International Agency for Research of Cancer (Iarc. 2012).

Pencatatan terhadap penyakit kanker banyak dilakukan guna mengantisipasi dan menganalisa pasien sejak dini agar dapat dilakukan pencegahan. Dengan mengetahui lebih dini penyakit kanker maka penanganannya pun akan semakin mudah karena sel kanker belum berkembang lebih jauh. Beberapa pencatatan menghasilkan data yang valid dan telah teruji secara klinis. Banyaknya record yang ada membuat data menjadi susah dibaca secara manual. Data yang besar bisa saja tidak berarti dan hanya akan menjadi sampah. Data mining dapat menjadikan data yang sebelumnya tidak berarti menjadi sebuah informasi atau pola dengan proses tertentu (I. H. Witten, Frank, and Hall 2011). Data mining juga memungkinkan terbentuknya sebuah model ataupun aturan dalam data yang sebelumnya tidak berharga (Prasetyo 2012). Salah satu peranan utama data mining adalah klasifikasi. Dengan melakukan klasifikasi data mining data lampau yang sebelumnya telah dikumpulkan dapat dijadikan sebuah pengetahuan baru. Beberapa teknik klasifikasi data mining terbukti baik dan menghasilkan akurasi yang tinggi.

Beberapa teknik klasifikasi data mining terbaik antara lain menggunakan algoritma K-Nearest Neighbour, Naive Bayes, serta menggunakan model Decission Tree C4.5 (Wu et al. 2007). Tipe data sangat mempengaruhi performa dan akurasi suatu algoritma (Amancio et al. 2013). Algoritma terbaik untuk sebuah tipe data belum tentu baik untuk tipe data yang lain (Patel, Vala, and Pandya 2014). Bahkan dimungkinkan suatu algoritma yang baik akan menjadi sangat buruk untuk tipe data yang lain

(Ragab et al. 2014) (Ashari, Paryudi, and Tjoa 2013). Penelitian ini membandingkan algoritma K-Nearest Neighbour, Naive Bayes dan Decission Tree C4.5 untuk klasifikasi penyakit kanker payudara.



Gambar 1. Persentase Penderita Jenis Kanker 5 tahun terakhir

Tabel 1. Data Penderita Kanker (Iarc. 2012)

Cancer	Incidence			Mortality			5-year prevalence		
	Number	(%)	ASR (W)	Number	(%)	ASR (W)	Number	(%)	Prop
Lip, oral cavity	300373	2.1	4	145353	1.8	1.9	702149	2.2	13.5
Nasopharynx	86691	0.6	1.2	50831	0.6	0.7	228698	0.7	4.4
Other pharynx	142387	1	1.9	96105	1.2	1.3	309991	1	6
Oesophagus	455784	3.2	5.9	400169	4.9	5	464063	1.4	8.9
Stomach	951594	6.8	12.1	723073	8.8	8.9	1538127	4.7	29.6
Colorectum	1360602	9.7	17.2	693933	8.5	8.4	3543582	11	68.2
Liver	782451	5.6	10.1	745533	9.1	9.5	633170	2	12.2

Gallbladder	178101	1.3	2.2	142823	1.7	1.7	205646	0.6	4
Pancreas	337872	2.4	4.2	330391	4	4.1	211544	0.7	4.1
Larynx	156877	1.1	2.1	83376	1	1.1	441675	1.4	8.5
Lung	1824701	13	23.1	1589925	19	19.7	1893078	5.8	36.5
Melanoma of skin	232130	1.7	3	55488	0.7	0.7	869754	2.7	16.8
Kaposi sarcoma	44247	0.3	0.6	26974	0.3	0.3	80395	0.2	1.5
Breast	**1671149**	**12**	**43.1**	**521907**	**6.4**	**12.9**	**6232108**	**19**	**240**
Cervix uteri	527624	3.8	14	265672	3.2	6.8	1547161	4.8	59.6
Corpus uteri	319605	2.3	8.3	76160	0.9	1.8	1216504	3.7	46.8
Ovary	238719	1.7	6.1	151917	1.9	3.8	586624	1.8	22.6
Prostate	1094916	7.8	30.7	307481	3.7	7.8	3857500	12	149
Testis	55266	0.4	1.5	10351	0.1	0.3	214666	0.7	8.3
Kidney	337860	2.4	4.4	143406	1.7	1.8	906746	2.8	17.5
Bladder	429793	3.1	5.3	165084	2	1.9	1319749	4.1	25.4
Brain, nervous system	256213	1.8	3.4	189382	2.3	2.5	342914	1.1	6.6
Thyroid	298102	2.1	4	39771	0.5	0.5	1206075	3.7	23.2
Hodgkin lymphoma	65950	0.5	0.9	25469	0.3	0.3	188538	0.6	3.6
Non-Hodgkin lymphoma	385741	2.7	5.1	199670	2.4	2.5	832843	2.6	16
Multiple myeloma	114251	0.8	1.5	80019	1	1	229468	0.7	4.4
Leukaemia	351965	2.5	4.7	265471	3.2	3.4	500934	1.5	9.6
All cancers excl. non-melanoma skin cancer	14067894	100	182	8201575	100	102	32455179	100	625

1.2 Landasan Teori

1.2.1 Data Mining

Data mining merupakan sebuah proses ekstraksi untuk mendapatkan suatu informasi yang sebelumnya tidak diketahui dari sebuah data (Ian H Witten, Frank, and Hall 2011). Data mining dapat menganalisa kasus lama untuk menemukan pola dari data dengan menggunakan teknik pengenalan pola seperti statistik dan matematika (Larose 2005). Data Mining atau sering juga disebut Knowledge Discovery in Database (KDD) adalah sebuah bidang ilmu yang banyak membahas tentang pola sebuah data. Serangkaian proses guna mendapatkan pengetahuan atau pola dari kumpulan data disebut dengan data mining (Ian H Witten. Eibe Frank. Mark A Hall 2011). Sebuah data yang besar bisa saja tidak berguna dan hanya akan menjadi sampah bila kita tidak dapat memanfaatkannya. Data mining menjawab masalah ini dengan menganalisa data yang besar tersebut kemudian membuat sebuah aturan, pola, ataupun model tertentu untuk mengenali data baru yang tidak berada dalam baris data yang tersimpan (Prasetyo 2012).

Data mining merupakan kegiatan yang meliputi pengumpulan dan pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data (Santosa 2007).Output dari data mining dapat dipakai untuk memperbaiki pengambilan sebuah keputusan di masa depan. Data ming memiliki kaitan dengan berbagai bidang ilmu yang lain seperti Machine Learning, Statistik, Visualisasi serta database

Walaupun tidak secara jelas membedakan data mining dengan disiplin ilmu lain, tetapi beberapa perbedaan dapat dilihat walau tidak terlalu tegas (Santosa 2007) seperti: Statistik lebih berdasarkan teori, lebih focus pada pengujian hipotesis. Machine Learning lebih bersifat heuristic, focus pada perbaikan performasi dari suatu teknik learning, juga

meliputi real-time learning dan robotic area yang tidak termasuk dalam data mining. Sedangkan data mining sendiri merupakan gabungan teori dan heuristik, focus pada seluruh proses penemuan knowledge / pola termasuk data cleansing, learning dan visualisasi dari hasilnya.

Beberapa peran utama data mining adalah: Estimation, Prediction, Classification, Clustering dan Association. Dari semua peranan data mining tersebut terbagi menjadi 2 berdasarkan metode pembelajarannya(Santosa 2007) yaitu Supervised Learning, Unsupervised Learning.Perbedaan dari kedua metode pembelajaran pada algoritma data mining tersebut adalah jika dalam supervised learning harus memiliki data sampel atau sering disebut juga dengan data training. Sedangkan dalam unsupervised learning tidak membutuhkan data training. Salah satu contoh peran data mining dengan metode supervised learning adalah klasifikasi.

1.2.2 Klasifikasi

Klasifikasi merupakan salah satu peranan utama data mining. Proses klasifikasi adalah proses menghitung data yang ada sebelumnya atau disebut juga data training dengan data baru atau data testing. Proses ini akan menghasilkan kemungkinan dalam data testing. Dalam data klasifikasi dataset yang digunakan harus memiliki label atau atribut tujuan. Beberapa algoritma dapat digunakan untuk perhitungan proses klasifikasi. Algoritma tersebut antara lain KNN, Naive Bayes serta Decission Tree C4.5.

1.2.3 Algoritma K-Nearest Neighbour

Algoritma K-Nearest Neighbour (KNN) merupakan algoritma yang pertama kali dikenalkan pada tahun 1967 (Cover and Hart 1967). Algoritma KNN merupakan pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama, yaitu berdasarkan pada pencocokan bobot dari sejumlah fitur yang ada (Kusrini and Taufiq 2009). K didalam k-NN merupakan jumlah tetangga yang akan diambil untuk menentukan keputusan.

1.2.4 Algoritma Naive Bayes

Metode Naive Bayes atau Naive Bayes Classifier (NBC) adalah salah satu metode yang digunakan untuk klasifikasi teks. Naive Bayes menggunakan teori probabilitas sebagai dasar teori. Bayesian classffiers mempunyai tingkat kecepatan dan akurasi yang tinggi ketikadiaplikasikan dalam database yang besar (Jiawei Han and Micheline Kaber 2006).

1.2.5 Algoritma C4.5

C4.5 Merupakan pengembangan dari algoritma ID3 (Larose 2005) yang dikembangkan oleh Quinlan (Han and Kamber 2006). Algoritma C4.5 ini banyak digunakan peneliti untuk melakukan tugas klasifikasi. Hasil perhitungan dari algoritma C4.5 adalah sebuah pohon keputusan atau decission tree. Algoritma C4.5 menjadi algoritma terpopuler dan terbaik dalam beberapa penelitian (Ragab et al. 2014) (Widiastuti 2007).

1.2.6 Cross Validation

Cross validation merupakan sebuah tindakan pembuktian dari sebuah metode atau performa suatu algoritma. Dalam proses pengujian data mining yang paling banyak digunakan adalah cross validation. Cross validation merupakan pembuktian dengan membagi data sebagian sebagai data training dan sebagian yang lain sebagai data testing dengan komposisi tertentu. Pembagian paling banyak digunakan dalam penelitian klasifikasi data mining adalah membagi data secara acak menjadi 10 bagian. Satu bagian menjadi data testing dan 9 bagian dijadikan data training. Validasi yang seperti ini disebut juga dengan 10folds cross validation (Ian H Witten. Eibe Frank. Mark A Hall 2011).

1.2.7 Confussion Matrix

Confusion Matrix merupakan sebuah hasil evaluasi dari sebuah klasifikasi data mining yang diwujudkan dalam sebuah tabel (Gorunescu 2011). Confusion Matrix berisi tentang perhitungan jumlah objek data testing yang diprediksikan kedalam sebuah kelas dengan klasifikasi yang sebenarnya. Bentuk confusion matrix secara umum dapat dilihat pada tabel 2.6. Dalam confusion matrix terdapat total record yang dipakai dalam dataset baik yang diprediksikan kedalam kelas positif ataupun negatif. Tupel / record dengan prediksi klasifikasi pesitif, prediksi klasifikasi negatif, serta kesalahan dalam klasifikasi dapat terlihat dalam matrix ini.

Tabel 2. *Confussion Matrix* (Gorunescu 2011)

Classification		*Predicted class*	
		Class: YES	*Class*: NO
Observed class	*Class* YES	***A (True Positive)***	***B (False Negative)***
	Class NO	***C (False Positive)***	***D (True Negative)***

2. METODE PENELITIAN

Metode penelitian yang akan digunakan dalam penyelesaian penelitian ini adalah metode eksperimen. Secara garis besar metode ini akan melakukan perhitungan terhadap dataset dengan menggunakan algoritma yang akan dikomparasi yaitu KNN, naive bayes dan C4.5. sub bab berikut akan menjelaskan secara lebih mendalam dan terperinci mengenai tahapan yang akan dilakukan.

2.1 Metode Pengumpulan Data

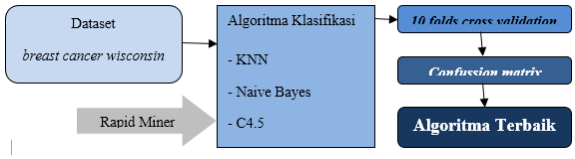
Dataset breast cancer wisconsin yang digunakan memiliki 699 record dengan 11 atribut. 1 atribut menjadi label yaitu atribut kelas dengan isian 2 untuk kanker jinak dan 4 untuk kanker ganas. 10 atribut lainnya antara lain atribut id, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single ephitelial cell size, bare nuclei, bland chromatin, normal nucleoli, serta mitoses.:

Tabel 3 Meta Data *breast cancer wisconsin*

ROLE	NAME	TYPE	STATISTICS	RANGE	MISSINGS
id	Sample code number	integer	avg = 1071704.099 +/- 617095.730	[61634.000 ; 13454352.000]	0.0
label	Class	binominal	mode = 2.0 (458), least = 4.0 (241)	2.0 (458), 4.0 (241)	0.0
regular	Clump Thickness	integer	avg = 4.418 +/- 2.816	[1.000 ; 10.000]	0.0
regular	Uniformity of Cell Size	integer	avg = 3.134 +/- 3.051	[1.000 ; 10.000]	0.0
regular	Uniformity of Cell Shape	integer	avg = 3.207 +/- 2.972	[1.000 ; 10.000]	0.0
regular	Marginal Adhesion	integer	avg = 2.807 +/- 2.855	[1.000 ; 10.000]	0.0
regular	Single Epithelial Cell Size	integer	avg = 3.216 +/- 2.214	[1.000 ; 10.000]	0.0
regular	Bare Nuclei	integer	avg = 3.545 +/- 3.644	[1.000 ; 10.000]	16.0
regular	Bland Chromatin	integer	avg = 3.438 +/- 2.438	[1.000 ; 10.000]	0.0
regular	Normal Nucleoli	integer	avg = 2.867 +/- 3.054	[1.000 ; 10.000]	0.0
regular	Mitoses	integer	avg = 1.589 +/- 1.715	[1.000 ; 10.000]	0.0
id	Sample code number	integer	avg = 1071704.099 +/- 617095.730	[61634.000 ; 13454352.000]	0.0
label	Class	binominal	mode = 2.0 (458), least = 4.0 (241)	2.0 (458), 4.0 (241)	0.0
regular	Clump Thickness	integer	avg = 4.418 +/- 2.816	[1.000 ; 10.000]	0.0
regular	Uniformity of Cell Size	integer	avg = 3.134 +/- 3.051	[1.000 ; 10.000]	0.0

2.2 Kerangka Pemikiran

Penelitian ini menggunakan validasi yaitu cross validation serta pengujian menggunakan *confussion matrix*. Adapun kerangka pemikiran dalam penelitian ini adalah sebagai berikut:



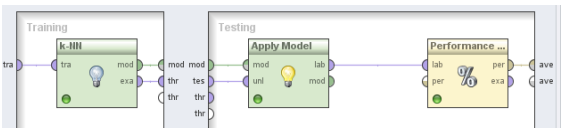
Gambar 2. Kerangka Pemikiran

3. Hasil dan Pembahasan

3.1 Penerapan algoritma Klasifikasi

Dari dataset yang sudah ada, tahapan berikutnya adalah melakukan perhitungan tingkat akurasi untuk setiap algoritma. Tahapan ini menggunakan aplikasi bantu yaitu rapid miner. Proses dalam tahapan ini antara lain dengan validasi menggunakan 10folds cross validation, proses ini berguna untuk membagi dataset menjadi 10 bagian secara acak lalu 9 bagian digunakan untuk data training sedangkan 1 bagian digunakan untuk data testing. Proses ini berulang sampai dengan 10 kali sampai dengan seluruh data mendapatkan porsinya atau mendapat giliran menjadi data testing.

Didalam proses validasi dilakukan perhitungan tingkat akurasi dengan menggunakan tabel kebingungan (confussion matrix). Tabel ini digunakan untuk mengukur tingkat akurasi dari algoritma. Dalam 10 kali percobaan keseluruhan tingkat akurasi dihitung rerata untuk mendapatkan tingkat akurasi dari algoritma tertentu. Gambar 3 merupakan hasil dari perhitungan algoritma KNN dengan menggunakan rapid miner.

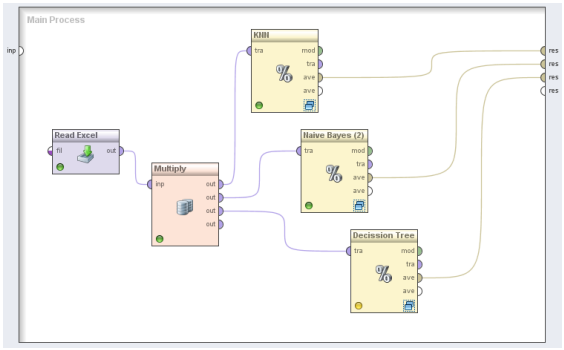


Gambar 3. Model Perhitungan Algoritma

Untuk perhitungan algoritma Naive Baye dan Decission Tree C4.5 dilakukan dengan tahapan yang sama dengan perhitungan KNN. Hanya saja algoritma yang digunakan adalah algoritma Naive Baye dan Decission Tree C4.5.

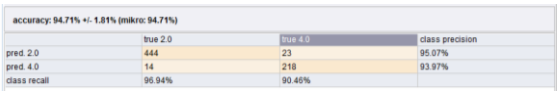
3.2 Hasil Komparasi Algoritma

Hasil komparasi dilakukan dengan menggabungkan semua perhitungan algoritma kedalam satu lembar kerja untuk mengetahui algoritma apa yang memiliki tingkat akurasi terbaik. Dambar 4 merupakan lembar kerja yang digunakan dengan satu dataset yang sama dan tiga algoritma yang akan dikomparasi.

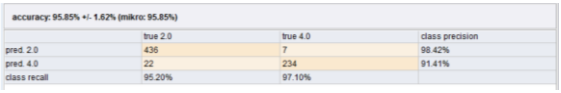


Gambar 4. Lembar kerja komparasi

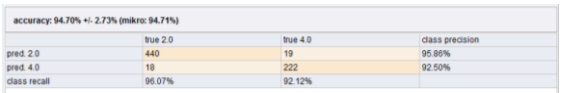
Dari hasil komparasi didapatkan nilai akurasi dari keseluruhan algoritma yang dikomparasi. Gambar 5 merupakan hasil akurasi dari algoritma KNN. Sedangkan gambar 6 merupakan hasil akurasi dari algoritma Naive Bayes dan gambar 7 merupakan hasil akurasi dari algoritma Decission Tree C4.5.



Gambar 5. Tingkat akurasi Algoritma KNN



Gambar 6. Tingkat akurasi Algoritma Naive Bayes



Gambar 7. Tingkat akurasi Algoritma Decission Tree C4.5

Setelah melakukan penelitian dan mendapatkan hasil tingkat akurasi dari keseluruhan algoritma. Tahap berikutnya adalah membandingkan dari ketiga hasil tersebut. Tabel 4 merupakan perbandingan hasil tingkat akurasi dari algoritma KNN, Naive Bayes serta Decission Tree C4.5.

Tabel 4. Hasil Komparasi

Algoritma.	Tingkat Akurasi	Micro
KNN	94,71	94,71
Naive Bayes	95,85	95,85
Decission Tree C4.5	94,70	94,71

4. Kesimpulan dan Saran

4.1 Kesimpulan

Penelitian ini menggunakan dataset publik yaitu data breast cancer wisconsin yang diambil dari UCI repository. Dataset ini sudah terbukti dan banyak digunakan peneliti untuk pemodelan dan aplikasi klasifikasi kanker payudara.

Dalam penelitian ini diketahui bahwa untuk klasifikasi penyakit kanker payudara algoritma Naive Bayes merupakan algoritma terbaik dengan tingkat akurasi sebesar 95,85%. Sedangkan algoritma KNN dan Decission Tree C4.5 hanya memperoleh tingkat akurasi masing masing 94,70% dan 94,71%./7

4.2 Saran

Penelitian ini menggunakan aplikasi bantu rapid miner untuk proses perhitungan dan pembuktian akurasi algoritma. Dalam penelitian berikutnya diharapkan dapat tercipta sebuah aplikasi yang dapat digunakan secara langsung oleh pihak terkait dalam klasifikasi atau deteksi penyakit kanker payudara.

5. Referensi

Amancio, D. R., C. H. Comin, D. Casanova, G. Travieso, O. M. Bruno, F. a. Rodrigues, and L. Da F. Costa. 2013. "A Systematic Comparison of Supervised Classifiers," October. http://arxiv.org/abs/1311.0202v1.

Ashari, Ahmad, Iman Paryudi, and A Min Tjoa. 2013. "Performance Comparison between Naïve Bayes , Decision Tree and K-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool" 4 (11): 33–39.

Christobel, Angeline, and D.r Sivaprakasam. 2011. "An Empirical Comparison of Data Mining Classification Methods" 3 (2): 24–28.

Cover, T M, and P E Hart. 1967. "Nearest Neighbor Pattern Classification" I.

Gorunescu, Florin. 2011. Data Mining: Concept, Models and Techniques. Vol 12. Berlin: Heidelberg: Springer Berlin Heidelberg.

Han, Jiawei, and Micheline Kamber. 2006. "Data Mining: Concepts and Techniques Second Edition" 40 (6). Elsevier: 9823. doi:10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Hastuti, Khafiizh. 2012. "Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif" 2012 (Semantik): 241–49.

Ian H Witten. Eibe Frank. Mark A Hall. 2011. Data Mining 3rd.

Iarc., International Agency for Research on Cancer. World Health Organization. 2012. "GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012." Globocan. doi:10.1002/ijc.27711.

Ivandari. 2014. "Improved Performance Algorithm K-Nearest Neighbor Classification in High Dimension Data." IC Tech IX-April 2: 5–9.

Jiawei Han and Micheline Kaber. 2006. "Data Mining:Concepts and Techniques." University of Illinois at Urbana-Champaign.

Kusrini, Sri Hartati, Retantyo Wardoyo, and Agus Harjoko. 2009. "Perbandingan Metode Nearest Neighbor Dan Algoritma c4.5 Untuk Menganalisis Kemungkinan Pengunduran Diri Calon Mahasiswa Di Stmik Amikom Yogyakarta" 10 (1).

Kusrini, and Luthfi Emha Taufiq. 2009. Algoritma Data Mining. Yogyakarta: Andi Offset.

Larose, Daniel T. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons.

Maimoon, Oded, and Lior Rokach. 2010. Data Mining and Knowledge Discovery Handbook. Vol. 40. Springer. doi:10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Maulina, Inas Ulvy, Mardji, and Edy Santoso. 2015. "Klasifikasi Kanker Payudara Menggunakan Decision Tree Dengan Algoritma Iterative Dichotomizer-3 (ID-3)" 3: 1–8.

Nursela, Dwi Ayu. 2014. "Penerapan Algoritma C4 . 5 Untuk Klasifikasi Tingkat Keganasan Kanker Payudara," 1–5.

Patel, Kanu, Jay Vala, and Jaymit Pandya. 2014. "Comparison of Various Classification Algorithms on Iris Datasets Using WEKA" 1 (1): 1–7.

Prasetyo, Eko. 2012. Data Mining Konsep Dan Aplikasi Menggunakan Matlab. Yogyakarta: Andi Offset.

Pudjianto, Tacbir Hendro, Faiza Renaldi, and Age Teogunadi. 2011. "Penerapan Data Mining Untuk Menganalisa Kemungkinan Pengunduran Diri Calon Mahasiswa Baru."

Rachman, Farizi, and Wulan Purnami. 2012. "Perbandingan Klasifikasi Tingkat Keganasan Breast Cancer Dengan Menggunakan Regresi Logistik Ordinal Dan Support Vector Machine (SVM)" 1 (1).

Ragab, Abdul Hamid M., Amin Y. Noaman, Abdullah S. Al-Ghamdi, and Ayman I. Madbouly. 2014. "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining." Proceedings of the 2014 Workshop on Interaction Design in Educational Environments - IDEE '14. New York, New York, USA: ACM Press, 106–13. doi:10.1145/2643604.2643631.

Santosa, Budi. 2007. Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis. Edisi Pert. Yogyakarta: Graha Ilmu.

Sugianti, Devi. 2012. "Algoritma Bayesian Classification Untuk Memprediksi Heregistrasi Mahasiswa Baru Di STMIK Widya Pratama," no. 2: 1–5.

Susanto, Sani, and Dedi Suryadi. 2010. Pengantar Data Mining: Menggali Pengetahuan Dari Bongkahan Data. Yogyakarta: Andi Offset.

Widiastuti, Dwi. 2007. "Analisa Perbandingan Algoritma SVM, Naïve Bayes, Dan Decission Tree Dalam Mengklasifikasikan Serangan (Attack) Pada Sistem Pendeteksi Intrusi." Jurusan Sistem Informasi Universitas Gunadarma, 1–8.

Witten, I. H., E. Frank, and M. A. Hall. 2011. Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition. Vol. 40. Elsevier. doi:10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.

Witten, Ian H, Eibe Frank, and Mark A. Hall. 2011. Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition. Elsevier.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2007. Top 10 Algorithms in Data Mining. Knowledge and Information Systems. Vol. 14. doi:10.1007/s10115-007-0114-2.