BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1. Penelitian Terkait

Kualitas Jasa dan Pelayanan merupakan salah satu unsur yang sangat penting dalam menciptakan kepuasan konsumen. Salah satu cara untuk menempatkan hasil pelayanan yang lebih unggul daripada pesaing adalah dengan memberikan pelayanan yang baik, efisien, dan cepat [11]. Umumnya konsumen melihat dari suatu pelayanan yang diberikan oleh perusahaan kepada pelanggan. Jika pelayanan yang diberikan sangat memuaskan dan mencapai tingkat kepuasan konsumen, maka konsumen akan merasa puas akan jasa atau pelayanan tersebut. Semua pelayanan dan fasilitas yang diberikan harus disesuaikan dengan kebutuhan konsumen dan dievaluasi melalui opini atau presepsi konsumen. Opini atau presepsi konsumen merupakan suatu penilaian terhadap kelebihan atau kekurangan suatu jasa atau pelayanan.

Atas dasar presepsi konsumen yang merupakan suatu penilaian terhadap jasa atau pelayanan suatu perusahaan, maka perusahaan harus melakukan peningkatan kualitas pelayanan yang diharapkan akan semakin meningkatnya konsumen untuk menggunakan jasa atau layanan yang di tawarkan oleh suatu perusahaan yang diharapkan dapat memenuhi kebutuhan setiap konsumen sehingga konsumen dapat merasa puas.

Berdasarkan kualitas jasa dan kepuasan pelanggan yang saling berkaitan maka dilakukanlah penelitian ini dengan terlebih dahulu melakukan studi kepuasan pelanggan dari penelitian-penelitian sebelumnya dan sumber lain. Dari penelitian-penelitian sebelumnya penulis menemukan beberapa penelitian yang membahas tentang topik yang terkait dengan penelitian penulis, antara lain adalah algortima yang akan digunakan oleh penulis pada penelitian ini.

Penelitian pertama dilakukan oleh Ibnu Fatchur Rochman [6] yang membuat penelitian pada kepuasan pelanggan perum damri menggunkan algoritma C4.5. Dari hasil pengujian algortima C4.5 dalam memprediksi

kepuasan pelanggan perum DAMRI atas 90 sample data pelanggan yang diuji dalam penelitian ini, menunjukan bahwa algoritma C4.5 memiliki tingkat akurasi yang cukup tinggi yaitu sebesar 93%. Pada penelitian ini peneliti membuat kuisioner dengan jumlah yang telah di tentukan sejumlah 150 kuisioner yang selanjutnya kuisioner tersebut akan menjadi bahan acuan awal untuk menentukan jumlah puas dan tidak puas pada prosentasi kepuasan pelanggan Bus Perum DAMRI. Hasil dari perhitungan manual jumlah puas dan tidak puas pada kuisioner tersebut dimasukkan kedalam rumus algoritma C4.5.

Kuisioner di bagi dalam tiga kategori yang masing-masing kategori memiliki beberapa pertanyaan, kategori kuisioner yang diajukan oleh peneliti meliputi Harga, Pelayanan, dan Fasilitas.

Peneliti melakukan 3 kali pengujian terhadap data pelanggan dengan jumlah data testing dan data training yang berbeda yaitu :

a. Data training 40% dan data testing 60%

b. Data training 60% dan data testing 40%

c. Data training 80% dan data testing 20%

Setelah dilakukan penelitian dan percobaan sebanyak 3 kali, maka dapat disimpulkan sebagai berikut :

a. Dari percobaan yang telah dilakukan penulis sebanyak 3x, maka dapat di ketauhi bahwa percobaan 1, 2, dan 3 ini dapat dikatakan baik dan berhasil, karena sudah terlihat jelas bahwa nilai akurasi yang terus bertambah dan semakin akurat.
b. Algoritma C4.5 pada kepuasan pelanggan di Perum DAMRI dapat diterapkan dengan baik.

Penelitian kedua dilakukan oleh Teguh Budi Santoso [7] Penelitian ini meneliti tentang prediksi loyalitas pelanggan data seluler menggunakan metode klasifikasi dengan Algoritma C4.5 dan hasil klasifikasi menggunakan algortima C4.5 menunjukan bahwa diperoleh akurasi mencapai 97,5% yang menunjukan bahwa algoritma C4.5 cocok digunakan untuk mengukur tingkat loyalitas pelanggan data seluler. Data yang digunakan adalah data

primer dari penyebaran kuisioner berupa penyataan embentukan model prediksi menggunakan metode C4.5. pada algoritma C4.5 dilakukan perhitungan entropy dan information gain dimana atribut loyalitas pelanggan sebagai atribut tujuan, sedangkan harga, pelayanan, promosi, citra perusahaan, dan kepercayaan sebagai atribut sumber untuk memperoleh node akar dan node lainnya.

Tahap pertama peneliti melakukan perhitungan nilai entropy dan information gain terhadap 40 sample, pada tahap selanjutnya peneliti membagi data dari hasil yang diperoleh dari konsep algoritma C4.5 menjadi 2 bagian yaitu data testing dan data training. Berdasarkan analisa penggunaan data minning dengan algoritma C4.5 dapat digunakan pada data set pelanggan kedalam kegiatan manajemen strategi sehingga dapat menahan selama mungkin pelanggannya dengan baik.

Selanjutnya, Penelitian yang dilakukan oleh David Hartanto dan Seng Hansun [8] meneliti tentang Tingkat Kelulusan Mahasiswa yang akan di prediksi menggunakan Algorita C4.5. Peneliti menggunakan 100 data yang diperoleh dari department IT Universitas multimedia nusantara program studi Teknik Informatika.

Dari hasil uji coba terhadap 100 data peneliti mendapatkan tingkat akurasi dari hasil prediksi kelulusan terhadap data testing sebesar 87.5%. Peneliti menyimpulkan bahwa IPS semester 6 merupakan attribute yang paling berpengaruh dari keputusan yang ada. Berdasarkan penelitian ini terbukti bahwa datamining dengan Algoritma C4.5 dapat di Implementasikan untuk memprediksi tingkat kelulusan mahasiswa. Dan hasil prediksi kelulusan dapat membantu bagian program studi untuk mengetahui status kelulusan mahasiswa.

Penelitian keempat dilakukan oleh Anik Andriani [9] dari AMIK BSI Jakarta meneliti tentang mahasiswa yang dinyatakan layak untuk melanjutkan studi atau harus dinyatakan putus kuliah atau *dropout* (DO). Penelitian ini menggunakan metode klasifikasi dengan menggunakan Algoritma C4.5 Dalam proses Klasifikasi peneliti menggunakan beberapa atribut data antara lain nama, nim, jenis kelamin, Usia masuk, Asal daerah, Jurusan SLTA,

status orangtua, penghasilan Orangtua/wali, waktu kuliah, IPK Semester 1, Kehadiran semester 1, Status Beasiswa, Biaya Studi, status Bekerja., peneliti menguji data menggunakan *confusion matrix* dan kurva ROC

Hasil evaluasi dan validasi dengan confussion matrix menunjukkan tingkat akurasi pada algoritma C4.5 sebesar 97,75%. Hasil dari penelitian ini menunjukan nilai lebih dari 0,9 sehingga penelitian ini dapat dikategorika sebagai excellent classification.

Penelitian yang dilakukan oleh Dyah Satiti, Sucipto dan Shyntia Atica [10] tentang analisis preferensi konsumen waralaba makanan cepat saji dengan menggunakan pendekatan data mining di restoran x Surabaya. Suatu Restoran pastinya membutuhkan strategi pemasaran yang tepat dengan mengetahui preferensi konsumen sebagai upaya mempertahankan posisi di tengah persaingan restoran cepat saji. Maka Dyah satiti melakukan penelitian ini yang bertujuan untuk mengetahui urutan atribut-atribut, preferensi pelanggan berdasarkan segmen, dan segmen priorita restoran X di Surabaya.

Peneliti menggunakan pendekatan data mining menggunakan metode K-means cluster analysis untuk memperoleh segmen konsumen berdasar karakterpenilaian preferensi dan neural network backpropgataion untuk membuat model pengenalan pola preferensi konsumen. Penelitian ini menunjukkan Hasil lima atribut yang penting bagi konsumen yaitu :

a. lokasi yang strategis
b. suasana yang bersih dan rapi
c. adanya areal parker
d. suasanya nyaman
e. serta rasa makanan yang lezat

Begitu juga ada tiga segmen konsumen yaitu :

a. konsumen penyuka pelayanan yang ramah
b. konsumen yang kritis dan,
c. konsumen penyuka suasana yang bersih dan nyaman.

Segmen prioritas restoran X adalah segmen ketiga (konsumen penyuka suasana yang bersih dan nyaman) dengan anggota terbanyak yakni 49,5%. Selain itu, segmen kedua (konsumen yang kritis) perlu dipertimbangkan

melihat anggotanya sebesar 45,5%. Dibutuhkan perbaikan kualitas kondisi restoran, menu makanan, pelayanan dan intensitas promosi untuk membidik dua segmen ini.

Tabel 2.1 Tabel Penelitian Terkait

No	Penulis	Topik	Metode	Variabel	Hasil
1	Ibnu Fatchur Rochman	Prediksi Kepuasan Pelanggan di perum DAMRI	Algoritma C4.5	Harga Fasilitas Pelayanan	Dari hasil pengujian algortima C4.5 dalam memprediksi kepuasan pelanggan perum DAMRI atas 90 sample data pelanggan yang diuji dalam penelitian ini, menunjukan bahwa algoritma C4.5 memiliki tingkat akurasi yang cukup tinggi yaitu sebesar 93%.
2	Teguh Budi Santoso	Analisa dan prediksi Loyalitas pelanggan data seluler	Algoritma C4.5	Usia Pelayanan Promosi Harga Citra Perusahaan Kepercayaan	hasil klasifikasi menggunakan algortima C4.5 menunjukan bahwa diperoleh akurasi mencapai 97,5% yang menunjukan bahwa algoritma C4.5 cocok digunakan untuk mengukur tingkat loyalitas pelanggan data seluler.
3	David Hartanto kamagi dan Seng Hansun	Implementasi data mining yang di implementasikan untuk memprediksi kelulusan mahasiswa	Algoritma C4.5	IPS1,IPS2 IPS3,IPS4 IPS5,IPS6 Jumlah SKS Ketepatan Lulus	Dari hasil uji coba terhadap 100 data peneliti mendapatkan tingkat akurasi dari hasil prediksi kelulusan terhadap data testing sebesar 87.5%. Peneliti menyimpulkan bahwa IPS semester 6 merupakan attribute yang paling berpengaruh dari keputusan yang ada.
4	Anik Andriani	Penerapan metode klasifikasi untuk mengklasifikasi	Algoritma C4.5	Waktu Kuliah IPK Smt 1 Kehadiran Smt 1	Hasil evaluasi dan validasi dengan confussion matrix menunjukkan tingkat akurasi pada algoritma C4.5 sebesar 97,75%. Hasil dari

		mahasiswa dropout		Status Orang tua Penghasilan Orang tua Beasiswa	penelitian ini menunjukan nilai lebih dari 0,9 sehingga penelitian ini dapat dikategorika sebagai excellent classification.
5	Dyah Satiti, Sucipto, Shyntia Atica Putri	Analisis preferensi konsumen waralaba terhadap makanan cepat saji	K-Means	Kondisi Restoran Menu makanan Jenis Pelayanan Bentuk Pemasaran	Terdapat 2 Segmen, Segmen prioritas restoran X adalah segmen ketiga (konsumen penyuka suasana yang bersih dan nyaman) dengan anggota terbanyak yakni 49,5%. Selain itu, segmen kedua (konsumen yang kritis) perlu dipertimbangkan melihat anggotanya sebesar 45,5%. Dibutuhkan perbaikan kualitas kondisi restoran, menu makanan, pelayanan dan intensitas promosi untuk membidik dua segmen ini.

2.2. Literatur yang Mendukung Penelitian

2.2.1 Kualitas Jasa dan Pelayanan

Pelayanan merupakan salah satu unsur yang sangat penting dalam menciptakan kepuasan konsumen. Agar harapan konsumen terpenuhi, perusahaan harus memberikan pelayanan yang berkualitas. Kualitas dapat diartikan sebagai pengukuran seberapa baik tingkat pelayanan yang diberikan dan sesuai dengan harapan konsumen, jadi dengan kata lain memberikan pelayanan berkualitas berarti menyesuaikan diri dengan harapan konsumen. Ini merupakan salah satu faktor keberhasilan dalam persaingan yang makin ketat. Pelayanan yang berkualitas adalah orientasi semua sumber daya manusia dalam suatu perusahaan terhadap kepuasan pelanggan [11].

Definisi kualitas jasa ada beberapa macam antara lain:

Menurut Wirasasmita, Sitorus dan Manurung [11], definisi kualitas jasa adalah:

"Suatu sifat atau ciri yang membedakan nilai dari suatu barang atau jasa dengan nilai dari barang atau jasa yang lain yang sejenis".

2.2.2 Kepuasan Konsumen

Kepuasan konsumen merupakan hal yang sangat penting dalam industri jasa. Karena dalam industri jasa, pelayanan yang dapat memuaskan konsumen akan memberikan imbalan yang menguntungkan, serta meningkatkan daya saing perusahaan.

Kotler [12] mendefinisikan kepuasan pelanggan adalah:

"Satisfaction is a person's feelings of pleasure or disappointment resulting from comparing a product's perceived performance (or outcome) in relation to his or her expectations. "

Secara umum kepuasan konsumen dan ketidakpuasan konsumen merupakan hasil dari perbedaan antara harapan dengan kinerja yang dirasakan oleh konsumen, Atau dengan kata lain ada dua kemungkinan yang akan terjadi, yaitu:

1. Kinerja yang dirasakan konsumen lebih besar dari yang diharapkan, artinya konsumen merasa puas dengan kualitas pelayanan yang diberikan oleh perusahaan
2. Kinerja yang dirasakan konsumen lebih kecil dari yang diharapkan, artinya konsumen tidak puas dengan kualitas pelayanan yang diberikan perusahaan.

2.2.3 Loyalitas Pelanggan

Loyalitas pelanggan secara umum dapat diartikan kesetiaan seseorang atas suatu produk, baik barang maupun jasa tertentu. Istilah loyalitas pelanggan menurut Swastha [21] sebetulnya berasal dari loyalitas merek yang mencerminkan loyalitas pelanggan pada merek tertentu. Pelanggan yang setia pada merek tertentu cenderung terikat pada merek tersebut dan akan membeli produk yang sama lagi sekalipun tersedia banyak alternatif lainnya.

2.2.4 Variabel Kuisioner

1. Harga

   Menurut Basu Swastha definisi dari harga adalah “sejumlah uang yang dibutuhkan untuk mendapat sejumlah kombinasi dari barang beserta pelayannya”[21].

2. Kualitas Pelayanan

   Kualitas Pelayanan adalah seberapa jauh perbedaan antara kenyataan dan harapan pelanggan atas layanan yang mereka terima. Terdapat lima dimensi dalam kualitas pelayanan yaitu tangibles, reliability, responsiveness, assurance,dan empathy yaitu:

   a. Berwujud (tangible)

      Yaitu kemampuan suatu perusahaan dalam menunjukkan eksistensinya kepada pihak eksternal. Penampilan dan kemampuan sarana dan prasarana fisik perusahaan yang dapat diandalkan keadaan lingkungan sekitarnya merupakan bukti nyata dari layanan yang diberikan oleh para pemberi jasa. Hal ini meliputi fasilitas fisik (contoh : gedung, gudang dan lain-lain), perlengkapan dan peralatan yang digunakan (teknologi) serta penampilan pegawainya.

   b. Keandalan (reliability)

      Yaitu kemampuan perusahaan untuk memberikan layanan sesuai dengan dijanjikan secara akurat dan terpercaya. Kinerja harus sesuai dengan harapan pelanggan yang berarti ketepatan waktu, layanan yang sama untuk semua pelanggan tanpa kesalahan, sikap yang simpatik dan dengan akurasi yang tinggi.

   c. Ketanggapan (responsiveness)

      Yaitu suatu kebijakan untuk membantu dan memberikan layanan yang cepat (responsive) dan tepat kepada pelanggan dengan penyampaian informasi yang jelas. Membiarkan konsumen menunggu, persepsi yang negatif dalam kualitas layanan.

   d. Jaminan dan kepastian (assurance)

      Yaitu pengetahuan, kesopansantunan dan kemampuan para pegawai perusahaan untuk menumbuhkan rasa percaya para pelanggan kepada

perusahaan. Hal ini meliputi beberapa komponen anatara lain komunikasi (communication), kredibilitas (credibility), keamanan (security), kompetensi (competence) dan sopan santun (courtesy).

e. Empati (empathy)

Yaitu memberikan perhatian yang tulus dan bersifat individual atau pribadi yang diberikan kepada para pelanggan dengan berupaya memahami keinginan konsumen. Dimana suatu perusahaan diharapkan memilki pengertian dan pengetahuan tentang pelanggan, memahami kebutuhan pelanggan secara spesifik, serta memiliki waktu pengoperasian yang nyaman bagi pelanggan.

3. Fasilitas

Menurut Kotler [12], mendefinisikan fasilitas yaitu segala sesuatu yang bersifat peralatan fisik dan disediakan oleh pihak penjual jasa untuk mendukung kenyamanan konsumen.

4. Loyalitas Konsumen

a. Behaviour

Keinginan konsumen untuk menggunakan taksi KOSTI di masa yang akan datang (Repurchase Behaviour)

Kecenderungan niat konsumen untuk selalu menggunakan taksi KOSTI disaat ingin menggunakan jasa transportasi taksi (Repeat Purchase Intensions)

b. Attitude

Niat konsumen untuk merekomendasikan taksi KOSTI kepada orang lain (word of mouth)

Niat konsumen untuk mengatakan hal-hal positif tentang taksi KOSTI kepada orang lain

Niat konsumen untuk mendorong orang lain agar menggunakan taksi KOSTI

c. Cognitive

Kerelaan konsumen untuk tetap menggunakan taksi KOSTI walaupun harga untuk menggunakan taksi KOSTI lebih mahal (Willingness to pay more)

Komitmen dari konsumen bahwa harga bukanlah masalah yang penting, dan akan tetap lebih memilih taksi KOSTI (Preference)

Kecenderungan niat konsumen untuk selalu menggunakan taksi KOSTI dan tidak mau menggunakan taksi merk lain (choice reduction behavior)

Kecenderungan untuk menempatkan taksi KOSTI sebagai pilihan utama (first choice in mind)

2.2.5 Desain Kuesioner dan Skala Pengukuran

Untuk memperoleh data tentang variable perlu menggunakan kuesioner. Kuesioner adalah alat ukur yang terdiri dari sejumlah pertanyaan atau pernyataan tertulis yang harus dijawab atau diisi oleh responden [19].Ada tiga macam format dasar yang digunakan dalam kuesioner yaitu :

1. Close Ended Questions

   Format ini berisi pertanyaan yang memberikan pilihan respon di dalam kuesioner.

2. Open Ended Questions

   Format pertanyaan yang tidak memberikan pilihan respon kepada responden. Responden diminta untuk mengisi pertanyaan dengan kata-kata nya sendiri.

3. Scale Response Questions

   Format ini menggunakan skala untuk mengukur respon konsumen atas pelayanan yang diberikan.

Dalam penelitian pengukuran perilaku reponden yang sifatnya subjektif tidak dapat diukur secara langsung karena menyangkut aspek mental, untuk itu digunakan skala. Skala tersebut akan menunjukkan hasil berupa angka yang diperoleh dari suatu proses pengukuran [20].

Ada 2 skala pengukuran yang dapat digunakan :

1. Skala Nominal

Skala yang paling sederhana dimana angka yang diberikan kepada suatu kategori lainnya, hanya berupa kode atau label

Contoh : *gender* atau status

2. Skala Interval

Skala yang memiliki jarak yang tetap antar respon yang ditawarkan, biasanya 1 unit skala [19].

2.2.6 Data Mining

Data mining [13] 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. Dalam data mining terdapat dua pendekatan metode pelatihan, yaitu [14]:

a. *Unsupervised learning*, metode ini diterapkan tanpa adanya latihan *(training)* dan tanpa ada guru *(teacher).* Guru di sini adalah label dari data.
b. Supervised learning, yaitu metode belajar dengan adanya latihan dan pelatih. Dalam pendekatan ini, untuk menemukan fungsi keputusan, fungsi pemisah atau fungsi regresi, digunakan beberapa contoh data yang mempunyai output atau label selama proses training.

Ada beberapa teknik yang dimiliki *data mining* berdasarkan tugas yang bisa dilakukan, setiap teknik memiliki algoritma masing-masing. Teknik dalam data mining terbagi menjadi enam kategori, yaitu [16] :

a. Deskripsi

   Para peneliti biasanya mencoba menemukan cara untuk mendeskripsikan pola dan trend yang tersembunyi dalam data.
b. Estimasi

   Estimasi mirip dengan klasifikasi, kecuali variabel tujuan yang lebih kearah numerik dari pada kategori.
c. Prediksi

   Prediksi memiliki kemiripan dengan estimasi dan klasifikasi. Hanya saja, prediksi hasilnya menunjukkan sesuatu yang belum terjadi (mungkin terjadi dimasa depan).

d. Klasifikasi

Dalam klasifikasi variabel, tujuan bersifat kategorik. Misalnya, kita akan mengklasifikasikan pendapatan dalam tiga kelas, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

e. Klastering

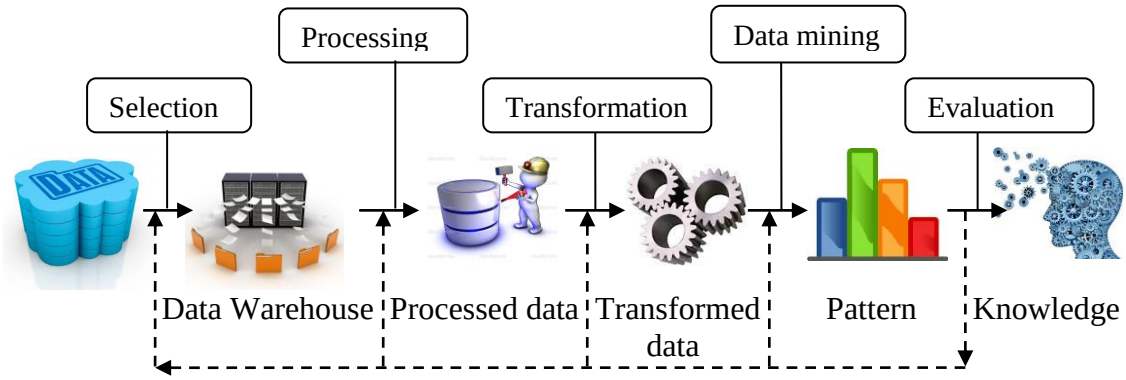
Clustering lebih ke arah pengelompokan *record*, pengamatan, atau kasusdalam kelas yang memiliki kemiripan.

f. Asosiasi

Mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu.

2.2.6.1 Tahap-tahap Data Mining

Sebagai suatu rangkaian proses, data mining dapat dibagi menjadi beberapa tahap. Tahap-tahap tersebut bersifat interaktif di mana pemakai terlibat langsung atau dengan perantaraan knowledge base [17].



Gambar 2.1 : Tahap – tahap data mining

Tahap-tahap data mining yaitu :

1. Pembersihan data (*data cleaning*)

Pembersihan data merupakan proses menghilangkan *noise* dan data yang tidak konsisten atau data tidak relevan. Pada umumnya data yang diperoleh, baik dari database suatu perusahaan maupun hasil eksperimen, memiliki isian-isian yang tidak sempurna seperti data yang hilang, data yang tidak valid atau juga hanya sekedar salah ketik.Selain itu, ada juga atribut-atribut data yang tidak relevan dengan hipotesa data mining yang dimiliki. Data-data yang tidak relevan itu juga lebih baik dibuang. Pembersihan data

juga akan mempengaruhi performasi dari teknik data mining karena data yang ditangani akan berkurang jumlah dan kompleksitasnya.

2. Integrasi data (*data integration*)

Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru. Tidak jarang data yang diperlukan untuk data mining tidak hanya berasal dari satu database tetapi juga berasal dari beberapa database atau file teks. Integrasi data dilakukan pada atribut-aribut yang mengidentifikasikan entitas-entitas yang unik seperti atribut nama, jenis produk, nomor pelanggan dan lainnya. Integrasi data perlu dilakukan secara cermat karena kesalahan pada integrasi data bisa menghasilkan hasil yang menyimpang dan bahkan menyesatkan pengambilan aksi nantinya. Sebgai contoh bila integrasi data berdasarkan jenis produk ternyata menggabungkan produk dari kategori yang berbeda maka akan didapatkan korelasi antar produk yang sebenarnya tidak ada. Dalam integrasi data ini juga perlu dilakukan transformasi dan pembersihan data karena seringkali data dari dua database berbeda tidak sama cara penulisannya atau bahkan data yang ada di satu database ternyata tidak ada di database lainnya.

3. Seleksi Data (*Data Selection*)

Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari database. Sebagai contoh, sebuah kasus yang meneliti faktor kecenderungan orang membeli dalam kasus market basket analysis, tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan saja.

4. Transformasi data (*Data Transformation*)

Beberapa teknik data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam data mining. Beberapa metode data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh beberapa metode standar seperti analisis asosiasi dan clustering hanya bisa menerima input data kategorikal. Karenanya data berupa angka numerik yang berlanjut perlu dibagi-bagi menjadi beberapa interval. Proses ini sering

disebut transformasi data.Transformasi dan pemilihan data ini juga menentukan kualitas dari hasil data mining nantinya karena ada beberapa karakteristik dari teknik-teknik data mining tertentu yang tergantung pada tahapan ini

5. Proses mining

Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

6. Evaluasi pola (*pattern evaluation*)

Untuk mengidentifikasi pola-pola menarik kedalam knowledge based yang ditemukan. Dalam tahap ini hasil dari teknik data mining berupa pola-pola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai hipotesa ada beberapa alternatif yang dapat diambil seperti menjadikannya umpan balik untuk memperbaiki proses data mining, mencoba metode data mining lain yang lebih sesuai, atau menerima hasil ini sebagai suatu hasil yang di luar dugaan yang mungkin bermanfaat.

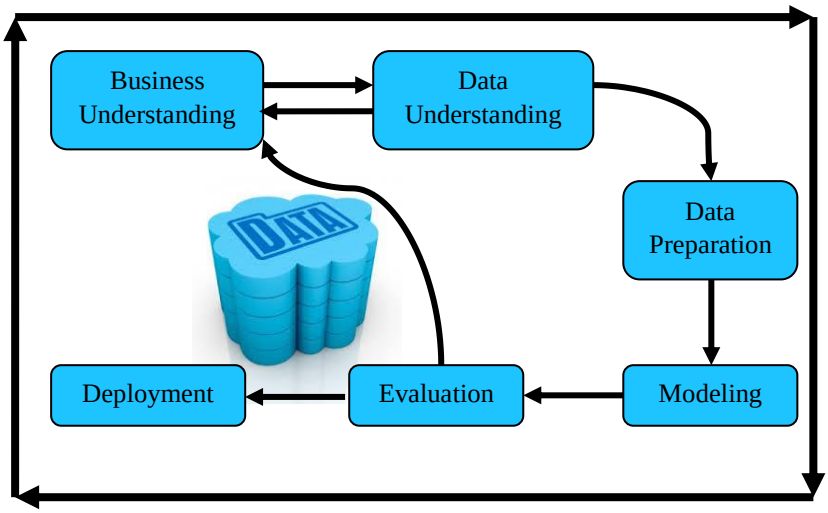
7. Presentasi pengetahuan (*knowledge presentation*)

Tahap terakhir dari proses data mining adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisis yang didapat. Visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.Ada kalanya hal ini harus melibatkan orang-orang yang tidak memahami data mining. Karenanya presentasi hasil data mining dalam bentuk pengetahuan yang bisa dipahami semua orang adalah satu tahapan yang diperlukan dalam proses data mining. Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil data mining.

2.2.7 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 2.2

menjelaskan tentang siklus hidup pengembangan *data mining* yang telah ditetapkan dalam CRISP-DM.



Gambar 2.2 : gambar siklus CRISP-DM

Berikut ini adalah enam tahap siklus hidup pengembangan *data mining* [16]:

1. *Business Understanding*

Tahap pertama adalah memahami tujuan dan kebutuhan dari sudut pandang bisnis, kemudian menterjemakan pengetahuan ini ke dalam pendefinisian masalah dalam *data mining*. Selanjutnya akan ditentukan rencana dan strategi untuk mencapai tujuan tersebut.

2. *Data Understanding*

Tahap ini dimulai dengan pengumpulan data yang kemudian akan dilanjutkan dengan proses untuk mendapatkan pemahaman yang mendalam tentang data, mengidentifikasi masalah kualitas data, atau untuk mendeteksi adanya bagian yang menarik dari data yang dapat digunakan untuk hipotesa untuk informasi yang tersembunyi.

3. *Data Preparation*

Tahap ini meliputi semua kegiatan untuk membangun *dataset* akhir (data yang akan diproses pada tahap pemodelan/*modeling*) dari data mentah. Tahap ini dapat diulang beberapa kali. Pada tahap ini juga mencakup pemilihan tabel, *record*, dan atribut-atribut data, termasuk proses

pembersihan dan transformasi data untuk kemudian dijadikan masukan dalam tahap pemodelan (*modeling*).

4. *Modeling*

Dalam tahap ini akan dilakukan pemilihan dan penerapan berbagai teknik pemodelan dan beberapa parameternya akan disesuaikan untuk mendapatkan nilai yang optimal. Secara khusus, ada beberapa teknik berbeda yang dapat diterapkan untuk masalah *data mining* yang sama. Di pihak lain ada teknik pemodelan yang membutuhan format data khusus. Sehingga pada tahap ini masih memungkinan kembali ke tahap sebelumnya.

5. *Evaluation*

Pada tahap ini, model sudah terbentuk dan diharapkan memiliki kualitas baik jika dilihat dari sudut pandang analisa data. Pada tahap ini akan dilakukan evaluasi terhadap keefektifan dan kualitas model sebelum digunakan dan menentukan apakah model dapat mencapat tujuan yang ditetapkan pada fase awal (*Business Understanding*). Kunci dari tahap ini adalah menentukan apakah ada masalah bisnis yang belum dipertimbangkan. Di akhir dari tahap ini harus ditentukan penggunaan hasil proses *data mining*.

6. *Deployment*

Pada tahap ini, pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Tahap *deployment* dapat berupa pembuatan laporan sederhana atau mengimplementasikan proses *data mining* yang berulang dalam perusahaan. Dalam banyak kasus, tahap *deployment* melibatkan konsumen, di samping analis data, karena sangat penting bagi konsumen untuk memahami tindakan apa yang harus dilakukan untuk menggunakan model yang telah dibuat.

2.2.8 Klasifikasi

Data Mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu : Deskripsi, Estimasi, Prediksi, Klasifikasi, Pengklusteran, dan Asosiasi.

Klasifikasi merupakan bagian dari algoritma data mining, klasifikasi ini adalah algoritma yang menggunakan data dengan target (*class/label*) yang berupa nilai kategorikal/nominal. Menurut Gorunescu [15] proses klasifikasi didasarkan pada empat komponen mendasar, yaitu:

1. Kelas (*Class*)

Variabel dependen dari model, merupakan variabel kategorikal yang merepresentasikan "label" pada objek setelah klasifikasinya. Contoh kelas semacam ini adalah: adanya kelas penyakit jantung, loyalitas pelanggan, kelas bintang (galaksi), kelas gempa bumi (badai), dll.

2. Prediktor (*Predictor*)

Variabel independen dari model, direpresentasikan oleh karakteristik (atribut) dari data yang akan diklasifikasikan dan berdasarkan klasifikasi yang telah dibuat. Contoh prediktor tersebut adalah : merokok, konsumsi alkohol, tekanan darah, frekuensi pembelian, status perkawinan, karakteristik (satelit) gambar, catatan geologi yang spesifik, kecepatan dan arah angin, musim , lokasi terjadinya fenomena , dll.

3. Pelatihan dataset (*Training dataset*)

Kumpulan data yang berisi nilai-nilai dari kedua komponen sebelumnya dan digunakan untuk melatih model dalam mengenali kelas yang cocok/sesuai, berdasarkan prediktor yang tersedia. Contoh set tersebut adalah: kelompok pasien yang diuji pada serangan jantung, kelompok pelanggan supermarket (diselidiki oleh intern dengan jajak pendapat), database yang berisi gambar untuk monitoring teleskopik dan pelacakan objek astronomi, database badai, database penelitian gempa.

4. Dataset Pengujian (*Testing Dataset*)

Berisi data baru yang akan diklasifikasikan oleh (*classifier*) model yang telah dibangun di atas sehingga akurasi klasifikasi (*model performance*) dapat dievaluasi.

Berikut beberapa model (metode) klasifikasi yang paling popular [15] :

1. *Decision/classification trees;*
2. *Bayesian classifiers/Naive Bayes classifiers;*
3. *Neural networks;*
4. *Statistical analysis;*
5. *Genetic algorithms;*
6. *Rough sets;*
7. *k-nearest neighbor classifier;*
8. *Rule-based methods;*
9. *Memory based reasoning;*
10. *Support vector machines.*

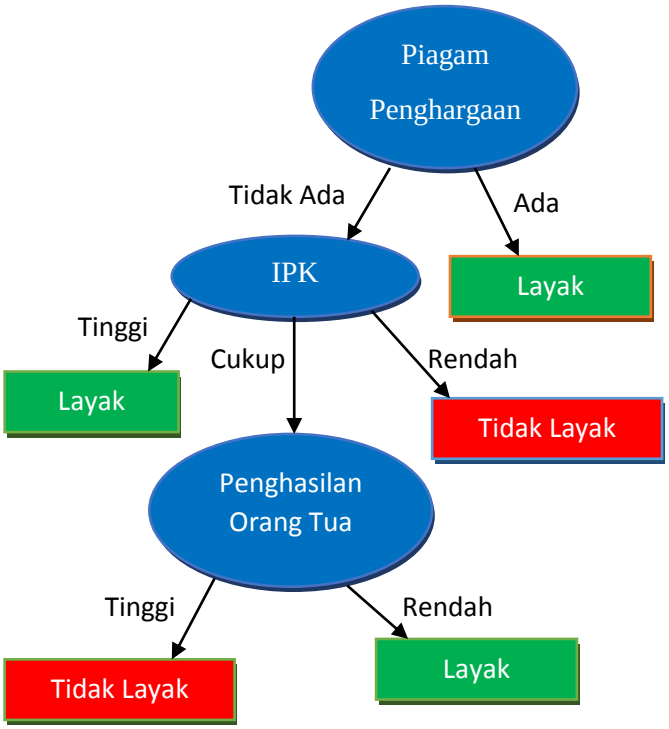
2.2.9 *Decision Tree* Algoritma C4.5

Pohon keputusan adalah salah satu metode klasifikasi yang kuat dan terkenal. Metode *Decision Tree* mengubah fakta besar menjadi pohon keputusan yang mewakili aturan, sehingga aturan tersebut dapat dengan mudah dipahami oleh manusia. *Decision Tree* juga berfungsi untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah variabel input dan variabel tujuan [14].

Model pohon keputusan terdiri dari satu set keputusan untuk membagi sejumlah populasi yang besar menjadi satu aturan yang kecil dengan memperhatikan target berupa objek. Objek target biasanya diklasifikasikan dan model pohon keputusan lebih fokus pada perhitungan probabilitas dari setiap *record* data dari beberapa kategori atau untuk mengklasifikasikan tiap *record* berdasarkan kelompok menjadi suatu kelas. Sebuah keputusan dapat dibangun dengan menerapkan salah satu algoritma *Decision tree* untuk memodelkan sekelompok data yang belum terklasifikasi. Konsep dari *Decision tree* adalah mengubah data menjadi pohon keputusan dan aturan keputusan.

Gambar 2.3 : Konsep *Decision Tree*

Dalam pohon keputusan sangat berhubungan dengan algoritma C4.5, karena dasar algoritma C4.5 adalah pohon keputusan. Algoritma data mining C4.5 merupakan salah satu algoritma yang digunakan untuk melakukan klasifikasi atau segmentasi atau pengelompokan yang bersifat prediktif. Cabang-cabang pohon keputusan merupakan pertanyaan klasifikasi dan daun-daunnya merupakan kelas-kelas atau segmen-segmennya.



Gambar 2.4 : Contoh Pohon Keputusan

Algoritma C4.5 merupakan salah satu algoritma machine learning. Dengan algoritma ini, mesin (komputer) akan diberikan sekelompok data untuk dipelajari yang disebut learning dataset. Kemudian hasil dari pembelajaran selanjutnya akan digunakan untuk mengolah data-data yang baru yang disebut test dataset. Karena algoritma C4.5 digunakan untuk melakukan klasifikasi, jadi hasil dari pengolahan test dataset berupa pengelompokkan data ke dalam kelas-kelasnya Umumnya, langkah-langkah algoritma C4.5 yang digunakan untuk membentuk pohon keputusan adalah [17].

a. Pilih atribut sebagai *root*.
b. Buat cabang untuk setiap nilai.

c. Bagi tiap cabang kedalam kelas.
d. Ulangi proses untuk setiap cabang sampai semua kasus pada tiap cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai *root,* didasarkan pada nilai *gain* tertinggi dari atribut yang tersedia.Sementara itu, untuk mendapat nilai *gain* tertinggi kita harus menghitung nilai *entropy* dari semua nilai didalam atribut.*Entropy* berperan sebagai parameter untuk mengukur varian dari data sampel. Setelah nilai *entropy* dalam data sampel diketahui, atribut yang paling berpengaruh akan menjadi pengukur dalam pengklasifikasian data, ukuran ini disebut sebagai *Information gain*.

Rumus menghitung entropy pada algoritma C4.5

$$\text{Entropi (S)} = \sum_{i=1}^{k} -pi * \log_2 pi$$

Keterangan :

- S adalah Himpunan (dataset) kasus
- k adalah banyaknya partisi S
- Pi adalah probabilitaas yang didapat dari Sum (Ya) atau Sum (Tidak) dibagi total kasus

Setelah mendapatkan entropi dari keseluruhan kasus, lakukan analisis pada setiap atribut dan nilai-nilainya dan hitung entropinya. Langkah berikutnya yaitu dengan menghitung Gain, rumus daripada Gain adalah sebagai berikut:

$$\text{Gain (A)} = \text{Entropi (S)} - \sum_{i=1}^{k} \frac{|S_i|}{|S|} x Entropi(S_i)$$

2.2.10 Confusion Matrix

Confusion Matrix adalah *tool* yang digunakan untuk evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah matrix dari prediksi yang akan dibandingkan dengan kelas yang asli dari inputan atau dengan kata lain berisi informasi nilai actual dan prediksi pada klasifikasi [17].

Tabel 2.2 : *Confusion Matrix* 2 kelas

Classification	Predicted class	
	Class = Yes	Class = No
Class=Yes	a (true positive-TP)	b (false negative-FN)
Class=No	c (false positive-FP)	d (true negative-TN)

vRumus untuk menghitung tingkat akurasi pada matriks adalah:

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN} = \frac{a+d}{a+b+c+d} \text{ x } 100\%$$

2.3 Rapid Miner

Rapid Miner merupakan perangakat lunak yang bersifat terbuka *(open source).* Rapid Miner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. Rapid Miner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. Rapid Miner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. Rapid Miner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. Rapid Miner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi.

Rapid Miner sebelumnya bernama YALE *(Yet Another Learning Environment),* dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari University of Dortmund. Rapid Miner didistribusikan di bawah lisensi AGPL (GNU Affero General Public License) versi 3. Hingga saat ini telah ribuan aplikasi yang dikembangkan mengunakan Rapid

Miner di lebih dari 40 negara. Rapid Miner sebagai software open source untuk data mining tidak perlu diragukan lagi karena software ini sudah terkemuka di dunia. Rapid Miner menempati peringkat pertama sebagai Software data mining pada polling oleh KDnuggets, sebuah portal data-mining pada 2010-2011.

Rapid Miner menyediakan GUI (Graphic User Interface) untuk merancang sebuah pipeline analitis. GUI ini akan menghasilkan file XML (Extensible Markup Language) yang mendefenisikan proses analitis keingginan pengguna untuk diterpkan ke data. File ini kemudian dibaca oleh Rapid Miner untuk menjalankan analis secara otomatis.

Rapid Miner memiliki beberapa sifat sebagai berikut:

a. Ditulis dengan bahasa pemrograman Java sehingga dapat dijalankan di berbagai sistem operasi.
b. Proses penemuan pengetahuan dimodelkan sebagai operator trees.
c. Representasi XML internal untuk memastikan format standar pertukaran data.
d. Bahasa scripting memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen.
e. Konsep multi-layer untuk menjamin tampilan data yang efisien dan menjamin penanganan data.
f. Memiliki GUI, command line mode, dan Java API yang dapat dipanggil dari program lain.

Beberapa Fitur dari Rapid Miner, antara lain:

a. Banyaknya algoritma data mining, seperti *decision tree* dan *self-organization map*.
b. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, *tree chart* dan 3D *Scatter plots*.
c. Banyaknya variasi plugin, seperti *text* plugin untuk melakukan analisis teks.
d. Menyediakan prosedur data mining dan *machine learning* termasuk: ETL (*extraction, transformation, loading*), data preprocessing, visualisasi, modelling dan evaluasi

e. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI
f. Mengintegrasikan proyek data mining Weka dan statistika R [14].

2.4 Java

Java adalah suatu teknologi di dunia *software* komputer, yang merupakan suatu bahasa pemrograman sekaligus suatu platform. Sebagai bahasa pemrograman, Java dikenal sebagai bahasa pemrograman tingkat tinggi yang berorientasi objek. Sebagai bahasa pemrograman Java dirancang agar dapat dijalankan di semua platform.

Java diciptakan oleh suatu tim yang dipimpin oleh Patrick Naughton dan James Gosling dalam suatu proyek dari Sun Microsystem yang memiliki kode Green dengan tujuan untuk menghasilkan bahasa komputer sederhana yang dapat dijalankan di peralatan sederhana dengan tidak terikat pada arsitektur tertentu.

Program yang ditulis menggunakan Java berjalan pada suatu Virtual Machine dengan nama *Java Runtime Environment* (JRE).

Pada Java, terdapat lima fase pada pembuatan dan eksekusi program. Fase pertama yaitu editing kode sumber (*source code*) Java menjadi file *.java pada penyimpanan sekunder (HDD). Fase kedua yaitu kompilasi *source code* *.java menjadi file dengan ekstensi *.class. Setelah terbentuk file dengan ekstensi *.class, dilakukan *class loading* pada fase ketiga kedalam memori primer (RAM) untuk dilakukan cek error sebelum dieksekusi. Setelah file *.class di*load* pada RAM, dilakukan *bytecode verification* pada fase empat. Setelah *bytecode* diverifikasi kemudian dieksekusi pada *Java Virtual Machine* (JVM) agar dapat digunakan oleh user [14].

2.5 Netbeans Integrated Development Environment (IDE)

Netbeans adalah sebuah *integrated development environment (IDE)* untuk pengembangan terutama dengan java, tetapi *netbeans* juga *support* bahasa pemrograman lain seperti di php tertentu, C/C++, dan html 5. Netbeans juga merupakan aplikasi *platform framework* untuk aplikasi desktop Java dan lainnya [15]. Beberapa karakteristik dari Netbeans IDE :

a. User Interface Framework

b. Data Editor

c. Customization Display

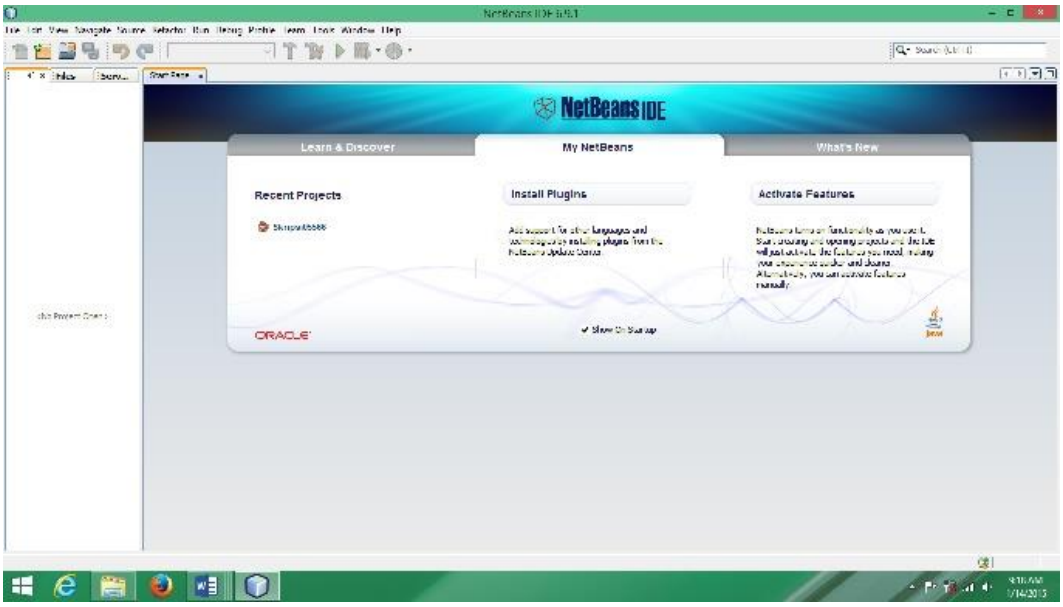
d. Wizard Framework

e. Data Systems

f. Internationalization

g. Help System

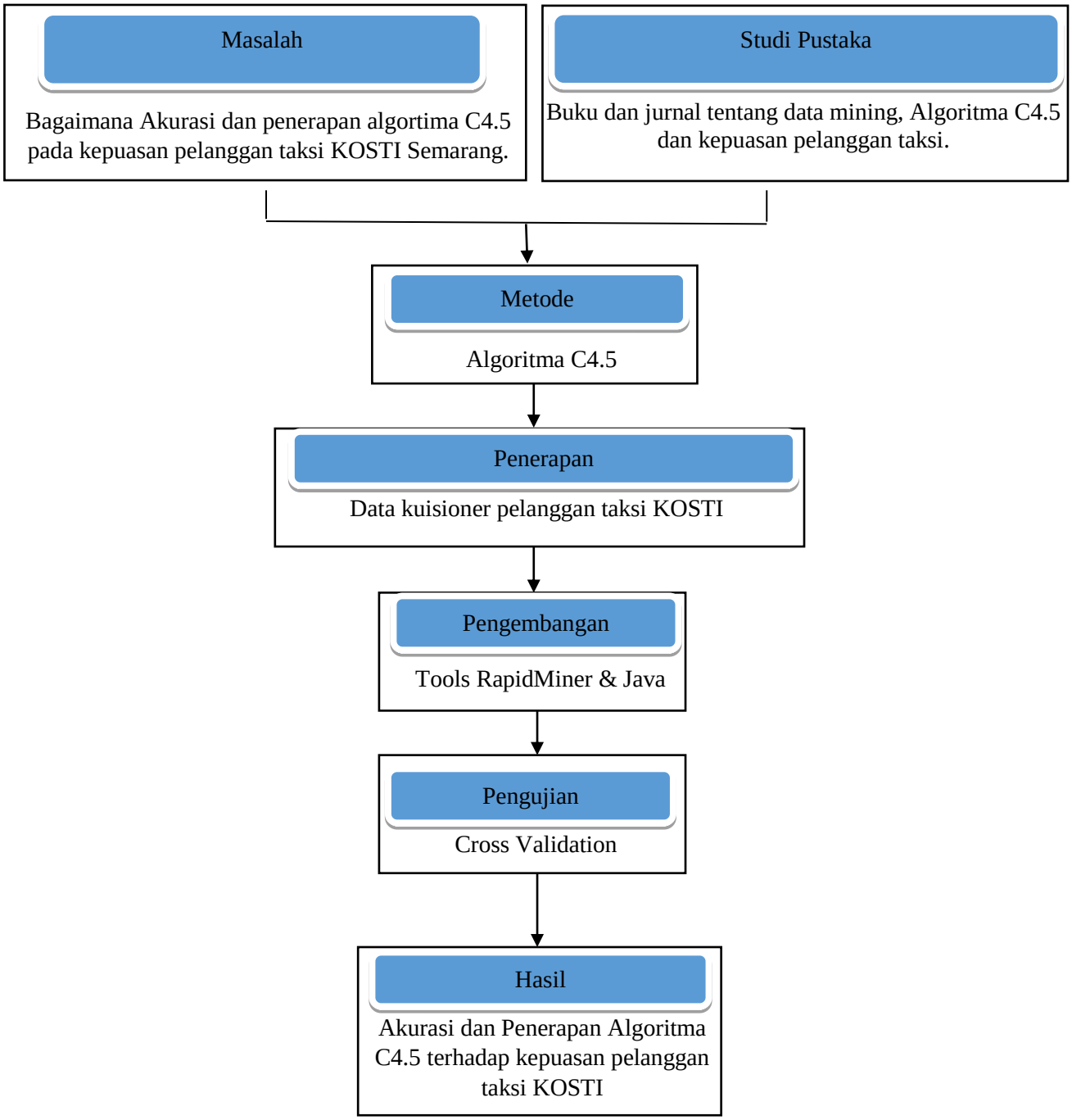
Fitur yang ditawarkan oleh Netbeans dapat dikostumisasi oleh pemrogram dengan mudah dan cepat dalam membangun software.



Gambar 2.5. Tampilan Awal Netbeans IDE

2.6 Kerangka Pemikiran

Penulis perlu membuat gambaran singkat sebagai alur penyusunan laporan ini dengan kerangka pemikiran sebagai berikut:



Gambar 2.6: kerangka pemikiran