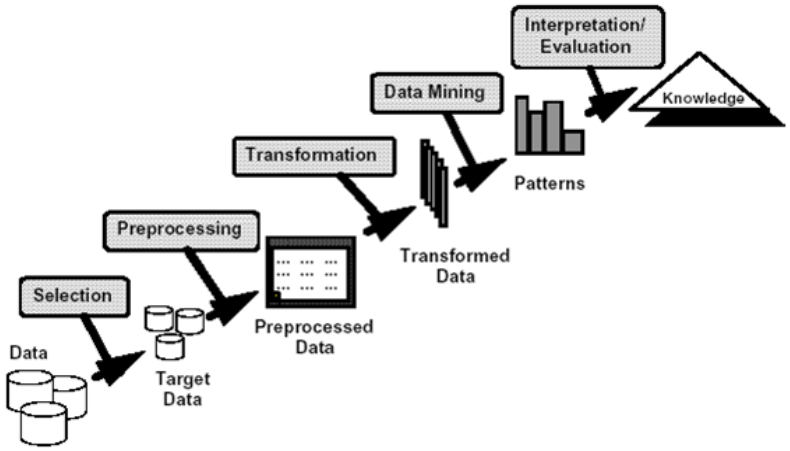
BAB II

LANDASAN TEORI

2.1 *Knowledge Discovery In Database (KDD)*

Menurut (retno tri vulandari,s.si., 2017:2) *Data Mining* adalah salah satu dari rangkaian ataupun tahapan dalam semua proses *knowledge discovery in database* (KDD). KDD memiliki hubungan dengan teknik dan penemuan ilmiah, penafsiran, dan gagasan dari pola-pola sejumlah koleksi data. Pada langkah ini, KDD menjelaskan pengetahuan dalam bentuk yang mudah dan dimengerti oleh pengguna. KDD adalah kegiatan yang melakukan pengumpulan data, mengolah data untuk mendapatkan hubungan atau pola yang teratur di dalam *database* dengan jumlah yang besar.



Gambar 2.1 Ilustrasi Proses Tahapan KDD

1. *Data selection* (seleksi data)

Penyeleksian data dari sekumpulan data yang ada pada *database* karena tidak semuanya dipakai dalam olah data. Data yang sesuai dari hasil penyeleksian akan disimpan untuk dianalisis yang akan diolah dan terpisah dari data-data yang yang tidak digunakan. Atau membuat data target dimana memfokuskan pada variabel atau sampel data.

2. *Preprocessing* (pembersihan data)

Sebelum proses *Data Mining* dilakukan, maka proses pembersihan data dilakukan pada data yang akan menjadi fokus KDD. Proses tersebut meliputi membuang data yang sama atau ganda, memeriksa data yang tidak konsisten, memperbaiki kesalahan pada data, misalnya kesalahan cetak,

3. *Transformation* (transformasi)

Merupakan proses transformasi atau perubahan pada data yang telah diseleksi, sehingga data tersebut sesuai dengan proses yang ada pada *Data Mining*. Proses ini sangat kreatif dan bergantung pada pola atau jenis informasi yang akan diambil dari *database*. Pada proses *transformation* ini berfungsi untuk memaparkan atau mempresentasikan data berdasarkan hasil yang akan dicapai dan didapat.

4. *Data Mining*

Pada proses *Data Mining* akan dilakukan pemilihan algoritma yang akan dipakai, pemilihan hasil dari proses KDD seperti klastering, klasifikasi, regresi. Pencarian pola atau informasi yang menarik dari data yang terseleksi dengan menggunakan teknik atau metode yang diambil. Metode atau teknik

pada *Data Mining* ada berbagai macam, sehingga pemilihan metode yang akan digunakan akan bergantung pada proses KDD.

5. *Interpretation* (evaluasi)

Pada proses ini, KDD akan melakukan pemeriksaan secara keseluruhan apakah informasi atau data yang di dapat berbanding terbalik dengan fakta atau hipotesa sebelumnya. Informasi atau pola yang ditemukan akan ditampilkan dalam bentuk yang mudah dimengerti dan dipahami oleh pengguna.

2.2 *Data Mining*

2.2.1 Definisi *Data Mining*

Data Mining adalah istilah yang digunakan untuk melakukan penemuan pengetahun baru di dalam *database*. *Data Mining* adalah peninjauan sekumpulan data untuk menemukan pola yang tidak diketahui dan merangkum data dengan teknik yang berbeda dan dapat dipahami dan berguna bagi pemilik data (Irfan, 2015). *Data Mining* diartikan sebagai suatu proses untuk mencari atau menemukan hubungan pola dan tren baru yang berguna dengan *filtering* data dengan skala besar yang tersimpan dalam database (Kamagi & Hansun, 2014).

Beberapa ahli juga memberikan pengertian tentang *Data Mining* yang dimuat dalam buku yang ditulisnya diantaranya;

Menurut (Dr.Suyanto, S.T., 2017:1) *Data Mining* adalah kumpulan sejumlah ilmu komputer yang diartikan sebagai proses menemukan pola-pola

baru dari kumpulan data yang besar meliputi metode-metode yang merupakan bagian dari kecerdasan buatan, mesin pembelajaran, statistik dan database. *Data Mining* adalah proses yang menggunakan teknik *statistic, mathematic ,artificial intelligence*, dan *machine learning* untuk mengetahui informasi yang berguna dan pengetahuan yang berhubungan dari *database* besar (kusrini & emha taufiq luthfi, 2009:3).

Data Mining adalah serangkaian proses atau langkah untuk menemukan nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu database (Retno Tri Vulandari,S.Si., 2017:1). Berdasarkan pengertian *Data Mining* yang diuraikan diatas maka dapat ditarik kesimpulan bahwa *Data Mining* merupakan proses penemuan pola yang baru dan bermakna dari sejumlah data yang besar didalam *database* dan bermanfaat bagi pemilik data.

2.2.2 Manfaat *Data Mining*

Menurut (Dr.suyanto, S.T., 2017:3) kegunaan *Data Mining* dibedakan menjadi dua yaitu prediktif dan deskriptif. Prediktif adalah *Data Mining* yang digunakan untuk membentuk sebuah model pengetahuan untuk melakukan proses prediksi. Sedangkan deskriptif adalah *Data Mining* yang digunakan untuk menemukan pola-pola yang dapat dimengerti manusia yang menjelaskan ciri-ciri data. Berdasarkan fungsinya, tugas-tugas *Data Mining* kedalam 6 kelompok berikut ini:

a. Klasifikasi (*classification*): menjabarkan struktur yang sudah diketahui untuk diterapkan ke dalam data-data baru

b. Klasterisasi (*clustering*): pengelompokan data yang tidak diketahui kelasnya kedalam sejumlah kelompok tertentu sesuai dengan kemiripannya.

c. Regresi (*regression):* menemukan suatu fungsi yang memodelkan data dengan alat seminimal mungkin.

d. Deteksi anomali (*anomaly detection):* mengidentifikasi data yang tidak biasa, perubahan yang mungkin sangat penting dan perlu dianalisa selanjutnya.

e. Pembelajaran aturan asosiasi (*association rule learning*) atau pemodelan kebergantungan (*dependency modeling*): mencari hubbungan antar variabel.

f. Perangkuman *(summarization*): menyediakan representasi data yang lebih sederhana, meliputi pembuatan laporan.

2.2.3 Pengelompokan *Data Mining*

Menurut (Kusrini & Emha Taufiq Luthfi, 2009:10) *Data Mining* dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan adalah:

1. Deskripsi

Kadang peneliti menganalisis secara sederhana ingin mencoba mencari cara untuk mengambarkan atau mendeskripsikan pola dan kemiripan yang terdapat dalam data.

2. Estimasi

Estimasi bisa dikatakan atau sama dengan *clasification*, kecuali variabel target estimasi lebih kearah numerik daripada kearah kategori. Model dibangun menggunakan laporan lengkap yang menyedikan nilai dari variabel target sebagai nilai prediksi.

3. Prediksi

Prediksi bisa juga dikatakan sama dengan klasifkasi dan estimasi, kecuali bahwa dalam prediksi nilai dari hasil akan ada di masa mendatang.

4. Klasifikasi

Didalam klasifikasi terdapat target variabel kategori. Misalnya pengklasifikasian pendapatan berdasarkan levelnya yaitu, pendapatan rendah, pendapatan sedang dan pendapatan tinggi.

5. Pengklusteran

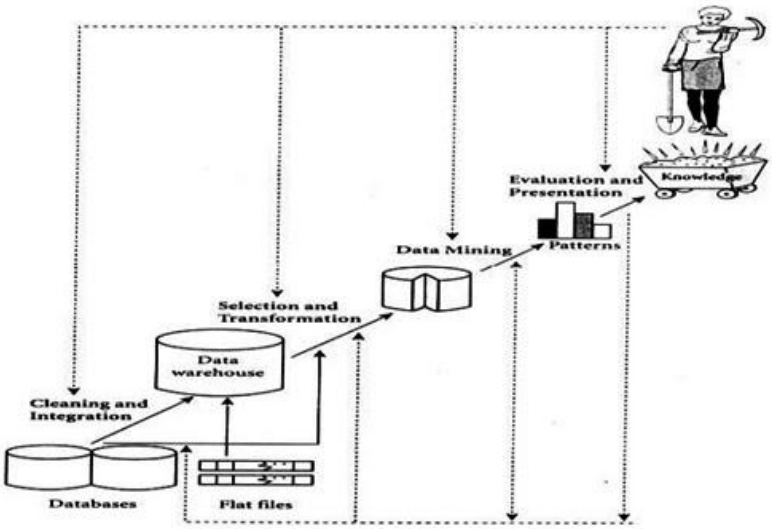
Kluster adalah sekumpulan laporan yang mempunyai kesamaaan antara yang satu dengan yang lainnya dan mempunyai ketidaksamaan dengan *record* dari kluster lain. Pengklusteran adalah pengelompokan *record* observasi atau memperhatikan dan membentuk kelas objek yang mempunyai kesamaan. Pengklusteran berbeda dengan klasifikasi yaitu tidak adanya variabel target akan tetapi pengklusteran melakukan pembagian terhadap seluruh data menjadi kelompok yang mempunyai kemiripan.

6. Asosiasi

Tugas asosiasi dalam *Data Mining* adalah menemukan atribut yang muncul dalam waktu yang bersamaan.

2.2.4 Tahapan *Data Mining*

Sebagai salah satu dari rangkaian proses, *Data Mining* dibagi menjadi beberapa tahapan proses yang akan dilakukan. Tahap-tahap tersebut bersifat saling berhubungan, pengguna ikut berpartisipasi langsung atau dengan perantara *knowledge base.*



Gambar 2.2 Gambaran Tahap-Tahap *Data Mining*.

Tahap-tahap *Data Mining* adalah sebagai berikut;

1. Pembersihan data (*data cleaning*)

   Pembersihan data adalah proses menghilangkan kotoran (*noise*) dan data yang tidak konsisten dan tidak relevan. Dalam *Data Mining* data yang tidak dibersihkan lebih dulu akan memberikan hasil kurang baik.

2. Integrasi Data (*Data Integration*)

   Perpaduan data dari beberapa *database* ke dalam *database* yang baru.

3. Seleksi Data (*Data Selection*)

Data yang ada pada *database* tidak semuanya digunakan hanya data yang akan dianalisis yang akan diambil dari *database* sehingga dilakukan seleksi data.

4. Transformasi Data (*Data Transformation*)

Data digabung atau diubah menjadi data yang memiliki format yang sesuai untuk diproses di *Data Mining*.

5. Proses Mining

Merupakan suatu proses utama saat metode-metode yang ada diterapkan untuk menemukan informasi berharga dan yang tidak diketahui sebelumnya dari data.

6. Evaluasi Pola *(Pattern*)

Untuk menemukan atau mengidentifikasi hubungan atau pola-pola yang menarik kedalam *knowledge base* yang dihasilkan.

7. Presentasi Pengetahuan (*Knowledge Presentation*)

Pemaparan pengetahuan tentang metode-metode yang digunakan untuk memperoleh informasi atau pengetahuan yang diperoleh oleh pengguna.

2.3 SNMPTN

SNMPTN (Seleksi Nasional Masuk Perguruan Tinggi Negeri) adalah salah satu cara atau jalur masuk ke PTN tanpa test dan murni berdasarkan hasil nilai raport siswa dalam 5 semester. Sekolah sebagai satuan pendidikan dan guru sebagai pendidik diharapkan selalu menjunjung tinggi kehormatan dan kejujuran dalam pendidikan. Oleh karena itu, sekolah berkewajiban

mengisi Pangkalan Data Sekolah dan Siswa (PDSS) yang berisikan rekam jejak kinerja sekolah dan prestasi akademik siswa dengan lengkap dan benar.. Siswa pendaftar yang berhak mengikuti SNMPTN adalah:

a. Siswa SMA/MA/SMK kelas 12 pada tahun 2019 yang memiliki prestasi
b. Memiliki NISN yang terdaftar di PDSS
c. Memiliki nilai rapor semester 1 s.d. 5 yang telah diisikan oleh sekolah.
d. Memiliki prestasi akademik.

2.4 Metode Naïve Bayes

Dalam proses *Data Mining* ada beberapa metode yang dipakai. Dalam penelitian ini, peneliti menggunakan metode naïve bayes, yang ditemukan oleh Thomas Bayes di abad ke-18. Menurut (kusrini & emha taufiq luthfi, 2009:189) bayes klasifikasi adalah pengklasifikasian yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas, terbukti memiliki akurasi dan kecepatan yang tinggi saat diiaplikasikan kedalam *database* dengan jumlah yang besar. Naïve bayes mengasumsikan bahwa nilai dari sebuah *input* tergantung dengan nilai atribut yang lain. Probabilitas atau peluang bersyarat dinyatakan sebagai:

$$P(H \mid X) = \frac{P(X|H)P(H)}{P(X)}$$ **Rumus 2.1** Naïve Bayes

Dimana:

X: Data dengan kelas yang belum diketahui

C: Hipotesis data X merupakan suatu kelas spesifik

P (C | X): Probabilitas hipotesis C berdasar kondisi X (probabilitas tidak lulus)

P(C): Probababilitas hipotesis C (probabilitas lulus)

P (X | C): Probablitas X berdasarkan kondisi pada Hipotesis C

P(X): Probabilitas X

Misalkan D adalah kumpulan *training set* yang berisi sejumlah tuple yang beridimension *n* dan dinyatakan sebgai X = X_1, X2. Xn yang didapat dari *n* atribut A_1, A2... An. Jika terdapat *m* kelas yaitu $C_{1,}$ $C_{2\ldots}C_{m.}$ untuk sebuah tuple masukan X, naïve bayes akan memprediksi bahwa tuple X termasuk kedalam kelas C*i* jika dan hanya jika:

$$P(C_i \mid X) > P(C_j \mid X)$$ **Rumus 2.2** Probabilitas

Jika berhadapan dengan jumlah data yang memiliki sangat banyak atribut, dapat mereduksi kompleksitas perhitungan dengan asusmsi naif yaitu: nilai-nilai atribut adalah saling independen. Maksudnya antar atribut adalah saling bebas tidak ada ketergantungan sama sekali. Untuk atribut yang berniali kategorik P (x_i | C_i) adalah jumlah tuple di kelas C_i dalam data yang memiliki nilai x_i pada atribut A_1 dibagi dengan jumlah semua kelas. Untuk menghitung probabilitas maka digunakan rumus:

$$P(X \mid C_i) / P(C_i)$$ **Rumus 2.3** Hitung Probabilitas

Dimana untuk probabilitas lulus dibagi dengan jumlah data dan probabilitas tidak lulus dibagi dengan jumlah data pada *training set data.*

2.6 Software Pendukung

Gambar 2.3 Logo Rapidminer

RapidMiner dikembangkan pada model inti terbukaRapid Miner adalah sebuah perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari *Institute of Technologi Blanchardstown* dan Ralf Klinkenberg dari rapid-i.com dengan tampilan GUI (*Graphical User Interface*) sehingga memudahkan pengguna dalam mengoperasikan perangkat lunak ini. Dengan menggunakan Rapid Miner, tidak dibutuhkan kemampuan koding khusus, karena semua fasilitas sudah disediakan. Rapid Miner dikhususkan untuk penggunaan *Data Mining*. Banyak metode yang disediakan oleh Rapid Miner mulai dari klasifikasi, klastering, asosiasi dan lain-lain.

2.7 Penelitian Terdahulu

Dalam penelitian ini, peneliti menggunakan beberapa jurnal dari penelitian terdahulu tentang kelulusan menggunakan *Data Mining* sebagai sumber referensi, yaitu:

1. menurut penelitian budanis Dwi Meiliani Dan Nofi Susanto dengan judul **"Aplikasi *Data Mining* Untuk Menghasilkan Pola Kelulusan Siswa Dengan Metode Naïve Bayes"** dengan memanfaatkan data induk siswa dan data kelulusan siswa sebagai sumber datanya,

diharapkan dapat menghasilkan informasi tentang pola tingkat kelulusan siswa melalui teknik *Data Mining*. Kategori kelulusan diukur dari nilai UNAS. Algoritma yang digunakan adalah naïve bayes. Pada aplikasi ini menghasilkan pola data kelulusan siswa sebelumnya berdasarakan atribut yang diujikan dan pola data baru yang dujikan berdasarkan data yang telah ada.

2. Menurut penelitian Langgeng Listiyoko, Rosella Wati, Achmad Farudin dengan judul **"Klasifikasi Siswa Untuk Meningkatkan Nilai Rata-Rata Kelas Menggunakan Metode *Data Mining*.** Salah satu usaha yang dapat ditempuh untuk meningkatkan nilai rata-rata kelas adalah dengan mengelompokkan siswa dengan kesamaan tertentu. Kelompok yang terbentuk diharapakan memiliki potensi yang relative sama. Dengan bantuan *Data Mining clustering* maka akan didapat paket kelompok siswa dengan kemampuan yang setara sehingga dapat diberikan penangan yang terarah.
3. Menurut penelitian Diasrina Dahri, Fahrul Agus, Dyna Marisa Khairina dengan judul **"Metode Naïve Bayes Untuk Penentuan Penerima Beasiswa Bidikmisi Universitas Mulawarman.** Ketidakkonsistenan pada sistem penentuan penerima menyebabkan tujuan penyelenggara menjadi kabur, tidak transparan dan tidak tepat sasaran. Penelitian ini bertujuan untuk membantu bagiaan proses seleksi dengan membuat aplikasi perangkat lunak sistem pendukung keputusan untuk penentuan penerima beasiswa bidikmisi Universitas Mulawarman.

Penentuan penerima beasiswa menggunakan beberapa kriteria antara lain: pekerjaaan orang tua, penghasilan orang tua, jumlah tanggungan, daya listrik (*watt*) dan nilai ujian nasional. Kelayakan calon penerima beasiswa bidikmisi ditentukan dengan menerapkan metode naïve bayes. Penelitian ini telah menghasilkaan apliksi sistem pendukung keputusan dengan tingkat akurasi 85.56%.

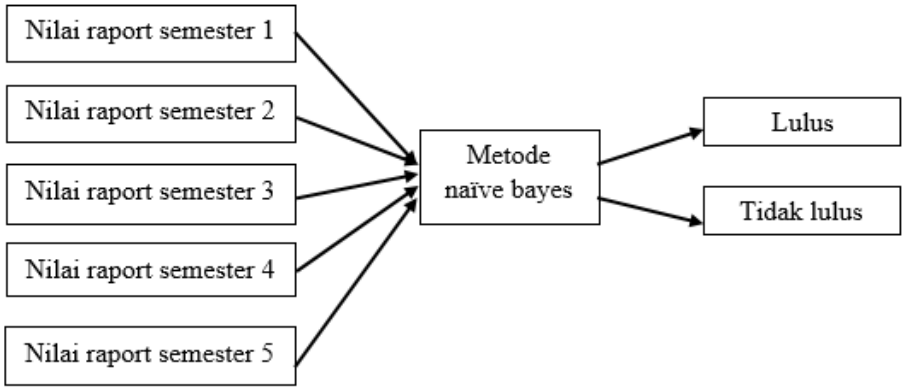
4. Menurut penelitian Windania Purba, Saut Tamba, Jepronel Saragih dengan judul **"*The Effect of Mining Data K-Means Clustering toward Students Profile Model Drop out Potensial*.** Tingginya keberhasilan dan rendahnya kegagalan mahasiswa dapat mencerminkan kualitas sebuah perguruan tinggi. Salah satu alasan mahasiswa gagal adalah berhenti kuliah. Untuk mengatasi masalah tersebut, maka diterapkann *Data Mining* dengan *K-means clustering*. Metode *K-means clustering* diimplementasikan mengelompokkan mahasiswa yang berpotensial untuk berhenti kuliah. Pertama, data hasilnya akan dikelompokkan untuk didapat informasi dari semua kondisi mahasiswa. Berdasarkan model yang diambil ditemukan bahwa mahasiswa yang berpotensi putus kuliah karena mahasiswa yang kurang menarik dalam belajar, orang tua yang tidak mendukung, mahasiswa yang tidak percaya diri dan perilaku mahasiswa yang kurang. Hasil dari proses *K-Means Clustering* dapat diketahui bahwa mahasiswa yang lebih berpotensi putus kuliah berada di kelompok cicilan uang kuliah, Kualitas Total, dan Indeks Prestasi Kumulatif (IPK).

5. Menurut penelitian Tina R. Patil, Mrs.S.S. Sherekar yang berjudul" ***Performance Analysis of Naïve Bayes and J48 Classification Algorithm for Data Classification.*** Klasifikasi adalah suatu yang penting dalam teknik *Data Mining* untuk mengklasifikasikan berbagai Jenis-jenis data yang digunakan di hampir setiap bidang kehidupan kita. Klasisfikasi digunakan untuk mengklasifikasikan laporan barang untuk fitur-fitur barang yang saling berhubungan dengan yang telah ditentukan sebelumnya dalam mengatur kelas-kelas. Penelitian ini menyoroti kinerja berdasarkan contoh yang benar dan yang salah menggunakan algoritma Naïve Bayes dan J48. Algoritma Naive Bayes berdasarkan pada probabilitas dan algoritma j48 berdasarkan pada keputusan pohon. Penelitian ini dibuat untuk melakukan pengujian komparatif. Pengklasifikasi naïve bayes dan J48 dalam konteks *dataset bank* untuk memaksimalkan nilai data yang benar dan meminimalkan nilai data yang salah dibandingkan mencapai keakuratan klasifikasi yang lebih tinggi menggunakan alat WEKA. Hasil dari penelitian ditunjukkan tentang akurasi klasifikasi, kepekaan dan kekhususan. Hasil dari penelitian ini juga menunjukkan bahwa efisiensi dan akurasi j48 lebih baik dibandingkan dengan naive bayes.

2.8 Kerangka Pemikiran

Untuk meningkatkan kelulusan siswa SNMPTN perlu dilakukan analisis data. Metode naïve bayes adalah metode yang sering digunakan untuk melakukann analisis data karena keakuratan dan kecepatan yang tinggi. Masuk SNMPTN yang diambil adalah nilai raport dari semester satu sampai semester lima, dalam penelitian nilai ini, nilai tersebut di seleksi dan diolah menggunakan aplikasi dengan metode naïve bayes dan menghasilkan *output* lulus dan tidak lulus

Dari latar belakang dan metode yang digunakan maka kerangka pemikiran dari penelitian adalah:



Gambar 2.4 Kerangka Pemikiran

Sumber: Penelitian (2018)