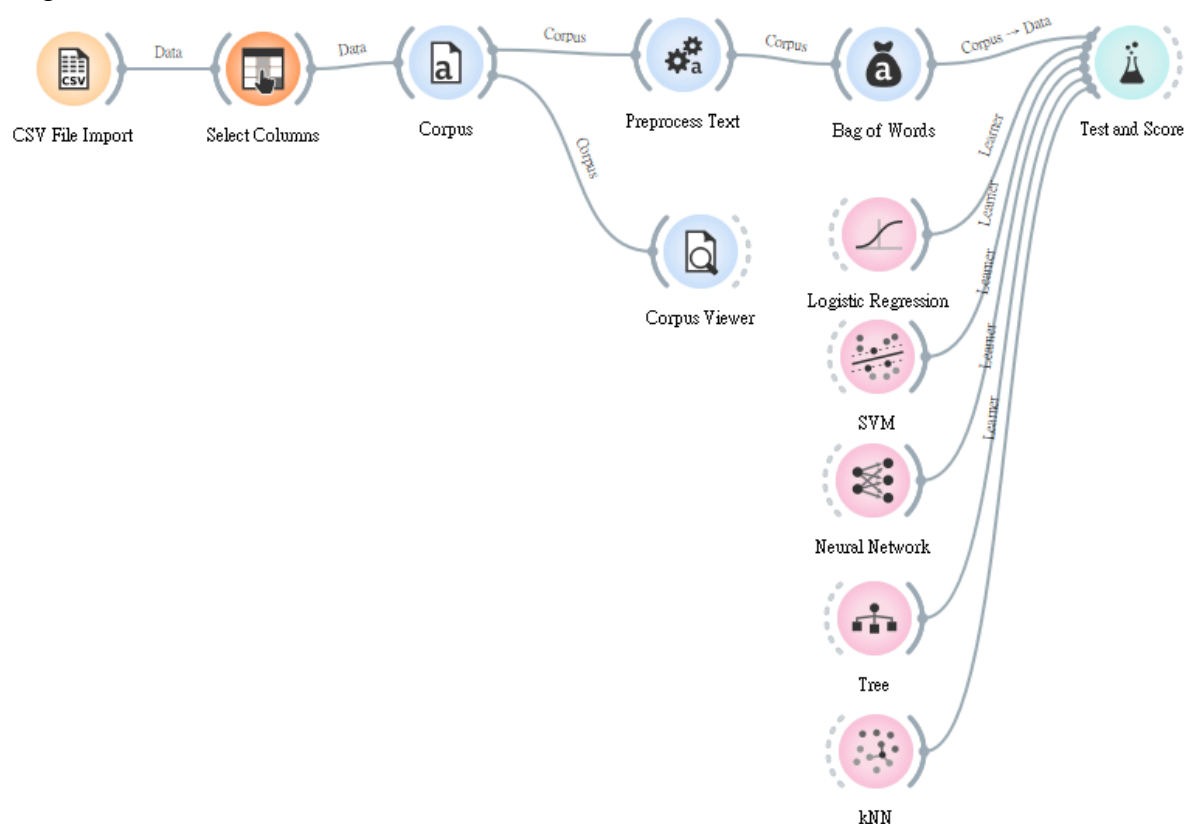
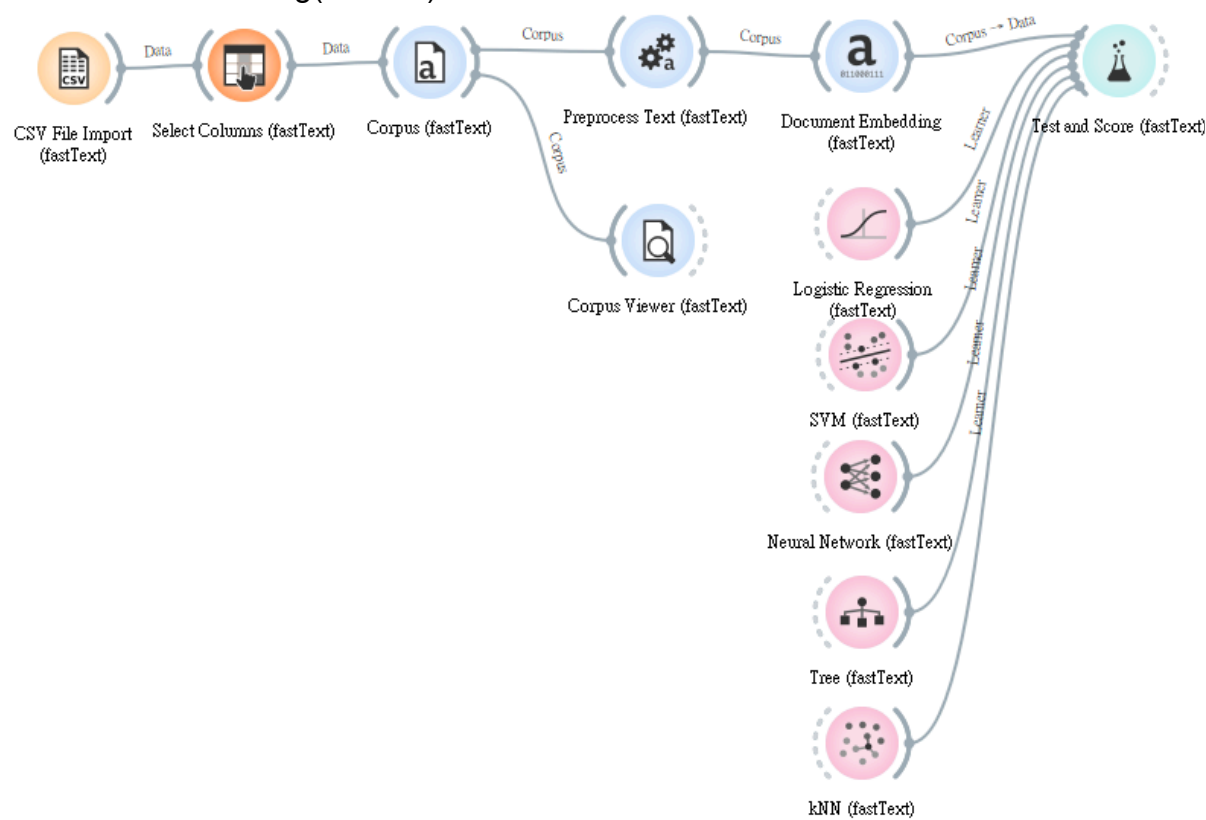


1. Workflow

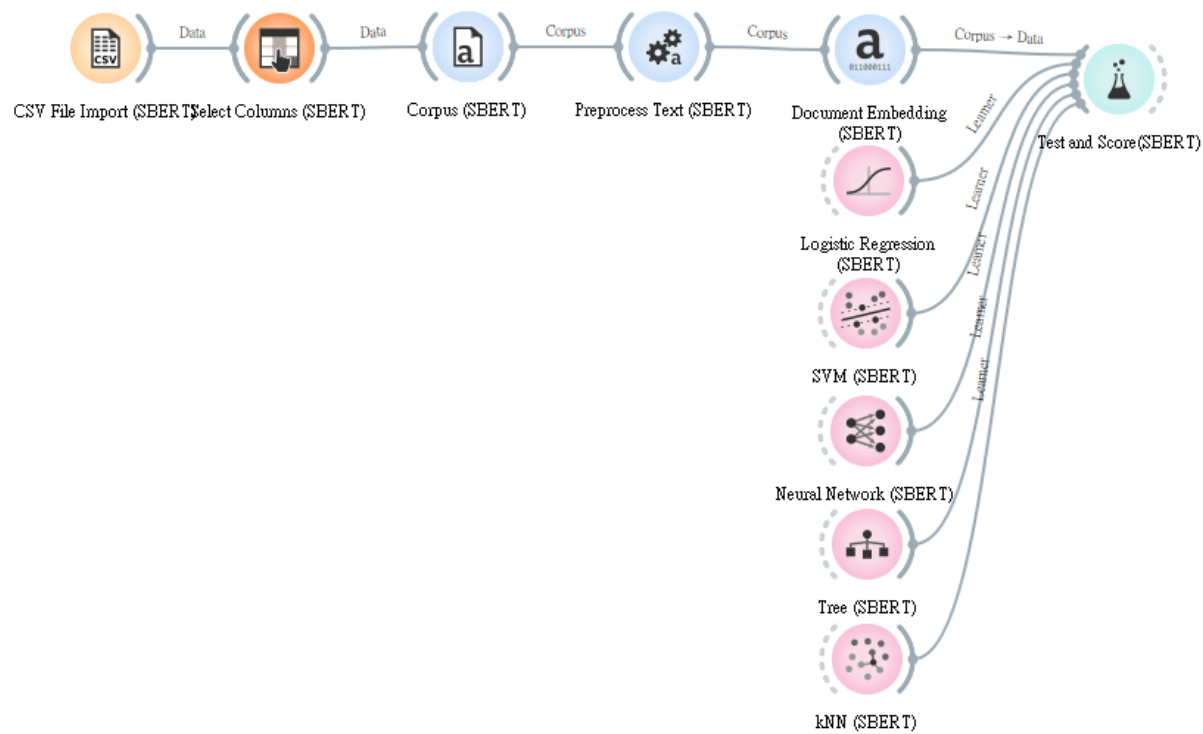
a. Bag of Words



b. Document Embedding(fastText)



c. Document Embedding(Multilingual SBERT)



2. Test and Score

a. Bag of Words

Test and Score - Orange

File Edit View Window Help

☒ Cross validation

Number of folds: 5

☒ 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
Logistic Regression	0.991	0.982	0.982	0.983	0.982	0.923
SVM	0.966	0.743	0.783	0.903	0.743	0.469
Neural Network	0.982	0.985	0.985	0.985	0.985	0.935
Tree	0.813	0.949	0.945	0.949	0.949	0.763
kNN	0.836	0.913	0.895	0.921	0.913	0.567

Compare models by: Area under ROC curve

☐ Negligible diff.: 0.1

	Logistic Re...	SVM	Neural Net...	Tree	kNN
Logistic Regression		0.978	0.961	1.000	1.000
SVM	0.022		0.078	1.000	0.999
Neural Network	0.039	0.922		1.000	1.000
Tree	0.000	0.000	0.000		0.173
kNN	0.000	0.001	0.000	0.827	

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.

5572 | 5572 | 5572

b. Document Embedding(fastText)

Test and Score (fastText) - Orange

File Edit View Window Help

☒ Cross validation
 Number of folds: 5
☒ 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
Logistic Regression (fastText)	0.984	0.959	0.957	0.957	0.959	0.813
SVM (fastText)	0.992	0.980	0.980	0.981	0.980	0.917
Neural Network (fastText)	0.993	0.985	0.985	0.985	0.985	0.934
Tree (fastText)	0.833	0.938	0.936	0.936	0.938	0.720
kNN (fastText)	0.977	0.960	0.961	0.964	0.960	0.842

Compare models by: Area under ROC curve ☐ Negligible diff.: 0.1

	Logistic ...	SVM (fas...	Neural N...	Tree (fast...	kNN (fas...
Logistic Regression (fastText)		0.021	0.007	1.000	0.914
SVM (fastText)	0.979		0.541	1.000	0.988
Neural Network (fastText)	0.993	0.459		1.000	0.983
Tree (fastText)	0.000	0.000	0.000		0.000
kNN (fastText)	0.086	0.012	0.017	1.000	

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.

5572 | 5572 | 5572

c. Document Embedding (Multilingual SBERT)

Test and Score(SBERT) - Orange

File Edit View Window Help

☒ Cross validation
 Number of folds: 5
☒ 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
Logistic Regression (SBERT)	0.996	0.987	0.987	0.987	0.987	0.944
SVM (SBERT)	0.997	0.990	0.990	0.990	0.990	0.957
Neural Network (SBERT)	0.995	0.990	0.990	0.990	0.990	0.957
Tree (SBERT)	0.815	0.935	0.932	0.932	0.935	0.703
kNN (SBERT)	0.981	0.975	0.975	0.975	0.975	0.892

Compare models by: Area under ROC curve ☐ Negligible diff.: 0.1

	Logistic R...	SVM (SB...	Neural N...	Tree (SBE...	kNN (SBE...
Logistic Regression (SBERT)		0.133	0.644	1.000	0.989
SVM (SBERT)	0.867		0.795	1.000	0.992
Neural Network (SBERT)	0.356	0.205		0.999	0.976
Tree (SBERT)	0.000	0.000	0.001		0.000
kNN (SBERT)	0.011	0.008	0.024	1.000	

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.

5572 | 5572 | 5572

3. Parameters

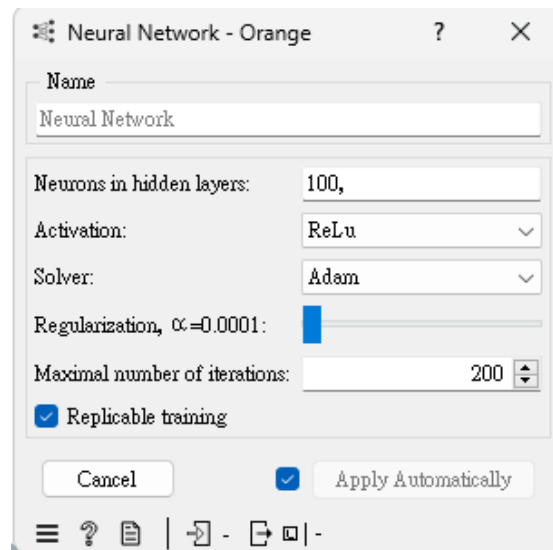
a. Logistic Regression

The screenshot shows the 'Logistic Regre...' window. The 'Name' field contains 'Logistic Regression'. The 'Regularization type' is set to 'Ridge (L2)'. The 'Strength' is represented by a slider between 'Weak' and 'Strong', with a blue bar at the center and the text 'C=1' below it. There is an unchecked checkbox for 'Balance class distribution'. At the bottom, there is a checked checkbox and the text 'Apply Automatically'. The window has a standard toolbar with icons for help, save, undo, redo, and zoom.

b. SVM

The screenshot shows the 'SVM - Orange' window. The 'Name' field contains 'SVM'. Under 'SVM Type', the 'SVM' radio button is selected, with 'Cost (C):' set to '1.00'. The 'Regression loss epsilon (ε):' is set to '0.10'. The 'ν-SVM' radio button is unselected, with 'Regression cost (C):' set to '1.00' and 'Complexity bound (ν):' set to '0.50'. Under 'Kernel', the 'RBF' radio button is selected, with the kernel formula $\text{Kernel: } \exp(-\gamma \|x - y\|^2)$ and the parameter γ set to 'auto'. The 'Linear' kernel is unselected, 'Polynomial' is unselected, and 'Sigmoid' is unselected. Under 'Optimization Parameters', 'Numerical tolerance:' is set to '0.0010' and 'Iteration limit:' is checked and set to '100'. At the bottom, there is a checked checkbox and the text 'Apply Automatically'. The window has a standard toolbar with icons for help, save, undo, redo, and zoom.

c. Neural Network



The 'Neural Network - Orange' dialog box is shown. It has a title bar with a question mark and a close button. The 'Name' field contains 'Neural Network'. The 'Neurons in hidden layers' field contains '100,'. The 'Activation' dropdown is set to 'ReLu'. The 'Solver' dropdown is set to 'Adam'. The 'Regularization, $\alpha=0.0001$:' slider is at the minimum. The 'Maximal number of iterations' spinner is set to '200'. The 'Replicable training' checkbox is checked. At the bottom, there are 'Cancel' and 'Apply Automatically' buttons, with the latter being checked. A toolbar with icons for help, save, undo, redo, and zoom is at the very bottom.

Name: Neural Network

Neurons in hidden layers: 100,

Activation: ReLu

Solver: Adam

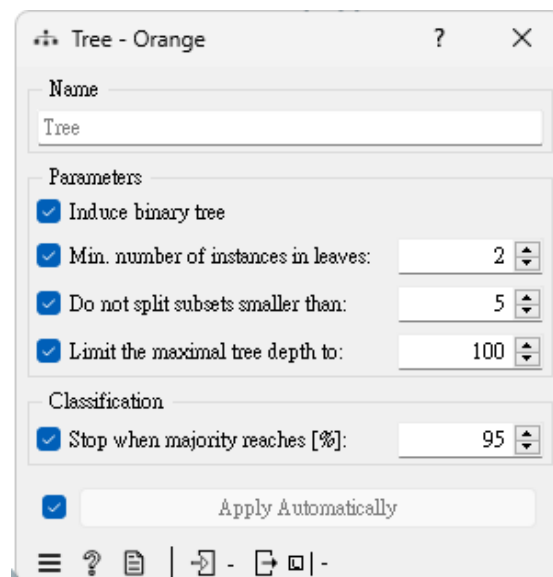
Regularization, $\alpha=0.0001$:

Maximal number of iterations: 200

☒ Replicable training

Cancel ☒ Apply Automatically

d. Tree



The 'Tree - Orange' dialog box is shown. It has a title bar with a question mark and a close button. The 'Name' field contains 'Tree'. Under the 'Parameters' section, 'Induce binary tree' is checked. 'Min. number of instances in leaves' is set to 2. 'Do not split subsets smaller than:' is set to 5. 'Limit the maximal tree depth to:' is set to 100. Under the 'Classification' section, 'Stop when majority reaches [%]:' is set to 95. At the bottom, there is an 'Apply Automatically' button which is checked. A toolbar with icons for help, save, undo, redo, and zoom is at the very bottom.

Name: Tree

Parameters

☒ Induce binary tree

☒ Min. number of instances in leaves: 2

☒ Do not split subsets smaller than: 5

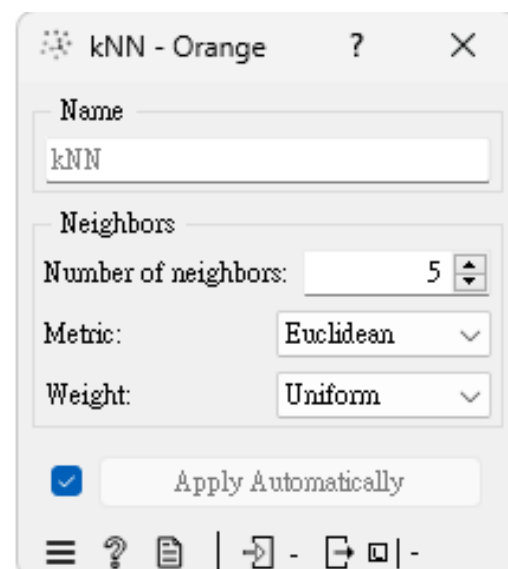
☒ Limit the maximal tree depth to: 100

Classification

☒ Stop when majority reaches [%]: 95

☒ Apply Automatically

e. kNN



The 'kNN - Orange' dialog box is shown. It has a title bar with a question mark and a close button. The 'Name' field contains 'kNN'. Under the 'Neighbors' section, 'Number of neighbors' is set to 5. 'Metric' is set to 'Euclidean'. 'Weight' is set to 'Uniform'. At the bottom, there is an 'Apply Automatically' button which is checked. A toolbar with icons for help, save, undo, redo, and zoom is at the very bottom.

Name: kNN

Neighbors

Number of neighbors: 5

Metric: Euclidean

Weight: Uniform

☒ Apply Automatically

4. 從上面的結果圖可以發現，在Document Embedding(fastText)以及Bag of Words只有SVM的 accuracy 相差較大，其他的 Model accuracy 相差不大。而三種方式裡面，Document Embedding(Multilingual SBERT) 所得到的平均 accuracy 是最高的，最高的甚至有到0.99，所以可能對於這份spam mail的資料，使用Document Embedding(Multilingual SBERT)搭配Logistic Regression、SVM或是Neural Network等方法可以達到最好的分類準確度。