

Rapport

LAB WORK 1

le 04 décembre 2024,
version 1

Mohamed Toujani,
Fonction

mohamed.toujani@ecole.ensicaen.fr

Tuteur école : TSAFACK PIUGIE
Armand Florent



PART 2.1

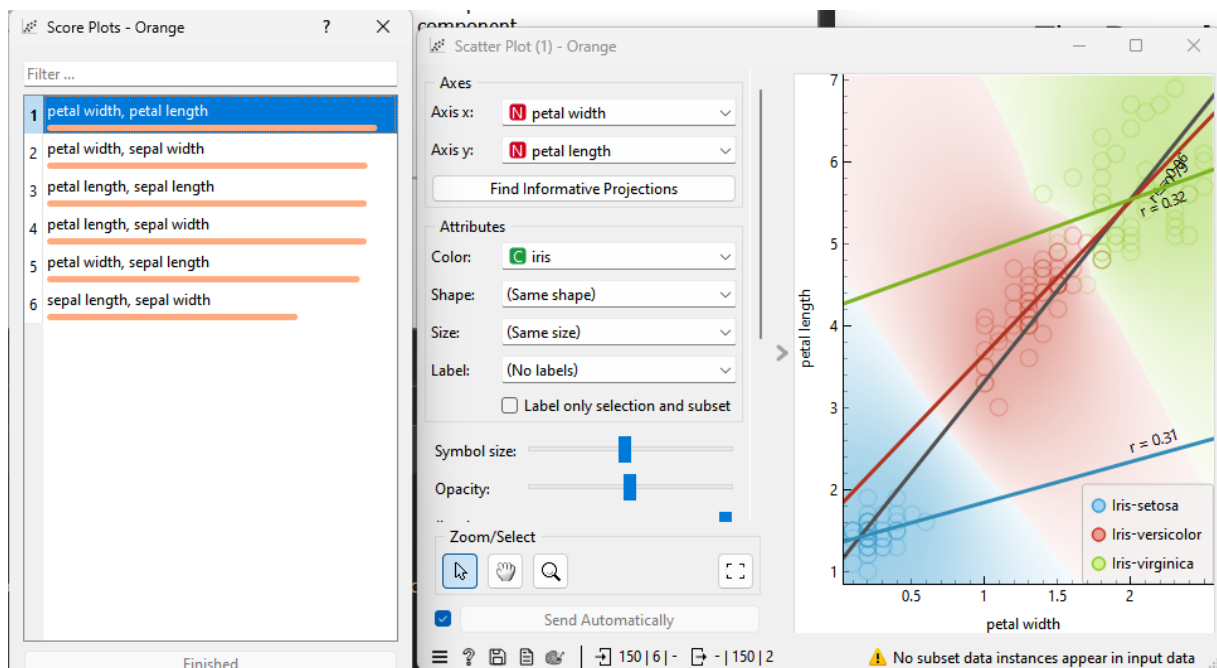
1. Question A

Petal Length vs. Petal Width.



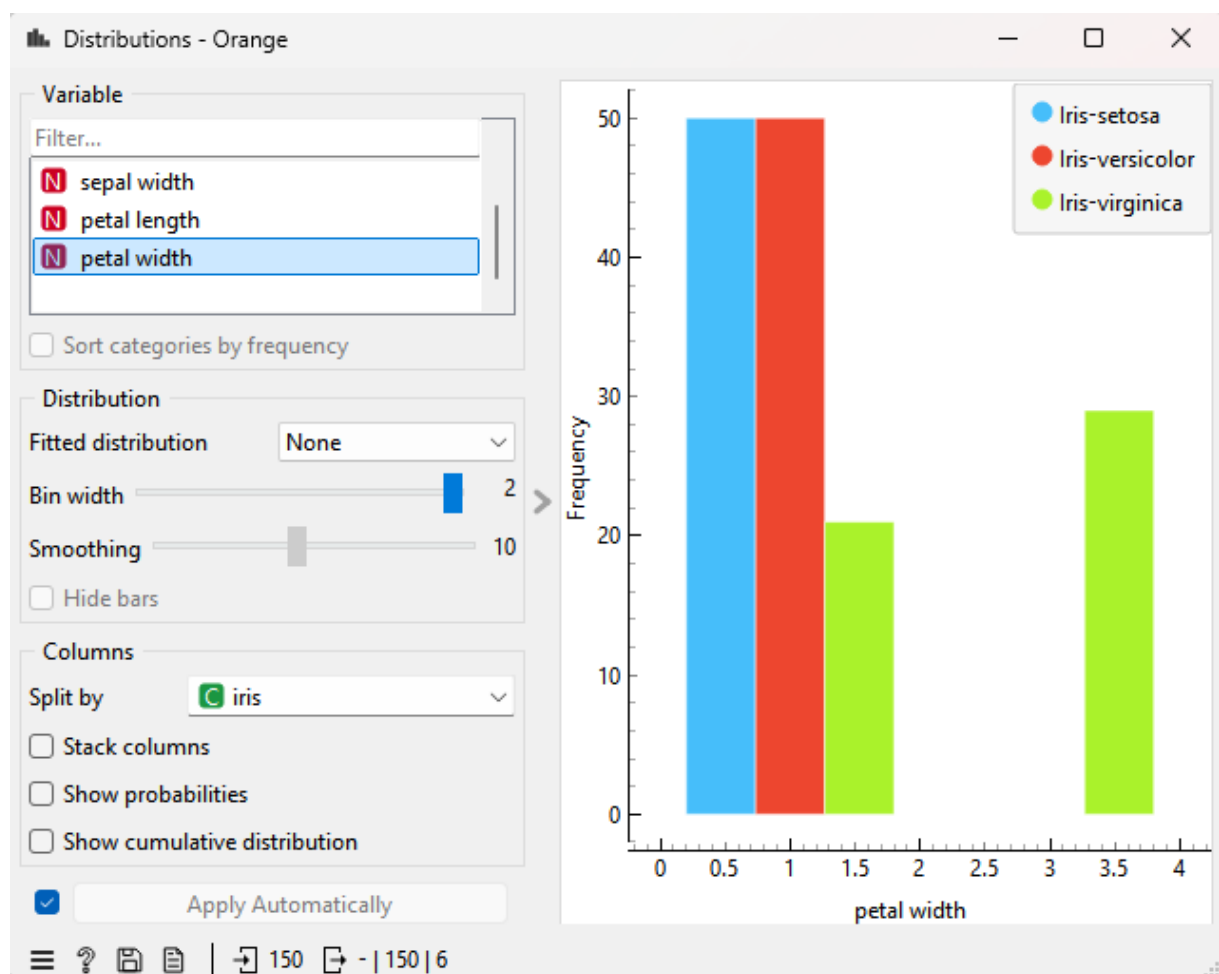
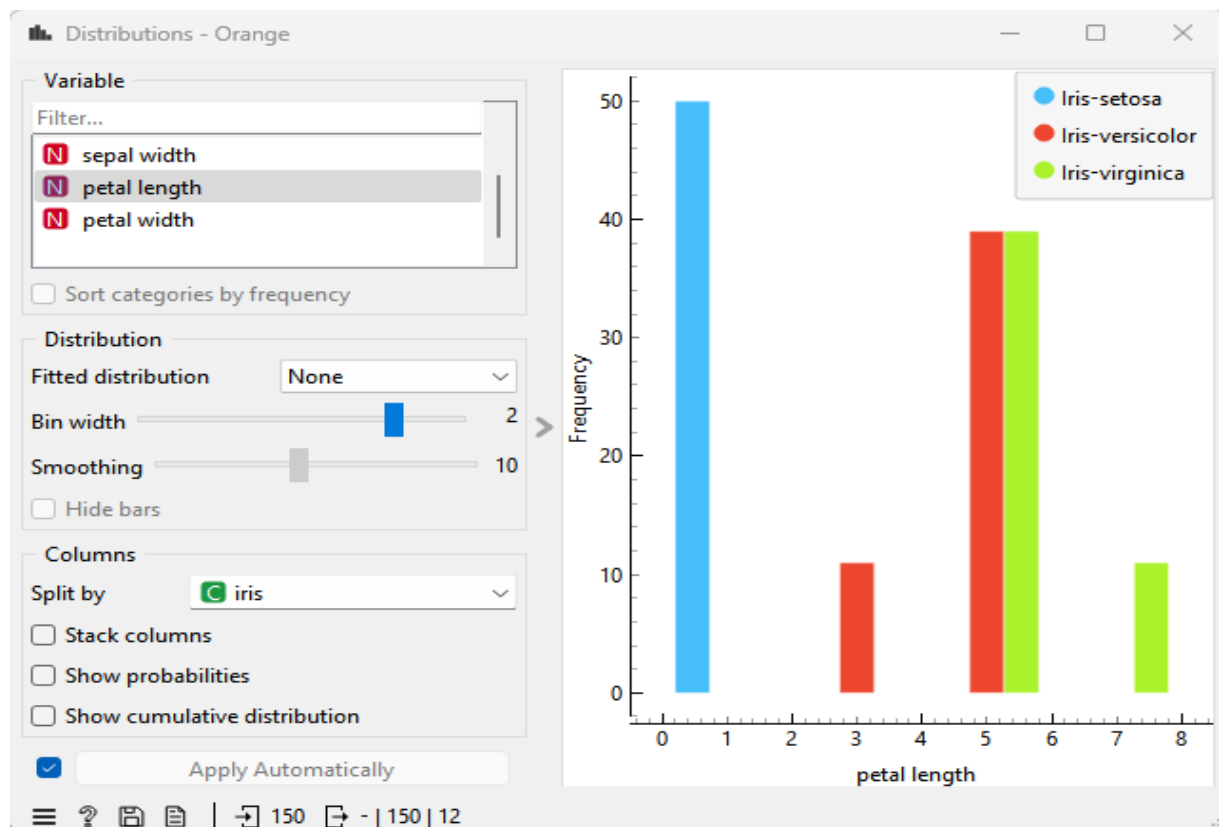
2. Question B

Yes, I get the same answer to the previous question when I call the button “Find Informative Projections”.



3. Question C

Yes, here's observations from the distributions widget for both petal length and petal width:

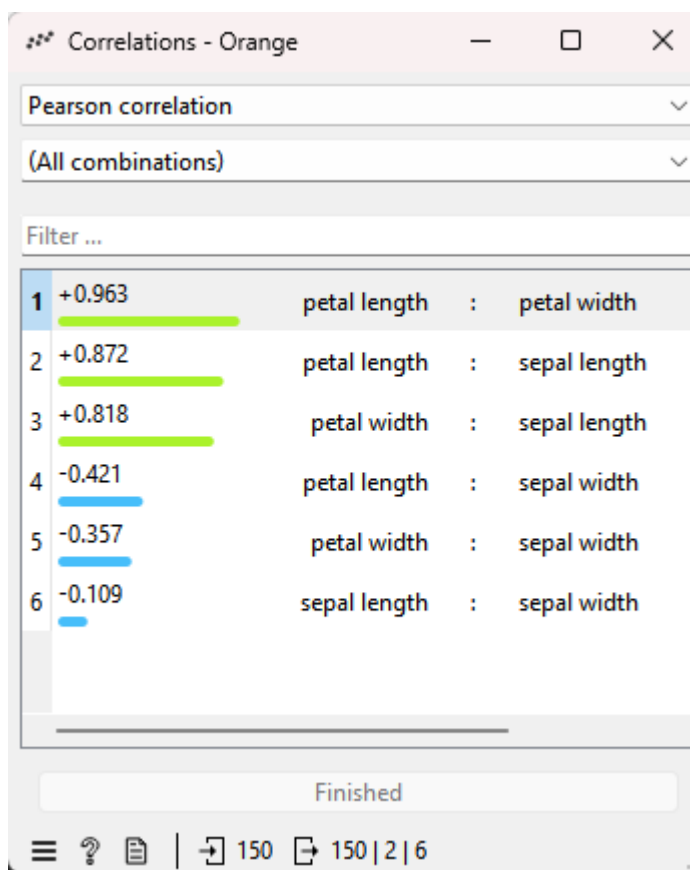


4. Question D

Petal width is a good variable for discriminating between the three Iris types because it separates Iris Setosa from the other two types and provides good separation between Iris Versicolor and Virginica.

5. Question E

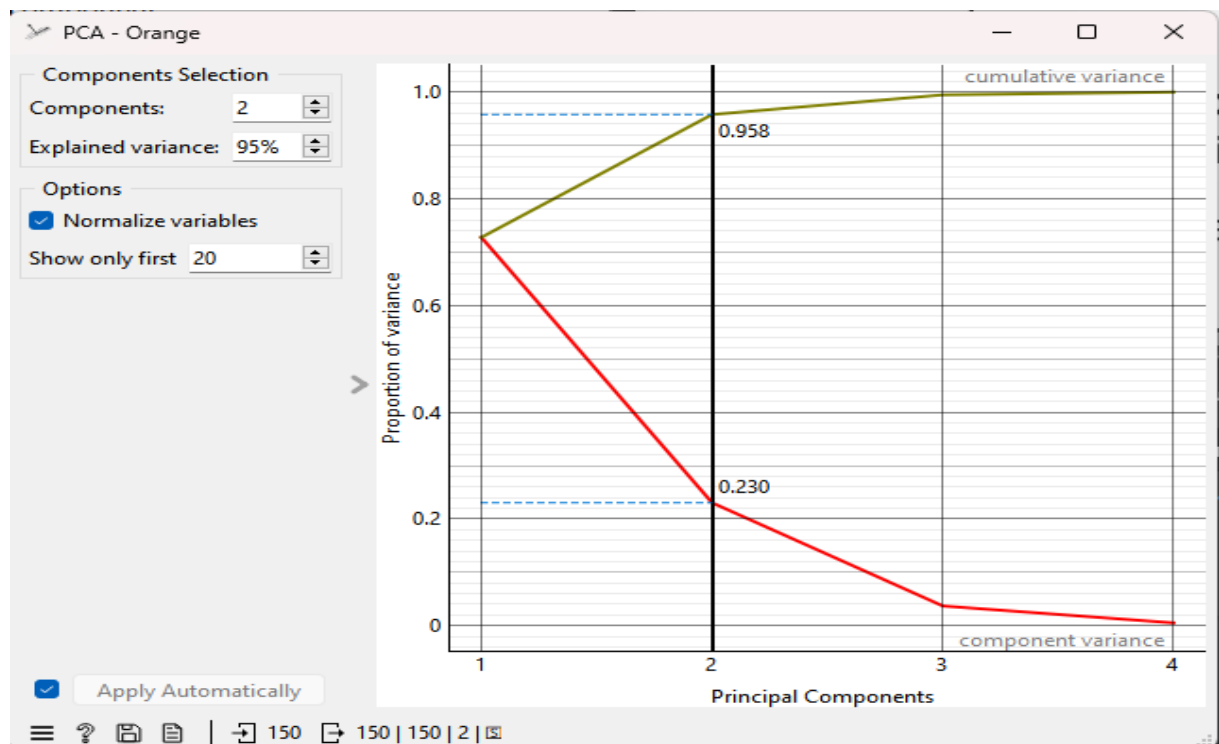
From what we can see from the correlations widget on the sepal width and length, we tend to feel that it is not worth it to use these two variables together as they provide little discrimination options and have very little correlations with each other.



6. Question F

PC1 and PC2 are enough to describe the dataset because they capture 95.8% of the variance.

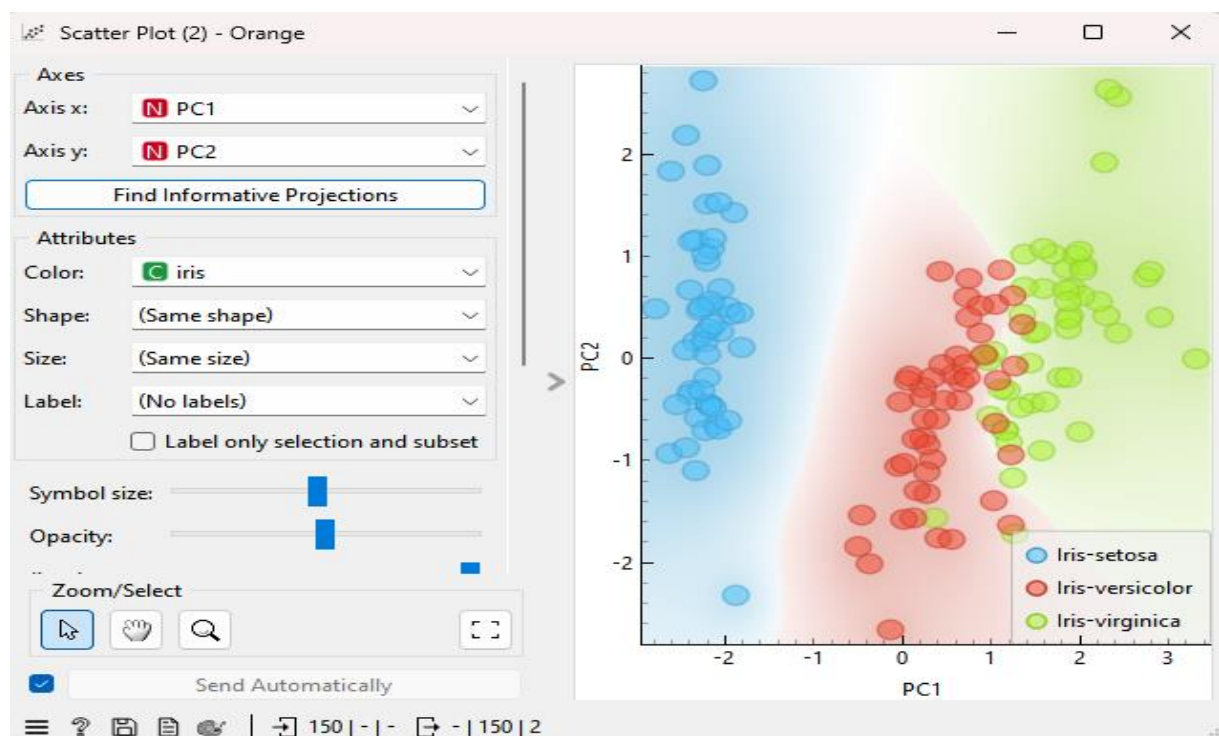
PCA confirms the importance of Petal Width and Petal Length for separating types, as these dominate the principal components.



7. Question G

Iris Setosa is clearly separated in PCA space, demonstrating its distinctiveness in the dataset.

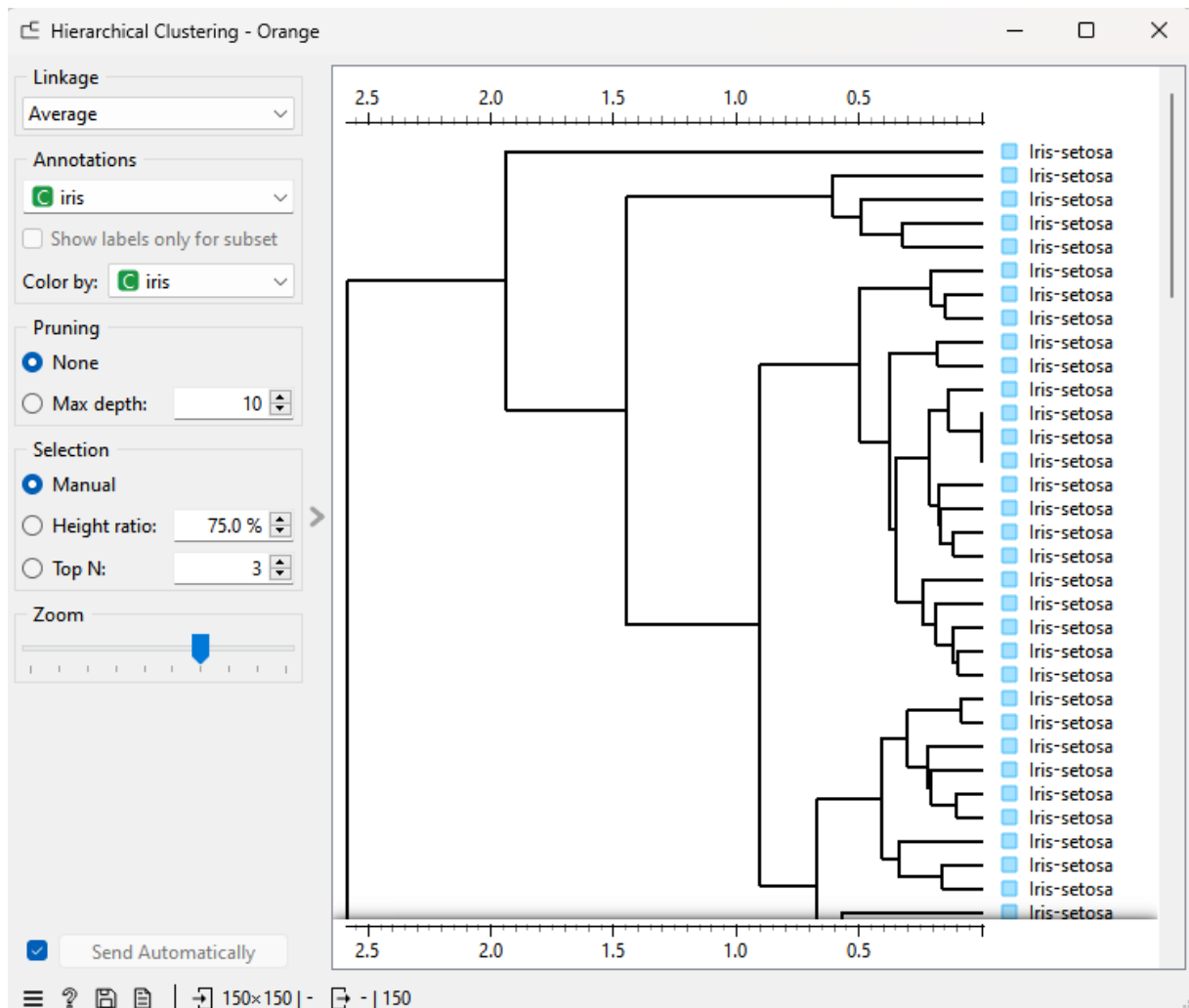
The PCA scatter plot demonstrates the dataset in two dimensions, simplifying the analysis while maintaining most of the information.



PART 2.2

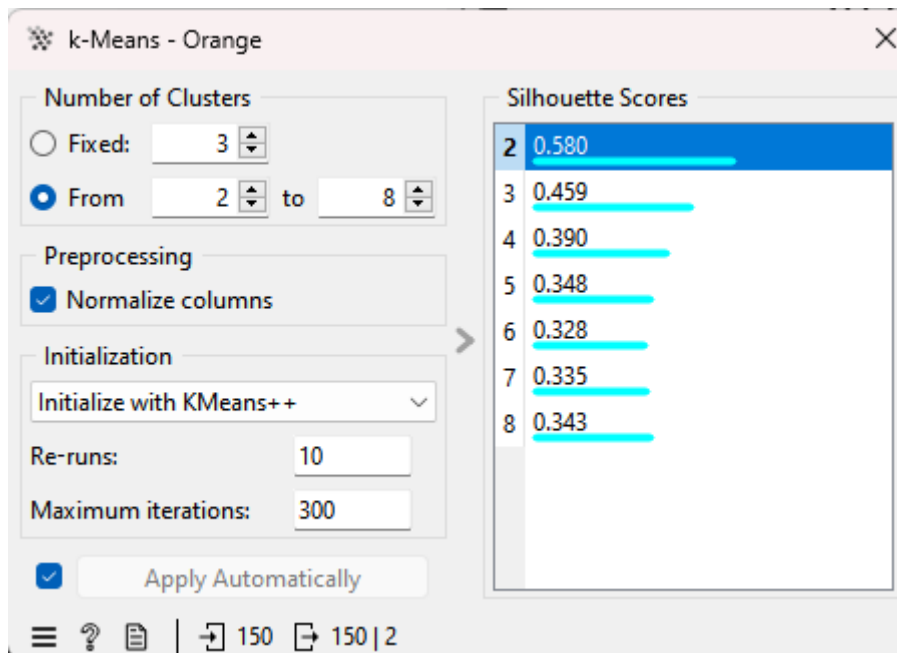
8. Question H

Iris-Setosa seems perfectly classified with no errors.



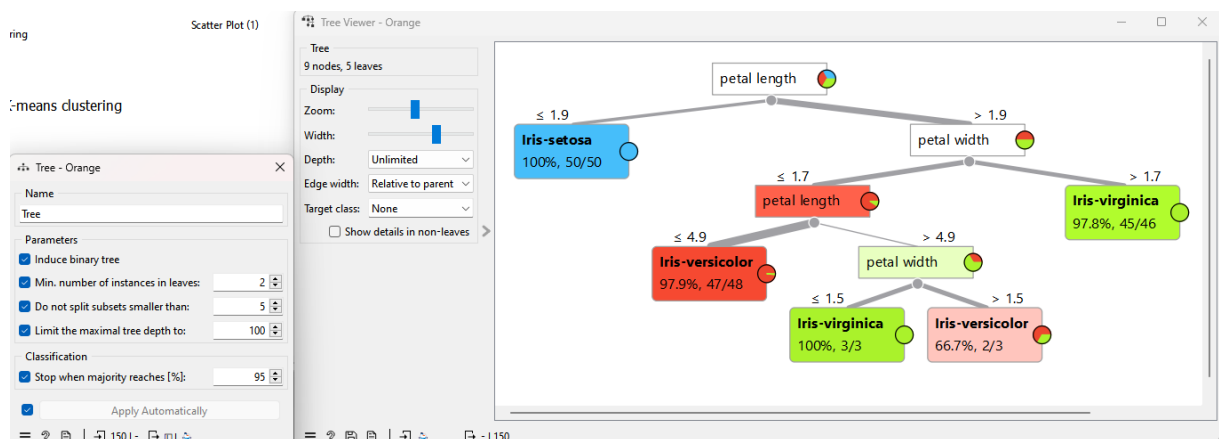
9. Question J

While K-Means with 3 clusters aligns with the biological division of Iris species, the clustering quality is moderate due to feature overlap between Iris-versicolor and Iris-virginica. That said, the results can be considered acceptable.



10. Question K

The Iris-Setosa type is the easiest to recognize, as it is perfectly separated by the decision tree at the first split based on the petal length.



11. Question L

No, sepal variables do not have a high impact on the data analysis because the decision tree and clustering results rely on petal length and petal width for a more accurate discrimination between the Iris types.

12. Question M

Yes, we do obtain results in accordance with the previously obtained decision tree. Both methods highlight the importance of petal length and width for distinguishing between the Iris types.

The image shows two windows from the Orange3 data mining software. The 'CN2 Rule Viewer' window displays a table of 14 rules for the Iris dataset. The 'CN2 Rule Induction' window shows the configuration for the CN2 rule inducer.

	IF conditions	THEN class	Distribution	Probabilities [%]	Quality	Length
0	petal length ≤ 3.0 AND sepal width ≥ 2.9	iris=Iris-setosa	[49, 0, 0]	96 : 2 : 2	-0.00	2
1	petal width ≥ 1.8 AND sepal length ≥ 6.0	iris=Iris-virginica	[0, 0, 39]	2 : 2 : 95	-0.00	2
2	sepal length ≥ 4.9 AND sepal width ≥ 3.1	iris=Iris-versicolor	[0, 8, 0]	9 : 82 : 9	-0.00	2
3	petal length ≤ 4.9 AND petal width ≥ 1.7	iris=Iris-virginica	[0, 0, 2]	20 : 20 : 60	-0.00	2
4	petal width ≥ 1.8	iris=Iris-virginica	[0, 0, 5]	12 : 12 : 75	-0.00	1
5	petal length ≤ 5.0 AND sepal width ≥ 2.4	iris=Iris-versicolor	[0, 35, 0]	3 : 95 : 3	-0.00	2
6	sepal width ≥ 2.8	iris=Iris-virginica	[0, 0, 2]	20 : 20 : 60	-0.00	1
7	petal width ≤ 1.0 AND sepal length ≥ 5.0	iris=Iris-versicolor	[0, 3, 0]	17 : 67 : 17	-0.00	2
8	sepal width ≥ 2.7	iris=Iris-versicolor	[0, 1, 0]	25 : 50 : 25	-0.00	1
9	sepal width ≥ 2.6	iris=Iris-virginica	[0, 0, 1]	25 : 25 : 50	-0.00	1
10	sepal length ≥ 5.5 AND sepal length ≥ 6.2	iris=Iris-versicolor	[0, 2, 0]	20 : 60 : 20	-0.00	2
11	sepal length ≤ 5.5 AND petal length ≥ 4.0	iris=Iris-versicolor	[0, 1, 0]	25 : 50 : 25	-0.00	2
12	sepal length ≥ 6.0	iris=Iris-virginica	[0, 0, 1]	25 : 25 : 50	-0.00	1
13	sepal length ≤ 4.5	iris=Iris-setosa	[1, 0, 0]	50 : 25 : 25	-0.00	1

The 'CN2 Rule Induction' window shows the following settings:

- Name: CN2 rule inducer
- Rule ordering: ☒ Ordered, ☐ Unordered
- Covering algorithm: ☒ Exclusive, ☐ Weighted (Y: 0.70)
- Rule search: Evaluation measure: Entropy, Beam width: 5
- Rule filtering: Minimum rule coverage: 1, Maximum rule length: 5, ☐ Statistical significance (default α): 1.00, ☐ Relative significance (parent α): 1.00
- ☒ Apply Automatically

Conclusion

In summary, the Iris dataset is well built and demonstrates clear divisions between the species, particularly for Iris-Setosa. However, distinguishing between Iris-Versicolor and Iris-Virginica remains more challenging, emphasizing the importance of selecting robust features like petal length and width.



Ecole Publique d'ingénieures et d'ingénieurs en 3 ans

6 boulevard Maréchal Juin, CS 45053

14050 CAEN cedex 04

