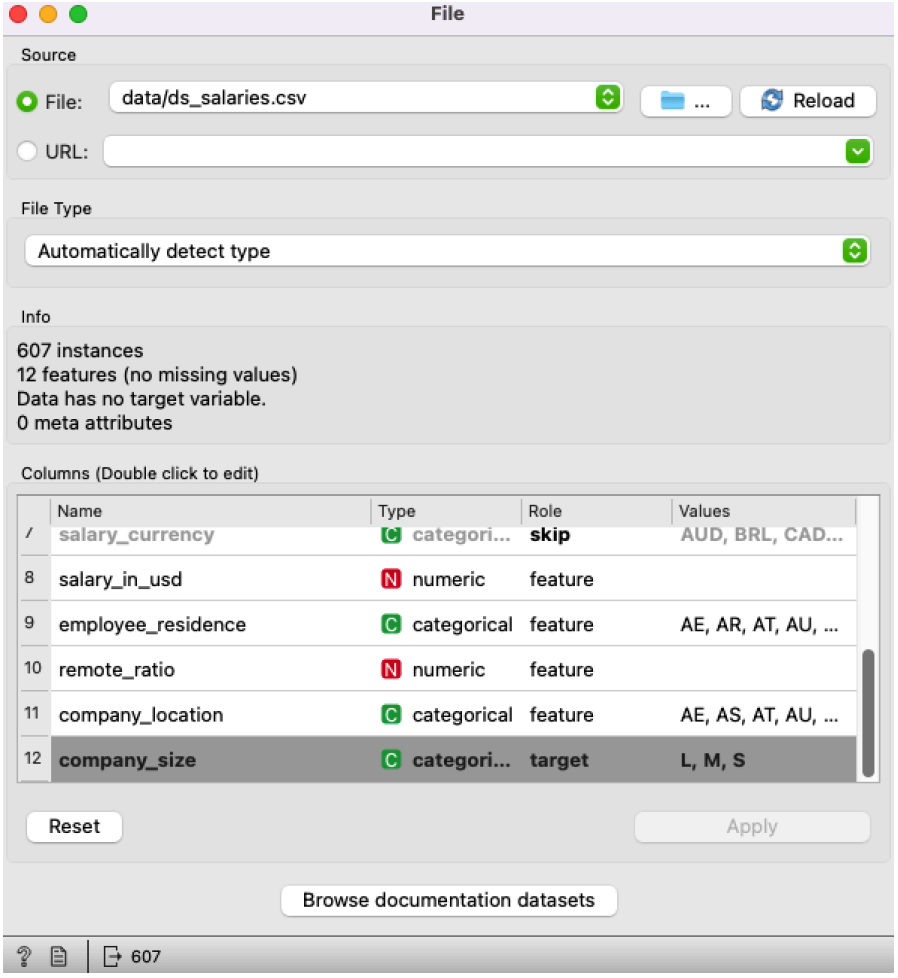
Preparation

Dataset:
ruchi798/data-science-job-salaries

File parameters

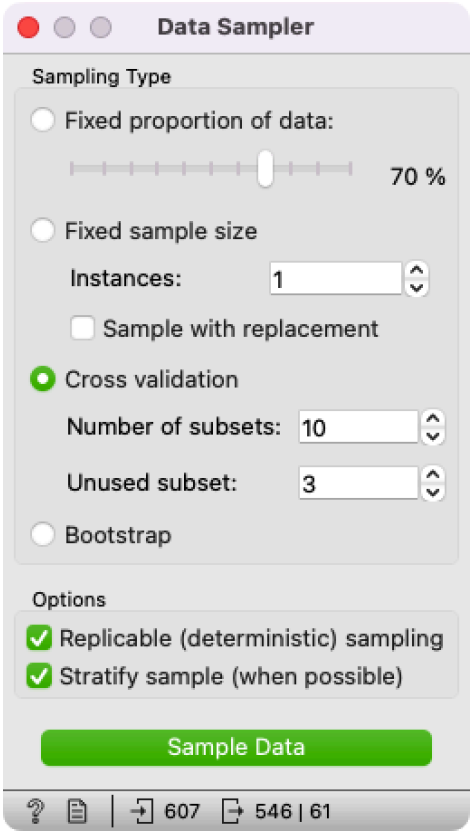
- Skip 2 parameters
- 1 for the target
- 9 for the features



data sampler

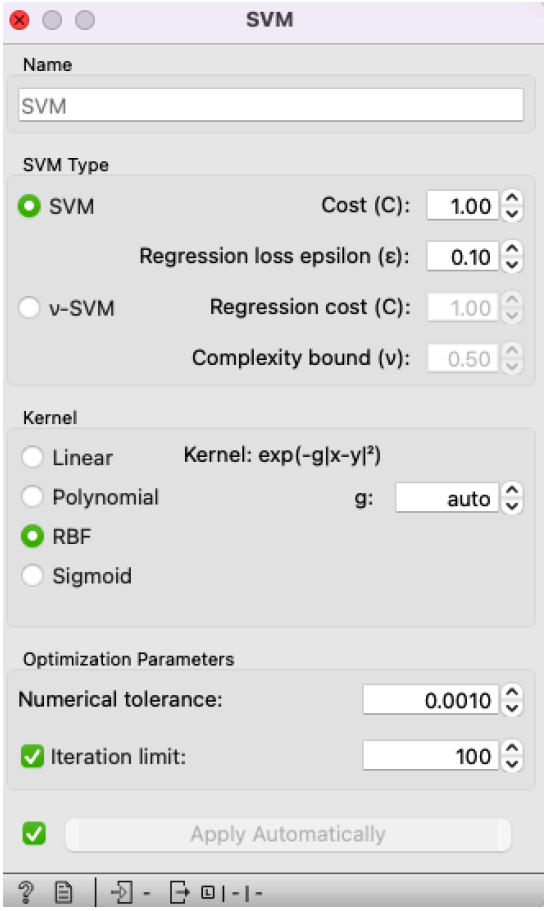
The original paper states:
“The models are tested and trained using ten-fold cross-validation. 30% of the data is used to test the learned model, and 70% is used to train the model, following the standard 7:3 dataset split.”

Thus, the settings are as follows:

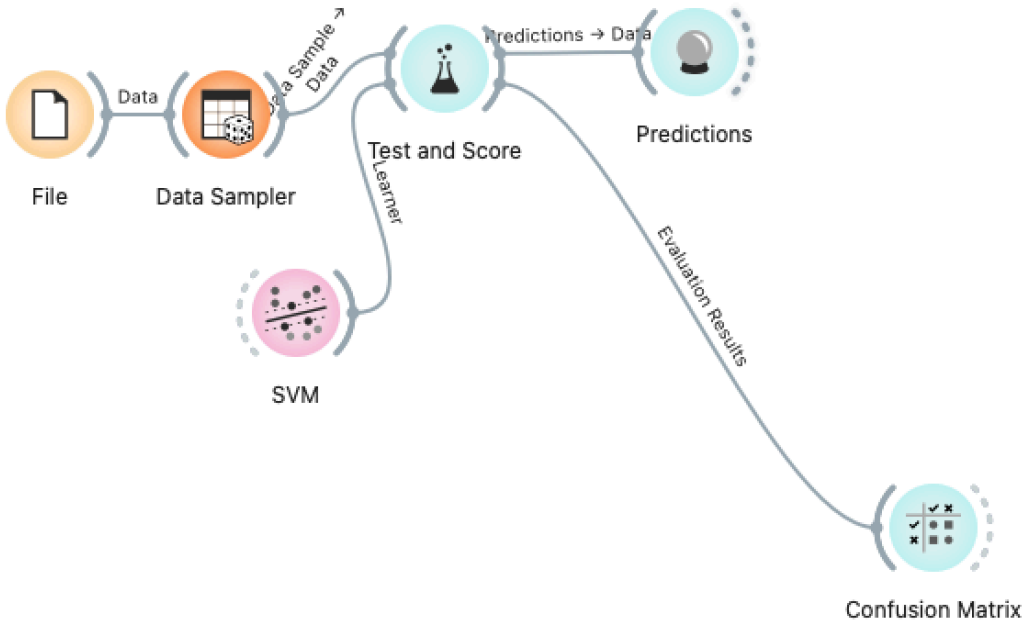


SVM setting

Use the default settings for both sets (meaning the results will remain the same).


SVM
Name
SVM
SVM Type
SVM
Cost (C):
1.00
Regression loss epsilon (ε):
0.10
ν-SVM
Regression cost (C):
1.00
Complexity bound (ν):
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 Automatically

Process



Result

Test and Score

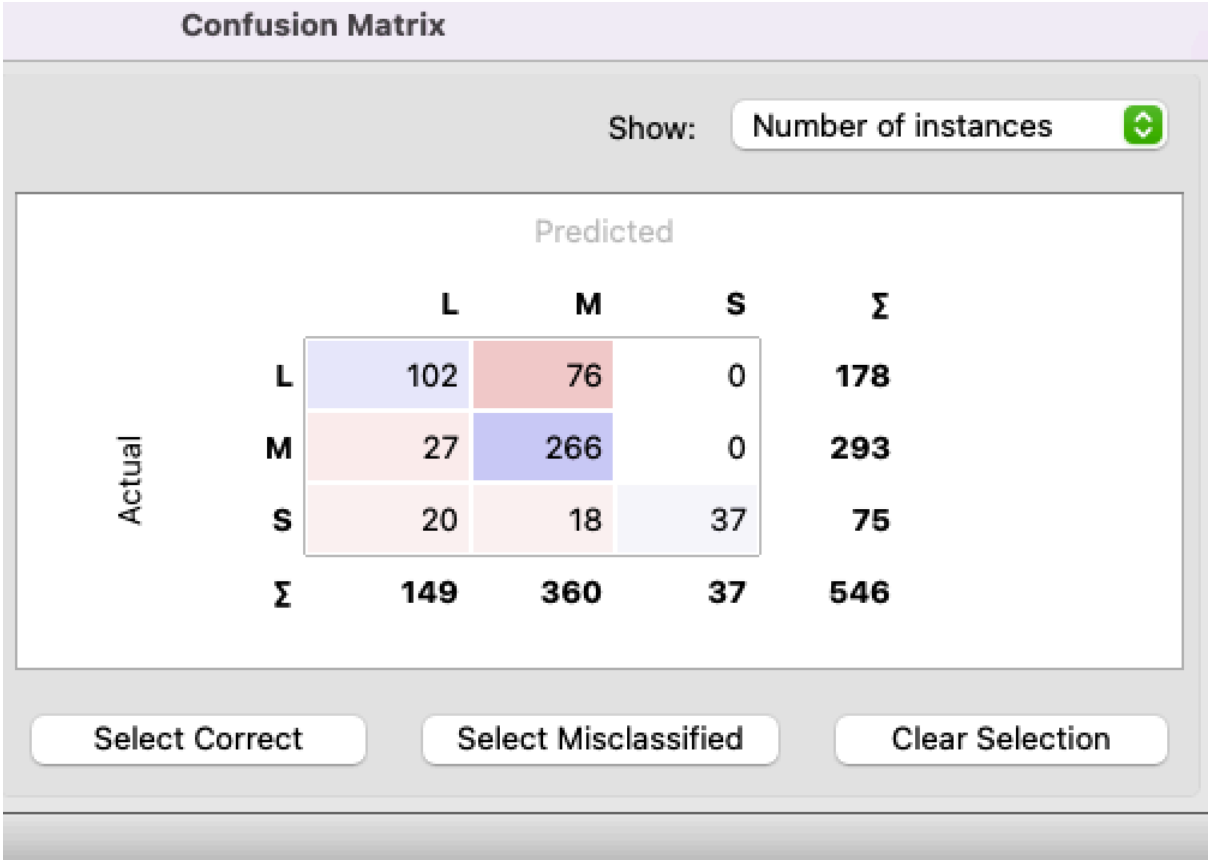
Test and Score

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

Model	AUC	CA	F1	Prec	Recall	MCC
SVM	0.865	0.742	0.731	0.757	0.742	0.542

confusion matrix

compare the predicted values and the actual values.



prediction errors

Prediction error is used to prove that the model and the way companies are put into groups is accurate.

Algorithm	Class		
	L	M	S
SVM	76	27	57

Conclusion

Support Vector Machine (SVM) and Company Size Prediction:

1. SVM Algorithm: SVM is a powerful machine learning algorithm used for classification and regression tasks. In this case, it's being used for multi-class classification to predict company size.
2. Application to Company Size Prediction:

- The model uses 9 parameters (features) to predict whether a company is Large (L), Medium (M), or Small (S).
- SVM works by finding the optimal hyperplane that best separates these three classes in a 9-dimensional space (one dimension for each parameter).

3. Performance Analysis:

- The model shows decent overall performance with an accuracy of 74.2%.
- It's particularly effective at identifying medium-sized companies (266 out of 293 correct).
- There's some confusion between Large and Medium companies, which might indicate similarity in some features for these categories.

4. SVM Strengths in This Context:

- Handling High-Dimensional Data: SVM can effectively handle the 9 parameters used for prediction.
- Non-linear Classification: If a kernel function is used, SVM can capture complex relationships between company features and size.
- Robustness: SVM is less prone to overfitting, especially with limited data.

In summary, SVM's ability to handle complex, multi-dimensional data makes it suitable for this task of predicting company size based on multiple parameters. The results suggest it's a viable approach, though there's room for refinement, particularly in distinguishing between Large and Medium companies and in identifying Small companies more accurately.