# Naïve Bayes untuk Klasifikasi Data : Studi Kasus Car Dataset

**Laporan Tugas Besar**

**Rizki**
**Mirza Dita A**
**Ali Ridho F**
**Luthfi Rendragiri**

**Program Studi Sarjana Teknik Informatika**
**Fakultas Informatika**
**Universitas Telkom**
**Bandung**
**2015**

# Bab 1

# Dasar Teori

## 1.1 Teorema Bayes

Teorema bayes, Hukum Bayes, atau Aturan Bayes merupakan teorema yang ditemukan oleh Thomas Bayes, ahli statistika dan filosofis Inggris.Teorema Bayes berkaitan tentang probabilitas kondisional. Secara umum, teorema Bayes dapat dinyatakan dengan persamaan matematis sebagai berikut :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Dimana A dan B merupakan *event*.

- $P(A)$ dan $P(B)$ merupakan peluang tiap masing-masing *event*.
- $P(A|B)$ merupakan probabilitas kondisional untuk $A$ jika $B$ terjadi.

Pada kecerdasan artifisial dan *Data Mining*, teorema Bayes digunakan dalam metode *learning* yang sering disebut dengan nama *Naïve Bayes Classifier*. Pada beberapa referensi, algoritma ini memiliki penamaan yang beragam.

Implementasi didunia nyata cukup beragam, misal untuk *text classification*, *spam filtering*, *document classification*, dan lain-lain. Berikut contoh sederhana untuk *Naïve Bayes Classifier* :

### 1.1.1 Contoh

Misal, terdapat dua kelas yaitu,

$$c1 = pria, c2 = wanita$$

diketahui, ada orang bernama "Drew" yang jenis kelaminnya diasumsikan kita tidak tahu. Berdasarkan kelas yang ada, probabilitas yang mungkin adalah, $P(pria|drew)$ atau $P(wanita|drew)$. Misal kita memiliki data sebagai berikut:

Diketahui berdasarkan Teorema Bayes,

$$P(A|B) = \frac{P(B|A)(P(A)}{P(B)}$$

sehingga, dengan data yang kita miliki,

$$P(pria|drew) = \frac{1/3 * 3/8}{5/8} = \frac{0.125}{0.625} = 0.2$$

| Nama | Jenis Kelamn |
|---|---|
| Drew | Pria |
| Claudia | Wanita |
| Drew | Wanita |
| Drew | Wanita |
| Alberto | Pria |
| Karin | Wanita |
| Sergio | Pria |

Tabel 1.1: Data sampel acak

$$P(wanita|drew) = \frac{2/5 * 5/8}{3/8} = \frac{0.250}{0.667} = 0.374$$

Maka, berdasarkan sampel acak dari tabel diatas, orang yang bernama Drew besar kemungkinan berjenis kelamin wanita.

# Bab 2

# Program dan Algoritma

## 2.1 Algoritma

Pada kasus tugas besar ini, digunakan dataset yang telah diberikan (*car.dat*) yang kemudian akan dibagi menjadi dua bagian yaitu data training dan data testing. Misal, 20% data testing 80% data training, atau kombinasi lain.

Algoritma komputasi pada program yang telah dibuat dapat dilihat pada method berikut:

**calcProbIndependent** menghitung *independent probability* pada kelas tertentu terhadap data training.

**CalcDependentProb** menghitung *dependent probability* atribut tertentu terhadap kelas tertentu. Dipanggil pada method *calcDependentOne*

**CalcDependentOne** menghitung *dependent probability* untuk satu *record* (seluruh atribut) terhadap seluruh kelas.

**calcTotalDependent** menghitung jumlah *dependent* suatu atribut terhadap kelas tertentu di data training. Dipanggil pada method *calcDependentProb*

**calcRecordTotal** menghitung jumlah *record* yang ada pada data training. Dipanggil pada method *calcDependentProb*

**classification** mengklasifikasikan suatu *record* pada kelas tertentu.

**calcAccuracy** menghitung akurasi hasil klasifikasi.

### 2.1.1 Algoritma Independent Probability

Berdasarkan pada method *calcProbIndependent*, algoritma untuk menghitung probabilitas independen adalah sebagai berikut:

$$P(kelas) = \frac{nKelas}{nData}$$

- nData adalah banyaknya datatraining.

Contoh : 50% data training, sehingga *independent probability* yang didapat adalah sebagai berikut:

$$P(unacc) = \frac{180}{864} = 0.2083$$

$$P(acc) = \frac{684}{864} = 0.7916$$

$$P(vgood) = \frac{0}{864} = 0$$

$$P(good) = \frac{0}{864} = 0$$

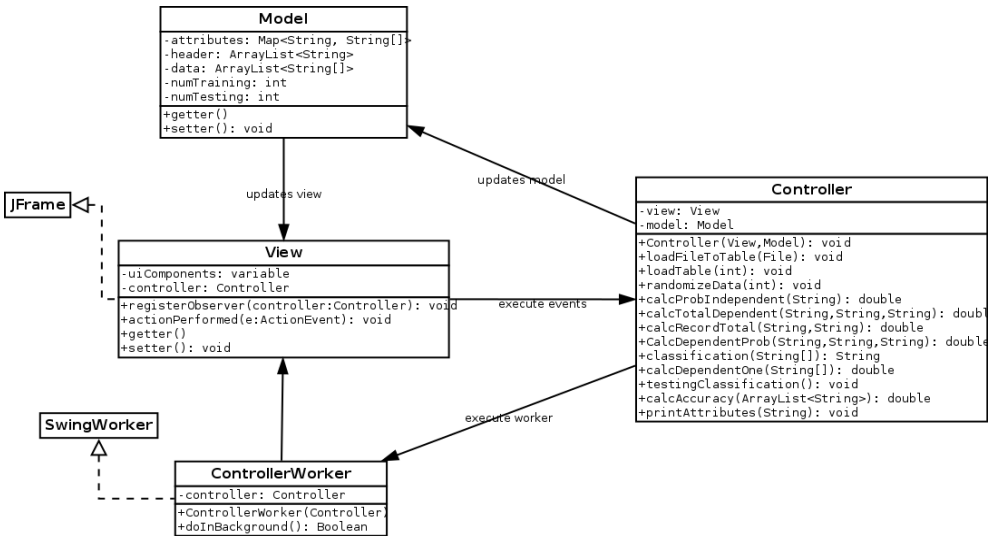
### 2.1.2 Algoritma Dependent Probability

Berdasaran pada method *calcDependenProb*, algoritma untuk menghitung probabilitas dependen adalah sebagai berikut:

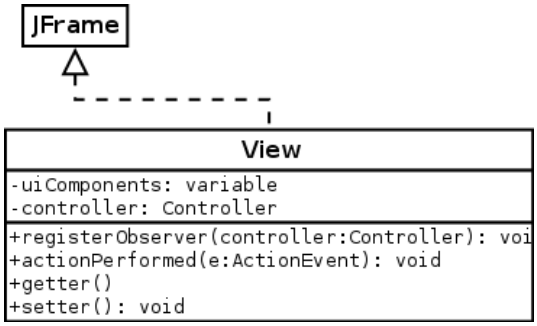
$$P(acc|x) = \frac{P(x_1|acc)P(x_2|acc)P(x_3|acc)P(x_4|acc)P(x_5|acc)P(x_6|acc)P(acc)}{P(x)}$$

dengan $x$ adalah satu *record* data, dalam kasus ini memiliki enam header / atribut.

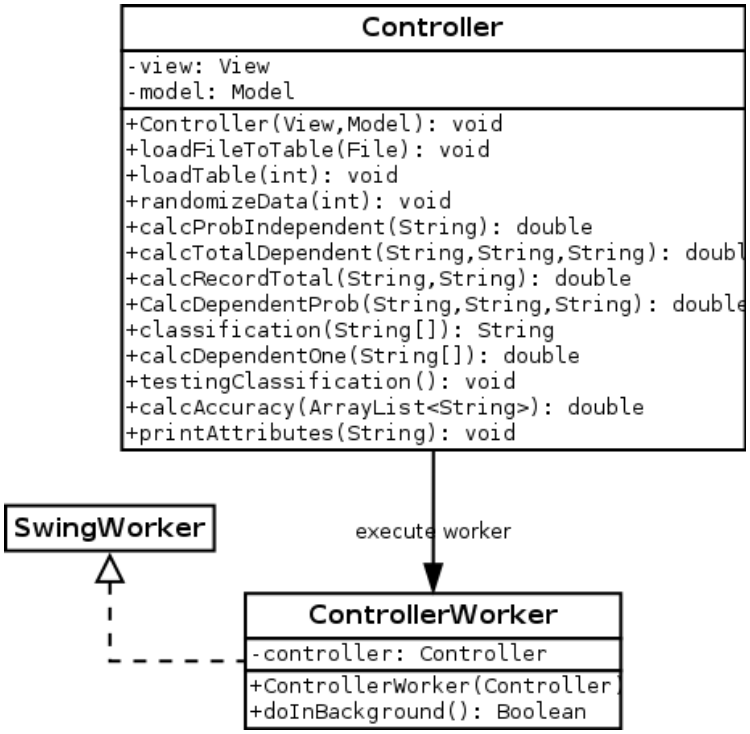
## 2.2 *Class Diagram*



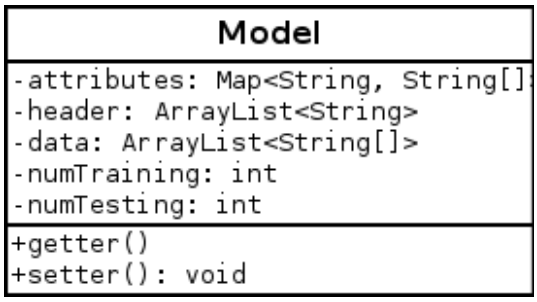

Gambar 2.1: Class Diagram




Gambar 2.2: View




Gambar 2.3: Controller




Gambar 2.4: Model

## 2.3 Fungsionalitas Program

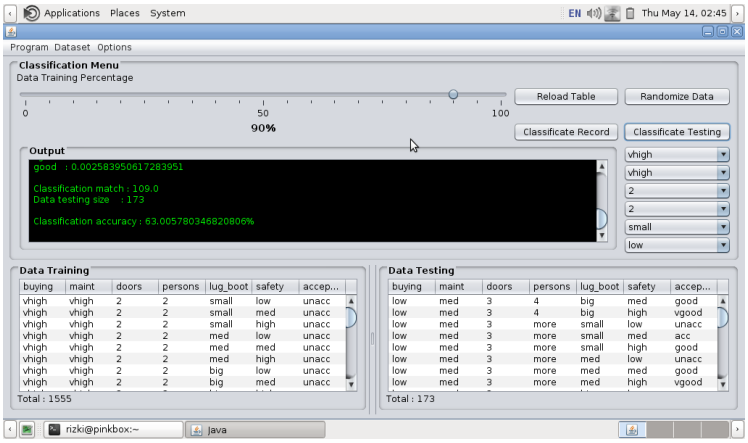
Selain fungsionalitas utama, terdapat fitur tambahan pada program meliputi:

- Tatap muka pengguna dengan *swing*
- *Randomize* atau acak data
- *Multi-threading* untuk komputasi berat (klasifikasi testing).
- *Console verbose* untuk melihat log jalannya program.
- *splitting* data training dan data testing dengan *slider*
- *load* dan *save* data dari *dataset* ke basis data dengan MySQL.

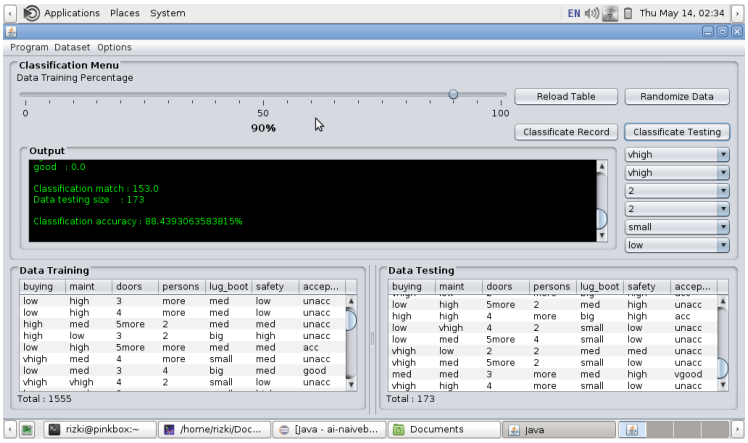
# Bab 3

# Analisis

## 3.1 Test Case




Gambar 3.1: Testcase 90:10 tanpa acak data




Gambar 3.2: Testcase 90:10 dengan data acak

Dengan program yang telah dibuat, dilakukan testcase dengan dataset *car.dat* yang telah disediakan. Adapun testcase akan diujikan pada perbandingan datat raining dan data testing sebesar 75:25, 50:50, 25:75, 10:90, dan 90:10. Kemudi-

an akan dilakukan analisis terhadap hasil testcase dan ditentukan saran terhadap permasalahan ini.

Setelah dilakukan testcase sebanyak lima kali untuk masing-masing perbandingan data training dan data testing yang telah diacak, dihasilkan nilai keakuratan yang fluktuatif dengan kisaran rata-rata terletak diantara 80%-90%. Namun, akurasi akan sangat buruk jika data tidak diacak. Hal ini normal, karena dataset yang disediakan adalah terurut. Selain itu *splitting* data dilakukan dengan cara membagi data dengan perbandingan yang telah ditentukan. Oleh karena itu, perlu dibuatkan method untuk mengacak data sebelum dilakukannya klasifikasi data testing.

## 3.2 Kesimpulan dan Saran

Berdasarkan hasil pengujian, kesimpulan yang didapatkan yaitu :

- Algoritma *Naïve Bayes Classifier* memiliki keakuratan yang cukup baik dengan implementasi yang tidak terlalu sulit.
- Urutan data berpengaruh terhadap nilai akurasi.
- Nilai akurasi bersifat fluktuatif dengan kisaran rata-rata yaitu antara 80% hingga 90%.
- Perbandingan ukuran data training dan data testing mempengaruhi nilai akurasi.

Adapun saran dari hasil pengujian ini antara lain :

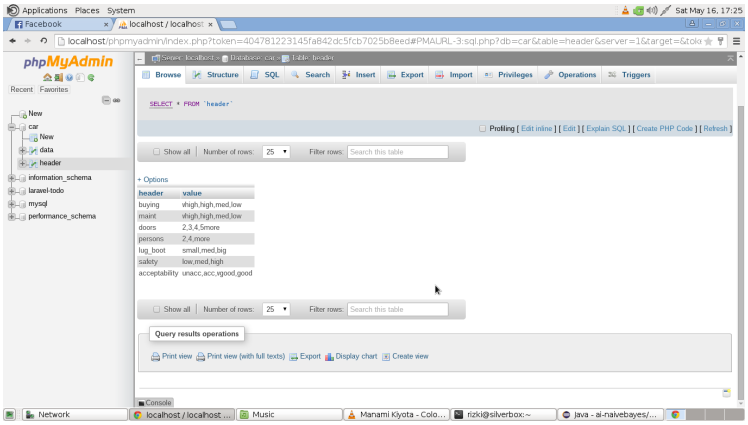
- Diperlukan kajian untuk mencari fungsi acak yang menghasilkan nilai akurasi optimal.
- Dilakukan klasifikasi dengan metode lain (optimasi algoritma *Naïve Bayes Classifier*).
- Dilakukan testcase yang lebih banyak agar hasil kajian akurat.

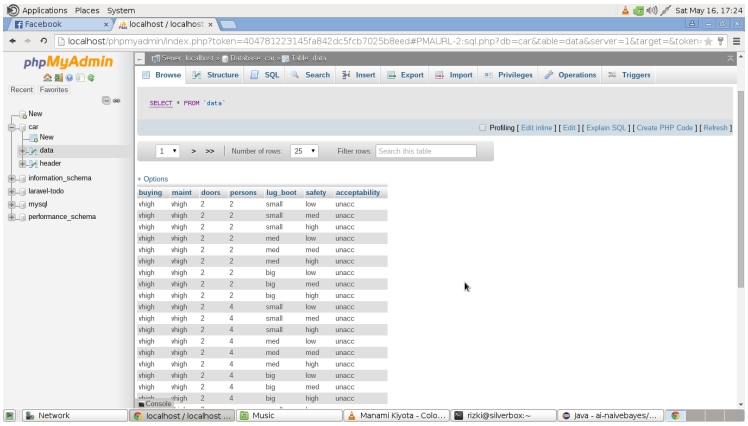
# Lampiran

## Catatan

- kode sumber dapat diunduh di https://github.com/rizkidoank/tugasbesar.
- *screenshot* hasil testcase lengkap dapat diunduh di alamat github diatas.

## Tangkapan Layar

# Bibliografi

[1] T. Calders. (2009) Classification : Naive bayes classifier evaluation. Eindhoven University of Technology.

[2] K. Chen. Naïve bayes classifier. University of Manchester.

[3] E. Keogh. Naive bayes classifier. UCR.

[4] unknown. [Online]. Available: stackoverflow.com

[5] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and Statistics for Engineers and Sciantist*, 9th ed. Prentice Hall, 2012.