BAB II
KAJIAN PUSTAKA

2.1 *Knowledge Discovery in Database* (KDD)

Pencarian pengetahuan di dalam *database* atau lebih dikenal dengan istilah *knowledge discovery in database* (KDD) merupakan rumpun ilmu yang memiliki sebuah proses yang dikenal dengan *data mining*. *Data mining* dan KDD umumnya sering disamakan, sebenarnya kedua istilah tersebut memiliki konsep yang berbeda. Proses KDD secara garis besar akan dijelaskan sebagai berikut (Kusrini, 2009):

1. *Data Selection*

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum memulai tahap penggalian informasi dalam KDD. Data hasil penyeleksian akan digunakan pada proses *data mining*, data tersebut nantinya akan di simpan di dalam suatu berkas dan terpisah dari *database* operasional.

2. *Pre-processing Cleaning*

Sebelum melakukan proses *data mining*, perlu dilakukan sebuah proses *cleaning* pada data yang akan menjadi fokus KDD. Proses ini mencangkup seperti membuang data yang redundan, pemeriksaan data yang inkonsisten, dan memperbaiki kesalahan yang ada pada data, seperti kesalahan penulisan (tipografi). Proses ini juga melakukan proses yang bernama *enrichment*, proses *enrichment* yaitu proses

“memperkaya” data yang sudah ada dengan data / informasi lainnya yang relevan dan diperlukan di dalam KDD, seperti data atau informasi eksternal.

3. *Transformation*

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* di dalam KDD merupakan proses yang tergantung pada jenis atau pola informasi yang akan dicari di dalam *database*.

4. *Data Mining*

Data mining adalah proses mencari pola atau informasi menarik dalam data yang telah terpilih dari proses sebelumnya dengan menggunakan sebuah metode tertentu. Teknik, metode, atau algroitma di dalam *data mining* sangat bervariasi seperti contohnya algoritma *naïve bayes classifier*. Pemilihan metode atau algoritma yang tepat bergantung pada tujuan dan proses KDD secara umum.

5. *Interpretation / Evaluation*

Pola informasi yang dihasilkan dari proses *data mining* harus ditampilkan dalam bentuk yang mudah dimengerti oleh *user* yang akan menggunakan hasil proses tersebut. Tahap ini merupakan salah satu bagian dari KDD yang disebut *interpretation*. Tahap ini mencangkup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.2 *Data Mining*

Data mining merupakan salah satu bagian dari Teknologi Informasi dan Komunikasi dan salah satu proses yang ada di dalam *knowledge discovery in database* (KDD). *Data mining* banyak digunakan dalam berbagai hal yang bersifat kolektif data yang besar. *Data mining* adalah sebuah aktivitas yang menggunakan suatu metode statistik, matematik, kecerdasan buatan dan penggunaan bahasa mesin yang digunakan untuk ekstraksi dan identifikasi suatu informasi yang bermanfaat yang berasal dari kumpulan data yang besar (Buaton, Sundari, & Maulita, 2016). *Data mining* menggunakan metode statistik yang digunakan dalam perhitungan secara empiris dari suatu *database* yang ada. Data yang di identifikasi dan di ekstraksi oleh *data mining* bukanlah data yang berbentuk kecil / sedikit, melainkan *data mining* digunakan apabila data tersebut merupakan suatu data yang besar / banyak sehingga apabila tanpa menggunakan *data mining* akan menyulitkan dalam pengolahannya.

Data mining sendiri adalah suatu istilah yang biasa dipakai dalam penguraian dan pencarian penemuan pengetahuan yang ada di *database*. Beberapa faktor yang mendorong kemajuan di dalam bidang *data mining* antara lain (Kusrini, 2009) :

1. Pertumbuhan yang cepat dalam kumpulan data.
2. Penyimpanan data dalam *data warehouse*, sehingga seluruh perusahaan memiliki akses ke dalam *database* yang andal.
3. Adanya peningkatan akses data melalui navigasi web dan intranet.

4. Tekanan kompetensi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
5. Perkembangan teknologi perangkat lunak untuk *data mining* (ketersediaan teknologi).
6. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

Berdasarkan penjelasan-penjelasan yang sudah di ulas sebelumnya, terdapat hal-hal penting yang terkait dengan *data mining*. Hal penting tersebut antara lain:

1. *Data mining* merupakan suatu proses otomatisasi terhadap data yang sudah ada.
2. Data yang akan diproses berupa data yang sangat besar.
3. Tujuan penggunaan *data mining* adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

Dalam poin ketiga dari hal penting yang ada di atas, terdapat kata hubungan / pola. Maksud dari hubungan yang dihasilkan dari *data mining* adalah mencari ketergantungan dari satu objek terhadap yang lain. Contoh dari hubungan yang ada di dalam *data mining* adalah, hubungan pembelian suatu produk terhadap produk lainnya. Selain itu, hubungan juga dapat dilihat antara dua atribut atau lebih dan dua atau lebih objek.

Sementara penemuan pola merupakan keluaran lain dari *data mining*. Misalkan sebuah perusahaan yang akan meningkatkan kartu kredit dari pelanggannya, maka perusahaan tersebut akan mencari sebuah pola dari pelanggan-pelanggan yang ada.

Pola tersebut digunakan untuk mencari dan mengetahui mana pelanggan yang berpontensi dan mana pelanggan yang tidak berpotensi.

Data mining merupakan salah satu dari rangkaian *Knowledge Discovery in Database* (KDD). KDD berhubungan dengan *integration and research, interpretation and visualitation* dari pola sejumlah data. Serangkaian proses tersebut memiliki tahapan antara lain (Vulandari, 2017):

1. Pembersihan data (untuk membuang data yang tidak konsisten dan *noise*).
2. Integrasi data (penggabungan data dari beberapa sumber).
3. Transformasi data (data diubah menjadi bentuk yang sesuai agar dapat di *mining*).
4. Aplikasi teknik *Data Mining*, pada tahap ini data di ekstrasi berdasarkan pola yang sudah ada.
5. Evaluasi pola yang ditemukan (proses interprestasi pola menjadi pengetahuan yang dapat digunakan dalam pengambilan keputusan).
6. Presentasi pengetahuan dengan menggunakan teknik visualisasi.

Beberapa alasan digunakannya *data mining* dalam kegiatannya dapat dilihat dari beberapa sudut pandang. Pada sudut pandang komersial *data mining* dilakukan karena beberapa alasan, diantaranya adalah:

1. Besarnya *volume* data yang di simpan di dalam *data warehouse* seperti *data warehouse* dari *e-commerce*, penjualan yang ada di *supermarket*, transaksi bank.
2. Proses komputasi yang dapat diupayakan dan dioptimalkan.

3. Kerasnya persaingan dalam menyediakan layanan yang lebih baik.

Selain alasan digunakannya *data mining* dalam kegiatannya dilihat dari sudut pandang komersial terdapat alasan dari sudut pandang yang lain. Pada sudut pandang keilmuan alasan digunakannya *data mining* adalah untuk dapat meng-capture, menganalisis serta menyimpan data yang bersifat *real time* dan sangat besar, misalnya:

1. *Remote* sensor yang ditempatkan pada satelit yang mengorbit di luar angkasa.
2. *Telescope* yang digunakan untuk memindai langit oleh para astronom.
3. Simulasi saintifik yang membangkitkan data dalam ukuran sangat besar yaitu ukutan *terabytes*.

2.3 Metode *Data Mining*

Dalam pengolahan sebuah data yang kecil ataupun data yang besar, terdapat sebuah cara atau teknik penyelesaian yang digunakan. Dalam *data mining* cara atau teknik dalam pengolahan data biasa disebut dengan metode. Terdapat beberapa metode yang dapat membantu *data mining* dalam pengolahan datanya. Metode-metode tersebut antara lain:

2.3.1 Metode Klasifikasi *Decision Tree*

Decision Tree merupakan metode yang memanfaatkan ilustrasi dari sebuah pohon. Ilustrasi pohon tersebut digunakan dalam pengambilan keputusan dengan metode klasifikasi. Pohon apabila diartikan dalam pemecahan masalah

pengambilan keputusan adalah penjabaran mengenai berbagai macam alternatif pemecahan masalah yang dapat ditarik dari permasalahan tersebut (Vulandari, 2017). Pohon tersebut juga dapat menampilkan faktor-faktor probabilitas yang dapat mempengaruhi alternatif keputusan tersebut disertai dengan hipotesis dari hasil yang akan di dapat apabila mengambil alternatif keputusan tersebut.

Decision tree merupakan salah satu algoritma dari metode klasifikasi yang paling populer, dikarenakan *decision tree* merupakan algoritma yang mudah diinterprestasi oleh manusia. Konsep dari *decision tree* adalah merubah data yang ada menjadi berbentuk sebuah pohon dengan beserta aturan-aturannya. Manfaat utama dalam penggunaan *decision tree* adalah *decision tree* memiliki kelebihan dalam mem-*break down* proses pengambilan keputusan yang kompleks menjadi pengambilan keputusan yang lebih ringkas dari sebelumnya.

2.3.2 Metode Klasifikasi Teorema *Bayes*

Teori keputusan *Bayes* adalah suatu pendekatan yang bersifat statistika di dalam *data mining* (Vulandari, 2017). Pendekatan ini berdasarkan kuantifikasi *trade-off* antara berbagai macam keputusan klasifikasi yang ada dengan menggunakan probabilitas. *Bayesian classification* adalah pengklasifikasian statistik yang dapat diterapkan dalam memprediksi probabilitas keanggotaan dalam suatu *class* (Kusrini, 2009). Metode ini memiliki kemampuan klasifikasi yang sejenis dengan *decission tree* dan *neural network*. Metode ini sangat cocok

diterapkan apabila penelitian tersebut memiliki *database* dengan data yang besar atau banyak.

Metode *naïve bayes classifier* (NBC) merupakan sebuah penglasifikasian atau pengelompokan dengan sifat probabilistik yang menghitung dari beberapa nilai probabilitas dengan menjumlahkan frekuensi dan kombinasi dari beberapa nilai yang diberikan dari *dataset* yang ada (Saleh, 2015). Algoritma yang dipakai menggunakan teori *Bayesian* dengan asumsi semua atribut bersifat *independen*.

Metode ini ditemukan oleh ilmuwan inggris bernama Thomas Bayes. Metode ini digunakan dalam memprediksi sebuah peluang di masa yang akan datang dengan didasari dari data-data yang berasal dari masa lampau. Metode ini didasari oleh penyederhanaan dimana atribut akan bersifat *independen* apabila atribut tersebut diberikan *output*. Metode *bayes* digunakan dalam melakukan inferensi induksi pada persoalan klasifikasi (Vulandari, 2017). Metode *bayes* menggunakan probabilitas sebagai landasarnnya. Dalam ilmu probabilitas bersyarat dinyatakan sebagai berikut:

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

Rumus 2.1 *Probabilitas Bayes*

Probabilitas X di dalam Y adalah probabilitas interseksi X dan Y terhadap probabilitas Y, atau dengan kata lain P(X|Y) adalah presentase banyaknya variabel X di dalam variabel Y.

Teorema *Bayes* memiliki bentuk umum rumus seperti berikut:

$$P(C|X) = \frac{P(X|c)P(c)}{P(X)}$$

Rumus 2.2 *Teorema Bayes*

Dalam hal ini X merupakan data dari *class* yang belum diketahui sebelumnya. Sedangkan C merupakan hipotesis yang muncul dari variabel X dan merupakan *class* yang spesifik. P(C|X) adalah probabilitas dari hipotesis C berdasarkan kondisi dari variabel X (*posteriori probability*). P(C) adalah probabilitas dari hipotesis C (*prior probability*). P(X|c) merupakan probabilitas dari variabel X terhadap kondisi pada hipotesis C. Sedangkan P(X) merupakan probabilitas variabel X.

2.4 *Software* Pendukung

Software merupakan sebuah sistem yang dapat dijalankan di dalam komputer yang dikerjakan secara otomatis. *Software* yang digunakan di dalam penelitian ini adalah WEKA. WEKA adalah sebuah paket *tools machine learning* yang praktis (Vulandari, 2017). WEKA atau WAIKATO *Environment for Knowledge Analysis* yang dibuat di Universitas Waikato, New Zealand yang diperuntukan sebagai penelitian pendidikan dan berbagai aplikasi. WEKA dianggap mampu menyelesaikan permasalahan *data mining* yang ada di dunia nyata, khususnya klasifikasi yang mendasari pendekatan-pendekatan *machine learning*.

WEKA mudah dalam penggunaannya dan banyak diterapkan pada beberapa tingkatan yang berbeda-beda. Pada WEKA tersedia implementasi dari algoritma pembelajaran *state-of-the-art* yang dapat diimplementasikan ke dalam *dataset* dari *command line*. WEKA juga mengandung *tools* yang dapat digunakan untuk pre-*processing* data, klasifikasi, regresi, *clustering*, aturan *association*, dan visualisasi.

Contoh dalam penggunaan WEKA adalah penerapan sebuah metode pembelajaran yang diimplementasikan ke dalam sebuah *dataset* dan di analisa sehingga *output* dari pemrosesan tersebut digunakan untuk memperoleh informasi tentang data, atau menerapkan beberapa metode dan dibandingkan performanya untuk nantinya akan dipilih.

Pengembangan WEKA sejalan dengan model yang dikeluarkan oleh LINUX, digit kedua yang genap pada nama *software* WEKA menandakan *release* yang stabil dan pada digit yang lainnya menandakan *release* pengembangan. Beberapa versi dari WEKA antara lain:

1. WEKA 3.0 : "versi buku" versi ini sesuai dengan yang di deskripsikan di dalam buku *data mining*.
2. WEKA 3.2 : "versi GUI" versi ini terdapat penambahan GUI (*Graphic User Interface*) dan CLI.
3. WEKA 3.3 : "versi pengembangan" versi ini sudah ditambahkan beberapa peningkatan.

2.5 Penelitian Terdahulu

Dalam melakukan penelitian tentang penerapan *data mining* dalam mengevaluasi nilai akademik mahasiswa, peneliti menggunakan beberapa referensi yang diambil dari jurnal-jurnal terdahulu yang menggunakan teknik *data mining*.

1. Alfa Saleh, (2015). "**Implementasi Metode Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga**". ISSN: 2354-

5771. Penelitian ini menjelaskan permasalahan tidak dapat memprediksi penggunaan listrik rumah tangga. Berbagai alat rumah tangga yang mengandalkan listrik sebagai sumber dayanya seperti lemari es, penanak nasi, kipas angin, *air conditioner*, televisi dan lainnya. Hal ini menjadikan permintaan akan listrik semakin meningkat namun permintaan ini berbanding terbalik dengan ketersediaan pasokan listrik yang semakin menipis. Penelitian ini menerapkan metode *naïve bayes* yang diharapkan mampu memprediksi besarnya penggunaan listrik rumah tangga sehingga dalam mengatur penggunaan listrik dapat menjadi lebih mudah. Pengujian dilakukan kepada 60 pengguna listrik dengan metode *naïve bayes*, didapatkan hasil presentase sebesar 78,3333% untuk keakuratan prediksi, dimana dari 60 data pemakaian listrik rumahan yang telah diuji, sebanyak 47 data pemakai listrik rumahan yang berhasil diklasifikasikan secara tepat.

2. Aline Embun Pramadhani dan Tedy Setiadi, (2014). "**Penerapan Data Mining Untuk Klasifikasi Prediksi Penyakit ISPA (Infeksi Saluran Pernafasan Akut) Dengan Algoritma Decission Tree (ID3)**". e-ISSN: 2338-5197. Dalam penelitian ini mengangkat permasalahan banyaknya data penderita penyakit ISPA di Klinik Dharma Husana yang hanya disimpan saja tidak menjadi representasi pengetahuan dari gejala penyakit ISPA sebelumnya. Sehingga perlu adanya klasifikasi penyakit yang paling banyak di klinik ini. Klasifikasi ini bertujuan membentuk sebuah pohon keputusan dimana pohon ini akan digunakan untuk memprediksi penyakit ISPA dari variabel yang paling

mempengaruhi penyakit ISPA dengan kategorinya. Dalam membentuk sebuah pohon keputusan penelitian ini menggunakan metode *decision tree* untuk pengolahan datanya. Dari pohon keputusan yang terbentuk dari 200 data pasien maka dapat diketahui bahwa jenis kelamin tidak berpengaruh terhadap penyakit ISPA.

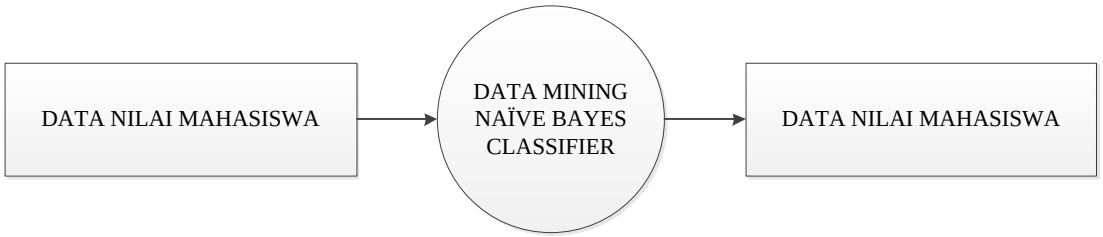
3. Andhika Novandya dan Isni Oktria, (2017). "**Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi**". ISSN: 2089-5615. Mengangkat permasalahan perkiraan cuaca di daerah Bekasi. Data yang digunakan dalam penelitian ini mengacu pada data dari *World Wheater Online*. Dengan menggunakan algoritma C4.5 penelitian ini diharapkan dapat menghasilkan pola klasifikasi cuaca. Hasil pengujian algoritma C4.5 menggunakan 10-*fold cross validation* dan dibuktikan dengan pembuatan aplikasi web untuk pengujian sehingga menghasilkan nilai akurasi sebesar 88.89%.
4. Katrina Shin dan Loganatan Muthu, (2015). "**Application of Big Data In Education Data Mining And Learning Analytics – A Literature Review**". ISSN: 2229-6956. Mengangkat permasalahan aktivitas *online* dari mahasiswa yang menimbulkan sejumlah data yang tidak terpakai saat mahasiswa tersebut *online* akan menjadi tidak berguna. Penelitian ini menggunakan metode *big data* dan beberapa variabel di dalam pendidikan untuk memproses sejumlah data yang tidak terpakai tadi oleh kegiatan *online* mahasiswa. Hasil penelitian ini dimasukan kedalam aplikasi *big data* untuk pendidikan sehingga dapat

menampilkan pembelajaran yang tersedia di dalam pendidikan *data mining* dan analisis.

5. P. Thangaraju, B. Deepa, dan T. Karthikeyan, (2014). "**Comparison of Data Mining Techniques for Forecasting Diabetes Mellitus**". ISSN: 2778-1021. Melakukan penelitian berdasarkan permasalahan penyakit diabetes. Pengambilan permasalahan penyakit diabetes dilakukan karena penyakit diabetes adalah penyakit yang tergolong sebagai penyakit berbahaya dan biasanya terjadi dengan periode yang lama. Penelitian ini menggunakan metode klustering *K-Means* untuk meramal penyakit diabetes. Hasil dari penelitian ini akan terbagi menjadi tiga pengelompokan, yaitu pengelompokan hierarki, pengelompokan kepadatan dan pengelompokan *K-Means*.

2.6 Kerangka Pemikiran

Di dalam melakukan penelitian, peneliti melakukan argumentasi sementara terhadap permasalahan yang diangkat. Kerangka berpikir digunakan untuk menjelaskan secara sementara terhadap suatu gejala yang menjadi objek permasalahan. Kerangka pikir penelitian ini disajikan dalam bentuk diagram sebagai berikut :



Gambar 2.1 Kerangka Pemikiran
Sumber Data Penelitian (2018)

Penjelasan dari kerangka pemikiran yang di sajikan di dalam diagram tersebut akan dijelaskan di sebagai berikut :

1. Tahapan *input* meliputi memasukan variabel penelitian sebagai sumber acuan ke dalam aplikasi WEKA untuk diproses. Variabel yang dipakai adalah nilai akademik mahasiswa.
2. Tahapan proses adalah tahap melakukan olah data menggunakan aplikasi WEKA dengan metode klasifikasi teorema *naïve bayes*.
3. Tahapan *output* adalah tahap dimana keluaran dari hasil olah data dengan menggunakan aplikasi WEKA. Hasil penelitian diharapkan sesuai dengan tujuan dari penelitian ini.