BAB II

LANDASAN TEORI

2.1 Definisi Sistem

Sistem secara fisik adalah kumpulan dari elemen-elemen yang beroperasi bersama-sama untuk menyelesaikan suatu sasaran (Gordon, 1991).

Jogianto (2005:2) mengemukakan bahwa sistem adalah kumpulan dari elemen-elemen yang berinteraksi untuk mencapai suatu tujuan tertentu.

2.1.1 Karakteristik Sistem

Jogianto (2005: 3) mengemukakan sistem mempunyai karekteristik atau sifat-sifat tertentu, yakni :

1. Komponen

   Suatu sistem terdiri dari sejumlah komponen yang saling berinteraksi, yang artinya saling bekerja sama membentuk satu kesatuan. komponen-komponen sistem atau elemen-elemen sistem dapat berupa suatu subsistem atau bagian-bagian dari sistem. setiap subsistem mempunyai sifat-sifat dari sistem untuk menjalankan suatu fungsi tertentu mempengaruhi proses sistem secara keseluruhan.
2. Batasan sistem

   Batasan sistem (boundary) merupakan daerah yang membatasi antara suatu sistem dengan sistem yang lainnya atau dengan lingkungan luarnya. batasan suatu sistem menunjukan ruang lingkup dari sistem tersebut.
3. Lingkungan Luar Sistem.

   Lingkungan luar (*evinronment*) dari suatu sistem adalah apapun diluar batas sistem yang mempengaruhi operasi. Lingkungan luar sistem dapat bersifat menguntungkan dan dapat juga bersifat merugikan sistem tersebut.
4. Penghubung Sistem

   Penghubung (interfance) merupakan media penghubung antara satu subsistem dengan subsistem yang lainya. melalui penghubung ini memungkinkan

sumber-sumber daya mengalir dari satu subsistem ke subsistem yang lainya. Dengan penghubung satu subsistem dapat berintegrasi dengan subsistem yang lainya membentuk satu kesatuan.

2.2 Indeks Kumulatif Prestasi (IPK)

IPK adalah singkatan dari Indeks Prestasi Kumulatif merupakan angka yang menunjukkan prestasi belajar mahasiswa atau Indeks Prestasi secara kumulatif mulai dari semester pertama sampai dengan semester paling akhir yang telah ditempuh. IP (Indeks Prestasi) yang diperoleh mahasiswa tiap semester digunakan dalam menentukan beban studi yang boleh diambil pada semester berikutnya. Penilaian ini meliputi semua mata kuliah yang direncanakan mahasiswa dalam KRS pada semester tersebut, dengan menggunakan rumus IP ssebagai berikut :

$$IP = \frac{\sum_{j=1}^{n} Nj \,.kj}{\sum_{j=1}^{n} kj} \qquad \text{.....(2.1)}$$

Keterangan :

IP = Indeks Prestasi

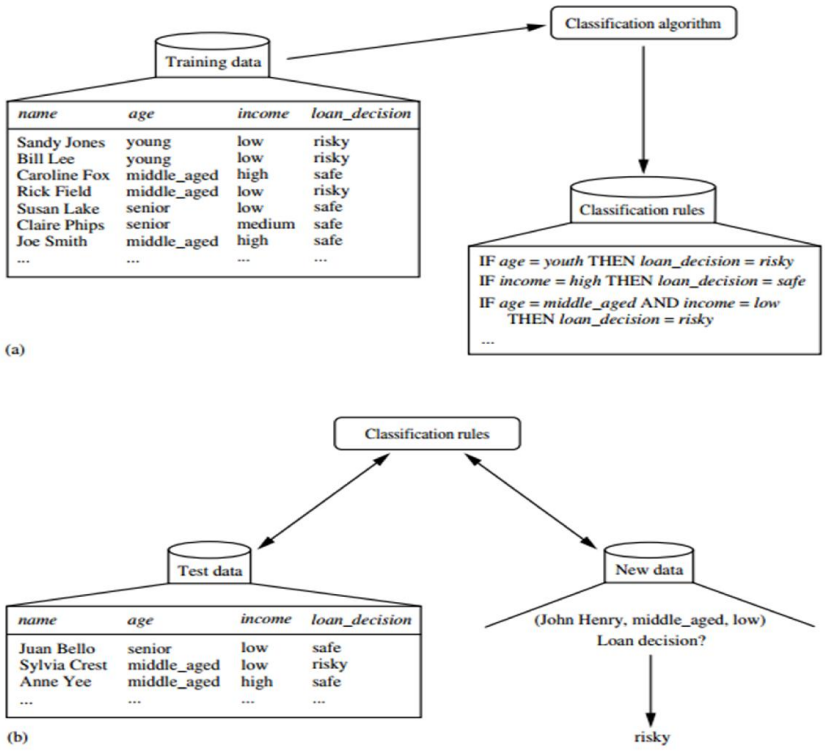
Nj = Nilai mata kuliah

n = Mata kuliah

kj = Bobot SKS mata kuliah

2.3 Klasifikasi

Menurut Mike Chapple (2008), klasifikasi adalah teknik data mining yang dilakukan untuk memprediksi kelas atau properti dari setiap instance data.



Gambar 2.1 Tahapan Klasifikasi Data Mining.

Tahapan dari klasifikasi dalam data mining menurut Han dan Kamber (2006) terdiri dari :

- Pembangunan Model

  Pada tahapan ini dibuat sebuah model untuk menyelesaikan masalah klasifikasi class atau attribut dalam data. Tahap ini merupakan fase pelatihan, dimana data latih dianalisis menggunakan algoritma klasifikasi, sehingga model pembelajaran direpresentasikan dalam bentuk aturan klasifikasi.

- Penerapan Model

  Pada tahapan ini model yang sudah dibangun sebelumnya digunakan untuk menentukan attribut / class dari sebuah data baru yang attribut / classnya belum diketahui sebelumnya. Tahap ini digunakan untuk memperkirakan keakuratan aturan klasifikasi terhadap data uji. Jika model dapat diterima, maka aturan dapat diterapkan terhadap klasifikasi data baru.

2.4 *Data Mining*

2.4.1 Pengertian *Data Mining*

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terakit dari berbagai database besar (Turban, dkk., 2005).

Tan (2006) mendefinisikan *data mining* sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar.

Menurut Han dan Kamber (2006), *data mining* adalah proses menemukan pola yang menarik dan pengetahuan dari data dalam jumlah besar.

Dari beberapa pernyataan tersebut, dapat disimpulkan bahwa *data mining* merupakan proses ekstraksi informasi dari database yang berukuran besar untuk mendapatkan pengetahuan yang tersimpan dari data tersebut.

Istilah *data mining* kadang disebut juga *Knowledge Discovery in Database* (KDD). Istilah *data mining* sering dipakai, mungkin karena istilah ini lebih pendek dari *Knowledge Discovery in Database*. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Data mining dianggap hanya sebagai suatu langkah penting dalam KDD. Proses KDD secara garis besar dapat dijelaskan sebagai berikut (Han dan Kamber, 2006) :

1. Pembersihan data, untuk menghilangkan *noise* dan data yang tidak konsisten.
2. Integrasi data, di mana beberapa sumber data dapat dikombinasikan. Sebuah tren populer di industri informasi adalah untuk melakukan pembersihan dan

integrasi data sebagai langkah preprocessing, dimana data yang dihasilkan akan disimpan dalam *data warehouse*.

3. Seleksi data, di mana data yang relevan dengan tugas analisis yang diambil dari database.
4. Data transformasi (dimana data diubah dan digabung ke dalam bentuk yang sesuai untuk pertambangan dengan melakukan ringkasan atau agregasi operasi) Terkadang transformasi data dilakukan sebelum proses seleksi data, khususnya dalam kasus *data warehouse*.
5. Data mining, merupakan proses esensial dimana metode cerdas diaplikasikan untuk mengekstrak data pola.
6. Evaluasi Pola, untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan.
7. Presentasi pengetahuan, dimana visualisasi dan teknik representasi pengetahuan digunakan untuk menyajikan pengetahuan hasil *data mining* kepada pengguna.

Langkah 1 sampai 4 merupakan berbagai bentuk preprocessing data, dimana data dipersiapkan untuk *data mining*. Hal ini, menunjukkan bahwa data mining sebagai salah satu langkah dalam proses KDD, karena dapat mengungkap pola-pola tersembunyi yang digunakan untuk evaluasi.

2.4.2 Tugas *Data Mining*

Secara umum, tugas *data mining* dapat diklasifikasikan kedalam 2 kategori (Han dan Kamber, 2006), yaitu :

a. Prediktif

   Tujuan dari tugas prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai atribut-atribut lain. Atribut yang diprediksi umumnya dikenal sebagai target atau variable tak bebas, sedangkan atribut-atribut yang digunakan untuk membuat prediksi dikenal sebagai *explanatory* atau *variable* bebas.

b. Deskriptif

Tujuan dari tugas deskriptif adalah untuk menurunkan pola-pola (korelasi, *trend, cluster,* teritori, dan anomali) yang meringkas hubungan yang pokok dalam data. Tugas *data mining* deskriptif sering merupakan penyelidikan dan seringkali memerlukan teknik *post-processing* untuk validasi dan penjelasan hasil.

2.4.3 Fungsi *Data Mining*

Fungsi *data mining* dan macam-macam pola yang dapat ditemukan menurut Han dan Kamber (2006), yaitu:

1. *Concept/Class Description*: *Characterization and Discrimination*

   Data characterization adalah ringkasan dari semua karakteristik atau fitur dari data yang telah diperoleh dari target kelas. Data yang sesuai dengan kelas yang telah ditentukan oleh pengguna biasanya dikumpulkan di dalam *database*. Misalnya, untuk mempelajari karakteristik produk perangkat lunak dimana pada tahun lalu seluruh penjualan telah meningkat sebesar 10%, data yang terkait dengan produk-produk tersebut dapat dikumpulkan dengan menjalankan sebuah *query SQL*.

   Data discrimination adalah perbandingan antara fitur umum objek data target kelas dengan fitur umum objek dari satu atau satu set kelas lainnya. target diambil melalui *query database*. Misalnya, pengguna mungkin ingin membandingkan fitur umum dari produk perangkat lunak yang pada tahun lalu penjualannya meningkat sebesar 10% tetapi selama periode yang sama seluruh penjualan juga menurun setidaknya 30%.

2. *Mining Frequent Patterns, Associations, and Correlations*

   Frequent Patterns adalah pola yang sering terjadi di dalam data. Ada banyak jenis dari *frequent patterns*, termasuk di dalamnya pola, sekelompok *item set*, *sub-sequence*, dan sub-struktur. Sebuah *frequent patterns* biasanya mengacu pada satu set item yang sering muncul bersama-sama dalam suatu kumpulan data transaksional, misalnya seperti

susu dan roti. *Frequent patterns* sering mengarah pada penemuan asosiasi yang menarik dan korelasi dalam data.

Associations Analysis adalah pencarian aturan-aturan asosiasi yang menunjukan kondisi-kondisi nilai atribut yang sering terjadi bersama-sama dalam sekumpulan data. Analisis asosiasi sering digunakan untuk menganalisa *Market Basket Analysis* dan data transaksi.

3. *Classification and Prediction*

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan memprediksikan kelas untuk data yang tidak diketahui kelasnya. Model yang diturunkan didasarkan pada analisis dari training data (yaitu objek data yang memiliki label kelas yang diketahui). Model yang diturunkan dapat direpresentasikan dalam berbagai bentuk seperti *If-then* klasifikasi, *decision tree*, naïve bayes, dan sebagainya.

Teknik *classification* bekerja dengan mengelompokkan data berdasarkan *data training* dan nilai atribut klasifikasi. Aturan pengelompokan tersebut akan digunakan untuk klasifikasi data baru ke dalam kelompok yang ada.

Dalam banyak kasus, pengguna ingin memprediksikan nilai-nilai data yang tidak tersedia atau hilang (bukan label dari kelas). Dalam kasus ini nilai data yang akan diprediksi merupakan data *numeric*. Disamping itu, prediksi lebih menekankan pada identifikasi *trend* dari distribusi berdasarkan data yang tersedia.

4. *Cluster Analysis*

Cluster adalah kumpulan objek data yang mirip satu sama lain dalam kelompok yang sama dan berbeda dengan objek data di kelompok lain. Sedangkan, *Clustering* atau Analisis *Custer* adalah proses pengelompokkan satu set benda-benda fisik atau abstrak kedalam kelas objek yang sama. Tujuannya adalah untuk menghasilkan pengelompokan

objek yang mirip satu sama lain dalam kelompok-kelompok. Semakin besar kemiripan objek dalam suatu *cluster* dan semakin besar perbedaan tiap *cluster* maka kualitas analisis *cluster* semakin baik.

5. *Outlier analysis*

   Outlier merupakan objek data yang tidak mengikuti perilaku umum dari data. *Outlier* dianggap sebagai noise atau pengecualian. Analisis *data outlier* dapat dianggap sebagai *noise* atau pengecualian. Analisis *data outlier* dinamakan *Outlier Mining*. Teknik ini berguna dalam *fraud detection* dan *rare events analysis*.

6. *Evolution Analysis*

   Analisis evolusi data menjelaskan dan memodelkan *trend* dari objek yang memiliki perilaku yang berubah setiap waktu. Teknik ini dapat meliputi karakterisasi, diskriminasi, asosiasi, klasifikasi, atau *clustering* dari data yang berkaitan dengan waktu.

2.5 Teorema Bayes

Pengklasifikasian adalah sebuah fungsi yang menugaskan data tertentu kedalam sebuah kelas. Dari sudut pandang peluang, berdasarkan aturan Bayes kedalam kelas c adalah :

$$P(c \mid E) = \frac{P(E \mid c)\ x\ P(c)}{P(E)} \quad \text{.....(2.2)}$$

Keterangan :

P(c | E) = Probabilitas akhir bersyarat (*conditional probability)* suatu hipotesis c terjadi jika diberikan bukti (*evidence)* E terjadi.

P(E | c) = Probabilitas sebuah bukti E terjadi akan mempengaruhi hipotesis c.

P(c) = Probabilitas awal hipotesis c terjadi tanpa memandang bukti apapun.

P(E) = Probabilitas awal bukti E terjadi tanpa memandang hipotesis atau bukti yang lain.

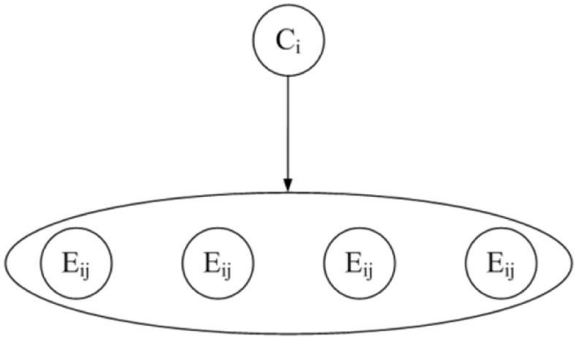
Untuk menentukan pilihan kelas, digunakan peluang maksimal dari seluruh c dalam C, dengan fungsi :

$$\underset{c \in C}{argmax} \frac{P(E|c)P(c)}{P(E)} \quad \text{.....(2.3)}$$

Karena P (E) adalah konstan untuk semua kelas, maka P (E) dapat diabaikan sehingga menghasilkan fungsi :

$$f_c(E) = \underset{c \in C}{argmax} P(E|c)P(c) \quad \text{.....(2.4)}$$

Pengklasifikasian menggunakan Teorema Bayes ini membutuhkan biaya komputasi yang mahal (waktu prosessor dan ukuran memory yang besar) karena kebutuhan untuk menghitung nilai probabilitas untuk tiap nilai dari perkalian kartesius untuk tiap nilai atribut dan tiap nilai kelas.



Gambar 2.2 Ilustrasi Teorema Bayes.

2.6 Naïve Bayes Classifier

Klasifikasi *Naïve Bayes* adalah metode yang berdasarkan probabilitas dan Teorema Bayes dengan asumsi bahwa setiap variabel bersifat bebas (*independence*) dan mengasumsikan bahwa keberadaan sebuah fitur tidak ada kaitannya dengan keberadaan fitur yang lain. Asumsi keidependenan atribut akan menghilangkan kebutuhan banyaknya jumlah data latih dari perkalian kartesius seluruh atribut yang dibutuhkan untuk mengklasifikasikan suatu data.

Formulasi Naïve Bayes untuk klasifikasi adalah

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i \mid Y)}{P(X)} \quad \text{.....(2.5)}$$

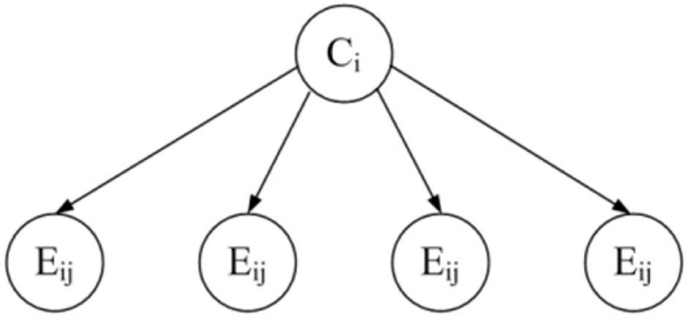
Keterangan :

P(Y|X) = probabilitas data dengan vektor X pada kelas Y

P(Y) = probabilitas awal kelas Y

$\prod_{i=1}^{q} P(X_i \mid Y)$ = probabilitas independen kelas Y dari semua fitur dalam vektor X

Karena P (X) selalu tetap, sehingga dalam perhitungan prediksi nantinya cukup hanya dengan menghitung $P(Y)\prod_{i=1}^{q} P(X_i \mid Y)$.



Gambar 2.3 Ilustrasi Naive Bayes.

Umumnya, Naïve Bayes mudah dihitung untuk fitur bertipe kategoris seperti pada contoh diatas. Namun untuk tipe numerik (kontinu), ada perlakuan khusus sebelum dimasukkan dalam Naïve Bayes, yaitu :

1. Melakukan diskretisasi pada setiap fitur kontinu dan mengganti nilai fitur kontinu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasi fitur kontinu ke dalam fitur ordinal.

2. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk fitur kontinu dan memperkirakan parameter distribusi dengan data pelatihan. Distribusi Gaussian biasanya dipilih untuk mempresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas $P(X_i|Y)$. Untuk setiap kelas y_j, probabilitas bersyarat kelas y_j untuk fitur X_i adalah

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}} \quad \text{.....(2.6)}$$

Keterangan :

μ_{ij} = mean sampel X_i ($\bar{x}$) dari semua data latih.

σ^2_{ij} = varian sampel (s^2) dari data latih.

2.6.1 Algoritma Klasifikasi Naïve Bayes

Algoritma Klasifikasi Naïve Bayes dihitung sesuai dengan rumus Naïve Bayes $P(Y)\prod_{i=1}^{q} P(X_i \mid Y)$, yang langkah-langkah perhitungannya dijelaskan sebagai berikut :

1. Menghitung nilai probabilitas kelas berdasarkan data latih → $P(Y)$

2. Menghitung nilai probabilitas tiap fitur berdasarkan data latih → $\prod_{i=1}^{q} P(X_i \mid Y)$

   - Untuk fitur bertipe numerik menggunakan rumus berikut

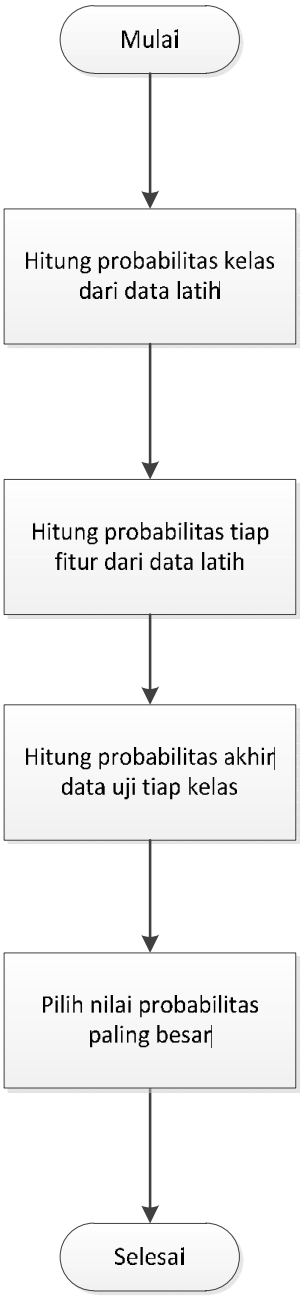
$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

   Fitur numerik berikut ini dihitung tiap data uji.

3. Menghitung nilai probabilitas akhir
   - Mengalikan hasil dari $P(Y)$ dan $\prod_{i=1}^{q} P(X_i \mid Y)$ pada masing-masing kelas dan data uji.

4. Data uji akan diklasifikasikan pada kelas dengan nilai probabilitas akhir terbesar.

Berikut flowchart dari naïve bayes seperti gambar 2.4 di bawah ini.



Gambar 2.4 *Flowchart* Naive Bayes.

2.7 Riset – Riset Terkait

Naïve Bayes merupakan metode popular yang banyak digunakan untuk klasifikasi. Beberapa riset yang telah dilakukan berkaitan dengan kasus prediksi yang menggunakan metode Naïve Bayes, antara lain :

Penelitian yang berjudul *"Data mining classification untuk prediksi lama masa studi mahasiswa berdasarkan jalur penerimaan dengan metode Naive Bayes"* oleh Jonh Fredrik Ulysses. Adapun data yang diambil dalam penelitian ini adalah data sampel dari 57 alumni mahasiswa STMIK Palangkaraya jurusan D3 Manajemen Informatika tahun kelulusan 2006-2008. Atribut yang digunakan adalah lama studi / semester dan dibagi menjadi 2 kelas, yaitu jalur khusus dan SPMB. Hasil penelitian menunjukkan bahwa mahasiswa yang masuk melalui Jalur Khusus memiliki kecenderungan untuk lulus lebih cepat dibandingkan mahasiswa melalui jalur SPMB.

Sri Kusumadewi melakukan penelitian untuk mengklasifikasikan status gizi menggunakan metode Naïve Bayes. Penentuan status gizi menggunakan pengukuran antropometri, yang meliputi penilaian terhadap usia dan berat badan, panjang badan, atau tinggi badan, dan lingkar lengan atas. Data akan diklasifikasikan sebanyak 5 kelas sesuai dengan nilai standar Indeks Massa Tubuh (IMT). Hasil penelitian menggunakan metode Naïve Bayes ini menunjukkan total kinerja sebesar 0,932 atau 93,2%.

Penelitian lain dilakukan oleh Dian Oktafia dan Crispina Pardede mengenai "*Perbandingan Kinerja Algoritma Decision Tree dan Naïve Bayes dalam Prediksi Kebangkrutan"*. Dalam penelitiannya, diuji coba dengan menggunakan sampel data sebanyak 33 perusahaan, terdiri dari 22 perusahaan yang masih aktif yang terdaftar di Bursa Efek Indonesia (BEI) dan 11 perusahaan yang sudah bangkrut diambil dari data yang digunakan pada penelitian sebelumnya yang dilakukan oleh Cindy Yoshiko Shirata (1998). Hasil penelitian menerangkan bahwa kinerja algoritma Naïve Bayes lebih baik dibandingkan dengan algoritma *Decision Tree*, terutama pada tipe data kategori. Algoritma *Decision Tree* mempunyai nilai

akurasi untuk data numerik adalah sebesar 96.97% dan data kategori sebesar 84.85%. Sedangkan algoritma Naive Bayes menghasilkan nilai yang lebih besar dibandingkan algoritma *Decision Tree* yaitu 100% untuk data numerik dan 87.88% untuk data kategori.

Selain itu, penelitian – penelitian yang terkait mengenai prediksi IPK mahasiswa, salah satunya adalah penelitian yang dilakukan oleh Abdul Nasrah dalam risetnya mengenai prediksi IPK mahasiswa ilmu komputer dengan menggunakan algoritma VF15 menjelaskan bahwa hasil penelitian dapat digunakan ketika proses penjurusan mahasiswa TPB mahasiswa sebagai pertimbangan bagi mahasiswa untuk masuk jurusan Ilmu Komputer berdasarkan nilai mata kuliah fisika, kalkulus, dan IP TPB dan menyimpulkan bahwa rata-rata akurasi yang dihasilkan pada angkatan 2001/2002 adalah 70.61% dan pada angkatan 2002/2003 adalah 60.03%.