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* univeristas banyak data yang tersimpan mulai dari karyawan, dosen, mahasiswa, administrasi di unversitas, dan data alumni unniversitas tersebut. Dari banyak data yang disimpan didalam database terkadang data-data yang berharga terbut tidak digunakan secara optimal dan tidak dianalisis lebih dalam lagi.

Menurut Penelitian (Firdaus, Putra, & Rosa, 2013) Analisis Business Intelligence terhadap Pengelohaan Data Alumni: Upaya Mendukung Monitoring Kualitas Alumni di Perguruan Tinggi (Studi Kasus di Fakultas Ilmu Komputer Universitas Sriwijaya). Jurnal Generic, 8(2), 221-229. Metode yang digunakan adalah Penerapan Electronic Business Intelligence System (E-BIS). Hasil akhirnya adalah proses data model sudah sangat baik untuk melanjutkan ke tahap perancangan. Analisis BI pada Fasilkom Unsri menggunakan business intelligence roadmap mencakup fase justification, planning, dan business analysis mengusulkan solusi BI yang dapat memenuhi kebutuhan informasi pihak eksekutif untuk monitoring data.

Penelitian (Nurjoko & Kurniawan, 2016) Aplikasi Datamining Untuk Memprediksi Tingkat Kelulusan Mahasiswa Menggunkan Algoritma APRIORI Di Ibi Darmajaya Bandar Lampung; Dengan memanfaatkan data induk mahasiswa dan data kelulusan mahasiswa, dapat diharapkan menghasilkan informasi tentang tingkat kelulusan dan data induk 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.

Menurut (Jananto, 2013), Penelitian tentang Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa dengan menggunakan teknik data mining khususnya klasifikasi untuk memprediksi dengan menggunakan algoritma naive bayes dilakukan terhadap ketepatan waktu studi dari mahasiswa berdasarkan data training yang ada. Algoritma naive bayes, menghitung perbandingan peluang antara jumlah dari masing-

masing kriteria nilai fields terhadap nilai hasil prediksi sesunggunya. Tinggi rendahnya tingkat kesalahan dapat disebabkan oleh jumlah record data dan tingkat konsistensi dari data training yang dgunakan. Sedangkan hasil prediksi dari ketepatan lama studi dari mahasiswa angkatan 2008 adalah sebesar 254 mahasiswa diprediksi ”Tepat Waktu” dan hanya 4 orang diprediksi ”Tidak Tepat Waktu”.

Menurut Penelitian (Eka Sabna and Muhardi, 2016), meneliti tentang Penerapan Data Mining Untuk Memprediksi Prestasi Akademik Mahasiswa Berdasarkan Dosen, Motivasi, Kedisiplinan, Ekonomi, dan Hasil Belajar. Metode data mining yang digunakan adalah metode klasifikasi dengan algoritma C4.5. Algoritma ini dapat menghasilkan pohon keputusan yang menjadi alat untuk mendukung keputusan memprediksi prestasi akademik mahasiswa. Dari hasil penelitian diperoleh bahwa variabel nilai rapor (hasil belajar masa lalu) menjadi node awal artinya dari 5 variabel yang menetukan prestasi akademik mahasiswa maka nilai rapor menjadi node yang terpilih sebagai penentu pertama terhadap prestasi akademik mahasiswa. Berdasarkan analisis data menggunakan algoritma Decision Tree untuk memprediksi prestasi akademik berdasarkan sosial ekonomi, motivasi, peran dosen, disiplin dan hasil belajar masa lalu diperoleh hasil : (1) variabel hasil belajar masa lalu adalah variabel yang menentukan potensi seseorang berhasil atau tidak dalam prestasi akademik. Hal ini dibuktikan bahwa Hasil Belajar menjadi node yang terpilih/awal. (2) Variabel Peran Dosen menjadi variabel kedua menentukan Prestasi akademik (3). Variabel Disiplin menjadi variabel ketiga menentukan Prestasi Akademik . (4) Hasil Akurasi klasifikasi menggunakan metode Area Under Curve (AUC) memperoleh nilai 65%.

Penelitian (Murtopo, 2015), Prediksi Kelulusan Tepat Waktu Mahasiswa STMIK YMI Tegal Menggunakan Algortima Naive Bayes. Faktor eksternal yang digunakan untuk menjadi penentu dalam model ini antara lain status kerja dan status perkawinan. Berdasar faktor tersebut apakah faktor eksternal berpengaruh pada kelulusan mahasiswa secara tepat waktu. Hasil dari penelitian ini adalah pengukuran akurasi, dimana sebelum didapatkan nilai akurasi dilakukan pengujian dengan memanfaatkan ROC Curva dan k-fold cross validation, pengujian dilakukan sebanyak 10 fold. Dari hasil pengujian didapat nilai akurasi rata-rata sebesar 91,29%, sedangkan nilai akurasi tertinggi dari hasil pengujian 10-fold cross validation sebesar 94,34%.

Menurut Penelitian (Hastuti, 2012), Analisis Komprasi Algoritma Klasifikasi data mining untuk prekdsi. Mahasiswa non aktif. Perlu diketahui faktor-faktor penyebab

mahasiswa memiliki status non aktif. Teknik klasifikasi data mining dapat digunakan untuk prediksi mahasiswa non aktif. Banyak algoritma klasifikasi data mining yang dapat digunakan, sehingga perlu dilakukan komparasi untuk mengetahui tingkat akurasi dari masing-masing algoritma. Algoritma yang digunakan adalah logistic regression, decision tree, naïve bayes dan neural network. Data yang digunakan sebanyak 3861 mahasiswa program studi Teknik Informatika, Sistem Informasi dan Desain Komunikasi Visual Universitas Dian Nuswantoro. Hasil dari proses klasifikasi dievaluasi dengan menggunakan cross validation, confusion matrix, ROC Curve dan T-Test untuk mengetahui algoritma klasifikasi data mining yang paling akurat untuk prediksi mahasiswa non aktif.

Penelitian (Fithri et al., 2014). Sistem Pendukung Keputusan untuk memprediksi Kelulusan Mahasiswa Menggunakan Metode Naive Bayes. Prediksi ini nantinya digunakan sebagai sumber informasi untuk menghasilkan sebuah keputusan. Pengolahan data mining mahasiswa dengan menggunakan metode naïve Bayes dimulai dari proses Data Gathering, Data Preprocessing, Proposed Model/Method, Method Test and Experiment, Result Evaluation and Validation. Dalam penelitian ini hasil yang dicapai memiliki akurasi untuk tepat waktu sebesar 93% dan akurasi untuk terlambat sebesar 71%.dengan menggunakan metode Naïve Bayes yang semakin optimal dengan menentukan mahasiswa lulus tepat waktu atau terlambat. Hasil dari Penelitian ini menghasilkan sistem pendukung keputusan dengan menggunakan metode Naïve Bayes dengan menggunakan beberapa parameter, yaitu jenis kelamin, rata-rata IPK, umur, alamat, status pekerjaan mahasiswa, status pernikahan mahasiswa, jumlah SKS dan status mahasiswa.

Menurut (Ridwan, Suyono, & Sarosa, 2013). Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. Hasil pengujian menunjukkan bahwa faktor yang paling berpengaruh dalam penentuan klasifikasi kinerja akademik mahasiswa yaitu Indeks Prestasi Komulatif (IPK), Indeks Prestasi (IP) semester 1, IP semester 4, dan jenis kelamin. Sehingga faktor-faktor tersebut dapat digunakan sebagai bahan evaluasi bagi pihak pengelola perguruan tinggi. Pengujian pada data mahasiswa angkatan 2005-2009, algoritma NBC menghasilkan nilai *precision, recall,* dan *accuracy* masing-masing 83%, 50%, dan 70%, sedangkan Output dari sistem ini berupa klasifikasi kinerja akademik mahasiswa yang diprediksi kelulusannya dan

memberikan rekomendasi untuk proses kelulusan tepat waktu atau lulus dalam waktu yang paling tepat dengan nilai optimal.

Penelitian (Yuda Septian, 2009). Data Mining Menggunakan Algoritma Naive Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro. Tujuan dari penelitian ini adalah untuk melakukan klasifikasi terhadap data mahasiswa Universitas Dian Nuswantoro Fakultas Ilmu Komputer angkatan 2009 berjenjang DIII dan S1 dengan memanfaatkan proses *data mining* dengan menggunakan teknik klasifikasi. Implementasi menggunakan *RapidMiner* 5.3 digunakan untuk membantu menemukan nilai yang akurat. Algoritma yang digunakan untuk klasifikasi kelulusan adalah algoritma *Naïve Bayes*. Hasil dari penelitian ini digunakan sebagai salah satu dasar pengambilan keputusan untuk menentukan kebijakan oleh pihak Fasilkom. Atribut yang digunakan adalah Tahun lulus, Nama, NIM, Program Studi, Jenjang, Jenis Kelamin, Provinsi Asal, SKS, dan IPK. Metode yang digunakan adalah *CRISP-DM* dengan melalui proses *business understanding*, *data understanding*, data *preparation*, *modeling*, *evaluation* dan *deployment*.

Menurut Penelitian (Amelia, Lumenta, & Jacobus, 2017). Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naive Bayes. Dengan menggunakan data mining khususnya klasifikasi untuk prediksi dengan algoritma naïve bayes dapat dilakukan prediksi terhadap ketepatan waktu studi dari mahasiswa berdasarkan data *training* yang ada.. Pengujian yang dipakai yaitu *k-fold cross validation 10-fold*. Hasil pengujian didapat nilai akurasi rata-rata sebesar 85.17 % sedangkan nilai akurasi tertinggi sebesar 88.96 %.

Maka yang menjadi *concern* penelitian ini berbeda seperti penelitian-penelitian sebelumnya yaitu bagaimana metode prediksi data mining mampu mengelompokan dan memprediksi masa tenggang alumni universitas muhammadiyah yogyakarta mendapatkan pekerjaan setelah menyelesaikan studi S1 berdasarkan data yang ada menggunakan algoritma naive bayes. *Algoritma Naive Bayesian* sendiri memiki beberapa kebihan yaitu konsisten terhadapat atribut yang tidak relevan, juga menangani data kuantitaf dan data diskrit, dan konsisten terhadap atribut yang tidak relevan. *Algoritma* ini hanya memilki 2 kekurangan yaitu data tidak berlaku jika probabilitasnya atau nilai pada atribut benilai *null*, apabila nol maka probabilitas prediksi akan bernilai nol juga dan kekurangan yang kedua yaitu meramalkan atau memperhitungkan variabel bebas. Oleh sebab itu, penulis menggunakan *algortima naive bayes classifier* untuk mudah dibaca dan dipahami saat dikelola pada *software RapidMiner*, serta penulis hanya memerlukan pengkodean yang

sederhana untuk melakukan perhitungan, dan juga *algoritma* ini tidak memelurkan banyak data pada data *testing* untuk melakukan klasifikasi.

2.2 Ladasan Teori

2.2.1. Data Mining

Data mining adalah sebuah *knowledge discovery in database* (KDD). Data Mining merupakan salah satu cara yang digunakan untuk mendapatkan pengetahuan baru dengan memanfaatkan jumlah data yang sangat besar. Beberapa teknik telah dikembangkan dan dimplementasikan untuk mengumpulkan pengetahuan yang mungkin berguna untuk pengambilan keputusan. Teknik-teknik yang digunakan untuk pengekstrakan pengetahuan dalam data mining adalah pengenalan pola, clustering, asosiasi, prediksi dan klasifikasi. (Defiyanti, 2015)

2.2.2. Pengelompokan Data Mining

Ada beberapa teknik yang dimiliki *data mining* berdasarkan tugas yang dilakukan, yaitu (Ridwan et al., 2013) :

1. Deskripsi

   Para peneliti biasanya mencoba untuk mendeskripsikan pola dan tren yang tersembunyi dalam data.

2. Prediksi

   Prediksi memiliki kemiripan dengan estimasi dan klasifikasi. Hanya saja, prediksi hasilnya menunjukan sesuatu yanng belum terjadi (mungkin terjadi di masa depan).

3. Estimasi

   Estimasi sama dengan klasifikasi, kecuali variable tujuannya lebih kearah numerik dari pada kategori.

4. Klasifikasi

   Dalam klasifikasi variabel target sudah diketahui, tujuannya bersifat kategorik. Misalnya, kita akan mengklasifikasikan pendapatan dalam 3 kelas, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah.

5. Clustering

   Clustering lebih ke arah pengelompokan record, pengamatan, atau kasus dalam kelas yang memiliki kemiripan.

6. Asosiasi

   Tugas asosiasi adalah mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu.

2.2.3. Tahapan-tahapan *Data Mining*

Dalam melakukan *proses mining* harus melewati beberapa tahapan seperti berikut ini (Ridwan et al., 2013):

1. Pembesihan Data (*data cleaning)*

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

2. Intergitas Data (*data integration)*

   Intergritas 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. Transformasi Data

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

5. Proses Mining

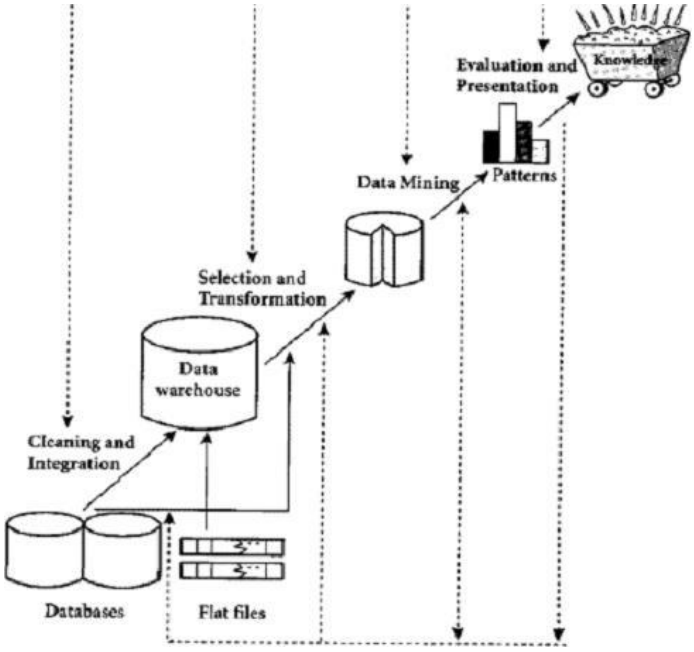
   Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

6. Evaluasi Pola (*puttern evaluation)*

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

7. Presentasi Pengetahuan (*knowledge presentation)*

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



Gambar 2. 1 *Gambar Proses Data Mining*

Klasifikasi adalah proses pencarian kumpulan model yang menggambarkan data dan membedakan kelas data dengan tujuan agar model tersebut dapat digunakan untuk memprediksikelas dari suatu obyek yang belum diketahu kelasnya (Setyawan & Nugroho, 2014).

2.2.4. Naive Bayes

Bayesian classification merupakan pengklasifikasi statistik yang dapat digunakan untuk memprediksi probalitas isi dari suatu *class. Bayesian classification* didasarkan pada teorema *Bayes* yang memiliki kemampuan klasifikasi serupa dengan *neural netrok* dan *decision tree. Bayesian classfication* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar. (Jananto, 2013). *Naive Bayes* memiliki persamaan seperti berikut ini :

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

X = Data dengan class yang belum diketahui

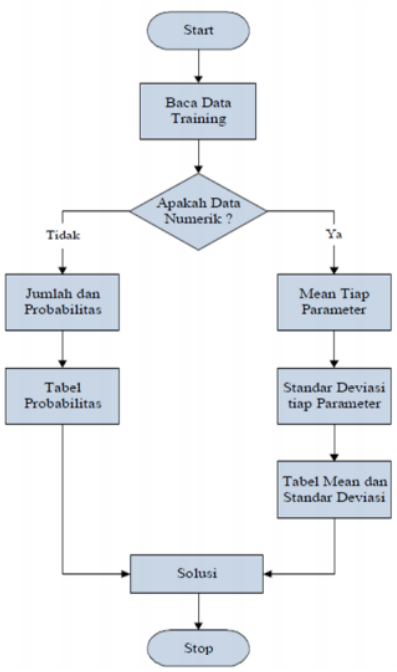
H = Hipotesis data X merupakan suatu class spesifik

P(H|X) = Probalitas hipotesis H berdasarkan kondisi x (posteriori prob.)

P(H) = Probalitas hipotesis H (prior prb.)

P(X|H) = Probalitas X berdasarkan kondisi tersebut P(X) = Probalitas dari X

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



Gambar 2. 2 *Alur Metode Naive Bayes*

Sumber: (Saleh, 2015)

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

$$\mu = \frac{\sum_i^n 1\ Xi}{n}$$

Atau

$$\mu = \frac{\mathrm{x1} + \mathrm{x2} + \mathrm{x3} + \cdots + \mathrm{Xn}}{n}$$

Di mana :

μ : rata-rata hitung (mean)

xi : nilai sample ke –i

n : jumlah sampel

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

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

Dimana :

σ : standar deviasi

Xi : nilai x ke –i

μ : rata-rata hitung

n : jumlah sampel

2.2.5. 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 preproccessing* dan visualisasi. *RapidMiner* juga adakah *software* yang berdiri sendiri untuk analisis data dan sebagai mesin *data mining* yang dapat diintegrasikan pada produknya sendiri. *RapidMiner* ditulis dengan menggunakan bahasa *java* sehingga dapat bekerja di semua sistem operasi (Wicaksana, n.d.).

2.2.6. Microsoft Excel

Microsoft exel adalah *software spreadsheet* paling terkenal di dunia bisnis dan perkantoran. *Exel* 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 grafis, serta mudah dipakai sehingga *excel* menjadi salah satu program komputer yang populer digunakan di PC hingga saat ini. Bahkanm saat *excel* merupakan program *spreadsheet* paling banyak digunakan, baik *platform* PC berbasis *windows* merupakan *plasform macintosh* berbasis *Mac OS* semenjak versi 5.0 yang keluar di tahun 1993.