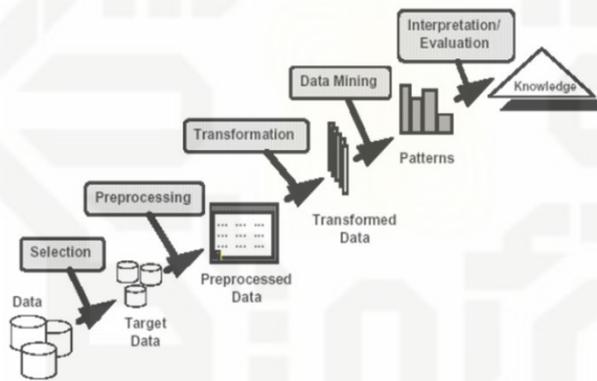


## BAB II

### LANDASAN TEORI

#### 2.1 Knowledge Discovery in Data (KDD)

Data Mining dan *Knowledge Discovery in Database* (KDD) sering kali digunakan untuk memberitahukan proses pencarian informasi yang tersembunyi di dalam *database* yang besar. Namun pada dasarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi saling berkaitan antara satu sama lain dan salah satu tahapan dalam proses KDD adalah *data mining*. Berikut akan dijelaskan proses KDD secara garis besar (Kusrini & Taufiq, 2009).



Gambar 2. 1 Proses KDD

##### 1. Data Selection

Merupakan pemilihan data dari sekumpulan data operasional yang perlu dilakukan sebelum penggalian informasi dalam KDD dimulai. Data dari hasil seleksi digunakan untuk proses *data mining*, disimpan pada suatu berkas, terpisah dari basis data operasional.

##### 2. Pre-processing/Cleaning

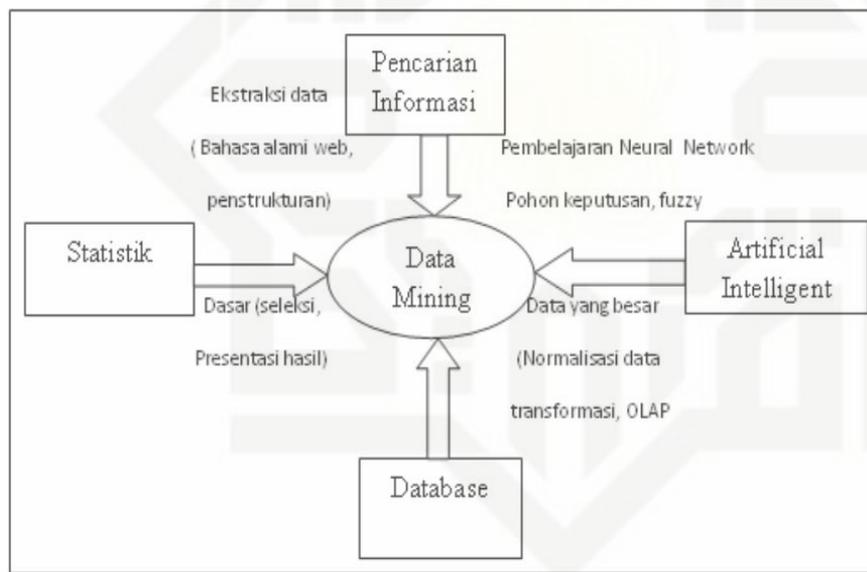
Tahap *cleaning* ini mencakup seperti membuang data duplikasi, memeriksa data yang *inkonsisten*, dan memperbaiki kesalahan pada data, seperti kesalahan dalam menulis (tipografi).



Data Mining juga merupakan suatu istilah yang digunakan untuk menemukan suatu pengetahuan yang tersembunyi di dalam *database* (Turban & Aronson, 2005).

Adapun dapat dikatakan sebagai data *mining* apabila terkait pada beberapa hal berikut.

1. *Data mining* merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses berupa data yang sangat besar. Biasanya data yang besar digunakan untuk membuat hasil yang lebih dipercaya.
3. Tujuan *data mining* yaitu mendapatkan hubungan atau pola yang tidak diketahui sebelumnya dan akan memberikan indikasi yang bermanfaat.



**Gambar 2. 2 Bidang Ilmu Data Mining**  
 (Kusrini & Taufiq, 2009)

### 2.2.1 Proses Data Mining

Proses data *mining* terbagi dalam tiga aktivitas yaitu (Gorunescu, 2011):

1. Eksplorasi data, terdiri dari aktivitas pembersihan data, transformasi data, pengurangan dimensi, pemilihan ciri dan lain-lain.
2. Membuat model dan pengujian validitas model, merupakan pemilihan terhadap model-model yang sudah dikembangkan yang cocok dengan kasus

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

yang dihadapi. Dengan kata lain, dilakukan pemilihan model secara kompetitif.

3. Penerapan model dengan data baru untuk menghasilkan perkiraan dari kasus yang ada. Tahap ini merupakan tahap yang menentukan apakah model yang dibangun dapat menjawab permasalahan yang dihadapi.

### 2.2.2 Teknik Data Mining

Ada beberapa teknik di dalam *data mining*. Dalam kasus tertentu teknik *data mining* dapat digabungkan untuk menjawab masalah yang dihadapi. Berikut ini beberapa teknik di dalam *data mining* yaitu (Bala, Devi, & Saravanan, 2012):

#### 1. Analisis Asosiasi

Analisis asosiasi ini merupakan penemuan aturan asosiasi yang menggambarkan kondisi atribut yang sering terjadi dalam satuan data tertentu. Analisis asosiasi biasa digunakan untuk analisa data pasar dan data transaksi.

#### 2. Klasifikasi dan Prediksi

Merupakan suatu proses untuk menemukan sebuah model (fungsi) yang menjelaskan atau mencirikan konsep atau kelas data untuk suatu kepentingan tertentu, yang bisa menggunakan permodelan untuk memprediksi kelas objek yang labelnya tidak diketahui. Model yang didapat mungkin bisa diwakili dalam berbagai format, seperti aturan klasifikasi IF-THEN, pohon keputusan, formula matematika, atau jaringan saraf tiruan. Klasifikasi bisa digunakan untuk memprediksi label kelas data objek pada data.

#### 3. Analisis *Clustering*

Berbeda dengan klasifikasi dan prediksi yang menganalisa pelabelan objek data, *clustering* menganalisis objek data tanpa mengkonsultasikan label kelas yang dikenal. Secara umum label kelas bukan didapat dalam pengolahan data sederhana, karena mereka tidak tahu bagaimana memulainya. *Clustering* bisa digunakan untuk meng-*generate* label. Objek yang di *cluster* (dikelompokkan) berdasarkan pada prinsip memaksimalkan persamaan dalam kelas dan meminimalkan kesamaan antar kelas.

#### 4. Analisis *Outlier*

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

Di dalam sebuah *database* mungkin berisi objek data yang tidak sesuai dengan kebiasaan umumnya yang disebut *Outlier*. Analisa *outlier* membantu dalam pendeteksian kesalahan dan nilai-nilai abnormal.

### 2.2.3 Pengelompokan *Data Mining*

Berdasarkan tugas yang dapat dilakukannya, *Data Mining* dibagi menjadi beberapa kelompok, yaitu (Larose, 2005):

#### 1. Deskripsi

Deskripsi adalah menggambarkan pola dan kecenderungan yang terdapat dalam data yang memungkinkan memberikan penjelasan dari suatu pola atau kecenderungan tersebut.

#### 2. Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan memprediksikan kelas untuk data yang tidak diketahui kelasnya.

#### 3. Estimasi

Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah *numeric* dari pada ke arah kategori. Model dibangun menggunakan *record* lengkap yang menyediakan nilai variabel target sebagai nilai prediksi.

#### 4. Prediksi

Prediksi hampir sama dengan klasifikasi dan estimasi, akan tetapi dalam prediksi nilai dari hasil akan ada di masa mendatang.

#### 5. Pengklusteran

Pengklusteran merupakan pengelompokan *record*, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan. Kluster adalah kumpulan *record* yang memiliki kemiripan satu dengan yang lainnya dan memiliki ketidakmiripan dengan *record-record* dalam kluster lain.

#### 6. Asosiasi

Asosiasi dalam *data mining* adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

## 2.3 Association Rules

*Association Rules* adalah teknik data *mining* untuk menemukan aturan asosiatif antara suatu kombinasi item. Contohnya pada analisis pembelian di suatu pasar swalayan adalah dapat diketahuinya berapa besar kemungkinan seorang pelanggan membeli roti bersama dengan susu. Dengan pengetahuan tersebut, pemilik pasar swalayan dapat mengatur penempatan barang atau merancang kampanye pemasaran untuk kombinasi barang tertentu. Analisis asosiasi menjadi terkenal dengan istilah *market basket analysis*.

*Association rule mining* adalah suatu prosedur pencarian hubungan antar item dalam suatu data set yang ditentukan (Han, Pei, & Kamber, Data Mining: Concepts and Techniques Third Edition, 2001). Penting tidaknya aturan asosiatif dapat diketahui dengan dua parameter, yaitu sebagai berikut :

1. *Support*, yaitu ukuran yang menunjukkan seberapa besar tingkat dominasi suatu item dari keseluruhan transaksi. (misal, dari keseluruhan transaksi yang ada, seberapa besar tingkat dominasi yang menunjukkan bahwa item A dan B dibeli bersamaan).
2. *Confidence*, yaitu ukuran yang menunjukkan hubungan antar dua item secara *conditional* (misal, seberapa sering item B dibeli jika orang membeli item A).

Kedua ukuran tersebut akan digunakan dalam menentukan *interesting association rules*, yaitu untuk dibandingkan dengan batasan yang ditentukan oleh *user*. Batasan tersebut umumnya terdiri dari *minimum support* dan *minimum confidence* yang digunakan pada proses pencarian *association rules*. Penentuan *minimum support* tujuannya untuk mereduksi jumlah pola akhir yang diperoleh, *support* yang terlalu kecil akan menyebabkan pola yang didapat sangat banyak sehingga menyulitkan pencarian pola, sebaliknya jika terlalu besar akan mengakibatkan informasi yang didapat hanya sedikit dan kemungkinan pola yang seharusnya bermanfaat malah akan terlewat.

Proses pencarian asosiasi terbagi menjadi dua tahap yaitu analisis *frequent itemset* dan pembentukan asosiasi (Han, Pei, & Kamber, Data Mining Concepts and Techniques Third Edition, 2011).

1. Analisis *frequent Item*

Tahapan ini mencari kombinasi item yang memenuhi syarat minimum dari nilai *support* dalam *database*. Nilai *support* sebuah item diperoleh dengan rumus 2.1 sebagai berikut:

$$Support(x) = \frac{Jumlah\ Transaksi\ yang\ mengandung\ item\ X}{Total\ transaksi} \times 100\% \dots\dots\dots(2.1)$$

Kemudian untuk mencari nilai *support* dari dua item dapat menggunakan rumus berikut:

$$Support(x,y) = \frac{\sum Transaksi\ yang\ mengandung\ item\ X\ dan\ Y}{\sum Total\ transaksi} \times 100\% \dots\dots\dots(2.2)$$

Setelah semua *frequent item* didapat, maka ditentukan syarat *minimum Confidence*. *Confidence* adalah nilai ukuran seberapa besar valid tidaknya suatu *association rules*. *Confidence* suatu rule R ( $x \rightarrow y$ ) adalah proporsi dari semua transaksi yang mengandung baik x maupun y dengan yang mengandung x.

2. Pembentukan *Assosiation Rule*

Setelah semua pola *frequent Itemset* ditemukan, kemudian dicari aturan asosiasi yang memenuhi syarat minimum untuk *confidence* dengan menghitung *confidence* aturan asosiatif  $x \rightarrow y$ . Nilai *confidence* dari aturan  $x \rightarrow y$  diperoleh dari rumus 2.3 berikut.

$$Confidence(x \rightarrow y) = \frac{\sum Transaksi\ yang\ mengandung\ item\ X\ dan\ Y}{\sum Transaksi\ yang\ mengandung\ X} \dots\dots\dots(2.3)$$

Keterangan :

x dan y : *item*

**2.4 Algoritma ECLAT**

Algoritma *Equivalence Class Transformation (ECLAT)* merupakan algoritma sederhana yang dapat digunakan untuk menemukan *itemset* yang paling sering muncul. Pada algoritma ini terdapat metode *frequent mining* dari kumpulan transaksi dengan menggunakan bentuk *TID-itemset* dimana *TID* merupakan ID transaksi dan *itemset* merupakan kumpulan dari barang yang dibeli pada transaksi (*TID List*). Pada dasarnya algoritma ECLAT melakukan pencarian dengan struktur kedalaman pada *database* dengan tata letak vertikal, jika *database* berbentuk

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

horizontal maka harus dikonversikan ke bentuk vertikal terlebih dahulu. (Kaur & Grag, 2014).

**Tabel 2. 1 Format Data Horizontal**

| TID  | Itemset    |
|------|------------|
| TI00 | T1, T2, T5 |
| TI01 | T2, T4     |
| TI02 | T1, T3, T5 |
| TI03 | T2, T3     |
| TI04 | T2, T4     |
| TI05 | T1, T2, T3 |

**Tabel 2. 2 Format Data Vertikal**

| Itemset | TID                          |
|---------|------------------------------|
| T1      | TI00, TI02, TI05             |
| T2      | TI00, TI01, TI03, TI04, TI05 |
| T3      | TI02, TI03, TI05             |
| T4      | TI01, TI04                   |
| T5      | T2, T4                       |

Proses pencarian dilakukan dari item yang paling sering muncul hingga yang paling jarang muncul tanpa harus memperhatikan urutan, sehingga proses *pemindaian* tidak perlu dilakukan secara berulang-ulang. Data setiap *itemset* disimpan di sebuah *Transaction Id List (TID List)*, kemudian *TID List* diurutkan berdasarkan transaksi yang mengandung *itemset* yang sama (*frequent itemset*). Selanjutnya *k-itemset* diatur ke dalam kelas-kelas berdasarkan kriteria tertentu yang

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

terbentuk dengan mempartisi suatu himpunan (*equivalence class*),  $(k+1)$ -*itemset* bisa didapat dengan menggabungkan pasangan *frequent k-itemset* dari kelas yang sama. Dalam prosesnya, algoritma ini dilakukan secara rekursif, dimana pencarian *itemset* akan terus dilakukan sepanjang masih ada *itemset* yang tersisa (pencarian menyeluruh). Secara garis besar algoritma ECLAT dibagi menjadi 3 fase (Adita et al., n.d.), yaitu:

1. Fase Inisialisasi: kontruksi perhitungan global untuk *frequent 2-itemset*
2. Fase Transformasi: mempartisi *frequent 2-itemset* dan menjadwalkan partisi atas *processor*. Fase ini juga melakukan vertikal transformasi pada *database*.
3. Fase *Asynchronous*: kontruksi *frequent k-itemset*, sesuai tujuan yang diinginkan.

**1) Fase Inisialisasi**

Proses pembentukan *itemset* pada algoritma ECLAT di mulai dengan fase inisialisasi yaitu mengubah bentuk transaksi, jika item pada transaksi berbentuk horizontal maka diubah menjadi bentuk vertikal dengan menggabungkan *TID List* pada transaksi yang memiliki item yang sama. Perubahan bentuk transaksi dari bentuk horizontal menjadi vertikal dapat dilihat seperti pada tabel 2.3 dan 2.4 berikut:

**Tabel 2. 3 Transaksi Awal**

| TID | Item       |
|-----|------------|
| 1   | a, b       |
| 2   | b, c, d    |
| 3   | b, c, d, e |
| 4   | a, d, e    |
| 5   | a, b, c    |





merupakan hasil dari operasi konjungsi antara anggota subset AB dan BC. Begitu juga dengan itemset seterusnya.

Kemudian dilakukan batas *minimum* pada *itemset* dengan menggunakan nilai *frequent* yang telah ditentukan (pada kasus ini *frequent* =2). Maka hasil akhir pada transaksi dengan 3-*itemset* adalah sebagai berikut

**Tabel 2. 8 Hasil Frequent 3-itemset**

| Itemset | TID List |
|---------|----------|
| (b,c,d) | (2,3)    |

Tabel 2.8 menunjukkan data transaksi yang memiliki frequent 3-itemset. Proses penyilangan dapat terus dilakukan sehingga *frequent k-itemset* ditemukan.

#### 4) Tahap Pembentukan Rule

Setelah hasil dari *frequent itemset* ditemukan maka selanjutnya dilakukan tahap pembentukan *rule* dengan melakukan perhitungan terhadap nilai *support* dan nilai *confidence* dari masing-masing *itemset* dengan menggunakan persamaan (2.2) dan persamaan (2.3) untuk menentukan nilai ukuran seberapa valid *asossiation rule* yang ditemukan.

**Tabel 2. 9 Nilai Support dan Confidence Frequent 2-itemset**

| Itemset | Support (%) | Confidence (%) |
|---------|-------------|----------------|
| (a→b)   | 40          | 66.667         |
| (b→c)   | 60          | 50             |
| (b→d)   | 40          | 50             |
| (c→d)   | 40          | 66.667         |
| (d→e)   | 40          | 66.667         |

Tabel 2.9 merupakan nilai *support* dan *confidence* dari *frequent 2-itemset*. Pada aturan asosiasi (a→b) diketahui bahwa nilai *support* atau nilai transaksi yang mengandung item a dan b dari total transaksi adalah 0.4%, sedangkan nilai *confidence* atau nilai proposisi semua transaksi yang mengandung a dan b dari semua transaksi yang mengandung a dari aturan asosiasi (a→b) adalah 0.66667%,

dari tabel 2.9 juga dapat dilihat bahwa aturan asosiasi yang mempunyai nilai *support* dan *confidence* tertinggi adalah (b→d). Untuk nilai *support* dan *confidence* pada aturan asosiasi dengan *frequent 3-itemset* dapat dilihat pada tabel 2.10 berikut.

**Tabel 2. 10 Nilai Support dan Confidence Frequent 3-itemset**

| <i>Itemset</i> | <i>Support (%)</i> | <i>Confidence (%)</i> |
|----------------|--------------------|-----------------------|
| b,c → d        | 40                 | 66.667                |
| b,d → c        | 40                 | 100                   |
| c,d → b        | 40                 | 100                   |
| b → c,d        | 40                 | 50                    |
| c → b,d        | 40                 | 66.667                |
| d → b,c        | 40                 | 66.667                |

Dari tabel 2.10 dapat dilihat bahwa transaksi yang mempunyai nilai *support* dan *confidence* tertinggi dari aturan asosiasi *frequent 3-itemset* adalah (b,d→c) dan (c,d→b) dengan nilai *support* 40% dan *confidence* 100%.

## 2.5 Penelitian Terkait

Terdapat beberapa penelitian terdahulu yang menjadi referensi untuk dijadikan acuan dalam penelitian ini. Adapun penelitian yang terkait dapat dilihat pada tabel 2.11.

**Tabel 2. 11 Penelitian Terkait**

| No. | Nama                                     | Tahun | Judul   | Hasil  | Metode                    |
|-----|--|-------|---|--|---------------------------|
| 1.  | Adita,<br>Muflikhah,<br>dan<br>Ratnawati | 2013  | Personalisasi<br><i>E-Library</i><br>Menggunakan<br>Kaidah<br>Asosiasi<br>Algoritma<br><i>ECLAT</i> | Dari penelitian ini<br>didapat rata-rata nilai<br><i>Lift Ratio</i> dari rule<br>yang dihasilkan adalah<br>3,43. Sedangkan titik<br>optimum dari<br><i>minimum support</i> | Algoritma<br><i>ECLAT</i> |

|    |                |      |  |  |  |
|----|----------------|------|--|--|--|
|    |                |      |  | <p>adalah 10% dan <i>minimum confidence</i> adalah 60% dengan jumlah transaksi 150 transaksi peminjaman.</p>   |  |
| 2. | Sinha & Ghosh  | 2014 | <p><i>Identification of Best Algorithm in Association Rule Mining Based on Performance</i></p> | <p>Hasil pada penelitian ini menunjukkan algoritma ECLAT merupakan algoritma terbaik dari segi waktu, karena memiliki waktu eksekusi yang lebih cepat dengan nilai <i>support</i> dan <i>confidence</i> yang lebih baik</p>  | <p><i>FP-Growth, Apriori</i> dan ECLAT</p> |
| 3. | Solanki & Sony | 2014 | <p><i>A Survey on Frequent Pattern Mining Methods Apriori, ECLAT, FP growth</i></p>            | <p>Hasilnya menunjukkan ECLAT hanya butuh satu kali <i>scanning database</i> untuk menemukan <i>frequent itemsets</i> dibandingkan dengan <i>Apriori</i> yang melakukan <i>scanning</i> secara berulang dan <i>FP growth</i> yang melakukan struktur data yang complex</p> | <p><i>FP-Growth, Apriori</i> dan ECLAT</p> |

- Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
    - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
    - b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
  2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

|    |                                  |      |  |   |  |
|----|----------------------------------|------|--|---|--|
| 4. | Sadat, Samuel, Patel, dan Zaiane | 2011 | <i>Fastest Association Rule Mining Algorithm Predictor (FARM-AP)</i>   | Penelitian ini memprediksi waktu tercepat dari ketiga algoritma asosiasi, yaitu <i>Apriori</i> , <i>FP-Growth</i> , dan <i>ECLAT</i> . Hasilnya didapati <i>ECLAT</i> sebagai algoritma yang paling baik dari ketiganya   | <i>FP-Growth</i> , <i>Apriori</i> dan <i>ECLAT</i> |
| 5. | Robi & Riri                      | 2015 | Implementasi <i>Data Mining</i> dengan Metode Algoritma <i>Apriori</i> dalam Menentukan Pola Pembelian Obat. | Pada penelitian tersebut didapat kecenderungan obat yang dibeli konsumen. Hasilnya ditentukan <i>minimum support</i> 50% dan <i>minimum confidence</i> 70% dan didapat outputnya jika membeli obat <i>amoxicilin</i> maka membeli obat <i>asamefenamat</i> dengan <i>confidence</i> 75% | Algoritma <i>Apriori</i>                           |

Hak Cipta Dilindungi Undang-Undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
- b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.