Прізвище: **КИРИЛЮК**
Ім'я: **Дмитро**
Група: **ПП-22**
Варіант: **08**
Дата захисту: **07.04.2025р**.

Кафедра: **САПР**
Дисципліна: **Системи інтелектуального аналізу та візуалізації даних**
Перевірив: **Андрій КЕРНИЦЬКИЙ**

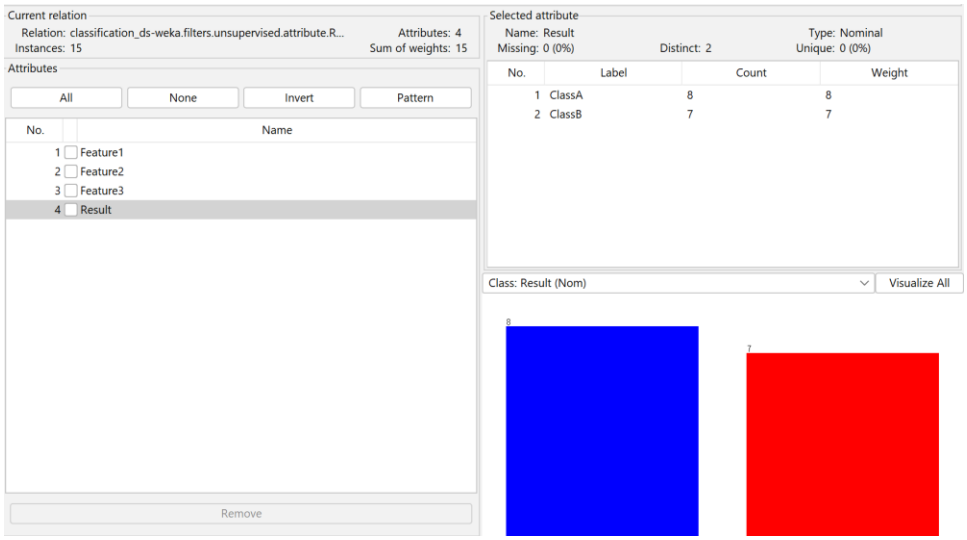
ЗВІТ

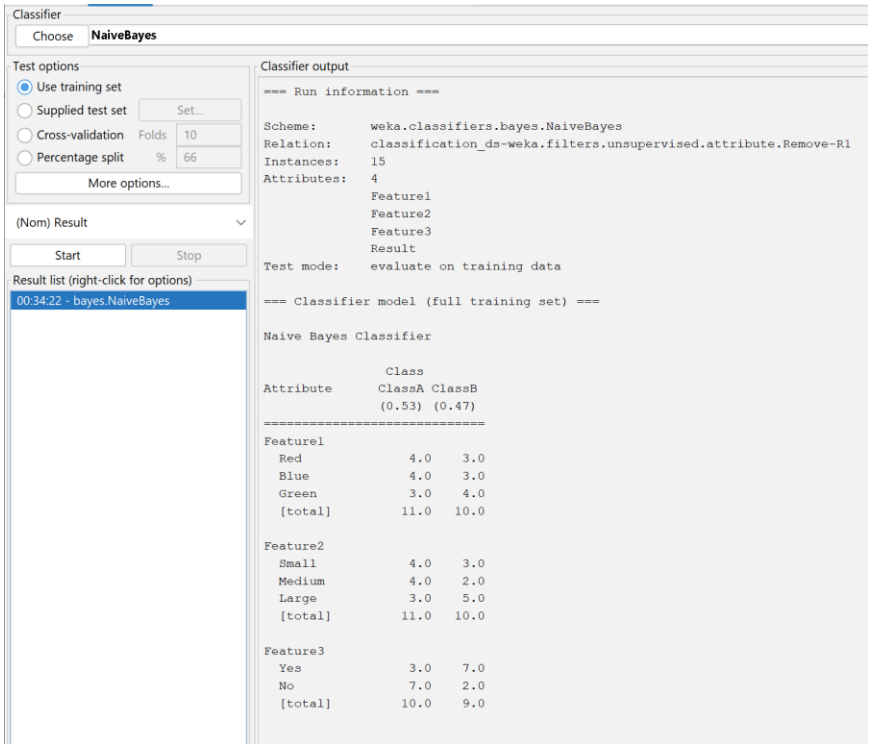
до лабораторної роботи №09
на тему **“Баєсівський класифікатор.”**

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.

Індивідуальне 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.

Індивідуальне завдання:
1 частина:






```
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          12               80      %
Incorrectly Classified Instances         3               20      %
Kappa statistic                          0.6018
Mean absolute error                      0.3216
Root mean squared error                  0.3754
Relative absolute error                 64.5755 %
Root relative squared error             75.2401 %
Total Number of Instances               15

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,750    0,143    0,857      0,750    0,800      0,607    0,884     0,893     ClassA
                 0,857    0,250    0,750      0,857    0,800      0,607    0,884     0,872     ClassB
Weighted Avg.    0,800    0,193    0,807      0,800    0,800      0,607    0,884     0,883

=== Confusion Matrix ===

 a b   <-- classified as
 6 2 | a = ClassA
 1 6 | b = ClassB
```

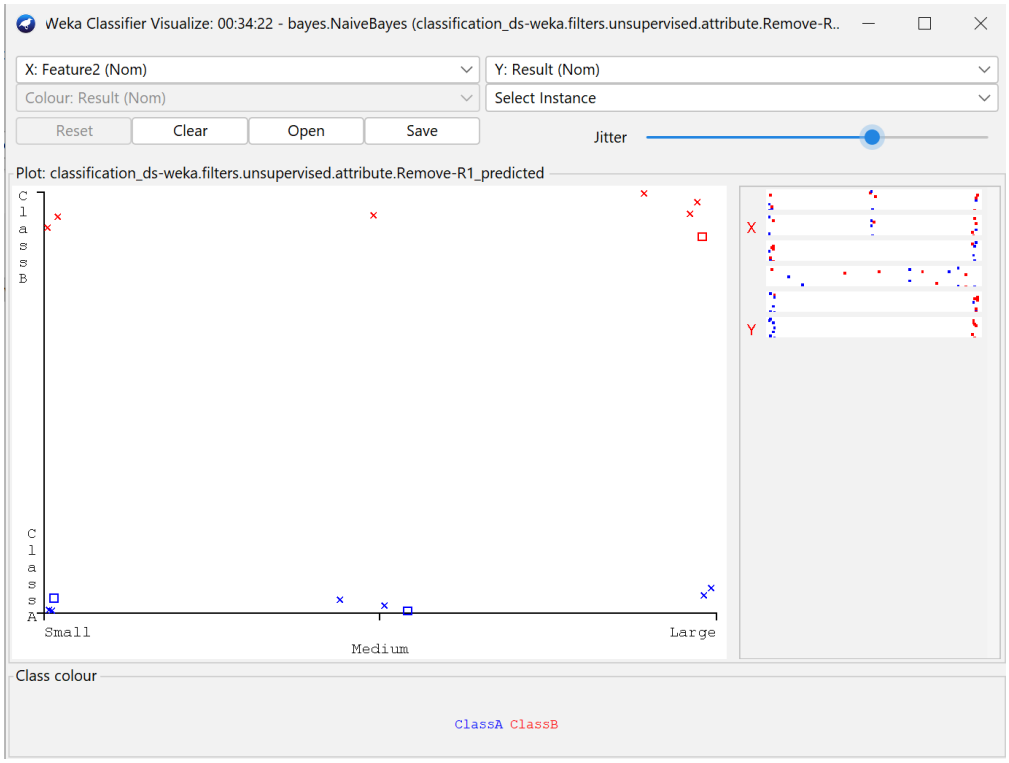
Виводиться матриця розміру 2 × 2, де 2 – кількість класів, ij-й елемент матриці дорівнює кількості об'єктів з i-го класу, які були віднесені до j-го. Кількість правильно класифікованих об'єктів дорівнює сумі елементів, що стоять на головній діагоналі. Правильно співвіднесено 12 об’єктів з 15.

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), 0,750. 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), 0,143. Значення Precision дорівнює відсотку правильно класифікованих об’єктів з об’єктів, віднесених алгоритмом до класу (відношення діагонального елемента до суми елементів стовпця), 0,857. Значення F-Measure обчислюється за такою формулою:
2 * Precision * Recall / (Precision + Recall), тобто це середнє гармонійне Precision (0,857) і Recall (0,750), результат — 0,800.

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.



2 частина:

No	Feature 1	Feature 2	Feature 3	Result
1	Red	Small	Yes	Class A
2	Blue	Large	No	Class B
3	Green	Medium	Yes	Class A
4	Red	Large	Yes	Class B
5	Blue	Small	No	Class A
6	Green	Large	Yes	Class B
7	Red	Medium	No	Class A
8	Blue	Large	No	Class A
9	Green	Small	Yes	Class B
10	Red	Large	No	Class A
11	Blue	Medium	Yes	Class B
12	Green	Medium	No	Class A
13	Red	Small	Yes	Class B
14	Blue	Small	No	Class A
15	Green	Large	Yes	Class B
Scoring	Blue	Medium	No	???

Sample size	15	Result					
		Class A	Class B				
	count	8	7				
	probability	0,53333333	0,46666667				
Feature 1	Red	0,375	0,28571429				
	Blue	0,375	0,28571429				
	Green	0,25	0,42857143				
Feature 2	Small	0,375	0,28571429				
	Medium	0,375	0,14285714				
	Large	0,25	0,57142857				
Feature 3	Yes	0,25	0,85714286				
	No	0,75	0,14285714				
No	Feature 1	Feature 2	Feature 3	Result			
1	Red	Small	Yes	Class A			
2	Blue	Large	No	Class B			
3	Green	Medium	Yes	Class A			
4	Red	Large	Yes	Class B			
5	Blue	Small	No	Class A			
6	Green	Large	Yes	Class B			
7	Red	Medium	No	Class A			
8	Blue	Large	No	Class A			
9	Green	Small	Yes	Class B			
10	Red	Large	No	Class A			
11	Blue	Medium	Yes	Class B			
12	Green	Medium	No	Class A			
13	Red	Small	Yes	Class B			
14	Blue	Small	No	Class A			
15	Green	Large	Yes	Class B	p'(scoring|Class A)	p'(scoring|Class B)	p'(Class A|scoring)
Scoring	Blue	Medium	No	Class A	0,05625	0,002721088	0,953857246

3 частина:

1. Результати у Weka:

- Використаний алгоритм: bayes.NaiveBayes
- Кількість правильно класифікованих об'єктів: 12 із 15
- Метрики:
  - True Positive Rate (Recall): 0.75
  - Precision: 0.857
  - F-Measure: 0.8
- Візуалізація даних: можливість аналізу розподілу точок та вибору атрибутів

2. Результати у Excel:

- Дані класифіковані на основі табличних розрахунків за Байєсовським методом
- Ймовірності для кожного класу обчислені за допомогою умовних частот
- Було знайдено p’(scoring|Class A), p’(scoring|Class B), p’(Class A|scoring)
- Використані формули COUNTIF, COUNTIFS, SUMIF для обчислення ймовірностей

3. Основні відмінності:

Характеристика	Weka	Excel
Автоматизація	Висока, миттєве навчання та тестування	Ручний розрахунок формул
Точність	80% (12/15)	Аналогічна, залежить від коректності обчислень
Гнучкість	Великий вибір параметрів, можливість змінювати алгоритми	Вимагає ручного налаштування
Візуалізація	Є інтерактивні графіки	Обмежена, лише табличні дані
Простота	Автоматизовано	Вимагає ручного розрахунку

Weka забезпечує швидке та ефективне навчання класифікатора з автоматичним обчисленням метрик. Excel дозволяє краще зрозуміти процес класифікації, але потребує значних ручних розрахунків. У цій задачі результати в обох системах подібні, однак Weka зручніша для масштабних задач.

Висновок: в ході виконання лабораторної роботи було проведено аналіз класифікації даних за допомогою двох підходів: автоматизованого методу у Weka та ручного розрахунку ймовірностей у Excel. Було сформовано набір даних із характеристиками гравців та їхньою матчевою ефективністю, після чого виконано обчислення ймовірностей для кожного класу та побудовано модель на основі методу Наївного Байєса. В результаті аналізу встановлено, що Weka забезпечує швидке навчання моделі, автоматичний розрахунок метрик точності та

можливість порівняння різних алгоритмів класифікації, тоді як Excel дозволяє вручну розраховувати ймовірності, що сприяє кращому розумінню процесу, але є менш ефективним для обробки великих обсягів даних. Отримані результати показали, що обидва підходи дають схожі значення точності, проте Weka значно перевершує Excel за зручністю використання та масштабованістю. Таким чином, можна зробити висновок, що для реальних задач аналізу даних та машинного навчання доцільніше використовувати спеціалізовані програмні засоби, такі як Weka, тоді як Excel є корисним для навчальних цілей і базового аналізу.