BAB II
TINJAUAN PUSTAKA DAN LANDASAN TEORI

2.1. Tinjauan Pustaka

Menurut Penelitian (Ridwan, M., Suyono, H., & Sarosa, M, 2013). Penerapan *Data Mining* Untuk Evaluasi Kinerja Akademik Mahasiswa menggunakan Algoritma *Naive Bayes Classifier*. Dalam penelitian ini bertujuan untuk memprediksi dan menganalisa mahasiswa yang bisa lulus tepat waktu dan tidak tepat waktu. Dalam proses *mining* menggunakan data akademik yang dibagi sebagai *data training* dan *data testing*. Hasil dari pengujian terdapat beberapa faktor dalam penentuan klasifikasi kinerja akademik yaitu IPK (Indeks Prestasi Kumulatif), IP (Indeks Prestasi) semester 1 dan semester 4 dan jenis kelamin yang paling berpengaruh dan pengujian dilakukan dengan beberapa percobaan untuk mendapatkan nilai *precision*, *recall,* dan *accuracy* yang dapat digunakan sebagai bahan evaluasi dalam memberikan rekomendasi untuk mendapatkan kelulusan tepat waktu.(Ridwan and Suyono 2013)

Menurut Penelitian (Nurliana Nasution, Khairani Djahara, Ahmad Samsuri, 2015). Evaluasi didalam kinerja akademik dapat dilihat dari semester 3 atau semester 6 untuk menentukan kelulusan tepat waktu dapat dibagi menjadi dua data yaitu data latih dan data uji dengan menggunakan metode penelitian CRISP-DM(*Cross Industry Standard Process for Data Mining)* yang memiliki 6 tahapan yaitu *Business Understanding* (Pemahaman Proses Bisnis), *Data Understanding* (Pemahaman Data), *Data Preparation* (Pengolahan Data), *Model Building* (Pembangunan Model), *Testing and Evaluation* (Pengujian dan Evaluasi), *Deployment* (Penyebaran). Atribut yang signifikan dalam penentuan kelulusan tepat waktu adalah Indeks Prestasi Kumulatif (IPK) menggunakan *naive bayes* dan beberapa atribut terpilih seperti NIM, Nama, Jenis Kelamin, Asal Sekolah, Kota Asal, Tempat Tanggal Lahir, Indeks Prestasi semester 1 sampai dengan 6 mendapatkan akurasi 76% dan meningkat menjadi 76,67% setelah menggunakan atribut terpilih dengan data latih dan data uji dengan porsi yang sama.(Nasution, Djahara, and Zamsuri, n.d.)

Menurut Penelitian (M. Syukri Mustafa, Muh Rizky Ramadhan, Angelina P.Thenata, 2017). Penelitian ini bertujuan untuk mengevaluasi kinerja akademik dalam menentukan proses kelulusan tepat waktu dengan menggunakan data latih dan *data testing* dari nilai mahasiswa angkatan 2008-2011 yang sudah dinyatakan lulus dengan data target mahasiswa angkatan 2013-2014 yang belum lulus. Faktor yang paling berpengaruh dari beberapa atribut yang digunakan adalah Indeks Prestasi (IP) semester 1,2,3,4 dan jenis kelamin dengan nilai akurasi 92,3 %.(Mustafa, Ramadhan, and Thenata 2018)

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 data mahasiswa angkatan 2009 Universitas Dian Nuswantoro Fakultas Ilmu Komputer dengan memanfaatkan proses *data mining* dan metode klasifikasi dengan algoritma *naive bayes*. Hasil penelitian digunakan sebagai dasar pengambilan keputusan untuk menentukan kebijakan Fasilkom dengan menggunakan atribut yaitu Tahun Lulus, Nama, NIM, Program Studi, Jenjang, Jenis Kelamin, Provinsi, Asal, SKS, dan IPK. Metode yang digunakan CRISP-DM (*Cross Industry Standard Process for Data Mining*).(Nugroho, n.d.)

Menurut Penelitian (Putri, 2017) Penerapan *Naive Bayes* Untuk Perangkingan Kegiatan Fakultas TIK Universitas Semarang. Tujuan dari penelitian adalah untuk melakukan perangkingan terhadap suatu acara yang ada di Fakultas TIK Universitas Semarang dengan menggunakan data yang diperoleh secara langsung melalui observasi dan wawancara terhadap Ketua Program Studi Teknik Informatika, Sistem Informasi dan Ilmu Komunikasi. Hasil dari penelitian dengan menganalisa perhitungan metode *naive bayes* dengan mengklasifikasikan 2 jenis favorit dan tidak favoritnya suatu acara yang ada di Fakultas TIK Universitas Semarang. Didapatkan hasil dengan menggunakan *RapidMiner* yaitu tingginya tingkat kegiatan yang tidak favorit di Fakultas TIK dengan rincian berdasarkan *chart* pada *RapidMiner* di Fakultas TIK bahwa Program Studi Ilmu Komunikasi menempati urutan Terfavorit dengan nilai 6, disusul urutan kedua yaitu Program

Studi Teknik Informatika dengan nilai 4, dan urutan terakhir adalah Program Studi Sistem Informasi dengan nilai 3.(Putri 2017)

Menurut Penelitian (Murtopo, 2016) Dalam Penelitian ini metode yang digunakan dalam memprediksi kelulusan tepat waktu mahasiswa pada Mahasiswa STMIK-YMI Tegal menggunakan algoritma *naive bayes*. Data yang digunakan sebanyak 510 dari mahasiswa yang lulus tahun 1999 sampai 2014 pada D3 Manajemen Informatika dan Komputer. Hasil dari penelitian untuk menghitung probabilitas kemungkinan tepat waktu atau tidak tepat waktu di evaluasi dengan *confusion matrix* dengan menggunakan *10-fold cross validation* dengan melibatkan faktor internal dan eksternal menghasilkan akurasi klasifikasi mahasiswa lulus tepat waktu sebesar 91.37 % sedangkan evaluasi pengujian menggunakan *10-fold cross validation* menghasilkan nilai tertinggi akurasi 94,34 % dengan rata-rata akurasi sebesar 91.29 sedangkan evaluasi kurva ROC dengan metode *AUC* sebesar 0.898 yang termasuk *good classification.* Dengan menggunakan metode *naive bayes* untuk menghitung probabilitas kelulusan tepat waktu atau terlambat menunjukkan peningkatan dibanding penelitian sebelumnya.(Murtopo 2016)

Menurut Penelitian (Marselina dan Suhartinah, 2010) Penerapan algoritma *naive bayes* dan C4.5 dalam prediksi kelulusan mahasiswa yang dapat lulus sesuai dengan waktu studi menggunakan *Java Netbeans*. Hasil pengujian dengan menggunakan algoritma *naive bayes* berdasarkan hasil pengujian dari 22 data yang diuji, terdapat 2 data hasil prediksi yang tidak sesuai dengan hasil data sebenarnya. Hasil pengujian yang didapat akurasi ketepatan hasil prediksi adalah 80,85% sementara persentase kesalahan sebesar 19,05%.(Suhartinah, n.d.)

Menurut Penelitian (Muhammad Ihsan Zul) Prediksi Hasil Penilaian Akhir Mahasiswa pada Matakuliah Tertentu dengan Menggunakan Algoritma *K-nn* dan *naive bayes*. Data yang digunakan adalah *data training* yang digunakan merupakan data penilaian untuk matakuliah tertentu yang dilaksanakan pada semester yang lalu. Artinya terdapat nilai akhir berupa nilai huruf sedangkan *data testing* dikumpulkan dari nilai semester yang telah berlangsung yang belum memiliki nilai angka akhir dan nilai huruf. Pengujian yang dilakukan adalah untuk memperoleh

persentase akurasi dari algoritma. Data dikatakan memiliki akurasi yang baik jika hasil prediksi yang diberikan memiliki nilai *confidence*. Poin-poin penilaian yang dijadikan acuan baku tersebut adalah Tugas, Ujian Tengah Semester, Ujian Akhir Semester, dan Presensi Kehadiran dengan akurasi pengujian yang mencapai 95%.(Zul, n.d.)

Menurut Penelitian (Rahman dan Firdaus, 2016) Penelitian ini bertujuan untuk mengetahui akurasi serta hasil presisi dan *recall* dari metode klasifikasi n*aive bayes* untuk memprediksi hasil belajar siswa sekolah menengah pertama (SMP). Data yang didapatkan berasal dari faktor internal dan eksternal seperti faktor demografi (pendidikan orang tua dan kemampuan ekonomi keluarga) selain itu juga data sekolah (tingkat kehadiran). Data yang digunakan sebanyak 263 data, pengujian yang dilakukan dengan menggunakan *10-fold validation*, sebanyak 85% data sebagai *data training* dan 15% sebagai *data testing*. Evaluasi penelitian ini menggunakan *confusion matrix* untuk melihat akurasi. Hasil yang didapat dengan metode *naive bayes* adalah akurasi sebesar 56,79%, presisi 62,80% dan *recall* 72,56%.(Rahman and Firdaus 2016)

Menurut (Jananto, 2013) Penelitian tentang Algoritma *Naive Bayes* Untuk Mencari Perkiraan Waktu Studi Mahasiswa. Data yang digunakan adalah Fakultas Teknologi Informasi UNISBANK dengan mengambil data mahasiswa yang telah lulus periode 2004-2007 dan nilai matakuliah sampai dengan semester 4 yang diperoleh sebanyak 266 *record data* dengan 6 atribut yaitu IPK (Indeks Prestasi Kumulatif), Jenis Kelamin, Kota Lahir, Tipe Sekolah dan Kota Sekolah. Berdasarkan data yang ada diperoleh hasil dengan tingkat kesalahan sebanyak 20 *record* dari total *data testing* sebanyak 66 *record* dengan tingkat kesalahan prediksi sebesar 20/66 atau 34%. Dilakukan uji coba lain dengan mengambil *data training* dan *testing* secara *random* yang selanjutnya diambil 75% *record* pertama atau 200 *record* dari keseluruhan data sebagai *data training* dan 25% sisanya sebagai *data testing* dan dilakukan uji coba sebanyak 5 kali. Hasil dari uji coba sebanyak 5 kali bahwa tingkat kesalahan prediksi memiliki nilai rata-rata sebesar 34% (33.59036033). Selain itu uji prediksi terhadap mahasiswa angkatan 2008/2009

untuk program studi S1 Sistem Informasi dan S1 Teknik Informatika dengan *data record* sebanyak 258. Hasil uji coba dengan *data training* yang inkonsistensi data yang tinggi diperoleh hasil hanya 1 yang lama prediksinya tidak tepat waktu sedangkan sisanya sebanyak 257 diprediksi tepat waktu. Sedangkan hasil prediksi saat menggunakan *data training 1-x* sebanyak 254 mahasiswa diprediksi Tepat Waktu dan hanya 4 orang diprediksi Tidak Tepat waktu.(Jananto 2013)

2.2. Landasan Teori

2.2.1. Data Mining

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. *Data Mining* atau *Knowledge Discovery in Database* (KDD). KDD adalah kegiatan seperti pengumpulan, pemakaian data, historis untuk menemukan keteraturan pola atau hubungan dalam set data berukuran besar. (Ridwan and Suyono 2013)

2.2.2. Pengelompokan Data Mining

Data mining menurut *Larose* dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu : (Nugroho, n.d.)

a. Deskripsi

   Para peneliti mencari pola untuk mendeskripsikan yang tersembunyi dalam data.

b. Estimasi

   Estimasi memiliki kemiripan dengan klasifikasi tetapi variabel tujuannya lebih numerik dari pada kategori.

c. Prediksi

   Prediksi memiliki kemiripan dengan klasifikasi dan estimasi, namun prediksi nilainya adalah hasil yang akan datang.

d. Klasifikasi

   Klasifikasi memiliki variabel yang bersifat kategori.

e. Pengklusteran

Clustering merupakan metode pengelompokkan data antara satu data dengan data yang lain yang memiliki kemiripan karakteristik.

f. Asosiasi

Menemukan dan menghubungkan atribut yang terdapat dalam satu waktu.

2.2.3. Tahapan Dalam *Data Mining*

Data Mining dalam melakukan prosesnya dibagi menjadi beberapa proses sebagai berikut : (Ridwan and Suyono 2013)

a. Pembersihan Data (*data cleaning)*

Pembersihan data adalah suatu proses menghilangkan data yang tidak konsisten.

b. Integrasi Data (*data integration)*

Dalam melakukan integrasi data adalah dengan menggabungkan beberapa data dari berbagai *database* menjadi satu *database*.

c. Seleksi Data (*data selection)*

Seleksi data dilakukan hanya untuk memilih data yang akan dipakai untuk dianalisis yang diambil dari *database*.

d. Transformasi Data

Data yang sudah dipilih kemudian diubah menjadi format yang sesuai dengan *data mining.*

e. Proses *Mining*

Proses utama dalam menerapkan metode yang tepat untuk menemukan pengetahuan dan tersembunyi dari data.

f. Evaluasi Pola

Mengidentifikasi pola-pola yang menarik ke dalam *knowledge based* yang ditemukan.

g. Presentasi Pengetahuan

Penyajian pengetahuan metode yang telah digunakan untuk mendapatkan pengetahuan yang ditemukan.

2.2.4. Algoritma Klasifikasi *Naive Bayes*

Naive Bayes Classifier adalah sebuah metode klasifikasi berdasarkan teorema *Bayes* yaitu suatu metode untuk memprediksi peluang berdasarkan pengalaman yang ada dengan menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuan Inggris bernama *Thomas Bayes*.(Mustafa, Ramadhan, and Thenata 2018)

Naive Bayes Classifier memiliki akurasi yang lebih baik di bandingkan dengan klasifikasi lain. Menurut penelitian yang dilakukan oleh *Xhemali*, *Hinde*, dan *Stone* dalam sebuah jurnal yang berjudul "*Naive Bayes vs Decision Tree vs Neural Networks in the classification of Training Web Pages". Naive Bayes* terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan kedalam sebuah *database* yang besar.(Mustafa, Ramadhan, and Thenata 2018)

Prediksi *Bayes* memiliki formula teorema *Bayes* dengan formula umum sebagai berikut : (Jananto 2013)

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

Keterangan :

X : data dengan sebuah class yang belum diketahui

H : Hipotesis data X adalah class yang spesifik

P(H|X) : Probabilitas hipotesis H terhadap kondisi X (*prosteriori prob.)*

P(H) : Probabilitas hipotesis H (*prior prob.)*

P(X|H) : Probabilitas X terhadap kondisi H

P(X) : Probabilitas dari X

Adapun perhitungan untuk data berkelanjutan adalah menggunakan *Distribusi Gauss :* (Mustafa, Ramadhan, and Thenata 2018)

$$g(x,\mu,\sigma)= \frac{1}{\sigma\sqrt{2\,\pi}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Keterangan:

g : *Distribusi Gauss*

μ : Rata-rata (*Mean)*

σ : *Standar Deviasi*

Sedangkan rumus yang digunakan dalam menghitung nilai rata-rata adalah sebagai berikut:

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

Keterangan:

μ : Rata-rata (*Mean*)

Xi : Nilai dari sample ke -i

n : Total jumlah sampel

Dan untuk rumus yang digunakan untuk menghitung *standar deviasi* adalah sebagai berikut:

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

Keterangan:

σ : *standar deviasi*

Xi : Nilai x yang ke i

μ : Rata-rata (*Mean*)

n : Total jumlah sampel

2.2.5. RapidMiner

RapidMiner adalah perangkat lunak yang bersifat terbuka(*open source*) untuk melakukan analisis terhadap *data mining, text mining* dan analisis prediksi. *RapidMiner* memiliki 500 operator *data mining* seperti operator *input, output, data preprocessing* dan visualisasi. *RapidMiner* menggunakan bahasa *java* sehingga dapat bekerja pada semua sistem operasi.

2.2.6. Microsoft Excel

Microsoft Excel adalah *software spreadsheet* paling terkenal yang hampir digunakan disemua bidang bisnis. Aplikasi *excel* memiliki fitur kalkulasi dan pembuatan grafik yang sangat mudah dipakai dan menjadi salah satu program komputer yang populer digunakan diberbagai PC hingga saat ini. Bahkan, merupakan program *spreadsheet* yang ada disemua sistem operasi *windows* maupun *platform macintosh* berbasis *Mac OS*.

2.2.7. Microsoft SQL Server

SQL Server adalah *Relational Database Management System* (RDMS) yang dirancang untuk aplikasi dengan arsitektur *client server*. Memiliki kemampuan membuat basis data dan menghubungkan satu dan beberapa server. Bahasa basis data yang digunakan adalah *Transact-SQL*. Bahasa *SQL server* yang berguna untuk mendapatkan satu atau kumpulan basis data dengan cara menjalankan perintah dari pernyataan suatu *SQL*.

2.2.8. MySQL

MySQL adalah *Relational Database Management System (DBMS)* yang didistribusikan secara gratis melalui lisensi GPL (*General Public License)* yang dimana setiap orang bebas menggunakannya. *MySQL* merupakan salah satu turunan dari konsep utama *database* yaitu SQL (*Structure Query Language*).