МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет "Львівська політехніка"

Інтелектуальний аналіз даних за допомогою програмного пакета WEKA та MS Excel.

Баєсівський класифікатор.

МЕТОДИЧНІ ВКАЗІВКИ
до лабораторної роботи № 9

з курсу **"Системи інтелектуального аналізу та візуалізації даних"**

для студентів за освітньою програмою Комп'ютерні науки (Проектування і програмування інтелектуальних систем та пристроїв)

Затверджено на засіданні кафедри
"Системи автоматизованого проектування"

Протокол N 1 від 28.08.2023р.

ЛЬВІВ 2023

1. МЕТА РОБОТИ

Мета роботи – навчитися класифікувати дані за допомогою використання баєсівського підходу. Вивчити теоретичні основи методу та для виконання аналізу даних навчитися використовувати програми WEKA та Excel.

2. КОРОТКІ ТЕОРЕТИЧНІ ВІДОМОСТІ

2.1. Наївний баєсівський класифікатор.

Наївний баєсівський класифікатор — один із найбільш використовуваних алгоритмів класифікації. Його математична основа — умовна ймовірність. Припустимо, що ряд незалежних атрибутів (x_1,x_2,...,x_n) можна використовувати для класифікації цільової змінної y. Маючи нову вибірку з певними атрибутами, наївний Баєс може передбачити ймовірність появи кожного можливого класу на основі атрибутів вибірки. Теорема Баєса викладена у рівнянні:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}, \qquad P(x) \neq 0 \tag{1}$$

де y та x є подіями.
$P(y)$ $P(x)$ є імовірностями y та x безвідносно одна до одної.
$P(y|x)$, умовна імовірність, є імовірністю події y за умови істинності x.
$P(x|y)$ є імовірністю x за умови істинності y.

Простіший спосіб запам’ятати теорему Баєса – це записати наступним чином:

$$P(y|x)P(x) = P(x|y)P(y). \tag{2}$$

Якщо є декілька незалежних атрибутів, $P(x) = P(x_1)P(x_2)\ldots P(x_n)$. Для k-го класу y_k (припустимо, що є m різних класів), ймовірність його появи для заданих (x_1,x_2,...,x_n) може бути виражена як

$$P'(y_k|x) = \frac{P(x_1|y_k)P(x_2|y_k)\cdots P(x_n|y_k)P(y_k)}{P(x)} \tag{3}$$

Після того, як ми обчислили кожен $P'(y_k|x)$, кінцевий $P(y_k|x)$ можна обчислити як

$$P(y_k|x) = \frac{P'(y_k|x)}{\sum_i^m P'(y_i|x)} \quad (4)$$

Зауважте, що у попередньому рівнянні і чисельник, і знаменник необхідно розділити на $P(x)$. Однак $P(x)$ скасовується, оскільки воно ділить і чисельник, і знаменник. Тому при обчисленні $P'(y_k|x)$ $P(x)$ також ігнорується. Розглянемо представлену концепцію на прикладі з таблиці 1.

Таблиця 1. Позначені класами навчальні кортежі з бази даних клієнтів AllElectronics.

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Jiawei Han. Data Mining: Concepts and Techniques. 2006 by Elsevier Inc.

Кортежі даних описуються атрибутами *age*, *income*, *student* і *credit rating*. Атрибут мітки класу buys computer має два різних значення (а саме {yes, no}). Нехай C_1 відповідає класу *buys computer* = yes, а C_2 відповідає *buys computer* = no. Кортеж, який ми хочемо класифікувати, є $\boldsymbol{X}$ = (*age = youth, income = medium, student = yes, credit rating = fair*)

Потрібно максимізувати $P(\boldsymbol{X}|Ci)P(Ci)$, для i = 1, 2. $P(Ci)$. Апріорна ймовірність кожного класу може бути обчислена на основі навчальних кортежів:

P(*buys computer = yes*) = 9/14 = 0.643

P(*buys computer = no*) = 5/14 = 0.357

Щоб обчислити $P(\boldsymbol{X}|Ci)$, для i = 1, 2, ми обчислюємо наступні умовні ймовірності:

P(*age = youth | buys computer = yes*) = 2/9 = 0.222

$P(age = youth \mid buys\ computer = no) = 3/5 = 0.600$
$P(income = medium \mid buys\ computer = yes) = 4/9 = 0.444$
$P(income = medium \mid buys\ computer = no) = 2/5 = 0.400$
$P(student = yes \mid buys\ computer = yes) = 6/9 = 0.667$
$P(student = yes \mid buys\ computer = no) = 1/5 = 0.200$
$P(credit\ rating = fair \mid buys\ computer = yes) = 6/9 = 0.667$
$P(credit\ rating = fair \mid buys\ computer = no) = 2/5 = 0.400$

Використовуючи наведені вище ймовірності, отримуємо

$$P(\boldsymbol{X}|buys\ computer = yes) = P(age = youth \mid buys\ computer = yes) * P(income = medium \mid buys\ computer = yes) * P(student = yes \mid buys\ computer = yes) * P(credit\ rating = fair \mid buys\ computer = yes) = 0.222*0.444*0.667*0.667 = 0.044.$$

Аналогічно,

$P(\boldsymbol{X}|buys\ computer = no) = 0.600*0.400*0.200*0.400 = 0.019.$

Щоб знайти клас Ci, який максимізує $P(\boldsymbol{X}|Ci)P(Ci)$, слід обчислити

$P(\boldsymbol{X}|buys\ computer = yes)P(buys\ computer = yes) = 0.044*0.643 = 0.028$

$P(\boldsymbol{X}|buys\ computer = no)P(buys\ computer = no) = 0.019*0.357 = 0.007$

Тому наївний байєсівський класифікатор передбачає *buys computer = yes* для кортежу $\boldsymbol{X}$.

Математична основа наївної класифікації Баєса відносно простаі це є важливою причиною, чому наївна класифікація Баєса має високу ефективність обчислень при великому розмірі даних. Хоча наївна класифікація Баєса робить «наївне» припущення, що всі атрибути є незалежними один від одного, наївний Баєс дійсно має достатню продуктивність порівняно з усіма іншими класифікаційними моделями інтелектуального аналізу даних.

2.2. Застосування Баєсівського класифікатора у програмі WEKA

Відкриємо файл *contact-lenses.arff* (див. рис.1).

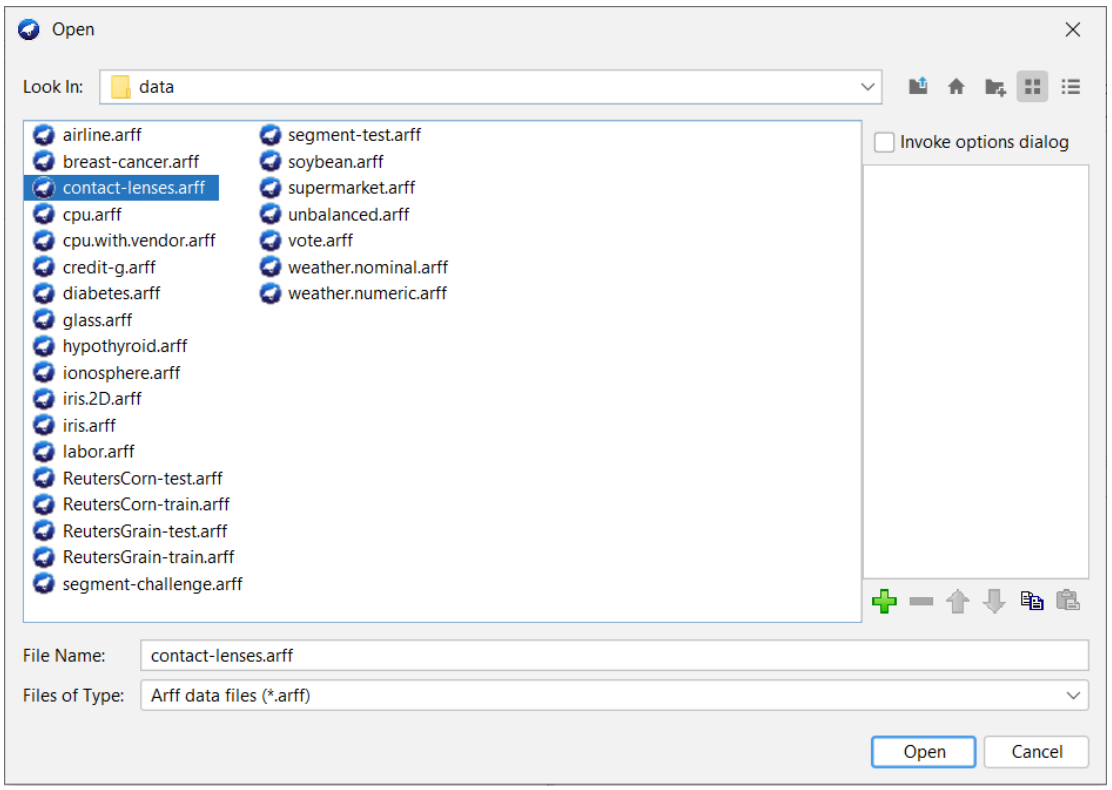

Рис. 1. Відкриття файлу у програмі WEKA

Після завантаження файлу на лівій нижній панелі *Attributes* можна вибрати ознаку, при цьому на правій панелі *Selected attribute* відображається гістограма значень ознаки. Зліва від цієї кнопки знаходиться компонент вибору цільової ознаки: яку ознаку вважати класом під час візуалізації (вибір NoClass відповідає тому, що всі об'єкти приписуються одному класу і для кожної ознаки показується гістограма розподілу всіх його значень). При натисканні на кнопку *Visualize All* з'являються гістограми за різними ознаками (рис. 2)

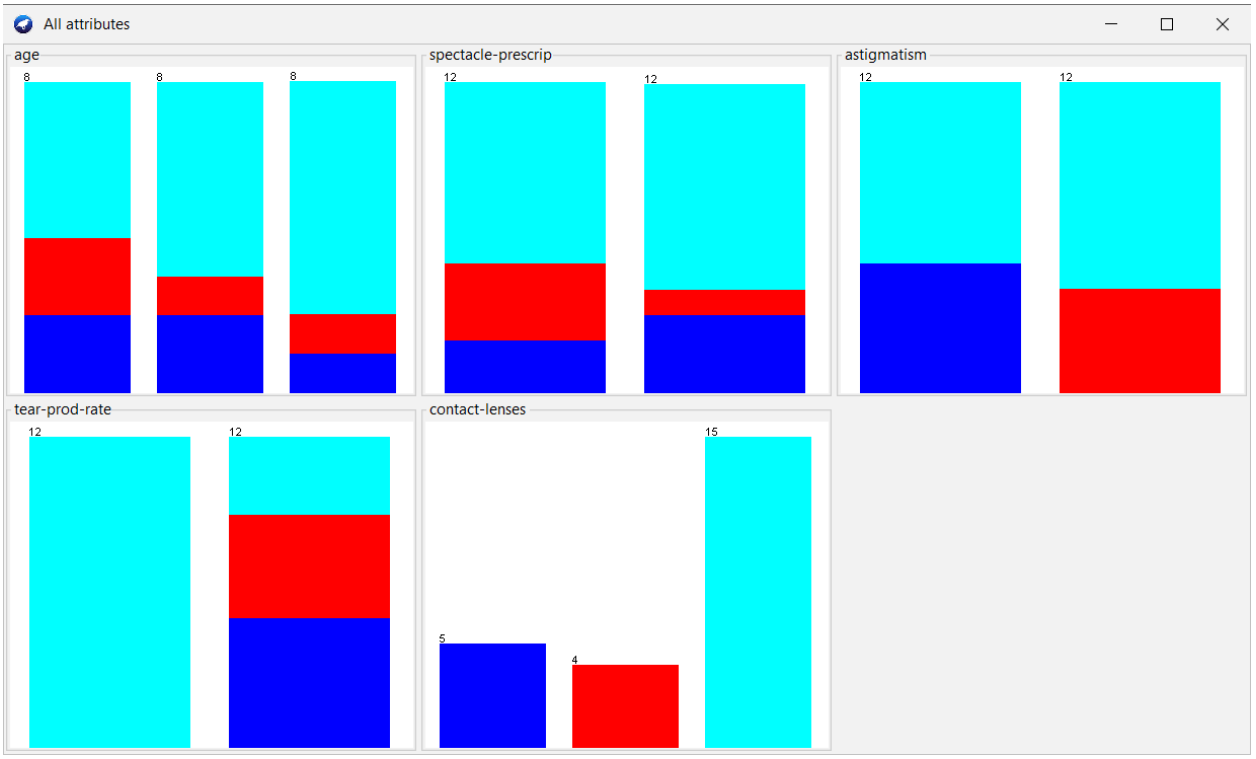

Рис. 2. Гістограми значень різних ознак

Так при відкритті файлу contact-lenses.arff з'являються 5 атрибутів:

- *age* (вік): *young* (молодий), *pre-presbyopic* (попередня далекозорість), *presbyopic* (далекозорість);
- *spectacle-prescription* (порушення зору): *myope* (короткозора людина), *hypermetrope* (далекозора людина);
- *astigmatism* (астигматизм): *yes* (так), *no* (ні);
- *tear-prod-rate* (сльозливість): *reduced* (знижена), *normal* (нормальна);
- *contact-lenses* (контактні лінзи): *soft* (м'які), *hard* (жорсткі), none (жодних).

Далі переходимо до класифікації з урахуванням баєсівського класифікатора (див. рис. 3). Для цього натискаємо кнопку *Choose* і вибираємо з ієрархічного списку *weka/classifiers/bayes/NaiveBayes*.

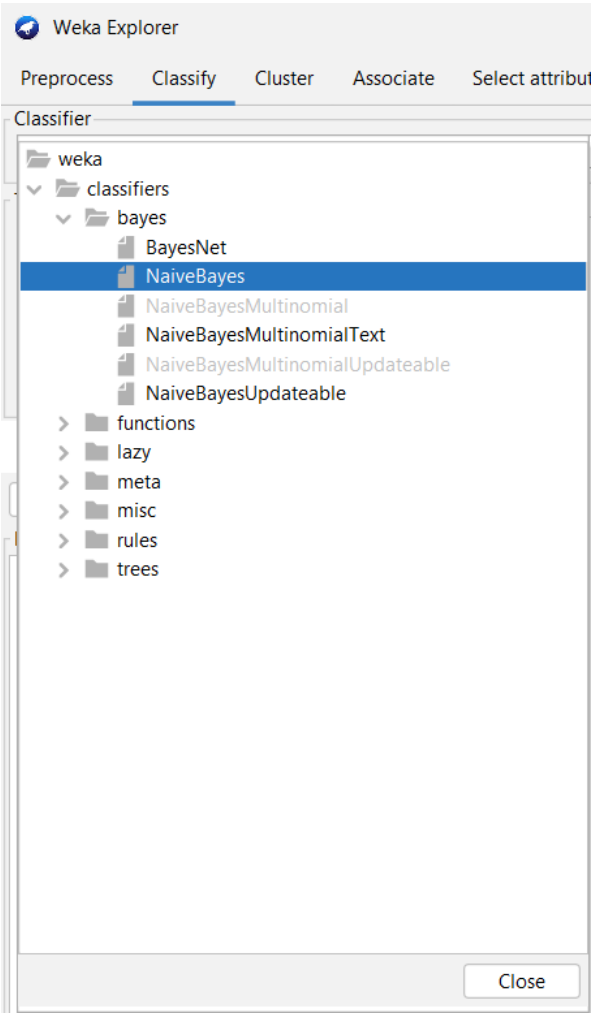

Рис. 3. Вибір методу

На панелі *Test options* визначено, як тестуватиметься класифікатор: на навчальній вибірці (*use training set*), на тестовій із окремого файлу (*supplied test set*), по блоках (*cross-validation*), за допомогою поділу вихідної вибірки на навчання та контроль (*percentage split*).

Нижче на панелі *Test options* знаходиться компонент для вибору цільової ознаки (що і вважатиметься класом у завданні). Натискання кнопки *Start* запускає навчання класифікатора. На найбільшій панелі *Classifier output* відображається звіт про навчання

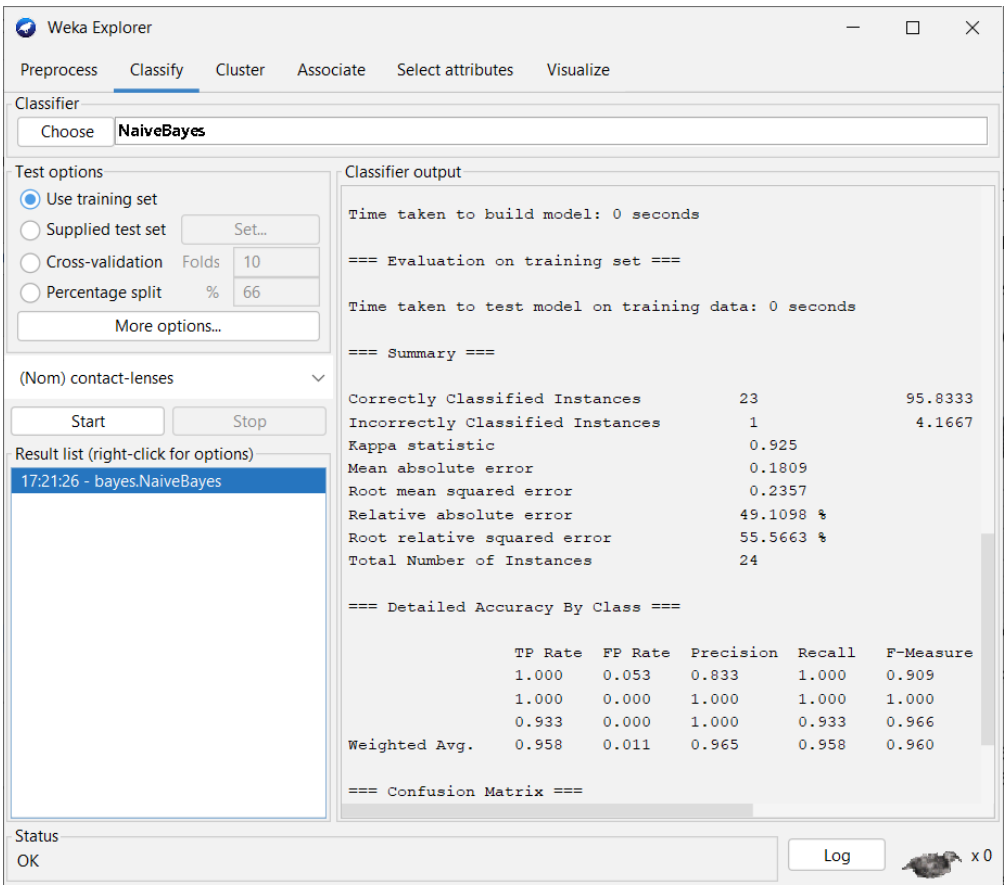

Рис. 4. Виведено звіт про налаштування та тестування алгоритму баєсівського класифікатора

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     contact-lenses
Instances:    24
Attributes:   5
              age
              spectacle-prescrip
              astigmatism
              tear-prod-rate
              contact-lenses
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

Naive Bayes Classifier

                         Class
Attribute                soft   hard   none
                        (0.22) (0.19) (0.59)
=============================================
age
  young                   3.0    3.0    5.0
  pre-presbyopic          3.0    2.0    6.0
  presbyopic              2.0    2.0    7.0
  [total]                 8.0    7.0   18.0
```

Потім параметри налаштованого класифікатора:

```
=== Classifier model (full training set) ===

Naive Bayes Classifier

                     Class
Attribute              soft   hard   none
                     (0.22) (0.19) (0.59)
=============================================
age
  young                 3.0    3.0    5.0
  pre-presbyopic        3.0    2.0    6.0
  presbyopic            2.0    2.0    7.0
  [total]               8.0    7.0   18.0

spectacle-prescrip
  myope                 3.0    4.0    8.0
  hypermetrope          4.0    2.0    9.0
  [total]               7.0    6.0   17.0

astigmatism
  no                    6.0    1.0    8.0
  yes                   1.0    5.0    9.0
  [total]               7.0    6.0   17.0

tear-prod-rate
  reduced               1.0    1.0   13.0
  normal                6.0    5.0    4.0
  [total]               7.0    6.0   17.0

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances          23               95.8333 %
Incorrectly Classified Instances         1                4.1667 %
Kappa statistic                          0.925
Mean absolute error                      0.1809
Root mean squared error                  0.2357
Relative absolute error                 49.1098 %
Root relative squared error             55.5663 %
Total Number of Instances               24

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.053    0.833      1.000    0.909      0.889    1.000     1.000     soft
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     hard
                 0.933    0.000    1.000      0.933    0.966      0.917    1.000     1.000     none
Weighted Avg.    0.958    0.011    0.965      0.958    0.960      0.925    1.000     1.000

=== Confusion Matrix ===

  a  b  c   <-- classified as
  5  0  0 |  a = soft
  0  4  0 |  b = hard
  1  0 14 |  c = none
```

Потім йде різна статистика роботи класифікатора, включаючи відсоток вірних відповідей контролю.

Виводиться матриця розміру $l \times l$, де l – кількість класів, ij-й елемент матриці дорівнює кількості об'єктів з i-го класу, які були віднесені до j-го. Кількість правильно класифікованих об'єктів дорівнює сумі елементів, що стоять на головній діагоналі. Правильно співвіднесено 23 об’єкти з 24.

У виведеній статистиці значення *True Positive* (*TP*) *rate* або *Recall* (для класу, що розглядається) дорівнює відсотку правильно класифікованих об’єктів класу (виходить розподілом діагонального елемента на суму елементів у його рядку), 0,958. Значення *False Positive* (*FP*) дорівнює відсотку об’єктів інших класів, які помилково занесені у клас (якщо з матриці викреслити рядок класу, що розглядається, то значення дорівнює сумі елементів стовпця цього класу, поділене на суму всіх елементів). Значення *Precision* дорівнює відсотку правильно класифікованих об’єктів з об’єктів, віднесених алгоритмом до класу (відношення діагонального елемента до суми елементів стовпця). Значення *F-Measure* обчислюється за такою формулою *2*Precision*Recall/(Precision+Recall),* тобто це середнє гармонійне *Precision* и *Recall*.

Вкладка *Visualize* дозволяє візуалізувати вибірку. Кнопка *Select Attributes* дозволяє вибрати ознаки для візуалізації: будуть побудовані картинки-проекції на різні пари цих ознак. Повзунок *PlotSize* вибирає розмір картинок, *PointSize* – розмір точок, що зображають об’єкти, *Jitter* – рівень шумів, які спеціально додаються до ознак.

Останній повзунок дуже важливий, оскільки для пари ознак, які набувають невеликої кількості значень (наприклад, k-значних) ціла група об’єктів зливається в одну точку, а повзунок дозволяє розсіяти цю точку в хмару.

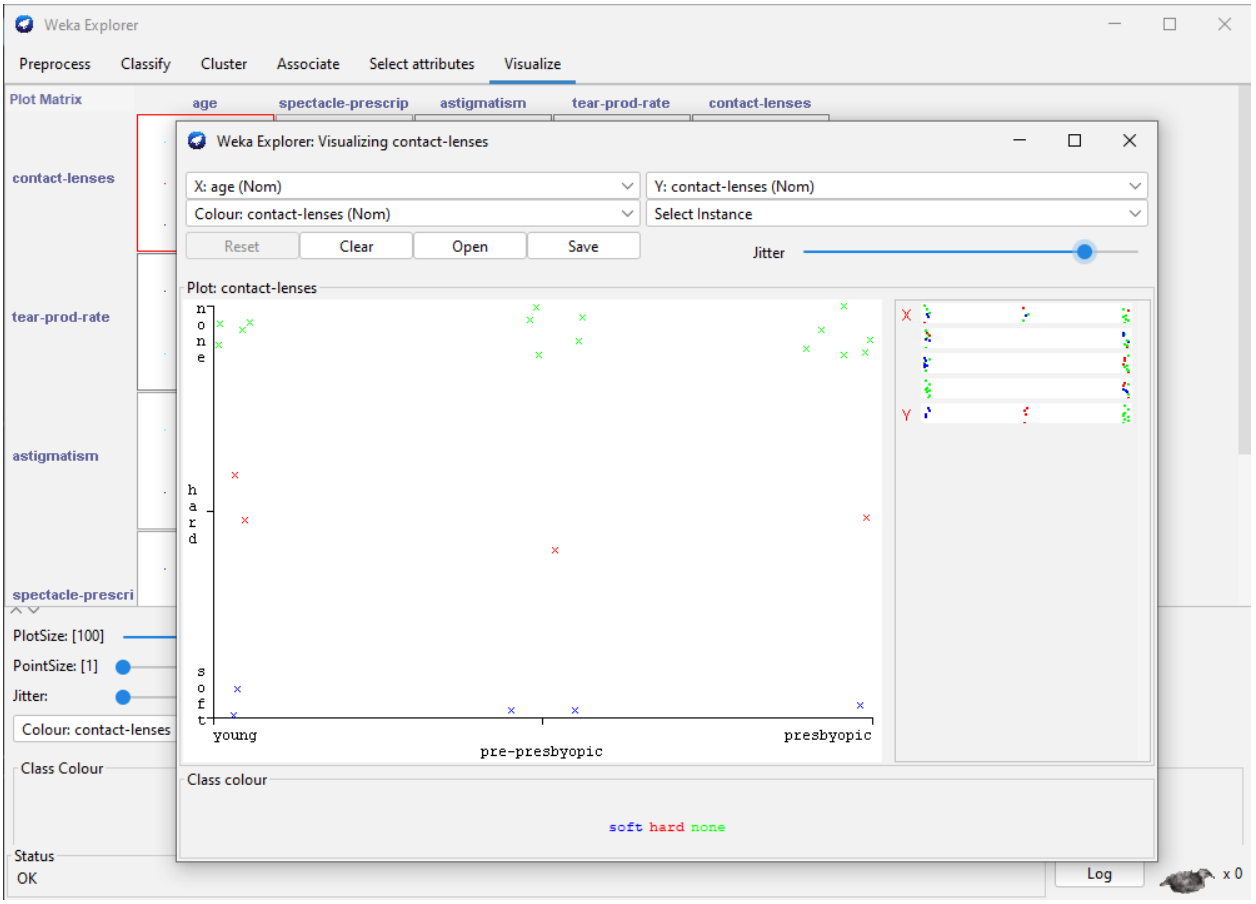

Рис. 5 Вкладка Visualize модуля Explorer

Таблиця 2. Параметри налаштування класифікатора

Метод	**Параметр**
NaiveBayes	*displayModelInOldFormat* – відображення побудованої моделі у старому форматі, що підходить, коли атрибут класу приймає багато значень. Новий формат краще у випадку, коли менше класів і багато атрибутів. *useKernelEstimator* – для оцінки числових атрибутів використовувати оціночну функцію відмінну від нормального розподілу. *useSupervisedDiscretization* – використовувати дискретизацію з учителем для перетворення числових атрибутів у номінальні

2.2. Побудова наївного баєсівського класифікатора в Excel.

У цьому розділі ми розглянемо два приклади в Excel. Перший приклад базується на таблиці 3.

Таблиця 3. Дані навчального набору .

No.	Size	Number	Thickness	Lung Cancer
1	Big	Low	Deep	Yes
2	Big	Over	Deep	Yes
3	Normal	Over	Shallow	No
4	Big	Low	Shallow	No
5	Big	Over	Deep	Yes
6	Normal	Over	Deep	Yes
7	Big	Low	Shallow	No
8	Big	Over	Shallow	Yes
9	Big	Over	Deep	Yes
10	Normal	Low	Deep	No
Scoring	Big	Over	Deep	???

Зображення на рис.6 містить ті самі дані, що й у таблиці 3. У нас є лише десять навчальних спостережень і один тестовий зразок. Причина зробити вибірку маленькою полягає у можлтвості кращої ілюстрації методу. Отримавши достатньо досвіду за допомогою цього простого прикладу, ви зможете працювати із складнішими наборами даних.

Щоб застосувати наївний метод Баєса для прогнозування класу зразку, нам спочатку потрібно побудувати модель на основі навчальних даних.

	A	B	C	D	E
1	Sample size		Lung Cancer		
2			yes	no	
3		count			
4		probability			
5		big			
6	Size	normal			
7		over			
8	Number	low			
9		deep			
10	Thickness	shallow			
11					
12	No.	Size	Number	Thickness	Lung Cancer
13	1	big	low	deep	yes
14	2	big	over	deep	yes
15	3	normal	over	shallow	no
16	4	big	low	shallow	no
17	5	big	over	deep	yes
18	6	normal	over	deep	yes
19	7	big	low	shallow	no
20	8	big	over	shallow	yes
21	9	big	over	deep	yes
22	10	normal	low	deep	no

Рис. 6 Підготовка даних для аналізу

Дотримуйтесь цих інструкцій:

1. Введіть 10 у комірку B1, оскільки існує лише десять навчальних записів.

2. Введіть формулу =COUNTIF($E$13:$E$22,C$2) у комірку C3. Ця формула підраховує кількість класу “yes” для цільової змінної Lung Cancer. Автозаповніть з клітинки C3 до клітинки D3. Комірка D3 зберігає кількість класів “no”.

3. Введіть формулу =C$3/$B$1 у комірку C4, і тоді автозаповніть з клітинки C4 до клітинки D4. Клітинки C4 і D4 зберігають *p*(yes) і *p*(no).

4. Введіть наступну формулу у клітинку C5:

=COUNTIFS($E$13:$E$22,C$2,$B$13:$B$22,$B5)/C$3

5. Автозаповніть з клітинки C5 до клітинки C6, і потім автохаповніть разом до клітинок D5:D6. Клітинки C5, C6, D5, і D6 представляти ймовірності *p*(big|yes), *p*(normal|yes), *p*(big|no), і *p*(normal|no), відповідно. *p*(big|yes) є умовною ймовірністю: враховуючи «yes», ймовірність Size = “big”.

На даний момент частина нашого аркуша виглядає так, як на рис. 7.

	A	B	C	D
1	Sample size	10	Lung Cancer	
2			yes	no
3		count	6	4
4		probability	0.6	0.4
5		big	0.8333333	0.5
6	Size	normal	0.1666667	0.5
7		over		
8	Number	low		
9		deep		
10	Thickness	shallow		
11				

Рис. 7 Наївний байєсівський аналіз частково завершено

6. Введіть наступну формулу в клітинку C7:

=COUNTIFS($E$13:$E$22,C$2,$C$13:$C$22,$B7)/C$3

7. Автозаповніть клітинки C7 до клітинки C8, а потім автозаповніть клітинки D7:D8. Клітинки C7, C8, D7 та D8 посилаються на ймовірності p(over|yes), p(low|yes), p(over|no), і p(low|no) відповідно.

8. Введіть наступну формулу в клітинку C9:

=COUNTIFS($E$13:$E$22,C$2,$D$13:$D$22,$B9)/C$3

9. Автозаповніть з клітинки C9 до клітинки C10, а потім автозаповніть клітинки D9:D10. Клітинки C9, C10, D9, і D10 посилаються на ймовірності p(deep|yes), p(shallow|yes), p(deep|no), і p(shallow|no), відповідно.

Тепер наш аркуш має виглядати так само, як на рис.8. Якщо є відмінності, перевірте формули в клітинках C3:D10.

	A	B	C	D	E
1	Sample size	10	Lung Cancer		
2			yes	no	
3		count	6	4	
4		probability	0.6	0.4	
5		big	0.8333333	0.5	
6	Size	normal	0.1666667	0.5	
7		over	0.8333333	0.25	
8	Number	low	0.1666667	0.75	
9		deep	0.8333333	0.25	
10	Thickness	shallow	0.1666667	0.75	
11					
12	No.	Size	Number	Thickness	Lung Cancer

Рис. 8 Підготовка даних для наївного байєсівського аналізу

Продовжуйте процес аналізу даних наївним Баєсом, дотримуючись наступних інструкцій:

10. Введіть “p'(scoring|yes)”, “p'(scoring|no)”, і “p(yes|scoring)” у комірки F22, G22, і H22, відповідно.

11. Введіть наступну формулу в клітинку F23:

=SUMIF($B$5:$B$10,$B23,C$5:C$10) *

SUMIF($B$5:$B$10,$C23,C$5:C$10) *

SUMIF($B$5:$B$10,$D23,C$5:C$10)*C$4

Попередня формула обчислює ймовірність того, що результат оцінки враховується як «yes». Він реалізує рівняння (3). Зауважте, що функція SUMIF має інший синтаксис від SUMIFS. Наприклад, SUMIF($B$5:$B$10,$B23,C$5:C$10) підсумовує клітинки в масиві C5:C10, лише якщо відповідні клітинки в масиві B5:B10 мають те саме значення, що й клітинка B23. Якщо ми хочемо використовувати функцію SUMIFS для тієї ж мети, тоді вираз має бути наступним SUMIFS(C$5:C$10,$B$5:$B$10,$B23).

12. Автозаповніть клітинки від F23 до клітинки G23. Клітинка G23 посилається на ймовірність запису оцінки за класом “no”.

13. Введіть у клітинку H23 формулу =F23/(F23+G23) яка реалізує рівняння (4).

14. Введіть у клітинку E23 формулу =IF(H23>0.5,"yes","no").

Це завершує нашу класифікацію використовуючи метод наївної класифікації Байєса. Наш результат має виглядати так само, як показано на рис. 9.

	A	B	C	D	E	F	G	H
1	Sample size	10	Lung Cancer					
2			yes	no				
3		count	6	4				
4		probability	0.6	0.4				
5		big	0.8333333	0.5				
6	Size	normal	0.1666667	0.5				
7		over	0.8333333	0.25				
8	Number	low	0.1666667	0.75				
9		deep	0.8333333	0.25				
10	Thickness	shallow	0.1666667	0.75				
11								
12	No.	Size	Number	Thickness	Lung Cancer			
13	1	big	low	deep	yes			
14	2	big	over	deep	yes			
15	3	normal	over	shallow	no			
16	4	big	low	shallow	no			
17	5	big	over	deep	yes			
18	6	normal	over	deep	yes			
19	7	big	low	shallow	no			
20	8	big	over	shallow	yes			
21	9	big	over	deep	yes			
22	10	normal	low	deep	no	p'(scoring|yes)	p'(scoring|no)	p(yes|scoring)
23	Scoring	big	over	deep	yes	0.347222222	0.0125	0.965250965

Рис. 9 Класифікація за допомогою наївного байєсівського аналізу завершена

3. ЛАБОРАТОРНЕ ЗАВДАННЯ

Ваше перше завдання для цієї лабораторної роботи - оцінити алгоритми класифікації наївний Баєс за допомогою Weka:

1. Для індивідуального завдання розв'яжіть задачу класифікації за допомогою алгоритму:

- наївна Баєсівська класифікація (bayes.NaiveBayes);

2. Змінюючи параметри налаштування алгоритмів, спробуйте досягти найвищої якості навчання класифікатора.

Ваше друге завдання – використати Excel для побудови моделі класифікації наївним Баєсом.

3. Порівняйте результати отримані в обидвох системах.

4. У звіті надайте результати роботи алгоритму, його налаштування.

4. КОНТРОЛЬНІ ЗАПИТАННЯ

1. У чому полягає задача класифікації? Наведіть практичний приклад.

2. Що таке навчання з учителем і без учителя? До якого типу належить задача класифікації?

3. Задача класифікації є описовою або прогнозуючою і чому?

4. Навіщо потрібні дві вибірки: навчальна і тестова?

5. Які існують підходи для поділу вихідної вибірки на навчальну і тестову?

6. Як оцінити якість побудованої моделі класифікації?

7. Поясніть роботу наївного баєсівського класифікатора.

8. Поясніть теорему Баєса?

9. Які типи наївного баєсівського класифікатора існують?

10. Де застосовується наївний Баєс?

5. ЗМІСТ ЗВІТУ

1. Тема і мета роботи.

2. Завдання до роботи.

3. Результати виконання завдань.

4. Висновки, що відображують результати виконання роботи та їх критичний аналіз.