

Rapport

# LAB WORK 2

le 04 décembre 2024,  
version 1

Mohamed Toujani,  
Fonction

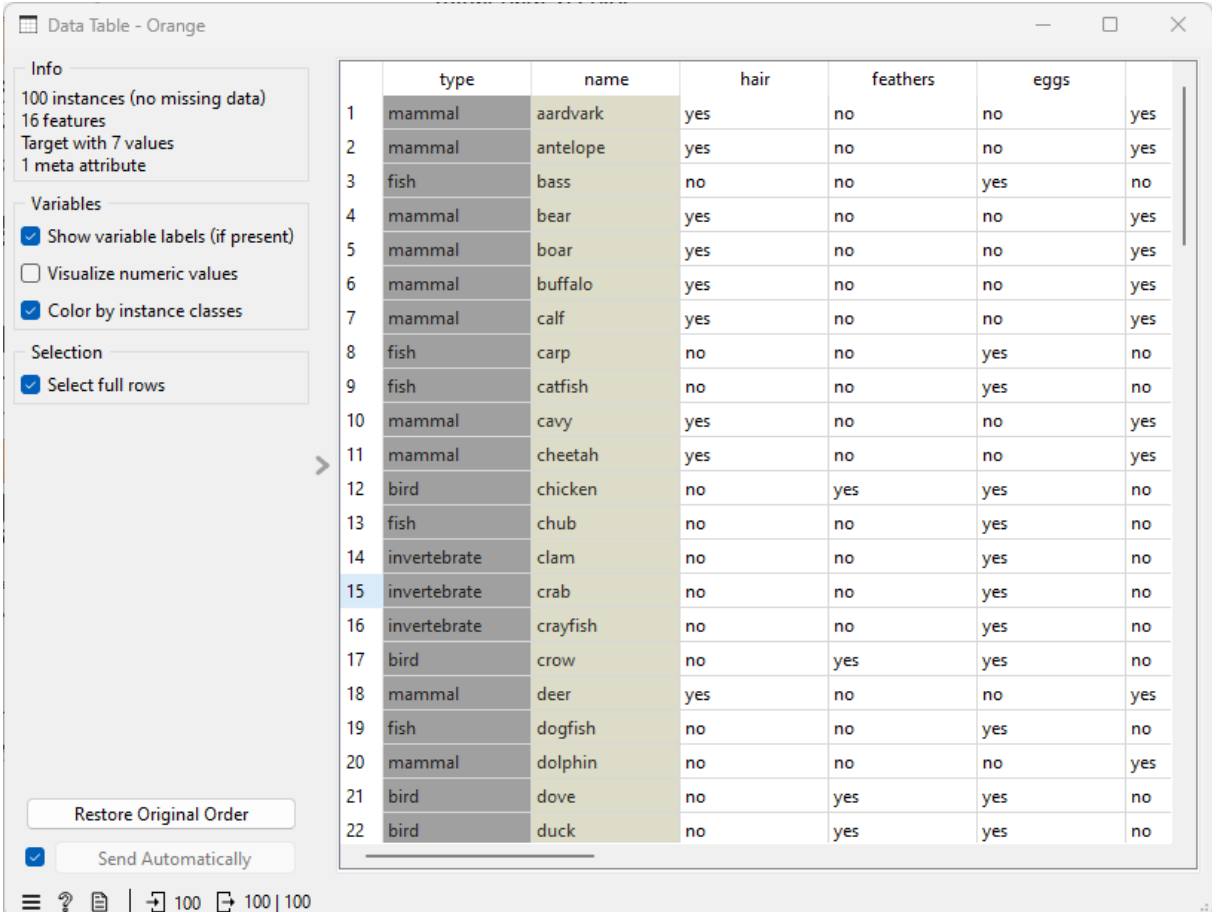
[mohamed.toujani@ecole.ensicaen.fr](mailto:mohamed.toujani@ecole.ensicaen.fr)

Tuteur école : TSAFACK PIUGIE  
Armand Florent



## 1. Question A

- Categorical variables: For attributes like hair, feathers, eggs, etc. (binary values: "yes" or "no").
- Target variable: type with 7 classes (e.g., mammal, bird, fish, etc.).

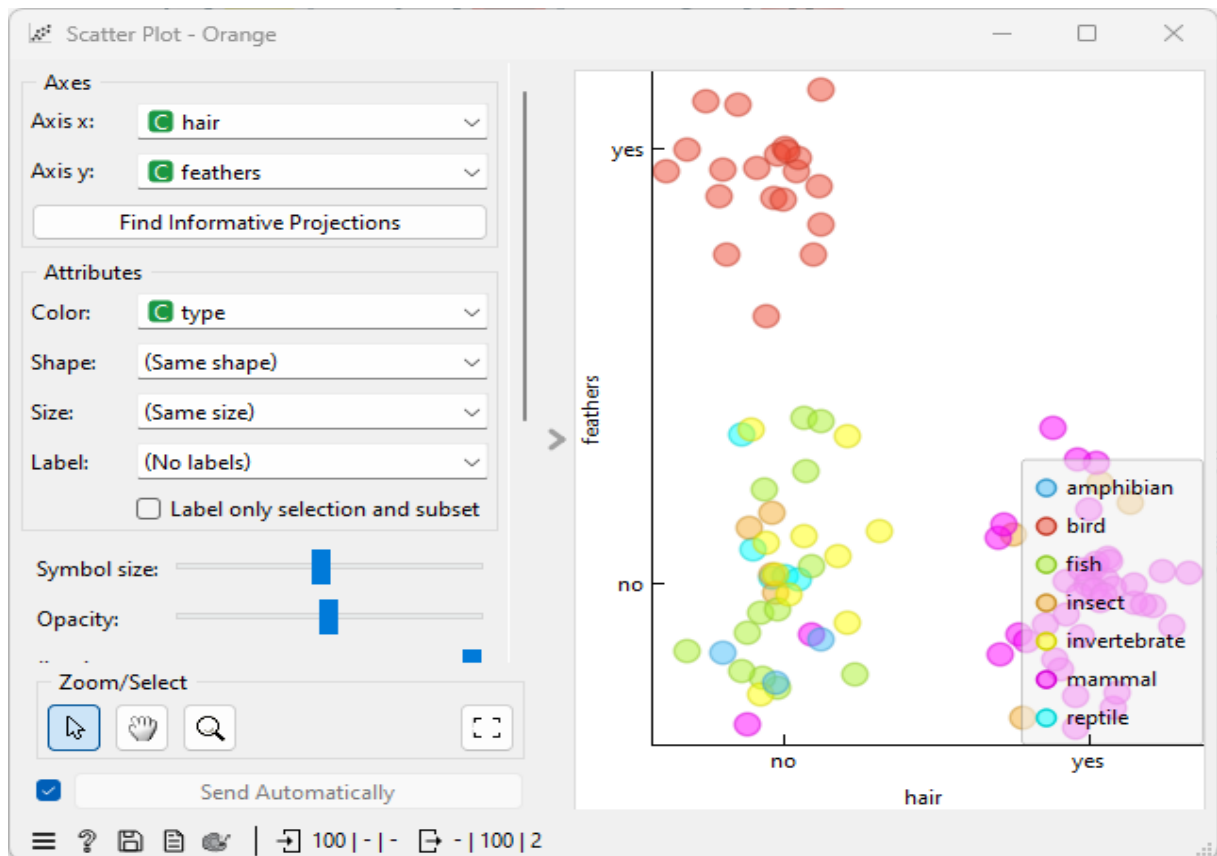


The screenshot shows the 'Data Table - Orange' window. On the left, the 'Info' panel indicates 100 instances, 16 features, a target with 7 values, and 1 meta attribute. The 'Variables' panel has 'Show variable labels (if present)' checked, 'Visualize numeric values' unchecked, and 'Color by instance classes' checked. The 'Selection' panel has 'Select full rows' checked. The main table displays 22 rows of data with columns: type, name, hair, feathers, eggs, and an unlabeled target column. The rows are color-coded by the 'type' variable: mammals (grey), birds (light green), fish (light blue), and invertebrates (yellow). The target column values are 'yes' for mammals and 'no' for birds, fish, and invertebrates.

	type	name	hair	feathers	eggs	
1	mammal	aardvark	yes	no	no	yes
2	mammal	antelope	yes	no	no	yes
3	fish	bass	no	no	yes	no
4	mammal	bear	yes	no	no	yes
5	mammal	boar	yes	no	no	yes
6	mammal	buffalo	yes	no	no	yes
7	mammal	calf	yes	no	no	yes
8	fish	carp	no	no	yes	no
9	fish	catfish	no	no	yes	no
10	mammal	cavy	yes	no	no	yes
11	mammal	cheetah	yes	no	no	yes
12	bird	chicken	no	yes	yes	no
13	fish	chub	no	no	yes	no
14	invertebrate	clam	no	no	yes	no
15	invertebrate	crab	no	no	yes	no
16	invertebrate	crayfish	no	no	yes	no
17	bird	crow	no	yes	yes	no
18	mammal	deer	yes	no	no	yes
19	fish	dogfish	no	no	yes	no
20	mammal	dolphin	no	no	no	yes
21	bird	dove	no	yes	yes	no
22	bird	duck	no	yes	yes	no

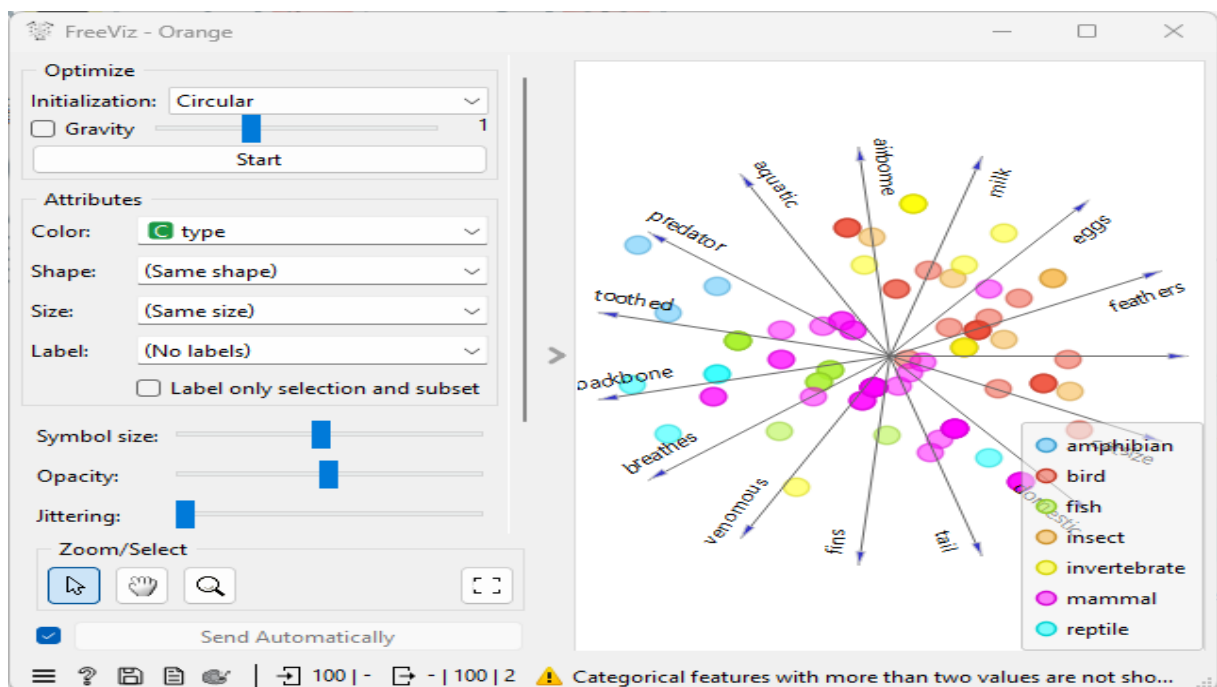
## 2. Question B

Using hair on the x-axis and feathers on the y-axis gives a clear view of the data, as it separates mammals (which have hair) from birds (which have feathers) quite well. It's a simple and effective way to visualize the differences between these groups.



### 3. Question C

The widget shows how features like milk, feathers, and fins separate classes (e.g., mammals, birds, fish). Optimization improves class grouping for clearer patterns.

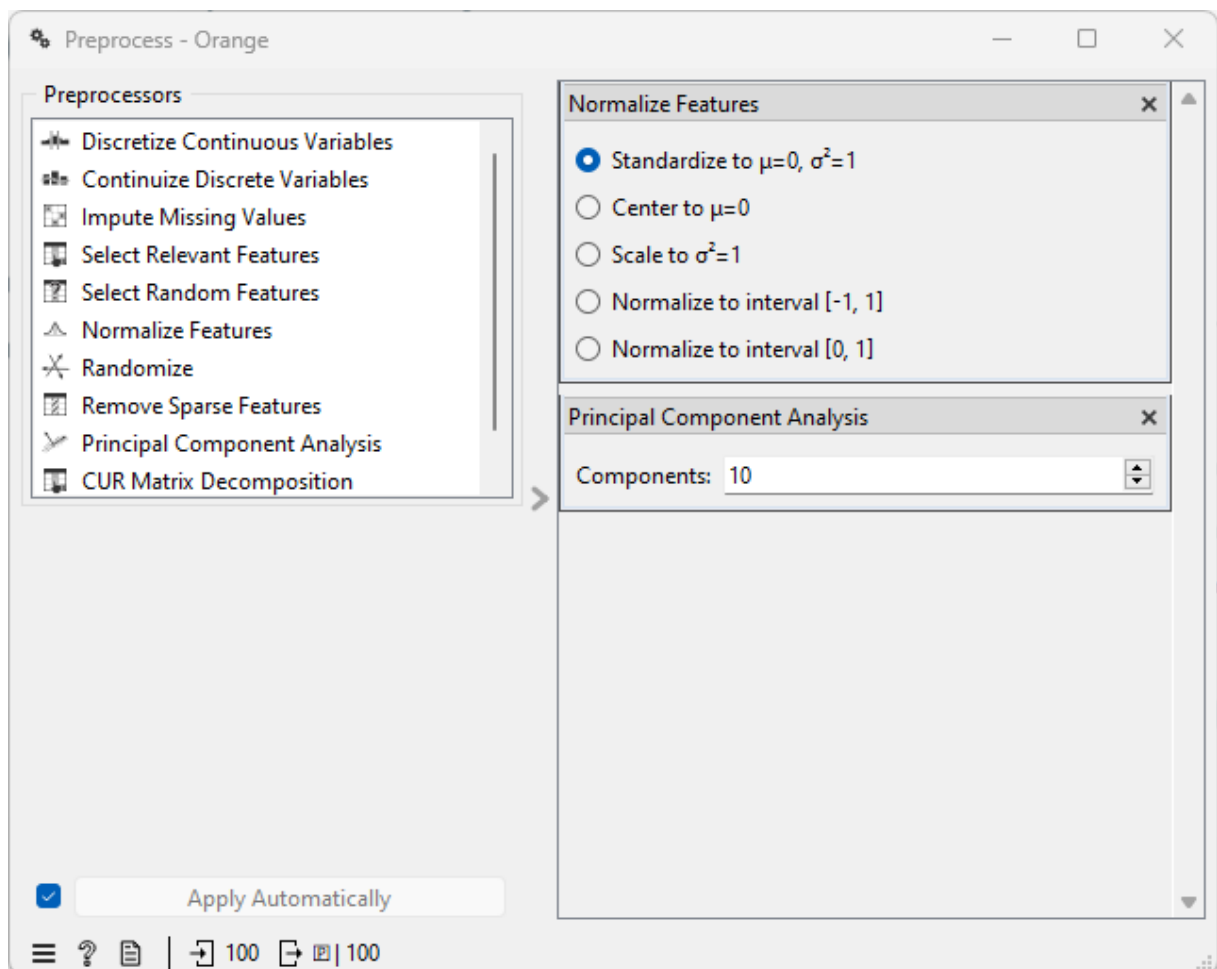


## 4. Question D

The PCA projection helps separate some classes like mammals and birds. However, classes like amphibians and reptiles or fish and invertebrates are harder to distinguish due to overlapping features.

## 5. Question E

- Normalization: Standardized to mean = 0, variance = 1.
- PCA: Reduced to 10 components for dimensionality reduction.



## 6. Question F

The dataset is divided into 10 equal parts. Each part is used once as the test set while the remaining 9 parts are used for training. This process repeats 10 times, and the results are averaged for a reliable evaluation of the model's performance.

Test & Score - Orange

☒ Cross validation

Number of folds: 10

☒ Stratified

☐ Cross validation by feature

☐ Random sampling

Repeat train/test: 10

Training set size: 66 %

☒ Stratified

☐ Leave one out

☐ Test on train data

☐ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
kNN	0.990	0.930	0.921	0.928	0.930	0.909
AdaBoost	0.975	0.960	0.960	0.961	0.960	0.947
Tree	0.945	0.920	0.920	0.922	0.920	0.894
SVM	0.993	0.930	0.924	0.935	0.930	0.907
Neural Network	0.989	0.970	0.968	0.968	0.970	0.960

Compare models by: Area und ☐ Negligible diff.: 0.1

	kNN	AdaB...	Tree	SVM	Neur...
kNN		0.814	0.957	0.355	0.746
AdaBoost	0.186		0.875	0.163	0.197
Tree	0.043	0.125		0.055	0.042
SVM	0.645	0.837	0.945		0.663
Neural Network	0.254	0.803	0.958	0.337	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

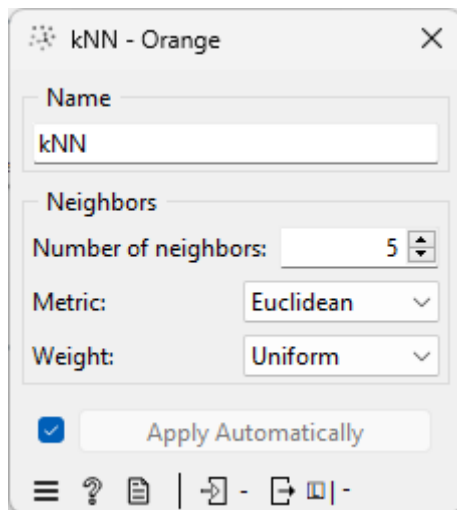
Can't run stratified 10-fold cross validation...

## 7. Question G

The Neural Network achieves the highest classification accuracy (CA = 97.0%), making it the best-performing algorithm in this workflow.

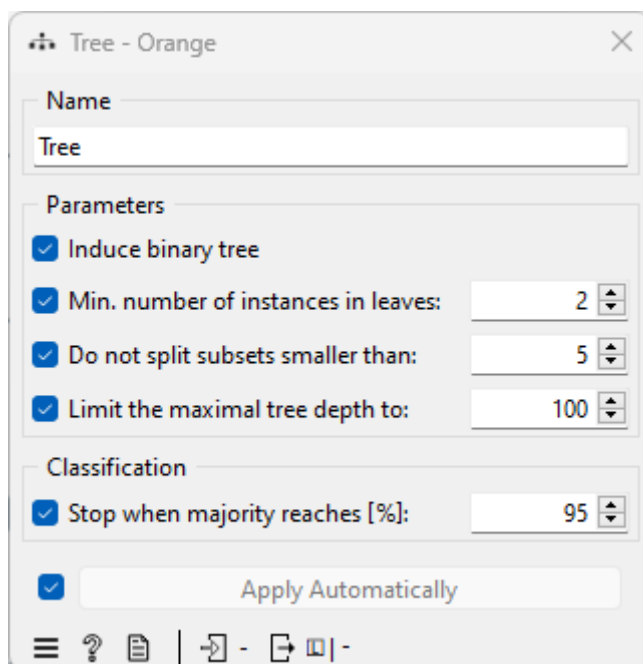
## 8. Question H

Increasing the number of neighbors, changing the metric (e.g., Manhattan or Chebyshev), or adjusting the weight (e.g., distance-based) may improve classification accuracy. Experiment with these settings and observe real-time updates.



## 9. Question I

The classification accuracy typically starts decreasing when the tree depth exceeds a certain point, such as 10 to 15 levels, due to overfitting. Exact results depend on the dataset.



## 10. Question J

The formula is:

Kernel:  $\exp(-g|x-y|^2)$

Yes, the RBF kernel provides excellent results for this dataset, as indicated by its high classification accuracy (SVM: 93.0%).

**SVM - Orange**

Name: SVM

**SVM Type**

☒ SVM Cost (C): 1.00  
 Regression loss epsilon ( $\epsilon$ ): 0.10

☐ v-SVM Regression cost (C): 1.00  
 Complexity bound ( $\nu$ ): 0.50

**Kernel**

☐ Linear Kernel:  $\exp(-g|x-y|^2)$

☐ Polynomial g: auto

☒ RBF

☐ Sigmoid

**Optimization Parameters**

Numerical tolerance: 0.0010

☒ Iteration limit: 100

☐ Apply

## 11. Question K

A decrease in classification accuracy is typically observed when the number of neurons in the hidden layer exceeds 150–200, as it may lead to overfitting.

**Neural Network - Orange**

Name: Neural Network

Neurons in hidden layers: 100

Activation: ReLu

Solver: Adam

Regularization,  $\alpha=0.0001$ : [slider]

Maximal number of iterations: 200

☒ Replicable training

Cancel ☒ Apply Automatically

## 12. Question L

Accuracy is calculated as:

Accuracy = correct predictions / total predictions

For kNN, correct predictions = 3 + 20 + 13 + 8 + 7 + 41 = 92

Total predictions = 100

Accuracy = 92%

Predictions - Orange

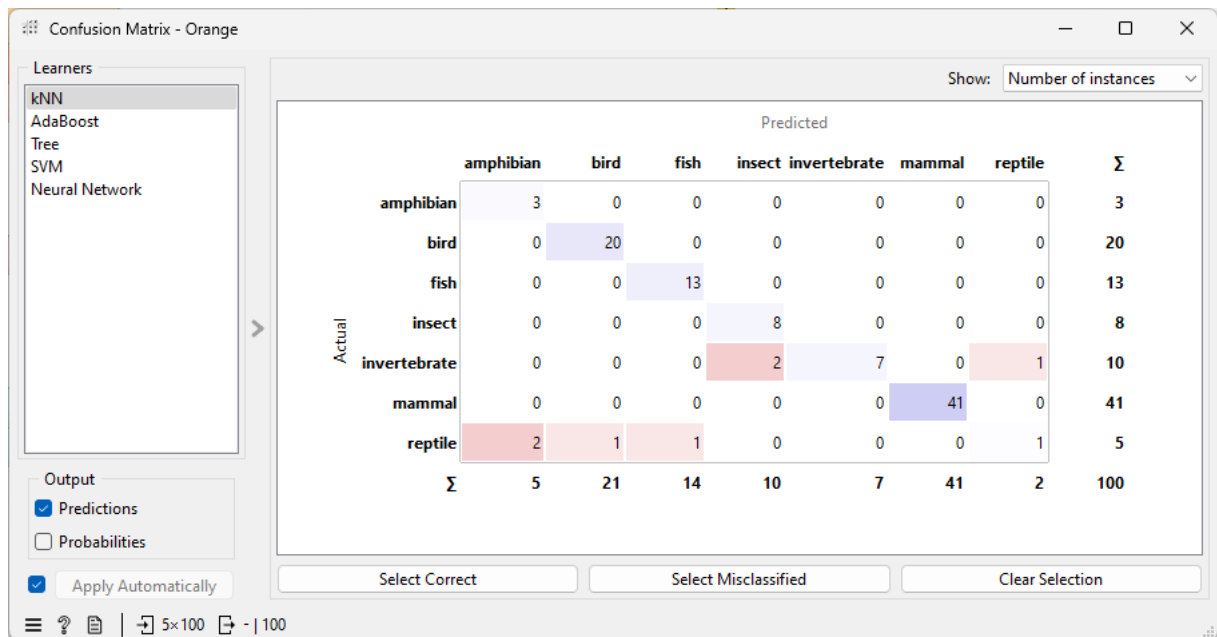
Show probabilities for: Classes in data ☒ Show classification errors [Restore Original Order](#)

	type	name	kNN	AdaBoost	Tree	
	mammal	aardvark	mammal	mammal	mammal	mamm
	mammal	cheetah	mammal	mammal	mammal	mamm
	mammal	elephant	mammal	mammal	mammal	mamm
	insect	housefly	insect	insect	insect	insect
	mammal	lion	mammal	mammal	mammal	mamm
	invertebrate	lobster	invertebrate	invertebrate	invertebrate	inverte
	mammal	opossum	mammal	mammal	mammal	mamm
	bird	rhea	bird	bird	bird	bird
	invertebrate	slug	insect	invertebrate	invertebrate	inverte
	mammal	squirrel	mammal	mammal	mammal	mamm
	mammal	boar	mammal	mammal	mammal	mamm
	fish	chub	fish	fish	fish	fish
	fish	dogfish	fish	fish	fish	fish
	mammal	goat	mammal	mammal	mammal	mamm
	mammal	gorilla	mammal	mammal	mammal	mamm
	fish	haddock	fish	fish	fish	fish

☒ Show performance scores Target class: (Average over classes)

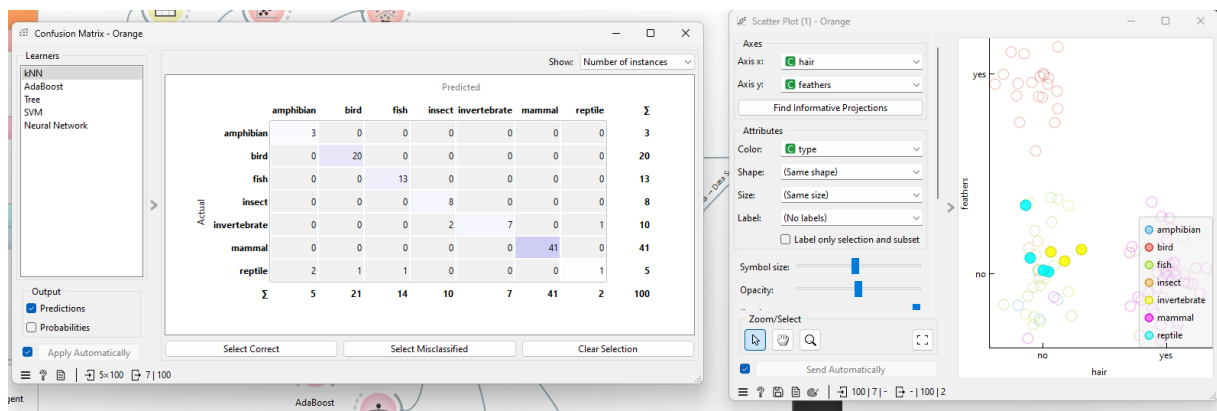
100 | - 100 | 100 | -





### 13. Question M

Misclassified data points are near overlaps between classes, like mammals and invertebrates, showing where KNN struggles with similar features.



### Conclusion

Through this lab, I explored various data visualization and machine learning techniques using Orange. Key takeaways include understanding how features like hair or feathers help separate classes, optimizing algorithms like kNN and neural networks for better accuracy, and recognizing challenges like overlapping classes. This practical work highlighted the strengths and limitations of different methods in analyzing and classifying data effectively.



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

6 boulevard Maréchal Juin, CS 45053

14050 CAEN cedex 04

