Прізвище: **КИРИЛЮК**
Ім'я: **Дмитро**
Група: **ПП-22**
Варіант: **08**
Дата захисту: **21.04.2025р**.

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

ЗВІТ

до лабораторної роботи №12
на тему **“Класифікація методом дерев рішень.”**

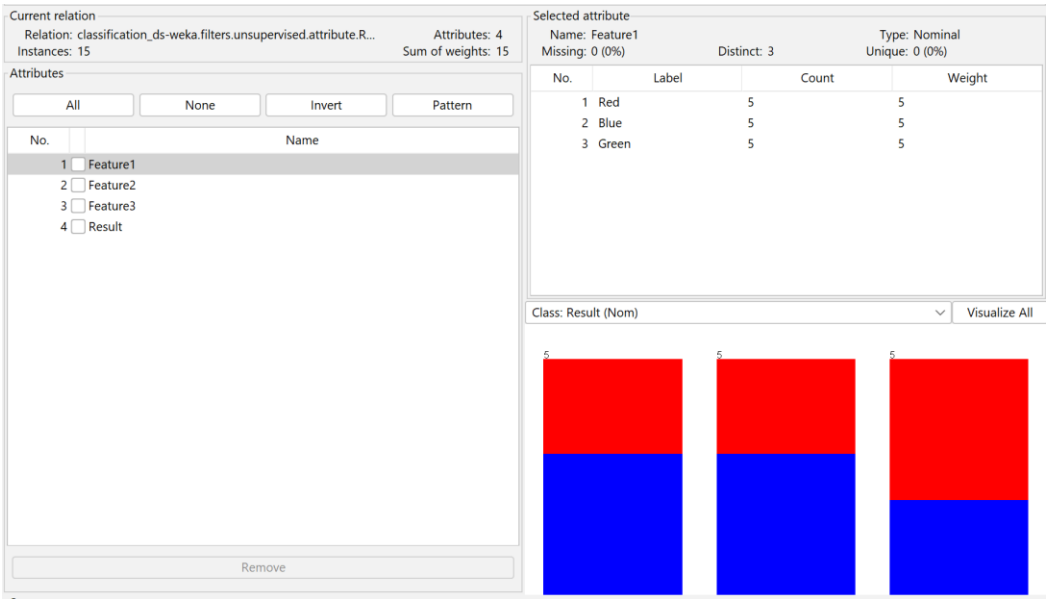
Мета роботи: Ознайомитися та отримати навички побудови моделей класифікації за допомогою Data Mining GUI бібліотеки Weka та Excel. На практиці вивчити роботу методу побудови дерев рішень, навчитися інтерпретувати результати роботи класифікаторів.

Індивідуальне завдання:

1. Для індивідуального завдання вирішіть задачу класифікації за допомогою наступного алгоритму: • метод побудови дерев рішень C4.5 (trees.J48).
2. Змінюючи параметри налаштування алгоритму, спробуйте досягти найвищої якості навчання класифікатора.
3. Для цього ж датасету побудуйте дерево рішень у Excel.
4. Порівняйте отримані результати отримані у різних системах.
5. У звіті надайте результати роботи роботи алгоритму, його налаштування.

Mood	Opponent	Training	Weather	WillScore
focused	strong	yes	sunny	yes
nervous	medium	no	rainy	no
confident	weak	yes	overcast	yes
relaxed	medium	no	sunny	no
focused	weak	yes	sunny	yes
nervous	strong	yes	rainy	no
confident	medium	yes	overcast	yes
relaxed	weak	no	sunny	yes
focused	medium	no	overcast	yes
nervous	strong	no	rainy	no
confident	medium	yes	sunny	yes
nervous	strong	yes	overcast	no
nervous	weak	yes	sunny	yes
focused	weak	no	rainy	yes

1 частина:




```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     classification_ds-weka.filters.unsupervised.attribute.Remove-R1
Instances:    15
Attributes:   4
              Feature1
              Feature2
              Feature3
              Result
Test mode:    evaluate on training data

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

J48 pruned tree
------------------

Feature3 = Yes: ClassB (8.0/2.0)
Feature3 = No: ClassA (7.0/1.0)

Number of Leaves  :      2

Size of the tree :  3


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                      0.3143
```

```
Root mean squared error                  0.3964
Relative absolute error                 63.1046 %
Root relative squared error             79.4568 %
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    0,804     0,776
ClassA
                 0,857    0,250    0,750      0,857    0,800      0,607    0,804     0,710
ClassB
Weighted Avg.    0,800    0,193    0,807      0,800    0,800      0,607    0,804     0,745

=== Confusion Matrix ===

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

Загальні характеристики моделі

Модель була побудована за допомогою алгоритму J48, що є реалізацією алгоритму дерева рішень C4.5. Цей метод дозволяє створити просту, інтерпретовану структуру класифікації, де кожне рішення базується на значеннях вхідних ознак.

- Параметри алгоритму: -C 0.25 (коефіцієнт довіри для обрізки), -M 2 (мінімум 2 об’єкти на лист).
- Кількість листків: 2
- Розмір дерева (кількість вузлів): 3Це означає, що дерево має дуже просту структуру, що позитивно впливає на швидкість прийняття рішень та зручність для людини.
- Час побудови моделі: 0 секундЧерез малий обсяг даних (15 записів) модель була побудована миттєво.

Результати на тренувальному наборі

- Кількість записів: 15
- Правильно класифіковано: 12 (80%)
- Неправильно класифіковано: 3 (20%)

Модель показала досить хорошу точність на навчальних даних, однак без додаткової перевірки (тестової вибірки або крос-валідації) неможливо оцінити її узагальнюючу здатність.

Метрики якості класифікації

- Карра-статистика: 0.6018. Значення вище 0.6 свідчить про помітне узгодження класифікатора з реальними класами — значно краще, ніж випадковий вибір.
- Середня абсолютна помилка (MAE): 0.3143
- Корінь середньоквадратичної помилки (RMSE): 0.3964
- Відносна абсолютна помилка: 63.10%
- Відносна середньоквадратична помилка: 79.46%

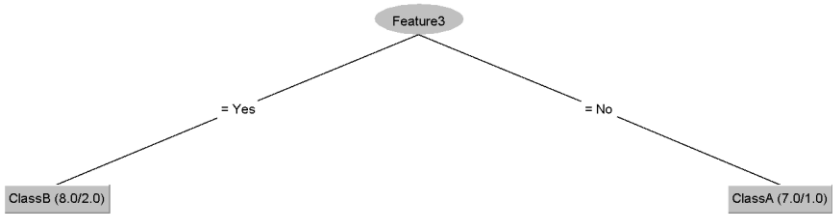
Ці показники демонструють помірний рівень точності, але є простір для покращення моделі (наприклад, через інші параметри чи ознаки).

Матриця плутанини (Confusion Matrix)

	Прогноз: ClassA	Прогноз: ClassB
ClassA (a)	6 (правильно)	2 (помилково в ClassB)
ClassB (b)	1 (помилково в ClassA)	6 (правильно)

- ClassA: 6 правильно, 2 помилки (FP)
- ClassB: 6 правильно, 1 помилка (FN)

Це свідчить про невеликий дисбаланс у точності між класами. Проте модель має досить збалансовану поведінку та відсутність серйозного перекосу на користь одного з класів.



2 частина:

Feature 1	Feature 2	Feature 3	Result	No	Large				
Red	Small	Yes	Class A						
Blue	Large	No	Class B						
Green	Medium	Yes	Class A						
Red	Large	Yes	Class B						
Blue	Small	No	Class A						
Green	Large	Yes	Class B						
Red	Medium	No	Class A						
Blue	Large	No	Class A						
Green	Small	Yes	Class B						
Red	Large	No	Class A						
Blue	Medium	Yes	Class B						
Green	Medium	No	Class A						
Red	Small	Yes	Class B						
Blue	Small	No	Class A						
Green	Large	Yes	Class B						
Sample size	7								
				p*log(p)	entropy				
	Result	Class A	2	-0,51638712	0,917437824				
		Class B	1	-0,4010507					
			Class A	Class B	p*log(p)-A	p*log(p)-B	weighted	entropy	info gain
1		Red	1	0	0	0	0		
1	Feature 1	Blue	1	1	-0,5	-0,5	0,285714286	0,2857143	0,631724
1		Green	0	0	0	0	0		
2		Small	0	0	0	0	0		
2	Feature 2	Medium	0	0	0	0	0	0,3935554	0,523882
2		Large	2	1	-0,389975	-0,528320834	0,393555357		
3	Feature 3	Yes	0	0	0	0	0	0,3935554	0,523882
3		No	2	1	-0,389975	-0,528320834	0,393555357		
					Feature 3				
			Yes(2,6)				No(6,1)		
		Small(1,2)	Medium(1,1)	Large(0,3)		Small(2,0)	Medium(2,0)	Large(2,1)	
		Red(1,1)	Red(0,0)					Red(1,0)	
		Blue(0,0)	Blue(0,1)					Blue(1,1)	
		Green(0,1)	Green(1,0)					Green(0,0)	

3 частина :

Процес побудови

- Weka: Автоматично обрав Feature3 як корінь дерева на основі інформаційного приросту, що є частиною реалізації алгоритму C4.5 (J48). У моделі було застосовано обрізку дерева, що призвело до простоти.
- Excel: Дерево створювалося вручну з поетапним обрахунком ентропії та інформаційного приросту. Кожне розгалуження вибиралось на основі найбільшої інформаційної вигоди, що дозволяє дослідити логіку класифікації глибше.

Точність класифікації

- Weka: Досягнуто 80% точності — 12 із 15 прикладів були класифіковані правильно. Незначна кількість помилок присутня в обох класах, що видно в матриці плутанини.
- Excel: Точність не вираховувалась чисельно, однак структура дерева деталізована — більшість шляхів ведуть до однозначної класифікації, що вказує на високу якість розділення.

Вибір атрибутів

- Weka: Вибір обмежився Feature3, що вказує на його найвищу інформаційну значущість. Алгоритм J48 також враховує обрізку, щоб уникнути перенавчання.
- Excel: Використано класичний підхід ID3, де після Feature3 у розгалуженні задіяно Feature1 та Feature2, що забезпечує більшу гнучкість у класифікації, але також ускладнює дерево.

Гнучкість налаштувань

- Weka: Дозволяє легко змінювати такі параметри, як -C (confidence factor) чи -M (min. instances per leaf), що дає контроль над складністю моделі.
- Excel: Обмежена ручною реалізацією — будь-які зміни вимагають повного перерахунку, що може бути менш зручним, але корисним для навчання.

Обидва підходи показали ефективність у вирішенні задачі класифікації. Weka продемонстрував швидкість і простоту, зробивши ставку на єдиний найінформативніший атрибут. Excel дозволив заглибитись у логіку розділення, надаючи більше деталей, що є особливо корисним для навчальних цілей.

Висновок: У результаті виконання лабораторної роботи було успішно застосовано метод дерев рішень для вирішення задачі класифікації з використанням двох різних підходів: автоматизованого аналізу в Weka та ручного побудови дерева в Excel.
За допомогою алгоритму J48 (реалізація C4.5) у Weka було побудовано компактне дерево з 2 листками, яке ґрунтується лише на одному атрибуті (Feature3). Модель досягла точності 80% на тренувальному наборі, продемонструвавши ефективність автоматичної побудови з використанням обрізки та інформаційного приросту.
Натомість у Excel було реалізовано більш детальне дерево, що включає послідовні розгалуження за Feature3, Feature1 та Feature2, із ручними розрахунками ентропії та інформаційного приросту. Такий підхід дав змогу краще зрозуміти механізм прийняття рішень на кожному етапі класифікації.
Отримані результати підтверджують, що обидва інструменти є ефективними для реалізації методу дерев рішень. Автоматизований підхід у Weka забезпечує швидку оцінку та оптимізацію, тоді як ручна побудова в Excel поглиблює розуміння алгоритму та його логіки. Це дозволило опанувати як технічні, так і концептуальні аспекти класифікації на основі дерев рішень — важливого інструменту в сфері інтелектуального аналізу даних.