

Preparation

Dataset:

ruchi798/data-science-job-salaries

File parameters

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

The screenshot shows a file selection window titled "File". It has two main sections: "Source" and "File Type".

Source: The "File:" option is selected, with the path "data/ds_salaries.csv" entered. There are buttons for "..." (file explorer) and "Reload". The "URL:" option is also available but not selected.

File Type: The dropdown menu is set to "Automatically detect type".

Info: This section provides summary statistics: "607 instances", "12 features (no missing values)", "Data has no target variable.", and "0 meta attributes".

Columns (Double click to edit): A table lists the columns with their names, types, roles, and sample values.

| | Name | Type | Role | Values |
|----|--------------------|-------------|---------|---------------------|
| / | salary_currency | categorical | skip | AUD, BRL, CAD... |
| 8 | salary_in_usd | numeric | feature | |
| 9 | employee_residence | categorical | feature | AE, AR, AT, AU, ... |
| 10 | remote_ratio | numeric | feature | |
| 11 | company_location | categorical | feature | AE, AS, AT, AU, ... |
| 12 | company_size | categorical | target | L, M, S |

At the bottom, there are "Reset" and "Apply" buttons, and a "Browse documentation datasets" button.

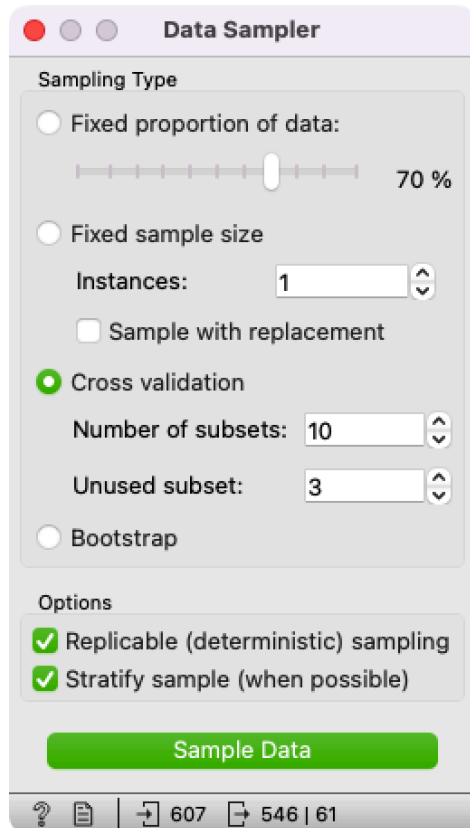
The bottom status bar shows a question mark icon, a document icon, and the text "607".

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:



The screenshot shows a window titled "Data Sampler" with a light purple header. Below the header, there are two main sections: "Sampling Type" and "Options".

Sampling Type

- ☐ Fixed proportion of data: A slider is positioned at 70 %.
- ☐ Fixed sample size: The "Instances:" field is set to 1. Below it, there is an unchecked checkbox for "Sample with replacement".
- ☒ Cross validation: The "Number of subsets:" field is set to 10, and the "Unused subset:" field is set to 3.
- ☐ Bootstrap

Options

- ☒ Replicable (deterministic) sampling
- ☒ Stratify sample (when possible)

At the bottom of the "Options" section is a green button labeled "Sample Data".

The bottom status bar of the window shows a question mark icon, a document icon, and the text "607 546 | 61".

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

☐ v-SVM

Regression cost (C): 1.00

Complexity bound (v): 0.50

Kernel

☐ Linear

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

☐ Polynomial

g: auto

☒ RBF

☐ Sigmoid

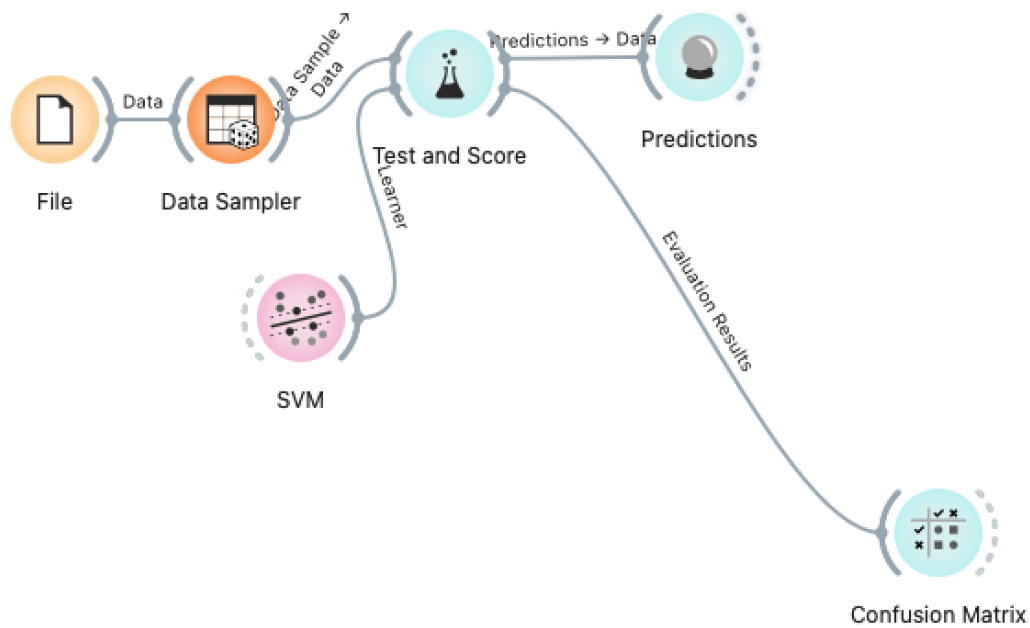
Optimization Parameters

Numerical tolerance: 0.0010

☒ Iteration limit: 100

☒ Apply Automatically

Process

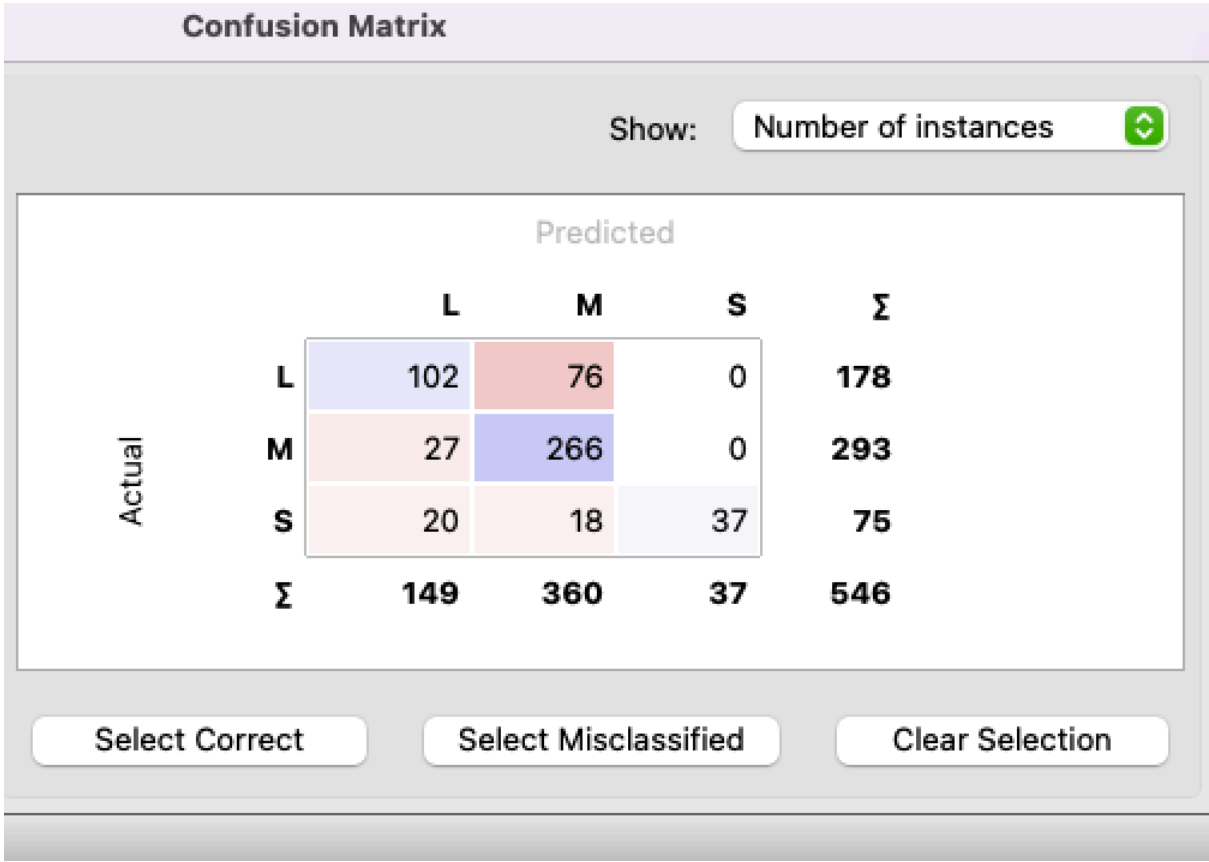


Result

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.