Tugas Individu Kecerdasan Buatan

Rangkuman Materi Algoritma Naïve Bayes

Dosen Pengampu: Khodijah Hulliyah M. Si.

Disusun Oleh
Muhammad Sigit Tri Pamungkas
11190910000013

Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Syarif Hidayatullah Jakarta
2021/2022

Algoritma Naïve Bayes

A. Pengertian

Algoritma Naive Bayes merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. Klasifikasi adalah salah satu teknik *data mining* untuk mengelompokkan data berdasarkan kelas atau label yang telah ditentukan. *Data mining* adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar. *Data mining* juga disebut sebagai serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data.

Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai Teorema *Bayes*. Pendekatan yang digunakan Teorema *Bayes* yaitu menghitung probabilitas sebuah kejadian pada kondisi tertentu (Lukito & Chrismanto, 2015). Dasar dari Teorema *Bayes* dinyatakan dalam persamaan (Bustami, 2013). Teorema tersebut dikombinasikan dengan *Naive* dimana diasumsikan kondisi antar atribut saling bebas. Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Metode ini menghitung sekumpulan peluang dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Terdapat dua asumsi utama pada metode Naïve Bayes:

1. Semua variabel memiliki prioritas yang sama pentingnya.
2. Semua variabel bersifat bebas secara statistik (nilai suatu variabel tidak terkait dengan nilai variabel lain).

Asumsi ini kebanyakan tidak selalu benar, namun dalam praktiknya walau asumsi tidak terpenuhi, metode Naïve Bayes tetap memberikan hasil yang baik. Karena asumsi metode Naïve Bayes adalah semua variabel bersifat bebas (saling independen/dapat berdiri sendiri). Rumus dari algoritma Naïve Bayes dapat dilihat pada persamaan 1.1 berikut.

$$p(H \mid X) = \frac{\mathrm{P(X \mid H)} \times \mathrm{P(H)}}{\mathrm{P(X)}} \qquad (1.1)$$

Keterangan:

X : Data dengan kelas yang belum diketahui

H : Hipotesis data, merupakan suatu kelas yang spesifik

P (H | X) : Probabilitas hipotesis H berdasar kondisi X (Probabilitas posterior)

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

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

P(X) : Probabilitas X

Algoritma Naïve Bayes sangat cocok digunakan untuk melakukan klasifikasi pada dataset bertipe *nominal*. Untuk dataset bertipe nominal, penghitungan algoritma Naïve Bayes menggunakan persamaan 1.1 diatas. Apabila dataset bertipe *numerik*, maka digunakan penghitungan distribusi *Gaussian*. Penghitungan distribusi Gaussian dapat dilihat dari persamaan 1.2, dimana dihitung terlebih dahulu nilai rata-rata μ sesuai persamaan 1.3, dan standard deviasi σ sesuai persamaan 1.4.

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\,.\,\sigma}\ exp^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{1.2}$$

$$\mu = \frac{\sum_i^n x_i}{\mathrm{n}} \tag{1.3}$$

$$\sigma = \sqrt{\frac{\sum_i^n (x_i\ -\ \mu)^2}{n-1}} \tag{1.4}$$

Langkah-langkah algoritma Naïve Bayes adalah sebagai berikut:

1. Siapkan dataset.
2. Hitung jumlah kelas pada data training.
3. Hitung jumlah kasus yang sama dengan kelas yang sama.
4. Kalikan semua hasil sesuai dengan data testing yang akan dicari kelasnya.
5. Bandingkan hasil per kelas, nilai tertinggi ditetapkan sebagai kelas baru.

B. Contoh Perhitungan Data Nominal Algoritma Naïve Bayes

Dataset yang digunakan pada perhitungan data nominal kali ini yaitu data nasabah asuransi. Dataset ini terdiri dari 8 atribut dan 1 klasifikasi dimana dataset tersebut dibagi menjadi dua yaitu *data training* dan *data testing*. Berikut merupakan langkah-langkah perhitungannya.

1. Siapkan dataset

Berikut merupakan dataset yang digunakan pada perhitungan data nominal kali ini.

Data Training

No	Nama	Jenis Kelamin	Usia	Status	Pekerjaan	Penghasilan	Masa Asuransi	Cara Pembayaran	Klasifikasi
1	Dani Lukman	Laki-Laki	30 - 40 Tahun	Kawin	Pns	< 25 Juta	>15 Tahun	Tahunan	Tidak Lancar
2	Evaliana	Perempuan	30 - 40 Tahun	Kawin	Pns	< 25 Juta	5 - 10 Tahun	Semesteran	Lancar
3	Rasyidah	Perempuan	20 - 29 Tahun	Kawin	Pegawai Swasta	< 25 Juta	5 - 10 Tahun	Triwulan	Tidak Lancar
4	Dina Saufika	Perempuan	30 - 40 Tahun	Belum Kawin	Pns	< 25 Juta	5 - 10 Tahun	Triwulan	Lancar
5	Wilsa Rizki	Laki-Laki	30 - 40 Tahun	Kawin	Wiraswasta	< 25 Juta	5 - 10 Tahun	Tahunan	Kurang Lancar
6	Irwanto	Laki-Laki	30 - 40 Tahun	Belum Kawin	Wiraswasta	> 50 Juta	11 - 15 Tahun	Semesteran	Lancar
7	Ade Gunawan	Laki-Laki	30 - 40 Tahun	Kawin	Pns	25 - 50 Juta	11 - 15 Tahun	Semesteran	Tidak Lancar
8	Fauziah	Perempuan	20 - 29 Tahun	Kawin	Wiraswasta	25 - 50 Juta	11 - 15 Tahun	Tahunan	Lancar
9	Zulaikha	Perempuan	20 - 29 Tahun	Kawin	Wiraswasta	< 25 Juta	11 - 15 Tahun	Triwulan	Tidak Lancar
10	Zulfahmi	Laki-Laki	20 - 29 Tahun	Kawin	Pns	< 25 Juta	11 - 15 Tahun	Triwulan	Kurang Lancar
11	Hidayatullah	Laki-Laki	30 - 40 Tahun	Belum Kawin	Wiraswasta	25 - 50 Juta	11 - 15 Tahun	Tahunan	Lancar
12	Nilam Sari	Perempuan	30 - 40 Tahun	Kawin	Wiraswasta	25 - 50 Juta	>15 Tahun	Tahunan	Kurang Lancar
13	Nahari Arifin	Laki-Laki	30 - 40 Tahun	Kawin	Wiraswasta	> 50 Juta	11 - 15 Tahun	Triwulan	Lancar
14	Yusnidar	Perempuan	>40 Tahun	Kawin	Pns	< 25 Juta	>15 Tahun	Semesteran	Kurang Lancar
15	Rizwan Hadi	Laki-Laki	20 - 29 Tahun	Belum Kawin	Pns	< 25 Juta	11 - 15 Tahun	Tahunan	Lancar
16	Rahmat Saputra	Laki-Laki	30 - 40 Tahun	Belum Kawin	Wiraswasta	< 25 Juta	11 - 15 Tahun	Semesteran	Lancar
17	M. Sahril	Laki-Laki	>40 Tahun	Kawin	Pegawai swasta	< 25 Juta	11 - 15 Tahun	Tahunan	Tidak Lancar
18	M. Irfan	Laki-Laki	30 - 40 Tahun	Kawin	Pegawai swasta	25 - 50 Juta	11 - 15 Tahun	Tahunan	Tidak Lancar
19	Tutri Wulandari	Perempuan	30 - 40 Tahun	Kawin	Wiraswasta	< 25 Juta	11 - 15 Tahun	Triwulan	Lancar
20	Leni Syamsiah	Perempuan	20 - 29 Tahun	Belum Kawin	Wiraswasta	25 - 50 Juta	5 - 10 Tahun	Bulanan	Tidak Lancar

Data Testing

No	Nama	Jenis Kelamin	Usia	Status	Pekerjaan	Penghasilan	Masa Asuransi	Cara Pembayaran	Klasifikasi
1	Syafi Arkan	Laki-Laki	30 - 40 Tahun	Kawin	wiraswasta	25 - 50 Juta	11 - 15 Tahun	Semesteran	????

2. Hitung jumlah kelas pada data training

Kelas pada *data training* diatas yaitu atribut klasifikasi yang terdiri dari 3 (tiga) klasifikasi yaitu Lancar, Kurang Lancar, dan Tidak Lancar, sehingga probabilitasnya adalah sebagai berikut.

Jumlah klasifikasi Lancar = 9

Jumlah klasifikasi Kurang Lancar = 4

Jumlah klasifikasi Tidak Lancar = 7

Maka:

P(Y = Lancar) = $\frac{9}{20}$ = 0,45

P(Y = Kurang Lancar) = $\frac{4}{20}$ = 0,20

P(Y = Tidak Lancar) = $\frac{7}{20}$ = 0,35

3. Hitung jumlah kasus yang sama dengan kelas yang sama

a. Atribut Jenis Kelamin

P(Jenis Kelamin = Laki-laki | Y = Lancar) = $\frac{5}{9}$ = 0,56

P(Jenis Kelamin = Laki-laki | Y = Kurang Lancar) = $\frac{2}{4}$ = 0,50

P(Jenis Kelamin = Laki-laki | Y = Tidak Lancar) = $\frac{4}{7}$ = 0,57

b. Atribut Usia

P(Usia = 30 - 40 Tahun | Y = Lancar) = $\frac{7}{9}$ = 0,78

P(Usia = 30 - 40 Tahun | Y = Kurang Lancar) = $\frac{2}{4}$ = 0,50

P(Usia = 30 - 40 Tahun | Y = Tidak Lancar) = $\frac{3}{7}$ = 0,43

c. Atribut status

P(Status = Kawin | Y = Lancar) = $\frac{4}{9}$ = 0,44

P(Status = Kawin | Y = Kurang Lancar) = $\frac{4}{4}$ = 1

P(Status = Kawin | Y = Tidak Lancar) = $\frac{6}{7}$ = 0,86

d. Atribut Pekerjaan

P(Pekerjaan = Wiraswasta | Y = Lancar) = $\frac{6}{9}$ = 0,67

P(Pekerjaan = Wiraswasta | Y = Kurang Lancar) = $\frac{2}{4}$ = 0,50

P(Pekerjaan = Wiraswasta | Y = Tidak Lancar) = $\frac{2}{7}$ = 0,29

e. Atribut Penghasilan

P(Penghasilan = 25 - 50 Juta | Y = Lancar) = $\frac{2}{9}$ = 0,22

P(Penghasilan = 25 - 50 Juta | Y = Kurang Lancar) = $\frac{1}{4}$ = 0,25

P(Penghasilan = 25 - 50 Juta | Y = Tidak Lancar) = $\frac{3}{7}$ = 0,43

f. Atribut Masa Asuransi

P(Masa Asuransi = 11 - 15 Tahun | Y = Lancar) = $\frac{7}{9}$ = 0,78

P(Masa Asuransi = 11 - 15 Tahun | Y = Kurang Lancar) = $\frac{1}{4}$ = 0,25

P(Masa Asuransi = 11 - 15 Tahun | Y = Tidak Lancar) = $\frac{4}{7}$ = 0,57

g. Atribut Cara Pembayaran

P(Cara Pembayaran = Semesteran | Y = Lancar) = $\frac{3}{9}$ = 0,33

P(Cara Pembayaran = Semesteran | Y = Kurang Lancar) = $\frac{1}{4}$ = 0,25

P(Cara Pembayaran = Semesteran | Y = Tidak Lancar) = $\frac{1}{7}$ = 0,14

4. Kalikan semua hasil sesuai dengan data testing yang akan dicari kelasnya

a. P(Laki-Laki| Lancar) × P(30-40 Tahun| Lancar) × P(Kawin| Lancar) × P(Wiraswasta| Lancar) × P(25-50 Juta|Lancar) × P(11-15 Tahun| Lancar) × P(Semesteran| Lancar) × P(Lancar)

= $\frac{5}{9} \times \frac{7}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{2}{9} \times \frac{7}{9} \times \frac{3}{9} \times \frac{9}{20}$

= 0,56 × 0,78 × 0,44 × 0,67 × 0,22 × 0,78 × 0,33 × 0,45

= **0,0032**

b. P(Laki-Laki| Kurang Lancar) × P(30-40 Tahun| Kurang Lancar) × P(Kawin| Kurang Lancar) × P(Wiraswasta| Kurang Lancar) × P(25-50 Juta| Kurang Lancar) × P(11-15 Tahun| Kurang Lancar) × P(Semesteran| Kurang Lancar) × P(Kurang Lancar)

$= \frac{2}{4} \times \frac{2}{4} \times \frac{4}{4} \times \frac{2}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{4}{20}$

= 0,50 × 0,50 × 1 × 0,50 × 0,25 × 0,25 × 0,25 × 0,20

= **0,0004**

c. P(Laki-Laki| Tidak Lancar) × P(30-40 Tahun| Tidak Lancar) × P(Kawin| Tidak Lancar) × P(Wiraswasta| Tidak Lancar) × P(25-50 Juta| Tidak Lancar) × P(11-15 Tahun| Tidak Lancar) × P(Semesteran| Tidak Lancar) × P(Tidak Lancar)

$= \frac{4}{7} \times \frac{3}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{3}{7} \times \frac{4}{7} \times \frac{1}{7} \times \frac{7}{20}$

= 0,57 × 0,43 × 0,86 × 0,29 × 0,43 × 0,57 × 0,14 × 0,35

= **0,0007**

5. **Bandingkan hasil per kelas, nilai tertinggi ditetapkan sebagai kelas baru**

Dari hasil diatas, dapat dilihat bahwa nilai probabilitas tertinggi yaitu pada kelas P(Lancar) maka dapat disimpulkan bahwa **status calon nasabah pada *data testing* tersebut masuk ke dalam klasifikasi "Lancar"**.

C. Contoh Perhitungan Data Numerik dan Nominal Algoritma Naïve Bayes

Jika contoh diatas hanya menggunakan data nominal saja, maka pada contoh dibawah ini akan digunakan data numerik dan nominal dengan penyelesaian menggunakan algoritma naïve bayes. Data bertipe numerik akan diselesaikan menggunakan persamaan 1.2 sedangkan data bertipe nominal akan diselesaikan menggunakan persamaan 1.1. Berikut merupakan langkah-langkah dalam penyelesaiannya.

1. **Siapkan dataset**

Dataset yang digunakan pada perhitungan kali ini yaitu dataset bermain golf yang dibagi menjadi dua bagian, yaitu *data training* dan *data testing* dimana datasetnya terdiri dari 4 (empat) atribut bertipe data nominal dan numerik serta 1 (satu) kelas bertipe data nominal.

Data Training

No	Cuaca	Temperatur	Kelembaban	Angin	Play
1	cerah	85	85	tidak	tidak
2	cerah	80	90	ada	tidak
3	mendung	83	78	tidak	ya
4	hujan	70	96	tidak	ya
5	hujan	68	80	tidak	ya
6	hujan	65	70	ada	tidak
7	mendung	64	65	ada	ya
8	cerah	72	95	tidak	tidak
9	cerah	69	70	tidak	ya
10	hujan	75	80	tidak	ya
11	cerah	75	70	ada	ya
12	mendung	72	90	ada	ya
13	mendung	81	75	tidak	ya
14	hujan	71	80	ada	tidak

Data Testing

No	Cuaca	Temperature	Kelembaban	Angin	Play
1	cerah	73	80	tidak	????

2. Hitung jumlah kelas pada data training

Kelas pada *data training* diatas terdiri dari 2 (dua) kelas yaitu bermain golf dan tidak bermain golf. Probabilitasnya dapat dilihat sebagai berikut.

Jumlah kelas bermain golf = 9

Jumlah kelas tidak bermain golf = 5

Maka:

P(play = ya) = $\frac{9}{14}$ = 0,64

P(play = tidak) = $\frac{5}{14}$ = 0,36

3. Hitung jumlah kasus yang sama dengan kelas yang sama

a. Atribut Cuaca

P(Cuaca = cerah | play = ya) = $\frac{2}{9}$ = 0,22

P(Cuaca = cerah | play = tidak) = $\frac{3}{5}$ = 0,60

b. Atribut Temperatur

μ_{play} = ya = $\frac{83 + 70 + 68 + 64 + 69 + 75 + 75 + 72 + 81}{9}$ = 73

μ_{play} = tidak = $\frac{85 + 80 + 65 + 72 + 71}{5}$ = 74,6

σ_{play} = ya = $\sqrt{\frac{(83-73)^2+(70-73)^2+\cdots+(81-73)^2}{9-1}}$ = 6,16

σ_{play} = tidak = $\sqrt{\frac{(85-74,6)^2+(80-74,6)^2+\cdots+(71-74,6)^2}{5-1}}$ = 7,89

P(temp = 73 | play = ya) = $\frac{1}{\sqrt{2\pi} \times 6,16}\ exp^{\frac{-(73-73)^2}{2\times(6,16)^2}}$ = 0,060

P(temp = 73 | play = tidak) = $\frac{1}{\sqrt{2\pi} \times 7,89}\ exp^{\frac{-(73-74,6)^2}{2\times(7,89)^2}}$ = 0,050

c. Atribut Kelembaban

μ_{play} = ya = $\frac{78 + 96 + 80 + 65 + 70 + 80 + 70 + 90 + 75}{9}$ = 78,22

μ_{play} = ya = $\frac{85 + 90 + 70 + 95 + 80}{5}$ = 84

σ_{play} = ya = $\sqrt{\frac{(78-78,22)^2+(96-78,22)^2+\cdots+(75-78,22)^2}{9-1}}$ = 9,84

σ_{play} = ya = $\sqrt{\frac{(85-84)^2+(90-84)^2+\cdots+(80-84)^2}{5-1}}$ = 9,62

P(lembab = 80 | play = ya) = $\frac{1}{\sqrt{2\pi} \times 9,84}\ exp^{\frac{-(80-78,22)^2}{2\times(9,84)^2}}$ = 0,040

P(lembab = 80 | play = tidak) = $\frac{1}{\sqrt{2\pi} \times 9,62}\ exp^{\frac{-(80-84)^2}{2\times(9,62)^2}}$ = 0,042

d. Atribut Angin

P(Angin = tidak | play = ya) = $\frac{6}{9}$ = 0,67

P(Angin = tidak | play = tidak) = $\frac{2}{5}$ = 0,40

4. Kalikan semua hasil sesuai dengan data testing yang akan dicari kelasnya

P(X | play = ya) = 0,22 × 0,060 × 0,040 × 0,67 = **0,00036**

P(X | play = tidak) = 0,60 × 0,050 × 0,042 × 0,40 = **0,00050**

P(play = ya | X) = 0,00036 × 0,64 = **0,00023148**

P(play = tidak | X) = 0,00050 × 0,36 = **0,00017850**

5. **Bandingkan hasil per kelas, nilai tertinggi ditetapkan sebagai kelas baru**

   Dari perhitungan probabilitas bermain golf dan probabilitas tidak bermain golf pada perhitungan diatas, maka dapat disimpulkan bahwa data cuaca = cerah, temperatur = 73, kelembaban = 80, dan angin = tidak, masuk ke dalam kelas **bermain golf**, karena probabilitas bermain golf (0,00023148) lebih tinggi dibandingkan probabilitas tidak bermain golf (0,00017850).

D. Contoh Penyelesaian Menggunakan Bahasa Pemrograman Python

Selain secara manual, kita juga bisa menggunakan bahasa pemrograman python untuk menyelesaikan perhitungan diatas. Kita bisa menggunakan *library* yang sudah disediakan yaitu *library* scikit-learn pada perhitungan kali ini. Dengan soal yang sama pada perhitungan data numerik dan nominal diatas maka langkah-langkah penyelesaiannya dapat dilihat sebagai berikut.

1. Pertama kita *import* beberapa *library* yang diperlukan seperti pandas, numpy, dan sebagainya. Kemudian kita *import* data latih yaitu dataset yang kita gunakan untuk melatih program. Untuk data nominal kita rubah ke bentuk numerik dengan ketentuan cuaca cerah = 0, mendung = 1, dan hujan = 2. Kemudian untuk tidak ada angin = 0 dan ada angin = 1. Untuk program dan *output*nya dapat dilihat pada gambar berikut.

```
[1]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     #memasukan data latih
     datalatih = pd.read_excel("data training numerik.xlsx")
     datalatih.head(15)

     #cuaca = 0 == cerah
     #cuaca = 1 == mendung
     #cuaca = 2 == hujan

     #angin = 0 == tidak
     #angin = 1 == ada
```

	Cuaca	Temperatur	Kelembaban	Angin	Play
0	0	85	85	0	tidak
1	0	80	90	1	tidak
2	1	83	78	0	ya
3	2	70	96	0	ya
4	2	68	80	0	ya
5	2	65	70	1	tidak
6	1	64	65	1	ya
7	0	72	95	0	tidak
8	0	69	70	0	ya
9	2	75	80	0	ya
10	0	75	70	1	ya
11	1	72	90	1	ya
12	1	81	75	0	ya
13	2	71	80	1	tidak

2. Kita bisa melihat info lengkap dari dataset yang sudah kita masukkan seperti dibawah ini.

```
[2] datalatih.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14 entries, 0 to 13
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Cuaca       14 non-null     int64 
 1   Temperatur  14 non-null     int64 
 2   Kelembaban  14 non-null     int64 
 3   Angin       14 non-null     int64 
 4   Play        14 non-null     object
dtypes: int64(4), object(1)
memory usage: 688.0+ bytes
```

3. Untuk proses *training* kita masukkan dataset ke dalam variabel x. Karena atribut Play digunakan sebagai klasifikasi maka atribut Play kita "*drop*" dan akan kita letakkan di variabel yang berbeda.

```
[3]  x = datalatih.drop(["Play"], axis=1)
     x.head(15)
```

	Cuaca	Temperatur	Kelembaban	Angin
0	0	85	85	0
1	0	80	90	1
2	1	83	78	0
3	2	70	96	0
4	2	68	80	0
5	2	65	70	1
6	1	64	65	1
7	0	72	95	0
8	0	69	70	0
9	2	75	80	0
10	0	75	70	1
11	1	72	90	1
12	1	81	75	0
13	2	71	80	1

4. Untuk atribut Play kita masukkan ke dalam variabel y sebagai klasifikasi dari dataset tersebut.

```
[4]  y = datalatih["Play"]
     y.head(15)
```

```
0     tidak
1     tidak
2       ya
3       ya
4       ya
5     tidak
6       ya
7     tidak
8       ya
9       ya
10      ya
11      ya
12      ya
13    tidak
Name: Play, dtype: object
```

5. Kemudian kita *training* data tersebut menggunakan fungsi GaussianNB yang terdapat pada *library* scikit-learn seperti berikut. Kita masukkan data *training* yang sudah kita masukkan ke dalam variabel x dan y.

```
[11] from sklearn.model_selection import train_test_split
     # Import Gaussian Naive Bayes model
     from sklearn.naive_bayes import GaussianNB
     # Mengaktifkan/memanggil/membuat fungsi klasifikasi Naive bayes
     modelnb = GaussianNB()
     # Memasukkan data training pada fungsi klasifikasi naive bayes
     nbtrain = modelnb.fit(x, y)
```

6. Selanjutnya kita *import* data terakhir pada dataset yang kita gunakan sebagai pengujian dari program. Kita set atribut Play = ya karena sesuai dengan perhitungan manual diatas dimana hasilnya yaitu "ya". Hasil tersebut nanti akan menjadi perbandingan dengan hasil prediksi dari program. Datanya dapat dilihat pada gambar berikut.

```
[12] datauji = pd.read_excel("data testing numerik.xlsx")
     datauji.head(2)
```

	Cuaca	Temperatur	Kelembaban	Angin	Play
0	0	73	80	0	ya

7. Kemudian kita buat variabel x_test berisi dataset yang ingin kita *test* kepada program. Seperti langkah sebelumnya kita *drop* atribut Play.

```
[13] x_test = datauji.drop(["Play"], axis=1)
     x_test.head(2)
```

	Cuaca	Temperatur	Kelembaban	Angin
0	0	73	80	0

8. Setelah itu kita bisa membuat prediksi dari dataset pengujian tersebut. Hasil pengujian kita letakkan ke dalam variabel Y_predict kemudian kita *output*. Hasil dari prediksi program adalah "ya", sesuai dengan perhitungan manual diatas.

```
[14] y_uji = datauji["Play"]
     y_uji.head(2)

     0    ya
     Name: Play, dtype: object
```

```
[15] Y_predict = nbtrain.predict(x_test)
     print("Prediksi Naive Bayes : ",Y_predict)

     Prediksi Naive Bayes :  ['ya']
```

9. Kemudian untuk akurasinya bisa kita hitung seperti berikut. Karena kita hanya menggunakan 1 data dan prediksinya benar maka hasil akurasinya yaitu “1.0”.

```
[16] from sklearn.metrics import accuracy_score
     accuracy= accuracy_score(y_uji, Y_predict)
     print("Akurasi Naive Bayes : ",accuracy)

     Akurasi Naive Bayes :  1.0
```

10. Kemudian untuk data akurasi yang lebih lengkap bisa kita lihat seperti gambar dibawah. Karena kita hanya menggunakan 1 data pengujian maka hasilnya sama yaitu “1.0”.

```
[17] # Menghitung nilai akurasi dari klasifikasi naive bayes
     from sklearn.metrics import classification_report
     print(classification_report(y_uji, Y_predict))

                   precision    recall  f1-score   support

               ya       1.00      1.00      1.00         1

         accuracy                           1.00         1
        macro avg       1.00      1.00      1.00         1
     weighted avg       1.00      1.00      1.00         1
```

Program diatas bisa kita gunakan untuk memprediksi data pengujian yang lebih banyak sehingga nilai akurasinya pun akan lebih beragam. Kemudian dengan menambahkan dataset *training* menjadi ratusan atau ribuan juga dapat menambah jumlah akurasi dari prediksi program tersebut.

Untuk *source code* program diatas dapat diakses pada: https://github.com/msigit26/Naive-Bayes-Classifier

REFERENSI

[1] Suntoro, Joko. 2019. *DATA MINING: Algoritme dan Implementasi Menggunakan Bahasa Pemrograman PHP*. Jakarta: Elex Media Computindo. ISBN: 978-602-04-9881-2.

[2] I. Cholissodin, E. Riyandani. 2016. *ANALISIS BIG DATA (Teori & Aplikasi) "Big Data vs Big Information vs Big Knowledge"* Versi 1.01. Malang: Fakultas Ilmu Komputer, Universitas Brawijaya.

[3] Bustami, 2013. *Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi*. Jurnal TECHSI Vol. 3 No. 2.