# BAB II
# TINJAUAN PUSTAKA

## 2.1 Data Mining

Data mining adalah suatu proses di mana data yang telah dipilih, dibersihkan, dan diubah menjadi bentuk yang lebih bermanfaat [5]. Metode klasifikasi adalah pendekatan untuk melakukan pengelompokan data dalam data mining yaitu menggolongkan data. Metode klasifikasi ini juga dapat digunakan untuk melakukan prediksi atas informasi yang belum diketahui sebelumnya [3].

*Association rule mining* adalah suatu prosedur untuk mencari hubungan antar item dalam suatu data set yang ditentukan. *Association rule* meliputi dua tahap:

a. Mencari kombinasi yang paling sering terjadi dari suatu itemset.
b. Mendefinisikan *Condition* dan *Results* (untuk *conditional association rule*).

Selain definisi diatas beberapa definisi juga diberikan seperti tertera dibawah ini : “Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.” “Data mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk pengenalan permasalahan pengambilan informasi dari database yang besar .” Kemajuan luar biasa yg terus berlanjut dalam bidang data mining didorong oleh beberapa faktor, antara lain :

1. Pertumbuhan yang cepat dalam pengumpulan data.
2. Penyimpangan data dalam data warehouse, sehingga seluruh perusahaan memiliki akses kedalam database yang handal
3. Adanya peningkatan akses data melalui navigasi web dan intranet.
4. Tekanan kompetisi bisnis untuk meningkatkan penguasaan pasar dalam globalisasi ekonomi.
5. Perkembangan teknologi perangkat lunak untuk data mining (ketersediaan teknologi).
6. Perkembangan yang hebat dalam kemampuan komputasi dan pengembangan kapasitas media penyimpanan.

Dari definisi-definisi yang telah disampaikan, hal penting yang terkait dengan data mining adalah:

1. Data mining merupakan suatu proses otomatis terhadap data yang sudah ada.
2. Data yang akan diproses berupa data yang sangat besar.
3. Tujuan data mining adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat.

Untuk melakukan data Mining, ada beberapa tahapan. Tahap-tahap tersebut bersifat interaktif di mana pemakai terlibat langsung atau dengan perantaraan knowledge base atau bisa disebut Data Preprocessing.

Data Preprocecssing merupakan sekumpulan Teknik yang diterapkan pada database untuk menghapus *noise*, *missing value*, dan data yang tidak konsisten. Data preprocessing dibagi menjadi beberapa langkah, yaitu cleaning data, data transformasi, integrasi data, dan data reduction. Data preprocessing ini digunakan karena dalam *data realtime database* seringkali tidak lengkap dan tidak konsisten sehingga mengakibatkan hasil data mining tidak tepat dan kurang akurat. Oleh karena itu, untuk meningkatkan kualitas data yang akan di analisis, perlu dilakukan langkah-langkah preprocessing data. Berikut langkah-langkah dari data preprocessing: [6]

a. Cleaning data

   Pada umumnya data yang diperoleh dari perusahaan memiliki data yang tidak sempurna seperti data yang hilang, data yang tidak valid. Sebaiknya data-data yang tersebut lebih baik dibuang karena keberadaannya dapat mengurangi mutu atau akurasi dari hasil data mining nantinya.

b. Integrasi data

   Tidak jarang data yang diperlukan untuk data mining tidak hanya berasal dari satu database tetapi juga berasal dari beberapa database atau file teks. Integrasi data perlu dilakukan secara hati hati dikarenakan kesalahan pada integrasi data dapat terjadi penyimpangan pada data keluaran proses data mining.

c. Transformasi data

   Beberapa teknik data mining membutuhkan format data yang khusus sebelum bisa diaplikasikan. Disini juga dilakukan pemilihan data yang diperlukan oleh teknik data mining yang dipakai. Transformasi dan pemilihan data ini juga

menentukan kualitas dari hasil data mining nantinya karena ada beberapa karakteristik dari teknik-teknik data mining tertentu yang tergantung pada tahapan ini.

d. Data Reduction

   Analisis data yang menggunakan dataset dalam ukuran besar akan sangat sulit dilakukan, oleh karena itu, perlu adanya Teknik data reduction dengan tujuan untuk meningkatkan efisiensi penyimpanan serta mengurangi biaya penyimpanan dan analisis data. Data reduction dibagi menjadi beberapa Teknik, yaitu salah satunya *Attribute Subset Selection*.

e. Aplikasi teknik data mining

   Aplikasi teknik data mining sendiri hanya merupakan salah satu bagian dari proses data mining. Gunakan teknik data mining yang sesuai dengan hasil yang diinginkan.

f. Evaluasi pola yang ditemukan

   Dalam tahap ini hasil dari teknik data mining berupa polapola yang khas maupun model prediksi dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai hipotesa ada beberapa alternatif yang dapat diambil seperti: menjadikannya umpan balik untuk memperbaiki proses data mining, mencoba teknik data mining lain yang lebih sesuai, atau menerima hasil ini sebagai suatu hasil yang di luar dugaan yang mungkin bermanfaat.

g. Presentasi pola

   Tahap terakhir dari proses data mining adalah bagaimana memformulasikan keputusan atau aksi dari hasil analisa yang didapat. Dalam presentasi ini, visualisasi juga bisa membantu mengkomunikasikan hasil data mining.

## 2.2 Rapid Miner

Rapid Miner adalah platfrom perangkat lunak yang dikembangkan oleh perusahaan dengan nama yang sama yang menyediakan lingkungan terpadu untuk ilmu data. Platform ini digunakan untuk pebelajaran mesin, termasuk pembelajaran mendalam, penambangan test dan analitik prediktif. Rapid Miner digunakan dalam berbagai konteks, baik dalam aplikasi bisnis dan komersial, maupun dalam

penelitian, pendidikan, pelatihan, serta pembuatan prototipe dan pengembangan aplikasi. Platform ini mendukung semua fase proses *machine learning*, termasuk persiapan data, visualisasi hasil validasi dan pengoptimalan. [5].

### 2.3 Algoritma Apriori

Algoritma Apriori merupakan suatu metode yang mengikuti aturan yang menghubungkan beberapa atribut atau dikenal juga sebagai analisis afinitas [7]. Analisis keranjang belanja (*market basket analisis*) adalah suatu metode analisis yang mempelajari perilaku konsumen dalam suatu kelompok tertentu dengan fokus yang spesifik. Data yang digunakan dalam analisis market basket umumnya digunakan sebagai awal dalam mencari pengetahuan dari data transaksi Ketika pola-pola spesifik belum diketahui. *Analisis market basket* diperlukan karena keakuratan dan manfaat yang dihasilkan, terutama dalam bentuk aturan asosiasi (*association rules*), yang merupakan pola keterkaitan dalam basis data [8].

Altgoritma Apriori adalah salah satu algoritma yang terkenal dalam menemukan pola frekuensi tinggi. Pola frekuensi tinggi merujuk pada pola item yang muncul secara sering dalam suatu database dengan frekuensi atau dukungan di atas ambang batas tertentu yang disebut sebagai *minimum support*. Setelah proses penemuan pola frekuensi tinggi, langkah selanjutnya melibatkan pembentukan aturan asosiasi [1].

Penerapan data mining dengan aturan asosiasi bertujuan menemukan informasi item-item yang saling terhubung dalam bentuk aturan atau rule. Aturan asosiasi adalah teknik data mining untuk menemukan aturan asosiasi suatu kombinasi item [1].

#### 2.3.1 Analisis Pola Frekuensi Tinggi Dengan Algoritma Apriori

Mencari kombinasi item yang memenuhi syarat minimum dari nilai support dalam basis data. Nilai support sebuah item diperoleh dengan menggunakan rumus berikut:

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

Sedangkan nilai support dari 2 item diperoleh dari rumus berikut:

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

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

### 2.3.2 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 ∪ B.

Nilai *confidence* dari aturan A ∪ B diperoleh dengan rumus berikut.

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

Dalam menentukan aturan asosiasi yang akan dipilih, langkahnya adalah dengan mengurutkannya berdasarkan nilai *support x confidence*. Aturan tersebut kemudian diambil dalam jumlah *n* aturan dengan hasil terbesar.

## 2.4 Association Rules

Aturan asosiasi, yang juga dikenal sebagai association rules, merupakan salah satu Teknik dalam data mining yang digunakan untuk mencari hubungan asosiatif antara item atau atribut dalam dataset. Aturan asosiasi dibentuk melalui analisis pola data yang sering muncul atau terjadi bersamaan, yang disebut sebagai pola frekuensi tinggi (*frequent pattern*), dengan menggunakan parameter *support* dan *confidence* untuk mengidentifikasi hubungan yang paling penting. *Support* mengindikasikan seberapa sering suatu item muncul dalam database. Sementara itu, *confidence* mengindikasikan seberapa sering pernyataan tersebut benar [3].

$$\text{Support} = P (X \cap Y) = \sum \frac{transaksi\ mengandung\ X\ dan\ Y}{Jumlah\ Transaksi}$$

$$\text{Confidence} = P (Y/X) = \frac{P (X \cap Y)}{\sum \quad Transaksi\ yang\ mengandung\ X}$$

### 2.4.1 Pattern Evaluation

Pada tahap ini dilakukan penetapan tren nilai data mining, dengan uji *lift ratio*. *Support* merupakan presentase kombinasi item yang berada pada database dan *confidance* dari kuatnya hubungan antar item maka *lift* adalah nilai yang menunjukan kevalidan proses transaksi dan memberikan informasi apakah benar item A dibeli bersamaaan dengan item B [9].

Sedangkan untuk nilai ketetapan dari *lift* adalah sebagai berikut:

a. Rule dikatakan valid apabila nilai *lift* > 1 yang artinya item A dan B dibeli secara bersamaan.
b. Rule dikatakan independent apabila nilai *lift* = 1.
c. Rule dengan nilai *lift* < 1 maka antar item tersebut tidak saling berkaitan antar *antecedent* (item A) dengan *consequent* (item B).

### 2.4.2 Tahapan Pada Apriori yang Dilakukan

Sebelum menggunakan bantuan tools Rapid Miner, terdapat beberapa langkah yang perlu dilakukan pada tahap Apriori, yaitu:

a. Memastikan bahwa data yang diperlukan berbentuk tabular dan dapat diakses dalam format excel.
b. Melakukan penyaringan data atau pembersihan data untuk menghilangkan duplikat, *noise*, atau nilai kosong yang tidak relevan.
c. Menentukan nilai minimum support dan confidence yang akan digunakan dalam analisis. Nilai ini akan menjadi parameter untuk mengidentifikasi aturan asosiasi yang signifikan.
d. Menampilkan hasil aturan asosiasi yang telah ditemukan setelah proses analisis. Hasil ini akan menunjukkan hubungan dan pola yang signifikan antara item atau atribut dalam dataset.

Setelah langkah-langkah tersebut dilakukan, barulah pengguna dapat menggunakan tools Rapid Miner untuk melaksanakan analisis dan menampilkan hasil aturan asosiasi yang relevan [5].

## 2.5 Penelitian Terdahulu

Penelitian sebelumnya yang telah dilakukan menjadi referensi bagi penulis dalam melaksanakan penelitian ini, dengan tujuan untuk memperkaya teori yang digunakan dalam mengkaji topik yang diteliti. Meskipun penulis tidak menemukan penelitian sebelumnya dengan judul yang sama, namun beberapa penelitian terdahulu telah digunakan sebagai referensi untuk memperluas pemahaman pada penelitian ini. Berikut ini adalah beberapa jurnal terkait yang menjadi referensi dalam penelitian ini.

Tabel 2.1 Penelitian Terdahulu

| No | Nama Peneliti | Judul Peneliti | Hasil Penelitian |
|---|---|---|---|
| **1** | Oscar Devinsen, 2022 | Pemanfaatan Data Tracer Study Dalam Membangun Business Intellegence Di Iib Darmajaya [10] | Penelitian ini menggunakan data dari tracer study yang mencakup informasi tentang tahun lulus, jurusan, pekerjaan, gaji, ketetapan kelulusan, IPK, jenis kelamin, dan bidang kerja alumni. Dalam implementasi tracer study, metode yang digunakan meliputi survei dan pengisian kuesioner melalui link google form oleh para responden, yang dilakukan oleh tim surveyor. |
| **2** | Indriyawati Henry, Winarti Titin, 2021. | Pemodelan Data Mining Pola Kelayakan Kemampuan Lulusan Dengan Kebutuhan Stakeholder Menggunakan Algoritma apriori [1] | Algoritma apriori pada perhitungan data mining dengan menggunakan data tracer study Universitas Semarang batasan dari minimum support adalah sebesar 50% dan minimum |

| | | | |
|---|---|---|---|
| | | | confidence nya adalah 100% sehingga membentuk 4 rules. |
| **3** | Indah Puji Astuti, 2019. | Algoritma Apriori Untuk Menemukan Hubungan Antara Jurusan Sekolah Dengan Tingkat Kelulusan Mahasiswa [11] | Pada penelitian ini penulis menggunakan algoritma apriori untuk menemukan hubungan antara jurusan yang diambil waktu sekolah tingkat SMA dengan tingkat kelulusan mahasiswa. Yang diukur berdasarkan lama studi dan IPK dengan menggunkan software Tanagra. |
| **4** | Irham Kurnawan, Fitri Marisa, Dwi Purnomo, 2018 | Implementasi data mining dengan algoritma apriori untuk memprediksi tingkat kelulusan mahasiswa [2] | Dalam hal ini menggunakan metode asosiasi dan algoritma Apriori. Metode ini menghitung nilai support yang ada nilai penunjang suatu butir dengan aturan emas besar 60% dari data nilai mata kuliah. Itu hasil penelitian ini membantu perguruan tinggi meningkatkan mutu pendidikan dan membantu dalam mengetahui informasi tentang tingkat kelulusan siswa berdasarkan nilai mata pelajaran dan indeks prestasi yang diperoleh mahasiswa program studi Teknik Informatika Universitas Widyagama Malang. |

| | | | |
|---|---|---|---|
| **5** | .Saputro, R. A. (2020). | Penerapan Association Rule Dengan Algoritma Apriori Untuk Menampilkan Informasi Tingkat kelulusan Mahasiswa Teknik Informatika S1 Fakultas Ilmu Komputer Universitas Dian Nuswantoro [12] | Dengan memanfaatkan data kelulusan dan data induk dari mahasiswa UDINUS diharapkan dapat menghasilkan informasi tentang tingkat kelulusan mahasiswa melalui teknik data mining. Kategori tingkat kelulusan di ukur dari lama studi dan IPK. Algoritma yang digunakan adalah algoritma Apriori, informasi yang ditampilkan berupa nilai support dan confidence dari masing - masing kategori tingkat kelulusan. |

Dari beberapa penelitian sebelumnya yang disajikan dalam tabel 1. Perbedaan yang mendasar dengan penelitian ini adalah terletak pada keterkaitan hubungan antara atributnya, pada penelitian ini atribut yang digunakan untuk menemukan pola dengan tingkat kelulusan yang dilihat berdasarkan NPM, Tahun Lulus, Prodi, IPK, Jenis kelamin, Lama studi, Status kelulusan, dan Umur.