BAB II

KAJIAN PUSTAKA

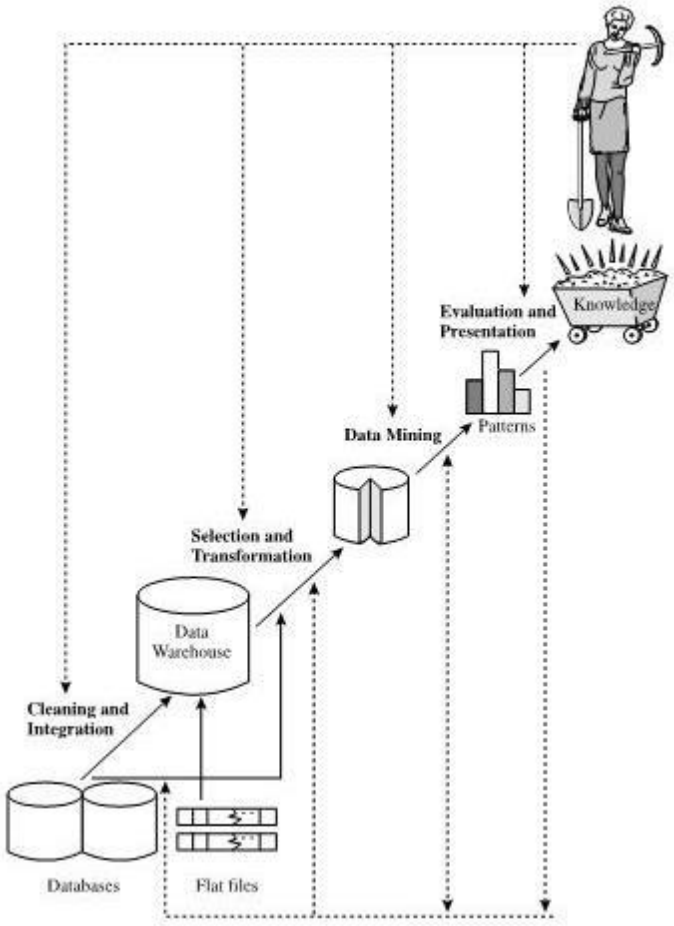
2.1. *Knowledge Discovery in Database* (KDD)

Pada zaman sekarang, data telah diproses secara komputerisasi, baik dalam bidang apapun yang berdampak memudahkan pengguna dalam mengumpulkan data. Namun sering terjadi keadaan dimana data yang disimpan telah menumpuk tidak dapat menghasilkan apapun dan menjadi tidak berguna. Keadaan ini disebut dengan isitilah *Rich of Data but Poor Information.*

Dalam beberapa tahun terakhir, data yang ada semakin heterogen dan kompleks dengan volume yang terus meningkat dengan cepat sekali. Menurut John Gantz, diperkirakan pada tahun 2020 nantinya, volume data yang ada bisa mencapai 44 *zettabytes* dan akan terus berkembang. Jika perusahaan dapat menganalisa data tersebut, perusahaan tersebut dapat melakuan pengambilan keputusan strategis dengan lebih baik dan lebih mudah (Suyanto, 2017).

Kebutuhan untuk memperoleh nilai tambah dari kumpulan data yang terkumpul menyebabkan lahirnya teknik-teknik analisa data yang akhirnya melahirkan metodologi *Data Mining*. *Data mining* merupakan salah satu tahapan dalam sebuah proses yang disebut *Knowledge Discovery and Data Mining* (KDD) yang dimana proses ini berfungsi untuk menggali dan menganalisis himpunan data sehingga informasi dan pengetahuan yang berguna dapat terekstrak dengan baik.

Terdapat beberapa proses iterasi pada proses KDD, yakni (Edelstein, 1999) :



Gambar 2.1 Tahapan pada KDD

1. *Data Selection* (Pemilihan Data)

   Proses pemilihan data bertujuan untuk memilih data penelitian yang akan diproses. Data yang dipilih akan disortir menjadi *data training* dan *data testing*.

2. *Data Cleaning* (Pembersihan Data) / *Preproccesing*

   Proses pembersihan data bertujuan untuk menghilangkan data yang sama (duplikasi) serta menghapus data yang salah.

3. Transformasi Data / *Transformation*

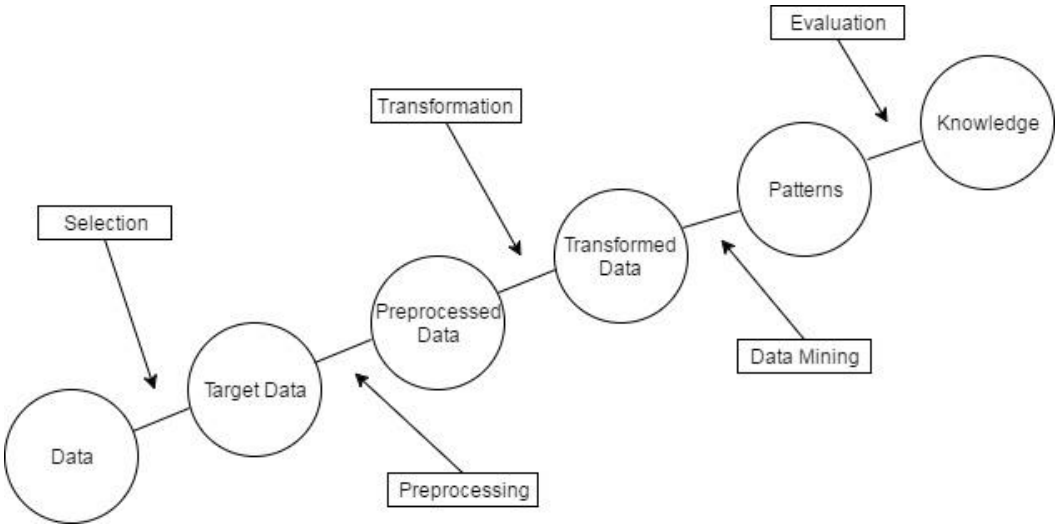
Proses transformasi data bertujuan untuk mengubah bentuk data, yaitu dari data kuantitatif menjadi data kualitatif.

4. *Data Mining*

   Proses data mining bertujuan untuk memproses data yang sudah lalu mendapatkan pola pohon keputusan yang akan menjadi solusi dari permasalahan yang ada.

5. *Evaluation*

   Proses ini merupakan proses pemeriksaan ketepatan pohon keputusan yang dibentuk melalui proses *data mining* sebelumnya dengan menggunakan algoritma yang bersangkutan.



Gambar 2.2 *Flowchart* KDD

KDD dapat diartikan sebagai seluruh proses *non-trivial* yang berhubungan dengan teknik integrasi dan penemuan ilmiah yang terjadi dalam mencari dan mengidentifikasi pola (*pattern*) dalam data, yang dimana pola yang ditemukan

nantinya bersifat sah, baru, dan dapat bermanfaat serta dapat dimengerti. (Sari & Sindunata, 2014).

2.2. *Data Mining*

Seperti yang dapat dilihat pada gambar 2.2, *data mining* merupakan bagian dari proses yang terdapat pada proses *Knowledge Discovery in Database* (KDD). Umumnya, *data mining* sering disamakan dengan KDD disebabkan sebagian besar pekerjaan pada KDD difokuskan pada proses *data mining* tersebut. (Lorena, Zarman, & Hamidah, 2014). *Data mining* sendiri memiliki pengertian suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam data yang jumlahnya sangat besar. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi and mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban, 1998). *Data mining* adalah suatu proses untuk mengetahui pola yang bermanfaat dalam sebuah kumpulan data yang besar (Larose & Larose, 2014). Dapat disimpulkan bahwa *data mining* merupakan metode untuk menemukan suatu pengetahuan dari data yang sangat besar atau dari sebuah database. Data mining adalah proses menggali dan menganalisa sejumlah data yang sangat besar untuk memperoleh sesuatu yang benar, baru, sangat bermanfaat dan akhirnya dapat menemukan suatu corak atau pola dalam data tersebut. *Data mining* memiliki beberapa karateristik sebagai berikut (Lorena et al., 2014) :

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

Ada beberapa faktor yang mempengaruhi kemajuan pada bidang *data mining* yang berdampak besar, yakni (Larose & Larose, 2014) :

1. Penggunaan kumpulan data yang mengalami pertumbuhan cepat.
2. Penggunaan *data warehouse* untuk menyimpan data.
3. Pemanfaatan Internet dalam peningkatan akses data.
4. Kompetisi bisnis yang mengakibatkan globalisasi ekonomi.
5. Pengembangan *software data mining*
6. Pengembangan kemampuan komputasi dan pengembangan media penyimpanan data.

Data mining dapat diterapkan dalam kehidupan kita sehari-hari dan dapat memberikan kita solusi mengenai masalah yang ada, seperti (Santoso, 2014) :

1. Mengetahui kondisi pasar

   Data mining dapat mengelompokan pembeli dengan kriteria dan model konsumen yang sama dengan menggunakan metode *clustering* dan dapat melakukan klasifikasi terhadap konsumen.

2. Mengetahui pola beli konsumen

   Data mining dapat melakukan klasifikasi terhadap pola beli konsumen yang akhirnya menghasilkan sebuah pengetahuan baru untuk dimanfaatkan di masa yang akan mendatang.

3. *Cross-market Analysis*

   Data mining dengan menggunakan metode apriori dapat menghasilkan pengetahuan berupa hubungan antara penjualan satu jenis produk dengan produk lainnya.

4. Profil Konsumen

   Data mining dengan metode *clustering* dapat mengelompokan konsumen berdasarkan profil yang dimiliki masing-masing.

2.3. Metode Data Mining

Secara umum, terdapat dua jenis metode pada *data mining*, yaitu metode prediksi (*predictive*) dan metode deskripsi (*descriptive*). Metode prediksi adalah proses untuk menemukan pola dari data yang menggunakan beberapa variabel untuk memprediksi variabel lainnya yang masih belum diketahui jenis ataupun nilainya, sedangkan yang dimaksud dengan metode deskripsi adalah proses untuk menemukan suatu karateristik penting dari sebuah data dalam *database*.

Di bawah ini merupakan pengertian dari jenis-jenis metode pada *data mining* yang paling umum, yaitu (Larose & Larose, 2014) :

1. Deskripsi / *Description*

   Peneliti mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data secara sederhana. Deskripsi sering memberikan kemungkinan penjelasan dalam sebuah pola atau kecenderungan dalam sebuah data yang besar.

2. Estimasi / *Estimation*

   Estimasi merupakan model yang dibangun menyerupai klasifikasi terhadap variabel yang bersifat numerik dengan berdasarkan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi.

3. Prediksi / *Prediction*

   Prediksi memiliki permodelan yang hampir sama dengan estimasi dan klasifikasi, kecuali pada prediksi, terdapat variabel waktu yaitu di masa mendatang.

4. Klasifikasi / *Classification*

   Klasifikasi hampir sama dengan estimasi, hanya letak perbedaannya adalah target variabel klasifikasi berupa variabel bersifat kategori sehingga tujuan dari peneliti cenderung ke membagi variabel ke dalam kategori-kategori.

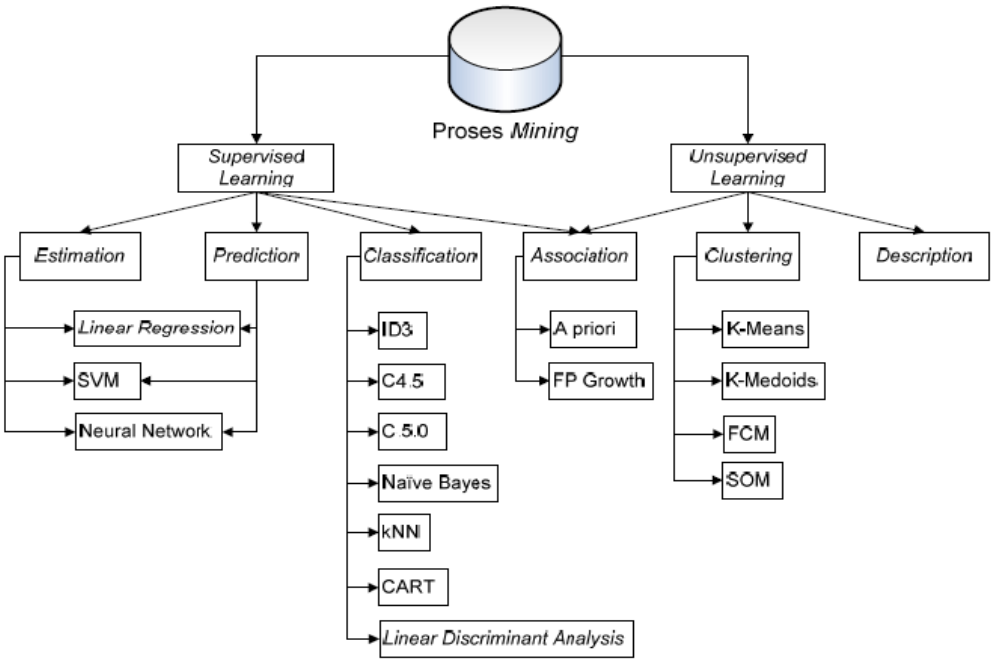
5. Klustering / *Clustering*

   Klustering lebih diartikan ke dalam pengelompokan sejumlah *record*, pengamatan atau kasus ke dalam kelas-kelas yang memiliki sifat yang hampir sama. Letak perbedaan klustering dengan klasifikasi adalah tidak adanya

variabel target pada klustering sehingga klustering lebih kepada melakukan pembagian seluruh data menjadi kelompok-kelompok dengan kemiripan.

6. Asosiasi / *Association*

Peran asosiasi adalah untuk mencari atau menemukan atribut yang muncul secara bersamaan atau atribut yang memiliki hubungan antar dua atau lebih.



Gambar 2.3 Model Metode *Data Mining*

Menurut Wu, Kumar (2009), model *data mining* terdiri dari seperangkat aturan atau persamaan yang kompleks yang digunakan untuk mengidentifikasi pola dari sebuah *dataset* yang bertujuan untuk memahami pengetahuan yang ada di dalamnya. Model metode *data mining* dikelompokan sebagai berikut (Bianglala et al., 2013) :

1. *Supervised Model*

   Supervised model adalah model yang telah diarahkan dengan tujuan untuk memprediksi suatu hal di masa yang akan mendatang. Metode model prediksi dikategorikan termasuk ke dalam model klasfikasi dan model estimasi.

   a. Klasifikasi

      Klasifikasi merupakan model dimana kelompok atau kelas telah diketahui di awal. Tujuan dari permodelan ini adalah untuk mengklasifikasikan kasus ke kelompok yang telah ditetapkan.

   b. Estimasi

      Estimasi memiliki model yang tidak jauh berbeda dengan klasifikasi, tapi ada perbedaan yang mencolok yaitu estimasi digunakan untuk memprediksi nilai secara berterusan / *continue* didasarkan pada nilai yang diamati.

2. *Unsupervised Model*

   Pada *unsupervised model*, tidak ada bidang *output* dalam permodelannya, hanya terdapat *input*. Pengenalan pola terarah namun tidak ada panduan oleh atribut tertentu.

a. Model *cluster*

   Pada permodelan *cluster*, kelompok tidak dikenal dari awal sehingga tujuan dari permodelan *cluster* ini adalah untuk menentukan *cluster* pada kasus yang ada.

b. Asosiasi dan *sequence model*

Model asosiasi juga termasuk ke dalam *unsupervised model* disebabkan model ini tidak langsung melibatkan prediksi secara langsung pada sebuah bidang khusus, melainkan setiap bidang yang ada memiliki peran ganda yaitu sebagai input dan juga output.

2.3.1. Metode Prediksi

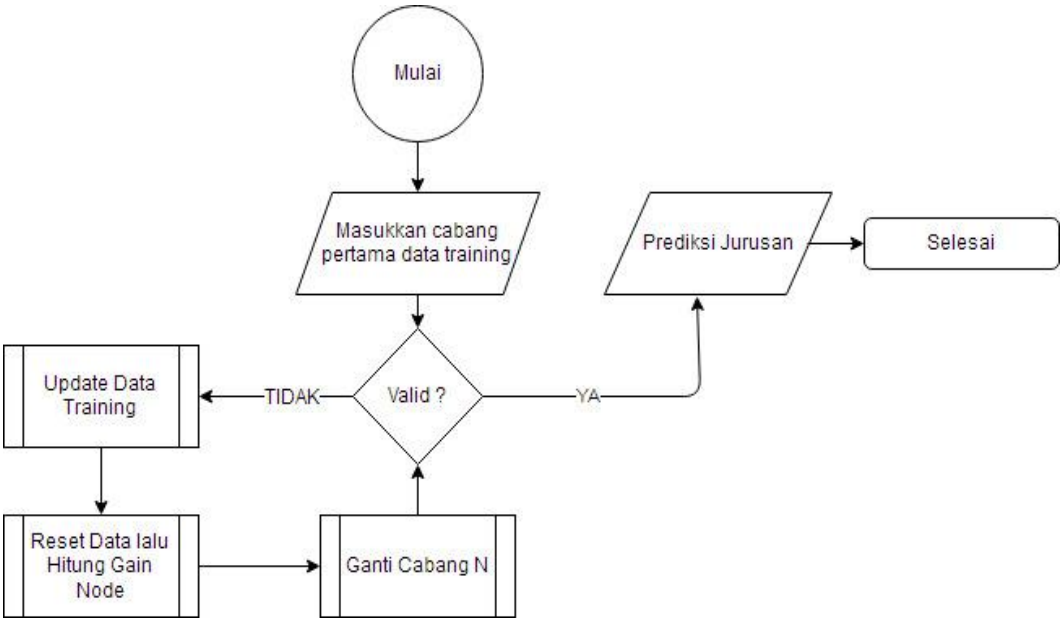
Seiring perkembangan zaman dan teknologi, penerapan *data mining* tidak lagi difokuskan pada bidan teknik informasi saja melainkan penggunaan *data mining* telah meluas dari bidang tersebut ke bidang-bidang dasar lainnya. Salah satu bidang yang telah menganut konsep *data mining* untuk memudahkan pekerjaannya yaitu bidang pendidikan. *Educational Data Mining* (EDM) merupakan suatu aliran yang baru dalam penelitian *data mining.* EDM menggunakan beberapa teknik yang bertujuan untuk mengeksplorasi jenis yang unik dari data-data pada pengelolaan pendidikan agar lembaga pendidikan dapat lebih memahami siswa (Kurniawan, Wibowo, & Astuti, 2007).

Prediksi adalah suatu proses memperkirakan sesuatu yang akan terjadi di masa yang akan datang. Prediksi dapat diperoleh melalui metode ilmiah yang nantinya dapat dimanfaatkan untuk melakukan klasifikasi, estimasi dan banyak lagi. Prediksi dapat menghasilkan *class* berdasarkan berbagai atribut yang telah disediakan. Teknik prediksi dapat dilakukan dengan beberapa metode yang salah satunya adalah dengan menggunakan *decision tree* menggunakan algoritma C4.5.

2.3.2. Algoritma C4.5

Algoritma yang digunakan pada penelitian adalah algoritma C4.5. Metode algoritma C4.5 merupakan pengembangan dari algoritma ID3 yang sama-sama diciptakan oleh seorang peneliti di bidang kecerdasan buatan / *Artificial Inteligence* (AI) yang bernama J. Rose Quilan pada akhir tahun 1970-an (Larose & Larose, 2014). Algoritma C4.5 merupakan algoritma yang menghasilkan pohon keputusan (Kusrini & Luthfi, 2009). Pohon keputusan akan berbentuk atas ke bawah dengan yang berada paling atas akan disebut dengan akar, sedangkan yang di bawah akan disebut dengan daun.

Gambar di bawah akan menjelaskan bagaimana alur algoritma C4.5 bekerja dalam penelitian ini secara sederhana :



Gambar 2.4 *Flowchart* Algoritma C4.5

Dalam pembuatan pohon keputusan dari algoritma C4.5 ini, ada beberapa tahapan yang harus dilalui (Kusrini & Luthfi, 2009) :

1. Mempersiapkan data *training*. Data *training* diambil dari hal-hal yang telah terjadi sebelumnya yang sudah dikelompokkan dalam bentuk kelas.
2. Menghitung akar dari pohon keputusan. Akar akan diambil dari atribut yang terpilih, dengan cara menghitung *gain* dari masing-masing atribut, lalu nilai *gain* yang paling tinggi akan menjadi akar pertama. Sebelum menghitung *gain*, kita harus mencari nilai *entropy* menggunakan rumus :

$$\text{Entropy}(S) = \sum_{i=1}^{n} - Pi * \log_2 Pi$$

Rumus 2.1. Perhitungan Entropi

dengan keteragan sebagai berikut :

S = Himpunan kasus yang ada

n = jumlah partisi S

Pi = proporsi S_i terhadap S

3. Kemudian gunakan rumus berikut untuk menghitung nilai *gain* tiap atribut.

$$\text{Gain}(S) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * \text{Entropy}(S)$$

Rumus 2.2. Perhitungan *Gain*

dengan keterangan sebagai berikut :

S : himpunan kasus

A : atribut

N : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi i

|S| : jumlah kasus dalam S

4. Ulangi langkah ke-2 hingga semua tupel terpartisi.
5. Proses partisi pohon akan berhenti jika :

a. Semua tupel dalam node N mendapat kelas yang sama

b. Tidak ada atribut dalam tupel untuk dipartisi lagi

c. Tidak ada tupel dalam cabang yang kosong

2.3.3. Pohon Keputusan

Salah satu cara menampilkan hasil *data mining* metode prediksi adalah dengan menggunakan pohon keputusan atau sering disebut dengan *decision tree*. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhierarki. Pembangunannya yang cepat dan hasil model yang dibangun mudah untuk dimengerti membuat teknik *decision tree* populer dan sangat diminati. *Decision tree* berbentuk seperti *flowchart* yang menyerupai struktur pohon dimana setiap *node* menunjukkan sebuah tes pada atribut dan tiap cabang menunjukkan hasil tes tersebut serta pembagian *class* yang ditunjukkan pada *leaf* pohon (Lorena et al., 2014).

Decision tree merupakan salah satu metode belajar yang sangat populer dan banyak digunakan secara praktis. Metode *decision tree* merupakan metode yang berusaha menemuka fungsi-fungsi pendekatan yang bernilai diskrit dan tahan terhadap data-data yang memiliki kesalahan. *Decision tree* merupakan sebuah struktur yang terdiri dari simpul akar (*root node*), simpul percabangan (*branch node*) dan simpul daun (*leaf node*) (Vulandari, 2017).

2.3.4. *Rule Based*

Menurut Han & Kamber, *rule based* adalah algoritma berbasis aturan yang merupakan cara terbaik dalam merepresentasikan sejumlah data atau pengetahuan. *Rule based* biasanya dituliskan dalam bentuk logika *IF-THEN* atau dalam bentuk persamaan *IF condition THEN conclusion*. Pernyataan *IF* dari persamaan tersebut dapat diartikan sebagai *rule antecedent* atau *precondition* sedangkan pada pernyataan *THEN* diartikan sebagai *rule consequent*. *Rule based* akan digunakan pada penelitian ini sebagai cara untuk memudahkan pembaca untuk memahami pohon keputusan yang akan diciptakan nantinya (Haryati, Sudarsono, & Suryana, 2015).

2.4. *Software* Pendukung

Data mining mempunyai beberapa perangkat lunak aplikasi atau *software* pendukung seperti WEKA, *Orange*, *Microsoft Analysis Service*, *Oracle Data Mining*, dan *RapidMiner*. Pada penelitian ini, peneliti akan menggunakan aplikasi *RapidMiner* untuk melakukan pengujian terhadap pohon keputusan.

RapidMiner adalah sebuah perangkat lunak atau aplikasi yang diciptakan oleh Dr. Markus Hofman dari *Institute of Technology Blanchardtown* dan Ralf Klinkenberg dengan tampilan GUI (*Graphical User Interface*) yang memudahkan pengguna dalam menggunakan aplikasi ini. Perangkat lunak ini bersifat *open source* dan dibuat dengan menggunakan program Java. Walaupun diciptakan

dengan menggunakan program Java, pengguna tidak perlu melakukan koding dalam penggunaan *RapidMiner* disebabkan semua fasilitas telah disediakan. *RapidMiner* adalah suatu aplikasi yang ditujukan secara khusus dalam pengelolaan *data mining* dengan fasilitas model-model *data mining* seperti Model *Bayes*, *Decision Tree*, *Neural Network*, dan banyak lagi (Haryati et al., 2015).

2.5. Penelitian Terdahulu

Ada beberapa penelitian yang digunakan penulis sebagai referensi dalam penelitian yang dilakukan. Referensi diambil karena adanya permasalahan yang berhubungan dengan penelitian penulis yang dibahas di dalamnya, antara lain :

1. Pada jurnal referensi yang ditulis oleh Obbie Kristanto yang berjudul **PENERAPAN ALGORITMA KLASIFIKASI DATA MINING ID3 UNTUK MENENTUKAN PENJURUSAN SISWA SMAN 6 SEMARANG**, membahas mengenai perancangan pembentukan sistem software dalam penentuan penjurusan yang menggunakan variabel nilai akademis IPA dan IPS. Setelah perhitungan dengan menggunakan algoritma ID3 yang tidak beda jauh dengan algoritma C4.5, menghasilkan akurasi program software yang memiliki akurasi 80% (Obbie Kristanto, 2013).
2. Pada jurnal referensi yang ditulis oleh Siska Haryati yang berjudul **IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA C4.5 (STUDI KASUS : UNIVERSITAS DEHASEN BENGKULU)**, membahas

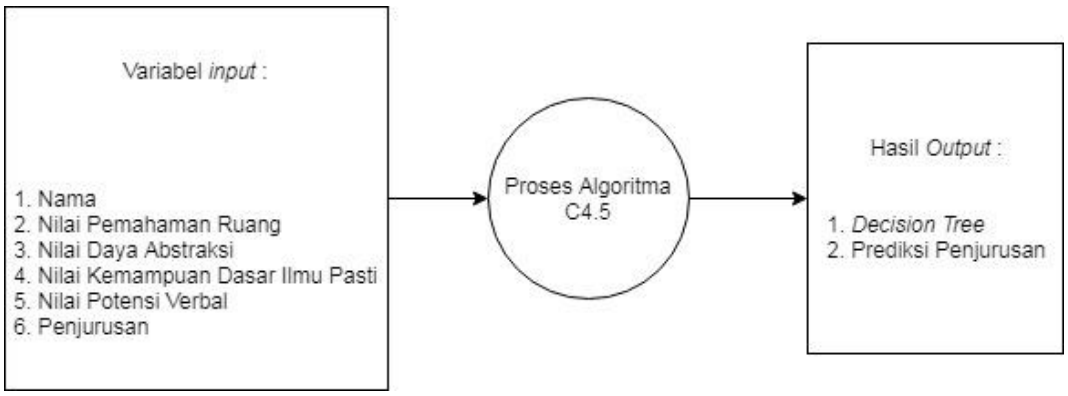
mengenai implementasi algoritma C4.5 dalam bentuk perancangan sistem baru untuk menggantikan sistem lama menggunakan program RapidMiner yang bertujuan menganalisa prediksi kelulusan mahasiswa dengan memanfaatkan klasifikasi data mining. Kesimpulan yang diambil adalah sistem baru lebih mudah, efektif dan efisien (Haryati et al., 2015).

3. Pada jurnal referensi yang ditulis oleh David Hartanto Kamagi yang berjudul **IMPLEMENTASI DATA MINING DENGAN ALGORTIMA C4.5 UNTUK MEMPREDIKSI TINGKAT KELULUSAN MAHASISWA**, membahas mengenai memprediksi kelulusan mahasiswa berdasarkan nilai IP total serta IP sementara yang dimiliki oleh mahasiswa yang dijadikan uji coba dengan menggunakan data mining algoritma C4.5. Kesimpulan yang diambil dari penelitian yang dilakukan adalah ketepatan yang mencapai 87.50% dari penelitian yang dilakukan sehingga dapat diimplementasikan (Kamagi & Hansun, 2014).
4. Pada jurnal referensi yang ditulis oleh Veronica Andriyana yang berjudul **PERBANDINGAN 3 METODE DALAM DATA MINING UNTUK PREDIKSI PENERIMA BEASISWA BERDASARKAN PRESTASI DI SMA NEGERI 6 SURAKARTA**, membahas mengenai perbandingan 3 metode prediksi data mining yaitu *Naives Bayes*, *Decision Tree Algorithm* ID3, dan Regresi Linear terhadap variabel yang berupa atribut nilai. Kesimpulan dari penelitian adalah bahwa algoritma ID3 menghasilkan nilai ketepatan paling tinggi dibanding metode lain (Andriyana, 2015).

5. Pada jurnal referensi yang ditulis oleh Sahirul Muklis yang berjudul **SISTEM PENDUKUNG KEPUTUSAN MINAT PEMILIHAN JURUSAN SMA DENGAN METODE K-MEANS CLUSTER ANALYSIS**, membahas mengenai penerapan metode K-Means Cluster Analysis dalam menentukan penjurusan dengan memperhatikan variabel nilai akademis per semester dan nilai Ujian Nasional. Dengan melakukan iterasi metode K-Means, maka didapatkan kesimpulan berupa penerapan pengambilan keputusan penjurusan dengan menggunakan metode K-Means Cluster Analysiss lebih mudah dan lebih efektif (Muklis, 2015).

2.6. Kerangka Pemikiran

Dalam penelitian ini, penulis akan menggunakan skema / *flowchart* berikut:



Gambar 2.5 Kerangka Pemikiran

Penjelasan dari gambar 2.5, awal proses penelitian diawali dengan data-data yang akan menjadi *input* dari penelitian yang dilakukan. Data-data akan

digabungkan menjadi sebuah *dataset* yang nantinya akan diproses melalui beberapa tahapan seperti *data selection*, *data cleaning* dan *data transformation*. Setelah melalui tahapan-tahapan tersebut, lalu data pada *dataset* akan melalui tahapan *data mining.* Tahapan *data mining* yang dilakukan menggunakan proses algoritma C4.5 yang nantinya akan menghasilkan sebuah pengetahuan baru yang dapat dimanfaatkan di masa mendatang. Hasil pengetahuan tersebut akan digambarkan dalam bentuk pohon keputusan atau *decision tree*.