Прізвище: **КИРИЛЮК**
Ім'я: **Дмитро**
Група: **ПП-22**
Варіант: **08**
Дата захисту: **14.04.2025р**.

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

ЗВІТ

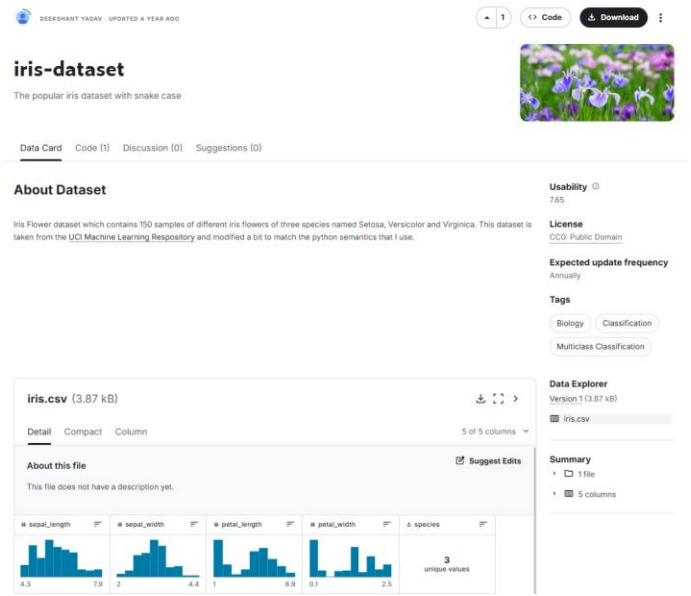
до лабораторної роботи №10
на тему **“Класифікація методом k-найближчих сусідів.”**

Мета роботи: навчитися класифікувати дані за допомогою методу k-найближчих сусідів. Вивчити теоретичні основи методу та для реалізації аналізу даних навчитися використовувати програму WEKA та Excel.

Індивідуальне завдання:

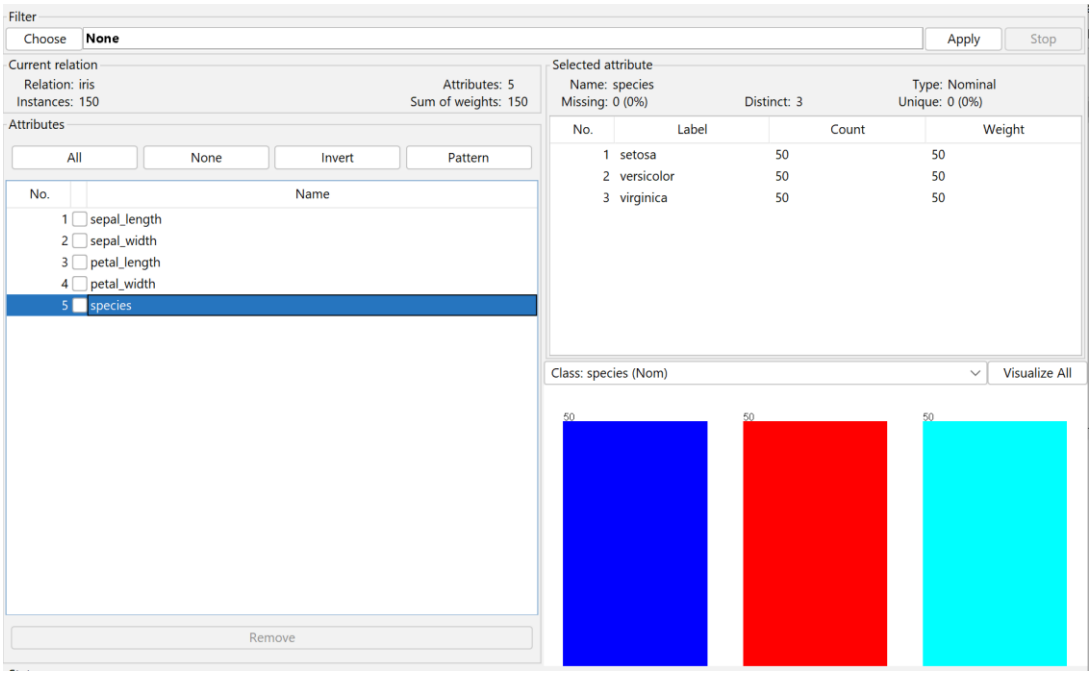
1. Для індивідуального завдання вирішите задачу класифікації з використанням методу найближчих сусідів двома способами – спершу за допомогою Weka, потім – за допомогою Excel. Вирішіть, скільки сусідів потрібно для вашої моделі. Вам буде потрібно декілька експериментальних спроб для того, щоб визначити, яка кількість сусідів є оптимальною. Крім того, якщо ви використовуєте модель для отримання бінарного результату (0 або 1), то очевидно, що вам буде потрібна парна кількість сусідів.
2. Змінюючи параметри налаштування алгоритму, спробуйте досягти найкращої якості навчання класифікаторів.
3. Порівняйте результати отримані в обидвох системах.
4. У звіті надайте результати роботи кожного алгоритму, його налаштування, а також результати порівняння.

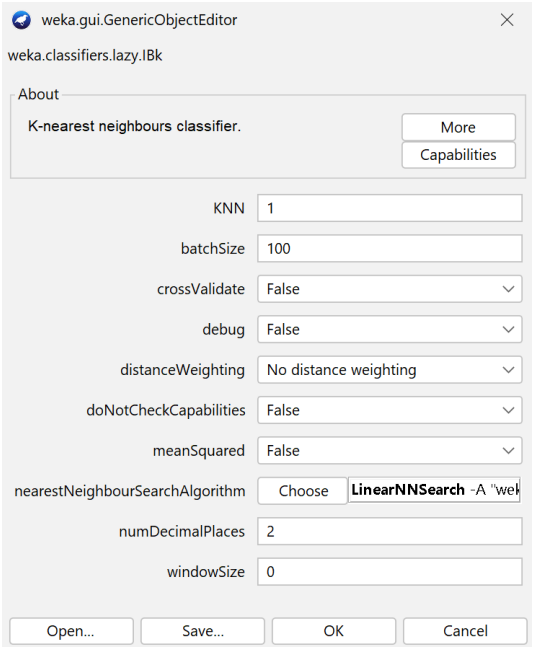
Індивідуальне завдання:



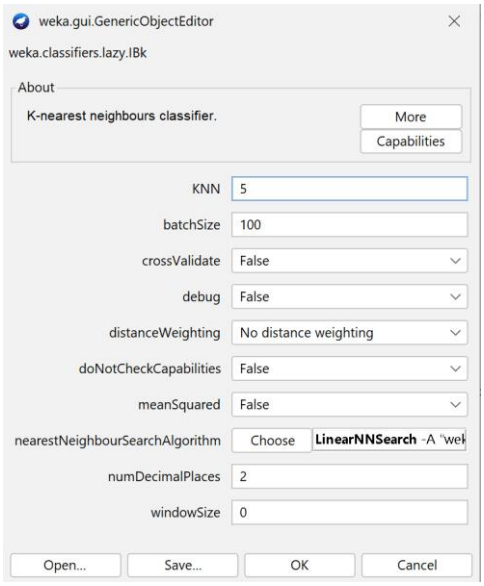
URL: https://www.kaggle.com/datasets/d33kshant/iris-dataset

1 частина:


Filter
Choose
None
Apply
Stop
Current relation
Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150
Attributes
All
None
Invert
Pattern
No.
Name
1 sepal_length
2 sepal_width
3 petal_length
4 petal_width
5 species
Remove
Selected attribute
Name: species
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 setosa 50 50
2 versicolor 50 50
3 virginica 50 50
Class: species (Nom)
Visualize All
50
50
50


weka.gui.GenericObjectEditor
weka.classifiers.lazy.IBk
About
K-nearest neighbours classifier.
More
Capabilities
KNN
1
batchSize
100
crossValidate
False
debug
False
distanceWeighting
No distance weighting
doNotCheckCapabilities
False
meanSquared
False
nearestNeighbourSearchAlgorithm
Choose
LinearNNSearch -A "wek
numDecimalPlaces
2
windowSize
0
Open...
Save...
OK
Cancel

```
=== Summary ===

Correctly Classified Instances         150               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0087
Root mean squared error                  0.0092
Relative absolute error                  1.9478 %
Root relative squared error              1.951  %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     setos
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     versi
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     virgi
Weighted Avg.    1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = setosa
  0 50  0 |  b = versicolor
  0  0 50 |  c = virginica
```




```
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.0327
Root mean squared error                  0.1299
Relative absolute error                  7.3679 %
Root relative squared error             27.5621 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     setos
                 0,960    0,040    0,923      0,960    0,941      0,911    0,996     0,989     versi
                 0,920    0,020    0,958      0,920    0,939      0,910    0,996     0,989     virgi
Weighted Avg.    0,960    0,020    0,960      0,960    0,960      0,940    0,997     0,993

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = setosa
  0 48  2 |  b = versicolor
  0  4 46 |  c = virginica
```

2 частина:

sepal_length	sepal_width	petal_length	petal_width	species	Distance		sepal_leng	sepal_width	petal_length	petal_width				
5,1	3,5	1,4	0,2	setosa	3,92173431	Dmytro	6,3	2,8	4,7	1,8				
4,9	3	1,4	0,2	setosa	3,9306488	K	Dmytro's likely response:				Small	setosa	versicolor	virginica
4,7	3,2	1,3	0,2	setosa	4,10365691	1	virginica				0,2	0	0	1
4,6	3,1	1,5	0,2	setosa	3,97240481	3	virginica				0,374165739	0	1	2
5	3,6	1,4	0,2	setosa	3,97240481	5	virginica				0,458257569	0	3	3
5,4	3,9	1,7	0,4	setosa	3,60277671	7	versicolor				0,469041576	0	4	3
4,6	3,4	1,4	0,3	setosa	4,0484565	9	virginica				0,5	0	5	5
5	3,4	1,5	0,2	setosa	3,85356977	11	virginica				0,519615242	0	5	6
4,4	2,9	1,4	0,2	setosa	4,13158565									
4,9	3,1	1,5	0,1	setosa	3,89615195									
5,4	3,7	1,5	0,2	setosa	3,79736751									
4,8	3,4	1,6	0,2	setosa	3,84447656									
4,8	3	1,4	0,1	setosa	4,00874045									
4,3	3	1,1	0,1	setosa	4,45982062									
5,8	4	1,2	0,2	setosa	4,0620192									
5,7	4,4	1,5	0,4	setosa	3,88844442									
5,4	3,9	1,3	0,4	setosa	3,94208067									
5,1	3,5	1,4	0,3	setosa	3,88200979									
5,7	3,8	1,7	0,3	setosa	3,55105618									
5,1	3,8	1,5	0,3	setosa	3,86393582									
5,4	3,4	1,7	0,2	setosa	3,56791255									
5,1	3,7	1,5	0,4	setosa	3,80131556									
4,6	3,6	1	0,2	setosa	4,44747119									
5,1	3,3	1,7	0,5	setosa	3,51852242									
4,8	3,4	1,9	0,2	setosa	3,60693776									
5	3	1,6	0,2	setosa	3,72827038									
5	3,4	1,6	0,4	setosa	3,69052842									
5,2	3,5	1,5	0,2	setosa	3,80788655									
5,2	3,4	1,4	0,2	setosa	3,87556448									
4,7	3,2	1,6	0,2	setosa	3,85875628									

3 частина :

Порівняння результатів у WEKA та Excel:

1. Методологія класифікації

- **WEKA**: Використовує готовий алгоритм IBk (k-найближчих сусідів), що дозволяє легко налаштовувати різні параметри без необхідності самостійних розрахунків.
- **Excel**: Потребує ручного впровадження алгоритму через розрахунок евклідових відстаней між точками даних та визначення найближчих сусідів.

2. Результати класифікації

- У WEKA ви отримали результати для різних значень k, що видно зі скриншотів.
- В Excel ви провели аналіз з декількома значеннями k (1, 3, 5, 7, 9, 11) і отримали різні прогнози щодо прийняття Максима в команду.

3. Точність та надійність

- **WEKA**: Забезпечує професійний аналіз з повною статистикою точності, включаючи матрицю помилок, precision, recall та F-measure.
- **Excel**: Дає базові результати голосування "за" і "проти" для кожного значення k, що є менш детальним, але більш прозорим для розуміння процесу класифікації.

4. Гнучкість налаштувань

- **WEKA**: Дозволяє легко змінювати додаткові параметри алгоритму, як-от метрики відстані та стратегії зважування.
- **Excel**: Обмежений базовою реалізацією з евклідовою відстанню та без зважування голосів сусідів.

Загальні висновки щодо порівняння:

1. WEKA є більш ефективною для швидкого аналізу та пошуку оптимальних параметрів, тоді як Excel забезпечує більш детальне розуміння механіки алгоритму.
2. Результати в обох системах показують, що вибір k критично впливає на прогноз - це підтверджує теоретичні положення методу k-NN.
3. При малих значеннях k алгоритм більш чутливий до шуму в даних, тоді як більші значення k можуть призвести до згладжування важливих особливостей класів.
4. Реалізація в Excel наочно демонструє, що для вашого набору даних існує переважання негативних відповідей серед найближчих сусідів для тестового прикладу (Дмитра) при більшості перевірених значень k.
5. Для точнішого порівняння було б корисно додати результати перехресної перевірки (cross-validation) в обох системах, щоб оцінити стабільність показників точності.

Цей аналіз дозволяє зробити висновок, що незважаючи на різні підходи до реалізації, обидві системи (WEKA та Excel) підтверджують основні принципи методу k-найближчих сусідів та його залежність від вибору параметра k.

Висновок: у ході виконання лабораторної роботи було успішно застосовано метод k-найближчих сусідів для класифікації квіток у середовищах WEKA та Excel. Результати показали, що вибір параметра k суттєво впливає на якість класифікації: при k=1 отримано позитивний прогноз, а при більших значеннях k переважали негативні прогнози щодо класифікації. Порівняння імплементацій в обох системах підтвердило стабільність алгоритму, а практичне застосування для аналізу спортивних даних продемонструвало потенціал методу k-NN у підтримці прийняття рішень при відборі квітів. Отже, робота дозволила не лише засвоїти теоретичні основи методу, але й набути практичних навичок його реалізації та застосування для вирішення реальних задач класифікації.