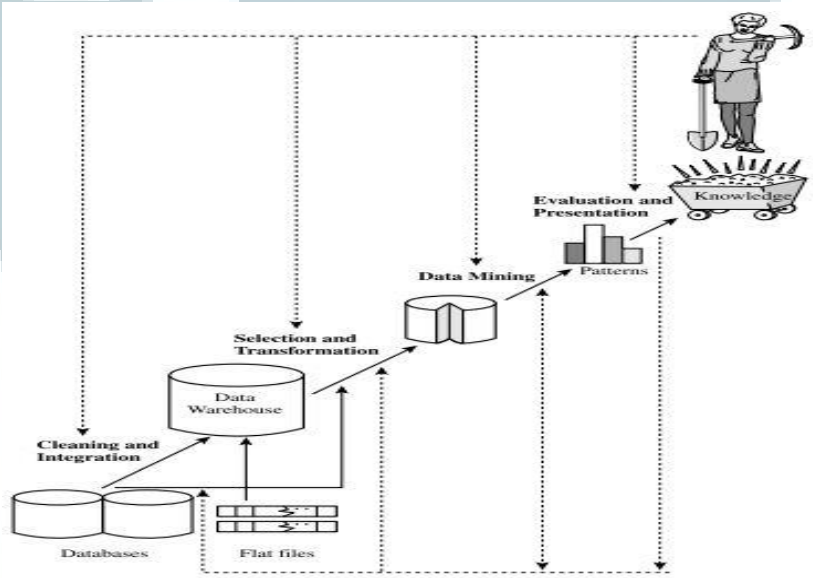
BAB II

LANDASAN TEORI

2.1 Knowledge Discovery of Data

Knowledge Discovery of Data (KDD) dan istilah tersebut cocok dengan tujuan dari data mining adalah pencarian pengetahuan dari sekumpulan dari data. Berikut adalah gambar proses dari *Knowledge Discovery of Data.*



Gambar 2.1 Proses *Knowledge Discovery of Data* (KDD)

Sumber: Han, *Data Mining Concepts and Techniques*, 2006

Proses penemuan pengetahuan yang dapat dilihat pada gambar 2.1, terdiri atas beberapa langkah dan berikut ini adalah penjelasannya:

1 *Data cleaning*

Data cleaning adalah proses mengubah dan menghapus data dalam database yang tidak benar, tidak lengkap, format tidak benar, atau tidak lengkap. Dalam proses cleaning perlu melakukan analisis terhadap data yang mempunyai Outlier. Data outlier adalah data yang tidak sesuai dengan rangkaian suatu data atau model data. Bila data outlier ada akan mengganggu proses analisis karena data akan menjadi bias atau informasi yang dihasilkan jadi tidak akurat.

2 *Data integration*

Integrasi data adalah proses penggabungkan data dari beberapa sumber berbeda dan disimpan ke dalam *data warehouse*. Sumber data yang dimaksud berasal dari beberapa *database, file* dan *segmentasi*. Dalam melakukan penggabungan akan menemukan redundansi data karena memiliki nama yang berbeda dalam database berbeda. Redundansi data dapat terjadi ketika atribut atau sekumpulan atribut tidak konsisten sehingga berpengaruh pada dataset.

3 *Data selection*

Data selection adalah proses penentuan jenis data yang sesuai dengan sumbernya. Dalam pengambilan data dari database adalah data yang ambil telah melalui *cleaning* dan *integration*. Proses pengambilan data yang sesuai dengan kebutuhan penelitian akan berdampak pada integritas data. Tujuan dari data selection adalah menentukan jenis data

yang sesuai, sumber, dan alat yang memungkinkan peneliti untuk menjawab pertanyaan penelitian secara benar.

4 *Data transformation*

Data transformation adalah proses melakukan mengubah data yang sudah dipilih sampai menemukan pola yang sesuai dengan tujuan penelitian yang menggunakan algoritma dan software data mining.

5 *Data Mining*

Pada tahap ini yaitu melakukan pembentukan model data mining yang meliputi pengumpulan, pemakaian data lama untuk menemukan keteraturan, pola dan hubungan dalam set data. Secara umum pada dalam metode data mining ada dua jenis, yaitu metode *predictive* dan metode *descriptive*. Metode *predictive* adalah proses untuk menemukan pola dari data yang menggunakan beberapa *variable* untuk memprediksi *variable* lain yang tidak diketahui atau jenis nilainya. Teknik yang ada didalam *predictive* mining antara lain Klasifikasi, *Regresi*, *Deviasi*.

Metode *descriptive* merupakan proses untuk menemukan suatu karakteristik penting dari data dalam suatu basis data. Teknik yang termasuk pada *descriptive mining* adalah *Clustering*, *Association*, dan *Sequential mining*.

6 *Pattern evaluation*

Pattern evaluation adalah proses analisis pola-pola untuk menemukan pola terbaik yang disesuaikan dengan tujuan penelitian. Hasil analisis ini pada umumnya terdiri dari banyak pola tetapi tidak semua pola

tersebut merupakan pola yang perlukan. Oleh sebab itu, diperlukan suatu teknik pengembangan teknik yang dapat menilai ketertarikan pola yang ditemukan. Proses ini diperlukan untuk memilih pola terbaik untuk menjawab tujuan penelitian.

7 *Knowledge representation*

Tahap terakhir dilakukan untuk menampilkan pengetahuan yang dihasilkan berbagai bentuk *visual*. Tujuannya ditampilkan ke dalam bentuk *visual* agar *user* yang melihat dapat memahaminya

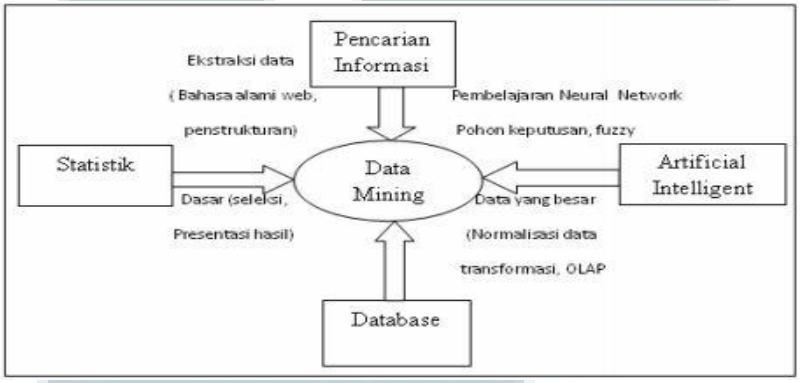
2.2 Data Mining

Data mining didefinisikan sebagai proses yang dilakukan untuk menemukan suatu pola dari sekumpulan data (Witten, Frank, & Hall, 2011). Pola yang telah didapatkan harus bisa memberikan keuntungan bagi yang menggunakan, salah satunya bidang pendidikan. Pengertian lain *data mining* adalah penggalian pengetahuan dari sekumpulan data yang jumlah sangat besar (Han & Kamber, 2006).

Data mining merupakan proses yang biasa digunakan pada teknik statistik, matematika, machine learning, dan kecerdasan buatan yang berfungsi untuk mengektraksi dan mengidenfikasi informasi yang bermanfaat dan pengetahuan yang terkait dari Data Warehouse (Turban, Aronson, & Liang, 2005). Pengetahuan yang diperoleh dari proses ekstraksi dan menghubungkan pola ini diharapkan dapat membantu dalam pengambilan keputusan.

Dari definisi yang telah disampaikan, hal penting yang terkait dengan *data mining* adalah menurut Kusrini & Luthfi (2009):

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



Gambar 2.2 Bidang Ilmu Data Mining

Sumber: (Kusrini & Luthfi, 2009)

2.3 Pengelompokkan Data Mining

Secara garis besar data mining dikelompokkan menjadi 2 kategori utama:

1. *Descriptive data mining*, yaitu menggambarkan data dengan cara yang ringkas dan menampilkan dalam bentuk yang menarik. Teknik data mining yang termasuk dalam *descriptive data mining* adalah *sequential mining*, *clustering*, dan *association*

2. *Predictive data mining*, yaitu menganalisis data dalam rangka untuk membangun satu set model yang digunakan untuk memprediksi perilaku dari set data yang dihasilkan. Teknik yang ada pada *predictive data mining* adalah klasifikasi.

2.4 Clustering

Clustering termasuk metode yang dikenal masih banyak digunakan didalam data mining dan digunakan untuk mencari data yang mempunyai kemiripan satu sama lain. Tujuan dari clustering sendiri adalah mengurangi fungsi yang ditetapkan pada proses *clustering*, yang secara garis besar berusaha mengurangi variasi dalam *cluster* dan memaksimalkan variasi antar *cluster*.

Pengertian lain *clustering* membagi data ke dalam kelompok sehingga data yang memiliki bentuk yang sama dipisahkan ke dalam satu *cluster* yang sama (Refaat, 2007). Dalam hal ini metode *clustering* berusaha menempatkan data yang mirip dalam *cluster* dan membuat jarak setiap *cluster* sejauh mungkin. Ini dapat diartikan data dalam satu *cluster* sangat mirip satu sama lain dan memiliki perbedaan dengan data dalam *cluster* lain.

2.4.1 K-means Clustering

K-means clustering adalah salah satu algoritma data clustering yang popular. algoritma K-Means bekerja dengan membagi data ke dalam k buah cluster yang telah ditentukan (Han & Kamber, 2006). Tujuan dari k-means adalah untuk mengklasifikasikan data ke dalam kelompok k dimana

k adalah parameter input. Terdapat dua tahap, pertama menentukan pusat k secara acak untuk setiap satu cluster. Tahap berikutnya menentukan jarak antara titik data dalam data.

2.5 Data Preprocessing

Data preproccesing adalah teknik data mining yang digunakan untuk mengubah data mentah menjadi format data yang dapat dimengerti dan melalui proses ini dapat meningkatkan kualitas dari data. Jika data yang dipakai untuk proses untuk mencari pengetahuan kualitasnya rendah, maka pengetahuan yang dihasilkan rendah pula (Han & Kamber, 2006). Data nyata yang akan digunakan terkadang tidak lengkap, tidak konsisten, dan banyak mengandung kesalahan dan melalui data preproccesing tahap mempersiapkan data mentah untuk proses lebih lanjut. Ada beberapa teknik dalam tahap data preproccesing, masing – masing teknik memiliki fungsi yang berbeda dalam menangani masalah.

Dibawah ini beberapa teknik yang terdapat pada tahap data preprocessing.

- Data *cleaning* dilakukan untuk menyelesaikan inkonsistensi dalam data dan mengisi nilai yang hilang pada suatu *instance*
- Data *integration* menggabungkan data yang berasal dari beberapa sumber berbeda, seperti data *warehouse.*
- Data *transformation* adalah proses yang dilakukan untuk merubah data, seperti nilai dan tipe data.

- Data *reduction* adalah proses yang dilakukan untuk mereduksi ukuran dari data dengan cara *clustering* dan *aggregating.*
- Data preproccesing, jika dilakukan sebelum tahap penggalian data akan meningkatkan kualitas dari pola atau prediksi dan waktu yang dibutuhkan dalam proses penggalian data (Han, 2006).

2.6 Perbandingan Proses Model Data Mining

Berdasarkan penelitian yang dilakukan yaitu dengan membandingkan 3 proses model data mining yang paling terkenal dan model proses ini *Knowledge Discovery Databases* (KDD) proses model, CRISP-DM, dan SEMMA (Shafique & Qasier, 2014). Ketiga proses model ini paling sering digunakan oleh para ahli dan peneliti untuk membantu pekerjaan yang mereka lakukan terutama berhubungan dengan data mining. Berikut ini adalah tabel yang berhasil dibuat atas perbandingan yang sudah dilakukan oleh Shafique dan Qasier.

Tabel 2.1 Summary KDD, CRISP-DM, dan SEMMA Proses

Sumber: (Shafique & Qaiser, 2014)

Data Mining Process Models	KDD	CRISP-DM	SEMMA
No. of Steps	9	6	5
Name of Steps	Developing and Understanding of the Application	Business Understanding	----------
	Creating a Target Data Set	Data Understanding	Sample
	Data Cleaning and Pre-processing		Explore
	Data Transformation	Data Preparation	Modify
	Choosing the suitable Data Mining Task	Modeling	Model
	Choosing the suitable Data Mining Algorithm		
	Employing Data Mining Algorithm		
	Interpreting Mined Patterns	Evaluation	Assessment
	Using Discovered Knowledge	Deployment	----------

Hasil penelitian yang dilakukan Shafique & Qaiser menyimpulkan bahwa para peneliti dan para ahli data mining lebih memilih mengikuti atau menggunakan knowledge Discovery Database (KDD) proses model lebih lengkap dan lebih akurat. Sebaliknya CRISP-DM dan SEMMA yang lebih berorientasi pada perusahaan, tetapi CRISP-DM masih lebih baik daripada SEMMA. Akan tetapi semua model proses ini tujuannya adalah untuk panduan dan membantu orang-orang dan para ahli untuk mengetahui bahwa bagaimana mereka dapat menerapkan data mining dalam skenario praktis (Shafique & Qaiser, 2014).

2.7 Perbandingan Aplikasi yang dipakai pada Data Mining

Aplikasi yang digunakan dalam membantu pekerjaan pada data mining dapat dikatakan bervariasi dan memiliki kelebihan dan kekurangan masing – masing. Pada bagian ini dibahas tiga tools data mining yang dipakai oleh Moghimipour & Ebrahimpour sebagai pembanding dengan menggunakan *method decision tree* dari proses penelitian yang mereka lakukan terhadap tiga tools itu yaitu SPSS-Clementine, RapidMiner,and Weka.

Menurut Moghimipour & Ebrahimpour (2014) kesimpulannya, program data mining, SPSS-Clementine, RapidMiner dan Weka dijelaskan perbedaan karena bahasa pemrograman yang berbeda ditekankan. Untuk mengidentifikasi mana yang lebih baik tergantung pada latar belakang Anda dan kebutuhan Anda

2.8 Data Masukan

Data masukan adalah data *training set* yang bersumber dari berkas ataupun *database* yang dapat digunakan. Aplikasi yang digunakan dapat mendukung format berkas sebagai masukan, yaitu ARFF, CSV, UCI, dan XLS. Data masukan harus memiliki relasi, atribut, dan *instance* agar sesuai dengan prosedur. Berikut ini adalah beberapa data masukan berdasarkan tipe berkas yang berbeda.

2.8.1 ARFF

Berkas ARFF (Attribute – Relation Berkas Format) adalah berkas text ASCII yang menggambarkan daftar instance dan atribut. Berkas ARFF dikembangkan oleh proyek machine learning di Departemen Ilmu Komputer dari University of Waikato untuk digunakan oleh perangkat lunak pembelajaran mesin Weka.

Berkas ARFF memiliki dua bagian yang berbeda. Bagian pertama adalah informasi mengenai nama relasi dan daftar atribut beserta nilainya, sedangkan bagian kedua adalah daftar data. Tiap baris informasi diawali dengan karakter '@', diikuti dengan nilai dari informasi tersebut.

```
% 1. Title: Iris Plants Database
%
% 2. Sources:
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
%
@RELATION iris

@ATTRIBUTE sepallength  NUMERIC
@ATTRIBUTE sepalwidth   NUMERIC
@ATTRIBUTE petallength  NUMERIC
@ATTRIBUTE petalwidth   NUMERIC
@ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

The **Data** of the ARFF file looks like the following:

```
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
```

Gambar 2.3 Format dalam Arff

Sumber: (http://www.cs.waikato.ac.nz/ml/weka/arff.html)

2.8.2 CSV

CSV (Comma Separated Value) adalah berkas teks ASCII yang penulisan nilanya dipisahkan oleh karakter koma. Nama dari berkas csv digunakan untuk mengisi informasi relasi. Pada berkas ini, terdapat n jumlah baris dimana baris pertama adalag daftar niali atribut, sedangkan baris kedua sampai baris terakhir adalah instance.

2.8.3 UCI

Format UCI adalah format standar yang digunakan oleh website penyedia data untuk keperluan proses data mining, yaotu website UCI dataset. UCI terdiri atas dua berkas, yaitu berkas .data dan berkas .names. Berkas .data ini hanya berisi daftar semua instance yang dipisahkan oleh karakter koma sedangkan informasi mengenai nama atribut dan nilainya ada di berkas lain dengan format .names sehingga berkas ini daoat digunakan sebagai masukan apabila berkas .names telah tersedia.

2.8.4 XLS

XLS adalah berkas yang dihasilkan oleh Microsoft Excel. Struktur dari berkas ini hampir sama dengan CSV. Terdapat sejumlah n baris dimana baris pertama berisi daftar semua atribut, sedangkan baris kedua sampai terakhir berisi instance. Lalu untuk membaca semua baris dari berkas XLS diperlukan *library* tambahan yang dikembangkan oleh Lars Vogel.

2.9 Pendidikan

Pada dasarnya pendidikan menurut UU SISDIKNAS No.20 tahun 2003.

"Pendidikan adalah usaha sadar dan terencana untuk mewujudkan suasana belajar dan proses pembelajaran agar peserta didik secara aktif mengembangkan potensi dirinya untuk memiliki kekuatan spiritual keagamaan, pengendalian diri, kepribadian, kecerdasan, akhlak mulia, serta keterampilan yang diperlukan dirinya dan masyarakat".

Dalam melaksanakan pendidikan ini tentu memliki tujuan sesuai undang-undang yang telah diatur pada UUD 1945 tentang pendidikan dituangkan dalam UU No. 20 tahun 2003 pasal 3 menyebutkan,

"Pendidikan nasional berfungsi mengembangkan kemampuan dan membentuk watak serta peradaban bangsa yang bermartabat dalam rangka mencerdaskan kehidupan bangsa, bertujuan untuk berkembangnya potensi peserta didik agar menjadi manusia yang beriman dan bertakwa kepada Tuhan Yang Maha Esa, berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, dan menjadi warga negara yang demokratis serta bertanggung jawab."

2.10 Bahasa Inggris

Bahasa Inggris merupakan bahasa yang digunakan sebagai bahasa komunikasi antara negara mulai dari untuk hal diplomasi sampai dengan wisawatan yang berwisata ditempat lain. Oleh karena ini, negara – negara

yang bahasa utamanya bukan bahasa Inggris mulai menerapkannya sebagai bahasa kedua yang wajib seperti di negara Indonesia.

Karakteristik dari bahasa Inggris dibagi menjadi empat keterampilan yaitu, mendengarkan, berbicara, membaca, dan menulis. Keempat keterampilan itu digunakan untuk memahami dan menghasilkan teks lisan atau tulisan yang dapat diartikan kemampuan wacana sehingga saat kemampuan tersebut telah dikembangkan dapat menanggapi dan menciptakan wacana dalam kehidupan bermasyarakat.

2.11 Manfaat Bahasa Inggris

Menurut English Domain - sebuah institusi non-formal yang mengajarkan bahasa inggris di Eropa, terdapat beberapa alasan untuk memperlajari bahasa inggris:

a. Berkomunikasi dengan orang lain. Bahasa inggris merupakan bahasa yang paling umum dan sering digunakan di seluruh dunia untuk berkomunikasi antar bangsa dan negara, sehingga memudahkan setiap warga negara untuk berpergian dan berinteraksi ke negara asing.
b. Mendorong karir jauh ke depan. Dapat berbicara bahasa inggris berarti terbukanya kesempatan untuk berkomunikasi dengan klien dan rekan asing sehingga meningkatkan nilai tambah di dalam dunia profesional.
c. Mendapatkan akses pengetahuan. Istilah dalam berbagai bidang pengetahuan dan akademis menggunakan bahasa inggris sebagai bahasa pengantar sehingga membantu dalam pengumpulan informasi dan ilmu pengetahuan.

d. Melihat dunia dengan perspektif berbeda. Bahasa inggris memungkinkan setiap orang untuk berkesempatan merasakan berbagai perbedaan budaya dari seluruh dunia, dan juga bahasa yang digunakan dalam berbagai macam media informasi dan hiburan.

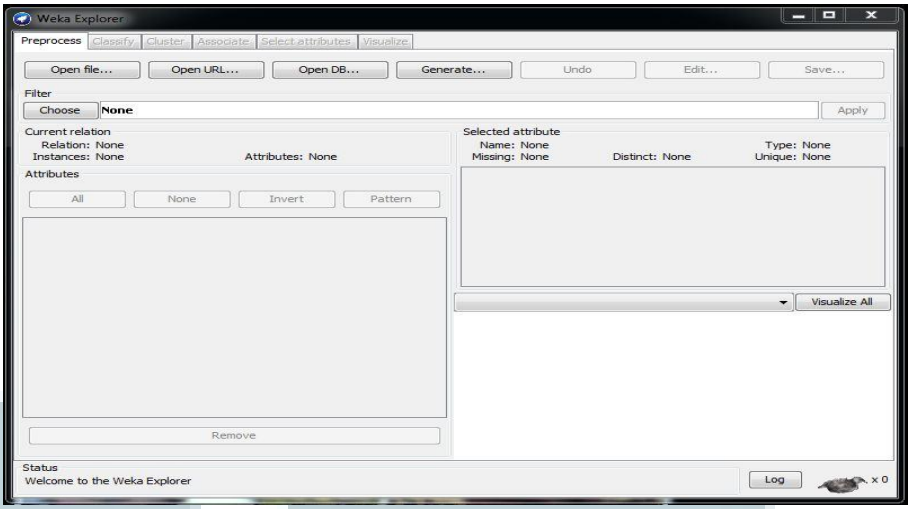
2.12 TOEFL

Menurut Cyssco (2013: 1), *Test of English as a Foreign Language* (TOEFL) adalah tes yang diselenggarakan oleh *Educational Testing Service* (ETS) di Amerika Serikat yang bertujuan untuk mengetahui tingkat penguasaan bahasa inggris secara lisan (*spoken English*) maupun tulisan (*written English*) bagi orang menggunakan bahasa inggris sebagai bahasa kedua untuk menempuh perkuliahan berbasis bahasa inggris atau di perguruan tinggi negara berbahasa inggris.

2.13 WEKA

Weka adalah sebuah perangkat lunak yang dibuat dengan menggunakan bahasa pemrograman Java. Algoritma *machine learning* yang terdapat pada Weka sangat banyak. Weka sendiri mempunyai kemampuan untuk menggambarkan hasil analisa data ke dalam bentuk *bar*, *tree*, *chart*, dan *scatterplot*.

Weka dapat menerima input dari berbagai sumber, seperti berkas XLS, CSV, ARFF, berbentuk *database*, ataupun dari URL langsung. Setelah itu berdasarkan sumber diatas, weka juga menyediaka fitur untuk melaksanakan *preprocessing* data.



Gambar 2.4 Aplikasi WEKA