4. Aplicación de un modelo Naive Bayes

Dado el siguiente modelo NB para clasificar las frutas en **Pera** o **Manzana** en base a los atributos **Color** y **Esfericidad:**

Atributo / Valor	Clase **Pera**	Clase **Manzana**
Color = Amarillo	0.9	0.1
Color = Mezcla	0.0	0.6
Color = Rojo	0.1	0.3
Esfericidad (μ)	0.6	0.8
Esfericidad (σ)	0.3	0.2

a) Si ambas clases son equiprobables (las probabilidades de clase a priori son P(**Pera**)=0.5 y P(**Manzana**)=0.5), y no se utiliza corrección de Laplace, indicar cómo clasificaría los siguientes 2 ejemplos, incluyendo los cálculos realizados, en la siguiente tabla:

P(x | Pera) = P(color = **Amarillo** ∩ Esfericidad = 0.6 | Pera)

Al ser una intersección de atributos independientes calculamos:

P(x | Pera) = P(color = Amarillo | Pera) * P(Esfericidad = 0.6 | Pera)
P(x | Pera) = 0.9 * pdf(x=0.6 | μ=0.6, σ=0.3)
P(x | Pera) = 0.9 * 1.32980760 = 1.19682684

P(x | Manzana) = P(color = **Amarillo** ∩ Esfericidad = 0.6 | Manzana)

Al ser una intersección de atributos independientes calculamos:

P(x | Manzana) = P(color = Amarillo | Manzana) * P(Esfericidad = 0.6 | Manzana)
P(x | Manzana) = 0.1 * pdf(x=0.6 | μ=0.8, σ=0.2)
P(x | Manzana) = 0.1 * 1.20985362 = 0.120985362

P(x | Pera) * P(Pera) = 1.19682684 * 0.5 = 0.59841342 = P(Pera | x)
P(x | Manzana) * P(Manzana) = 0.120985362 * 0.5 = 0.060492681 = P(Manzana | x)

P(x | Pera) = P(color = **Mezcla** ∩ Esfericidad = 0.8 | Pera)

Al ser una intersección de atributos independientes calculamos:

P(x | Pera) = P(color = Mezcla | Pera) * P(Esfericidad = 0.8 | Pera)
P(x | Pera) = 0 * pdf(x=0.6 | μ=0.8, σ=0.3)
P(x | Pera) = 0

P(x | Manzana) = P(color = **Mezcla** ∩ Esfericidad = 0.8 | Manzana)

Al ser una intersección de atributos independientes calculamos:

P(x | Manzana) = P(color=mezcla | Manzana) * P(Esfericidad = 0.8 | Manzana)
P(x | Manzana) = 0.6 * pdf(x=0.8 | μ=0.8, σ=0.2)
P(x | Manzana) = 0.6 * 1.9947114 = 1.196826841

P(x | Pera) * P(Pera) = 0 * 0.5 = 0
P(x | Manzana) * P(Manzana) = 1.19682684 * 0.5 = 0.598413420

Color	Esfericidad	P(x| Pera)	P(x| Manzana)	P(x | Pera) * P(Pera)	P(x| Manzana) * P(Manzana)	Predicción
Amarillo	0.6	1.19682684	0.120985362	0.59841342	0.060492681	Pera
Mezcla	0.8	0	1.196826841	0	0.598413420	Manzana

b) Si las probabilidades de clase a priori son P(**Pera**)=0.01 y P(**Manzana**)=0.99, y no se utiliza corrección de Laplace, indicar cómo clasificaría los siguientes 2 ejemplos, incluyendo los cálculos realizados en la siguiente tabla:

P(x | Pera) = P(color = **Amarillo** ∩ Esfericidad = 0.6 | Pera)

Al ser una intersección de atributos independientes calculamos:

P(x | Pera) = P(color = Amarillo | Pera) * P(Esfericidad = 0.6 | Pera)
P(x | Pera) = 0.9 * pdf(x=0.6 | μ=0.6, σ=0.3)
P(x | Pera) = 0.9 * 1.32980760 = 1.19682684

P(x | Manzana) = P(color = **Amarillo** ∩ Esfericidad = 0.6 | Manzana)

Al ser una intersección de atributos independientes calculamos:

P(x | Manzana) = P(color = Amarillo | Manzana) * P(Esfericidad = 0.6 | Manzana)
P(x | Manzana) = 0.1 * pdf(x=0.6 | μ=0.8, σ=0.2)
P(x | Manzana) = 0.1 * 1.20985362 = 0.120985362

P(x | Pera) * P(Pera) = 1.19682684 * 0.01 = 0.01196826 = P(Pera | x)
P(x | Manzana) * P(Manzana) = 0.120985362 * 0.99 = 0.11977550 = P(Manzana | x)

P(x | Pera) = P(color = **Mezcla** ∩ Esfericidad = 0.8 | Pera)

Al ser una intersección de atributos independientes calculamos:

P(x | Pera) = P(color = Mezcla | Pera) * P(Esfericidad = 0.8 | Pera)
P(x | Pera) = 0 * pdf(x=0.6 | μ=0.8, σ=0.3)

P(x | Pera) = 0

P(x | Manzana) = P(color = **Mezcla** ∩ Esfericidad = 0.8 | Manzana)

Al ser una intersección de atributos independientes calculamos:

P(x | Manzana) = P(color=mezcla | Manzana) * P(Esfericidad = 0.8 | Manzana)
P(x | Manzana) = 0.6 * pdf(x=0.8 | μ=0.8, σ=0.2)
P(x | Manzana) = 0.6 * 1.9947114 = 1.196826841

P(x | Pera) * P(Pera) = 0 * 0.01 = 0
P(x | Manzana) * P(Manzana) = 1.19682684 * 0.99 = 1.18485857

Color	Esfericidad	P(x| Pera)	P(x| Manzana)	P(x | Pera) * P(Pera)	P(x| Manzana) * P(Manzana)	Predicción
Amarillo	0.6	1.19682684	0.120985362	0.01196826	0.11977550	Manzana
Mezcla	0.8	0	1.196826841	0	1.18485857	Manzana

5. Generación de un modelo NB

a) En base a los datos del archivo **estrellas.xslx**, generar un modelo de NB para clasificar su tipo espectral (F o K), sin utilizar corrección de Laplace. Incluir sus cálculos.

Temperatura (°K) μ
Clase F = (6200 + 7500 + 6600 + 6100) / 4 = 6600
Clase K = (3000 + 1900 + 2300 + 1400 + 2500 + 3450) / 6 = 2425

varF = ((6200 - 6600)^2 + (7500 - 6600)^2 + 0 + (6100 - 6600)^2) / (4 - 1) = 406666,6667
varK = ((3000 - 2425)^2 + (1900 - 2425)^2 + (2300 - 2425)^2 + (1400 - 2425)^2 + (2500 - 2425)^2 + (3450 - 2425)^2) / (6-1) = 545750

Temperatura (°K) σ
Clase F = √(406666,6667) = 637.704215683
Clase K = √(545750) =738.74894247

Planeta Habitable Cercano = Si
Clase F = 1/4
Clase K = 2/6

Planeta Habitable Cercano = No
Clase F = 3/4
Clase K = 4/6

Luminosidad (Relativa al Sol) μ
Clase F = (22 + 6 + 3 + 16) / 4 = 11.75
Clase K = (0.9 + 0.2 + 0.3 + 6 + 2 + 5) / 6 = 2.4

varF = ((22 - 11.75)^2 + (6 - 11.75)^2 + (3 - 11.75)^2 + (16 - 11.75)^2) / (4 - 1) = 77.5833
varK = ((0.9 - 2.4)^2 + (0.2 - 2.4)^2 + (0.3 - 2.4)^2 + (6 - 2.4)^2 + (2 - 2.4)^2 + (5 - 2.4)^2) / (6-1) = 6.276

Luminosidad (Relativa al Sol) σ
Clase F = √(77.5833) = 8.80813828229
Clase K = √(6.276) = 2.50519460322

P(Clase F) = 4/10 = 0,4
P(Clase K) = 6/10 = 0,6

Atributo / Valor	Clase F	Clase K
Temperatura (°K) μ	6600	2425
Temperatura (°K) σ	637.704215683	738.74894247
Planeta Habitable Cercano = Si	1/4	2/6
Planeta Habitable Cercano = No	3/4	4/6
Luminosidad (Relativa al Sol) μ	11.75	2.4
Luminosidad (Relativa al Sol)σ	8.80813828229	2.50519460322

P(Clase F)	0.4
P(Clase K)	0.6

b) Utilizando el modelo anterior, clasifique los **3 primeros ejemplos** del conjunto de datos **estrellas.xslx**. Utilice una tabla como la siguiente para realizar los cálculos. Utilizar notación científica con 2 decimales para escribir las probabilidades para números menores a 0.01. Por ejemplo, 0.0000436 se escribe como 4.36e-5. y 7.3214123e-12 se redondea a 7.32e-12.

Primera fila:
P(x | F) = P(Temperatura = 6200 | F) * P(Planeta Habitante Cerca = No | F) * P(Luminosidad = 22 | F)
P(x | F) = pdf(6200 | μ = 6600, σ = 637.704215683) * 3/4 * pdf(22| μ = 11.75, σ = 8.80813828229)
pdf(6200 | μ = 6600, σ = 637.704215683) = 0.00051387 = 5.13e-4
pdf(22| μ = 11.75, σ = 8.80813828229) = 0.02301268 = 2.30e-2

P(x | F) = 0.00051387 * 3/4 * 0.02301268 = 8.87e-6
P(x | K) = P(Temperatura = 6200 | K) * P(Planeta Habitante Cerca = No | K) * P(Luminosidad = 22 | K)
P(x | K) = pdf(6200 | μ = 2425, σ = 738.74894247) * 4/6 * pdf(22| μ = 2.4, σ = 2.50519460322)
P(x | K) = 1.15e-9 * 4/6 * 8.13e-15 = 6.23e-24

P(x | F)*P(F) = 8.87e-6 * 0.4 = 0.000003548 = 3.54e-6 = P(F|x)
P(x | K)*P(K) = 6.23e-24 * 0.6 = 3.73e-24 = P(K|x)

Segunda fila:
P(x | F) = P(Temperatura = 7500 | F) * P(Planeta Habitante Cerca = No | F) * P(Luminosidad = 6 | F)
P(x | F) = pdf(7500 | μ = 6600, σ = 637.704215683) * 3/4 * pdf(6| μ = 11.75, σ = 8.80813828229)
pdf(7500 | μ = 6600, σ = 637.704215683) = 0.00023109 = 2.31e-4
pdf(6| μ = 11.75, σ = 8.80813828229) = 0.03660057 = 3.66e-2
P(x | F) = 0.00023109 * 3/4 * 0.03660057 = 6.34e-6

P(x | K) = P(Temperatura = 7500 | K) * P(Planeta Habitante Cerca = No | K) * P(Luminosidad = 6 | K)
P(x | K) = pdf(7500| μ = 2425, σ = 738.74894247) * 4/6 * pdf(6| μ = 2.4, σ = 2.50519460322)
pdf(7500| μ = 2425, σ = 738.74894247)
P(x|K) = 3.05e-14 * 4/6 * 0.05670971566820744 = 1.15e-15

P(x | F)*P(F) = 6.34e-6 * 0.4 = 0.000002536 = 2.53e-6 = P(F|x)
P(x|K)*P(K) = 1.15e-15 * 0.6 = 6.9e-16 = P(K|x)

Tercera fila:
P(x | F) = P(Temperatura = 3000| F) * P(Planeta Habitante Cerca = Si | F) * P(Luminosidad = 0.9 | F)
P(x | F) = pdf(3000| μ = 6600, σ = 637.704215683) * 1/4 * pdf(0.9| μ = 11.75, σ = 8.80813828229)
pdf(3000| μ = 6600, σ = 637.704215683)
P(x|F) = 7.51e-11 * 1/4 * 0.021209646682249787 = 3.98e-13

P(x | K) = P(Temperatura = 3000 | K) * P(Planeta Habitante Cerca = Si | K) * P(Luminosidad = 0.9 | K)
P(x | K) = pdf(3000 | μ = 2425, σ = 738.74894247) * 2/6 * pdf(0.9| μ = 2.4, σ = 2.50519460322)
pdf(3000| μ = 2425, σ = 738.74894247) = 0.00039890 = 3.98e-4
pdf(0.9| μ = 2.4, σ = 2.50519460322) = 0.13311284 = 1,33e-1

P(x | K) = 0,00039890 * 2/6 * 0,13311284 = 1,77e-5

P(x|F)*P(F) = 3.98e-13 * 0.4 = 1.59e-13 = P(F|x)
P(x | K)*P(K) = 1.77e-5 * 0.6 = 0.00001062 = 1.06e-5 = P(K|x)

Temperatura	Planeta Habitable Cerca	Luminosidad	P(x | F)	P(x | K)	P(x | F) * P(F)	P(x | K) * P(K)	Predicción
6200	No	22	8.87e-6	6.23e-24	3.54e-6	3.73e-24	F
7500	No	6	6.34e-6	1.15e-15	2.53e-6	6.9e-16	F
3000	Si	0.9	3.98e-13	1.77e-5	1.59e-13	1.06e-5	K

c) Las predicciones del modelo ¿coinciden con las etiquetas del dataset? Calcule las predicciones para el resto de los ejemplos (sin llenar la tabla, utilizando RapidMiner).

Luego, calcule el **accuracy** (porcentaje de ejemplos clasificados correctamente) del modelo para ese conjunto de datos. Por ejemplo, dados 10 ejemplos, si el modelo acierta la clase de 7 de ellos, entonces el accuracy es 7/10=0.7 o 70%.

Las predicciones del modelo coinciden con el dataset.
Las predicciones en el rapidminer quedaron así:

Row No.	Clase Espec...	prediction(C...	confidence(F)	confidence(K)	Temperatur...	Planeta Habi...	Luminosida...
1	F	F	1	0	6200	No	22
2	F	F	1.000	0.000	7500	No	6
3	K	K	0.000	1.000	3000	Si	0.900
4	F	F	1.000	0.000	6600	Si	3
5	K	K	0.000	1.000	1900	No	0.200
6	K	K	0.000	1.000	2300	No	0.300
7	K	K	0.000	1.000	1400	No	6
8	F	F	1.000	0.000	6100	No	16
9	K	K	0.000	1.000	2500	No	2
10	K	K	0.000	1.000	3450	Si	5

Tiene un 100% de precisión

	true F	true K	class precision
pred. F	4	0	100.00%
pred. K	0	6	100.00%
class recall	100.00%	100.00%	

d) Se agrega un ejemplo al conjunto de datos:

Temperatura	Planeta Habitable Cerca	Luminosidad	Clase
8500	No	4	K

Vuelva a generar el modelo, ahora incluyendo este ejemplo (no es necesario incluir los cálculos). ¿Cómo afecta al modelo de cada atributo o clase?

Atributo / Valor	Clase F	Clase K
Temperatura (°K) μ	6600	3292.8571428571
Temperatura (°K) σ	637.704215683	2.393,120099268
Planeta Habitable Cercano = Si	1/4	2/7
Planeta Habitable Cercano = No	3/4	5/7
Luminosidad (Relativa al Sol) μ	11.75	2.62857142857
Luminosidad (Relativa al Sol)σ	8.80813828229	2.37837479525

P(Clase F)	0,3636
P(Clase K)	0.6363

e) Vuelva a calcular el accuracy para los datos y el modelo del punto **d)** ¿Cambió? Si es así, ¿Qué ejemplo vio su predicción cambiar? ¿Por qué?

	true F	true K	class precision
pred. F	3	0	100.00%
pred. K	1	7	87.50%
class recall	75.00%	100.00%	

El accuracy se modificó, antes tenía 100% y ahora 90.91%.
El ejemplo que cambió es el de la fila 4, que tenía como predicción “F”, pero después de agregar el nuevo ejemplo su predicción cambió erróneamente a K.
El nuevo valor afectó los valores de la clase K.
Repercute en gran medida a la media de temperatura, ya que la media es un valor muy sensible y el máximo valor de temperatura que tenía K era de 3450, por lo que 8500 es un valor extremo que lo hace crecer en gran medida.
El resto de medias, y desviaciones lógicamente también cambiaron, pero es un cambio menor.