

BAB II

LANDASAN TEORI

2.1 Data Mining

2.1.1 Definisi Data Mining

Data Mining adalah langkah analisis terhadap proses penemuan pengetahuan didalam basisdata atau *knowledge discovery in databases* yang disingkat KDD. Pengetahuan bisa berupa pola data atau relasi antar data yang valid (yang tidak diketahui sebelumnya). *Data Mining* merupakan gabungan sejumlah disiplin ilmu komputer yang didefinisikan sebagai proses penemuan pola-pola baru dari kumpulan-kumpulan data sangat besar, meliputi metode-metode yang merupakan irisan dari AI (*artificial intelligence*), *machine learning*, *statistics*, dan *database systems* (Suyanto, 2017).

Data Mining ditujukan untuk mengekstrak (mengambil intisari) pengetahuan dari sekumpulan data sehingga didapatkan struktur yang dapat dimengerti manusia serta meliputi basisdata dan manajemen data, pemrosesan data, pertimbangan model dan inferensi, ukuran ketertarikan, pertimbangan kompleksitas, pasca pemrosesan terhadap struktur yang ditemukan, visualisasi, dan *online updating* (Suyanto, 2017).

2.1.2 Pengelompokan Data Mining

Dalam (Kusrini and Luthfi, 2009) *Data Mining* dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu:

a. Deskripsi

Terkadang peneliti dan analisis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Deskripsi dari pola kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

b. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Model dibangun

menggunakan baris data (record) lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

c. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang. Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

d. Klasifikasi

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

e. Pengklasteran (*Clustering*)

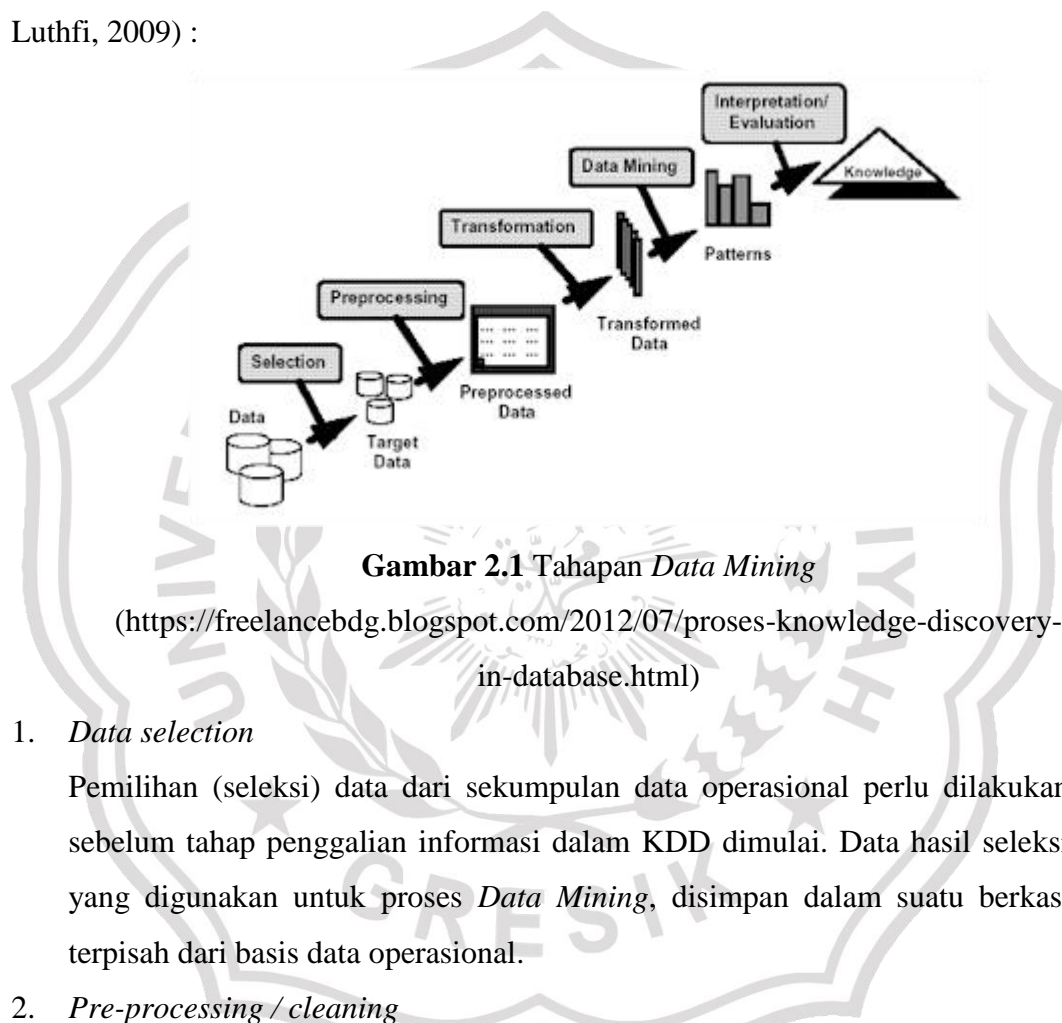
Pengklasteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas obyek-obyek yang memiliki kemiripan. Klaster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan *record* dalam klaster yang lain. Berbeda dengan klasifikasi, pada pengklasteran tidak ada variabel target. Pengklasteran tidak melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target, akan tetapi, algoritma pengklasteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (homogen), yang mana kemiripan *record* dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal.

f. Asosiasi

Tugas asosiasi dalam *Data Mining* adalah untuk menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja (*Market Basket Analysis*).

2.2 Tahapan-tahapan *Data Mining*

Tahapan yang dilakukan pada proses *Data Mining* diawali dari seleksi data dari data sumber ke data target, tahap *preprocessing* untuk memperbaiki kualitas data, transformasi, *Data Mining* serta tahap interpretasi dan evaluasi yang menghasilkan output berupa pengetahuan baru yang diharapkan memberikan kontribusi yang lebih baik. Secara *detail* dijelaskan sebagai berikut (Kusrini and Luthfi, 2009) :



Gambar 2.1 Tahapan *Data Mining*

(<https://freelancebdg.blogspot.com/2012/07/proses-knowledge-discovery-in-database.html>)

1. *Data selection*

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang digunakan untuk proses *Data Mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing / cleaning*

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.

3. *Transformation*

Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *Data Mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. *Data Mining*

Data Mining adalah 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*

Pola informasi yang dihasilkan dari proses *Data Mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya.

2.3 Binerisasi dan Diskretisasi

Analisis asosiasi membutuhkan data dengan atribut biner yang asimetris karena dalam analisis asosiasi hanya ada atribut dengan nilai 1 yang dianggap penting (Andi, 2012).

2.4 Association rule

2.4.1 Pengertian Association rule

Association rule adalah suatu prosedur yang mencari hubungan atau relasi antara satu *item* dengan *item* lainnya. *Association rule* biasanya menggunakan “*if*” dan “*then*” misalnya “*if A then B and C*”, hal ini menunjukkan jika A maka B dan C. Dalam menentukan *association rule* perlu ditentukan *support* dan *confidence* untuk membatasi apakah *rule* tersebut *interesting* atau tidak (Morgan Kauffman, 2016).

Association rule berguna untuk menemukan hubungan penting antar *item* dalam setiap transaksi, hubungan tersebut dapat menandakan kuat tidaknya suatu aturan dalam asosiasi, Tujuan *association rule* adalah untuk menemukan keteraturan dalam data. *Association rule* dapat digunakan untuk mengidentifikasi *item-item* produk yang mungkin dibeli secara bersamaan dengan produk lain, atau dilihat secara bersamaan saat mencari informasi mengenai produk tertentu. Dalam pencarian *association rule*, diperlukan suatu variabel ukuran kepercayaan (*interestingness measure*) yang dapat ditentukan oleh *user*, untuk mengatur batasan sejauh mana dan sebanyak apa hasil *output* yang diinginkan oleh *user*.

2.4.2 Ukuran Kepercayaan *Rule* (*Interestingness Measure*)

Menurut (I. Anshor, 2016) terdapat dua ukuran kepercayaan yang menunjukkan kepastian dan tingkat kegunaan *rule* yang ditemukan yaitu:

1. *Support*

Support (dukungan) merupakan suatu ukuran yang menunjukkan seberapa besar dominasi suatu *item* atau *itemset* dari keseluruhan transaksi.

2. *Confidence*

Confidence (tingkat kepercayaan) adalah suatu ukuran yang menunjukkan hubungan antar *item* secara *conditional* (misalnya seberapa sering *item B* dibutuhkan jika perusahaan membutuhkan *item A*).

Untuk menemukan aturan asosiasi seperti yang diharapkan maka harus menemukan nilai dari *support* yang telah ditentukan. *Support* tersebut merupakan jumlah *item* pada setiap transaksi yang ada didalam *database*. Untuk dapat menemukan nilai *support* kita dapat mencari semua aturan yang jumlah *support* \geq *minimum support*. Dalam hal ini dapat digunakan sebagai cara untuk menemukan sebuah nilai *confidence*. Nilai *confidence* ditentukan dari nilai *support* suatu aturan dalam sebuah transaksi.

Jika *itemset* pada setiap transaksi tidak sering muncul (*infrequent*), maka kandidat yang tidak sesuai dengan nilai *support* \geq *minimum support*

tersebut harus segera dipangkas tanpa harus menghitung *confidencenya*. Strategi umum digunakan oleh banyak algoritma penggalian aturan asosiasi adalah memecahkan masalah ke dalam dua pekerjaan utama, yaitu:

1. *Requent Itemset Generation*

Tujuannya adalah mencari semua *itemset* yang memenuhi ambang batas *minimum support*. *Itemset* itu disebut *itemset frequent* (*Itemset* yang sering muncul)

2. *Rules Generation*

Tujuannya adalah mengekstrak aturan dengan *confidence* tinggi dari *itemset frequent* yang ditemukan dalam langkah sebelumnya. Aturan ini kemudian disebut aturan yang kuat (*Strong rules*) (Andi, 2012).

2.5 Algoritma Apriori

Pada bagian ini akan dijelaskan tentang algoritma Apriori sebagai metode yang digunakan dalam tugas akhir ini, yang meliputi definisi, langkah-langkah dan contoh kasus dengan menggunakan algoritma Apriori.

2.5.1 Pengertian Algoritma Apriori

Apriori adalah suatu algoritma yang sudah sangat dikenal dalam melakukan pencarian *frequent itemset* dengan menggunakan teknik *association rule*. Algoritma Apriori menggunakan *knowledge* mengenai *frequent itemset* yang telah diketahui sebelumnya, untuk memproses informasi selanjutnya. Pada algoritma Apriori untuk menentukan kandidat-kandidat yang mungkin muncul dengan cara memperhatikan *minimum support* (Andi, 2009).

Algoritma apriori termasuk jenis aturan asosiasi pada *Data Mining*. Selain algoritma apriori, yang termasuk pada golongan ini adalah metode *Generalized Rule Induction* dan Algoritma *Hash Based*. Aturan yang menyatakan asosiasi antara beberapa atribut sering disebut *affinity analysis* atau *market basket analysis*. Analisis asosiasi atau *association rule mining* adalah teknik *Data Mining* untuk menemukan aturan asosiatif antara suatu

kombinasi *item*. Metodologi dasar analisis asosiasi terbagi menjadi dua tahap yaitu :

1. Analisis pola frekuensi tinggi

Tahap ini mencari kombinasi *item* yang memenuhi syarat *minimum* dari nilai *support* dalam *database*. Nilai *support* sebuah *item* diperoleh dengan rumus berikut : (Andi, 2009).

$$\text{Support}(A) = \frac{\text{Jumlah transaksi mengandung } A}{\text{Total transaksi}} \dots (2.1)$$

Gambar 2.2 Rumus Nilai *Support* 1 *item*

Nilai *support* dari 2 *item* diperoleh dengan menggunakan rumus:

$$\text{Support}(A, B) = \frac{\sum \text{Transaksi mengandung } A \text{ dan } B}{\sum \text{transaksi}} \dots (2.2)$$

Gambar 2.3 Rumus Nilai *Support* 2 *item*

Frequent itemset menunjukkan *itemset* yang memiliki frekuensi kemunculan lebih dari nilai *minimum* yang ditentukan (ϕ). Misalkan $\phi = 2$, maka semua *itemsets* yang frekuensi kemunculannya lebih dari atau sama dengan 2 kali disebut *frequent*. Himpunan dari *frequent k-itemset* dilambangkan dengan F_k .

2. Pembentukan Aturan Asosiasi

Setelah semua pola pola frekuensi tinggi ditemukan, barulah dicari aturan asosiasi yang memenuhi syarat *minimum* untuk *confidence* dengan menghitung *confidence* aturan asosiatif $A \rightarrow B$. Nilai *Confidence* dari aturan $A \rightarrow B$ diperoleh rumus berikut.

$$\text{Confidence} = P(B|A) = \frac{\sum \text{Transaksi mengandung } A \text{ dan } B}{\sum \text{Transaksi mengandung } A} \quad (2.3)$$

Gambar 2.4 Rumus Nilai *Confidence*

Untuk menentukan aturan asosiasi yang akan dipilih maka harus diurutkan berdasarkan $\text{Support} \times \text{Confidence}$. Aturan diambil sebanyak n -aturan yang memiliki hasil terbesar.

2.5.2 Proses Utama Algoritma Apriori

Proses Utama Algoritma Apriori untuk meningkatkan efisiensi dari pencarian *k-itemset*, dapat digunakan suatu metode tambahan yang dinamakan Apriori *Property*. Metode ini dapat mengurangi lingkup pencarian sehingga waktu pencarian dapat dipersingkat.

Menurut (Morgan Kauffman, 2016) terdapat dua proses utama yang dilakukan dalam algoritma Apriori, yaitu:

1. *Join* (Penggabungan).

Pada proses ini setiap *item* dikombinasikan dengan *item* yang lainnya sampai tidak terbentuk kombinasi lagi. Untuk menemukan L_k , suatu set dari kandidat *k-itemset* dihasilkan dengan cara men-*join*kan L_{k-1} dengan dirinya sendiri. Set kandidat hasil *join* ini nanti akan dinotasikan sebagai C_k . Adapun aturan dari *join* ini adalah setiap kandidat yang dihasilkan tidak boleh mengandung kandidat yang kembar antara satu dengan yang lainnya.

2. *Prune* (Pemangkasan).

Pada proses ini, hasil dari *item* yang telah dikombinasikan tadi lalu dipangkas dengan menggunakan *minimum support* yang telah ditentukan oleh *user*. Semua $(k-1)$ -*itemset* yang tidak *frequent* tidak mungkin dapat menjadi subset dari *frequent k-itemset*. Oleh karena itu, jika ada $(k-1)$ subset dari kandidat *k-itemset* yang tidak termasuk dalam L_{k-1} , maka kandidat tidak mungkin *frequent* juga dan oleh karena itu dapat dihapus dari C_k .

2.5.3 Langkah-langkah dari proses Algoritma Apriori

Langkah-langkah algoritma Apriori untuk mendapatkan *rules* yang diinginkan oleh *user*, antara lain:

1. Melakukan *scan database* untuk mendapat kandidat 1-*itemset*, yaitu C_1 (Himpunan *item* yang terdiri dari 1 *item*) dan menghitung nilai *support*-nya. Bandingkan nilai *support* dengan *minimum support* yang sudah ditentukan, jika nilainya lebih besar atau sama dengan *minimum support*, maka *itemset*

tersebut termasuk dalam *large itemset* yaitu L1 (*Large itemset* dengan 1 *itemset*).

2. *Itemset* yang tidak termasuk dalam *large itemset* tidak disertakan dalam iterasi selanjutnya (dilakukan *pruning*).
3. Himpunan L1 hasil iterasi pertama akan digunakan untuk iterasi selanjutnya. Pada L1 dilakukan proses *join* terhadap dirinya sendiri untuk membentuk kandidat 2 *itemset* (C2). Bandingkan lagi *support* dari *item-item* C2 dengan *minimum support*, bila tidak kurang dari *minimum support*, maka *itemset* tersebut masuk dalam *large itemset* L2. Pada iterasi selanjutnya, hasil *large itemset* pada iterasi sebelumnya (Lk-1) akan dilakukan proses *join* terhadap dirinya sendiri untuk membentuk kandidat baru (Ck), dan *large itemset* baru (Lk). Setelahnya dilakukan proses *pruning* pada *itemset* yang tidak termasuk dalam Lk-.
4. Dari seluruh *large itemset* yang memenuhi *minimum support* (*frequent itemset*) dibentuk *association rule* dan nilai *confidencenya*. Aturan-aturan yang nilai *confidencenya* lebih kecil dari *minimum confidence*, tidak termasuk dalam *association rule* yang dipakai.

2.6 Korelasi Lift

Proses *mining* Apriori ditandai dengan terbentuknya kekuatan hubungan kombinasi *itemset* dengan alat ukur asosiasi final. Namun untuk mengukur *valid* atau tidaknya asosiasi final tersebut maka dapat menggunakan *lift ratio* (K.E.Dewi, 2014). *Lift ratio* adalah alat ukur penting dalam aturan asosiasi. Fungsinya adalah mengukur ketepatan dan kecermatan suatu alat ukur (*support dan confidence*) agar dapat dipercaya sepenuhnya. Dalam penelitian ini *lift ratio* memastikan bahwa apakah penggunaan media A digunakan secara bersamaan dengan media B. Rumus perhitungan *lift ratio* dapat dirujuk pada peneliti (I.Anshor, 2016). Pada akhirnya sebuah kombinasi *itemset* dinyatakan *valid* dan kuat jika nilai *lift ratio* > 1 .

Nilai korelasi dapat diketahui dengan menggunakan rumus persamaan sebagai berikut:

$$Lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)} \quad (2.4)$$

$Lift(A, B)$ = Korelasi antara A dan B

$P(A \cup B)$ = Jumlah kemunculan antara A dan B dibagi dengan total transaksi

$P(A) P(B)$ = Jumlah kemunculan A dikali jumlah kemunculan B pada total transaksi.

Apabila dari perhitungan tersebut menghasilkan nilai dibawah 1 maka terdapat korelasi negatif. Untuk perhitungan yang menghasilkan nilai diatas 1 maka terdapat korelasi positif. Namun apabila menghasilkan nilai sama dengan 1 maka tidak ada korelasi antara X dan Y.

2.7 Penelitian Sebelumnya

Beberapa riset yang telah dilakukan berkaitan dengan kasus asosiasi yang menggunakan metode apriori antara lain:

Penelitian yang berjudul “*Perancangan Aplikasi Data Mining Untuk Menganalisa Produk Kosmetik dan Kecantikan Menggunakan Algoritma Apriori Pada PT. Rajawali Mutiara Sejahtera*” Oleh Zulkifli, Bayu Dedy Irwanto Penelitian ini bertujuan untuk menemukan hubungan khusus antar produk yang dibeli bersamaan. Berdasarkan hubungan tersebut, pemilik usaha dapat mengetahui barang yang dimungkinkan untuk dibeli secara bersamaan oleh konsumen. Sangat dimungkinkan pemilik Usaha melakukan promosi barang dengan pola keterikatan barang tersebut. Konsumen yang membeli barang akan tertarik untuk membeli barang yang lain yang biasa dibelinya. Bila konsumen tidak membeli barang yang ada dalam pola penjualan barang, Perusahaan dapat menawarkan barang yang ada dalam pola penjualan barang. Salah satu algoritma penemuan kombinasi pola barang adalah algoritma apriori. Penggunaan metode asosiasi dalam pencarian pola keterikatan untuk promosi produk, diharapkan dapat meminimalkan promosi barang yang mempunyai tingkat penjualan rendah. Dengan meminimalkan promosi barang yang tidak terbeli, konsumen tidak akan terganggu dengan

promosi barang yang tidak mempunyai pola keterikatan, sehingga promosi akan lebih efektif.

Penelitian selanjutnya yaitu “*Implementasi Data Mining Asosiasi Untuk menentukan Rekomendasi Paket Pembelian Produk kebutuhan syar’i Menggunakan Algoritma Apriori (Studi kasus:Toko Tiara Scarf)*” Oleh M. Habibi Abdillah Almasri. Pakaian syar’i merupakan kebutuhan sehari-hari bagi kaum wanita khususnya Khimar (kerudung), Gamis, Jubah, Mukenah, dan Kebutuhan lainnya. Banyaknya macam-macam produk mengharuskan pemilik toko untuk mengolah penjualan dengan baik sehingga dapat mempertahankan penjualan kepada konsumen. Salah satunya adalah dengan cara membuat paket penjualan, Proses penentuan paket penjualan dapat diselesaikan dengan *Association rule* (aturan asosiasi) menggunakan Algoritma apriori untuk menentukan pasangan antar produk syar’i. Data transaksi penjualan akan dilakukan *preprocessing* agar diketahui nilai dari masing-masing produk yang nantinya akan diproses menggunakan algoritma apriori.

Penelitian selanjutnya oleh: Ana Nur Hayati “*Penerapan Asosiasi Rule Untuk Menentukan Pola penjualan Pada percetakan CV. Karya Duta Dengan Algoritma Apriori*” Dalam persaingan dunia bisnis sekarang ini menurut para pelakunya untuk senantiasa mengembangkan bisnis mereka dan juga agar selalu bertahan dalam persaingan.Salah satu nya yaitu dengan mengetahui pola penjualan produk yang sering di pesan oleh konsumen. Sehingga pemilik dapat mementukan strategi bisnis yang akan dilakukan. Algoritma apriori termasuk jenis aturan asosiasi pada *Data Mining*. Aturan yang menyatakan asosiasi atau *association rule* mining adalah teknik *Data Mining* untuk menemukan aturan suatu kombinasi *item*. Salah satu tahap analisis asosiasi pola frekuensi tinggi (*frequent pattern mining*). Penting tidaknya suatu asosiasi dapat diketahui dengan dua tolak ukur, yaitu *support* dan *confidence*. Nilai penunjang (*support*) adalah persentase kombinasi *item* tersebut dalam *database* sedangkan nilai kepastian (*confidence*) adalah kuatnya hubungan antaritem dalam aturan asosiasi.