

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Penelitian yang dilakukan oleh Dinda Setiawati Devi dengan menggunakan metode Apriori untuk analisa keranjang pasar untuk 100 data transaksi dan 55 jenis barang menghasilkan 8 asosiasi data dengan parameter *min-confidence* >- 10% dan *min-support* >- 5% [7].

Penelitian yang dilakukan oleh S.Moertini Veronica dan Yulita Marsela dengan algoritma *hash-based* pada data transaksi penjualan obat di apotik menyimpulkan kinerja *hash-based* ternyata bergantung kepada bilangan prima yang digunakan pada fungsi *hashing*. Semakin besar bilangan ini (hingga di atas dan paling dekat dengan jumlah item), semakin cepat waktu komputasinya [19].

Penelitian yang dilakukan oleh Afif Syaifullah Muhammad, melakukan implementasi data mining algoritma Apriori pada system penjualan menyimpulkan teknik data mining dengan algoritma Apriori dapat diimplementasikan pada system penjualan [1].

Penelitian yang dilakukan Erwin, melakukan analisis *market basket* dengan algoritma Apriori dan FP-Growth menyimpulkan bahwa algoritma Apriori membutuhkan waktu komputasi yang lama untuk mendapatkan *frequent itemsets*, karena berulang kali melakukan pemindahan data dan membutuhkan memori yang besar sedangkan FP-Growth hanya membutuhkan dua kali *scanning database* dalam mencari *frequent itemsets* sehingga waktu yang dibutuhkan menjadi relatif singkat dan efisien [8].

Penelitian yang dilakukan Febriana Santi Wahyuni , mencari kaidah asosiasi antar kombinasi suatu item barang dengan algoritma K-Means dan Fuzzy C-Means menyimpulkan bahwa data yang di-*cluster* akan menghasilkan kaidah yang mempunyai rata – rata nilai *confidence* yang lebih tinggi, yaitu 85% pada nilai minimum *support* 20 dan 40, dibandingkan dengan data yang tidak melalui proses *clustering*. Data yang melalui proses *clustering* dengan K-Means menghasilkan rata

– rata nilai *confidence* yang lebih tinggi dibandingkan dengan Fuzzy C-Means (83%) pada nilai minimum *support* 20 dan 40 [20].

Penelitian yang dilakukan oleh Andreas, Gregorius Satia Budhi, Hendra Rusly , meneliti asosiasi pembelian barang di *supermarket* dengan metode *market basket analysis* menyimpulkan [11] :

- Menemukan *frequent itemset* dan *association rule* yang memenuhi syarat *minimum support*, berdasarkan item yang ada dalam transaksi.
- Dapat mengetahui asosiasi barang apa saja yang sering dibeli bersamaan oleh konsumen.
- Semakin kecil *minimum support* dan *confidence* yang ditentukan maka semakin banyak *rules* yang dapat dihasilkan.
- Pada proses pembuatan 2 – *itemset candidate* membutuhkan waktu yang paling lama, semakin banyak jumlah 1-*itemset candidate* yang memenuhi *minimum support*, maka semakin banyak pula jumlah 2-*itemset* yang harus dihasilkan.

Dalam aturan asosiasi pertambahan tinjauan terbaru dari Sotiris Kotsiantis, Dimitris Kanellopoulos , menyimpulkan algoritma konvensional aturan asosiasi penemuan berlangsung dalam dua langkah. Semua *itemset* sering ditemukan pada langkah pertama. *The Frequent itemset* adalah *itemset* yang disertakan dalam setidaknya transaksi *minimum support*. Asosiasi aturan dengan kepercayaan pada *minimum confidence* setidaknya dihasilkan pada langkah kedua [14].

Penelitian Christian Borgelt , menyimpulkan pada tahap *preprocessing* dilakukan penghapusan *item* transaksi yang tidak sering muncul sehingga waktu lebih efisien. Implementasi algoritma FP-pertumbuhan melebihi algoritma Apriori dan algoritma Eclat [4].

Penelitian Christian Borgelt, Thorsten Meinl dan Michael Berthold , menyimpulkan algoritma didasarkan pada pencarian pertama kedalaman dalam pohon substruktur, yang membuat pendekatan algoritma ini efisien adalah pemangkasan struktural yang mencoba meminimalkan jumlah pencarian [5].

Penelitian Mohammad El-Hajj dan Osmar R.Zaiane , menyimpulkan pendekatan lompatan *traversal* secara signifikan mengurangi jumlah calon *itemsets* dan cocok sebagai kerangka kerja yang baik untuk pertambangan dan pemrosesan parallel, serta menunjukkan penghematan drastis dalam penggunaan memori [9].

Penelitian Xun Zhu, Hongtao Deng, Zheng Chen , menyimpulkan aturan asosiasi pertambangan pola yang sering muncul mengalami perkembangan yang cukup pesat yang telah diterapkan untuk berbagai aplikasi domain, seperti pengindeksan dan pencarian kesamaan kompleks data terstruktur, pertambangan multimedia, dan pertambangan web [2].

Tabel 2.1 . Rangkuman penelitian terkait analisis data transaksi penjualan barang

No	Nama Peneliti dan Tahun	Masalah	Metode	Hasil
1	Dinda Setiawati Devi, 2009	Penggalian asosiasi data transaksi penjualan.	Algoritma Asosiasi Apriori, dengan 100 data transaksi penjualan.	Ada 8 Asosiasi, dengan <i>min-support</i> 5% dan <i>min-confidence</i> 10%
2	S.Moertini Veronica dan Yulita Marsela, 2004	Penggalian asosiasi data transaksi penjualan.	<i>Algoritma Hash-Based</i>	Bilangan prima yang paling efektif untuk digunakan dalam fungsi <i>hashing</i> adalah yang terdekat dan lebih besar dari jumlah seluruh produk.
3	Afif Syaifullah Muhammad. 2009	Penggalian asosiasi data transaksi penjualan.	Algoritma Apriori	Teknik <i>data mining</i> dengan algoritma Apriori dapat diimplementasikan pada system penjualan.

Tabel 2.1 Rangkuman penelitian terkait analisis data transaksi penjualan barang (lanjutan)

No	Nama Peneliti dan Tahun	Masalah	Metode	Hasil
4	Erwin,2009	Penggalian asosiasi data transaksi penjualan.	Algoritma Apriori dan FP-Growth	algoritma Apriori membutuhkan waktu komputasi yang lama untuk mendapatkan <i>frequent itemsets</i> , karena berulang kali melakukan pemindahan data dan membutuhkan memori yang besar sedangkan FP-Growth hanya membutuhkan dua kali <i>scanning database</i> dalam mencari <i>frequent itemsets</i> sehingga waktu yang dibutuhkan menjadi lebih singkat dan efisien.
5	Santi Wahyuni Febriana, 2010	Penggalian asosiasi data transaksi penjualan.	Algoritma K-Means dan Fuzzy C Means	Data yang di- <i>cluster</i> mempunyai rata – rata nilai <i>confidence</i> lebih tinggi yaitu 85% pada nilai <i>min-support</i> 20 dan 40, dari pada data yang tidak melalui proses <i>clustering</i> . Data yang melalui proses <i>clustering</i> dengan K-Means menghasilkan rata – rata nilai <i>confidence</i> yang lebih tinggi dibandingkan dengan Fuzzy C- Means (83%) pada nilai minimum <i>support</i> 20 dan 40.

Tabel 2.1 . Rangkuman penelitian terkait analisis data transaksi penjualan barang (lanjutan)

No	Nama Peneliti dan Tahun	Masalah	Metode	Hasil
6	Andreas Handojo, Gregorius Satia Budhi, Hendra Rusli, 2004.	Menggali asosiasi item barang	Metode <i>Market Basket Analysis</i>	<ul style="list-style-type: none"> - Menemukan <i>frequent itemset</i> dan <i>association rule</i> yang memenuhi syarat <i>minimum support</i>, berdasarkan item yang ada dalam transaksi. - Dapat mengetahui asosiasi barang apa saja yang sering dibeli bersamaan oleh konsumen. - Semakin kecil <i>minimum support</i> dan <i>confidence</i> yang ditentukan maka semakin banyak <i>rules</i> yang dapat dihasilkan. - Pada proses pembuatan 2 – <i>itemset candidate</i> membutuhkan waktu yang paling lama, semakin banyak jumlah 1-<i>itemset candidate</i> yang memenuhi <i>minimum support</i>, maka semakin banyak pula jumlah 2-<i>itemset</i> yang harus dihasilkan.

Tabel 2.1 . Rangkuman penelitian terkait analisis data transaksi penjualan barang (lanjutan)

No	Nama Peneliti dan Tahun	Masalah	Metode	Hasil
7	Sotiris Kotsiantis, Dimitris Kanellopoulos, 2006	Aturan Asosiasi Pertambahan	Algoritma konvensional aturan asosiasi.	Algoritma konvensional aturan asosiasi penemuan berlangsung dalam dua langkah. Semua <i>itemset</i> sering ditemukan pada langkah pertama. <i>The Frequent itemset</i> adalah <i>itemset</i> yang disertakan dalam setidaknya transaksi <i>minimum support</i> . Asosiasi aturan dengan kepercayaan pada <i>minimum confidence</i> setidaknya dihasilkan pada langkah kedua.
8	Christian Borgelt, 2005	Penambahan <i>itemset</i> yang sering muncul dari <i>database</i> transaksi.	Algoritma Fp-Growth, algoritma Apriori dan algoritma Eclat.	Pada tahap <i>preprocessing</i> dilakukan penghapusan <i>item</i> transaksi yang tidak sering muncul sehingga waktu lebih efisien. Implementasi algoritma FP-Growth melebihi algoritma Apriori dan algoritma Eclat.
9	Christian Borgelt, Thorsten Meini, Michael Berthold, 2005	Pertambahan sub struktur molekul yang paling sering muncul dalam satu set molekul.	<i>Search Tree Traversal</i>	Algoritma didasarkan pada pencarian pertama kedalaman dalam pohon substruktur, yang membuat pendekatan ini efisien adalah pemangkasan struktural yang mencoba meminimalkan jumlah pencarian.

Tabel 2.1 . Rangkuman penelitian terkait analisis data transaksi penjualan barang (lanjutan)

No	Nama Peneliti dan Tahun	Masalah	Metode	Hasil
10	El-Hajj Mohammad dan Osmar R.Zaiane, 2005	Menemukan maksimal pola dengan memperkenalkan metode baru menghitung dukungan kandidat berdasarkan dukungan dari pola kandidat lainnya.	<i>Leap Traversal</i> dengan dua pendekatan, pertama mengurangi pola pohon yang tidak sering muncul , yang kedua partisi ruang transaksi menggunakan Cofi-pohon.	Pendekatan lompatan traversal secara signifikan mengurangi jumlah calon <i>itemssets</i> dan cocok sebagai kerangka kerja yang baik untuk pertambangan dan pemrosesan parallel, serta menunjukkan penghematan drastis dalam penggunaan memori.
11	XunZhu, Hongtao Deng , Zheng Chen .2011	Asosiasi aturan pertambangan pola yang sering muncul.	Algoritma Apriori, Fp-Growth, Eclat.	Aturan asosiasi pertambangan pola yang sering muncul mengalami perkembangan yang cukup pesat yang telah diterapkan untuk berbagai aplikasi domain, seperti pengindeksan dan pencarian kesamaan kompleks data terstruktur, pertambangan multimedia, dan pertambangan web

2.2. Landasan Teori

2.2.1. Konsep Dasar Aturan Asosiasi Algoritma.

Asosiasi aturan pertambangan untuk mencari korelasi antara *item* dalam *dataset* telah banyak mendapat perhatian terutama sejak publikasi AIS dan algoritma Apriori [2][3] yang awal penelitian sebagian besar didorong oleh analisis data keranjang pasar, hasil yang memungkinkan perusahaan lebih memahami perilaku pembelian sehingga berdampak pada pencapaian target pasar. Analisis asosiasi pertambangan

adalah proses dua bagian. Pertama identifikasi set *item* atau *itemsets* dalam *dataset*. Kedua derivasi selanjutnya dari kesimpulan dari *itemsets* tersebut. Sebagian besar penelitian terkait sampai saat ini difokuskan pada penemuan efisiensi *itemsets* sebagai tingkat kompleksitas jauh lebih besar dibandingkan generasi *inferensi*. [6]

Selama satu dekade terakhir berbagai algoritma telah dikembangkan melalui penyempurnaan strategi pencarian, teknik pemangkasan, struktur data dan penggunaan organisasi *dataset* alternatif. Sementara sebagian besar algoritma fokus pada penemuan eksplisit dari semua kesimpulan untuk *dataset* tertentu, pertimbangan peningkatan sedang diberikan kepada algoritma khusus yang berusaha untuk meningkatkan waktu pemrosesan. Untuk mengurangi waktu pengolahan bisa dilakukan dengan teknik pemangkasan dinamis mengurangi *dataset* selama pengolahan, membuang *item* yang tidak perlu. [6]

Kendala yang paling umum digunakan untuk mengurangi eksplorasi selama identifikasi *itemsets* adalah spesifikasi dari *minsup* ambang batas dukungan, dimana dukungan berkaitan dengan berapa kali *itemsets* akan muncul. Ini secara efektif mengurangi identifikasi *itemsets* dengan penemuan hanya *itemsets* yang melebihi *minsup* kemunculan dianggap sebagai *itemsets* yang valid. [6]

Meningkatkan Apriori [6] :

- Mengurangi melewati *scan database* transaksi.
- Kecilkan jumlah calon / *itemsets*.
- Memfasilitasi dukungan menghitung calon.
- Menghapus transaksi yang jarang muncul dari *k-itemset* dari *database* untuk *scan* lebih lanjut.

Ada dua langkah dasar yang penting untuk aturan asosiasi, dukungan (*s*) dan kepercayaan (*c*). Karena *database* besar dan kekhawatiran pengguna tentang barang yang sering dibeli, biasanya ambang dukungan dan kepercayaan yang telah ditetapkan oleh pengguna untuk menetapkan aturan – aturan yang menarik atau berguna. Dukungan (*s*) dari suatu aturan asosiasi didefinisikan sebagai prosentasi yang berisi *X U Y* terhadap total jumlah *record* dalam *database*. Misalkan dukungan

dari *item* adalah 0.1%, itu berarti hanya 0,1 persen dari transaksi mengandung pembelian *item* ini [14].

Keyakinan atau kepercayaan adanya aturan asosiasi didefinisikan sebagai prosentasi ber transaksi yang mengandung $X \cup Y$ dengan jumlah total transaksi yang mengandung X . Keyakinan adalah ukuran kekuatan aturan asosiasi, misalnya keyakinan dari aturan asosiasi $X \longrightarrow Y$ adalah 80%, itu berarti bahwa 80% dari transaksi yang mengandung X juga mengandung Y bersama-sama [14].

Umumnya sebuah asosiasi aturan algoritma pertambangan berisi langkah-langkah sebagai berikut [14] :

- Himpunan kandidat *k-itemset* yang dihasilkan oleh 1-ekstensi besar (k-1) – *itemsets* dihasilkan dalam iterasi sebelumnya.
- Mendukung untuk calon *k-itemsets* yang dihasilkan dari *database*.
- *Itemsets* yang tidak memiliki dukungan minimal dibuang dan sisanya disebut *k-itemsets*.

Algoritma konvensional aturan asosiasi menemukan dua langkah pertambangan asosiasi, langkah pertama menemukan semua *itemset*. *The frequent itemset* adalah *itemset* yang disertakan dalam setidaknya transaksi *minsupport*. Asosiasi aturan dengan kepercayaan pada *minconfidence* setidaknya dihasilkan pada langkah kedua [14].

2.2.2. Teknik *Clustering*

Teknik *cluster* termasuk teknik yang sudah cukup dikenal dan banyak dipakai dalam *data mining*. Usaha – usaha untuk menghitung jumlah *cluster* yang optimal dan peng-*cluster*-an yang paling baik masih terus dilakukan. Dengan demikian menggunakan metode yang sekarang, kita tidak bisa menjamin hasil peng-*cluster*-an sudah merupakan hasil yang optimal. Namun hasil yang dicapai biasanya sudah cukup bagus dari segi praktis. [22]

Tujuan utama dari metode *cluster* adalah pengelompokkan sejumlah data/obyek ke dalam *cluster* (group) sehingga dalam setiap *cluster* akan berisi data yang semirip mungkin. Dalam *clustering* kita berusaha untuk menempatkan obyek yang mirip (

jaraknya dekat) dalam satu *cluster* dan membuat jarak antar *cluster* sejauh mungkin. Ini berarti obyek dalam satu *cluster* sangat mirip satu sama lain dan berbeda dengan obyek dalam *cluster – cluster* yang lain.[22]

2.2.3. Algoritma Klasifikasi K-Means.

Algoritma klasifikasi K-Means merupakan metode *clustering* berbasis jarak yang mempartisi data ke sejumlah kelompok dan bekerja pada atribut numerik. Algoritma ini dimulai dengan pemilihan jumlah kelompok (k) secara acak serta pengambilan sebagian populasi sejumlah (k) untuk dijadikan sebagai titik pusat awal. Salah satu metode perhitungan jarak yang bisa digunakan adalah *Euclidean Distance*. Perhitungan jarak menggunakan metode *Euclidean* dinyatakan sebagai berikut :

$$d (x,y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \dots\dots\dots(1)$$

Dimana :

x ; obyek ke – 1

y ; obyek ke – 2

n ; banyaknya atribut obyek ke – 1, obyek ke – 2.

Prosedur dasar *clustering* K-Means adalah sebagai berikut :

1. Menentukan k sebagai jumlah *cluster* yang ingin dibentuk.
2. Membangkitkan k *centroid* (titik pusat *cluster*) awal secara random.
3. Menghitung jarak setiap data ke masing – masing *centroids*.
4. Setiap data memilih *centroid* yang terdekat.
5. Menentukan posisi *centroid* baru dengan cara menghitung nilai rata – rata dari data – data yang berada pada *centroid* yang sama.
6. Kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama.

Ada dua pendekatan dalam *clustering* yaitu *partitioning* dan *hirarki*. Dalam *partitioning* kita mengelompokkan obyek ke dalam k *cluster*. Ini bisa dilakukan dengan menentukan pusat *cluster* awal, lalu dilakukan realokasi obyek berdasarkan kriteria tertentu sampai dicapai pengelompokkan yang optimum. Dalam *cluster hirarki*, kita mulai dengan membuat m *cluster* dimana setiap *cluster* beranggotakan satu obyek dan berakhir dengan satu *cluster* dimana anggotanya adalah m obyek.

Dalam *clustering hirarki* kita hitung jarak masing – masing obyek dengan setiap obyek yang lain. Selanjutnya kita temukan pasangan obyek yang jaraknya terdekat. Sehingga tiap obyek akan berpasangan dengan satu obyek atau kelompok obyek yang lain yang paling dekat jaraknya. Langkah – langkah yang perlu dilakukan untuk melakukan *clustering hirarki* adalah [21] :

1. Kelompokkan setiap obyek ke dalam kelompok / *clusternya* sendiri.
2. Temukan pasangan paling mirip untuk dimasukkan ke dalam *cluster* yang sama dengan melihat data dalam matriks kemiripan.
3. Gabungkan kedua obyek dalam satu *cluster*.
4. Ulangi sampai tersisa hanya satu *cluster*.

2.2.4. Algoritma Apriori

Analisis asosiasi atau *association rule mining* adalah teknik *data mining* untuk menemukan aturan asosiatif antara suatu kombinasi item. Biasanya digunakan untuk analisis pembelian barang di pasar swalayan dengan maksud untuk mengetahui berapa besar kemungkinan seorang konsumen membeli satu jenis barang bersamaan dengan jenis barang yang lainnya. Sehingga perusahaan dapat mengatur strategi promosi dengan penempatan barang yang saling berhubungan ditempatkan secara berdekatan dan menetapkan strategi harga promosi untuk barang – barang tertentu yang saling berhubungan.

Metodologi dasar analisis asosiasi terbagi menjadi dua tahap [15] :

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 :

$$\text{Support } A = \frac{\text{Jumlah transaksi mengandung } A}{\text{Total transaksi}} \dots\dots\dots(2)$$

Sementara itu, nilai *support* dari 2 item diperoleh dari rumus 2 berikut :

$$\text{Support } (A,B) = P (A \cap B)$$

$$\text{Support } (A,B) = \frac{\text{Jumlah transaksi mengandung } A \text{ dan } B}{\text{Jumlah transaksi}} \dots\dots\dots(3)$$

2. Pembentukan Aturan Asosiasi

Setelah semua 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 dari rumus berikut :

$$\text{Confidence} = P (B | A) = \frac{\text{Jumlah transaksi mengandung } A \text{ dan } B}{\text{Jumlah transaksi mengandung } A} \dots\dots\dots(4)$$

2.2.5. Confusion Matrix

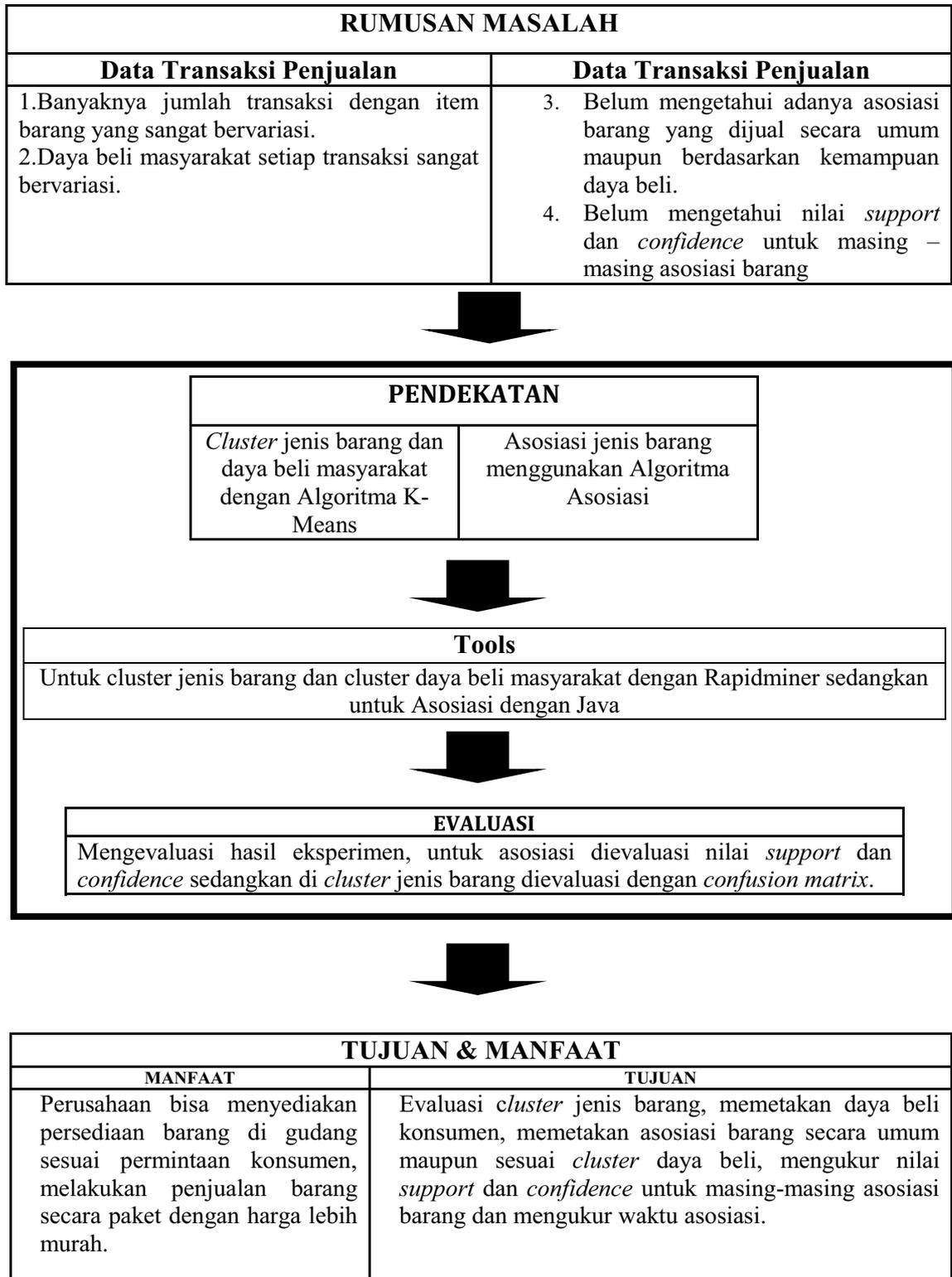
Confusion Matrix menunjukkan jumlah prediksi yang benar dan salah yang dibuat oleh model klasifikasi dibandingkan dengan hasil yang sebenarnya (nilai target) dalam data. *Matrix* adalah $n \times n$, dimana n adalah jumlah nilai target (kelas). Kinerja model seperti ini biasanya dievaluasi dengan menggunakan data dalam *matrix*. Tabel berikut menampilkan *confusion matrix* 2×2 untuk dua kelas (positif dan negatif).[12]

Confusion Matrix		Tar get		
		Positive	Negative	
Model	Positive	a	b	Positive Predictive Value $a/(a+b)$
	Negative	c	d	Negative Predictive Value $d/(c+d)$
		sensitivity	Specificity	Accuracy = $(a+d)/(a+b+c+d)$
		$a/(a+c)$	$d/(b+d)$	

Akurasi : proporsi jumlah prediksi yang benar.[12]

- Positif presisi nilai prediktif atau proporsi kasus positif yang diidentifikasi dengan benar.
- Negatif prediktif value : proporsi kasus negatif yang diidentifikasi dengan benar.
- Sensitivitas : proporsi kasus positif sebenarnya yang diidentifikasi dengan benar.
- Kekhususan : proporsi kasus negatif yang sebenarnya yang diidentifikasi dengan benar.

2.3.Kerangka Pemikiran



Bagan 2.1. Kerangka Pemikiran