PERBANDINGAN KINERJA ALGORITMA C4.5, NAÏVE BAYES, K-NEAREST NEIGHBOR, LOGISTIC REGRESSION, DAN SUPPORT VECTOR MACHINES UNTUK MENDETEKSI PENYAKIT KANKER PAYUDARA

Taghfirul Azhima Yoga Siswa[1], Prihandoko[2]

[1]Universitas AMIKOM Yogyakarta
[2]Universitas Gunadarma
[1]taghfirul.yoga@yahoo.co.id, [2]prihandoko@gmail.com

Abstrak

Mengevaluasi perbandingan kinerja terbaik metode klasifikasi data mining algoritma *C4.5, Naïve Bayes, K-Nearest Neighbor, Logistic Regression,* dan *Support Vector Machines* untuk mendeteksi kanker payudara menggunakan uji 10 *fold Cross Validation* dengan perbandingan nilai akurasi, *precision,* dan *recall* menggunakan *confusion matrix*. Dataset kanker payudara yang digunakan berjumlah 699 record dengan 11 parameter indikator yang terdiri dari Code *Number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses,* dan *Class* yang didapat dari http://archive.ics.uci.edu. Data diolah menggunakan software Rapid Miner Versi 9. Hasil penelitian ini didapatkan bahwa presentase kinerja masing – masing algoritma klasifikasi yang dianalisis yaitu antara lain Algoritma C4.5 (akurasi 93.70%, *precision* 94.26%, *recall* 87.86%), Algoritma *Naïve Bayes* (akurasi 96.19%, *precision* 92.25%, *recall* 97.50%), Algoritma *K-Nearest Neighbor* (akurasi 95.61%, *precision* 94.99%, *recall* 92.43%), Algoritma *Logistic Regression* (akurasi 96.77%, *precision* 95.93%, *recall* sebesar 94.98%), dan Algoritma Support Vector Machines (akurasi 96.78%, *precision* 94.83%, *recall* sebesar 96.20%). Hasil kinerja terbaik yang diuji menggunakan T-Test didapatkan bahwa algoritma *Logistic Regression* dan *Support Vector Machines* memiliki nilai akurasi tertinggi yang sama yaitu sebesar 0,968.

Kata Kunci: ***Data Mining, C45, Naive Bayes, K-Nearest Neighbor, Logistic Regression, Support Vector Machines***

Abstract

Evaluate the best performance comparison of C4.5, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, and Support Vector Machines classification methods for detecting breast cancer using a 10 fold Cross Validation test by comparing the values of accuracy, precision, and recall using confusion matrix . The breast cancer dataset used was 699 records with 11 indicator parameters consisting of Code Number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and Classes obtained from http://archive.ics.uci.edu. The data was processed using Rapid Miner Version 9 software. The results of this study found that the percentage of performance of each classification algorithm analyzed, that is C4.5 Algorithm (accuracy 93.70%, precision 94.26%, recall 87.86%), Naïve Bayes Algorithm (accuracy 96.19 %, precision 92.25%, recall 97.50%), K-Nearest Neighbor Algorithm (95.61% accuracy, precision 94.99%, recall of 92.43%), Logistic Regression Algorithm (accuracy 96.77%, precision 95.93%, recall 94.98%), and Support Vector Machines algorithm (accuracy 96.78%, precision 94.83%, recall 96.20%). The best performance results tested using T-Test found that the Logistic Regression and Support Vector Machines algorithm has the same highest accuracy value that is equal to 0.968.

Keywords: Data Mining, C45, Naive Bayes, K-Nearest Neighbor, Logistic Regression, Support Vector Machines

1. Pendahuluan

Saat ini kanker payudara menjadi jenis kanker yang sangat menakutkan bagi perempuan diseluruh dunia, hal ini juga berlaku di Indonesia. Kanker payudara adalah tumor ganas yeng terbentuk dari sel - sel payudara yang tumbuh dan berkembang tanpa terkendali sehingga dapat menyebar di antara jaringan atau organ di deket payudara atau ke bagian tubuh lainya.

Berdasarkan data rutin Subdit Kanker Direktorat Penyakit Tidak Menular, Direktorat Jenderal Pengendalian Penyakit dan Penyehatan Lingkungan, Kementerian Kesehatan RI, sampai dengan tahun 2013, program deteksi dini kanker payudara baru

diselenggarakan pada 717 Puskesmas dari total 9.422 Puskesmas di 32 provinsi. Dengan demikian, dapat dilihat bahwa Puskesmas yang memiliki program deteksi dini masih sangat sedikit atau sekitar 7,6%. Estimasi jumlah penderita kanker payudara diketahui bahwa Provinsi Jawa Timur, Jawa Tengah dan Jawa Barat memiliki estimasi jumlah penderita kanker payudara terbesar, sementara itu Provinsi Gorontalo dan Papua Barat memiliki estimasi jumlah penderita terkecil dari seluruh provinsi. Kanker yang diketahui sejak dini memiliki kemungkinan untuk mendapatkan penanganan lebih baik. Oleh karena itu, perlu dilakukan upaya pencegahan untuk meningkatkan kesadaran masyarakat dalam mengenali gejala dan risiko penyakit kanker sehingga dapat menentukan langkah-langkah pencegahan dan deteksi dini yang tepat.

1.1 Kanker Payudara

Kanker adalah kelompok penyakit, dimana sel tubuh berkembang, berubah, dan menduplikasikan diri diluar kendali. Biasanya, nama kanker diberikan berdasarkan bagian tubuh dimana kanker pertama kali tumbuh. Jadi, kanker payudara adalah tumor ganas yang telah berkembang dari sel-sel yang ada di dalam payudara. Kanker payudara merujuk pada pertumbuhan serta perkembangbiakan sel abnormal yang muncul pada jaringan payudara (Chyntia, 2009).

Kanker payudara adalah suatu penyakit dimana terjadi pertumbuhan berlebihan atau perkembangan tidak terkontrol dari sel-sel (jaringan) payudara. Kanker bisa mulai tumbuh di dalam kelenjar susu, saluran susu, jaringan lemak maupun jaringan ikat pada payudara (Rahayu, 1991).

1.2 Data Mining

Data mining, sering disebut juga sebagai Knowledge Discovery in Database (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data-data yang berukuran besar (Santosa, 2007). Ouput atau keluaran dari Data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan. Sehingga istilah pattern recognition sekarang jarang digunakan karena sudah termasuk bagian dari Data mining.

1.3 Metode Klasifikasi

Salah satu bagian penting dalam data mining adalah teknik klasifikasi, yaitu bagaimana mempelajari sekumpulan data sehingga dihasilkan aturan yang bisa mengklasifikasi atau mengenali data-data baru yang belum pernah dipelajari. Klasifikasi dapat didefinisikan sebagai proses untuk menyatakan suatu objek data sebagai salah satu kategori (kelas) yang telah didefinisikan sebelumnya (Zaki et al. 2013). Klasifikasi banyak digunakan dalam berbagai aplikasi, di antaranya adalah deteksi kecurangan (fraud detection), pengelolaan pelanggan, diagnosis medis, prediksi penjualan, dan sebagainya.

Klasifikasi Data mining menurut Vercellis (2009) adalah suatu metode pembelajaran, untuk memprediksi nilai dari sekelompok attribut dalam menggambarkan dan membedakan kelas data atau konsep yang bertujuan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui.

1.4 Algoritma C4.5

Metode C4.5 merupakan salah satu metode klasifikasi menarik yang melibatkan konstruksi pohon keputusan, koleksi node keputusan, terhubung oleh cabang-cabang, memperpanjang bawah dari simpul akar samapai berakhir di node daun. Dimulai dari node root, yang oleh konvensi ditempatkan dibagian atas dari diagram pohon keputusan, atribut diuji pada node keputusan, dengan setiap hasil yang mungkin menghasilkan cabang. Setiap cabang kemudian mengarah ke node lain baik keputusan atau ke node daun untuk mengakhiri (Larose, 2005).

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 (Kusrini & Luthfi, 2009), yaitu :

1 Menyiapkan data training. Data training biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.

2 Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih dengan cara menghitung nilai Gain dari masing-masing atribut, nilai Gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai Gain dari atribut, hitung dahulu nilai entropy yaitu :

$$Entropy(S) = \sum_{i=1}^{n} - pi * \log_2 pi$$

Keterangan :
S : himpunan kasus
A : atribut
n : jumlah partisi S
P_i : proporsi dari S_i terhadap S

3. Kemudian hitung nilai Gain dengan metode *information gain* :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i)$$

Keterangan :
S : himpunan kasus
A : atribut
n : jumlah partisi atribut A
|Si| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

4. Ulangi langkah ke-2 hingga semua tupel terpartisi.
5. Proses partisi pohon keputusan akan berhenti saat:
   a. Semua tupel dalam node N mendapat kelas yang sama.
   b. Tidak ada atribut di dalam tupel yang dipartisi lagi.
   c. Tidak ada tupel di dalam cabang yang kosong.

1.5 Algoritma Naïve Bayes

Klasifikasi *Bayesian* adalah klasifikasi statistik yang bisa memprediksi probabilitas sebuah class. Klasifikasi Bayesian ini dihitung berdasarkan Teorema Bayes. *Teorema Bayes* adalah perhitungan statistik dengan menghitung probabilitas kemiripan kasus lama yang ada dibasis kasus dengan kasus baru. *Teorema Bayes* memiliki tingkat akurasi yang tinggi dan kecepatan yang baik ketika diterapkan pada database yang besar (Han, 2012).

Persamaan dari teorema Bayes adalah sebagai berikut :

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} (1)$$

Keterangan :
X : Kriteria suatu kasus berdasarkan masukan
Ci : Kelas solusi pola ke-i, dimana i adalah jumlah label kelas
P(Ci|X) : Probabilitas kemunculan label kelas Ci dengan kriteria masukan X
P(X|Ci) : Probabilitas kriteria masukan X dengan label kelas Ci
P(Ci) : Probabilitas label kelas Ci

1.6 Algoritma K-Nearest Neighbor

Menurut (Suyanto, 2017) algoritma *K-Nearest Neighbor* (k-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut, Ketepatan algoritma k-NN ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi.

Menurut Kusrini (2009) langkah – langkah algoritma kNN ditunjukkan sebagai berikut:

1. Tentukan nilai latih k, yaitu jumlah tetangga terdekat
2. Menghitung kuadrat jarak euclidian (euclidean distance) masing-masing objek terhadap data sampel yang diberikan

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Dimana
d : jarak kedekatan
X : data training
Y : data testing
n : jumlah attribut individu antara 1 sampai n
f : fungsi similitary atribut I antara kasus x dan kasus y
Wi : bobot yang diberika pada atribut ke-i

Jarak antara objek x dan y didefinisikan sebagai D_{xy}, dimana xi merupakan *record* yang akan diprediksi dan yi merupakan record data pola sedangkan nilai n didefinisikan sebagai jumlah atribut dan nilai I merujuk pada record ke-i

3. Mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak euclid terkecil
4. Mengumpulkan kategori Y (Klasifikasi nearest neighbor)
5. Dengan menggunakan kategori mayoritas, maka dapat diprediksikan nilai query instance yang telah dihitung.

1.7 Algoritma Logistic Regeression

Menurut Santoso (2007) Logistic Regression Regresi logistik (Logistic regression) adalah bagian dari analisis regresi yang digunakan ketika variabel dependen (respon) merupakan variabel dikotomi. Variabel dikotomi biasanya hanya terdiri atas dua nilai yang mewakili kemunculan atau tidak adanya suatu kejadian yang biasanya diberi angka 0 atau 1. Tidak seperti regresi linier biasa, regresi logistik tidak mengasumsikan hubungan antara variabel independen dan dependen secara linier. Regresi logistik merupakan regresi non linier dimana model yang ditentukan akan mengikuti pola kurva linier.

$$Log(P/1\text{-}P) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K$$

Hosmer dan Lemeshow (2000) menjelaskan bahwa model regresi logistik biner dibentuk dengan menyatakan nilai P(Y=1|x) sebagai π (x), yang dinotasikan sebagai berikut :

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}$$

Dimana xi adalah variabel prediktor, i=1,2,…,n dengan n = ukuran sampel, π (x) adalah peluang terjadinya kejadian sukses, β0 adalah konstanta, dan β p adalah nilai koefisien regresi ke-j dengan j=1,2,…, p.

1.8 Algoritma Algoritma Support Vector Machine (SVM)

SVM adalah sebuah algoritma yang diusulkan oleh Vapnik pada tahun 1995. SVM tergolong metode klasifikasi baru dan telah banyak dijadikan metode dalam sejumlah penelitian, seperti pattern recognition, regresi, dan estimasi. SVM menggunakan masukan atau input dari data training untuk menemukan hyperplane yang dapat mengklasifikasikan dua atau lebih tipe data untuk kemudian memproses atribut-atribut dari problem klasifikasi (Nisa, 2013). SVM dapat diterapkan pada data yang bersifat linear maupun non-linear. Untuk kasus klasifikasi, dimana datanya tidak linear dapat menggunakan metode Kernel.

Problem optimasi menggunakan SVM untuk kasus klasifikasi dengan dua kelas dimana data tidak dapat dikelompokkan secara benar dapat dirumuskan sebagai berikut :

$$min\ \frac{1}{2}\ \|\omega\|^2 + C\sum_{i=1}^{\lambda} t_i$$

Subject to

$$y_i(wx_i + b) \geq 1, i = 1, \ldots, l$$

Dimana

xi = data input

yi = output dari data *x*

wib = parameter yang akan dicari nilainya

C = parameter yang akan dicari oleh user

Untuk mengatasi permasalahan yang bersifat tidak linier, dapat digunakan metode kernel. Dengan metode kernel suatu data x diinput space di mapping ke feature space F yang lebih tinggi. Suatu kernel map mengubah problem yang tidak linier menjadi linier dalam space baru. Fungsi kernel yang biasanya dipakai dalam literatur SVM (Haykin,1999) dalam Santosa (2007)

1. Linier : $x^y x$
2. Polynomial : $(x^y x^i + 1)^p$
3. Radial basis function (RBF) :

$$\exp\left(-\frac{1}{2\pi 2}\|x - x_I\|^2\right)$$

4. Tangent hyperbolic (sigmoid) :
   $\tanh((ßx^y x_i + ß_i))$

Pemilihan jenis fungsi kernel yang akan digunakan untuk substitusi dot product di feature space akan sangat bergantung pada data.

1.9 Evaluasi dan Pengujian

Model validasi yang digunakan pada penelitian ini adalah 10 fold cross validation. 10 fold cross validation digunakan untuk mengukur kinerja model prediksi. Setiap dataset secara acak dibagi menjadi 10 bagian dengan ukuran yang sama. Selama 10 kali, 9 bagian untuk melatih model (data training) dan 1 bagian digunakan untuk menguji (data testing) yang lainya setiap kali dilakukan pengujian. Pegukuran pada evaluasi kinerja klasifikasi bertujuan untuk mengetahui seberapa akurat model klasifikasi dalam prediksi kelas dari suatu baris data Han & Kamber (2012).

Bramer (2007) metode *confusion matrix* merepresentasikan hasil evaluasi model dengan menggunakan tabel matriks, jika dataset terdiri dari dua kelas, kelas pertama dianggap positif, dan kelas kedua dianggap negative. Evaluasi menggunakan *confusion matrix* menghasilkan nilai akurasi, presisi, *recall*. Akurasi dalam klasifikasi merupakan presentase ketepatan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi. *Precision* atau *confidence* merupakan proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. *Recall* atau *sensitivity* merupakan proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar.

Tabel 1 Confusion Matrix

Class		Actual	
		True	False
Predic	True	True Positif(TP)	False Negative (FN)
	False	False Positive (FP)	True negative (TN)

True positive (tp) merupakan jumlah *record* positif dalam data set yang diklasifikasikan *positive. True negative* (tn) merupakan jumlah *record negative* dalam data set yang diklasifikasikan *negative. False positive* (fp) merupakan jumlah *record* negatif dalam data set yang diklasifikasikan potitif. *False negative* (fn) merupakan jumlah *record positive* dalam data set yang diklasifikasikan *negative.*

Berikut adalah persamaan model *confusion matrix:*

a. Nilai akurasi (acc) adalah proporsi jumlah prediksi yang benar. Dapat dihitung dengan menggunakan persamaan:

$$akurasi = \frac{tp + tn}{tp + tn + fp + fn}$$

b. *Sensitivity* atau *recall* digunakan untuk membandingkan proporsi tp terhadap tupel yang positif, yang dihitung dengan menggunakan persamaan:

$$Sensitivity = \frac{tp}{tp + fn}$$

c. *Specificity* digunakan untuk membandingkan proporsi tn terhadap tupel yang negatif, yang dihitung dengan menggunakan persamaan:

$$Specificity = \frac{tn}{tn + fp}$$

d. PPV (*positive predictive value)* atau *precision* adalah proporsi kasus dengan hasil diagnosa positif, yang dihitung dengan menggunakan persamaan:

$$PPV = \frac{tp}{tp + fp}$$

e. NPV (*negative predictive value)* adalah proporsi kasus dengan hasil diagnosa negatif, yang dihitung dengan menggunakan persamaan:

$$NPV = \frac{tn}{tn + fn}$$

Lewandowski (2009) hasil akurasi juga dapat dilihat dengan melakukan perbandingan klasifikasi menggunaakan *curva Receiver Operating Characteristic* (ROC) dari hasil *confusion matrix*. ROC menghasilkan dua garis dengan bentuk true positif yang ditandai dengan garis vertical dan false positive yang ditandai dengan garis horiozontal. ROC adalah grafik antara sensitivitas true positive rate pada sumbu X dan sumbu Y. Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix.*

Tingkat akurasi dapat di diagnosa sebagai berikut (Gorunescu, 2011):

a. Akurasi 0.90 – 1.00 = Excellent classification
b. Akurasi 0.80 – 0.90 = Good classification

c. Akurasi 0.70 – 0.80 = Fair classification
d. Akurasi 0.60 – 0.70 = Poor classification
e. Akurasi 0.50 – 0.60 = Failure

1.10 Penelitian Terkait

Penelitian Durairaj & Deepika (2015) membandingkan tiga tradisional model algoritma klasifikasi seperti Naive Bayes, k-NN (lazy classifiers) dan Decision Tree berdasarkan nilai performa akurasi dan time execution pada dataset kanker leukimia yang datasetnya terdiri dari 7.130 atribut dan 72 records. Penelitian ini membuktikan pada algoritma Naive Bayes memiliki nilai performa akurasi yang terbaik yaitu 91,17% daripada model algoritma klasifikasi lainnya yaitu Decision Tree dan K-NN.

Penelitian (Sartika, 2017) membandingkan algoritma klasifikasi Nearest Neighbour, Naive Bayes dan Decision Tree yang digunakan pada studi kasus pengambilan keputusan pemilihan pola pakaian, menyatakan bahwa perbandingan menunjukkan metode Decision Tree memiliki tingkat akurasi tertinggi dibandingkan algoritma Naive Bayes dan Nearest Neighbour yaitu mencapai 75.6%. Algoritma Decision Tree yang digunakan adalah algoritma J48 dengan pruned yang menghasilkan model Decision Tree dengan daun sebanyak 166 dan pohon keputusan yang besarnya 255.

Penelitian Alverina (2018) membandingkan akurasi prediksi kategori Indeks Prestasi (IP) semester pertama mahasiswa Fakultas Teknologi Informasi (FTI) Universitas Kristen Duta Wacana (UKDW). Hasil penelitian ini algoritma C4.5 dan CART memiliki akurasi yang sama untuk memprediksi kategori IP mahasiswa baru pada jalur prestasi (data non numerik), yaitu sebesar 86,86%. Untuk memprediksi kategori IP mahasiswa baru pada jalur non-prestasi (data numerik), algoritma CART memberikan akurasi lebih baik daripada C4.5, yaitu 63,16% berbanding 61,54%.

Penelitian Kashyap (2016) melakukan analisis komparatif dari Data Mining yang berbeda dengan teknik klasifikasi dalam industri ritel untuk analisis data pemasaran, analisis data keuangan, analisis data pendidikan serta untuk analisis data biomedis menggunakan WEKA. Algoritma Decision Tree C5.0, ID3 memberikan hasil yang akurat untuk klasifikasi dataset pendidikan terstruktur namun untuk klasifikasi data pendidikan tidak terstruktur diakomodir oleh (SVM). Klasifikasi Naïve Bayes serta Neural Network (NN) memberikan hasil yang akurat dalam hal beberapa parameter seperti kecepatan, efisiensi dan data biomedis. Neural Network (NN) classifier memberikan hasil yang lebih akurat dibandingkan dengan Decision Tree dan Naïve Bayes dalam hal efisiensi analisis Mammographic Mass dataset. Di Bank direct marketing, Support Vector Machine (SVM) menyediakan hasil yang lebih akurat dalam hal kecepatan dan efisiensi dibandingkan dengan Classifier lainnya.

Penelitian (Keerthana, 2016) mengidentifikasi metode image mining yang lebih baik dalam analisis citra medis berdasarkan analisis kinerja meliputi metode klasifikasi, Regresi Tree (CART), K-Means, Naive Bayes (NB), Decision Tree (DT) K-Nearest Neighbor dan Support Vector Machine (SVM). Dihasilkan tingkat akurasi algoritma SVM sebesar 87%, CART sebesar 91%, k- means menunjukkan 96%, Naïve bayes 90% dan akhirnya Decision Tree menunjukkan 59%. Keakuratan di atas dalam klasifikasi citra adalah gagasan utama untuk mengevaluasi kinerja dalam algoritma data mining.

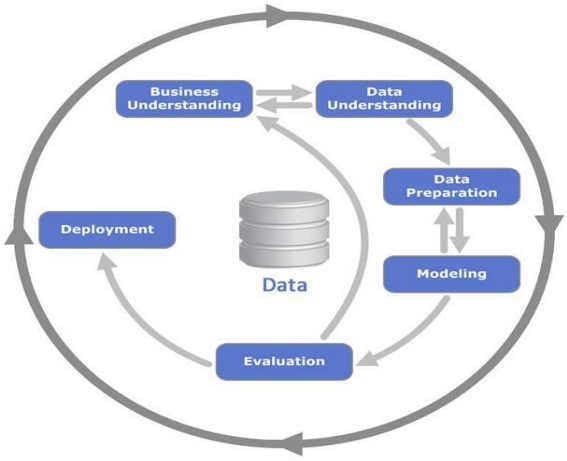
2. Pembahasan

2.1 Metodologi

Jenis penelitian yang dilakukan adalah eksperimen, investigasi hubungan sebab akibat dengan menggunakan ujicoba yang dikontrol oleh peneliti. Penelitian yang dilakukan dengan menerapkan serangkaian tindakan untuk membuktikan suatu konsep dalam hal ini adalah perbandingan evaluasi kinerja terbaik data mining.

Metode analisis data dalam penelitian ini mengacu pada tahapan proses CRISP-DM. CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) merupakan suatu konsorsium perusahaan yang didirikan oleh Komisi Eropa pada tahun 1996 dan telah ditetapkan sebagai proses standar dalam data mining yang dapat diaplikasikan di berbagai sektor industri.

Gambar 1 CRISP-DM



2.2 Pembahasan

a) Business Understanding.
Dalam penelitian ini fokus pada pendeteksian kanker payudara dengan menggunakan perbandingan 5 algoritma klasifikasi data mining yaitu C4.5, Naïve Baye, *K-Nearest Logistic Regression, dan Support Vector Machines.*

b) Data Understanding.
Dataset kanker payudara yang diambil pada http://archive.ics.uci.edu, data Breast Cancer dari Dr. William H. Woldberg (1989-1991)

University of Wisconsin Hospital, Madison, USA. Dataset kanker payudara ini berjumlah 699 dengan 11 parameter indikator yang akan diuji pada dataset kanker payudara yang diambil pada http://archive.ics.uci.edu antara lain: *Sample Code Number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses,* dan *Class* (atribut hasil prediksi).

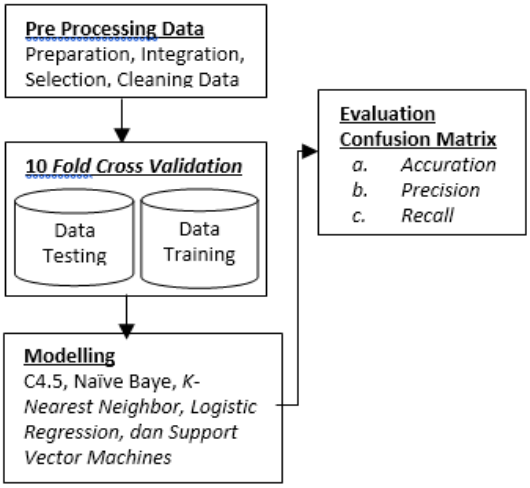
c) Data Preparation.
Dalam penelitian ini dilakukan pemilihan data seluruh indikator dalam membentuk dataset kanker payudara. Selanjutnya dilakukan pembobotan nilai yang secara default sudah ada http://archive.ics.uci.edu.

Tabel 2 Struktur Dataset Kanker Payudara

No	Indikator	Keterangan	Nilai	Type
1	*Sample Code Number*	Nomor kode sampel penderita	Nomor ID	Integer
2	*Clump Thickness*	Menentukan apakah sel berlayer atau tidak, karena sel kanker jinak (begin celss) cenderung hanya mempunya satgu layer (monolayer) sedangkan sel kanker ganas cenderung mempunya banyak (multilayer)	1-10	Integer
3	*Uniformity of Cell Size*	Menentukan konsistensi dalam ukuran sel kanker payudara	1-10	Integer
4	*Uniformity of Cell Shape*	Menentukan kesamaan bentuk sel kanker payudara	1-10	Integer
5	*Marginal Adhesion*	Menentukan apakah sel-sel memiliki ikatan dalam bentuk bersama sama atau tidak, karena sel ganas cenderung tidak memiliki kemampuan ini.	1-10	Integer
6	*Single Epithelial Cell Size*	Menentukan apakah sel *Epithelial* cenderung membesar atau tidak	1-10	Integer
7	*Bare Nuclei*	Menentukan apakah sel dikelilingi sitoplasma (sisa sel) atau tidak	1-10	Integer
8	*Bland Chromatin*	Menentukan tingkat tekstur dari sel kromatin	1-10	Integer
9	*Normal Nucleoli*	Menentukan bentuk dari sel *nucleoli*	1-10	Integer
10	*Mitoses*	Menentukan seberapa sel kanker membagi, membelah atau memperbanyak dirinya	1-10	Integer
11	*Class*	Menentukan apalah kanker payudara yang diderita termasuk kategori jinak atau ganas	2 kategoti jinak, dan 4 kategori ganas	Binominal

d) Modelling.

Gambar 2 Model Penelitian



e) Evaluation

Pada phase ini akan dilakukan proses evaluasi dari phase sebelumnya. Phase evaluasi ini akan dilakukan perbandingan quantitatif dengan mempertimbangkan nilai komparasi confusion matrix dan AUC dengan pengukuran berupa *Accuracy*, *Precision* dan *Recall*.

f) *Deployment*.

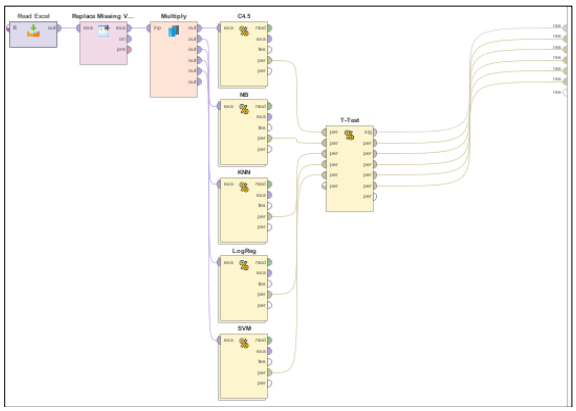
Tahapan penentuan model klasifikasi yang memiliki nilai kinerja terbaik dari hasil komparasi model data mining C4.5, Naïve Baye, *K-Nearest Logistic Regression, dan Support Vector Machines* Kemudian dibuat rekomendasi model mana yang terbaik yang akan diterapkan pada pendeteksian kanker payudara.

2.3 Hasil Analisis

2.3.1 Main Model Proses

Hasil konfigurasi model pada Rapidminer versi 9 dengan perbandingan 5 metode klasifikasi data mining C4.5, Naïve Baye, *K-Nearest Logistic Regression, dan Support Vector Machines* menggunakan uji beda T-Test untuk mencari kinerja terbaik.

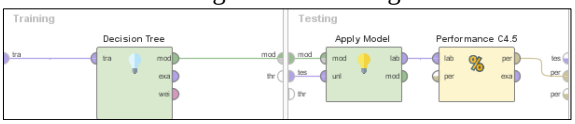
Gambar 3 Main Model Penelitian



2.3.2 Algoritma C4.5

Hasil model konfigurasi Algoritma C4.5 pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.

Gambar 3 Konfigurasi Model Algoritma C4.5

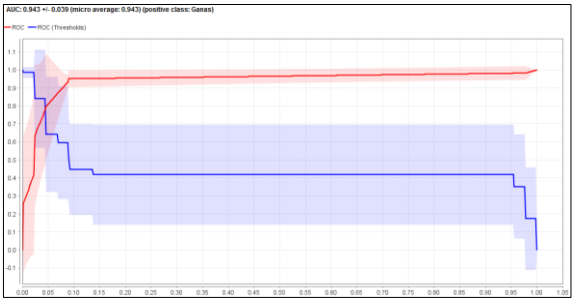


Hasil pengujian dan validasi melalui *confusion matrix* dengan pengukuran akurasi, *precision*, dan *recall* Algoritma C4.5 pada Rapidminer versi 9 tervisualisasi pada gambar 4 sedangkan pada gambar 5 adalah hasil kurva ROC Algoritma C4.5.

Gambar 4 Hasil Confusion Matrix dan AUC Algoritma C4.5

```
accuracy: 93.72% +/- 3.25% (micro average: 93.70%)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  424     23
Ganas:  20      216
precision: 91.78% +/- 4.68% (micro average: 91.53%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  424     23
Ganas:  20      216
recall: 90.38% +/- 8.34% (micro average: 90.38%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  424     23
Ganas:  20      216
AUC (optimistic): 0.968 +/- 0.037 (micro average: 0.968) (positive class: Ganas)
AUC: 0.943 +/- 0.039 (micro average: 0.943) (positive class: Ganas)
AUC (pessimistic): 0.918 +/- 0.048 (micro average: 0.918) (positive class: Ganas)
```

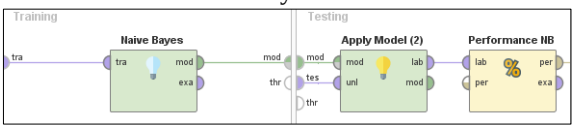
Gambar 5 Kurva ROC Algoritma C4.5



2.3.3 Algoritma Naïve Bayes

Hasil model konfigurasi Algoritma Naïve Bayes pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.

Gambar 6 Konfigurasi Model Algoritma Naïve Bayes

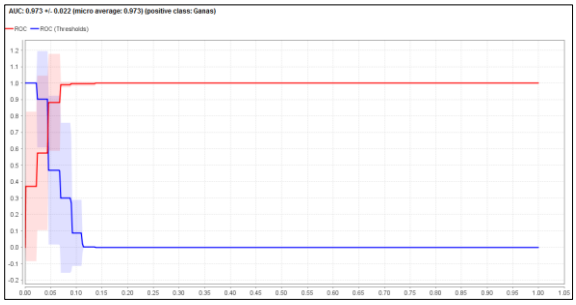


Hasil pengujian dan validasi melalui *confusion matrix* dengan pengukuran akurasi, *precision*, dan *recall* Algoritma Naïve Bayes pada Rapidminer versi 9 tervisualisasi pada gambar 7 sedangkan pada gambar 8 adalah hasil kurva ROC Algoritma Naïve Bayes.

Gambar 7 Hasil Confusion Matrix dan AUC Algoritma Naïve Bayes

```
accuracy: 96.19% +/- 1.64% (micro average: 96.19%)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  424     6
Ganas:  20      233
precision: 92.25% +/- 4.05% (micro average: 92.09%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  424     6
Ganas:  20      233
recall: 97.50% +/- 2.04% (micro average: 97.49%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  424     6
Ganas:  20      233
AUC (optimistic): 0.995 +/- 0.004 (micro average: 0.995) (positive class: Ganas)
AUC: 0.973 +/- 0.022 (micro average: 0.973) (positive class: Ganas)
AUC (pessimistic): 0.973 +/- 0.022 (micro average: 0.973) (positive class: Ganas)
```

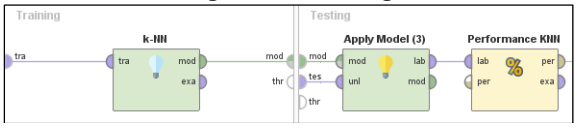
Gambar 8 Kurva ROC Algoritma Naïve Bayes



2.3.4 Algoritma K-Nearest Neighbor

Hasil model konfigurasi Algoritma K-Nearest Neighbor pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.

Gambar 9 Konfigurasi Model Algoritma K-NN

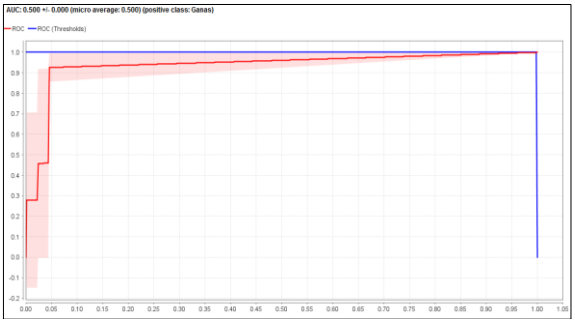


Hasil pengujian dan validasi melalui *confusion matrix* dengan pengukuran akurasi, *precision*, dan *recall* Algoritma K-Nearest Neighbor pada Rapidminer versi 9 tervisualisasi pada gambar 10, sedangkan pada gambar 11 adalah hasil kurva ROC Algoritma Naïve Bayes.

Gambar 10 Hasil Confusion Matrix dan AUC Algoritma K-NN

```
accuracy: 95.61% +/- 2.69% (micro average: 95.61%)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  432     18
Ganas:  12      221
precision: 94.99% +/- 3.60% (micro average: 94.85%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  432     18
Ganas:  12      221
recall: 92.43% +/- 7.02% (micro average: 92.47%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  432     18
Ganas:  12      221
AUC (optimistic): 0.998 +/- 0.003 (micro average: 0.998) (positive class: Ganas)
AUC: 0.500 +/- 0.000 (micro average: 0.500) (positive class: Ganas)
AUC (pessimistic): 0.899 +/- 0.070 (micro average: 0.899) (positive class: Ganas)
```

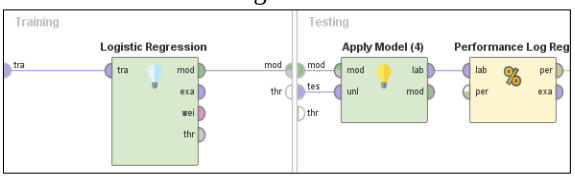
Gambar 11 Kurva ROC Algoritma K-NN



2.3.5 Algoritma Logistic Regression

Hasil model konfigurasi Algoritma Logistic Regression pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.

Gambar 12 Konfigurasi Model Algoritma Logistic Regression

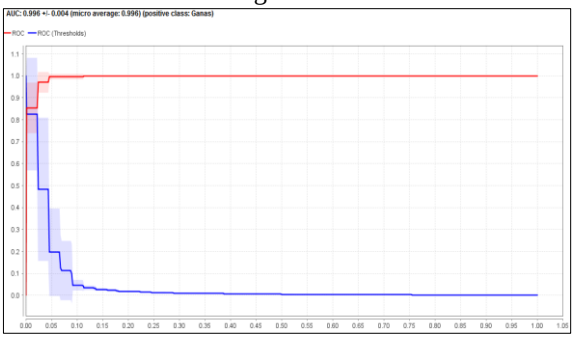


Hasil pengujian dan validasi melalui *confusion matrix* dengan pengukuran akurasi, *precision*, dan *recall* Algoritma Logistic Regression pada Rapidminer versi 9 tervisualisasi pada gambar 13, sedangkan pada gambar 14 adalah hasil kurva ROC Algoritma Naïve Bayes.

Gambar 13 Hasil Confusion Matrix dan AUC Algoritma Logistic Regression

```
accuracy: 96.77% +/- 1.11% (micro average: 96.78%)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  434     12
Ganas:  10      227
precision: 95.93% +/- 3.11% (micro average: 95.78%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  434     12
Ganas:  10      227
recall: 94.98% +/- 3.63% (micro average: 94.98%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  434     12
Ganas:  10      227
AUC (optimistic): 0.996 +/- 0.004 (micro average: 0.996) (positive class: Ganas)
AUC: 0.996 +/- 0.004 (micro average: 0.996) (positive class: Ganas)
AUC (pessimistic): 0.996 +/- 0.004 (micro average: 0.996) (positive class: Ganas)
```

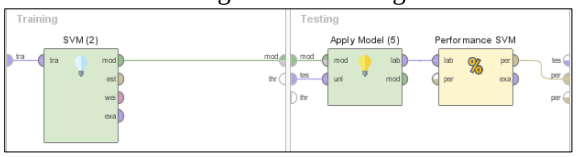
Gambar 14 Kurva ROC Algoritma Logistic Regression



2.3.6 Algoritma Support Vector Mechines

Hasil model konfigurasi Algoritma Support Vector Mechines pada Rapidminer versi 9 dengan *performance* 10-*fold Cross Validation*.

Gambar 15 Konfigurasi Model Algoritma SVM

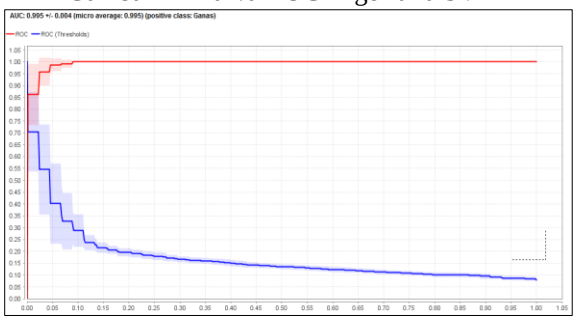


Hasil pengujian dan validasi melalui *confusion matrix* dengan pengukuran akurasi, *precision*, dan *recall* Algoritma Support Vector Mechines pada Rapidminer versi 9 tervisualisasi pada gambar 10, sedangkan pada gambar 11 adalah hasil kurva ROC Algoritma Naïve Bayes.

Gambar 16 Hasil Confusion Matrix dan AUC Algoritma SVM

```
accuracy: 96.78% +/- 1.70% (micro average: 96.78%)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  431     9
Ganas:  13      230
precision: 94.83% +/- 3.43% (micro average: 94.65%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  431     9
Ganas:  13      230
recall: 96.20% +/- 4.46% (micro average: 96.23%) (positive class: Ganas)
ConfusionMatrix:
True:   Jinak   Ganas
Jinak:  431     9
Ganas:  13      230
AUC (optimistic): 0.995 +/- 0.004 (micro average: 0.995) (positive class: Ganas)
AUC: 0.995 +/- 0.004 (micro average: 0.995) (positive class: Ganas)
AUC (pessimistic): 0.995 +/- 0.004 (micro average: 0.995) (positive class: Ganas)
```

Gambar 17 Kurva ROC Algoritma SVM



2.3.7 Hasil Perbandingan Kinerja Algoritma

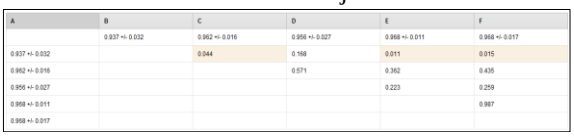
Tabel 2 Hasil Komparasi Kinerja Algoritma

Item	C.45	NB	KNN	LR	SVM
Accuratio n	93.70 %	96.19 %	95.61 %	96.77 %	96.78 %
Precision	94.26%	92.25 %	94.99 %	95.93 %	94.83 %
Recall	87.86 %	97.50 %	92.43 %	94.98 %	96.20 %
AUC	0, 931	0,973	0,500	0,996	0,995

Dari hasil komparasi 5 algoritma klasifikasi yang dianalisis diantaranya algoritma C4.5, Naïve Baye, *K-Nearest Logistic Regression, dan Support Vector Machines* dengan pengukuruan *accuration, precision, recall*, dan AUC dapat disimpulkan bahwa diagnosa akurasi berada pada rentang 0.90 – 1.00 masuk kategori *Excellent classification.*

Selain itu mengacu pada hasil uji beda T-Test pada gambar 18 mangasilkan bahwa nilai T-Test didapatkan algoritma *Logistic Regression* dan *Support Vector Machines* memiliki nilai akurasi tertinggi dari 3 algoritma yang lain, dengan nilai perolehan yang sama yaitu sebesar 0,968.

Gambar 18 Hasil Uji T-Test



3. Kesimpulan dan Saran

3.1 Kesimpulan

Berdasarkan pembahasan dan hasil analisis data pada bab sebelumnya, maka diperoleh kesimpulan sebagai berikut :

1. Hasil penelitian ini didapatkan bahwa presentase kinerja masing – masing algoritma klasifikasi yang dianalisis yaitu Algoritma C4.5 (akurasi 93.70%, *precision* 94.26%, *recall* 87.86%), Algoritma *Naïve Bayes* (akurasi 96.19%, precision 92.25%, recall 97.50%), Algoritma *K-Nearest Neighbor* (akurasi 95.61%, *precision* 94.99%, *recall* 92.43%), Algoritma *Logistic Regression* (akurasi 96.77%, precision 95.93%, recall 94.98%), dan Algoritma *Support Vector Machines* (akurasi 96.78%, precision 94.83%, recall 96.20%).
2. Hasil kinerja terbaik yang diuji menggunakan T-Test didapatkan bahwa algoritma Logistic Regression dan Support Vector Machines memiliki nilai akurasi tertinggi yang sama nilainya yaitu sebesar 0,968.

3.2 Saran

Berdasarkan hasil penelitian diatas, maka dapat di buat beberapa saran berikut :

1. Dibutuhkan jumlah data yang lebih besar, atribut yang lebih kompleks, bahkan menggunakan sampel penyakit lain yang sifatnya memiliki struktur data baru sehingga hasil pengukuran yang dihasilkan akan lebih berguna dan lebih handal akurasinya.
2. Dilakukan metode optimasi untuk meningkatkan kinerja algoritma seperti Ant Particle Swarm Optimization (PSO) Colony Optimization (ACO), Genetik Algorithm (GA), dan lain sebagainya.
3. Melakukan pengujian dan perbandingan pada algoritma lain ataupun menggunakan metode hybrid untuk mendapatkan pengetahuan komparasi yang lebih luas.
4. Melakukan pengembangan pada preprocessing data dengan menggunakan metode seleksi atribut yang lain seperti chi-square, information index

dan sebagainya untuk ketepatan penyeleksian atribut.

Daftar Pustaka

Bramer, Max. (2007). Principles of Data Mining. London: Springer. ISBN-10: 1-84628-765-0, ISBN-13: 978-1-84628-765-7.

Chyntia, E. (2009). Akhirnya aku sembuh dari kanker payudara. Maximus, Yogyakarta

Dea Alverina, A. R. Chrismanto, and R. G. Santosa, 2018, Perbandingan Akurasi Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa, Jurnal Teknologi dan Sistem Komputer, vol. 6, no. 2, Apr. 2018. doi: 10.14710/jtsiskom.6.2.2018.76-83, ISSN:2338-0403.

Dewi Sartika, Dana Indra Sensuse, 2017, Perbandingan Algoritma Klasifikasi Naive Bayes, Nearest Neighbour, dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian. Jatisi, Vol. 1 No. 2, ISSN: 1978-1520

Durairaj, M., & Deepika, R, 2015, Comparative Analysis of Classification Algorithms for the Prediction of Leukemia Cancer. International Journal of Advanced Research in Computer Science and Software Engineering; Volume 5, No. 8, pp. 787-791.

Geeta Kashyap, Ekta Chauhan, 2016, Parametric Comparisons of Classification Techniques in Data Mining Applications, International Journal of Engineering Development and Research, Vol 4, Issue 2, ISSN: 2321-9939

Gorunescu, F. (2011). Data Mining Concepts, Models and Techniques, Springer, Verlag Berlin Heidelrbeg

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques.

Hosmer dan Lemeshow. 2000. Applied Logistic Regression. USA. John Wiley and Sons.

Keerthana, P et al, 2016, Performance Analysis Of Data Mining Algorithms For Medical Image Classification. International Journal of Computer Science and Mobile Computing, Vol.5 Issue.3, pg. 604-609, ISSN 2320–088X

Kusrini dan Emha Taufiq Lutfi, 2009, Algorima Data Mining, Andi Offset, Yogyakarta.

Larose, D.T. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Willey & Sons, Inc

Nisa, U. Z., "Model Prediksi Finansial Distress Pada Perusahaan Manufaktur Go Public di Indonesia", Thesis Program Magister Bidang Optimasi Sistem Industri Jurusan Teknik Industri Institut Teknologi Sepuluh Nopember,2013.

Rahayu-Tjioe, A. (1991). Kanker payudara. Yayasan Kanker Wisnuwardhana, Surabaya

Santosa, B. 2007. Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis. Graha Ilmu, Yogyakarta.

Santoso. (2007). Data Mining Terapan dengan Matlab. Yogyakarta: Graha Ilmu.

Suyanto, 2017, Data Mining Untuk Klasifikasi Dan Klasterisasi Data, Informatika, Bandung

Vercellis, C. (2009). Data Mining and Optimization for Decision Making. Italy: WILEY.

Zaki, M.J., & Meira, M.J. 2013. Data Mining and Analysis: Fundamental Concepts and Algorithms.