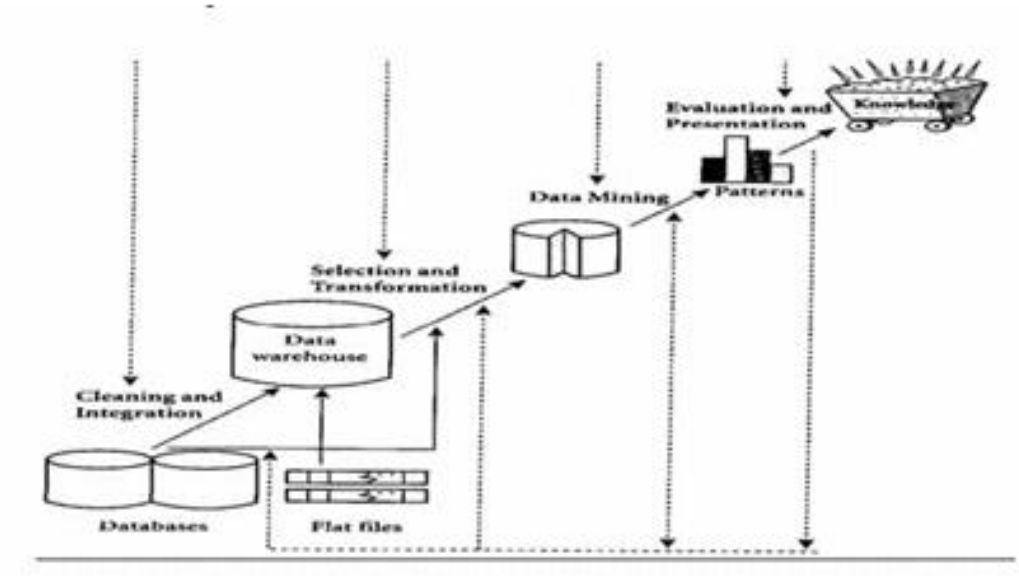
BAB 2

TINJAUAN PUSTAKA

2.1 Data Mining

Data mining merupakan suatu metode menemukan suatu pengetahuan dalam suatu database yang cukup besar. *Data mining* adalah proses menggali dan menganalisa sejumlah data yang sangat besar untuk memperoleh sesuatu yang benar, baru, sangat bermanfaat dan akhirnya dapat dimengerti suatu corak atau pola dalam data tersebut. Data mining adalah bagian integral dari penemuan pengetahuan dalam *database* (KDD), yang merupakan proses keseluruhan mengubah data mentah menjadi informasi yang bermanfaat, seperti yang ditunjukkan pada Gambar 2.1.



Gambar 2.1 *Data Mining* sebagai dari proses *knowledge discovery.*

Gambar 2.1 menunjukkan proses penjelajahan pengetahuan di mulai dari beberapa *database* dilakukan proses *cleaning* dan *integration* sehingga menghasilkan *data warehaouse.* Dilakukan proses *selection* dan *transformation* yang kemudian disebut sebagai *data mining* sehingga menemukan pola dan memperoleh pengetahuan dari data *(knowledge).*

Terdapat teknik data mining yang sering disebut dalam literatur. Namun ada 3 teknik *data mining* yang populer (Siska Haryati, Aji Sudarsono and Program 2015), yaitu:

1) *Association Rule Mining*

   Association Relu Mining adalah teknik mining untuk menemukan asosiatif antara kombinasi atribut.Contoh dari aturan asosiatif dari analisa pembelian di suatu pasar swalayan dapat mengatur penempatan barangnya atau merancang strategi pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu.

2) *Cluster*

   Clustering juga berbeda dengan *association rule mining* dan *klasifikasi* dimana kelas data telah ditentukan sebelumnya, *clustering* bisa di pakai untuk memberikan label pada kelas data yang belum diketahui. Karena itu *clustering* sering digolongkan sebagai metode *unsupervised learning.* Prinsip *clustering* adalah memaksimalkan kesamaan antara *cluster. Clustering* dapat dilakukan pada data yang memiliki beberapa atribut yang dipetakan sebagai ruang multidimensi.

3) *Klasifikasi*

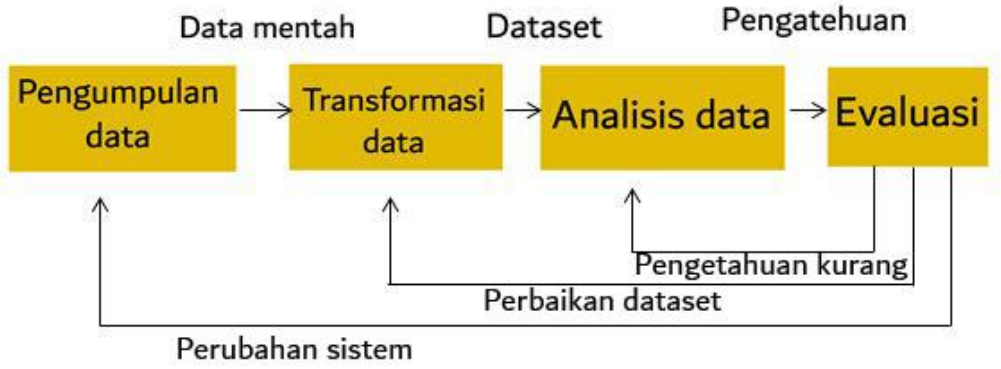
   Klasifikasi terdapat target variabel kategori. Sebagai contoh, pengolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, pendapatan rendah.

2.2 Data mining pada pendidikan *(Education Data Mining)*

Education Data Mining (EDM) adalah bidang ilmu baru yang mengeksploitasi statistik, *machine learning*, dan algoritma *data mining* (DM) pada berbagai jenis data pendidikan yang tujuan utamanya yaitu menganalisis jenis data dalam menyelesaikan masalah-masalah penelitian pendidikan. Pengembangan metode yang berkaitan dalam pengaturan pendidikan untuk mengeksplorasi jenis data yang unik. (M. Mahaputra Hidayat, Diana Purwitasari 2013)

Diagram yang menggambarkan aliran informasi dalam proses data mining (M. Mahaputra Hidayat, Diana Purwitasari 2013). Proses *data mining* pada gambar tersebut ditunjukkan sebagai proses yang iteratif. Hasil evaluasi

pengetahuan yang dihasilkan data mining dapat menimbulkan kebutuan pengetahuan yang lebih lengkap, perbaikan kumpulan data (dataset) atau perusahan pada sistem.



Gambar 2.2 Aliran Informasi dalam *Data Mining*

Banyak penelitian terkait dengan *educational* data mining seperti (Mayadewi and Rosely 2015) prediksi nilai proyek akhir mahasiswa menggunakan algoritma klasifikasi *data mining* dengan menggunakan motede algoritma ID3, CHAID serta Naïve Bayes. Hasil dari metode ini nilai akurasi nya adalah 62,66%. Yang kedua Algoritma klasifikasi *data mining* untuk memprediksi siswa dalam memperopeh bantuan dana pendidikan menggunakan metode Algoritma Klasifikasi *Data Mining* . Hasil yang di peroleh nilai *Accuracy Algoritma* C4.5 adalah sebesar 98,80% (HENDRIAN 2018).

2.3 Rough Set

Rough set adalah sebuah teknik matematika yang dikembangkan oleh *Pawlack* pada tahun 1980. Teknik ini digunakan untuk menangani masalah *Uncertainty*.(*Missing data, Inclompleted Data* dan *Inclonsistency Data Imprecision* dan *Vagueness*) dalam aplikasi *Artifical Intelligence*(AI).

Rough Set merupakan teknik yang efisien untuk *Knowledge Discovery* dalam *Database* (KDD) proses dan *Data Mining*. Secara umum, teori *Rough Set* telah digunakan dalam banyak aplikasi seperti *medicine, pharmacology, business, banking, engineering design, image processing* dan *decision analysis*. *Rough Set*

merupakan teknik yang efisien untuk KDD proses dan *Data Mining* (Prajana 2016)

2.3.1 Tabel Informasi dan Tabel Informasi Keputusan

Berdasarkan sudut pandang teori *rough set*, sistem informasi terdiri dari 4 tuple dari persamaan *S = (U ,A,V ,f)*, dimana $U = \{u_1, u_2, \ldots, u_{|u|}\}$ merupakan kumpulan objek yang terbatas dan tidak kosong, $A=\{a_1, a_2, \ldots, a_{|A|}\}$ adalah kumpulan atribut terbatas dan tidak kosong, $V=U_{a \in A} V_a$, V_a adalah domain (set nilai) dari atribut *a, f:* $U \times A \rightarrow V$ adalah fungsi informasi yang sedemikian rupa sehingga $f(u,a) \in V_a$, $(u,a) \in U \times A$, fungsi untuk informasi (pengetahuan)(Sutoyo 2018).

Jika *U* berisi setidaknya satu objek dengan nilai yang tidak diketahui atau hilang, maka *S* disebut sistem informasi tidak lengkap. Nilai yang tidak diketahui atau hilang dinotasikan dengan simbol "*" didalam sistem informasi yang tidak lengkap. Dalam tulisan ini, digunakan 4 tuple $S^*=(U,A,V_*, f)$ digunakan untuk menunjukkan sistem informasi yang tidak lengkap. Setelah mempresentasikan gagasan sistem informasi di atas(Sutoyo 2018).

2.3.2 *Indiscernible Relation*

Hubungan yang tidak dapat dibedakan (*Indiscernible relation)* disebabkan oleh himpunan atribut *B*, sehingga dapat dilambangkan dengan *IND(B)*, yang merupakan relasi yang ekivalen. Hubungan ekivalensi dapat menyebabkan partisi yang berbeda antar partisi lainnya. Partisi *U* yang disebabkan oleh *IND(B)* di *S = (U,A,V,f)* dapat dilambangkan dengan *U|B* dan ditunjukkan dengan $[x]_B$. Misalnya, *B* adalah himpunan bagian dari *A* didalam *S* dan *X* menjadi bagian dari *U*, sehingga *B-lower approximation* dari *X* dapat dilambangakan dengan *Ḅ(X)*, dan *B-upper approximation* dari *X* dapat dilambangkan *Ḃ(x)*(Sutoyo 2018).

2.3.3 *Set Approximation*

Akurasi untuk aproksimasi dari setiap himpunan bagian dari $X \subseteq U$ terhadap $B \subseteq A$, dapat dilambangkan dengan $a_B(X)$ dan dapat diukur menggunakan formula $a_B(X)=|Ḅ(X)|/|Ḃ(X)|$,dimana $|X|$ melambangkan kardinalitas dari *X* . Setiap himpunan yang kosong ω , dapat didefinisikan menjadi $a_B(\omega) = 1$ [20]. Sehingga, $0 \leq a_B(X) \leq 1$. Jika *X* adalah gabungan dari beberapa

kelas yang ekivalen dari *U,* maka $a_B(X) = 1$. Himpunan *X* adalah himpunan yang *crisp,* dan jika *X* bukan merupakan gabungan dari beberapa kelas yang ekivalen dari *U,* maka $a_B(X) < 1$. Himpunan *X* adalah *rough* (tidak jelas atau samar) terhadap *B* (Sutoyo 2018).

2.3.4 *Discernible Matrix*

Definisi *Discerniblity Matrix:* Diberikan sebuah *IS A=(U,A)* and *B U A, discernibility matrix* dari A adalah MB, dimana tiap-tiap *entryMB(I,j)* terdiri dari sekumpulan *attribute* yang berbeda antara objek X_i dan X_j(Sembiring and Azhar 2013).

2.3.5 *Discernibilty Matrix Modulo D*

Didefinisikan seperti berikut dimana MB(I,j) adalah sekumpulan *attribute* yang berbeda antara objek $X_i$dan X_j dan juga berbeda *attribute* keputusan. Diberikan sebuah DS *A=(U,A{d{)* dan subset dari *attributeB A, Discrenibility Matrix ModuloD dari A,MBd*(Sembiring and Azhar 2013).

2.3.6 *Reduct*

Reduct adalah penyeleksian *attribute* minimal *(interesting attribute)* dari sekumpulan *attribute* kondisi dengan menggunakan *Prime Implicant* fungsi Boolean. Kumpulan dari semua *Prime Implicant* mendeterminasikan *sets of reduct*(Sembiring and Azhar 2013).

2.3.7 *Generating Rules*

Proses selanjutnya yaitu mendapatkan pengetahuan dalam *database* melalui ekstraksi aturan dari sistem keputusan. Hasil keputusan tersebut didasarkan pada proses *reduct*(Sembiring and Azhar 2013).

2.4 Pengukuran Kualitas Aturan

Pengukuran kualitas aturan (*quality measurement for rules*) ialah pengukuran yang dilakukan terhadap item rule yang dihasilkan (Soelaiman, Anggraeni, and Setiawan 2008). Terdapat beberapa instrumen pengukuran yakni: yang meliputi *support, strength, accuracy* dan *coverage*.

1) *Support.*

   Suport dari *decision rules* ialah jumlah objek dari *decision rule* yang memiliki *antecedent* (f) *and conclusion* (g) yang sesuai.

$$Support(f \rightarrow g) = card(|f \cap g|)$$

2) *Accuracy.*

   Accuracy dari *decision rules* merupakan rasio perbandingan dari objek yang memenuhi *antecedent* dan juga memenuhi *conclusion* terhadap objek yang memenuhi *antecedent* saja.

$$accuracy(f \rightarrow g) = \frac{Support(f \rightarrow g)}{|f|}$$

3) *Coverage*

   Coverage merupakan rasio perbandingan dari objek yang memenuhi *antecendent* dan juga memenuhi *conclusion* terhadap objek yang memenuhi *conclusion* saja.

$$coverage(f \rightarrow g) = \frac{Support(f \rightarrow g)}{|g|}$$

2.5 Validasi

Tabel 2.1 *Confusion Matrix*

TT	LT
TL	LL

TT = Tepat di prediksi Tepat

LT = Telat di prediksi Tepat

TL = Tepat di prediksi Telat

LL= Telat di Prediksi Telat

- Tepat di prediksi Tepat (TT) : kita memprediksi mahasiswa Tepat dan memang benar mahasiswa tersebut Tepat
- Telat di prediksi Tepat (LT) : kita memprediksi mahasiswa Telat dan ternyata prediksi salah, ternyata mahasiswa tersebut Tepat