BAB II

TINJAUAN PUSTAKA

2.1 *Knowledge discovery in database* (KDD)

Knowledge discovery in database adalah suatu rangkaian proses yang digunakan oleh suatu lembaga untuk mendeskripsikan, memperbaharui, menerangkan, dan menyebarluaskan informasi untuk digunakan kembali, diketahui, dan dipelajari di dalam organisasi. Kegiatan ini biasanya terkait dengan objektif organisasi dan ditujukan untuk mencapai suatu hasil tertentu seperti wawasan bersama, pengembangan kinerja, keberhasilan daya saing, atau tingkat pembaharuan yang lebih baik (Sistem, Fakultas, Komputer, & Sriwijaya, 2016).

Knowledge Discovery in Database (KDD) adalah proses yang melingkupi pengambilan, penggunaan database untuk memperoleh kesamaan, rule atau keterkaitan dalam set data yang bermedium besar. Di dalam jurnal yang berjudul "Pembentukan Cluster dalam *Knowledge in Database* dengan Algoritma K-*Means*". *Knowledge Discovery in Database* (KDD) didefenisikan sebagai estraksi inforasi potensial, implisit dan tidak dikenal dari sekumpulan data.

Istilah *data mining* dan *knowledge discovery in database* (KDD) sering kali digunakan secara bergantian untuk memahami proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah *data mining* (Studi, Informasi, Ilmu, Universitas, & Kuning, 2016).

2.1.1 Proses *Knowledge Discovery in Database*

Proses *knowledge discovery in databases (KDD)* secara garis besar dapat dijelaskan sebagai berikut(Pujiono, Budiono, & Fahmi, 2014) :

1. *Data selection*

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD. Data hasil seleksi yang akan 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 focus KDD. Proses cleaning mencakup antara lain membuang penggandaan data, memeriksa data yang *inkonsisten*, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Juga dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan yang diperlukan untuk KDD, seperti data atau informasi eksternal.

3. Integrasi data

Integrasi data merupakan penggabungan data dari berbagai database kedalam database baru.

4. 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.

5. Data Mining

*Data minin*g 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.

6. Interpretation/Evaluation

Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dipahami 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.

7. Presentasi Pengetahuan *(knowledge*)

Merupakan visualisasi dan penyajian pengetahuan mengenai teknik yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna. Tahap terakhir dari proses *data mining* adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisa yang didapat.

2.2 *Data Mining*

Data mining adalah kegiatan mengolah dengan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mendistribusikan dan

mendeskripsikan informasi yang berguna dan wawasan yang berhubungan dari berbagai data yang disimpan. Pada saat kita dihadapkan kepada sekumpulan data dari suatu objek atau kejadian, kita perlu mengolahnya untuk mendapatkan manfaat dari data itu. Kita perlu mengenali polanya sehingga kita akan menemukan kecenderungan dari data tersebut (Lunak, 2017).

Kemudian kita dapat juga melakukan perkiraan atas apa yang akan terjadi pada jangka waktu kedepannya berdasar data masa sebelumnya berkaitan dengan data jumlah penjualan tersebut. Jadi pengenalan pola adalah suatu disiplin ilmu yang mempelajari bagaimana kita mengelompokkan obyek ke berbagai kelas dan bagaimana dari data bisa kita temukan kecenderungannya. Yang pertama mengacu pada kasus klasifikasi dan yang kedua mengacu pada regresi. *Data mining* juga mengacu pada langkah- langkah menentukan variable atau fitur yang penting untuk dipakai dalam klasifikasi dan regresi. *Data mining* memegang peranan penting dalam bidang industry, keuangan, cuaca, ilmu dan teknologi.

Berikut ini adalah beberapa contoh yang memperlihatkan masalah-masalah dalam data mining (Lunak, 2017) :

1. Memprediksi harga suatu saham dalam beberapa bulan kedepan berdasarkan performansi perusahaan dan data-data ekonomi.
2. Memprediksi apakah seorang pasien yang diopname akan mendapat serangan jantung berikutnya berdasarkan catatan kesehatan sebelumnya dan pola makanannya.
3. Memprediksi permintaan semen dalam beberapa tahun mendatang berdasarkan data permintaan semen di tahun-tahun sebelumnya.

4. Memprediksi apakah akan terjadi tornado berdasarkan informasi dari sebuah radar tentang kondisi angin dan kondisi atmosfir yang lain.

5. Identifikasi apakah sudah terjadi penipuan terhadap kartu kredit dengan melihat catatan transaksi yang tersimpan dalam database perusahaan kartu kredit (Lunak, 2017).

Data mining menetapkan *set size* dan *set size frequency* pada data transaksi sehingga mengetahui ukuran dan frekuensi transaksi yang terjadi yang berisikan menggunakan algoritma apriori klasik yang sudah dikembangkan sebelumnya dan belum menggunakan teknik optimasi untuk memperoleh aturan asosiasi yang lebih efisien. *Data mining* dapat dibagi menjadi beberapa kelompok yaitu deskripsi, estimasi, prediksi, klasifikasi, klustering dan asosiasi (Jaya, 2018).

Data mining adalah suatu proses menemukan hubungan pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika, *Data mining* juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data (Pujiono et al., 2014).

Karakteristik *data mining* sebagai berikut(Pujiono et al., 2014):

1. *Data mining* berhubungan dengan penemuan sesuatu yang tersembunyi dan pola data tertentu yang tidak diketahui sebelumnya.
2. *Data mining* biasa menggunakan data yang sangat besar, biasanya data yang besar digunakan untuk membuat hasil lebih dipercaya.
3. *Data mining* berguna untuk membuat keputusan yang kritis.

2.2.1. Pengelompokan *Data Mining*

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Tampubolon et al., 2013) :

1. *Description* (Deskripsi)

Terkadang peneliti dan analis secara sederhana ingin mencoba mencari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Sebagai contoh, petugas pengumpulan suara mungkin tidak dapat menemukan keterangan atau fakta bahwa siapa yang tidak cukup profesional akan sedikit didukung dalam pemilihan presiden.

2. *Estimation*

Estimasi hampir sama dengan klasifikasi, kecuali variable target estimasi lebih kearah numerik dari pada kearah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi. Sebagai contoh akan dilakukan estimasi tekanan darah sistolik pada pasien rumah sakit berdasarkan umur pasien, jenis kelamin, indeks berat badan, dan level sodium darah. Hubungan antara tekanan darah sistolik dan nilai variabel prediksi dalam proses pembelajaran akan menghasilkan model estimasi. Model estimasi yang dihasilkan dapat digunakan untuk kasus baru lainnya.

3. Prediksi.

Prediksi hampir sama dengan klasifikasi dan estimasi, kecuali bahwa dalam predikasi nilai dari hasil akan ada dimasa mendatang. Contoh prediksi bisnis dan penelitian adalah:

a. Prediksi harga beras dalam tiga bulan yang akan datang.

b. Prediksi persentasi kenaikan kecelakaan lalu lintas tahun depan jika batas bawah kecepatan dinaikkan.

Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi dapat pula digunakan (untuk keadaan yang tepat) untuk prediksi.

4. Klasifikasi

Dalam klasifikasi, terdapat target variable kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori , yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Contoh lain klasifikasi dalam bisnis dan penelitian adalah:

a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau tidak.

b. Memperkirakan apakah suatu pengajuan hipotek oleh nasabah merupakan suatu kredit yang baik atau buruk.

c. Mendiagnosis penyakit seorang pasien untuk mendapatkan termasuk kategori penyakit apa.

5. Pengklusteran (*Clustering*)

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. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target dalam pengklusteran. Pengklusteran tidak mencoba untuk melakukan klasifikasi, mengestimasi, atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki kemiripan (*homogeny*), yang mana kemiripan dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan *record* dalam kelompok lain akan bernilai minimal. Contoh pengklusteran dalam bisnis dan penelitian adalah:

a. Mendapatkan kelompok-kelompok konsumen untuk target pemasaran dari satu suatu produk bagi perusahaan yang tidak memiliki dana pemasaran yang besar.

b. Untuk tujuan audit akuntansi, yaitu melakukan pemisahan terhadap perilaku *financial* dalam baik dan mencurigakan.

c. Melakukan pengklusteran terhadap ekspresi dari gen, untuk mendapatkan kemiripan perilaku dari gen dalam jumlah besar.

6. Asosiasi

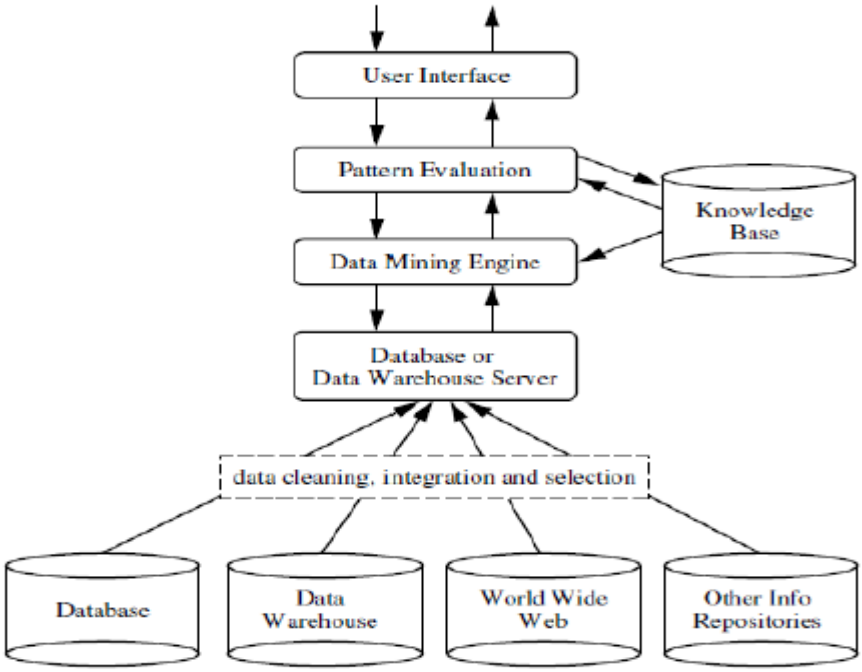
Tugas asosiasi dalam *data mining* adalah menemukan *attribut* yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja. Contoh asosiasi dalam bisnis dan penelitian adalah:

a. Meneliti jumlah pelanggan dari perusahaan telekomunikasi seluler yang diharapkan untuk memberikan respon positif terhadap penawaran *upgrade* layanan yang diberikan.

b. Menentukan barang dalam supermarket yang dibeli secara bersamaan dan yang dibeli secara bersamaan dan yang tidak pernah dibeli secara bersamaan.

2.2.2 Arsitektur Sistem *Data Mining*

Gambar berikut merupakan bagian-bagian dari arsitektur sistem *data mining*,



Gambar 2.1 Arsitektur Sistem *Data Mining*

Penjelasan dari bagian-bagian arsitektur sistem data mining diatas adalah sebagi berikut:

1. Basis data, data warehouse atau media penyimpanan lainnya

Media dalam hal ini dapat berupa basis data, data warehouse, spreadsheets, atau jenis-jenis penampungan informasi lainnya. Pembersihan data, integrasi data, dan seleksi data dilakukan pada bagian tersebut.

2. Server basis data/data *warehouse*

Server basis data/data *warehouse* bertanggung jawab dalam menyediakan data yang relevan berdasarkan permintaan pengguna *data mining*.

3. Basis pengetahuan

Pengetahuan yang digunakan dalam pencarian hubungan dari pola yang dihasilkan, seperti *concept merarchies* digunakan untuk mengorganisasikan nilai atribut atau atribut-atribut ke dalam level abstraksi yang berbeda.

4. Mesin *data mining*

Mesin *data mining* merupakan bagian dari perangkat lunak yang menjalankan program berdasarkan algoritma yang ada.

5. Model evaluasi pola

Model evaluasi pola merupakan bagian dari perangkat lunak yang berfungsi untuk menentukan pola-pola yang terdapat dalam basis data yang diolah sehingga nantinya proses *data mining* dapat menemukan pengetahuan yang sesuai.

6. GUI

Bagian ini merupakan sarana antar pengguna dan sistem data mining untuk berkomunikasi, dimana pengguna dapat berinteraksi dengan sistem melalui *data mining query*, untuk menyediakan informasi yang dapat membantu dalam pencarian pengetahuan. Bagian ini mengijinkan pengguna untuk melakukan browsing pada basis data dan data warehouse. Menevaluasi pola tersebut dengan tampilan yang berbeda-beda.

2.3 Algoritma *Association*

Algoritma aturan asosiasi akan menggunakan data latihan, sesuai dengan definisi *data mining*, untuk menghasilkan pengetahuan. Pengetahuan seperti apa akan dihasilkan dalam aturan asosiasi? Pengetahuan untuk mengetahuan *item-item* belanja yang sering dibeli secara bersamaan dalam suatu waktu. Aturan asosiasi yang berbentuk "*if...then*…" atau "jika…maka…" merupakan pengetahuan yang dihasilkan dari fungsi aturan asosiasi (Studi et al., 2016).

Metodologi dasar analisis asosiasi terbagi menjadi dua tahap :

1. Analisa 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:

Support (A) = aksi TotalTransndungA saksiMengaJumlahTran Sedangkan nilai dari *support* 2 item diperoleh dari rumus berikut :

Support (A,B) = P(A B) = aksi TotalTransndungAdanB saksiMengaJumlahTran

2. Pembentukan aturan assosiatif

Setelah semua pola frekuensi tinggi ditemukan, barulah dicari aturan assosiatif yang memenuhi syarat minimum untuk *confidence* dengan menghitung *confidence* aturan assosiatif A Nilai *confidence* dari aturan A diperoleh dari rumus berikut :

Confidence = P(B|A) = ndungA saksiMengaJumlahTran

2.4 *Algoritma Apriori*

Algoritma apriori masuk dalam kategori kaidah asosiasi pada *data mining*, *Algoritma apriori* adalah *algoritma* yang paling terkenal untuk menemukan pola frekuensi tinggi. *algoritma apriori* dibagi menjadi beberapa tahap yang disebut narasi atau pass (Studi et al., 2016):

1. Pembentukan kandidat *itemset*, kandidat k-*itemset* dibentuk dari kombinasi (k-1)-*itemset* yang didapat dari iterasi sebelumnya. Satu cara dari *algoritma apriori* adalah adanya pemangkasan kandidat k-*itemset* yang *subset*-nya yang berisi k-1 item tidak termasuk dalam pola frekuensi tinggi dengan panjang k-1.
2. Penghitungan *support* dari tiap kandidat k-*itemset*. *Support* dari tiap kandidat k-*itemset* didapat dengan men*scan database* untuk menghitung jumlah transaksi yang memuat semua *item* didalam kandidat k-*itemset* tersebut. Ini adalah juga ciri dari *algoritma apriori* dimana diperlukan penghitungan dengan cara seluruh *database* sebanyak k-*itemset* terpanjang.
3. Tetapkan pola frekuensi tinggi. Pola frekuensi tinggi yang memuat k *item* atau k-*itemset* ditetapkan dari kandidat k-*itemset* yang *support*nya lebih besar dari minimum *support*.
4. Bila tidak didapat pola frekuensi tinggi baru maka seluruh proses dihentikan. Bila tidak, maka k ditambah satu dan kembali bagian 1.

2.5 Analisa Keranjang Belanja

Analisis keranjang belanja mengacu pada berbagai teknologi yang mempelajari komposisi keranjang belanja yang terdiri atas produk-produk yang dibeli pada satu kejadian belanja. Teknik ini telah diterapkan secara luas dalam berbagai operasi pasar swalayan. Data keranjang belanja dalam bentuknya yang paling mentah adalah daftar transaksi pembelian oleh pelanggan, yang mengindikasikan hanya barang yang dibeli bersamaaan (Listriani, Setyaningrum, & A, 2016).

2.6 *Software* Pendukung

Tanagra adalah *software Data Mining* terbuka bagi akademik dan penelitian ini mengajukan beberapa metode *data mining* dari analisis pengembangan data, pembelajaran statistik, pembelajaran mesin dan daerah database.Tanagra adalah "proyek *open source*" karena setiap peneliti dapat mengakses ke kode sumber, dan menambahkan algoritma sendiri, sejauh dia setuju dan sesuai dengan lisensi distribusi perangkat lunak (Badrul, Studi, & Informasi, 2016).

Tujuan utama dari proyek Tanagra adalah memberikan peneliti dan mahasiswa yang mudah untuk menggunakan perangkat lunak *data mining*, sesuai dengan norma-norma yang hadir dari pengembangan perangkat lunak dalam domain ini (terutama dalam desain GUI dan cara untuk menggunakannya), dan memungkinkan untuk menganalisis baik data yang nyata atau sintetis.

Tujuan kedua Tanagra adalah untuk mengusulkan kepada peneliti arsitektur yang memungkinkan mereka untuk dengan mudah menambahkan metode penambangan data mereka sendiri, untuk membandingkan kinerja mereka. Tanagra bertindak lebih sebagai platform eksperimental untuk membiarkan mereka pergi ke penting dari pekerjaan mereka, pengeluaran mereka untuk berurusan dengan bagian menyenangkan dalam *programmation* semacam ini alat pengelolaan data.

Tujuan ketiga dan terakhir, arah pengembang pemula, terdiri dalam menyebarkan metodologi yang mungkin untuk membangun perangkat lunak semacam ini. Mereka harus mengambil keuntungan dari akses gratis ke kode sumber, untuk melihat bagaimana perangkat lunak semacam ini dibangun, masalah untuk menghindari, langkah-langkah utama dari proyek ini, dan alat-alat dan perpustakaan kode yang digunakan untuk. Dengan cara ini, Tanagra dapat dianggap sebagai alat pedagogis untuk belajar teknik pemrograman.

2.7 Penelitian Terdahulu

Berikut merupakan penelitian-penelitian terkait yang telah dibuat oleh peneliti sebelumnya, yaitu:

1. Penelitian yang dilakukan oleh (Prisilla, Farmadi, & Candra, 2014) yang berjudul **IMPLEMENTASI METODE ALGORITMA APRIORI PADA SISTEM PENDUKUNG KEPUTUSAN**, pada penelitian ini membahas tentang algoritma apriori dapat diperolehnya beberapa rule dapat menentukan sistem pendukung keputusan yang dapat digunakan sebagai acuan. Hasil dari penelitian

ini adalah berupa keputusan bahwa barang yang sering dibeli konsumen kemungkinan berkisar antara rok dan blus, sehingga untuk kedepannya pada proses order barang pihak manajemen toko dapat menggunakan keputusan dari sistem yang telah dibangun.

2. Penelitian yang dilakukan oleh (Jaya, 2018) yang berjudul **ANALISIS PENJUALAN PRODUK RETAIL DENGAN METODE DATA MINING ASOSIASI,** pada penelitian ini membahas tentang penyediaan barang memiliki sifat yang saling berhubungan atau membentuk aturan asosiasi sehingga dapat membantu dalam analisis penjualan produk retail yang efektif. Hasil dari penelitian ini adalah pembentukan aturan apriori menunjukan keterkaitan antar produk yang sering dibeli secara bersamaan antara lain tolak angin sachet, farm house sosis sapi, dll.

3. Penelitian yang dilakukan oleh (Nursikuwagus et al., 2016) yang berjudul **IMPLEMENTASI ALGORITMA UNTUK ANALISIS PENJUALAN DENGAN BERBASIS WEB,** pada penelitian ini membahas tentang aplikasi berdasarkan algoritma apriori dalam penjualan dengan berbasis web yang dapat membantu dalam memahaminnya. Hasil penelitian ini memberikan paparan mengenai hasil dari implementasi dari algoritma apriori dengan bahasa pemrograman berbasis web. Beberapa hasil menunjukan setiap proses yang dilakukan oleh tahapan algoritma apriori.

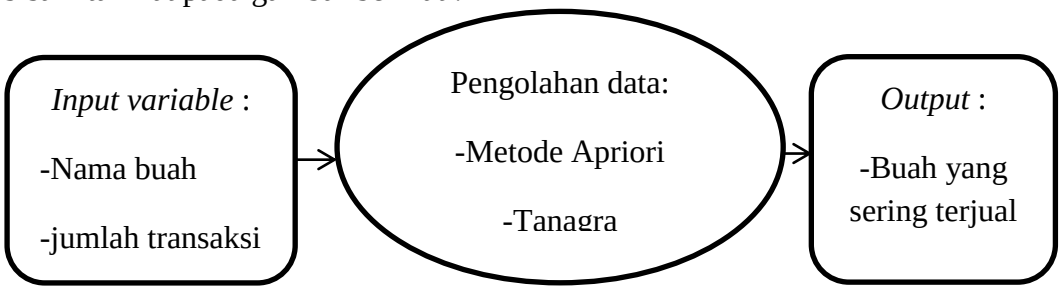
4.Penelitian yang dilakukan oleh (Apriori, 2016) yang berjudul **IMPLEMENTASI DATA MINING UNTUK PENENTUAN POSISI BARANG PADA RAK MENGGUNAKAN METODE APRIORI PADA PT**

MIDI UTAMA INDONESIA, pada penelitian ini membahas membantu pihak alfamidi untuk menyusun penempatan produk yang dijual berdasarkan penentuan posisi barang pad rak dengan menggunakan data mining dalam pengolahannya. Hasil penelitian ini adalah perhitungan dengan algoritma apriori yang berupa kombinasi itemsets atau keterkaitan barang dapat digunakan untuk melakukan penataan barang dalam rak atau etalase.

5. Penelitian yang dilakukan oleh (Handoko, 2016) yang berjudul **PENERAPAN DATA MINING DALAM MENINGKATKAN MUTU PEMBELAJARAN PADA INSTANSI PERGURUAN TINGGI MENGGUNAKAN METODE K-MEANS CLUSTERING (STUDI KASUS DI PROGRAM STUDI TKJ AKADEMI KOMUNITAS SOLOK SELATAN),** pada penelitian ini membahas tentang penerapan *data mining* dengan menggunakan metode clustering k-means untuk meningkatkan mutu pembelajaran pada instansi perguruan tinggi studi kasus program studi TKJ akademi komunitas solok selatan. Hasil penelitian ini adalah penerapan *data mining* dengan metode clustering k-means sehingga menghasilkan cluster-cluster dalam meningkatkan mutu pembelajaran.

2.8 Kerangka Pemikiran

Kerangka pemikiran yang dapat penulis gambarkan untuk penelitian ini bisa kita lihat pada gambar berikut :



Gambar 2. 2 Kerangka Pemikiran

Pada gambar 2.2, dapat dijelaskan bahwa proses awal dari penelitian ini adalah adanya variabel yang telah dipilih berdasarkan nama buah dan jumlah transaksi akan diseleksi terlebih dahulu, agar tidak ada tipografi. Selanjutnya data ini akan diproses dengan metode Apriori lalu diuji dengan Tanagra untuk menemukan buah yang paling banyak terjual dan untuk menentukan apakah hasil dari *software* penguji sama dengan cara manual. Hasil akhir yaitu bisa mengetahui buah yang sering terjual, pada hasil akhir inilah yang akan diterapkan pada toko tersebut sehingga membantu dalam pemilihan *stock.*