BAB II
KAJIAN LITERATUR

2.1 Data Mining

Data Mining adalah kajian yang meliputi kegiatan pengumpulan, pembersihan, pemrosesan, dan analisa sekumpulan data sehingga dengan kegiatan tersebut dapat diperoleh pemahaman yang mendalam akan data. Data mining telah banyak menarik perhatian di dunia sistem informasi dan di masyarakat secara keseluruhan dalam beberapa tahun ini, karena ketersediaan luas dalam jumlah besar data dan kebutuhan segera untuk mengubah data tersebut menjadi informasi yang berguna dan pengetahuan. Informasi dan pengetahuan yang diperoleh dapat digunakan untuk aplikasi mulai dari pasar analis, deteksi penipuan, dan retensi pelanggan, untuk pengendalian produksi dan ilmu pengetahuan eksplorasi. Banyaknya data, ditambah dengan kebutuhan untuk alat analisis data yang kuat, telah digambarkan sebagai kaya data tapi miskin informasi. Jumlah data yang tumbuh secara cepat, dikumpulkan dan disimpan dalam repositori data yang besar dan banyak, telah jauh melampaui kemampuan mannusia untuk memahami data-data tersebut tanpa mampu mengelola data tersebut. Akibatnya, data yang dikumpulkan dalam repositori data yang besar menjadi "kuburan data" . Data mining adalah disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar. Data mining adalah disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar [9].

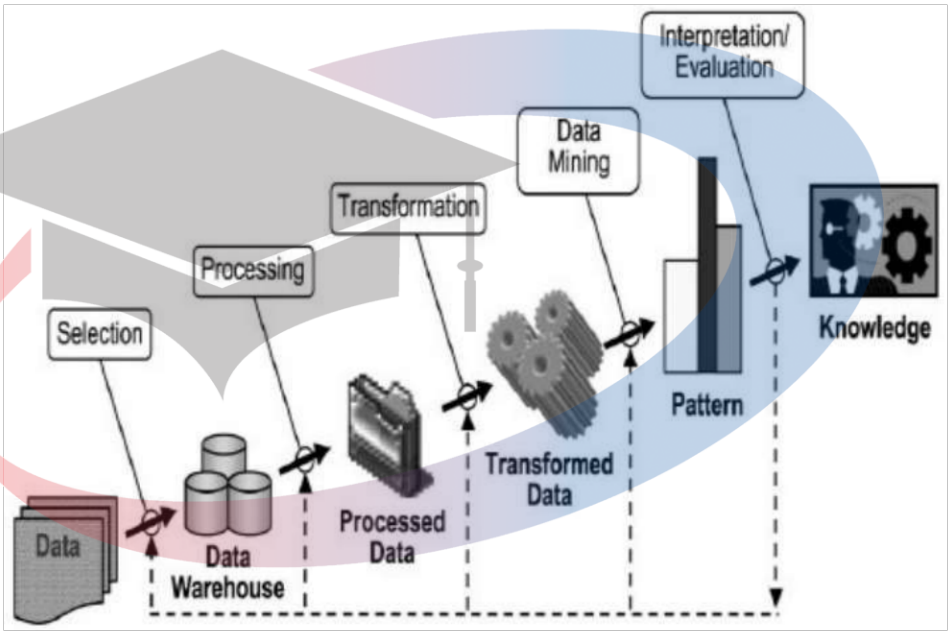
Gambar 2.1 Konsep Data Mining

Data mining yang disebut juga dengan *Knowledge Discovery in Database* (KKD), juga dapat didefinisikan sebagai suatu proses ekstraksi atau penggalian data yang belum diketahui sebelumnya, namun dapat dipahami dan berguna dari *database* yang besar serta digunakan untuk membuat suatu keputusan bisnis yang sangat penting. KDD sendiri memiliki beberapa tahapan

yang perlu dilakukan untuk memperoleh hasil berupa pola, pengetahuan ataupun informasi tertentu dari sebuah *database*. Tahapan-tahapan tersebut yakni *Selection*, *Preprocessing*, *Transformation*, Data Mining, *Interpretation and Evaluation,* Data *Visualization*. . Peneliti menggunakan tahapan-tahapan yang sama dalam melakukan penelitian ini untuk memperoleh hasil akhir dari data yang telah dikumpulkan [10].

2.1.1 Tahapan-tahapan Data Mining

Adapun tahapan tahapan data mining adalah sebagai berikut [11]:



Gambar 2.2 Tahapan Data Mining

Berikut ini adalah penjelasan tahapan data mining berdasarkan gambar 2.2:

1. Data *selection*

   Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang 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 KDD. Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data.

3. *Transformation*

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. Data mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. Interpretation/evalution

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 KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.1.2 Pengelompokan Data Mining

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas dan fungsinya yaitu [12] [13].

1. Deskripsi

   Menggambarkan pola dan kecenderungan yang ada pada data, deskripsi dari pola dan kecenderungan sering memberikan penjelasan untuk suatu pola atau kecenderungan

2. Estimasi

   Estimasi hampir serupa dengan klasifikasi, kecuali variabel target dalam estimasi lebih ke arah numerik daripada kearah 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.

3. Prediksi

Didalam prediksi, nilai dari hasil akan ada dimasa mendatang. Prediksi memiliki kemiripan dengan klasifikasi dan estimasi. Beberapa algoritma dan teknik yang digunakan dalam klasidikasi dan estimasi dapat pulak digunakan dapam prediksi (untuk kondisi yang tepat).

4. Klasifikasi

   Men-generalisasi struktur yang diketahui untuk diaplikasikan pada data-data baru. Dalam klasifikasi, terdapat target variabel kategori.

5. Pengklusteran

   Adalah suatu pengelompokan *record,* pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang mempunyai kesamaan.

6. Asosiasi

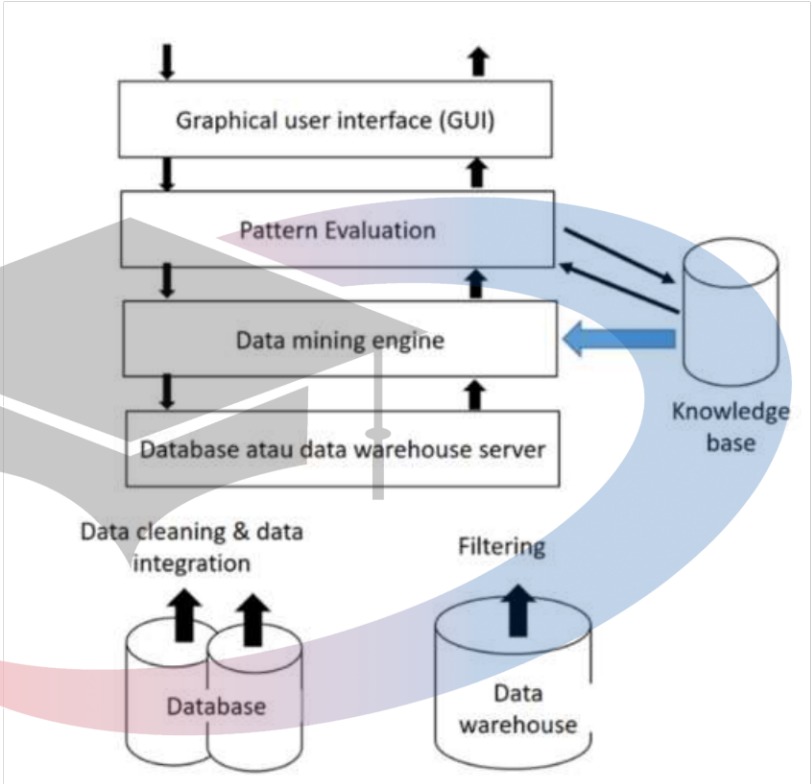
   Menemukan atribut yang muncul dalam suatu waktu. Dalam dunia bisnis lebih umum disebut keranjang belanja.

2.1.3 Arsitektur Data Mining

Data mining merupakan proses penggalian pengetahuan dari data-data yang ada dalam jumlah dan ukuran yang besar yang tersimpan dalam basis data (*database*). Arsitektur data mining memiliki komponen-komponen utama yaitu [14]:

1. Data *warehouse* yaitu tempat penyimpanan informasi lainnya.
2. Data *warehouse* server adalah komponen yang bertanggung jawab dalam pengambilan dalam relevan data berdasarkan kebutuhan penggunanya.
3. Basis pengetahuan, komponen ini merupakan domain *knowledge* yang digunakan untuk proses pencairan atau mengevaluasi pola-pola. Dapat berupa kepercayaan pengguna yang dapat digunakan untuk menarik pola-pola yang diperoleh.
4. Data mining *engine,* yaitu bagian penting dalam arsitektur data mining yang berisikan modul-modul fungsional data mining seperti : asosiasi, klasifikasi, *cluster*, dan lainnya.
5. Modul evaluasi pola, komponen ini berinteraksi dengan modul data mining *engine* dalam proses penarikan pola-pola. Modul ini menggunakan *threshold* untuk memperoleh dan memfilter pola yang akan diperoleh.
6. Antar muka pengguna grafis/*user interface*, modul ini adalah komponen untuk berinteraksi atau berkomunikasi dengan pengguna. Modul ini juga menyediakan

informasi yang memfokuskan pada pencarian dan eksplorasi data mining. Antar muka ini juga dapat memungkinkan pengguna mencari (*browser*) *database* dan skema data *warehouse*, mengevaluasi pola yang diperoleh dan visualisasi pola dalam berbagai bentuk.



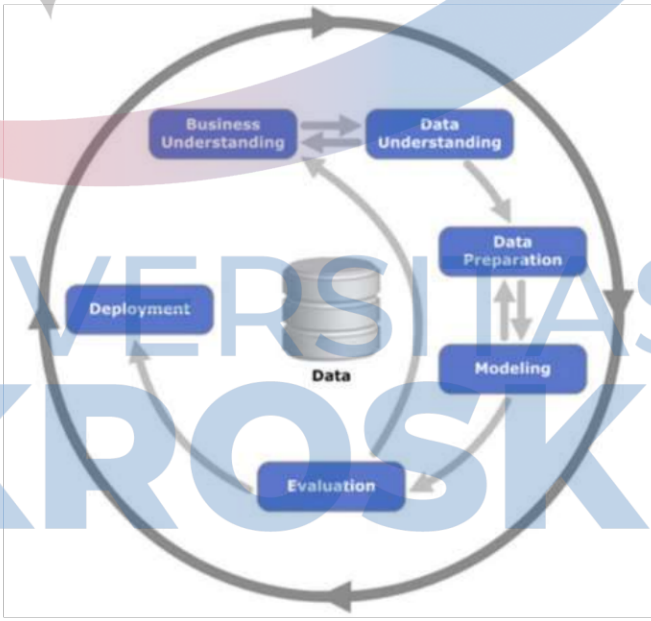
Gambar 2.3 Arsitektur sistem data mining

2.1.4 Siklus Hidup Data Mining

Dalam pengembangan data mining, siklus hidup dalam data mining dibagi menjadi enam tahapan yaitu :

1. *Busines understanding*, pada tahapan merupakan tahap untuk memahami tujuan dan analisa kebutuhan bisnis yang kemudian diubah kedalam pengetahuan untuk menyusun dan menentukan rencana dan strategi
2. Data *understanding*, tahapan ini merupakan tahapan untuk proses pengumpulan data, mengidentifikasi data untukmenemukan pla yang menarik yang dapat digunakan untuk membuat hipotesis.

3. Data *preparation*, tahapan ini tahap membangun dataset yang akan dilanjutkan ke tahap pemodelan dari data yang telah terkumpul. Tahapan ini meliputi pemilihan tabel, atribut data termasuk proses pembersihan (*cleansing*).
4. *Modeling*, pada tahapan ini pemilihan teknik data mining dipilih dan disesuaikan dengan parameter untuk mendapatkan hasil nilai yang optimal. Teknik yang sama dapat dipergunakan untuk menyelesaikan permasalah yang berbeda.
5. *Evaluation*, model yang sudah terbentuk diharapkan menghasilkan nilai yang berkualitas. Untuk mengetahui hasil tersebut maka dilakukan proses evaluasi. Proses evaluasi dilakukan untuk melihat keefektifan dan kualitas model yang digunakan.
6. *Deployment*, tahapan ini menyajikan hasil pengetahuan atau informasi yang diperoleh secara khusus sehingga dapat dipergunakan oleh *user*/pengguna. Tahapan ini dapat berupa proses pembuatan laporan sederhana.



Gambar 2.4 Siklus Hidup Data Mining

2.2 Algoritma C4.5

2.2.1 Pengertian Algoritma C4.5

Algoritma C4.5 merupakan salah satu dari beberapa metode klasifikasi dimana algoritma ini dalam keputusan akhir memakai *decision tree* atau sebuah pohon keputusan. Dari keputusan

itulah akan dibentuk sebuah cabang yang memperlihatkan hasil dari setiap perhitungan menggunakan algoritma C4.5. Pohon keputusan akan menenunjukkan beberapa hubungan antara tiap item atribut yang tersembunyi. Model ini biasa disebut dengan prediksi hirarki yang membantu sebuah pohon yang bukan memiliki daun tetapi dalam bentuk *node*.

2.2.2 Tahapan Algoritma C4.5

Penyelesaian menggunakan algoritma ini dilakukan beberapa tahapan diantaranya adalah sebagai berikut yaitu [3]:

1) Menyiapkan data yang akan dilakukan pengujian atau disebut juga dengan penyiapan data *training* dari data yang lama yaitu data atau histori yang sudah pernah terjadi sebelumnya dengan data yang baru sebagai bahan pertimbangan dalam pengambilan keputusan dari tinjauan data yang pernah ada.
2) Menentukan akar dari sebuah pohon keputusan yang didapat dari mencari rumus atau nilai *Gain*, dimana nilai *Gain* dapat disimpulkan jika sudah mendapatkan nilai *Entropy* sebelumnya. *Entropy* diambil dengan menggunakan rumus berikut:

$$Entropy\ (s) = \Sigma_{i}^{\Pi}{}_{=1} - pi * log_2 pi$$

Keterangan :

s : himpunan dari sebuah kasus

n : jumlah

pi : nilai keseimbangan dari Si terhadap S

a) Mencari nilai *Gain*

$$Gain\ (S, A) = Entropy\ (s) - \Sigma_{i}^{\Pi}{}_{=1} \frac{si}{s} * Entropy$$

Keterangan :

Si : jumlah kasus pada ke-*i*

S : jumlah kasus dalam atribut S

A : atribut

N : jumlah atribut A

b) Lakukan seluruh proses mulai dari pengelompokan terhadap pemilihan atribut hasil, pencarian nilai *entropy* hingga nilai *gain* hingga keseluruhan data terpastisi.

c) Lakukan pembentukan pohon keputusan jika sudah mendapatkan nilai *gain*.

Secara garis besar langkah-langkah yang dilakukan oleh Algoritma C.45 dalam membentuk pohon keputusan adalah sebagai berikut [15]:

1. Melakukan simpul akar untuk pohon yang dibuat
2. Jika semua sampel positif, berhenti dengan suatu pohon dengan satu simpul akar, dan berikan tanda (+)
3. Jika semua sampel negatif, berhenti dengan suatu pohon dengan satu simpul akar, dan berikan tanda (-)
4. Apabila atribut kosong, berhenti dengan suatu pohon dengan suatu simpul akar, dengan label sesuai nilai yang terbanyak yang ada dalam label *training*
5. Untuk yang lain, mulai
    a. A --- atribut yang mengklasifikasikan sampel dengan hasil terbaik (berdasarkan *gain* rasio)
    b. Atribut keputusan untuk simpul akar --- A
    c. Pada setiap nilai, vi, yang mungkin untuk A
        1) Tambahkan cabang dibawah akar yang berkaitan dengan A=vi
        2) Tentukan sampel S_{vi} sebagai subset dari sampel yang mempunya nilai vi untuk atribut A
        3) Jika sampel S_{vi} kosong
            i. Dibawah cabang tambahkan simpul daun dengan label = nilai yang terbanyak yang ada pada label training.
            ii. Yang lain tambah cabang baru dibawah cabang yang sekarang.
    d. Berhenti.

2.2.3 Kriteria Algoritma C4.5

Beberapa kriteria yang dimiliki oleh algoritma induktif C4.5 adalah [16]:

1) Atribute-value description

Himpunan data yang digunakan untuk menganalisa harus dapat direpresentasikan dalam bentuk himpunan atribut. Tiap atribut dapat memiliki nilai diskrit maupun kontinu.

2) *Predefined classes*

Kategori yang akan diberikan kepada tiap sampel harus ditentukan terlebih dahulu.

3) Kelas diskrit

Sebuah kasus atau sampel harus tergolong atau tidak tergolong kedalam sebuah kelas tertentu dan jumlah sampel harus jauh lebih besar dari pada jumlah kelas yang ada.

4) Jumalah data yang mencukupi

Jumlah data yang dibutuhkan dipengaruhi oleh jumlah atributdan kelas serta kompleksitas dari model klasifikasi yang digunakan.

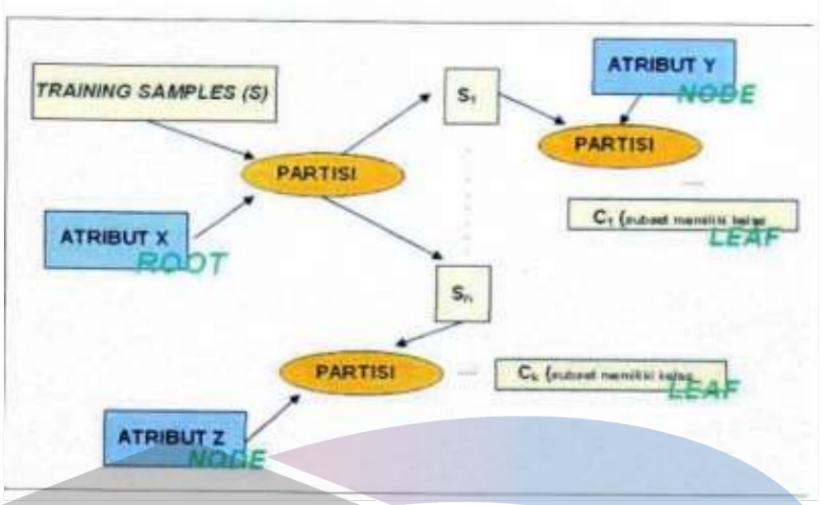
5) Model klasifikasi logis

Pendekatan induktif digunakan untuk membangun *classifer* yang dapat diekspresikan sebagai pohon keputusan atau aturan keputusan.

Misal *training* samples T, atribut ($A_1, A_2, A_3, A_4, \ldots$) dan kelas terdiri dari ($K_1, K_2, K_3, K_4, \ldots$), maka kerangka utama dari algoritma C4.5 dapat diuraikan sebagai berikut:

1) Jika T tidak kosong dan semua sampel yang ada didalamnya memiliki kelas K_1, yang sama maka pohon keputusan untuk T adalah sebuah simpul daun (*leaf node*) dengan label K_1.
2) Jika atribut kosong, maka pohon keputusan berisi sebuah simpul daun dengan label K_j, dimana Kj adalah kelas yang paling dominan pada *training* sampel T.
3) Jika T terdiri dari sampel yang memiliki kelas yang berbeda-bedamaka T dipartisi kedalam $T_1, T_2, T_3, \ldots, T_n$, *training* sampel 7 dipartisi berdasarkan distint value dari atribut Ak yang pada saat itu menjadi *node parent*. Misalkan Ak terdiri dari 4 jenis nilai n_1, n_2, n_3, n_4 maka T akan dipartisi kedalam 3 subset yang itu nilai $A_k = n_1$, $A_K = n_2$, $A_k = n_3$ dan $A_k = n_4$.

Proses ini dilakukan terus secara rekursif dengan *base case* langkah 1 dan 2. Cara untuk mencari atribut yang akan menjadi *node parent* pada suatu iterasi dilakukan dengan menghitung sebuah kriteria yang disebut *gain*. *Gain* berfungsi untuk memilih atribut yang akan diuji berdasarkan konsep teori informasi entropi. Dibawah ini merupakan gambar proses algoritma C4.5.



Gambar 2.5 Proses Algoritma C4.5

2.2.4 Komponen penyusun Algoritma C4.5

Berikut ini akan dijelaskan komponen-komponen yang menyusun algoritma C4.5 dalam membentuk pohon keputusan [16] :

a) Entropi

Entropi merupakan distribusi probabilitas dalam teori informasi dan diadopsi kedalam algoritma C4.5 yang difungsikan untuk mengukur tingkat homogenitas distribusi kelas dataset. Semakin tinggi tingkat entropi dari sebuah dataset maka semakin homogen distribusi kelas pada dataset tersebut. Jika distribusi probabilitas dari kelas didefenisikan dengan $P = (P_1, P_2, P_3, \ldots, Pk)$ maka entropi dapat dituliskan sebagai persamaan dari :

$$E(p1, p2, pk) = \sum_{i=1}^{k} (pi * log2(pi)) \quad (2.1)$$

Persamaan 2. 1 sama dengan persamaan *Info(T)* sebagai berikut :

$$Info(T) = -\sum_{i=1}^{k} \left(\left(\frac{frekuensi(Ci,T)}{|T|} \right) * \log 2 \left(\frac{frequency(Ci,T)}{|T|} \right) \right.$$

Dimana frekuensi (*Ci, T)* adalah jumlah sampel dihimpunan T yang memiliki kelas C_1, C_2, C_3, ... C_k. sebagai contoh, distribusi kelas (0.5, 0.5) lebih homogen bila dibandingkan dengan distribusi (0.67, 0.33) sehingga distribusi (0.5, 0.5) memiliki entropi yang lebih tinggi dari distribusi (0.67, 0.33). Hal ini dapat dibuktikan sebagai berikut :

$$E(0.5, 0.5) = -0.5x \log2(0.5) - 0.5x\log2(0.5) = 1$$

$$E(0.67, 0.33) = -0.67 \times \log2(0.67) - 0.33 \times \log2(0.33) = 0.91$$

Setelah T dipartisi kedalam sejumlah subset *T1, T2, T3, ... Tn berdasarkan* atribut *X* maka perhitungan info dilakukan dengan menggunakan himpunan *training* data yang merupakan hasil partisi sebagai berikut :

$$Info_X(T) = \sum_{i=1}^{n}\left(\frac{T_i}{T}\right) * \text{Info}(T_i))$$

1) *Information Gain*

Entropi dari dataset akan berubah setelah membagi dataset berdasarkan sebuah atribut kedalam subset yang lebih kecil. Perubahan entropi ini dapat digunakan guna membentuk bagus atau tidaknya pembagian data yang telah dilaksanakan. Perubahan entropi ini disebut dengan *information gain* yang diukur dengan menghitung selisih antara entropi dataset sebelum dan sesudah pembagian (*splitting)* dilakukan. Pembagian yang terbaik akan menghasilkan entropi subset yang paling kecil, dengan *informasi gain* yang besar. Jika sebuah dataset D dipartisi berdasarkan nilai dari atribut X dan menghasilkan subset (*Tl, T2,...,Tn)* maka *information gain* dapat dihitung dengan persamaan : *Gain (x) =Info$_x$(T)*. Dimana *info (T)* adalah entropi dari data set sebelum berpartisi berdasarkan atribut X, dan *Info$_X$ (T)* adalah info dari subset setelah dilakukan pemartisian berdasarkan atribut X.

2) *Gain Ratio*

Gain Ratio merupakan normalisasi dari *information gain* yang memperhitungkan entropi dari distribusi probabilitas subset setelah dilakukan proses partisi. Secara matematis, *Gain ratio* dihitung sebagai berikut :

$$GainRatio(X) = \frac{Gain(x)}{Split\inf o(x)} \quad \ldots\ldots (2.5)$$

dimana :

$$SplitInfo(x) = -\sum_{i=1}^{n}\left(\left|\frac{T_1}{T}\right| * \log_2\left(\left|\frac{T_1}{T}\right|\right)\right)$$

$|T_i|$ adalah kardinalitas dari subset *TI* yang berada dalam training data T.

2.3 Confusion Matrix

Confusion Matrix adalah salah satu cara untuk mengukur kinerja suatu learner. Metode ini dapat diterapkan dalam masalah yang mempunyai prediksi binary maupun *multi-class. Output* yang dihasilkan dari *confusion matrix* untuk *n* kelas berupa tabel dengan n^2 sel. Empat kondisi

yang mungkin terjadi pada *confusion* untuk $n = 2$, namun untuk $n > 2$ maka tidak ada kondisi *True Negatif* [17]. Untuk itu perlu dilakukan reduksi, yang mana proses reduksi ini berfungsi untuk mengelompokkan kelas pada confusion matrix multi class kedalam kumpulan kelas seperti confusion matrix binary class. Dengan begitu, nilai TP TN dan metrik lainnya dapat ditentukan [18].

Secara formal, membandingkan klasifikasi aktual dengan klasifikasi yang diprediksi mengungkapkan empat hasil berbeda [19]:

1) Klasifikasi yang sebenarnya adalah positif, seperti klasifikasi yang diperkirakan. Ini dikenal sebagai *'true positive'*, disingkat TP, karena *classifier* mengidentifikasi sampel positif dengan benar.
2) Klasifikasi sebenarnya adalah negatif, dan prediksi klasifikasi adalah negatif. Ini adalah hasil "*true negative* " (TN) karena pengklasifikasi mengidentifikasi sampel negatif dengan benar.
3) Klasifikasi prediksi positif, sedangkan klasifikasi aktual negatif. Ini adalah hasil *'false positive'* (FP) karena classifier salah mengidentifikasi sampel negatif sebagai positif.
4) Klasifikasi prediksi negatif, sedangkan klasifikasi aktual positif. Ini adalah hasil *'false negative'* (FN) karena classifier salah mengidentifikasi sampel positif sebagai negatif.

Keempat hasil ini, dengan interpretasi di atas, sebenarnya berkaitan dengan kelas positif, asalkan kelas ini sangat penting dan pantas mendapat penekanan; itu mengakomodasi apa yang bisa disebut sampel 'relevan', sedangkan kelas negatif dianggap 'tidak relevan'. Hasil TP, TN, FP, dan FN sangat penting dan dirujuk sebagai 'blok bangunan' karena mereka digunakan untuk merumuskan semua ukuran kinerja.

Konsep dari *confusion matrix* dan ukuran kinerja yang relevan yang dibahas dalam untuk klasifikasi biner dapat dengan mudah diperluas ke klasifikasi multi-kelas. Sebagai langkah awal dalam proses generalisasi ini, kami mempertimbangkan masalah klasifikasi 3 kelas (n=3). *Confusion matrix* memiliki tiga baris dan tiga kolom, seperti yang ditunjukkan pada Gambar 2.6. Kelas-kelas tersebut diberi label A, B, dan C. Sebuah sel *matrix* di persimpangan baris i^{th} dan kolom j^{th} diberi nama C_{ij}.

Nilai sel yang ditunjukkan pada Gambar.2.6 hanyalah contoh untuk hasil klasifikasi; ada total N = 150 sampel yang diuji dalam sembilan sel.

Untuk kelas A, kami mendefinisikan:

1) TP_A: Sampel Kelas-A diklasifikasikan dengan benar sebagai kelas A. ini adalah nilai sel c_{11} saja, di persimpangan baris A dan kolom A. Lihat Gambar 2.6.
2) TN_A: Sampel bukan kelas-A (yaitu sampel kelas B atau kelas C) yang diklasifikasikan dengan benar atau salah sebagai bukan kelas A. Ini adalah jumlah dari empat sel c_{22}, c_{23}, c_{32}, dan c_{33}, bagian dari *matrix* yang tersisa setelah menghapus baris A dan kolom A.
3) FP_A: Bukan sampel kelas-A yang salah diklasifikasikan sebagai kelas A. Ini adalah jumlah dari dua sel c_{21}, dan c_{31}, bagian kolom A yang tersisa setelah menghapus sel c_{11}.
4) FN_A: Sampel Kelas-A salah mengklasifikasikan A sebagai tidak kelas A. Ini adalah jumlah dari dua sel c_{12}, dan c_{13}, the bagian dari baris A yang tersisa setelah menghapus sel c_{11} (TP_A).

		Predicted A	Predicted B	Predicted C	
Actual	A	c_{11} 32	c_{12} 10	c_{13} 8	←Row A
	B	c_{21} 9	c_{22} 38	c_{23} 4	←Row B
	C	c_{31} 12	c_{32} 9	c_{33} 28	←Row C
		Column A	Column B	Column C	

Gambar 2.6 Confusion Matrix Untuk Klasifikasi 3 Kelas

		Predicted A	Predicted B	Predicted C
Actual	A	TP_A	FN_A	FN_A
	B	FP_A	TN_A	TN_A
	C	FP_A	TN_A	TN_A

Gambar 2.7 Confusion Matrix Untuk Kelas A

		A	B	C
Actual	A	TN_B	FP_B	TN_B
	B	FN_B	TP_B	FN_B
	C	TN_B	FP_B	TN_B

Gambar 2.8 Confusion Matrix Untuk Kelas B

		A	B	C
Actual	A	TN_C	TN_C	FP_C
	B	TN_C	TN_C	FP_C
	C	FN_C	FN_C	TP_C

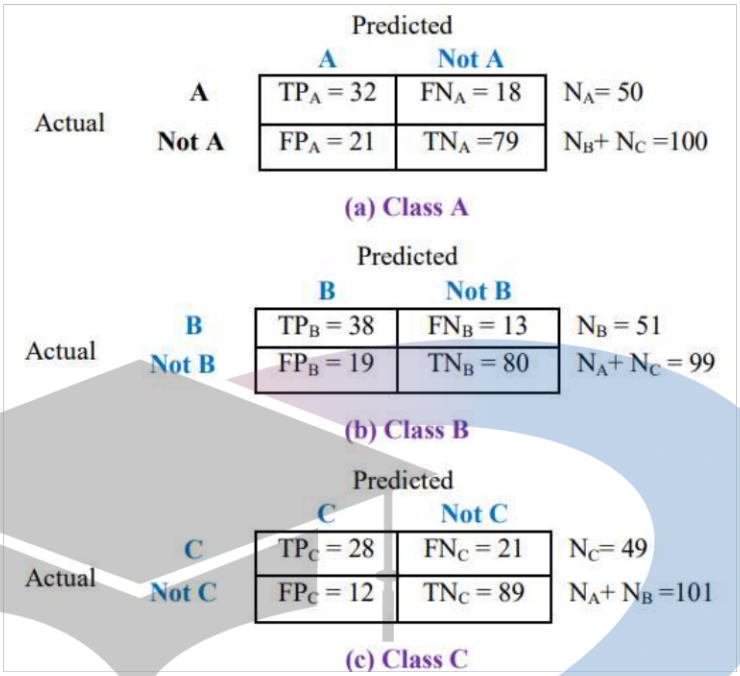
Gambar 2.9 Confusion Matrix Untuk Kelas C

Gambar 2.7 menunjukkan, pada *confusion matriks*, tabel TP_A, TN_A, FP_A, dan FN_A, untuk kelas A. Melalui argumen serupa, kami mendefinisikan tabel untuk kelas B dan C. Untuk kenyamanan, ketika mempertimbangkan satu kelas, kami menganggap kelas ini positif dan dua kelas lainnya negatif.Tabel untuk tiga kelas A, B, dan C dirangkum dalam Gambar. 2.10. Jumlah TP + TN + FP + FN untuk setiap kelas, seperti yang diharapkan, sama dengan jumlah total sampel, N = 150. Sebagai jelas pada Gambar 1, N_A = 50, N = 51_B, dan N_C= 49.

Yang menarik adalah fakta bahwa *confusion matrix* 3x3 dapat didekomposisi menjadi tiga *confusion matrix* 2x2 (analog dengan yang didefinisikan untuk klasifikasi biner). Dengan menggunakan nilai tabel untuk kelas A, B, dan C yang ditunjukkan pada Gambar.2.10, *confusion matrix* 3x3 dari Gambar.2.6 didekomposisi menjadi tiga *confusion matrix* komponen 2x2 untuk kelas A, B, dan C, masing-masing, seperti yang ditunjukkan pada Gambar 2.11.

	TP	TN	FP	FN	
Class A	32	79	21	18	N = 150
Class B	38	80	19	13	N = 150
Class C	28	89	12	21	N = 150

Gambar 2.10 Tabel Untuk Kelas Dalam Klasifikasi 3 Kelas Dari Gambar 2.6



Gambar 2.11 Matrix 3 Komponen 2x2 Untuk Matrix 3x3 Pada Gambar 2.6

Nilai sel dari tiga *matrix* kebingungan 2X2 dari Gambar.2.11 menghasilkan informasi yang sama dengan tiga baris tabel dari Gambar.2.10 tabel dari masing-masing kelas. Di sini, dalam arti tertentu, kita dapat memvisualisasikan bahwa masalah klasifikasi 3 kelas diubah menjadi tiga masalah klasifikasi *biner*.

Rumus untuk mencari nilai akurasi [20]:

a. Akurasi :menghitung keakuratan sistem mengklasifikasikan data dengan tepat .

$$\text{Akurasi} \quad \frac{TPA+TPB+TPC}{N}$$

b. *Recall* : menghitung seberapa banyak nilai kebenaran (positif) dari dataset yang memang bernilai benar (positif) muncul.

$$\text{Recall} \quad \frac{TP}{TP+FN}$$

c. *Precision* : ketepatan nilai kebenaran dalam klasifikasi sesuai nilai kebenaran yang sesungguhnya.

$$\textit{Precision} \quad \frac{TP}{TP+Fp}$$

2.4 Performa Mahasiswa

Performa dapat diartikan sebagai tindakan dalam melaksanakan rangkaian tugas sebagai misi atau disebut juga sebagai hasil dari suatu kegiatan selama suatu periode waktu tertentu. Performa akademik sebagai keyakinan pada diri individu untuk mengarahkan dirinya melalui kemampuan mengontrol proses berfikir, motivasi, dan tindakannya yang bertujuan untuk menguasai situasi tertentu yang mengarah pada ilmu pengetahuan yang bersifat ilmiah dan telah diuji kebenarannya, sehingga bisa diukur baik berupa nilai maupun yang biasanya.

Adapun ciri-ciri performa akademik dapat dilihat melalui tiga dimensi yaitu: *magnitude* atau tingkat kesulitan tugas yang dihadapi individu; *generality* atau kemampuan individu dalam menguasai suatu tugas; dan *strength* atau kekuatan dan kemantapan individu terhadap keyakinannya. Adapun secara terperincinya adalah sebagai berikut [21]:

a) *Magnitude*, Dimensi ini berkaitan dengan derajat kesulitan tugas yang ditujukan kepada individu. Apabila tugas-tugas yang dibebankan pada individu disusun 3 menurut tingkat kesulitannya, maka perbedaan efikasi diri secara individual mungkin terbatas pada tugas-tugas yang tergolong sederhana, menengah atau berat.
b) *Generality*, Dimensi ini berkaitan dengan penguasaan individu terhadap bidang atau tugas pekerjaan. Individu dapat menyatakan dirinya memiliki efikasi diri pada aktivitas yang luas, atau terbatas pada fungsi *domain* tertentu saja.
c) *Strength*, Dimensi yang ketiga ini lebih menekankan pada tingkat kekuatan atau kemantapan individu terhadap keyakinannya. Efikasi diri menunjukkan bahwa tindakan yang dilakukan individu akan memberikan hasil yang sesuai dengan yang diharapkan individu.

Ada 3 aspek yang dapat mempengaruhi performa mahaiswa antara lain:

1) aspek kognitif, efikasi diri performa akademik memengaruhi individu dalam memandang kemampuan dirinya, terutama dalam penyelesaian tugas-tugas yang sulit. Individu yang memiliki keyakinan terhadap kemampuannya dalam menyelesaikan tugastugas akademik tidak akan menghindar dari tugas-tugas yang sulit, dan merasa percaya diri mampu menyelesaikan tugas-tugas tersebut meskipun sifatnya cenderung suli.

2) Aspek afektif, individu yang memiliki keyakinan akademik dan kepercayaan diri tinggi, akan berusaha keras, gigih, dan ulet, serta sangat yakin dapat menyelesaikan tugas-tugas akademik yang sulit.
3) Aspek perilaku, efikasi diri akademik dan kepercayaan diri yang tinggi akan membuat individu tetap menampilkan kemampuan terbaiknya dalam berbagai situasi dan kegiatan akademik.

2.5 Dataset

1) Defenisi Dataset

Dataset adalah suatu *database* dalam memory. Dataset mempunyai semua karakteristik, fitur dan fungsi dari *database* biasa. Dataset memiliki banyak tabel, dan tabel-tabel memiliki hubungan (*relationship*). Tabel-tabel pada suatu dataset dapat memiliki *foreign key* dan integritas referensial. Dataset adalah objek yang merepresentasikan data dan relasinya pada *memory*. Struktur dalam dataset sama dengan data yang ada pada *database*. Dataset berisi koleksi dari data tabel dan data [22].

2) Jenis-jenis Dataset

Dataset memiliki 2 macam jenis antara lain yaitu *private* dataset (dataset primer) dan *public* dataset (dataset *sekunder*).

a) *Private* dataset (dataset primer)

Private dataset adalah dataset yang dapat diambil dari organisasi yang kita jadikan tempat atau objek penelitian. Adapun contoh-contohnya seperti instansi, rumah sakit, pabrik, perusahaan jasa dan lain-lain.

b) *Public* dataset (dataset *sekunder*)

Public dataset adalah dataset yang dapat diambil dari *repository* publik yang telah disepakati oleh para peneliti. Adapun contoh-contohnya seperti : *UCI Machine learning*, *kaggle* dan *student performance.*

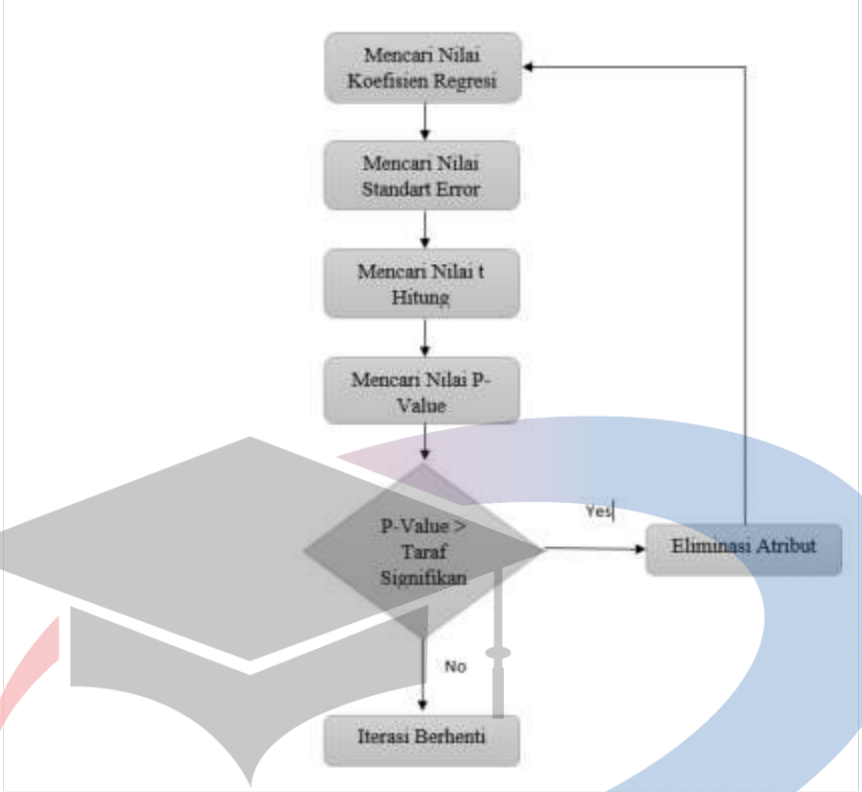
2.6 Backward Elimination

Backward Elimination merupakan metode yang bisa menghapus atribut yang tidak signifikan dari model. Berfungsi sebagai seleksi atribut dengan memanfaatkan regresi statistik untuk mengetahui korelasi setiap kombinasi atribut dengan target. Semakin kecil *significance* level, maka semakin ketat pemilihan atribut yang akan terpilih sehingga semakin sedikit atribut yang terpilih sebagai model. Metode *Backward Elimination* merupakan metode regresi yang baik

karena pada metode ini menggambarkan variabel perilaku respon yang terbaik dengan memilih variabel penjelas dari banyak variabel penjelas yang tersedia pada data. Membentuk persamaan regresi linier dengan metode *Backward Elimination* [23]:

a) Meregresikan seluruh variabel bebas dengan variabel terikat, selanjutnya lihat
b) pada uji parsial dan bandingkan nilai p *value* setiap variabel dengan nilai taraf signifikan.
c) Keluarkan variabel yang memiliki nilai p *value* < taraf signifikan dari persamaan regresi.
d) Meregresikan variabel bebas yang tersisa dengan variabel terikat.
e) Perhatikan hasil uji parsialnya. Jika terdapat nilai p *value* dari variabel bebas yang < taraf signifikan maka kembali ke langkah b. Sedangkan jika nilai p value setiap variabel bebas > taraf signifikan maka lanjutkan ke langkah berikutnya.
f) Menentukan model regresi yang signifikan secara parsial dan serentak.
g) Melakukan pengujian asumsi klasik untuk regresi linier berganda.
h) Menentukan nilai koefisien determinasi untuk model dengan metode *Backward Elimination.*

Dibawah ini adalah gambar dari alur *Backward Elimination*:



Gambar 2.12 Alur Backward Elimination

Berikut adalah penjelasan alur perhitungan Teknik *Backward Elimination*:

1. Mencari nilai Koefisien Regresi
2. Menghitung nilai Standart Error pada masing-masing atribut
3. Mencari nilai t hitung untuk masingmasing atribu
4. Selanjutnya menentukan nilai *P-Value* berdasarkan pada nilai t table
5. Menentukan nilai *P-Value* terbesar dari masing-masing atribut
6. Selanjutnya membandingkan nilai antara *P-Value* terbesar dengan taraf signifikan, apabila nilai *P-Value* lebih besar dari taraf signifikannya maka atribut dengan *P-Value* terbesar tersebut akan dieliminasi dan tahap selanjutnya adalah mengulangi langkah pertama kembali sampai nilai *P-Value* pada setiap atribut yang tersisa kurang dari taraf signifikannya.

2.7 Penelitian Terdahulu

Penelitian terdahulu merupakan acuan yang dibutuhkan seorang peneliti untuk melakukan penelitian. Penelitian terdahulu pada penelitian ini diambil berdasarkan kesamaan metode yang

digunakan yaitu Algoritma C4.5. Banyak penelitian menggunakan metode ini dalam berbagai kasus yang dirangkum dalam tabel dibawah ini, yaitu : [5], [6], [10], [15], [24]:

Tabel 2.1 Penelitian Terdahulu

No	Peneliti & Tahun Penelitian	Judul Penelitian	Kesimpulan
1	Dini Rizky Sitorus P, Agus Perdana windarto, Dedy Hartama, Irfan Sudahri Damanik (2019).	Penerapan klasifikasi C4.5 dalam meningkatkan sistem pembelajaran mahasiswa	Berdasarkan hasil penelitian yang dilakukan dapat disimpulkan bahwa penerapan klasifikasi C4.5 pada peningkatan sistem pembelajaran mahasiswa di STIKOM Tunas Bangsa dapat diterapkan. Atribut yang digunakansebagai parameter peningkatan sistem pembelajaran mahasiswa antara lain: C1 (Sistem Pengajaran), C2 (Alat Peraga),C3 (Lingkungan), C4 (Sarana Prasarana) dan C5 (Pemberian Tugas). Hasil perhitungan menyebutkan atribut C5 (Pemberian Tugas) merupakan variabel yang paling berpengaruh terhadap peningkatan sistem pembelajaran mahasiswa. Pengujian juga dilakukan untuk membuktikan apakah metode C4.5 dapat diterapkan pada kasus peningkatan sistem

			pembelajaran mahasiswa dengan menggunakan bantuan software *Rapidminer* dan diperoleh akurasi 95%.
2	Khairunnissa Fanny Irnanda, Dedy Hartama, Agus Perdana Windarto (2021).	Analisa Klasifikasi C4.5 Terhadap Faktor Penyebab Menurunnya Prestasi Belajar Mahasiswa Pada Masa Pandemi.	Berdasarkan hasil penelitian yang telah dilakukan dapat disimpulkan bahwa penerapan klasifikasi C4.5 pada faktor penyebab menurunnya prestasi belajar mahasiswa di masa pandemi dapat diterapkan. Atribut yang digunakan sebagai parameter peningkatan sistem pembelajaran mahasiswa antara lain: Cara Belajar (C1), Waktu Belajar (C2), Pemahaman Materi (C3), Pemberian Tugas (C4), Lingkungan (C5). Hasil perhitungan menyebutkan atribut C3 (Pemahaman Materi) merupakan variabel yang paling berpengaruh terhadap menurunnya prestasi.
3	Daniel Sinaga, Edwin J. Solaiman, Fergie Joanda Kaunang (2021).	Penerapan Algoritma *Decision Tree* C4.5 Untuk Klasifikasi Mahasiswa Berpotensi *Drop Out* di Universitas Advent Indonesia.	Berdasarkan hasil tersebut, atribut IP Semester merupakan hal yang sangat berpengaruh dalam klasifikasi mahasiswa berpotensi *drop out*. Dapat dilihat bahwa mahasiswa yang memiliki IP semesterdengan kategori *Under*

			yaitu dengan IP 2.33 – 2.75 diklasifikan sebagai mahasiswa berpotensi *drop out* sedangkan mahasiswa dengan kategori Sangat Memuaskan dan Dengan Pujian tidak berpotensi *drop out*.
4	Eka pandu cynthia, Adi Ismanto (2018).	Metode *decision tree* algoritma C4.5 dalam mengklasifikasi data penjualan bisnis gerai makanan cepat saji.	Dari hasil percobaan pencarian pohon hasil keputusan dari data penjualan gerai makanan cepat saji menggunakan algoritma C4.5 dihasilkan nilai *entropy* dan *gain* tertinggi yaitu 1,501991 pada atribut-atribut Menu Makanan pada perhitungan manual. Sedangkan menggunakan aplikasi *Rapidminer* diperoleh hasil pohon keputusan seperti terlihat pada Gambar 3.2. Harga – Jumlah Terjual – Menu Makanan (Rice Bento = Kurang Laris, Dada = Laris) dengan bobot (weight) masing-masing atribut : Harga (0,738), Jenis Menu (0,067), Jumlah Terjual (0,156), Status Penjualan (0,040).
5	Sirli Fahriah, Wikta sari (2021).	Algoritma C4.5 Berbasis *Forward Selection* Untuk Klasifikasi Bidang Minat Studi Mahasiswa Teknik Informatika.	Hasil pengujian dari kinerja algoritma C4.5 dan C4.5 berbasis *forward selection* dengan 540 data menghasilkan akurasi tertinggi pada algoritma C4.5

			berbasis *forward selection* dengan hasil akurasi 85% dengan meningkatkan akurasi pada metode algoritma C4.5 yang sebelumnya 84,33% dan C4.5 berbasis *forward selection* 85%.