BAB II

LANDASAN TEORI

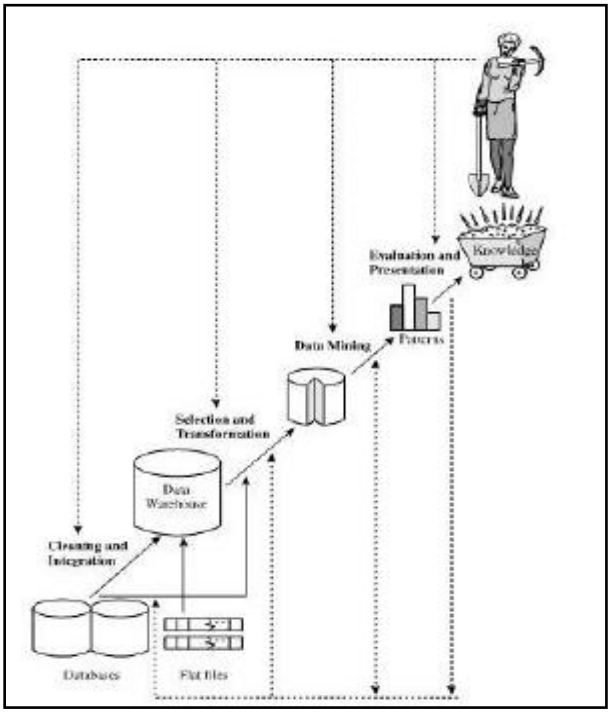
2.1 Studi Pustaka

Tahapan pengumpulan data pada penelitian ini dilakukan dengan cara mengumpulkan data dan informasi dari buku dan jurnal yang terkait dengan pokok bahasan penelitian.

2.2 *Data Mining*

Nama *data mining* sebenarnya mulai dikenal sejak tahun 1990, ketika pekerjaan pemanfaatan data menjadi sesuatu yang penting dalam berbagai bidang, mulai dari bidang akademik, bisnis , hingga medis . *Data mining* dapat diterapkan pada berbagai bidang yang mempunyai sejumlah data, tetapi karena wilayah penelitian dengan sejarah yang belum lama, dan belum melewati masa 'remaja', maka *data mining* masih diperdebatkan posisi bidang pengetahuan yang memilikinya. Maka Daryl Pregibon menyatakan bahwa "*data mining* adalah campuran dari statistik, kecerdasan buatan, dan riset basis data " yang masih berkembang (Gorunescu , 2011) . *Data mining* sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar (Tan et al, 2006) . *Data mining* juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan. Istilah *data mining* kadang disebut juga *knowledge discovery* (Prasetyo, 2012) . Salah satu teknik yang dibuat dalam *data mining* adalah bagaimana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan.

Kebutuhan untuk prediksi juga dapat memanfaatkan teknik ini. Dalam *data mining*, pengelompokan data juga bisa dilakukan. Tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada. Anomali data transaksi juga perlu dideteksi untuk dapat mengetahui tindak lanjut berikutnya yang dapat diambil. Semua hal tersebut bertujuan mendukung kegiatan operasional perusahaan sehingga tujuan akhir perusahaan diharapkan dapat tercapai. *Data mining* merupakan bagian dari proses *Knowledge Discovery from Data* (KDD). Dibawah ini digambarkan skema dari proses KDD.



Gambar 2.1 *Data mining* sebagai dari proses *knowledge discovery* (Han dan Kamber ,2006)

Gambar 2.1 menunjukkan proses penjelajahan pengetahuan dimulai dari beberapa *database* dilakukan proses *cleaning* dan *integration* sehingga menghasilkan *data warehouse*. Dilakukan proses *selection* dan *transformation* yang kemudian disebut sebagai data mining hingga menemukan pola dan memperoleh pengetahuan dari data (*knowledge*).

Terdapat beberapa teknik data mining yang sering disebut dalam literatur (Haryati et al, 2015). Namun ada 3 teknik data mining yang populer ,yaitu:

1. *Association Rule Mining*

Association Rule Mining adalah teknik mining untuk menemukan asosiatif antara kombinasi atribut.

Contoh dari aturan asosiatif dari analisa pembelian di suatu pasar swalayan dapat mengatur penempatan barangnya atau merancang strategi pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu.

2. *Clustering*

Berbeda dengan *association rule mining* dan klasifikasi dimana kelas data telah ditentukan sebelumnya, *clustering* dapat dipakai untuk memberikan label pada kelas data yang belum di ketahui. Karena itu *clustering* sering digolongkan sebagai metode *unsupervised learning*. Prinsip

clustring adalah memaksimalkan kesamaan antar *cluster*. *Clustering* dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi.

3. Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, pendapatan rendah.

2.3 Pohon Keputusan (*Decision Tree)*

Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki (Haryati et al, 2015). Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan. Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record*. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan *tree*. Misalkan untuk menentukan main tenis, kriteria yang digunakan adalah cuaca, angin, iklim dan temperatur.

Manfaat utama menggunakan pohon keputusan adalah kemampuannya untuk *membreak down* proses pengambilan keputusan yang kompleks menjadi lebih simpel sehingga pengambilan keputusan akan menjadi lebih menginterprestasikan solusi permasalahan. Pohon keputusan juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan sebuah variabel target. Pohon keputusan memadukan antara eksplorasi data dan pemodelan sehingga sangat bagus sebagai langkah awal pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain.

2.4 Algoritma Klasifikasi

Algoritma klasifikasi menggunakan metode pohon keputusan (Haryati et al, 2015) . Pohon keputusan adalah metode klasifikasi dan prediksi yang sudah terbukti *powerfull* dan sangat terkenal. Metode ini berfungsi mengubah fakta menjadi pohon keputusan yang merepresentasikan aturan yang dapat mudah dimengerti dengan bahasa alami. Proses dari pohon keputusan ini dimulai dari *node* akar hingga *node* daun yang dilakukan secara rekursif dimana setiap percabangan menyatakan kondisi dan setiap ujung pohon akan menyatakan keputusan.

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan, dibutuhkan algoritma C4.5. Di akhir tahun 1970 hingga di awal tahun 1980-an, J. Ross Quinlan seorang peniliti dibidang mesin pembelajaran

mengembangkan sebuah model pohon keputusan yang dinamakan ID3 (*Iterative Dichotomiser*), walaupun sebenarnya proyek ini telah dibuat sebelumnya oleh E.B.Hunt, J.Marin, dan P.T. Stone. Kemudian Quinlan membuat algoritma dari pengembangan ID3 yang dinamakan C4.5 yang berbasis *supervised learning*. Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 (Kusrini dan Lutfhi, 2009), yaitu :

1. Menyiapkan data *training*. Data *training* biasanya dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon akar akan diambil dari atribut yang terpilih dengan cara menghitung nilai *Gain* dari masing-masing atribut, nilai *Gain* yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai *Gain* dari atribut, hitung dahulu nilai *entropy* yaitu :

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

Keterangan:

S : himpunan kasus

A : atribut

n : jumlah partisi S

pi : proporsi dari Si terhadap S

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

$$\mathrm{Gain}\ (\mathrm{S}, \mathrm{A}) = Entropy(\mathrm{S}) - \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{|\mathrm{si}|}{|\mathrm{s}|} * Entropy \quad \ldots (2)$$

Keterangan:

S : himpunan kasus

A : atribut

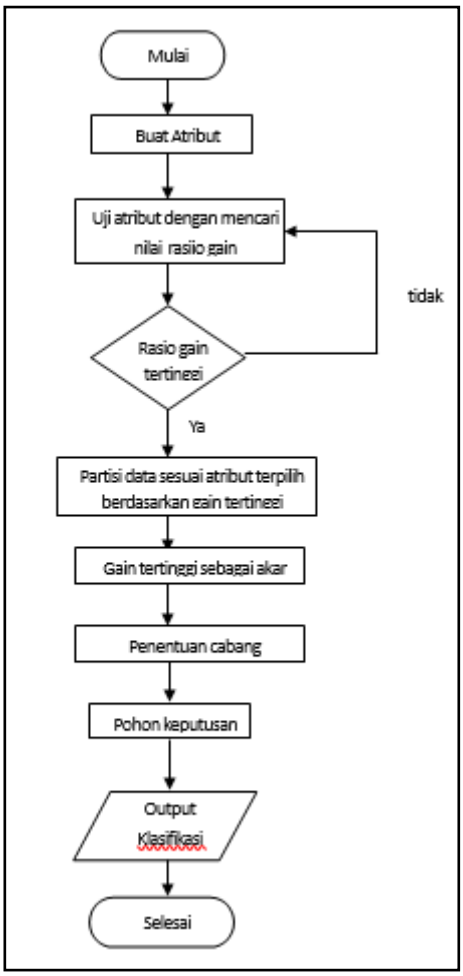
n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

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

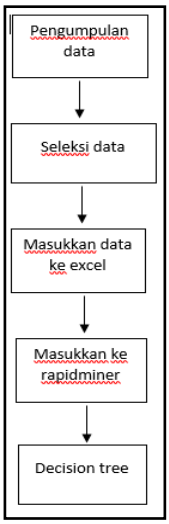
Penjabaran dari proses perhitungan algoritma C45 dapat digambarkan dengan *flowchart* sebagai berikut :



Gambar 2.2 *Flowchart* Algoritma C4.5

2.5 Klasifikasi *Rule Based*

Rule based atau algoritma berbasis aturan merupakan cara terbaik untuk merepresentasikan sejumlah bit data atau pengetahuan (Han dan Kamber, 2006). *Rule based* biasanya dituliskan dalam bentuk logika *IF-THEN* atau jika dibuat persamaannya yaitu: *IF condition THEN conlusion* contoh sebuah *rule* yaitu : *IF age = youth AND student = yes THEN buys_computer = yes* Pernyataan *IF* dari persamaan di atas dikenal sebagai *rule antecedent* atau *precondition* sedangkan pernyataan *THEN* disebut sebagai *rule consequent* . Dalam *rule antecedent* biasanya menyertakan satu atau lebih atribut (misalnya atribut *age* dan *student*) dan menggunakan logika *AND* jika menggunakan lebih dari satu atribut. *Rule consequent* merupakan prediksi kelas, dalam contoh di atas prediksinya yaitu membeli komputer atau *buys_computer = yes* (Han dan Kamber, 2006). Aturan-aturan dalam *rule based* dapat diturunkan dari pohon keputusan yang telah terbentuk. Karena pohon keputusan yang besar, terkadang sulit untuk menginterpretasikan pohon bentuk keputusan (Han dan Kamber, 2006). Agar pohon keputusan ini dapat lebih mudah dipahami oleh manusia, maka perlu diinterpretasikan dalam bentuk aturan-aturan atau *rule based*. Dalam kasus ini tidak digunakan logika OR, karena aturan-aturan diekstraksi langsung dari pohon keputusan yang disebut *mutually exclusive* dan *exhaustive*. Dengan *mutually exclusive* artinya tidak ada aturan yang berbenturan atau konflik karena tidak boleh ada dua aturan dalam tupel yang sama . Sedangkan *exhaustive* artinya dalam satu *set* aturan merupakan kombinasi nilai yang mungkin, artinya setiap aturan pasti menggambarkan kombinasi atribut dan nilai yang mungkin (Han dan Kamber, 2006) .



Gambar 2.3 Kerangka eksekusi pengujian

2.6 *Rapid Miner*

Rapid Miner merupakan perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari *Institute of Technologi Blanchardstown* dan Ralf Klinkenberg dari *rapid-i.com* dengan tampilan GUI (*Graphical User Interface*) sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini (Haryati et al,2015). Perangkat lunak ini bersifat *open source* dan dibuat dengan menggunakan program *Java* di bawah lisensi *GNU Public Licence* dan *Rapid Miner* dapat dijalankan disistem operasi manapun. Dengan menggunakan *Rapid Miner,* tidak dibutuhkan kemampuan koding khusus, karena semua fasilitas sudah disediakan.

Rapid Miner dikhususkan untuk penggunaan data mining. Model yang disediakan juga cukup banyak dan lengkap, seperti *Model Bayesian, Modelling, Tree Induction, Neural Network* dan lain-lain. Banyak metode yang disediakan oleh *Rapid Miner* mulai dari klasifikasi, *clustering*, asosiasi dan lain-lain. Jika tidak ada model atau model algoritma yang tidak ada dalam *Rapid Miner*, pengguna boleh menambahkan modul lain, karena *Rapid Miner* bersifat *open source,* jadi siapapun dapat ikut mengembangkan perangkat lunak ini.

2.7 Studi Literatur

Studi literatur dimaksudkan sebagai bahan pembanding peneliti dalam klasifikasi data penawaran pemeliharaan perkapalan menggunakan metode C45. Studi literatur dilakukan dengan mengumpulkan data dari beberapa jurnal penelitian yang telah dipublikasi yang berkesesuaian atau memiliki hubungan dengan penelitian ini. Ringkasan identitas literatur, isi literatur, serta perbandingan antara literatur dengan penelitian ini dapat dilihat pada Tabel 2.1.

Tabel 2.1 Ringkasan Studi Literatur

No	Sitasi	Algoritma	Jenis Data	Jumlah data latih
1.	Penerapan algoritma klasifikasi C4.5 dalam rekomendasi penerimaan mitra penjualan studi kasus : PT. Atria Artha Persada Arifin and Fitrianah, 2018	C45	Data mitra penjualan	107

Tabel 2.1 Ringkasan Studi Literatur *(lanjutan)*

No	Sitasi	Algoritma	Jenis data	Jumlah data latih
2.	Analisa algoritma C4.5 untuk memprediksi penjualan motor pada PT. Capella Dinamik Nusantara Cabang Muka Kuning Azwanti, 2018	C45	Data penjualan motor	2592
3.	C4.5 Algorithm Application for Prediction of Self Candidate New Students in Higher Education Darmawan,2018	C45	Prediksi Calon Mahasiswa Baru di Perguruan Tinggi	Tidak ada data latih
4.	Prediksi profit pada perusahaan dengan klasifikasi algoritma C4.5 Elisa, 2018	C45	Data konsultan konstruksi	Tidak ada data latih
5.	Metode decision tree algoritma C.45 dalam mengklasifikasi data penjualan bisnis gerai makanan cepat saji Cynthia and Ismanto, 2018	C45	Data pejualan bisnis gerai makanan cepat saji	Tidak ada data latih
6.	Analisa dan Penerapan Algoritma C4.5 dalam *Data Mining* Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Konstruksi PT. Arupadhatu Adisesanti Elisa, 2017	C45	Data kecelakaan kerja konstruksi	12

Tabel 2.1 Ringkasan Studi Literatur *(lanjutan)*

No	Sitasi	Algoritma	Jenis data	Jumlah data latih
7.	Penerapan Algoritma C4.5 untuk Memprediksi Penerimaan Calon Pegawai Baru di PT.WISE Harryanto and Hansun,2017	C45	Data penerimaan calon pegawai baru	Tidak ada data latih
8.	Prediksi kebangkrutan perusaahaan menggunakan algoritma C4.5 berbasis forward selection Saleh, 2017	C45	Data kebangkrutan perusahaan	Tidak ada data latih
9.	Penerapan *data mining* untuk menganalisa jumlah pelanggan aktif dengan menggunakan algoritma C4.5 Jamhur, 2016	C45	Data jumlah pelanggan aktif	19
10.	Implementasi data mining untuk memprediksi masa studi mahasiswa menggunakan algoritma C4.5 (Studi kasus : Universitas Dehasen Bengkulu) Haryati et al, 2015	C45	Data memprediksi masa studi mahasiswa	Tidak ada data latih
11.	Algoritma C45 dalam menganalisa kelayakan kredit (Studi kasus di koperasi pegawai Republik Indonesia (KP-RI) Lengayang Pesisir Selatan, Painan, Sumatera Barat Lusinia, 2014	C45	Data menganalisa kelayakan kredit	Tidak ada data latih

Berikut rincian studi literatur:

1. Dalam penelitian ini Arifin dkk pada tahun 2018 adanya permasalahan yang sering mucul dalam bisnis pada penjualan dengan sistem pembayaran kredit tempo adalah antara lain kredit macet, order fiktif, dan penipuan, maka hasil penelitian data dilakukan klasifikasi dan menggunakan algoritma C45 dalam proses pengklasifikasian.
2. Dalam penelitian ini Azwanti pada tahun 2018 Strategi pencapaian dalam penjualan membuat persaingan bisnis semakin tajam. Faktor utamanya adalah ketelitian konsumen dalam memilih suatu produk yang tidak hanya dilatarbelakangi oleh harga yang ekonomis, namun dapat membantu kegiatan sehari hari. Persaingan dunia bisnis ini terjadi pada seluruh perusahaan baik barang maupun jasa, termasuk di perusahaan penjualan motor, maka hasil penelitian menggunakan C45 untuk memprediksi penjualan motor.
3. Dalam penelitian ini Darmawan pada tahun 2018 *Data Mining* memiliki latar belakang dengan kondisi kelimpahan data (data yang berlebihan) dan informasi ledakan yang dihadapi oleh perusahaan, lembaga atau organisasi yang disimpan selama bertahun-tahun. Situasi ini juga dihadapi di beberapa perguruan tinggi yang menyimpan berbagai macam data terutama database penerimaan mahasiswa baru.
4. Dalam penelitian ini Elisa pada tahun 2018 dalam kegiatan proyek konstruksi, perencanaan dipergunakan sebagai acuan bagi pelaksana pekerjaan dan menjadi standar pelaksanaan proyek,meliputi dokumen, spesifikasi teknik, jadwal dan anggaran, maka hasil penelitian mengguanakan algoritma c45 dalam menentukan profit perusahaan.
5. Dalam penelitian ini Cynthia dkk pada tahun 2018 pada bisnis gerai makanan cepat saji XYZ dan diharapkan dapat memberikan informasi berupa klasifikasi penjualan menu makanan yang paling digemari pelanggan dan kurang digemari (laris dan tidak laris). Sehingga kedepannya pemilik bisnis ini dapat melakukan analisa menu mengikuti trend dan kegemaran pelanggannya, maka hasil penelitian menggunakaan algoritma C45. Persamaan dari apa yang saya buat adalah sama-sama menggunakan algoritma C45. Perbedaan dari data penjualan pada sebuah gerai makanan cepat saji dengan apa yang saya buat adalah data penjualan pada sebuah gerai makanan cepat saji mengumpulkan data dalam satu bulan (30 hari) sedangkan saya mengumpulkan data dari tahun 2012 sampai tahun 2018 .
6. Dalam penelitian ini Elisa pada tahun 2017 kecelakaan merupakan suatu kejadian yang tidak terencana begitu pun pada sebuah proyek konstruksi dimana kecalakaan sering terjadi hal

ini disebabkan oleh berbagai faktor, maka hasil penelitian menggunakan algoritma C45 untuk mengidentifikasi penyebab terjadinya kecelakaan kerja yang nantinya hasil penelitian ini dapat digunakan sebagai panduan untuk menghindari risiko kecelakaan.

7. Dalam penelitian ini Harryanto dkk pada tahun 2017 penerimaan calon pegawai baru merupakan sebuah tahap dimana sebuah perusahaan melakukan rektmen terhadap orangorang yang melamar ke perusahaan tersebut dan menentukan apakah orang tersebut memenuhi kriteria dan kebutuhan unit kerja pada perusahaan tersebut, maka hasil penelitian menggunakan C45 untuk memprediksi proses penerimaan calon pegawai baru.
8. Dalam penelitian ini Saleh pada tahun 2017 memprediksi kebangkrutan perusahaan adalah upaya yang penting dalam mengatasi masalah manajemen perusahan, maka hasil penelitian meggunakan algoritma C45.
9. Dalam penelitian ini Jamhur pada tahun 2016 penelitan digunakan untuk menganalisa jumlah pelanggan aktif di PT. Multidaya Prima, dengan menggunakan algoritma C45.
10. Dalam penelitian ini Haryati dkk pada tahun 2015 tujuan dr penelitian ini adalah dengan menggunakan pohon keputusan berbasis algoritma c45 dan diimplementasikan ke suatu aplikasi yaitu rapidminer diharapkan dapat meningkatkan keakuratan analisa massa studi mahasiswa.
11. Dalam penelitian ini Lusinia pada tahun 2014 algoritma C45 adalah algoritma klasifikasi pohon keputusan (decision tree). Pohon keputusan algoritma C45 dibangun dengan tiga tahap yaitu pemilihan atribut sebagai akar, membuat cabang untuk tiap-tiap nilai dan membagi kasus dalam cabang.

Perbedaan dari apa yang saya buat adalah data yang menggunakan laporan rencana *docking* pemeliharaan kapal dan pengelolaan data menggunakan aplikasi *rapidminer* untuk mengklasifikasi data dengan menggunakan metode algoritma C45. Untuk membangun *decision tree* (pohon keputusan) sehingga dapat melihat klasifikasi data.