Jurnal Edik Informatika
Penelitian Bidang Komputer Sains dan Pendidikan Informatika
V2.i2(213-219)
ISSN : 2407-0491
E-ISSN : 2541-3716

Data Mining : Klasifikasi Menggunakan Algoritma C4.5

Yuli Mardi
Dosen Akademi Perekam dan Informasi Kesehatan (APIKES) Iris Padang
Jl. Gajah Mada No. 23 Padang, Sumatera Barat
adimardi@yahoo.com

ABSTRAK

Data mining merupakan bagian dari tahapan proses *Knowledge Discovery in Database (KDD)*. Dengan data mining, kita dapat melakukan pengklasifikasian, memprediksi, memperkirakan dan mendapatkan informasi lain yang bermanfaat dari kumpulan data dalam jumlah yang besar. Klasifikasi dalam data mining dapat dilakukan dengan menggunakan algoritma C4.5. Dengan algoritma C4.5, akan didapatkan sebuah pohon keputusan yang mudah dipahami dan mudah dimengerti.

Kata kunci : Data mining, Klasifikasi, Algoritma C4.5, Pohon keputusan

PENDAHULUAN

Database yang tersimpan di media penyimpanan jarang sekali dimanfaatkan oleh sebagian besar penggunanya dan bahkan dalam jangka waktu tertentu data-data tersebut dihapus karena dianggap sampah dan hanya memenuhi media penyimpanan saja. Anggapan tersebut tidak sepenuhnya benar, karena sesungguhnya database dalam ukuran yang besar dapat memberikan informasi yang dibutuhkan untuk berbagai kepentingan, baik untuk kepentingan bisnis dalam mengambil keputusan maupun untuk ilmu pengetahuan dan penelitian.

Knowledge Discovery In Database(KDD) merupakan metode untuk memperoleh pengetahuan dari database yang ada. Dalam database terdapat tabel - tabel yang saling berhubungan / berelasi. Hasil pengetahuan yang diperoleh dalam proses tersebut dapat digunakan sebagai basis pengetahuan *(knowledge base)* untuk keperluan pengambilan keputusan.

Istilah *Knowledge Discovery in Database (KDD)* dan data mining seringkali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain, dan salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Proses KDD secara garis besar dapat dijelaskan sebagai berikut[1]:

1. *Data Selection*
   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam *Knowledge Discovery in Database(KDD)* dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas terpisah dari basis data operasional.

2. *Pre-processing / Cleaning*
   Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus *Knowledge Discovery in Database (KDD).* Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak. Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk *Knowledge Discovery in Database (KDD)*, seperti data atau informasi eksternal lainnya yang diperlukan.

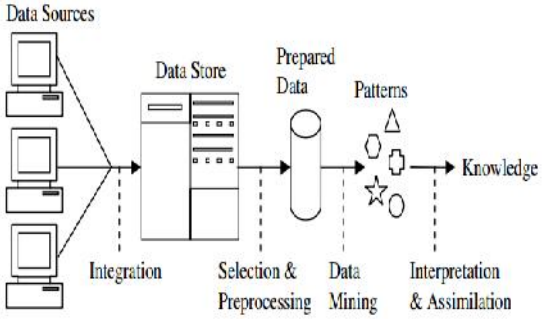
3. *Transformation*
   Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses *coding* dalam *Knowledge Discovery in Database (KDD)* merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

ISSN : 2407-0491
E-ISSN : 2541-3716
Jurnal Edik Informatika
Penelitian Bidang Komputer Sains dan Pendidikan Informatika
V2.i2(213-219)

4. *Data Mining*
   Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik-teknik, metode-metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *Knowledge Discovery in Database (KDD)* secara keseluruhan.

5. *Interpretation / Evaluation*
   Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses *Knowledge Diccovery in Database (KDD)* yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

Sementara itu menurut[2], proses *Knowledge Diccovery in Database (KDD)* dapat digambarkan sebagai berikut :



Gambar 1 : Proses Knowledge Discovery in Database

DATA MINING

Menurut Gartner Group,data mining adalah proses menemukan hubungan baru yang mempunyai arti, pola dan kebiasaan dengan memilah-milah sebagian besardata yang disimpan dalam media penyimpanan dengan menggunakan teknologi pengenalan pola seperti teknik statistik dan matematika. Data mining merupakan gabungan dari beberapa disiplin ilmu yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar[3].

Data mining menurut David Hand, Heikki Mannila, dan Padhraic Smyth dari MIT adalah analisa terhadap data (biasanya data yang berukuran besar) untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut[4].

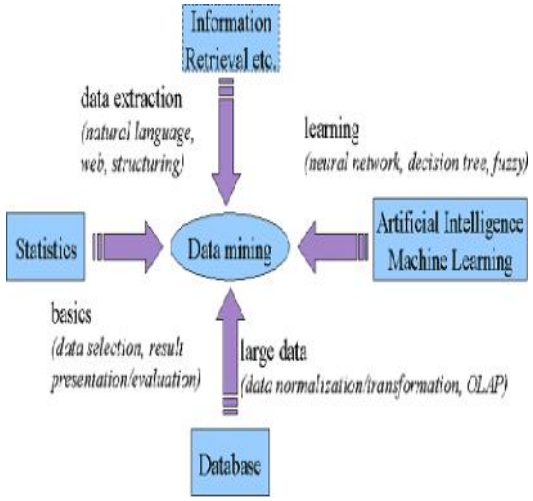
Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. Data mining merupakan serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual[1].

Dari definisi-definisi yang telah disampaikan, hal penting yang terkait dengan data miningmenurut[1]:

1. Data mining merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses merupakan data yang sangat besar.
3. Tujuan data mining adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

Data mining bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan data mining adalah kenyataan bahwa data mining mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang dulu sudah mapan terlebih dulu. Gambar 2 menunjukkan bahwa data mining memiliki akar yang panjang dari bidang ilmu yang berbeda seperti kecerdasan buatan (artificial intelligent), machine learning, statistik, database, dan juga information retrieval[1].

ISSN : 2407-0491
Jurnal Edik Informatika E-ISSN : 2541-3716
Penelitian Bidang Komputer Sains dan Pendidikan Informatika
V2.i2(213-219)



Gambar 2 : Bidang Ilmu Data Mining

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

1. *Description* (Deskripsi)

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. *Estimation* (Estimasi)

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik daripada ke arah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh yaitu estimasi nilai indeks prestasi kumulatif mahasiswa program pasca sarjana dengan melihat nilai indeks prestasi mahasiswa tersebut pada saat mengikuti program sarjana.

3. *Prediction* (Prediksi)

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada dimasa mendatang.Contoh prediksi dalam bisnis dan penelitian adalah :

a. Prediksi harga beras dalam tiga bulan yang akan datang.
b. Prediksi tingkat penganguran lima tahun akan datang.
c. Predisksi persentase kanaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikan.

Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

4. *Classification* (Klasifikasi)

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.Contoh lain klasifikasi dalam bisnis dan penelitian adalah :

a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.
b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.
c. Mendiagnosis penyakit seorang pasien untuk mendapatkan termasuk penyakit apa.

5. *Clustering* (Pengklusteran)

Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan record yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan record-record dalam kluster lain.

Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.Contoh pengklusteran dalam bisnis dan penelitian adalah :

a. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari

ISSN : 2407-0491
Jurnal Edik Informatika E-ISSN : 2541-3716
Penelitian Bidang Komputer Sains dan Pendidikan Informatika
V2.i2(213-219)

suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.

b. Untuk tujuan audit akuntansi, yaitu melakukan pemisahan terhadap perilaku finansial dalam baik dan mencurigakan
c. Melakukan pengklusteran terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar

6. *Association* (Asosiasi)

Tugas asosiasi dalam data mining adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.Contoh asosiasi dalam bisnis dan penelitian adalah :

a. Meneliti jumlah pelanggan dari perusahaan telekomunikasi seluler yang diharapkan untuk memberikan respons positif terhadap penawaran upgrade layanan yang diberikan.
b. Menemukan barang dalam supermarket yang dibeli secara bersamaan dan barang yang tidak pernah dibeli secara bersamaan.

KLASIFIKASI

Salah satu tugas yang dapat dilakukan dengan data mining adalah pengklasifikasian. Klasifikasi pertama kali diterapkan pada bidang tanaman yang mengklasifikasi suatu spesies tertentu, seperti yang dilakukan oleh Carolus von Linne (atau dikenal dengan nama Carolus Linnaeus) yang pertama kali mengklasifikasi spesies berdasarkan karakteristik fisik. Selanjutnya dia dikenal sebagai bapak klasifikasi[4].

Dalam klasifikasi terdapat target variabel kategori. Metode-metode / model-model yang telah dikembangkan oleh periset untuk menyelesaikan kasus klasifikasi antara lain[4]:

a. Pohon keputusan
b. Pengklasifikasi bayes/*naive bayes*
c. Jaringan saraf tiruan
d. Analisis statistik
e. Algoritma genetik
f. *Rough sets*
g. Pengklasifikasi *k-nearest neighbour*
h. Metode berbasis aturan
i. *Memory based reasoning*
j. *Support vector machine*

1. POHON KEPUTUSAN

Diantara beberapa metode yang dapat digunakan untuk klasifikasi adalah metode pohon keputusan atau *decission tree*. Metode pohon keputusan merupakan sebuah metode yang dapat mengubah fakta yang sangat besar menjadi sebuah pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami[1].

Sebuah pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan-kumpulan data yang besar menjadi himpunan-himpunan *record* yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dengan masing-masing rangkaian pembagi, anggota himpunan hasil menjadi mirip satu dengan yang lainnya. Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record*. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon. Misalkan untuk menentukan main tenis, kriteria yang diperhatikan adalah cuaca, angin dan temperatur. Salah satu atribut merupakan atribut yang menyatakan data solusi per *item* data yang disebut target atribut. Atribut memiliki nilai-nilai yang dinamakan dengan *instance*. Misalkan atribut cuaca mempunyai *instance* berupa cerah, berawan dan hujan. Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule*, dan menyederhanakan *rule*[1].

Banyak algortima yang bisa digunakan dalam pembentukan pohon keputusan, antara lain ID3, CART, dan C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID.

ALGORITMA C4.5

Salah satu algoritma yang dapat digunakan untuk membuat pohon keputusan (*decission tree)* adalah algoritma C4.5. Algoritma C4.5 merupakan algoritma yang sangat populer yang digunakan oleh banyak peneliti di dunia, hal ini dijelaskan oleh Xindong Wu dan Vipin Kumar dalam bukunya yang berjudul The Top Ten Algorithms in Data Mining. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang di ciptakan oleh J. Rose Quinlan.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut[1]:

ISSN : 2407-0491
E-ISSN : 2541-3716
Jurnal Edik Informatika
Penelitian Bidang Komputer Sains dan Pendidikan Informatika
V2.i2(213-219)

a. pilih atribut sebagai akar
b. buat cabang untuk tiap-tiap nilai
c. bagi kasus dalam cabang
d. ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan persamaan 1.

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

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

Sementara itu, perhitungan nilai *entropy* dapat dilihat pada persamaan 2

$$Entropy\ S = \sum_{i=1}^{n} - p * \log_2 pi \quad (2)$$

Keterangan :
S = himpunan kasus
A = fitur
n = jumlah partisi S
p_i = proporsi dari S_i terhadap S

Untuk lebih jelasnya langkah-langkah dalam pembuatan pohon keputusan, berikut diberikan contoh kasus yang dapat diklasifikasi menggunakan algoritma C4.5. Dari Tabel 1, hitung apakah seorang nasabah bermasalah dalam kredit atau tidak.

Data *training* pada Tabel 1 adalah untuk menentukan apakah seorang nasabah bermasalah atau tidak yang ditentukan oleh kolom *predictor* simpanan, aset, dan pendapatan. Kolom resiko kredit adalah kelas dari masing-masing record.

Tabel 1. Tabel Data untuk Klasifikasi Resiko Kredit

Pelanggan	Simpanan	Aset	Pendapatan	Resiko Kredit
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

Adapun langkah untuk membuat pohon keputusan, yaitu :

1. Untuk atribut pendapatan yang bernilai angka, dibuat dalam bentuk kategori, yaitu pendapatan <=25, pendapatan >25, pendapatan <=50, pendapatan >50, pendapatan ,=75, dan pendapatan >75
2. Hitung nilai *entropy*. Dari data *training* diketahui jumlah kasus ada 8, yang beresiko kredit *good* 5 *record* dan *bad* 3 *record* sehingga didapat entropy :

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

   $= (-5/8*\log_2(5/8)) + (-3/8*\log_2(3/8))$

   = 0,9544

3. Hitung nilai *gain* untuk tiap atribut, lalu tentukan nilai *gain*tertinggi. Yang mempunyai nilai *gain* tertinggi itulah yang akan dijadikan akar dari pohon. Misalkan untuk atribut simpanan dengan nilai *low* didapat nilai *gain* :

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

   = 0,9544 – (2/8(1) + 3/8(0) +3/8(0,9183)

   = 0,3601

ISSN : 2407-0491
E-ISSN : 2541-3716
Jurnal Edik Informatika
Penelitian Bidang Komputer Sains dan Pendidikan Informatika
V2.i2(213-219)

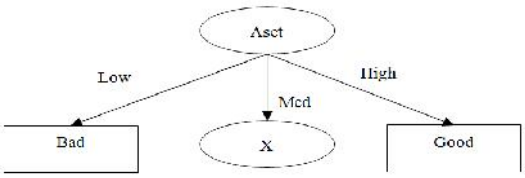
Hasil perhitungan *gain* untuk tiap atribut dapat terlihat pada Tabel 2, nilai gain tertinggi akan menjadi akar dari pohon.

Tabel 2. Nilai Entropy dan Gain untuk Menentukan Simpul Akar

Simpul			**Data**	**Good**	**Bad**	**Entropy**	**Gain**
Akar	Total		8	5	3	0,9544	
	Simpanan						0,3601
		Low	3	1	2	0,9183	
		Medium	3	3	0	0	
		High	2	1	1	1	
	Aset						0,5488
		Low	2	0	2	0	
		Medium	4	3	1	0,8113	
		High	2	2	0	0	
	Pendapatan						0,1589
		<=25	3	1	2	0,9183	
		>25	5	4	1	0,7219	
							0,3476
		<=50	5	2	3	0,971	
						0	
		>50	3	3	0	0	
							0,0924
		<=75	7	4	3	0,9852	
		>75	1	1	0	0	

Terlihat dari Tabel 2 bahwa atribut aset mempunyai nilai *low*, *medium*, dan *high*. Nilai *low* dan *high* masing-masing sudah menjadi satu klasifikasi karena pada data *training*, semua aset menghasilkan keputusan yang sama yaitu *bad* untuk nilai *low* dan *good* untuk nilai *high*. Sedangkan untuk simpul dengan nilai *medium* perlu dipartisi lagi.

Gambar 3 adalah hasil pembentukan pohon keputusan berdasarkan perhitungan yang terdapat pada Tabel 2. Dari hasil perhitungan didapat nilai *gain* tertinggi untuk atribut aset, maka aset menjadi akar pohon keputusan. Untuk menentukan akar dari atribut *medium*, dilakukan lagi perhitungan *gain*.

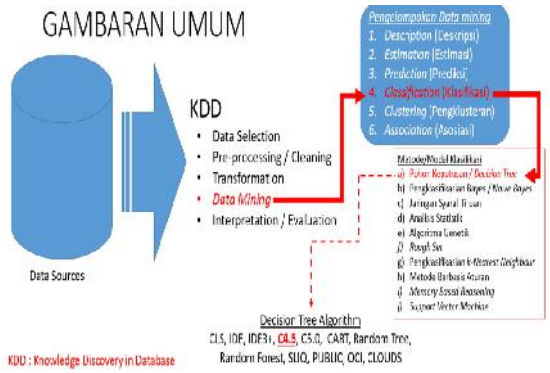


Gambar 3 Pohon Keputusan C4.5 dengan Simpul Akar Aset

KESIMPULAN

Secara keseluruhan, proses data mining hingga menghasilkan sebuah pohon keputusan yang dapat memberikan informasi yang diperlukan, dapat dilihat dari Gambar 4[5].

ISSN : 2407-0491
Jurnal Edik Informatika E-ISSN : 2541-3716
Penelitian Bidang Komputer Sains dan Pendidikan Informatika
V2.i2(213-219)



Gambar 4 Proses Klasifikasi menggunakan Algoritma C4.5

Dari gambar 4 dapat di jelaskan proses data mining hingga menghasilkan sebuah pohon keputusan adalah sebagai berikut :

1. Sumber data, merupakan database yang didalamnya terdapat informasi yang bisa diambil dan dimanfaatkan untuk kepentingan bisnis dan penelitian
2. Proses KDD, merupakan proses yang dilakukan untuk mengambil informasi yang terdapat dalam database, di antara proses tersebut terdapat proses data mining
3. Data mining, data mining merupakan bagian dari proses kdd, apa yang dapat dilakukan dengan data mining dapat dilihat digambar 4, diantaranya adalah klasifikasi
4. Beberapa model dapat digunakan untuk melakukan klasifikasi dan dalam pembahasan ini kita menggunakan model pohon keputusan
5. Algoritma yang dapat dilakukan untuk membuat pohon keutusan salah satunya dalah algoritma C4.5

DAFTAR PUSTAKA

Bramer, Max (2007)*Principles of Data Mining*, Springer Science

Kusrini dan Emha Taufiq Luthfi (2009) *Algoritma Data Mining*, Andi Offset

Larose, Daniel T (2005) *Discovering Knowledge in Data Mining An Introduction to Data Mining*, Wiley Interscience

Mardi, Yuli (2014) *Analisa Data Rekam Medis untuk Menentukan Penyakit Terbanyak Berdasarkan International Classification Of Disease (ICD) Menggunakan Decision Tree C4.5 (Studi Kasus : RSU. CBMC Padang)*. UPI YPTK Padang

Widodo *et al* (2013) Penerapan Data Mining dengan Matlab, Rekayasa Sains