BAB II

TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1. Tinjauan Pustaka

Dalam proses pendataan dan penyimpanan data yang besar di suatu universitas dibutuhkan *database* untuk menyimpan semua data tersebut. Didalam *database* unversitas banyak data yang tersimpan mulai dari, karyawan, dosen, mahasiswa dan administrasi di universitas. Dari banyaknya data yang disimpan didalam database terkadang data-data yang berharga tersebut tidak dipergunakan secara optimal dan tidak dianalisis lebih dalam lagi.

Jananto, A. (2013) *Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa.* Dalam melakukan penelitian tersebut Jananto,A. (2013) menggunakan Query pada MySQL dalam menganilisis data menggunakan teknik *data mining*. Data yang dianalisis adalah nilai mahasiswa semester I sampai IV pada angkatan 2009 dan 2011, kota sekolah, kota lahir, jenis kelamin dan tipe sekolah. Dari hasil penelitian tersebut didapatkan 254 mahasiswa diprediksi "Tepat waktu" dan sisanya 4 orang diprediksi "tidak tepat waktu".

Ridwan, M., Suyono, H., & Sarosa, M. (2013). *Penerapan Data mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier.* Dalam penelitian tersebut Ridwan, M., Suyono, H., & Sarosa, M. (2013) menggunakan tool Rapidminer dalam menganilisis data menggunakan teknik *data mining*. Dalam penelitian ini, data yang dianalisis mahasiswa angkatan

2005 - 2009 yang sudah dinyatakan lulus akan digunakan sebagai data *traning* dan *testing.* Sedangkan mahasiswa angkatan 2010 - 2011 dan belum lulus akan digunakan sebagai target. Hasil pengujian faktor yang paling berpengaruh dalam kalsifikasi penenentuan kinerja akademik yaitu indeks Prestasi komulatif (IPK), Indeks Prestasi (IP) semester I sampai IV dan jenis kelamin.

Oenunu, D. M., Widyastuti, N., & Hamzah, A. (2015). PREDIKSI LAMA STUDI MAHASISWA DENGAN MENGGUNAKAN METODE K-NNPREDICTION OF STUDY TIMING PERIOD USING K-NNMETHOD. Data diambil dari *database* akademik mahasiswa jurusan Teknik Informatika S1, Institut Sains & Teknologi AKPRIND Yogyakarta (Tahun Akademik 2004 s/d 2009 yang telah lulus) sebanyak 216 mahasiswa dengan 34 atribut diantaranya Jenis Kelamin, Usia Saat Mendaftar, Agama, Asal Sekolah, Pekerjaan Orang Tua, IP Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, Jumlah SKS Semester 1, Jumlah SKS Semester 2, Jumlah SKS Semester 3, Jumlah SKS Semester 4, Nilai Setiap Mata Kuliah Semester 1, Nilai Setiap Mata Kuliah Semester 2, Nilai Setiap Mata Kuliah Semester 3, Nilai Setiap Mata Kuliah Semester 4. Menentukan prediksi lama studi mahasiswa dengan hasil prediksi sebagai berikut : Studi mahasiswa <5 Tahun dinyatakan “Sangat Tepat Waktu”, Studi mahasiswa > 5s/d6 Tahun dinyatakan “Tepat Waktu”, Studi mahasiswa >6 Tahun dinyatakan “Tidak Tepat Waktu.”

Dari penelitian yang dilakukan seperti diatas penulis ingin melakukan penelitian yang sama dengan menggunakan tools yang berbeda yaitu:

1. *Rapidminer*
2. *SQL Server 2012*

2.2. Landasan Teori

2.2.1. *Data Mining*

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam basis data. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data besar (Kusrini & Emha Taufiq Luthfi, 2009:25).

2.2.2. Pengelompokan *Data Mining*

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu (Larose dalam Kusrini, 2009):

1. Deskripsi

   Terkadang peneliti dan analis secara sederhana ingin mencoba mancari cara untuk menggambarkan pola dan kecenderungan yang terdapat dalam data. Deskripsi dari pola dan kecenderungan sering memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.

2. Estimasi

   Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih kearah numerik daripada kearah kategori. Model dibangun menggunakan record lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi. Selanjutnya pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilau variabel prediksi.

3. Prediksi

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

4. Klasifikasi

   Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, sedang dan rendah.

5. Pengklusteran

   Pengklusteran merupakan pengelompokkan 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 melakukan klasifikasi, estimasi atau memprediksi nilai dari variabel target. Akan tetapi, algoritma pengklusteran mencoba untuk melakukan pembagian terhadap keseluruhan data menjadi kelompok-kelompok yang memiliki

kemiripan (homogen), yang mana kemiripan record dalam satu kelompok akan bernilai maksimal, sedangkan kemiripan dengan record dalam kelompok lain akan bernilai minimal.

6. Asosiasi

   Tugas asosiasi adalah menemukan atribut yang muncul dalam satu waktu. Dalam dunia bisnis lebih umum disebut analisis keranjang belanja.

2.2.3 Tahapan-tahapan *Data Mining*

Dalam melakukan proses *mining* harus melewati melalui tahap-tahap seperti berikut ini:

1. Pembersihan data (*data cleaning*)

   Pembersihan data merupakan proses menghilangkan noise dan data yang tidak konsisten atau data tidak relevan.

2. Integrasi data (*data integration*)

   Intergrasi data merupakan penggabungan data dari berbagai *database* kedalam suatu *database* baru.

3. Seleksi data (*data selection*)

   Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database.*

4. Transfomasi data

   Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining.*

5. Proses Mining

   Merupakan suatu proses utama saat metode diterapkan untuk menemukan penegtahuan berharga dan tersembunyi dari data.

6. Evaluasi pola (*pattern evaluation*)

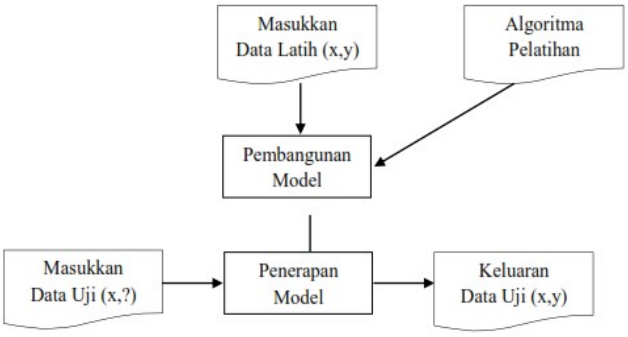
   Untuk mengidentifikasi pola-pola menarik ke dalam *knowledge based* yang ditemukan.

7. Presentasi pengetahuan (*knowledge presentation*)

   Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

2.2.4. Klasifikasi

Klasifikasi adalah proses pencarian sekumpulan model yang menggambarkan dan membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu obyek yang belum diketahui kelasnya (Herman Aldino, Naam, Julfriadif , 2012). Proses klasifikasi tersebut terlihat pada gambar 2.1.



Gambar 2. 1 Proses klasifikasi

2.2.5. Indeks Prestasi Komulatif (IPK)

IPK adalah angka yang menunjukkan prestasi atau keberhasilan studi mahasiswa dari semester pertama sampai dengan semester terakhir yang telah ditempuh secara kumulatif. IPK digunakan untuk:

a. Menentukan beban studi yang dapat diambil mahasiswa pada semester berikutnya.
b. Evaluasi Akademik per semester.
c. Evaluasi hasil studi pada akhir program.

2.2.6. Naive Bayes

Bayesian classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediski probabilitas keanggotaan suatu *class*. *Bayesian classification* didasarkan pada teorema *Bayes* yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network*. *Bayesian classification* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar. (Jananto, A. 2013). *Naive bayes* memiliki persamaan seperti berikut ini:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)}$$

Dimana:

X :Data dengan *class* yang belum diketahui

H : Hipotesis data merupakan suatu *class* spesifik

P(H|X) :Probabilitas hipotesis *H* berdasar kondisi *X* (posteriori probabilitas)

P(H) : Probabilitas hipotesis *H* (prior probabilitas)

P(X|H) :Probabilitas *X* berdasarkan kondisi pada hipotesis *H*

P(X) : Probabilitas *X*

Sedangkan untuk menghitung data yang bersifat kontinyu, maka menggunakan rumus *Densitas Gauss*:

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2 ij}}$$

Di mana :

P : Peluang

Xi : Atribut ke *i*

xi : Nilai atribut ke *i*

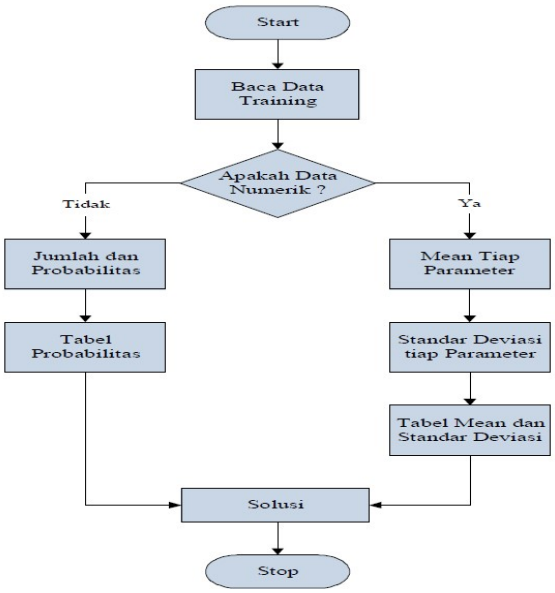
Y : Kelas yang dicari

yi : Sub kelas *Y* yang dicari

μ : *mean*, menyatakan rata – rata dari seluruh atribut

σ :Deviasi standar, menyatakan varian dari seluruh atribut.

Alur dari metode *Naive Bayes* dapat dilihat pada gambar 2.2 sebagai berikut:



Gambar 2. 2 Alur Metode *Naive Bayes*

Sumber: (Alfa Saleh, 2015)

Rumus yang digunakan untuk menghitung nilai rata-rata dari setiap data (*mean)* dapat dilihat sebagai berikut:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

atau

$$\mu = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

di mana :

μ : rata – rata hitung (*mean*)

x_i : nilai sample ke -*i*

n : jumlah sampel

Dan rumus untuk menghitung simpangan baku (standar Deviasi) dapat dilihat sebagai berikut:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}$$

di mana :

σ : standar deviasi

x_i : nilai x ke -*i*

μ : rata-rata hitung

n : jumlah sampel

2.2.7. *Rapidminer*

RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*). *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap *data mining*, *text mining* dan analisis prediksi. *RapidMiner* menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. *RapidMiner* memiliki kurang lebih 500 operator *data mining*, termasuk operator untuk *input, output*, *data preprocessing* dan visualisasi. *RapidMiner* merupakan *software* yang berdiri sendiri untuk analisis data dan sebagai mesin *data mining* yang dapat diintegrasikan pada produknya

sendiri. *RapidMiner* ditulis dengan munggunakan bahasa *java* sehingga dapat bekerja di semua sistem operasi (dkk Aprilla, C. Dennis.2013).

RapidMiner memiliki beberapa sifat sebagai berikut (dkk Aprilla, C. Dennis.2013):

- Ditulis dengan bahasa pemrograman *java* sehingga dapat dijalankan di berbagai sistem operasi.
- Proses penemuan pengetahuan dimodelkan sebagai *operator trees.*
- Representasi XML internal untuk memastikan format standar pertukaran data.
- Bahasa *scripting* memungkinkan untuk eksperiman skala besar dan otomatisasi eksperimen.
- Konsep *multi-layer* untuk menjamin tampilan data yang efisien dan menjamin penangan data.
- Memiliki GUI, *command line mode* dan *Java API* yang dapat dipanggil dari program lain.

Beberapa fitur dari *RapidMiner*, antara lain [5]:

- Banyaknya algoritma *data mining*, seperti *decision tree* dan *self-organization map.*
- Bentuk grafis yang canggih, seperti tumbang tindih *diagram histogram*, *tree chart* dan 3D *scatter plots.*
- Banyaknya variasi *plugin*, seperti *text plugin* untuk melakukan analisis teks.

- Menyediakan prosedur *data mining* dan *machine learning* termasuk: ETL (*extraction, transformation, loading) data preprocessing*, visualisasi, modeling dan evalualisasi.
- Proses *data mining* tersusun atas operator-operator yang *nestable*, dideskripsikan dengan XML, dan dibuat dengan GUI.
- Mengintegrasikan proyek *data mining* Weka dan statistic R.

2.2.8. Microsoft SQL

SQL Server merupakan *Relational Database Management System* (RDMS) yang menghubungkan pengguna dengan data untuk pengelolaan basis data. SQL Server dapat digunakan untuk menghubungkan satu ataupun beberapa server. Bahasa basis data yang digunakan *SQL Server* adalah *Transact-SQL*. *Transact-SQL* merupakan bahasa *SQL* yang dimiliki oleh *SQL Server* yang berguna bagi pengguna untuk mendapatkan satu atau kumpulan data pada basis data dengan cara menjalankan perintah dari suatu pernyataan *SQL* (Suhadi, 2016).

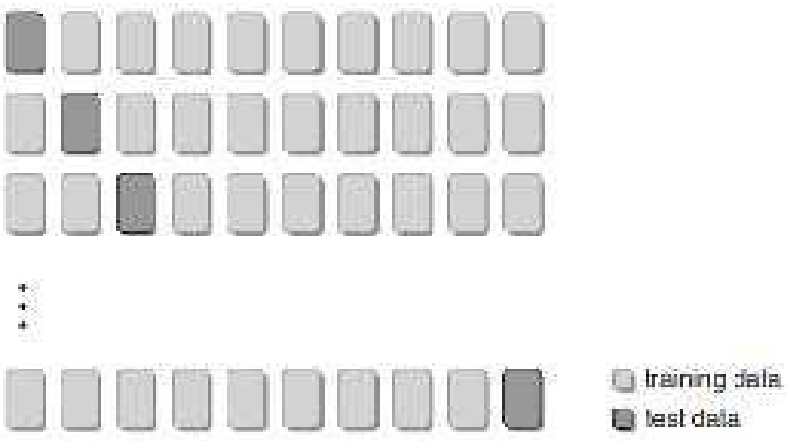
2.2.9. Microsoft Excel

Microsoft excel adalah *software spreadsheet* paling terkenal di dunia bisnis dan perkantoran. *Excel* digunakan hampir semua bidang bisnis. *Excel* dapat dijumpai di mana-mana dan bisa dikatakan sebagai aplikasi yang *universal* dan dipakai semua orang. Aplikasi *excel* memiliki fitur kalkulasi dan pembuatan grafik, serta mudah dipakai sehingga *excel* menjadi salah satu program komputer yang populer digunakan di PC hingga saat ini. Bahkan, saat ini *excel* merupakan program

spreadsheet paling banyak digunakan, baik *platform* PC berbasis *windows* maupun *platform macintosh* berbasis *Mac OS* semenjak versi 5.0 yang keluar di tahun 1993.

2.2.10. *Cross Validation*

Cross Validation merupakan salah satu teknik untuk menilai atau memvalidasi keakuratan sebuah model yang dibangun berdasarkan *dataset* tertentu. Pembuatan model biasanya bertujuan untuk melakukan prediksi maupun klasifikasi terhadap suatu data baru yang boleh jadi belum pernah muncul di dalam *dataset*. Data yang digunakan dalam proses pembangunan model disebut data latih atau *training*, sedangkan data yang akan digunakan untuk memvalidasi model disebut sebagai data *test*. *Cross Validation* mempunyai banyak model, diantaranya adalah *k-fold cross validation* dan *leave-one-out validation validation.* Biasanya, 10 *fold cross validation* direkomendasikan untuk akurasi prdiksi karena relatif rendah nilai bias dan varian (J.Han, M. Kamber and J. Pci, 2012).



Gambar 2. 3 *K-fold Validation*