BAB II
KAJIAN PUSTAKA

2.1 *Knowladge Discovery in Database*

Istilah *knowledge discovery in database*(KDD) dan data mining sering digunakan untuk menjabarkan proses pencarian informasi yang tersembunyi dari suatu basis data yang besar. Adapun tahapan proses KDD secara garis besar dapat dijelaskan sebagai berikut (Vulandari, 2017):

1. Data *Selection*

Seleksi data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan ke dalam suatu berkas, terpisah dari basis data operasional.

2. Pre-processing/Cleaning

Pemprosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan *noise* dilakukan. Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).

3. Transformation

Proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. Data mining

Proses *Data mining* yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. Interpretation/evaluation

Penerjemahan pola-pola yang dihasilkan dari data mining. Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah mimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa 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 dipakai dalam berbagai hal yang bersifat kolektif data yang besar. *Data mining* adalah suatu 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* merupakan suatu cara penelusuran data yang ada untuk membangun sebuah model, yang kemudian menggunakan model data tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang disimpan tersebut (Siska Haryati, 2015). *Data mining* menggunakan metode statistik yang digunakan dalam perhitungan secara empiris dari suatu database yang ada. *Data mining* merupakan suatu proses yang menemukan hubungan yang berarti, pola, dan kecendrungan dengan memeriksa dalam sekumpulan data besar yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola (Maharani, Mesran, & Suginam, 2017). Data yang diidentifikasi dan diekstraksi oleh *data mining* bukanlah data yang berbentuk sedikit, melainkan data mining digunakan apabila data tersebut merupakan suatu data yang banyak sehingga apabila tanpa menggunakan data mining akan menyulitkan dalam pengolahannya.

Data mining adalah proses interaktif dan iteratif untuk menemukan model baru yang shahih (sempurna), bermanfaat dan dapat dimengerti dalam suatu *database* yang sangat besar (*massive database)* (Hermawati, 2013).

1. Sahih: sesuatu dapat digereneralisasi untuk masa yang akan datang.
2. Baru: sesuatu yang sedang tidak diketahui.
3. Bermanfaat: digunakan untuk melakukan sebuah tindakan.
4. iteratif: memerlukan beberapa proses yang harus diulang.

5. Interaktif: sesuatu hal yang memerlukan interaksi manusia dalam suatu proses.

Data mining berisi pencarian pola yang diinginkan dalam database yang besar untuk membantu dalam proses pengambil keputusan di waktu yang akan datang. Pola ini dapat dikenali oleh perangkat tertentu yang bisa memberikan suatu analisa data yang berguna, bermanfaat dan berwawasan yang kemudian dapat dipelajari dengan lebih teliti, yang mungkin menggunakan perangkat pendukung keputusan yang lainnya.

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:

1. Proses pertumbuhan yang cepat di dalam kumpulan data.
2. Penyimpanan data di dalam data warehouse, sehingga seluruh perusahaan bisa memiliki akses ke dalam database yang andal.
3. Adanya suatu peningkatan akses data yang melalui navigasi web dan intranet.
4. Tekanan kompetensi bisnisyang digunakan untuk meningkatkan penguasaan pasar di dalam globalisasi ekonomi.
5. Perkembangan sebuah teknologi perangkat lunak untuk data mining.
6. Perkembangan di 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 sebuah proses otomatisasi terhadap data yang 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.

Tahapan proses dalam penggunaan data mining yang merupakan proses *Knowledge Discovery In Database* (KDD) dapat di urakikan sebagai berikut (Hermawati, 2013):

1. Memahami domain aplikasi untuk mengetahui dan menggali pengetahuan awal serta apa sasaran pengguna.
2. Membuat target dat-set yang meliputi pemilihan data dan fokus pada sub-set data.
3. Pembersihan dan transformasi data meliputi eliminasi derau, *outliers, missing value* serta pemilihan fitur dan reduksi dimensi.
4. Pengggunaan algoritma data mining yang terdiri dari asosiasi, sekuensial, klasifikasi, klasterisasi, dll.
5. Interpretasi, evaluasi dan visualisasi pola untuk melihat apakah ada sesuatu yang baru dan menarik dan dilakukan iterasi jika diperlukan.

Secara sistematis, ada tiga langkah utama dalam data mining (Prasetyo, 2012):

1. Eksplorasi / pemprosesan awal data

   Eksplorasi awal data terdiri dari pembersihan data, normalisasi data, trasformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur, dan sebagainya.

2. Membangun model dan melakukan validasi terhadapnya

   Membangun model dan melakukan validasi terhadapnya berarti melakukan analisis berbagai model dan memilih model dengan kinerja prediksi yang terbaik. Dalam langkah ini digunakan metode-metode seperti klasifikasi, regresi, analisis cluster, deteksi anomali, juga masuk dalam langkah eksplorasi. Akan tetapi, deteksi anomaly juga dapat digunakan sebagai algoritma utama, terutama untuk mencari data-data yang special.

3. Penerapan

   Penerapan berarti menerapkan model pada data yang baru untuk menghasilkan perkiraan/perdiksi masalah yang diinvestasikan.

2.3 Metode Data Mining

Terdapat beberapa metode yang digunakan pada *data mining* dalam mengelolah data. Metode-metode tersebut antara lain:

1. Metode *Klasifikasi Decision Tree*

Decision tree atau pohon keputusan adalah pohon yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari masalah yang dimasukkan (Prasetyo, 2014). Pohon keputusan merupakan sebuah struktur pohon, dimana setiap node pohon merepresentasikan atribut yang telah diuji, setiap cabang merupakn suatu pembagian hasil uji, dan node daun merepresentasikan kelompok kelas tersebut (Julianto, Yunitarini, & Sophan, 2014). Pohon keputusan juga berguna untuk mengekplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variable input dengan sebuah variable target. Karena pohon

keputusan memadukan antara ekspolrasi data dan pemodelan, pohon keputusan sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai model akhir dari beberapa teknik lain.

Sebuah pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Dengan masing-masing rangkaian pembagi, anggota himpunan hasil menjadi mirip satu dengan yang lain (Kusrini, 2009). Pohon keputusan banyak digunakan untuk menyelesaikan kasus penentuan keputusan seperti di bidang psikologi (teori pengambilan keputusan), ilmu komputer (struktur data), bidang kedokteran (diagnosis penyakit pasien), dan sebagainya. Banyak algoritma yang bisa digunakan dalam pembentukan pohon keputusan, antara lain ID3, CART, dan C4.5.

2. Metode Algoritma C4.5

Algoritma C4.5 diperkenalkan oleh Quinlan (1996) sebagi versi perbaikan dari ID3. Dalam ID3, induksi decision tree hanya bisa dilakukan pada fitur bertipe katogerikal (nominal atau ordinal), sedangkan tipe numerik (interval atau rasio) tidak dapat digunakan (Prasetyo, 2014). Perbaikan yang membedakan algoritma C4.5 dengan ID3 adalah dapat menangani fitur dengan tipe numeric, melakukan pemotongan (pruning) pohon keputusan, dan penurunan (deriving) rule set. Algoritma C4.5 juga menggunakan kriteria gain dalam menentukan fitur yang menjadi pemecah node pada pohon yang di induksi.

Yang menjadi hal terpenting dalam induksi pohon keputusan adalah bagaimana menyatakan syarat pengujian pada node. Ada 3 kelompok penting dalam syarat pengujian node:

1. Fitur biner

Fitur yang hanya memiliki dua nilai berbeda disebut dengan fitur biner. Syarat pengujian ketika fitur ini menjadi node (akar maupun internal) hanya punya dua pilihan cabang.

2. Fitur bertipe katogerikal

Untuk fitur yang bertipe katogerikal (nominal atau ordinal) bisa mempunyai beberapa nilai berbeda.

3. Fitur bertipe numeric

Untuk fitur bertipe numeric, syarat pengujian dalam node (akar maupun internal) dinyatakan dengan pengujian perbandingan (A<v) atau (A>v) dengan hasil biner, atau untuk multi dengan hasil berupa jangkauan nilai dalam bentuk $v_i \leq A < v_{i+1}$, untuk i = 1, 2, …, k.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

1. Pilih atribut sebagai akar.
2. Buat cabang untuk masing-masing nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan rumus seperti tertera dalam persamaan 1 berikut:

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

Rumus 2.1 *Gain*

Keterangan:

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

Sementara itu, perhitungan nilai entropsi dapat dilihat pada persamaan 2 berikut:

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

Rumus 2.2 *Entropy*

Keterangan:

S : himpunan kasus

A : fitur

n : jumlah partisi S

p_i : property dari S_i terhadap S

Kriteria yang paling bnyak digunakan untuk memilih fitur sebagai pemecah dalam algoritma C4.5 adalah rasio *gain,* yang diformulasikan oleh persamaan berikut:

$$RasioGain\ (s,j) = \frac{Gain\ (s,j)}{SplitInfo\ (s,j)}$$

Rumus 2.3 *RasioGain*

Keterangan:

S : himpunan kasus

J : fitur ke-j

Sedangkan untuk split info dapat diperoleh dari persamaan berikut:

$$SplitInfo\ (s,j) = -\sum_{i=1}^{k} p(Vi \mid s) log_2\ p(Vi \mid s)$$

Rumus 2.4 *SplitInfo*

Keterangan:

K : jumlah pemecahan

2.4 *Software* Pendukung

Software yang digunakan pada penelitian ini yaitu *software* WEKA. Weka merupakan sebuah paket *tools machine learning* yang praktis (Vulandari, 2017). WEKA merupakan kepanjangan dari *Waikato Environment for Knowledge Analysis* yang dibuat di Universitas Waikato, New Zealand yang digunakan sebagai pendidikan, penelitian, dan berbagai aplikasi. WEKA dianggap bisa menyelesaikan masalah-masalah tentang data mining di dunia, khususnya klasifikasi yang mendasari pendekatan *machine learning.*

WEKA mudah dalam penggunaannya dan banyak di terapkan pada beberapa tingakat yang berbeda. Pada WEKA tersedia implementasi dari algoritma-algoritma pembelajaran *state-of-the-art* yang dapat digunakan pada dataset dari *command line*. WEKA juga mengandung *tools* yang digunakan untuk *pre-processing* data, regresi, klasifikasi, *clustering*, visualisasi dan aturan asosiasi. Pengguna bisa melakukan preprocess pada data, memasukkannya ke dalam proses skema pembelajaran, dan menganalisa classifier yang dihasilkan dan performansinya tanpa harus mengetikkan kode programnya sama sekali.

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. WEKA berkembang mengikuti model *releases* Linux. Beberapa versi awal dari WEKA adalah sebagai berikut:

1. WEKA 3.0 : "versi buku" versi ini sesuai dengan deskripsi yang ada di dalam buku data mining.
2. WEKA 3.2 : "versi GUI" versi ini yang menambahkan GUI dari CLI awal.
3. WEKA 3.3 : "versi pengembangan" versi ini bersudah ditambah peningkatan.

2.5 Penelitian Terdahulu

Berikut beberapa penelitian terdahulu yang menjadi acuan penulisan dalam melakukan penelititan sehingga peneliti dapat memperkaya teori yang digunakan dalam mengkaji penelititan yang dilakukan.

1. Chandra Purnmaningsih, Ristu Saptono, dan Abdul Aziz (vol 3. No 1. Juni 2014) ISSN : 2301-7201 yang berjudul **Pemanfaatan Metode K-Means Clustering dalam Penentuan Jurusan Siswa SMA.** Penelitian ini menjelaskan dalam penentuan penjurusan siswa SMA dilakukan berdasarkan nilai akademik yang menjadi ciri dari masing-masing jurusan IPA/IPS, dengan demikian kemungkinan bagi siswa untuk memenuhi kriteria di terima di

jurusan IPA/IPS atau di tolak di jurusan tersebut. Cara untuk mempermudah penentuan jurusan ini dalah dengan cara pengelompokan (*clustering*) data siswa tersebut. Untuk mengelompokkan data siswa tersebut, metode yang digunakan adalah *K-Means Clustering*. Hasil penelitian pengujian terbaik pada praprocessing clustering K-Means IPA dengan hasil akurasi 0.905882, tingkat kesesuaian hasil prediksi dengan data sebenarnya (recall) 1, ketepatan hasil pengujian dalam memprediksi clustering (sensitivity) 0.876923, kesesuaian prediksi negatif terhadap aktual negatif (specificity) 0.714285. Sedangkan pengujian terbaik juga pada praprocessing clustering K-Means IPS didapatkan akurasi 0.905882, recall 0.714285, sensitivity 1, dan specificity 1. Hasil perbandingan clustering terbaik pada praprocessing clustering K-Means IPA dengan praprocessing clustering K-Means IPS menunjukkan bahwa tidak ada siswa yang diterima di dua jurusan IPA/IPS atau siswa ditolak di keduanya.

2. Adhika Novandya dan Isni Oktria (vol. 6 no 2 tahun 2017) ISSN : 2089-5615 dengan judul **Penerapan Algoritma Klasifikasi Data Mining C4.5 Pada Dataset Cuaca Wilayah Bekasi.** Dalam penelitian ini menjelaskan permasalahan dalam perkiraan cuaca wilayah Bekasi dengan penggunaan ini pengetahuan dan teknologi dalam memperkirakannya. Data yang dipakai pada penelitian ini mengacu pada World Weather Online. Penelitian ini menggunakan algoritma C4.5 menggunakan 10-fold cross validation dan dibuktikan dengan pembuatan aplikasi web untuk pengujian sehingga menghasilkan nilai akurasi sebesar 88.89%.

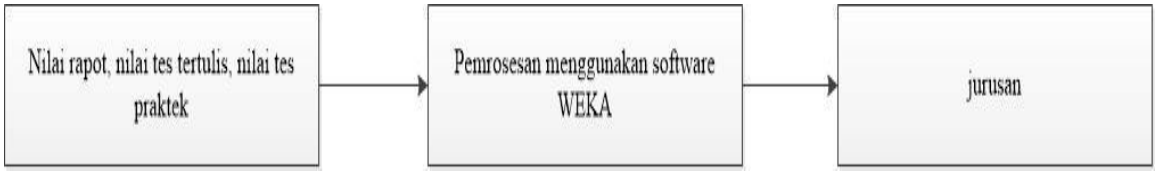
3. Kennedi Tampubolon, Hoga Saragih, dan Bobby Reza (vol 1. No 1. Oktober 2013) ISSN : 2339-210X dengan judul **Impelmentasi Data Mining Algoritma Apriori Pada Sistem Persediaan Alat-Alat Kesehatan.** Penelitian ini menjelaskan Data Mining telah diimplementasikan ke berbagai bidang, diantaranya bidang bisnis atau perdangangan, bidang pendidikan, dan telekomunikasi. Dibidang bisnis misalnya hasil implementasi data mining menggunakan algoritma Apriori dapat membantu para pebisnis dalam kebijakan pengambilan keputusan terhadap apa yang berhubungan dengan persediaan barang. Misalnya pentingnya sistem persediaan barang di suatu Apotek dan jenis barang apa yang menjadi prioritas utama yang harus di stok untuk mengantisipasi kekosongan barang. Karena minimnya stok barang dapat berpengaruh pada pelayanan konsumen dan pendapatan Apotek. Oleh sebab itu ketersediaan berbagai jenis alat-alat kesehatan di Apotek sebagai salah satu upplier alatalat kesehatan, mutlak untuk mendukung kelancaran penyalurannya kepada konsumen, sehingga aktivitas pelayanan konsumen berjalan dengan baik. Dalam penelitian ini membahas tentang implementasi *data mining* menggunakan algoritma apriori.
4. Katrina Sin dan Loganathan Muthu (vol.5 Juli 2015) ISSN: 2229-6956 dengan judul ***Application Of Big Data In Education Data Mining And Learning Analytics – A Literature Review.*** Penelitian ini menjelaskan penggunaan sistem manajemen pembelajaran dalam pendidikan yang telah meningkat saat ini. Saat ini banyak siswa yang menggunakan ponsel, baik untuk kegiatan sehari-hari maupun mengakses konten online. Aktifitas online siswa tersebut

menghasilkan sejumlah data yang besar yang tidak digunakan terkadang terbuang sia-sia karena tidak adanya kemampuan untuk memprosesnya. Penelitian ini mengacu kepada aplikasi data yang besar dalam pendidikan dan menampilkan literatur pendidikan data mining dan analisis pembelajaran.

5. P. Thangaraju, B. Deepa dan T. Karthikeyan (vol.3 Agustus 2014) ISSN: 2278-1021 dengan judul ***Comparison Of Data Mining Techniques For Forecasting Diabetes Mellitus***. Penelitian ini menjelaskan *data mining* sangat memainkan peran penting dalam industri perawantan kesehatan. *Data mining* paling sering digunakan dalam industri perawatan kesehatan untuk proses peramalan penyakit. Penelitian ini melakukan peramalan penyakit diabetes menggunakan teknik *clustering* dan menggunakan *software* WEKA.

2.6 Kerangka Pemikiran

Kerangka berfikir merupakan suatu model konseptual tentang bagai mana teori berhubungan dengan berbagai factor yang telah didefinisikan sebagai masalah dalam penelitian. Berdasarakan teori-teori yang telah di deskripsikan pada bagian sebelumnya, selanjutnya dianalisis secara kritis dan sistematis sehingga menghasilkan sistema atau kesimpulan tentang hubungan antara variable yang diteliti. Kerangka pemikiran penelitian ini di sajikan dalam bentuk diagram di bawah ini:



Gambar 2.1 Kerangka Pemikiran
Sumber: Data Peneliti(2018)

Penjelasan dari kerangka berfikir yang telah disajikan pada diagram diatas, sebagai berikut:

1. Tahap pertama meliputi memasukkan variabel penelitian sebagai sumber acuan ke dalam aplikasi WEKA untuk dilakukan pemrosesan. Variabel tersebut terdiri dari nilai rapot, nilai tes tertulis, dan nilai tes praktek.
2. Tahap selanjutnya yaitu melakukan pengolahan data menggunakan aplikasi WEKA dengan menggunakan algoritma C4.5.
3. Tahap terakhir adalah tahap keluaran dari hasil olah data dengan menggunakan aplikasi WEKA. Hasil penelitian diharapkan sesuai dengan tujuan dari penelitian.