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

Rapport

LAB WORK 1

le 04 décembre 2024,

version 1

Mohamed Toujani,
Fonction
mohamed.toujani@ecole.ensicaen.fr

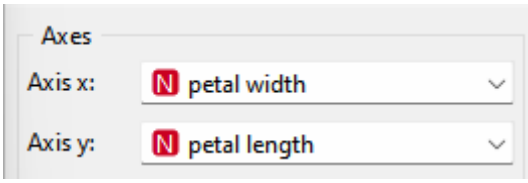
Tuteur école : TSAFACK PIUGIE Armand Florent

ENSI CAEN
ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

PART 2.1

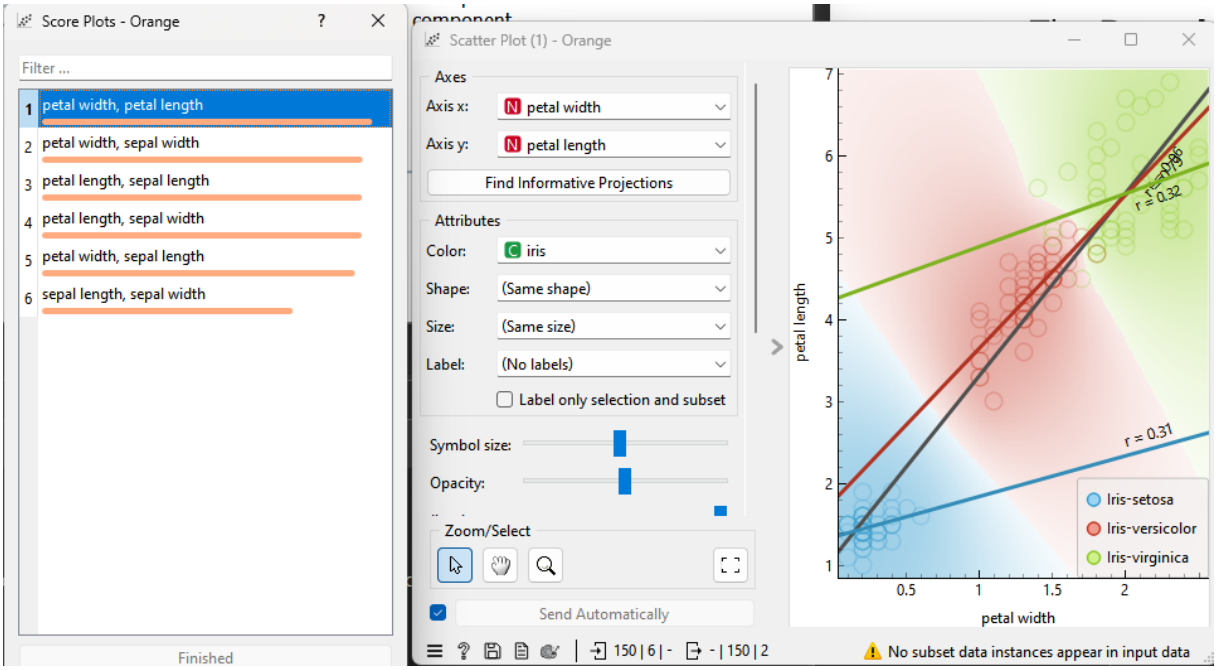
1. Question A

Petal Length vs. Petal Width.



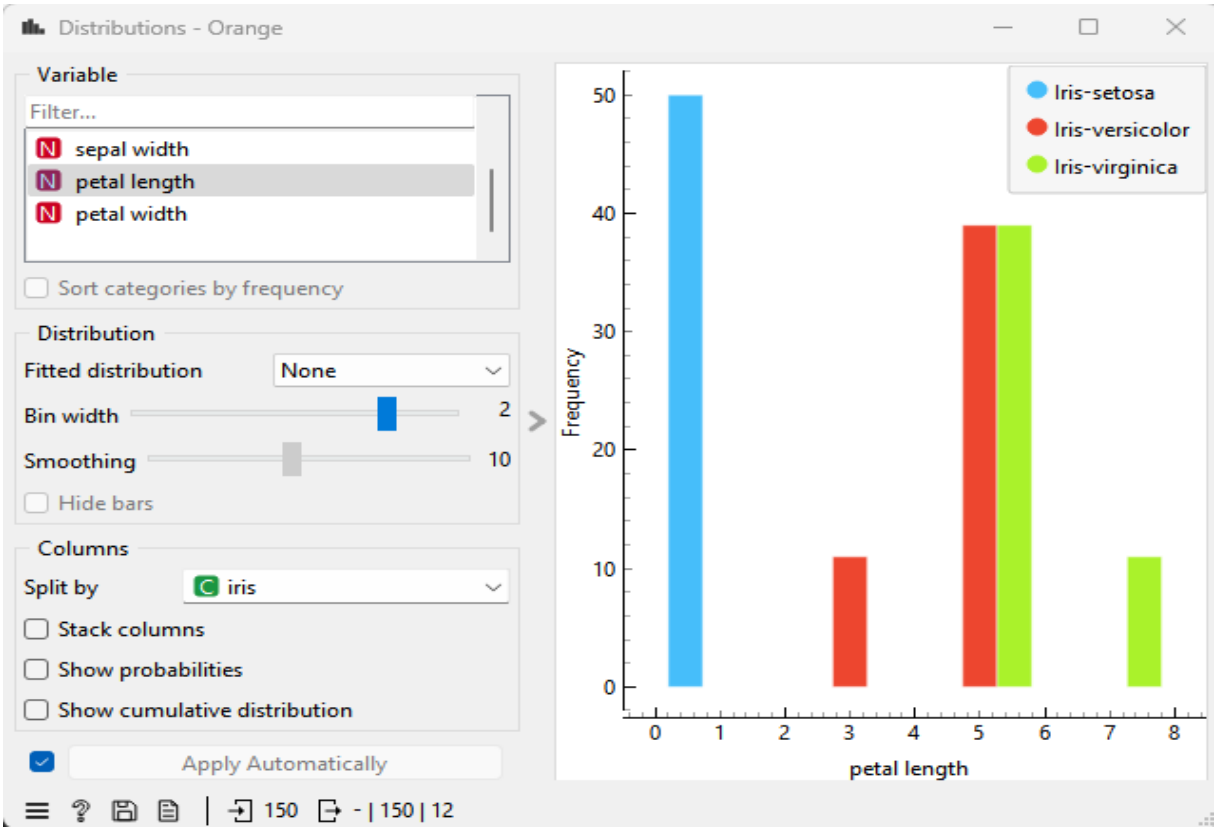
2. Question B

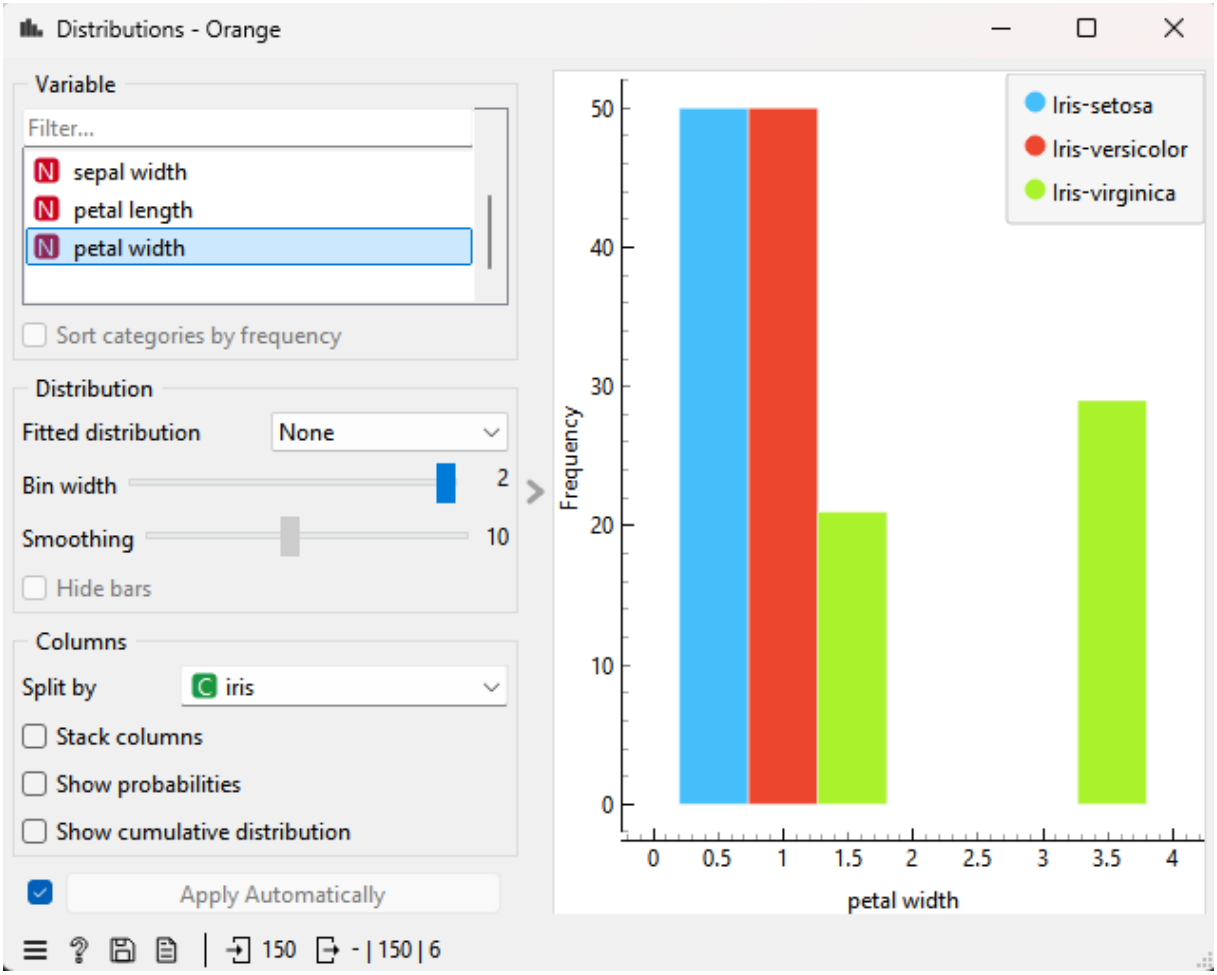
Yes, I get the same answer to the previous question when I call the button “Finnd Informative projections”.



3. Question C

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


Distributions - Orange
Variable
Filter...
sepal width
petal length
petal width
Sort categories by frequency
Distribution
Fitted distribution
None
Bin width
2
Smoothing
10
Hide bars
Columns
Split by
iris
Stack columns
Show probabilities
Show cumulative distribution
Apply Automatically
150
- | 150 | 12
Iris-setosa
Iris-versicolor
Iris-virginica
Frequency
petal length

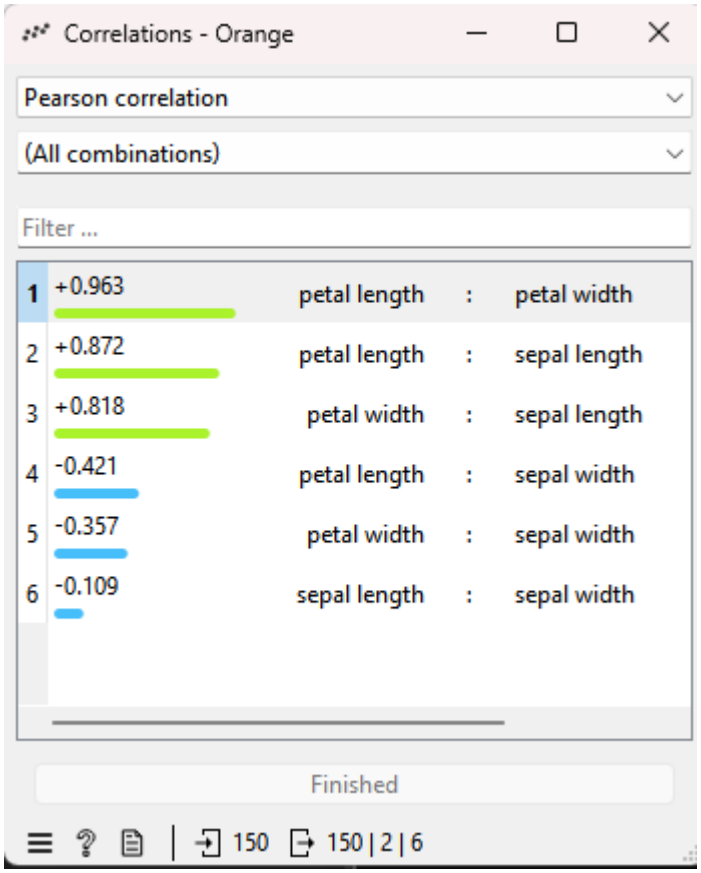

Distributions - Orange
Variable
Filter...
sepal width
petal length
petal width
Sort categories by frequency
Distribution
Fitted distribution
None
Bin width
2
Smoothing
10
Hide bars
Columns
Split by
iris
Stack columns
Show probabilities
Show cumulative distribution
Apply Automatically
150
- | 150 | 6
Iris-setosa
Iris-versicolor
Iris-virginica
Frequency
petal width

4. Question D

Pedal 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

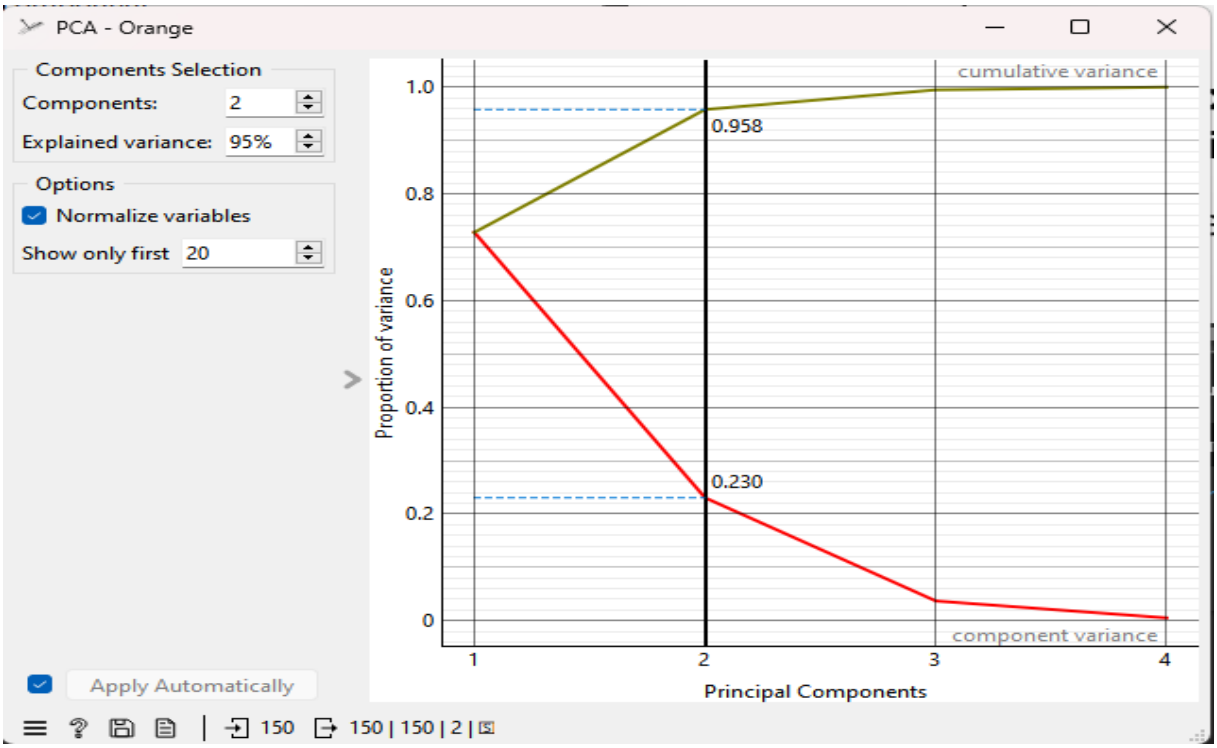
From what we can see from the correlations widget on the sepal width and length, we tend to feel that it is not worth it to use these two variables together as they provide little discrimination options and have very little correlations with each other.



6. Question F

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

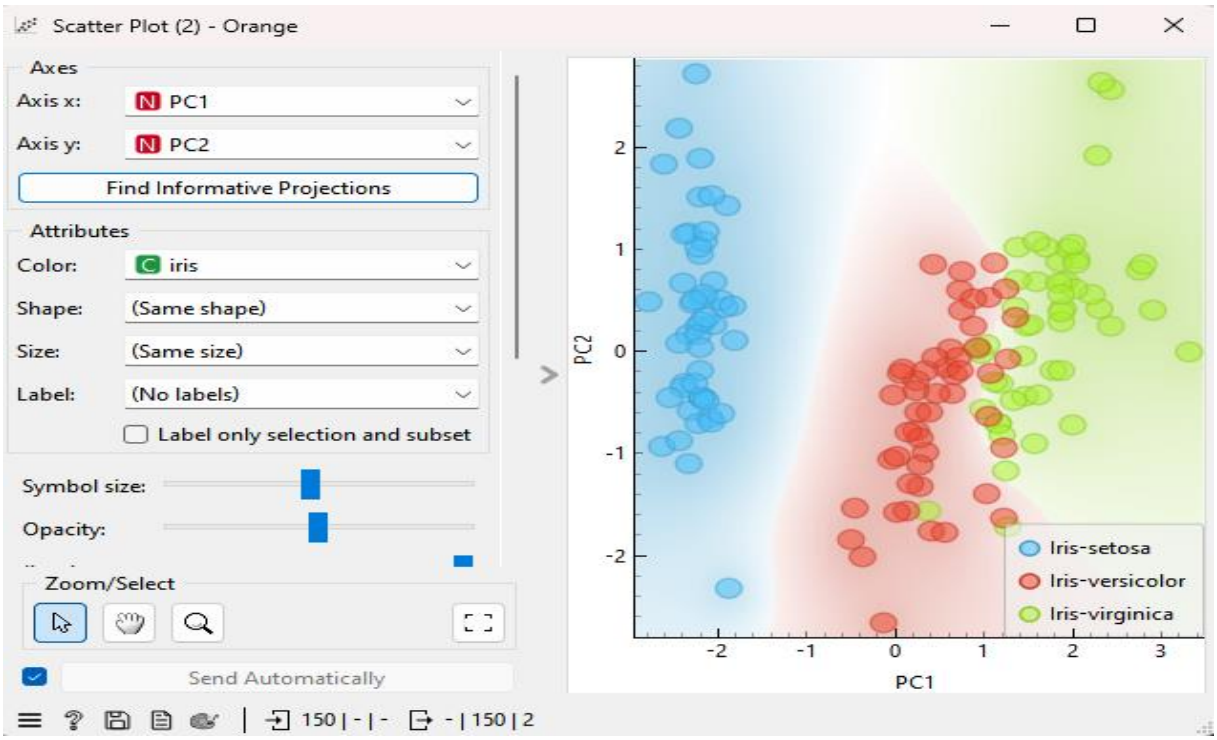
PCA confirms the importance of Petal Width and Petal Length for separating types, as these dominate the principal components.



7. Question G

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

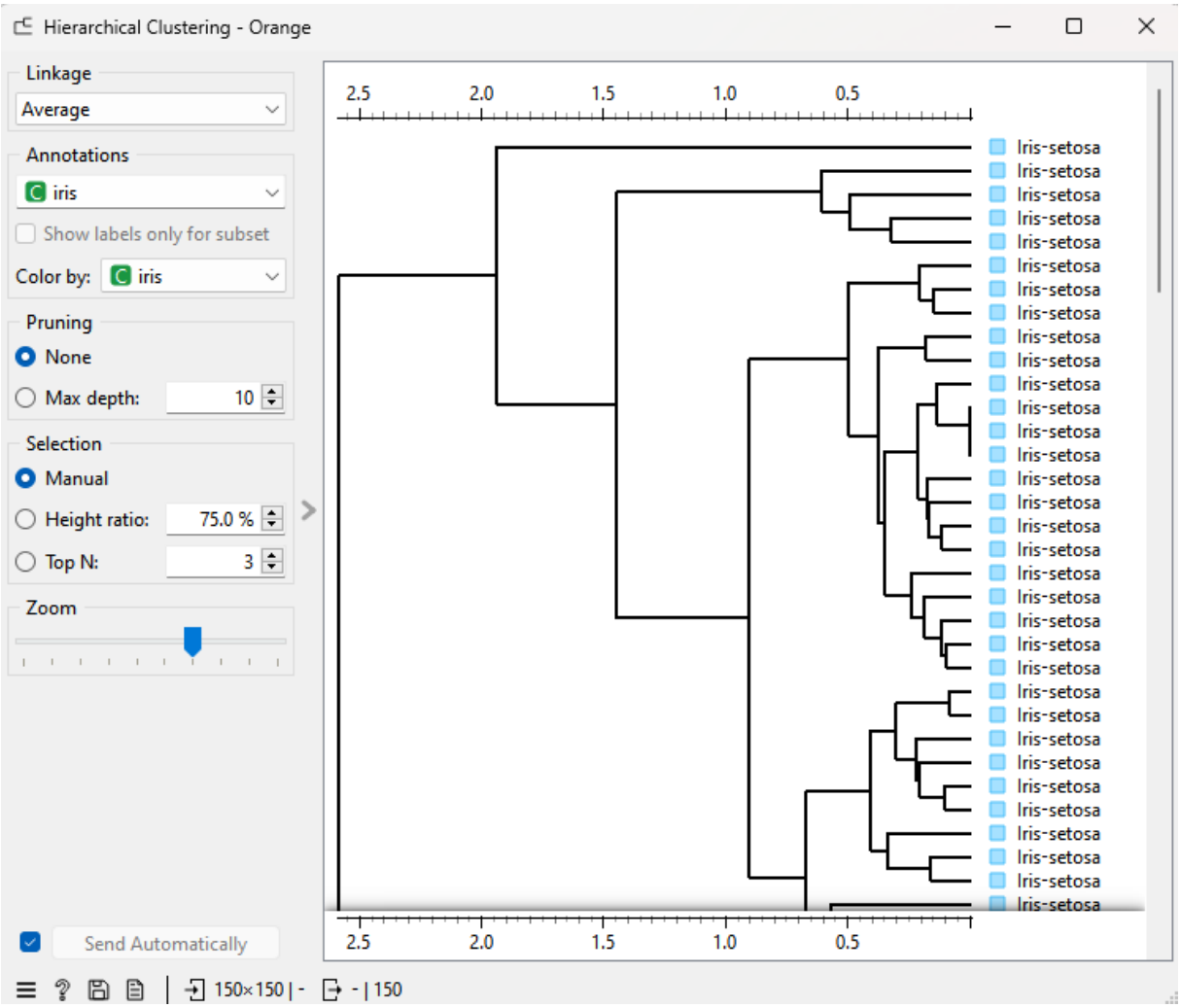
The PCA scatter plot demonstrates the dataset in two dimensions, simplifying the analysis while maintaining most of the information.



PART 2.2

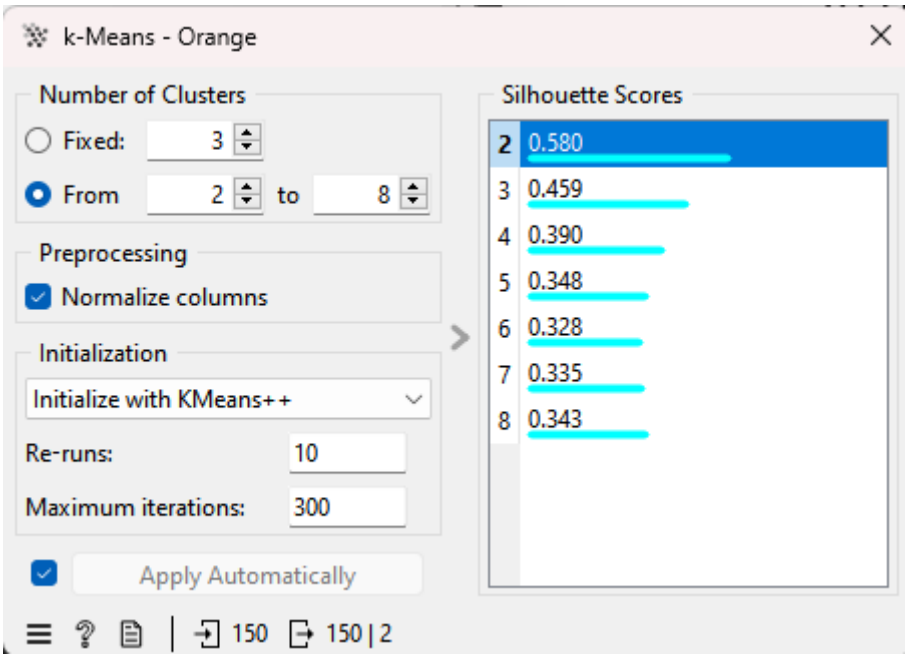
8. Question H

Iris-Setosa seems perfectly classified with no errors.



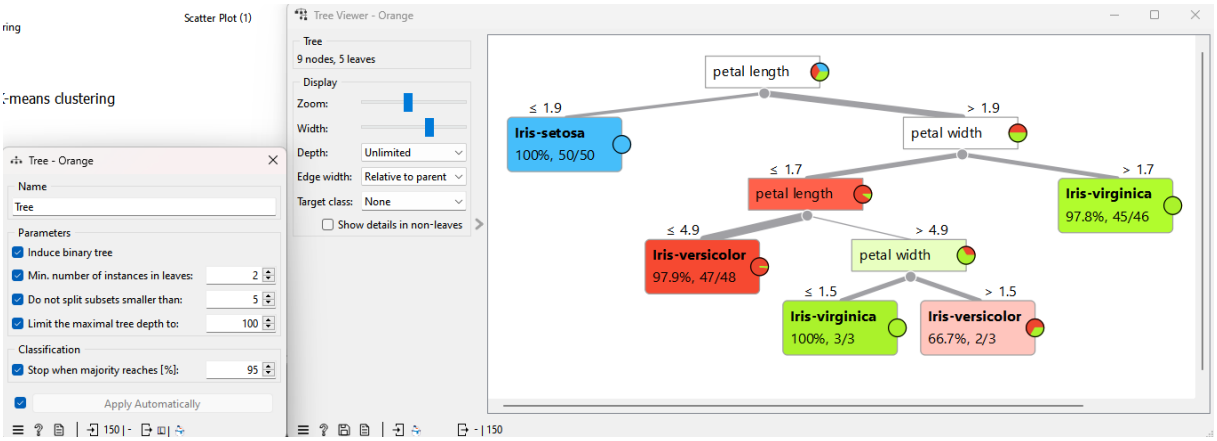
9. Question J

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.

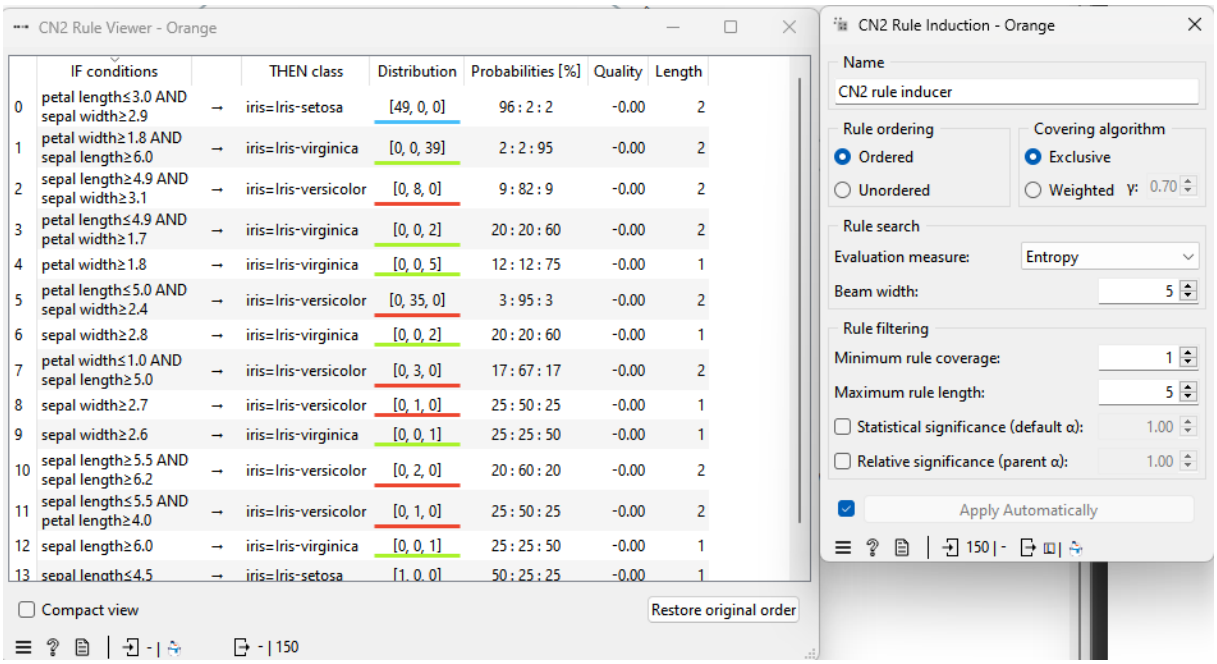


11. Question L

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.



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.

ENSI CAEN

ÉCOLE PUBLIQUE D'INGÉNIEURS
CENTRE DE RECHERCHE

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

6 boulevard Maréchal Juin, CS 45053
14050 CAEN cedex 04