BAB II

KAJIAN PUSTAKA

2.1 *Knowlegde discovery in database* (KDD)

Menurut (Luvia, Windarto, Solikhun, & Hartama, 2017) *data mining* dan *knowledge discovery in databases* (KDD) dapat digunakan secara bergantian yang berguna untuk menjelaskan proses mengekstraksi data yang tersembunyi dari *database* besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi saling terhubung dan saling berhubungan. Tahapan keseluruhan proses KDD adalah *data mining*.

(Mardi, 2017) *Knowledge Discovery In Database* (KDD) merupakan metode untuk mendapat dan memperoleh pengetahuan dari sebuah dari *database* yang ada. Dalam database terdapat record-record yang saling berhubungan satu sama lain. Hasil pengetahuan yang diperoleh dalam proses tersebut dapat digunakan sebagai basis pengetahuan untuk pengambilan keputusan dan penggalian informasi tersembunyi dalam suatu *database*.

Menurut (D. W. T. Putra, 2016) terdapat 4 tahapan menggunakan proses KDD, antara lain :

1. Pengumpulan data

Sumber data utama yang digunakan dalam melakukan penelitian merupakan data yang diperoleh dari data penjualan motor PT. Capella Dinamik Cabang Muka Kuning pada tahun 2019, dimana data tersebut akan dianalisis untuk memperolah

masalah-masalah yang ada untuk digunakan sebagai sampel dalam melanjutkan penelitian.

2. Penyeleksi data

Pada tahapan ini analisis penelitian diperoleh dari 15 jurnal serta 2 judul buku sebagai referensi sehingga pada tahapan penyeleksi sumber data, data yang akan diambil harus sesuai dengan referensi yang sudah ditetapkan. Pada tahapan ini data akan dikelompokkan sesuai kelas untuk difokuskan pada subset variable dimana variabel tersebut akan menjadi data penelitian yang akan digunakan pada proses *data mining* lalu disimpan dalam sebuah berkas yang akan dipisahkan dari basis data yang tidak dibutuhkan.

3. Preprocessing / Cleaning

Tujuan utama dari *Preprocessing / Cleaning* adalah untuk memperoleh sampel data yang akan digunakan sebagai acuan dalam penelitian, membersikan keseluruhan data yang tidak dibutuhkan, memeriksa isi data serta memperbaiki semua kesalahan yang ada pada data.

4. Transformasi data

Transformasi ini sangat berguna untuk mengintegrasikan data yang belum valid sehingga menjadikan data yang belum mempunyai entitas akan diubah keseluruhan menjadi sebuah data yang valid karena dalam proses *data mining* harus menggunakan data yang valid.

2.2 Data mining

Proses mendapatkan suatu informasi yang berguna dari sebuah database merupakan manfaat *data mining*, dari sebuah database yang besar akan di ekstrak dan diambil informasi baru yang berguna untuk membantu dalam pengambilan suatu keputusan, sering juga disebut *knowledge discovery* (Haryati, Sudarsono, & Suryana, 2015).

Menurut (Luvia et al., 2017), berikut ini adalah tahap-tahap *data mining* :

1. Pembersihan data (*data cleaning*)

   data cleaning merupakan proses membuang data yang tidak konsisten, *noise* dan data yang tidak relevan.

2. Integrasi data (*data integration*)

   data integration merupakan proses menggabungkan data dari beberapa *database* kedalam satu *database* baru yang lebih efektif.

3. Seleksi data (*data selection*)

   Hanya data yang dibutuhkan saja yang perlu diambil dari database untuk dianalisis karena tidak semua data bisa digunakan, untuk itu dilakukan seleksi data.

4. Transformasi data (*data transformation*)

   data transformation bertugas untuk mengubah dan menggabungkan data menjadi nilai kedalam format yang sesuai untuk selanjutnya diproses kedalam *data mining*, untuk data yang cakupannya terlalu luas akan dilakukan pengelompokan menjadi beberapa kelompok kecil (Putri & Waspada, 2018).

5. Proses *mining*

   Menemukan pengetahuan baru yang berharga dan tersembunyi dari berbagai *database* yang digabungkan merupakan istilah dari proses *mining*.

6. Evaluasi pola

   Untuk mengidentifikasikan pola-pola menarik yang didapatkan dari *database* kedalam *knowledge based* yang ditemukan.

7. Persentasi pengetahuan

   Visualisasi dan penyajian pengetahuan tentang metode yang dapat digunakan untuk memperoleh pengetahuan yang didapat pengguna.

Menurut (Mardi, 2017) pengelompokan *data mining* terdiri dari beberapa kelompok berdasarkan metode dan tugas pemrosesan nya adalah sabagai berikut.

1. Deskripsi

Deskripsi merupakan sebuah cara dalam menggambarkan dan menemukan sebuah pola dalam sebuah data. Dalam hal ini sering kali peneliti memiliki kebiasaan hanya menjelaskan suatu pola saja, hal ini dikarenakan karna sulitnya menentukan keterangan atau fakta mengenai data yang sudah diperoleh.

2. Estimasi

Sebuah model yang sudah dibangun menggunakan record data yang lengkap, estimasi hampir sama seperti klasifikasi, perbedaannya hanya divariabel target estimasi yang lebih cenderung mengarah ke numerik dibanding kearah kategori.

3. Prediksi

Prediksi merupakan suatu keadaan dimana nilai sebuah data dihasilkan belum diketahui kepastiannya, mirip dengan klasifikasi dan estimasi, kecuali dalam prediksi nilai dari hasil akan ada dimasa yang akan datang.

4. Klasifikasi

Teknik klasifikasi digunakan sebagai pendekatan secara sistematis untuk mengelompokan data berdasarkan dari kemiripan tertentu dengan menggunakan target variabel berfungsi sebagai pemisah atau sebagai penentu golongan pada tiap kategori (Gunawan, 2016). Misalnya, penggolongan pendapatan dipisahkan kedalam tiga kategori/kelas, yaitu pendapatan rendah, pendapatan sedang, dan pendapatan tinggi.

5. Pengklasteran

Pengklasteran bisa dikatakan sebagai pengelompokan record, memperhatikan atau pengamatan objek-objek yang mempunyai kemiripan yang hampir sama dengan objek lain, tetapi didalam pengklasteran tidak memiliki variabel target seperti pada klasifikasi.

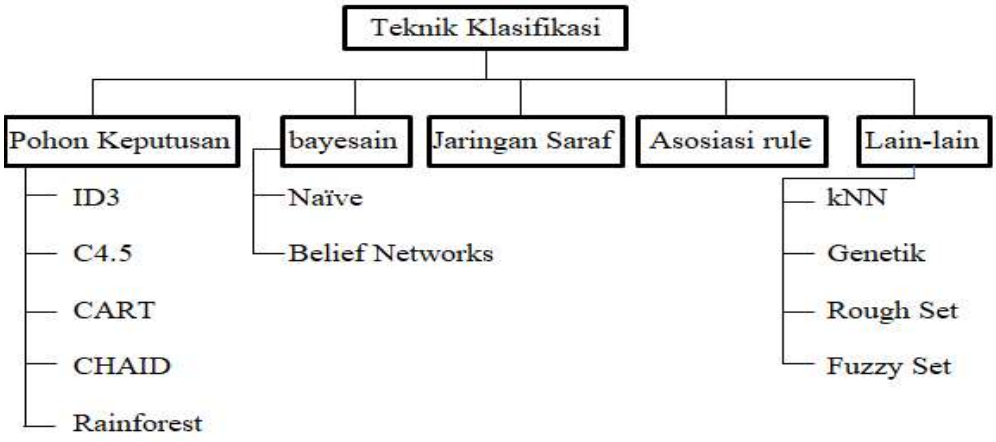
6. Asosiasi

Asosiasi berfungsi untuk menemukan suatu atribut atau mengidentifikasi hubunngan antara berbagai atribut yang muncul dalam suatu waktu.

2.3 Metode klasifikasi

Klasifikasi adalah proses untuk mengetahui suatu objek yang sudah didefinisikan sebelumnya. Klasifikasi berfungsi untuk memprediksi kelas dari

suatu objek yang atribut belum diketahui. Dengan memanipulasi data yang sudah didapatkan teknik ini dapat memberikan klasifikasi pada data baru yang telah diklasifikasi dan hasilnya dapat digunakan untuk memberikan sejumlah aturan (Eka Ratnawati, ., & Muflikhah, 2014). Klasifikasi pertama kali diterapkan pada bidang tanaman seperti yang dilakukan oleh Carolus von Linne merupakan peneliti pertama yang mengklasifikasi spesies tanaman berdasarkan karakteristik fisik. Setelah penelitian tersebut dia dikenal sebagai bapak klasifikasi(Mardi, 2017).



Gambar 2.1 Pengelompokan Teknik klasifikasi
Sumber : Data Penelitian, 2019

Algoritma klasifikasi berfungsi untuk menemukan hubungan antara nilai-nilai prediksi dan nilai atribut target (Brevik et al., 2016). Metode klasifikasi mempunyai banyak algoritma, apabila algoritma klasifikasi berbeda berarti harus menggunakan teknik yang juga berbeda untuk mencari hubungan atara nilai-nilai prediksi. Klasifikasi model dapat diuji dengan membandingkan nilai-nilai diprediksi nilai atribut target dikenal dalam satu set data uji (Han, Kamber, & Pei,

2012). Data untuk klasifikasi biasanya dibagi menjadi dua set data, satu untuk membangun model, yang lain untuk pengujian model. (P. Putra, Informatika, & 2018, 2018)

1. Pohon Keputusan

Didalam algoritma C4.5 ada beberapa metode yang dapat digunakan untuk klasifikasi tetapi yang paling sering digunakan adalah pohon keputusan (*decision tree*). Pohon keputusan merupakan sebuah metode yang dapat mengubah fakta yang sangat besar yang sudah didapatkan menjadi sebuah pohon keputusan yang merepresentasikan aturan (*rule*) yang dapat dengan mudah dipahami dengan bahasa alami (Mardi, 2017).

Manfaat utama yang didapat dengan menggunakan pohon keputusan adalah kemampuannya mengambil keputusan yang kompleks dan mudah dipahami menjadi lebih simpel (*break downproces*) sehingga pengambilan keputusan dapat mehasilkan solusi permasalahan. Algoritma C4.5 dan pohon keputusan merupakan dua model yang saling berhubungan dan tidak terpisahkan, karena untuk membangun sebuah pohon keputusan harus membutuhkan algoritma C4.5 (Haryati et al., 2015).

Secara umum ada beberapa langkah untuk membangun pohon keputusan (Rani, 2015) yakni :

1. Pilih atribut sebagai akar
2. Buat cabang untuk tiap-tiap atribut
3. Bagi kasus dalam cabang

Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

2. Algoritma C4.5

Algoritma C4.5 pertama kali diperkenalkan oleh J. Ross Quinlan yang merupakan pengembangan dari algoritma ID3, algoritma ID3 digunakan untuk membuat sebuah pohon keputusan yang berguna untuk mengekspolari data dan menemukan hubungan-hubungan tersembunyi antara sejumlah calon variabel-variabel input dengan variabel target. Pada dasarnya konsep dari algoritma C4.5 digunakan untuk mengubah data menjadi pohon keputusan dan menghasil aturan-aturan (*rule*) keputusan (Mukminin & Riana, 2017).

Algoritma C4.5 menggunakan konsep information gain atau entropy reduction untuk memilih pembagian akar atau node dari sebuah pohon keputusan. Tahapan dalam membuat pohon keputusan dengan algoritma C4.5 (Septiani, 2017) yaitu:

1. Mempersiapkan data training, dapat diambil dari data histori atau data penjualan yang pernah terjadi sebelumnya serta mewawancarai pihak yang terkait dalam penelitian guna mendapatkan sampel data yang dibutuhkan, selanjutnya data yang sudah didapatkan kemudian dikelompokkan dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon keputusan yang didapatkan dengan menghitung nilai gain yang tertinggi dari masing-masing *entropy* yang didapatkan dari tiap-tiap partisi atribut yang akan dijadikan sebagai akar dalam pohon keputusan dengan rumus sebagai berikut :

$$Entropy\ (s) = -\sum_{i=1}^{n} -pi * \log 2\ pi$$

Dimana :

S : himpunan kasus

A : fitur

N : jumlah partisi

Pi : proporsi dari Si terhadap S

3. Hitung nilai gain dengan menggunakan rumus sebagai berikut :

$$Gain(S, A) = Entropy\ (S) \sum_{i=1}^{p} \frac{Si}{S} * \text{Entropy (Si)}$$

Dimana :

S : himpunan kasus

A : atribut

N ; jumlah partisi ke A

Si : jumlah kasus pada partisi ke_i

S : jumlah kasus dalam S

4. Ulangi kembali langkah ke-2 hingga semua record terpartisi mendapatkan nilai entropy dan gain. Proses dari partisi pohon keputusan akan selesai dan berhenti apabila:

a. Semua tupel dalam record dalam simpul N mendapat kelas yang sama.

b. Tidak ada atribut dalam record yang dipartisi lagi.

c. Tidak ada record di dalam cabang yang kosong.

2.4 *Software* Pendukung

2.4.1 *Rapidminer*

Data mining mempunyai beberapa perangkat lunak aplikasi atau software pendukung seperti *WEKA, Orange, Microsoft Analysis Service, Oracle Data Mining*, dan *RapidMiner* dimana aplikasi-aplikasi ini dapat membantu dalam pengolahan *data mining*. Pada penelitian ini, peneliti akan menggunakan aplikasi *RapidMine*r untuk melakukan pengujian.

RapidMiner adalah *software* yang dibuat oleh Dr. Markus *Hofmann dari Institute of Technologi Blanchardstown* dan Ralf Klinkenberg dari rapid-i.com dengan tampilan GUI sehingga memudahkan pengguna dalam mengolah data menggunakan perangkat lunak ini.

Perangkat lunak ini bersifat open source yang dapat dijalankan di sistem operasi manapun dan dibuat dengan menggunakan program *Java*. Dengan menggunakan *RapidMiner*, tidak dibutuhkan kemampuan menggunakan bahasa pemograman, karena semua fasilits sudah disediakan. *RapidMiner* dikhususkan untuk penggunaan *data mining*. Model yang disediakan, seperti Model Bayesian, Modelling, Tree Induction, Neural Network dan lain-lain. (Haryati et al., 2015)

2.5 Penelitian Terdahulu

Terdapat beberapa penelitian yang digunakan peneliti sebagai referensi dalam melakukan penelitian yang berupa jurnal, referensi-referensi tersebut diambil karena adanya permasalahan atau metode yang berhubungan dengan penelitian yang dibahas didalamnya , antara lain sebagai berikut:

1. Pada jurnal referensi yang ditulis oleh (P. Putra et al., 2018) yang berjudul **PENGEMBANGAN APLIKASI PERHITUNGAN PREDIKSI STOCK MOTOR MENGGUNAKAN ALGORITMA C 4.5 SEBAGAI BAGIAN DARI SISTEM PENGAMBILAN KEPUTUSAN,** meneliti tentang inventaris sepeda motor sesuai dengan kebutuhan konsumen. Berdasarkan proses perhitungan yang telah dilakukan beberapa tahap selama implementasi pemrosesan data C4.5 diperoleh pohon keputusan dengan 12 aturan-aturan (*rule*) dalam menentukan prediksi jumlah persediaan stock motor pada dealer saudara motor.
2. Pada jurnal referensi yang ditulis oleh (Harahap, 2015) yang berjudul **PENERAPAN *DATA MINING* DALAM MEMPREDIKSI PEMBELIAN CAT FITRIANA,** meneliti tentang Departement Penjualan Home Smart Medan mengolah data perusahaan hanya menggunakan aplikasi Microsoft excel sehingga terkadang tidak dapat bersaing dengan perusahaan lain dalam mengolah data untuk dijadikan informasi sebagai strategi untuk penjualan. Berdasarkan hasil penelitian yang dilakukan bahwa pembelian cat dengan menggunakan metode *data mining* khususnya algoritma C4.5 akan bermanfaat sekali dalam proses pengambilan keputusan dalam pembelian cat pada Home Smart Medan.
3. Pada jurnal referensi yang ditulis oleh (Nababan & Tanlim, 2019) yang berjudul **ANALIS PERSEDIAAN STOK BARANG MENGGUNAKAN ALGORITMA C4.5 PADA CV HARAPAN JAYA,** meneliti tentang jumlah penjualan yang fluktuatif mengakibatkan stok barang yang tersedia

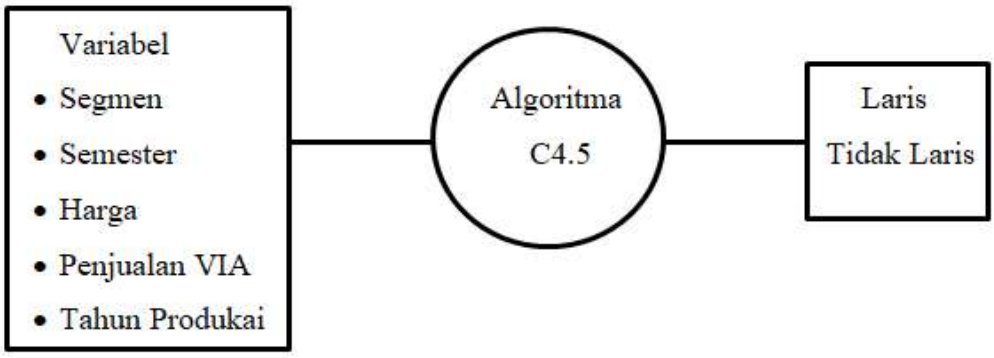
tidak stabil dan dapat berdampak langsung ke retailer. Berdasarkan penelitian dan perhitungan menggunakan algoritma C4.5 maka diperoleh pohon keputusan dengan 9 aturan-aturan (*rule*) dalam menentukan prediksi jumlah persediaan stok barang pada CV Harapan Jaya.

4. Pada jurnal referensi yang ditulis oleh (Jurnal & Informasi, 2016) yang berjudul **PENERAPAN *DATA MINING* UNTUK MEMPREDIKSI PENJUALAN WALLPAPER MENGGUNAKAN ALGORITMA C4.5** meneliti tentang cara mempermudah penjual memilih wallpaper mana yang banyak diminati konsumen agar disediakan stok maka perlu dilakukan prediksi untuk penjualan wallpaper terbanyak dengan metode klasifikasi. Berdasarkan penelitian yang dilakukan yang menjadi faktor utama yang mempengaruhi penjualan adalah faktor jumlah model wallpaper sedangkan faktor harga, ukuran, kualitas bahan dan warna tidak mempengaruhi pembelian.
5. Pada jurnal referensi yang ditulis oleh (Permasalahan & Masalah, 2017) yang berjudul **PERANCANGAN APLIKASI TREND PENJUALAN DAN STOK BARANG MENGGUNAKAN WAREHOUSE DAN *DATA MINING*** meneliti tentang penggunaan data warehouse perusahaan untuk mendapatkan ramalan penjualan dengan *data mining*. Dari hasil penelitian yang dilakukan, penggunaan aplikasi analisa *trend* penjualan ini dapat dijadikan acuan pada perusahaan dalam mengambil kebijakan dan strategi tepat yang dibutuhkan agar target penjualan semakin meningkat dimasa mendatang.

6. Pada jurnal referensi yang ditulis oleh (Elisa, 2017) yang berjudul **ANALISA DAN PENERAPAN ALGORITMA C4.5 DALAM *DATA MINING* UNTUK MENGIDENTIFIKASI FAKTOR-FAKTOR PENYEBAB KECELAKAAN KERJA KONTRUKSI PT.ARUPADHATU ADISESANTI** meneliti tentang penggunaan algoritma C.45 untuk mengidentifikasi penyebab kecelakaan kerja agar dapat digunakan sebagai panduan untuk menghindari resiko kecelakaan (zero accident), dari hasil penelitian yang dilakukan diambil sebuah kesimpulan bahwa metode Algoritma C4.5 atau pohon keputusan lebih efektif dan fleksibel jika digunakan pada proses pengklasifikasian.
7. Pada jurnal referensi yang ditulis oleh (Jamhur, 2016) yang berjudul tentang **PENERAPAN *DATA MINING* UNTUK MENGANALISA JUMLAH PELANGGAN AKTIF DENGAN MENGGUNAKAN ALGORITMA C4.5** meneliti tentang tingkat kepuasan pelanggan dengan mengetahui jumlah pelanggan aktif Untuk itu perlu adanya pengolahan data tentang pelanggan aktif yaitu dengan menggunakan algoritma C4.5. Hasil dari penelitian dapat menjadi kriteria evaluasi untuk pelanggan aktif dan tidak aktif menggunakan algoritma C4.5 dan membuat aturan yang dapat menggambarkan proses yang terkait dengan pelanggan aktif dan tidak aktif.

2.6 Kerangka Pemikiran

Penelitian dilakukan menggunakan tahapan-tahapan kegiatan dengan mengikuti kerangka pemikiran yang meliputi metode pengumpulan data, analisis data, dan pengujian hasil. Berikut ini adalah kerangka pemikiran penulis dalam melakukan penelitian.



Gambar 2.2 Kerangka Pemikiran
Sumber : Data Penelitian, 2019

Pada gambar 2.2, penelitian ini menggunakan 4 variabel sebagai input data yaitu segmen, Harga, semester, penjualan VIA dan tahun produksi kemudian data tersebut akan diolah menggunakan *algoritma C4.5* dengan teknik klasifikasi untuk mendapat *output* laris dan tidak laris, dari situ kita akan membuat pohon keputusan (*decision tree*) yang akan menghasilkan rule atau aturan berdasarkan pohon keputusan untuk mendapatkan hasil dan menyimpulkan penelitian yang telah dilakukan.