BAB II
LANDASAN TEORI

2.1. Tinjauan Pustaka

Putra (2018) Melakukan penelitian dengan judul Klasifikasi Nilai Gizi Balita Menggunakan *Naive Bayes Classification* (Studi Kasus Posyandu Ngudi Luhur) untuk memudahkan penentuan Nilai gizi balita menggunakan data mining dengan algoritma *Naive Bayes classification* (NBC). Sistem dibangun dengan bahasa pemrograman PHP dan database MySQL. Penelitian ini menggunakan data 17 balita dengan rentang waktu 2 tahun pengukuran. Total data yang digunakan berjumlah 408 data. Dilakukan dua kali pengujian data, pertama dengan perbandingan 60:40 dan kedua 80:20 data training dan data testing. Hasil penelitian menunjukkan akurasi sebesar 93,1%. Dengan kata lain NBC dikategorikan baik untuk pengujian Nilai gizi balita. (Putra, 2018)

Sulastri dan Nugroho (2017) pada jurnal yang berjudul "Penerapan Data Mining Untuk Prediksi Rating Penjualan Buku Menggunakan Metode Nai e Bayes" dijelaskan bahwa melihat banyaknya buku yang beredar di pasaran dengan segala kelebihan dan kelemahannya maka diperlukan sebuah sistem prediksi untuk mengetahui faktor yang mempengaruhi rating penjualan buku untuk menghasilkan sebuah buku yang memenuhi kriteria pembaca. Teknik yang digunakan dalam prediksi rating penjualan buku ini memakai algoritma Naive Bayes. Naive Bayes digunakan untuk mencari nilai probabilitas paling besar pada setiap variabel yang sudah ada. (Sulastri dan Nugroho, 2017)

Kevin Moniarga dan Suharta (2018) dalam judul "Penerapan Algoritma Naive Bayes Classifier Untuk Mengetahui Minat Beli Pelanggan Terhadap Sofa (Studi Kasus Di Mebel Kelumer Bayau)" menyebutkan permasalahan utama yang dihadapi oleh Mebel Kelumer Bayau adalah bagaimana memprediksi minat beli pelanggan dari penjualan barang Mebel Kelumer Bayau pada masa mendatang dari data yang telah diperoleh sebelumnya. Penelitian ini menggunakan metode Algoritma Naive Bayes Classifier yang mempunyai akurasi dan kecepatan untuk dilakukan penggalian pengetahuan yang sudah ada pada database. (Kevin Moniarga dan Suharta, 2018)

Dwi Retnoningsih, Maslichah Raichatul Janah, Sri Ernawati (2016) yang Berjudul "Analisis Sistem Pendeteksian Plagiatisme Karya Ilmiah Di Universitas Sahid Surakarta" Analisis PIECES merupakan salah satu metode analisis untuk mengidentifikasi masalah. Analisis ini terdiri dari analisis terhadap kinerja, informasi, ekonomi, keamanan aplikasi, efisiensi dan pelayanan yang lebih dikenal dengan PIECES analysis (performance, Information, economy, Control, efficiency dan Services). Hasil analisis PIECES adalah dokumen kelemahan sistem yang menjadi rekomendasi untuk perbaikan-perbaikan yang harus dibuat pada sistem yang akan dikembangkan. (Dwi Retnoningsih, 2016)

Wahyudi, Dahlan Susilo, Mochtar Yunianto (2010) yang berjudul "Si tem Pendukung Keputusan Untuk Seleksi Penyedia Pengadaan Barang dan Jasa di Departemen Pekerjaan Umum Direktorat Jenderal Sumber Daya Air Yogyakarta" menyebutkan sistem pendukung keputusan untuk seleksi penyedia pengadaan barang dan jasa dapat mempercepat dan mempermudah di dalam proses seleksi penyedia pengadaan barang dan jasa. (Wahyudi, Dahlan Susilo, Mochtar Yunianto, 2010)

2.2. Kerangka Pemikiran

Kerangka pemikiran dalam penelitian ini dijelaskan sebagai berikut:

1. Latar Belakang Masalah

   Belum adanya suatu sistem yang pasti untuk melakukan klasisfikasi nilai peserta CAT CPNS untuk tahun berikutnya. Rumusan Masalah

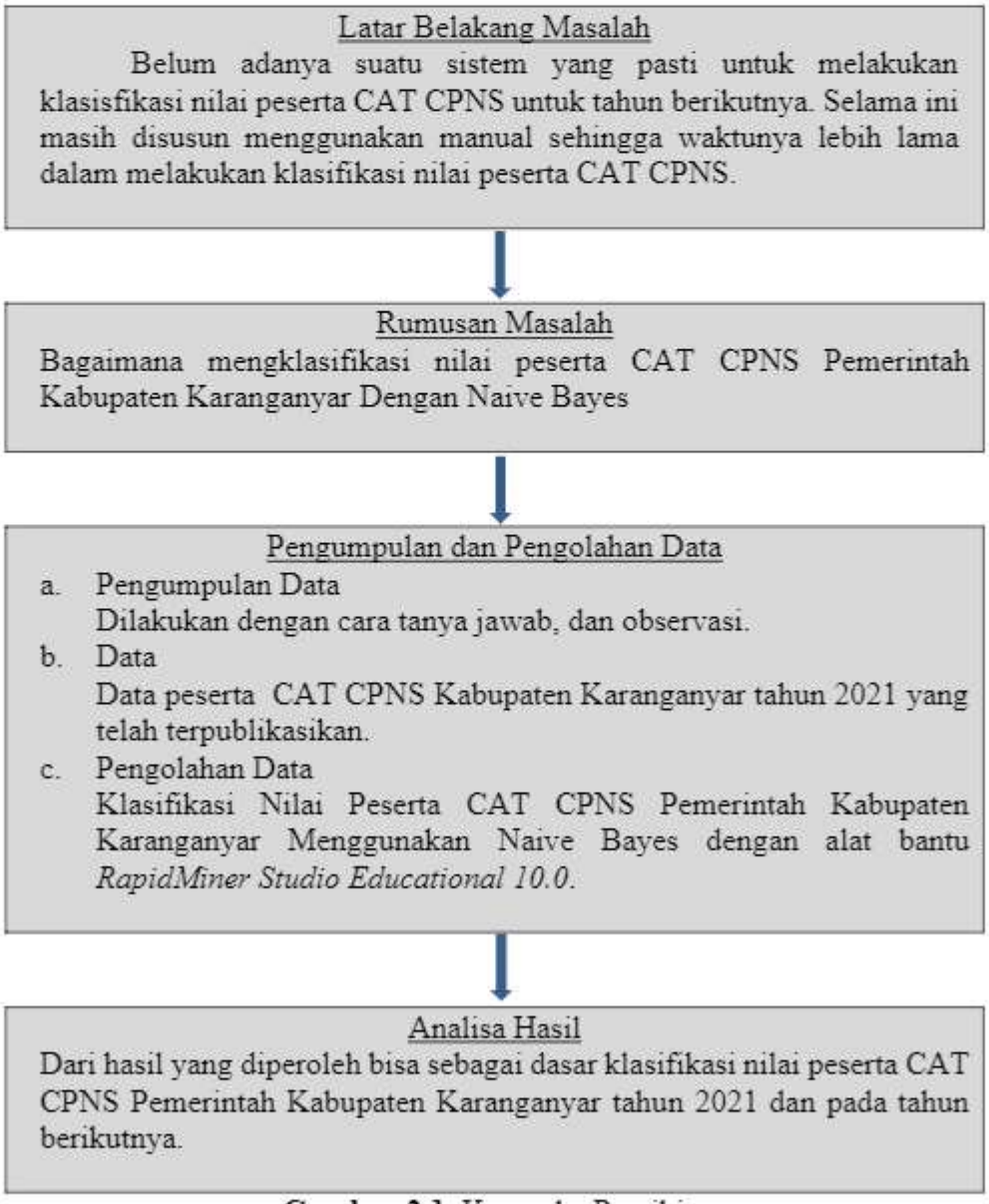
   Bagaimana Mengklasifikasi nilai peserta CAT CPNS Pemerintah Kabupaten Karanganyar Dengan Naive Bayes?

2. Pengumpulan dan Pengolahan Data
   a. Pengumpulan data dilakukan dengan tanya jawab dan observasi
   b. Data penelitian yang akan digunakan merupakan data nilai SKD peserta CAT CPNS Kabupaten Karanganyar tahun 2021 yang telah terpublikasikan.

c. Data yang siap diolah akan diproses perhitungan prediksi menggunakan algoritma Naive Bayes dengan alat bantu *RapidMiner Studio Educational 10.0*.

3. Analisis Hasil

Tahap ini dilakukan analisisa hasil pengolahan data berupa Klasifikasi Nilai Peserta CAT CPNS sehingga data yang diperoleh bisa dijadikan dasar keputusan yang tepat untuk memprediksi kelulusan peserta dengan klasifikasi nilai SKD peserta CAT CPNS pada tahun berikutnya.



Gambar 2.1. Kerangka Pemikiran

2.3. Landasan Teori

2.3.1. Klasifikasi

Klasifikasi adalah sebuah proses untuk mencari model atau fungsi yang menjelaskan dan membedakan kelas atau konsep dari data, dengan tujuan untuk menggunakan model dan melakukan prediksi dari kelas suatu objek dimana tidak diketahui label dari kelas tersebut. Model yang ada berasal dari analisis kumpulan *training* data (objek data dimana kelas label diketahui) (Han & M, 2006). Algoritma yang dapat digunakan untuk klasifikasi antara lain Naive Bayes, Adaptive Bayes Network, Decision Tree, dan Support Vector Machine. Tabel 2.1 adalah tabel perbandingan keempat algoritma tersebut.

Tabel 2.1. Perbandingan Algoritma Klasifikasi (Witania, Riani, & Mulyawan, 2009)

Fitur	Naive Bayes	Adaptive Bayes Network	Decision Tree	Support Vector Machine
Kecepatan	Sangat cepat	Cepat	Cepat	Cepat
Ketepatan	Baik Disemua data	Baik disemua data	Signifikan	ik disemua aturan
Transparansi	Tanpa aturan	Aturan khusus	Tanpa aturan	Aturan

Model Klasifikasi terdiri dari:

a. Pemodelan Deskriptif

   Dapat bertindak sebagai suatu alat yang bersifat menjelaskan untuk membedakan antara objek dengan kelas yang berbeda.

b. Pemodelan Prediktif

   Model klasifikasi juga dapat menggunakan prediksi label kelas yang belum diketahui recordnya.

2.3.2. Penerimaan Calon Pegawai Negeri Sipil

Penerimaan calon pegawai negeri sipil diatur dalam Peraturan Pemerintah (PP) Nomor 11 Tahun 2017 tentang Manajemen Pegawai Negeri Sipil sebagaimana telah diubah dengan PP Nomor 17 Tahun 2020 tentang Perubahan atas PP Nomor 11 Tahun 2017 tentang Manajemen Pegawai Negeri Sipil, Peraturan Menteri PANRB Nomor 27 Tahun 2021 tentang Pengadaan Pegawai Negeri. Setiap Warga Negara Indonesia yang memenuhi persyaratan mempunyai kesempatan yang sama untuk melamar menjadi Pegawai Negeri Sipil. Proses Pengadaan Pegawai Negeri ini diperlukan perencanaan, data, dan informasi yang benar dan jelas agar tidak terjadi kesalahan-kesalahan yang dapat merugikan pelamar Pegawai Negeri Sipil.

2.3.3. Seleksi CAT SKI dan SKB

Kompetensi Dasar adalah kemampuan dan karakteristik dalam diri seseorang berupa pengetahuan, keterampilan, dan perilaku. Kompetensi Bidang adalah kemampuan dan karakteristik dalam diri seseorang berupa pengetahuan, keterampilan, perilaku yang diperlukan dalam pelaksanaan tugas jabatannya sehingga individu mampu menampilkan unjuk kerja yang tinggi dalam suatu Jabatan. Seleksi Kompetensi Dasar yang selajutnya disingkat SKD adalah seleksi yang mengukur kemampuan dan karakteristik dalam diri seseorang berupa pengetahuan, keterampilan, dan perilaku. Seleksi Kompetensi Bidang yang selanjutnya disingkat SKB adalah seleksi yang mengukur kemampuan dan karakteristik dalam diri seseorang berupa pengetahuan, keterampilan, perilaku yang diperlukan dalam pelaksanaan tugas jabatannya sehingga individu mampu menampilkan unjuk kerja yang tinggi dalam suatu Jabatan tertentu. Computer Assisted Test disingkat CAT adalah suatu metode seleksi/tes dengan menggunakan komputer. Sistem Seleksi Calon Aparatur Sipil Negara yang disingkat SSCASN sebagai portal pelamaran terintegrasi berbasis internet yang digunakan dalam Pengadaan ASN.

2.3.4. Penilaian

Nilai TWK (Tes Wawasan Kebangsaan), Nilai TIU (Tes Intelegensia Umum), nilai TKP (Tes Karakteristik Pribadi), nilai total SKD (Seleksi Kompetensi Dasar), nilai Total SKB (Seleksi Kompetensi Bidang, dan Integrasi nilai SKD dan SKB. Ambang batas Nilai SKD, Nilai SKB, dan Integrasi nilai SKD dan SKB berdasarkan pada Peraturan Menteri PANRB Nomor 27 Tahun 2021 tentang Pengadaan Pegawai Negeri.

2.3.5. Data Mining

Data mining adalah sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar, yang dapat juga diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Data mining merupakan proses yang menggunakan berbagai teknik dan alat analisis data untuk menemukan hubungan dan pola yang tersembunyi (Anjar & et, 2020)

Ada beberapa tugas yang dapat dilakukan oleh Data Mining dalam proses pemecahan masalah dan pencarian pengetahuan baru di antaranya adalah sebagai berikut:

1. Klastering (*Clustering*)

   Digunakan untuk mengelompokkan atau mengidentifikasi data yang memiliki karakteristik tertentu. Contoh algoritma: K-Means, K-Medoids.

2. Klasifikasi (*Classification*)

   Digunakan untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Contoh algoritma: Naive Bayes, C4.5, K-Nearest Neighbor.

3. Asosiasi (*Association*)

   Digunakan untuk mengatasi masalah bisnis yang khas, yakni dengan menganalisa tabel transaksi penjualan yang mengidentifikasi produk-produk yang sering kali dibeli bersamaan oleh *customer*, misalnya apabila orang

membeli sambal, biasanya juga dia membeli kecap. Contoh algoritma: Apriori, Frequent Pattern Growth.

4. Estimasi (*Estimation*)

   Digunakan untuk memperkirakan atau menilai suatu hal yang belum pernah ada sebelumnya yang disajikan dalam bentuk hasil kuantitatif. Contoh algoritma: Regresi Linier, Confidence Interval Estimations.

5. Prediksi (*Predictions*)

   Digunakan untuk memperkirakan atau meramalkan suatu kejadian yang belum pernah terjadi. Contoh algoritma: Decision Tree, K-Nearest Neighbor.

Data mining merupakan suatu proses pencarian pola dari data-data dengan jumlah yang sangat banyak yang tersimpan dalam suatu tempat penyimpanan dengan menggunakan teknologi pengenal pola, teknik statistik, dan matematika. (Saputra & Sibarani2, Agustus 2020)

Karakteristik data mining adalah sebagai berikut (Saputra & Sibarani, 2020):

1. Data mining berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
2. Data mining bisa menggunakan data yang sangat besar.
3. Biasanya data yang besar digunakan untuk membuat hasil lebih dipercaya.
4. Data mining berguna untuk membuat keputusan yang kritis, terutama dalam strategi.

Data mining sering juga disebut *knowledge discovery in database* (KDD), adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining bisa dipakai untuk memperbaiki pengambilan keputusan dimasa depan.

Proses data mining dijelaskan pada Gambar 2.2 sebagai berikut (Saputra & Sibarani, 2020):

Gambar 2.2. Proses Data Mining

Tahapan-tahapan proses data mining, di antaranya:

1. Seleksi data *(Data selection)*

   Seleksi data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi tersebut akan disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. Integrasi Data *(Data Integrasion)*

   Integrasi data merupakan penggabungan data dari berbagai database ke dalam satu database baru.

3. *Pre-processing / Cleaning*

   Proses cleaning dilaksanakan sebelum proses data mining pada data yang menjadi fokus KDD. Proses cleaning diantaranya membuang duplikasi data dan memperbaiki kesalahan pada data. Selain itu dilakukan juga proses enrichment, yaitu proses memperkaya data yang sudah ada dengan data atau informasi yang relevan dan diiperlukan untuk KDD.

4. Transformasi *(Transformation)*

   Pengkodean adalah proses transformasi pada data yang telah dipilih. Proses pengkodean dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

5. Data Mining

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

6. Interpretasi/Evaluasi (*Interpretation/Evaluation*)

   Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap-tahap ini disebut dengan interpretasi yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

7. Presentasi Pengetahuan (*Knowledge Presentation*)

   Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

2.3.6. Naive Bayes Classification

Klasifikasi Naive Bayes adalah klasifikasi berdasar teorema Bayes dan digunakan untuk menghitung probabilitas tiap kelas dengan asumsi bahwa antar satu kelas dengan kelas yang lain tidak saling tergantung (independen). Pada metode ini, semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atribut yang sama penting dan setiap atribut saling bebas satu sama lain. (S Kusumadewi, 2009)
Persamaan dari teorema Bayes adalah

$$P(H|X) = \frac{P(X|H}{P(X)} . P(H)$$

Di mana :

X : Data dengan class yang belum diketahui

H : Hipotesis data merupakan suatu class spesifik

P(H|X) : Probabilitas hipotesis H berdasar kondisi X (posteriori probabilitas)

P(H) : Probabilitas hipotesis H (prior probabilitas)

P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H

P(X) : Probabilitas X

Untuk menjelaskan teorema Naive Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut.

2.3.7. Confusion Matrix

Confusion matrik adalah salah satu teknik yang dapat digunakan untuk mengevaluasi *performance* algoritma dari *Machine Learning (ML)*. Pada dasarnya *confusion matrix* mempresentasikan prediksi dan kondisi sebenarnya dari data yang dihasilkan oleh algoritma ML. Berdasarkan *confusion matrik* kita bisa menentukan nilai *Accuracy*, *Precision* dan *Recall*. (Nugroho, 2019)

a. *Accuracy* menggambarkan seberapa akurat model dalam mengklasifikasikan dengan benar.
b. *Precision* menggambarkan akurasi antara data yang diminta dengan hasil prediksi yang diberikan oleh model.
c. *Recall* menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi.

Confusion matrik berbentuk tabel yang menggambarkan lebih detail tentang jumlah data yang diklasifikasikan dengan benar maupun salah. Tabel *confusion matrix* dijelaskan pada Gambar 2.3 sebagai berikut:

Gambar 2.3. *Confusion Matrix*

Ada empat nilai yang dihasilkan didalam tabel confusion matri.:, diantaranya:

a. *True Positive (1P)* : Jumlah data yang bernilai Positif dan diprediksi benar sebagai Positif.

b. *False Positive (FP)* : Jumlah data yang bernilai Negatif tetapi diprediksi sebagai Positif.
c. *False Negative (FN)* : Jumlah data yang bernilai Positif tetapi diprediksi sebagai Negatif.
d. *True Negative (TN)* : Jumlah data yang bernilai Negatif dan diprediksi benar sebagai Negatif.

2.3.8. RapidMiner

RapidMiner adalah *platform* perangkat lunak ilmu data yang dikembangkan oleh perusahaan ternama sama dengan yang menyediakan lingkungan lingkungan terintegrasi untuk persiapan data, pemebelajaran mesin, pembelajaran dalam, penambangan teks, dan analisis prediktif. Sejarah *RapidMiner*, sebelumnya dikenal sebagai YALE (*Yet Another Learning Environment*), mulai dikembangkan pada tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer dari Unit Kecerdasan Buatan Universitas Teknik Dortmund. Mulai Tahun 2006, perkembangannya didorong oleh Rapid-I, sebuah perusahaan yang didirikan oleh Ingo Mierswa dan Ralf Klinkenberg. Pada tahun 2007, nama perangkat lunak itu berubah dari YALE menjadi *RapidMiner*.
Pada tahun 2013, perusahaan melakukan rebranding dari Rapid-I menjadi *RapidMiner* (Ramdhan et al., 2020).

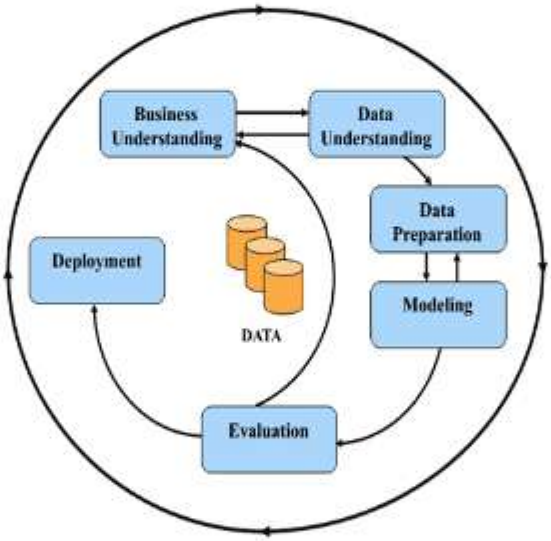
RapidMiner dibangun menggunakan bahasa Java sehingga dapat dijalankan di berbagai sistem operasi seperti: Windows, Linux, Unix serta Mac Os. Sebagai rekomendasi sebaiknya menggunakan sistem 64 bit, dikarenakan jumlah maksimum yang dapat digunakan oleh *RapidMiner* terbatas pada sistem operasi dengan sistem 32 bit yaitu hanya sebesar 2GB (Ramdhan et al., 2020).

RapidMiner dikembangkan dengan model open core. Terdapat 2 versi edisi dari *RapidMiner*, yang pertama adalah *RapidMiner* Basic Edition (gratis), yang dibatasi kemampuannya pada 1 prosesor logis dan maksimal 10,000 baris data, tersedia dengan lisensi AGPL. Versi gratis dapat diunduh melalui tautan https://rapidminer.com/get-started/. Kedua adalah *RapidMiner* versi komersial dengan harga mulai dari $2.500.

2.3.9. Metode CRIPS-DM

Metode *Cross Industry Standart Process for Data Mining* (CRIPS-DM) merupakan sebuah metodologi yang menerapkan pendekatan terstruktur untuk perencanaan proyek data mining yang sangat ampuh dan sudah teruji dengan baik. Metode ini sangat umum digunakan karena sangat praktis, fleksibel, dan aplikatif untuk memecahkan isu bisnis yang sulit sekalipun. Metode ini merupakan metode andalan yang dapat dijalankan dihampir semua persoalan bisnis data mining.

Proses data mining berdasarkan CRIPS-DM terdiri dari enam fase. Fase-fase metode CRIPS-DM dapat dilihat pada Gambar 2.4 sebagai berikut (Kuncoro, 2021):



Gambar 2.4. Proses CRIPS-DM

Proses data mining berdasarkan CRIPS-DM terdiri dari enam fase yaitu:

1. Fase Pemahaman Bisnis (*Business Understanding*)

   Pada fase ini dibutuhkan pemahaman tentang substansi dari kegiatan data mining yang akan dilakukan. Kegiatannya antara lain; menentukan sasaran atau tujuan bisnis, memahami situasi bisnis, menentukan tujuan data mining dan membuat perencanaan strategi serta jadwal penelitian.

2. Fase Pemahaman Data (*Data Understanding*)

Adalah fase mengumpulkan data awal, mempelajari data untuk bisa mengenal data yang akan dipakai. Fase ini mencoba mengidentifikasikan masalah yang berkaitan dengan kualitas data.

3. Fase Persiapan Data (*Data Preparation*)

   Kegiatan yang dilakukan dalam fase ini antara lain: memilih *table* dan *field* yang akan ditransformasikan ke dalam *database* baru untuk bahan data mining.

4. Fase Pemodelan (*Modelling*)

   Adalah fase menentukan teknik data mining yang digunakan, menentukan *tools* data mining, teknik data mining, algoritma data mining.

5. Fase Evaluasi (*Evaluation*)

   Adalah fase interpretasi terhadap hasil data mining yang ditunjukkan dalam proses pemodelan pada fase sebelumnya. Evaluasi dilakukan secara mendalam dengan tujuan menyesuaikan model yang didapat agar sesuai dengan sasaran yang ingin dicapai dalam fase pertama.

6. Fase Penyebaran (*Deployment*)

   Adalah fase penyusunan laporan atau presentasi dari pengetahuan yang didapat dari evaluasi pada proses data mining.