Data Mining
Naïve Bayes Classifier
Week 5

Ardytha Luthfiarta, M.Kom, MCS

Introduction to Bayesian Theory

- Statistical Method for classification
- Supervised learning method
- Can solve problem involving both categorical and continues valued attributes
- Named after Thomas Bayes, who proposed the Bayes Theorem

Naïve Bayes Classifier

- Simple Probabilistic Technic based on Bayesian Theorm
  - Merupakan teknik prediksi (klasifikasi data) berbasis probabilistic sederhana dengan menggunakan teorema bayes.
- Independence Feature Based
  - Teorema bayes menggunakan asumsi independensi (ketidaktergantungan) sehingga naïve bayes menggunakan model fitur yang independen

Naïve Bayes Classifier

- Prediksi Bayes didasarkan pada teorema bayes dengan formula umum sebagai berikut :
- P(H|E) = $\frac{P(E|H)\ x\ P(H)}{P(E)}$, dimana
  - P(H|E) adalah probabilitas akhir bersyarat (conditional probability) suatu hipotesis H terjadi jika diberikan bukti (evidence) E terjadi.
  - P(E|H) adalah sebuah bukti E terjadi akan mempengaruhi H.
  - P(H) adalah probabilitas awal (priori) Hipotesis H tanpa memandang bukti apa pun.
  - P(E) adalah probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti yang lain.

Contoh

- Dalam suatu peramalan cuaca untuk memperkirakan terjadinya hujan, ada factor yang mempengaruhi terjadinya hujan, yaitu mendung.
- Jika diterapkan dalam Naïve Bayes, probabilitas terjadinya hujan, jika bukti mendung sudah diamati, dinyatakan dengan
- P(H | E) = $\frac{P(E|H)\ x\ P(H)}{P(E)}$ →
- P(Hujan | Mendung) = $\frac{P(Mendung|Hujan)\ x\ P(Hujan)}{P(Mendung)}$

Contoh

- Untuk contoh diatas jika ditambahkan bukti suhu udara dan angin, bentuknya berubah menjadi :
- P(Hujan | Mendung, Suhu, Angin) =

$$\frac{P(Mendung|Hujan)\ x\ P(Suhu|Hujan)\ x\ P(Angin|Hujan) x\ P(Hujan)}{P(Mendung)\ x\ \ P(Suhu)\ \ x\ \ P(Angin)}$$

Naïve Bayes Classifier – Konsep Dasar

- Ide dasar dari aturan Bayes adalah bahwa hasil dari hipotesis atau peristiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati. Ada beberapa hal penting dari aturan Bayes tersebut, yaitu
  - Sebuah probabilitas awal/priori H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati
  - Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Data training “All Electronics customer database”

Id	Age	Income	Student	Credit_rating	Class: buys_computer
1	<=30	High	No	Fair	No
2	<=30	High	No	Excellent	No
3	31..40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31..40	Low	Yes	Excellent	Yes
8	<=30	Medium	No	Fair	No
9	<=30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	<=30	Medium	Yes	Excellent	Yes
12	31..40	Medium	No	Excellent	Yes
13	31..40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

Keterangan

- Terdapat dua class dari klasifikasi yang dibentuk yaitu :
  - C1 => buys_computer = yes
  - C2 => buys_computer = no
- Misal terdapat data X (belum diketahui *class*-nya).
- X = (age="<=30", income="Medium",

  student="Yes", credit_rating="Fair")

Penyelesaian (1)

LANGKAH 1

- Hitung P(Ci) untuk i=1, 2.
- P(Ci) merupakan prior probability untuk setiap class berdasar data contoh:
  - P(buys_computer="yes") = 9/14 = 0.643
  - P(buys_computer="no") = 5/14 = 0.357

Penyelesaian (2)

- **LANGKAH 2**

  Hitung P(X | Ci), untuk i =1, 2

  - P(age="<=30" | buys_computer="yes") = 2/9= **0.222**
  - P(age="<=30" | buys_computer="no") = 3/5 = **0.600**

  - P(income="medium" | buys_computer="yes")=4/9= **0.444**
  - P(income="medium" | buys_computer="no")=2/5= **0.400**

  - P(student="yes" | buys_computer="yes") = 6/9= **0.667**
  - P(student="yes" | buys_computer="no") = 1/5= **0.200**

  - P(credit_rating="fair" | buys_computer="yes")=6/9= **0.667**
  - P(credit_rating="fair" | buys_computer="no")=2/5= **0.400**

Penyelesaian (2)

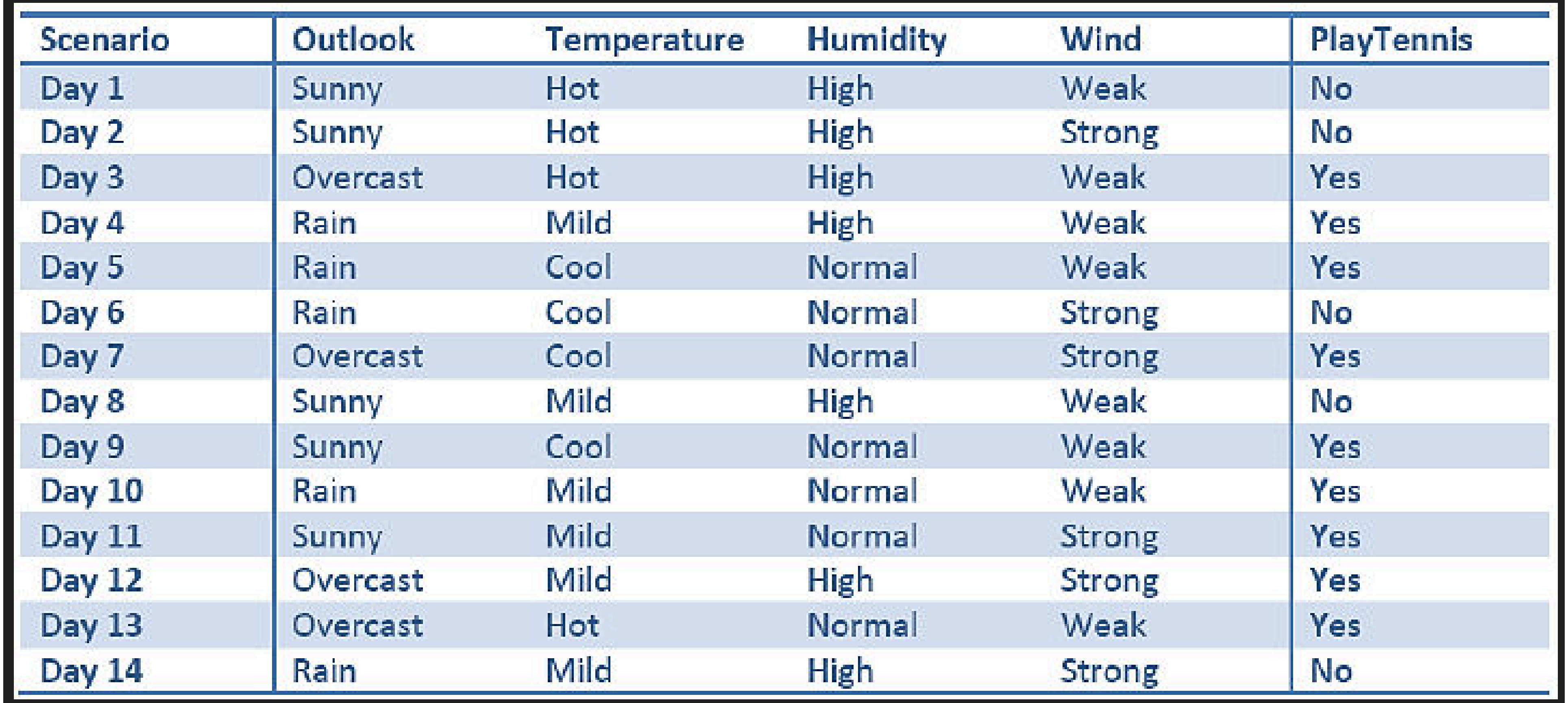
- **Langkah ke-3 --**
  - P(X | buys_computer=“yes”) = 0.222*0.444*0.677*0.677 = **0.044**
  - P(X | buys_computer=“no”) = 0.600*0.400*0.200*0.400= 0.019
- **Langkah ke-4--**
  - P(X | buys_computer=“yes”)*P(buys_computer=“yes”) = 0.044 * 0.643 = **0.028**
  - P(X | buys_computer=“no”)*P(buys_computer=“no”) = 0.019 * 0.357 = 0.007
- **Langkah ke-5 --**

  Kesimpulan: ***buys_computer = “yes”***

Latihan

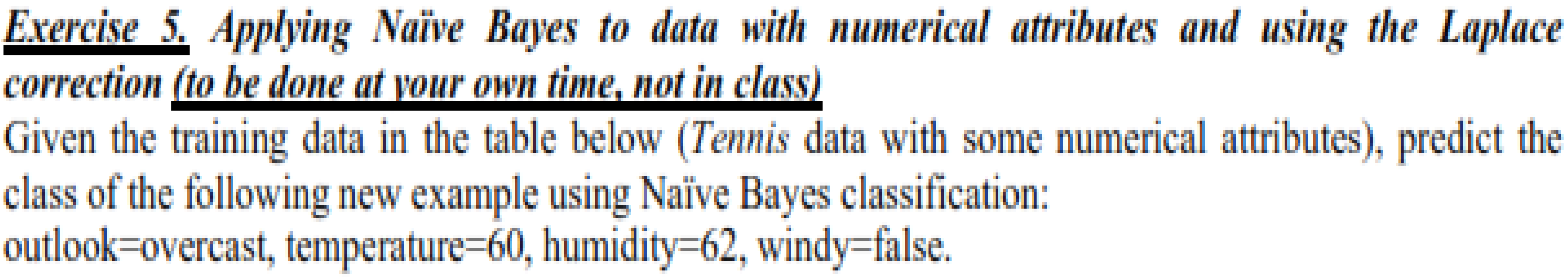
Tentukan klas label dari X:

X = (Outlook=Rain, Temperature=Cool, Humidity=High, Wind=Weak)



Scenario	Outlook	Temperature	Humidity	Wind	PlayTennis
Day 1	Sunny	Hot	High	Weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 3	Overcast	Hot	High	Weak	Yes
Day 4	Rain	Mild	High	Weak	Yes
Day 5	Rain	Cool	Normal	Weak	Yes
Day 6	Rain	Cool	Normal	Strong	No
Day 7	Overcast	Cool	Normal	Strong	Yes
Day 8	Sunny	Mild	High	Weak	No
Day 9	Sunny	Cool	Normal	Weak	Yes
Day 10	Rain	Mild	Normal	Weak	Yes
Day 11	Sunny	Mild	Normal	Strong	Yes
Day 12	Overcast	Mild	High	Strong	Yes
Day 13	Overcast	Hot	Normal	Weak	Yes
Day 14	Rain	Mild	High	Strong	No

Soal



Exercise 5. ***Applying Naïve Bayes to data with numerical attributes and using the Laplace correction (to be done at your own time, not in class)***

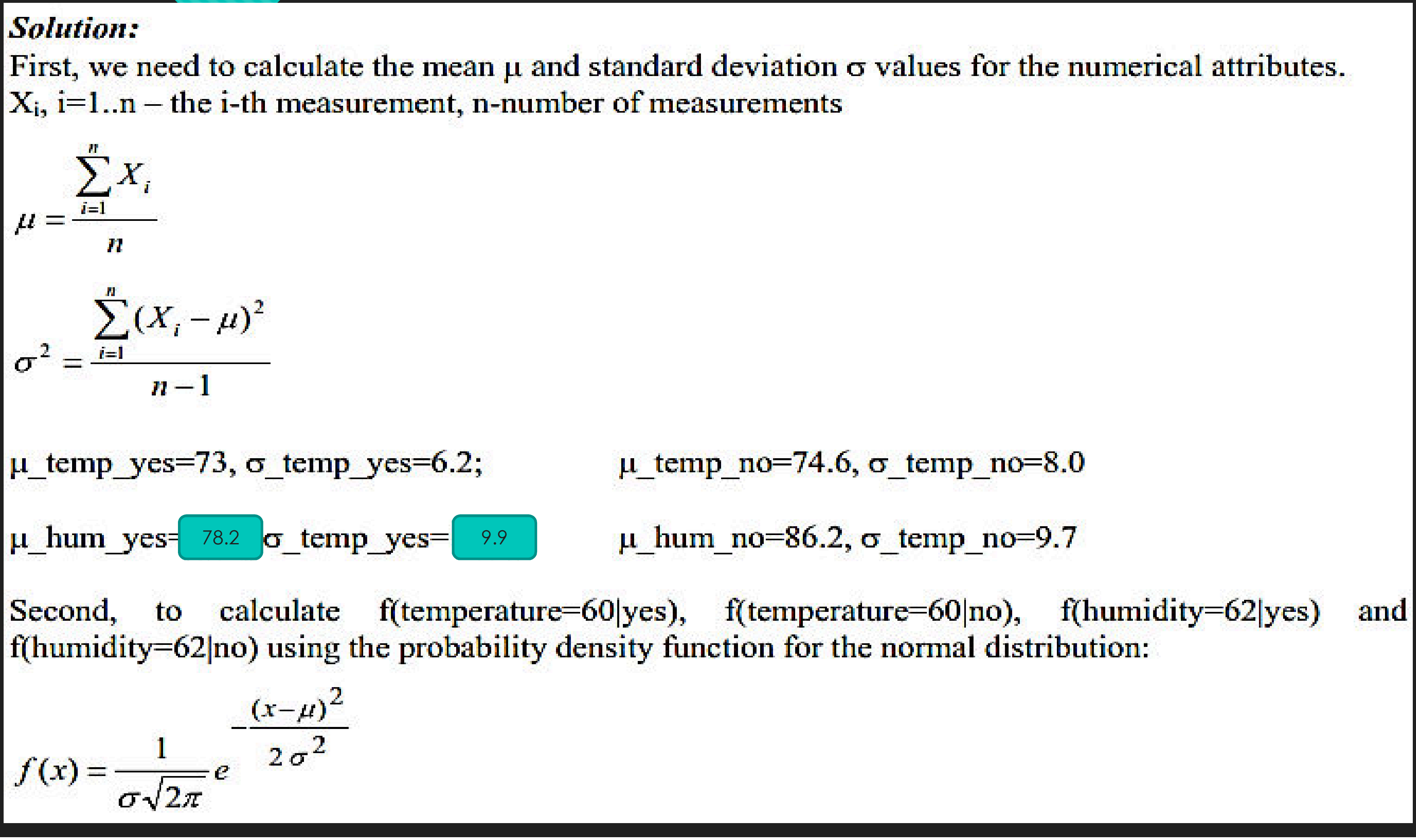
Given the training data in the table below (*Tennis* data with some numerical attributes), predict the class of the following new example using Naïve Bayes classification:
outlook=overcast, temperature=60, humidity=62, windy=false.

Dataset dengan Atribut Numerik

OUTLOOK	TEMPERATURE	HUMIDITY	WIND	PLAY
Sunny	85	85	FALSE	no
Sunny	80	90	TRUE	no
Overcast	83	78	FALSE	yes
Rain	70	96	FALSE	yes
Rain	68	80	FALSE	yes
Rain	65	70	TRUE	no
Overcast	64	65	TRUE	yes
Sunny	72	95	FALSE	no
Sunny	69	70	FALSE	yes
Rain	75	80	FALSE	yes
Sunny	75	70	TRUE	yes
Overcast	72	90	TRUE	yes
Overcast	81	75	FALSE	yes
Rain	71	91	TRUE	no

Langkah Pertama

Outlook = overcast, temperature = 60, humidity = 62, windy = false, class ???



Solution:

First, we need to calculate the mean μ and standard deviation σ values for the numerical attributes.
X_i, i=1..n – the i-th measurement, n-number of measurements

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

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

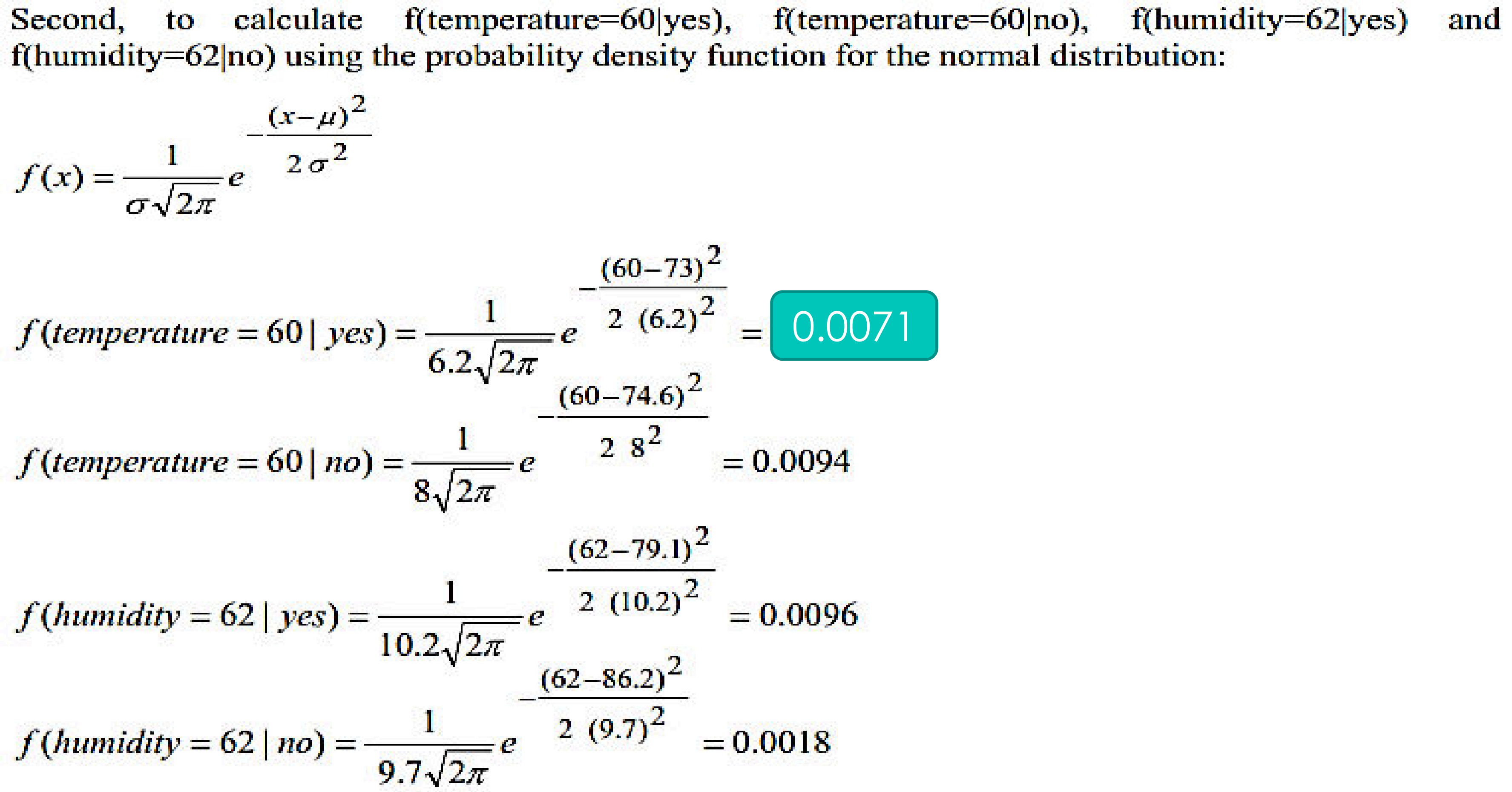
μ_temp_yes=73, σ_temp_yes=6.2; μ_temp_no=74.6, σ_temp_no=8.0

μ_hum_yes= 78.2 σ_temp_yes= 9.9 μ_hum_no=86.2, σ_temp_no=9.7

Second, to calculate f(temperature=60|yes), f(temperature=60|no), f(humidity=62|yes) and f(humidity=62|no) using the probability density function for the normal distribution:

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

Langkah Kedua



Second, to calculate f(temperature=60|yes), f(temperature=60|no), f(humidity=62|yes) and f(humidity=62|no) using the probability density function for the normal distribution:

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

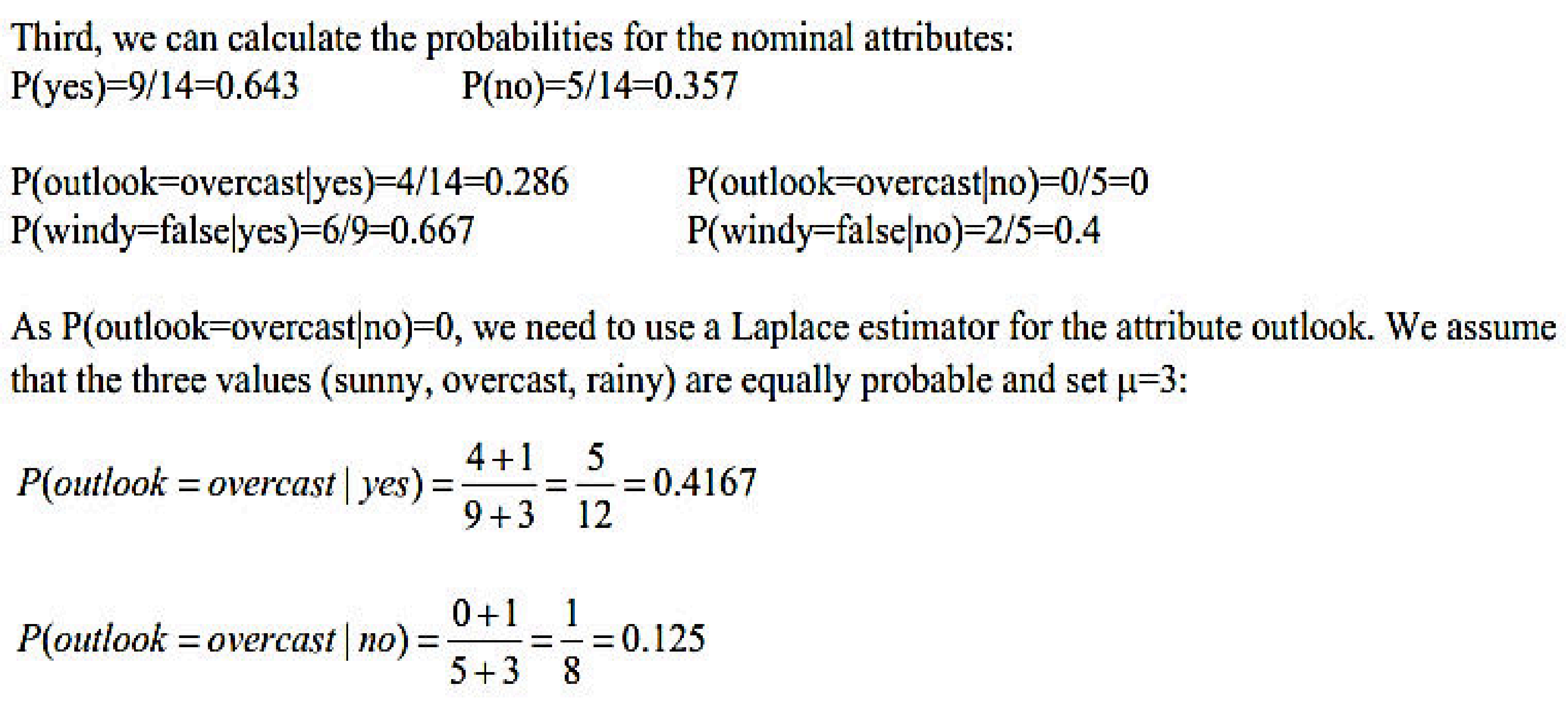
$$f(temperature = 60 \mid yes) = \frac{1}{6.2\sqrt{2\pi}} e^{-\frac{(60-73)^2}{2\ (6.2)^2}} = \boxed{0.0071}$$

$$f(temperature = 60 \mid no) = \frac{1}{8\sqrt{2\pi}} e^{-\frac{(60-74.6)^2}{2\ 8^2}} = 0.0094$$

$$f(humidity = 62 \mid yes) = \frac{1}{10.2\sqrt{2\pi}} e^{-\frac{(62-79.1)^2}{2\ (10.2)^2}} = 0.0096$$

$$f(humidity = 62 \mid no) = \frac{1}{9.7\sqrt{2\pi}} e^{-\frac{(62-86.2)^2}{2\ (9.7)^2}} = 0.0018$$

Langkah Ke-tiga



Third, we can calculate the probabilities for the nominal attributes:

P(yes)=9/14=0.643 P(no)=5/14=0.357

P(outlook=overcast|yes)=4/14=0.286 P(outlook=overcast|no)=0/5=0

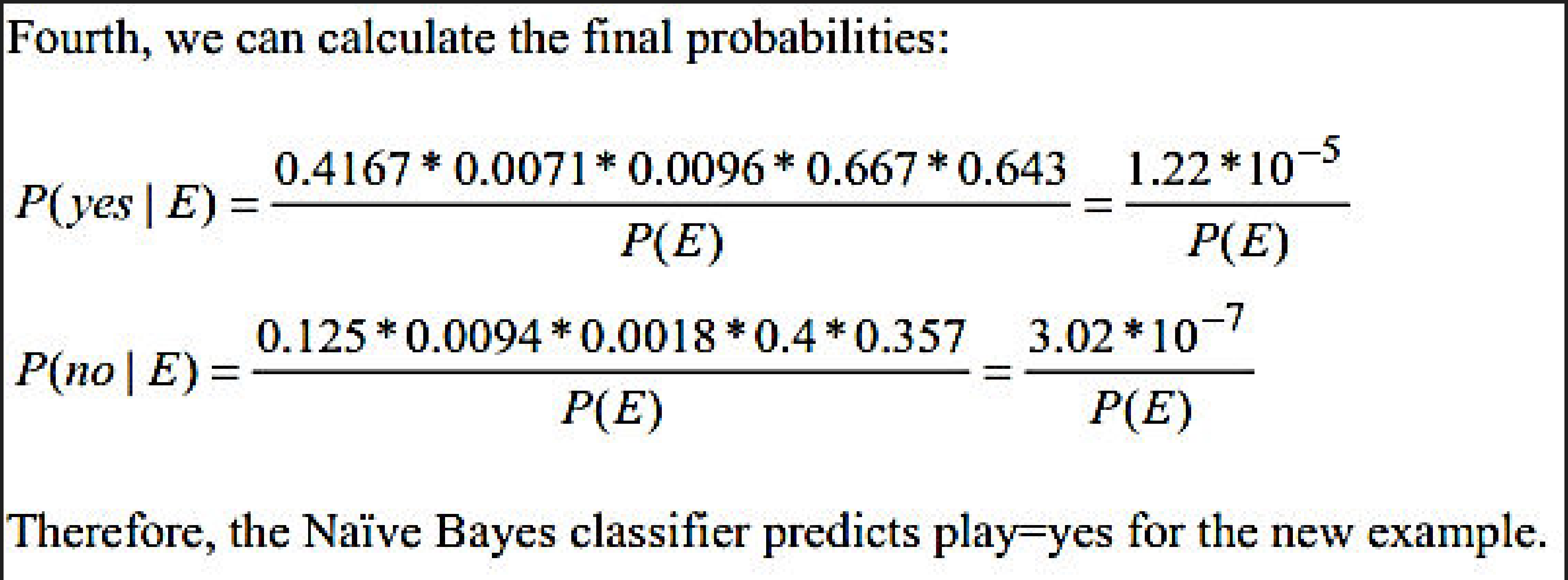
P(windy=false|yes)=6/9=0.667 P(windy=false|no)=2/5=0.4

As P(outlook=overcast|no)=0, we need to use a Laplace estimator for the attribute outlook. We assume that the three values (sunny, overcast, rainy) are equally probable and set μ=3:

$$P(outlook = overcast \mid yes) = \frac{4+1}{9+3} = \frac{5}{12} = 0.4167$$

$$P(outlook = overcast \mid no) = \frac{0+1}{5+3} = \frac{1}{8} = 0.125$$

Langkah ke-Empat



Fourth, we can calculate the final probabilities:

$$P(yes \mid E) = \frac{0.4167 * 0.0071 * 0.0096 * 0.667 * 0.643}{P(E)} = \frac{1.22 * 10^{-5}}{P(E)}$$

$$P(no \mid E) = \frac{0.125 * 0.0094 * 0.0018 * 0.4 * 0.357}{P(E)} = \frac{3.02 * 10^{-7}}{P(E)}$$

Therefore, the Naïve Bayes classifier predicts play=yes for the new example.

Dataset dengan Atribut Numerik

x : outlook = rain, temp = 72, hum = 75, wind = true, play ??

outlook	temperature	humidity	wind	play
Sunny	85	85	FALSE	no
Sunny	80	90	TRUE	no
Overcast	83	78	FALSE	yes
Rain	70	96	FALSE	yes
Rain	68	80	FALSE	yes
Rain	65	70	TRUE	no
Overcast	64	65	TRUE	yes
Sunny	72	95	FALSE	no
Sunny	69	70	FALSE	yes
Rain	75	80	FALSE	yes
Sunny	75	70	TRUE	yes
Overcast	72	90	TRUE	yes
Overcast	81	75	FALSE	yes
Rain	71	91	TRUE	no