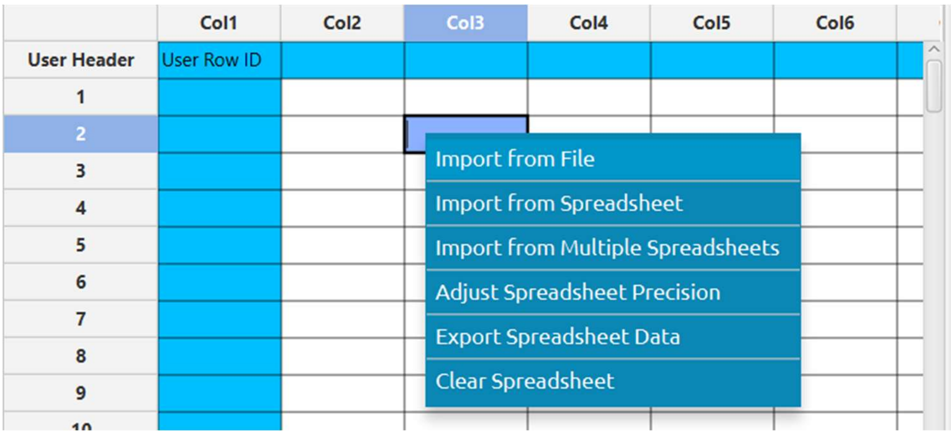
Email Spam Classification Dataset CSV

The csv file, which can be found in https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv/data contains 5172 rows, each row for each email. The first column indicates Email name. The name has been set with numbers and not recipients' name to protect privacy. The last column has the labels for prediction: 1 for spam, 0 for not spam. The remaining columns are the most common words in all the emails, after excluding the non-alphabetical characters/words.

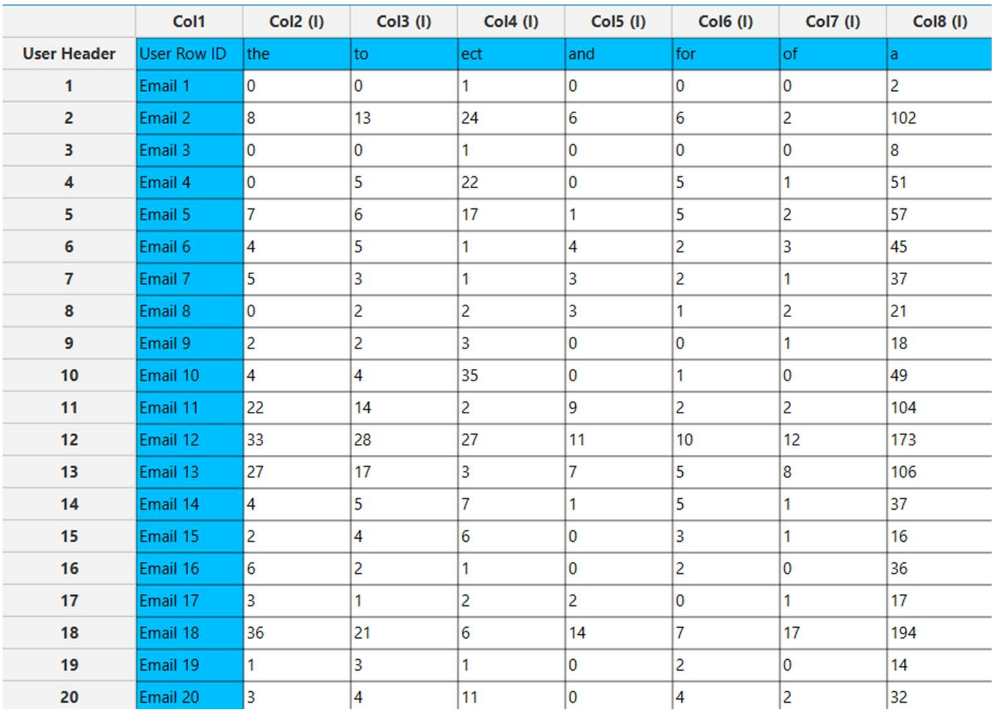
Isalos version used: 2.0.6

Step 1: Import data from file

Right click on the input spreadsheet (left) and choose the option "Import from File". Then navigate through your files to load the one with the spam emails data.



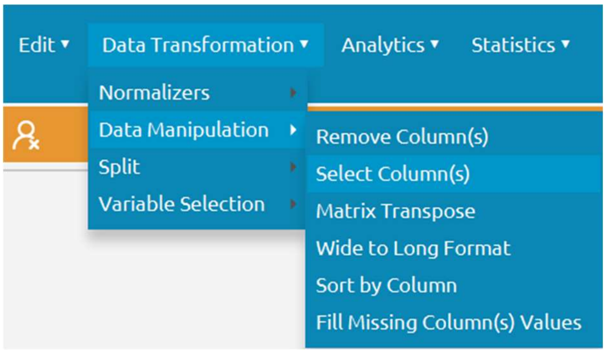
The data will appear on the left spreadsheet.

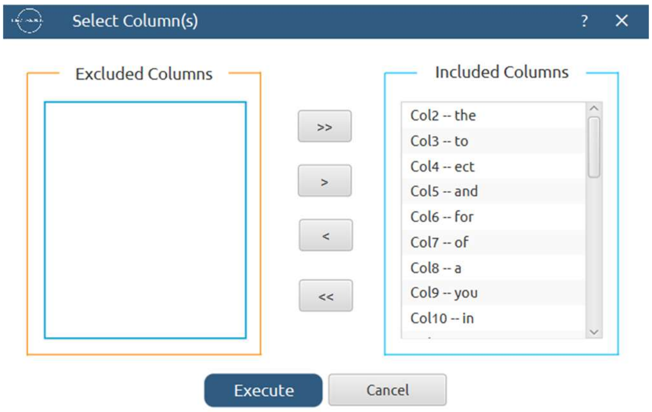


	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (I)	Col6 (I)	Col7 (I)	Col8 (I)
User Header	User Row ID	the	to	ect	and	for	of	a
1	Email 1	0	0	1	0	0	0	2
2	Email 2	8	13	24	6	6	2	102
3	Email 3	0	0	1	0	0	0	8
4	Email 4	0	5	22	0	5	1	51
5	Email 5	7	6	17	1	5	2	57
6	Email 6	4	5	1	4	2	3	45
7	Email 7	5	3	1	3	2	1	37
8	Email 8	0	2	2	3	1	2	21
9	Email 9	2	2	3	0	0	1	18
10	Email 10	4	4	35	0	1	0	49
11	Email 11	22	14	2	9	2	2	104
12	Email 12	33	28	27	11	10	12	173
13	Email 13	27	17	3	7	5	8	106
14	Email 14	4	5	7	1	5	1	37
15	Email 15	2	4	6	0	3	1	16
16	Email 16	6	2	1	0	2	0	36
17	Email 17	3	1	2	2	0	1	17
18	Email 18	36	21	6	14	7	17	194
19	Email 19	1	3	1	0	2	0	14
20	Email 20	3	4	11	0	4	2	32

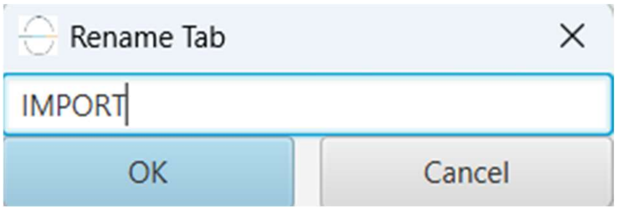
Step 2: Manipulate data

In our dataset there are not any empty values or categorical features, so we can select all the columns to be used. On the menu click on *Data Transformation → Data Manipulation → Select Column(s)* and select all columns.





All of the data will appear in the output (right) spreadsheet. This tab can be renamed "IMPORT" by right-clicking on it and choosing the "Rename" option.

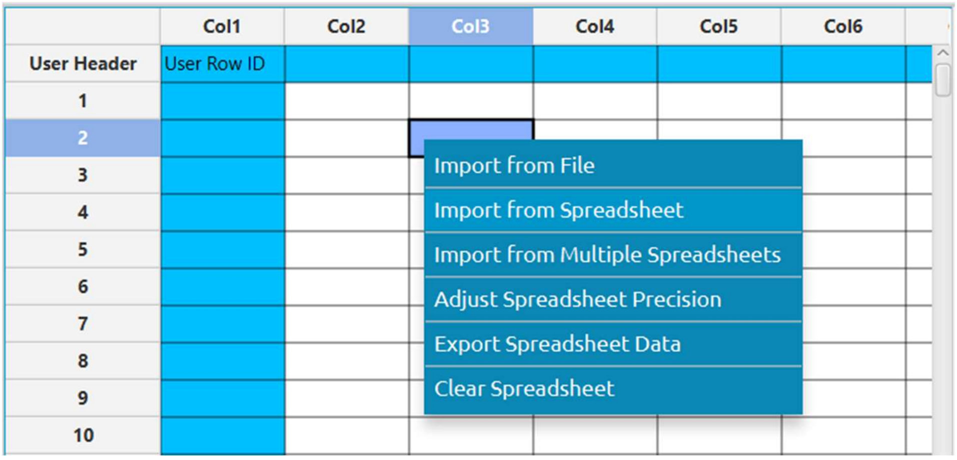

Rename Tab

IMPORT

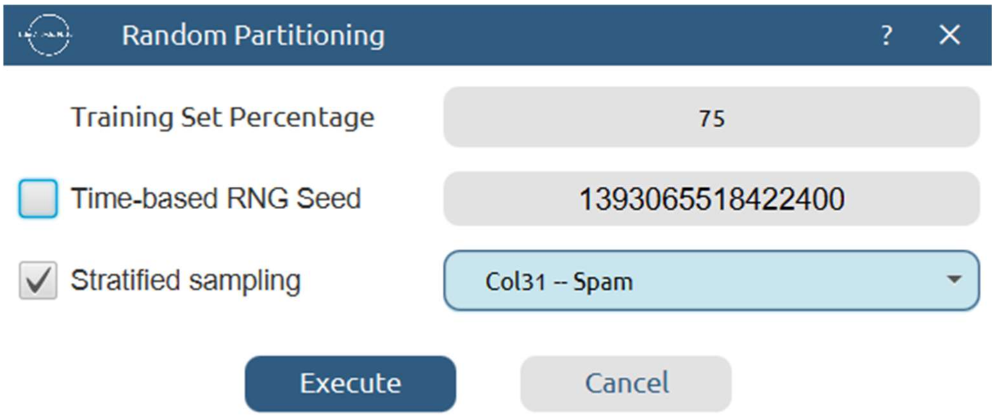
Step 3: Split data

Create a new tab by pressing the "+" button on the bottom of the page with the name "TRAIN_TEST_SPLIT" which we will use for splitting the train and test set.

Import data into the input spreadsheet of the "TRAIN_TEST_SPLIT" tab from the output of the "IMPORT" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".



Split the dataset by choosing *Data Transformation → Split → Random Partitioning*. Then choose the "Training set percentage" and the column for the sampling as shown below:



Training Set Percentage: 75

☐ Time-based RNG Seed: 1393065518422400

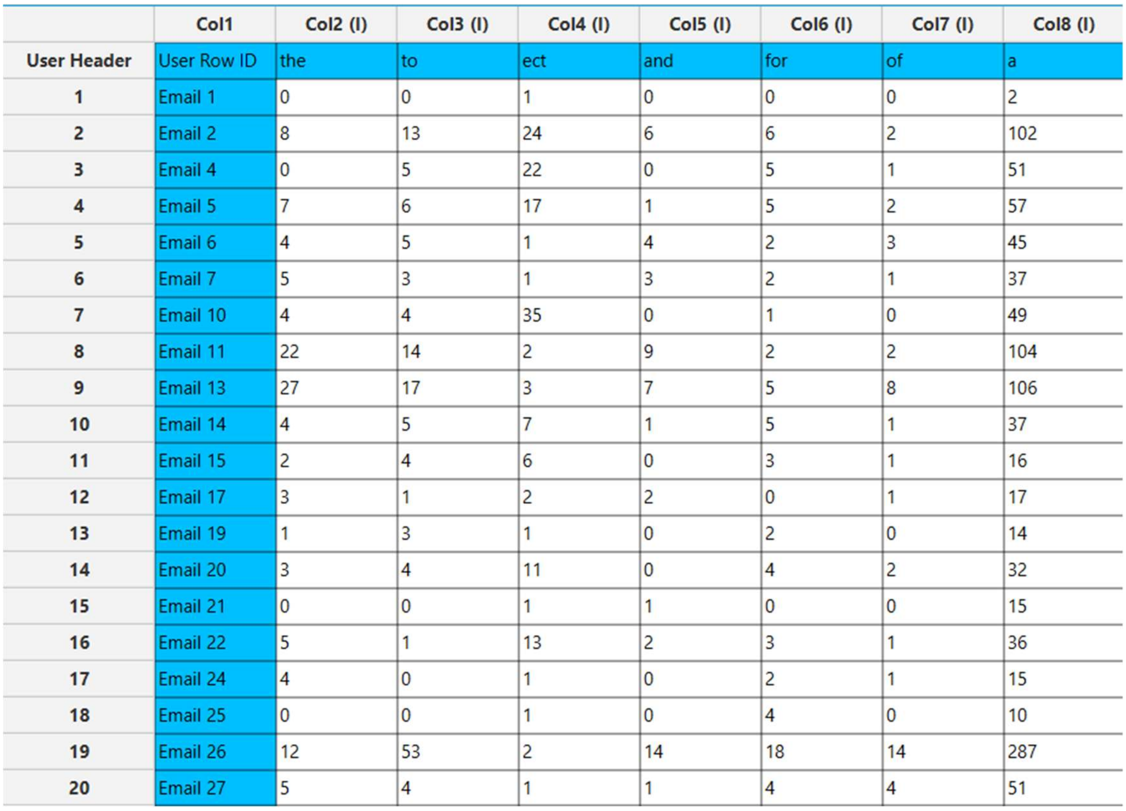
☑ Stratified sampling: Col31 -- Spam

The results will be two separate spreadsheets, "TRAIN_TEST_SPLIT: Training Set" and "TRAIN_TEST_SPLIT: Test Set", which will be available to import into the next tabs.

Step 4: Normalize the training set

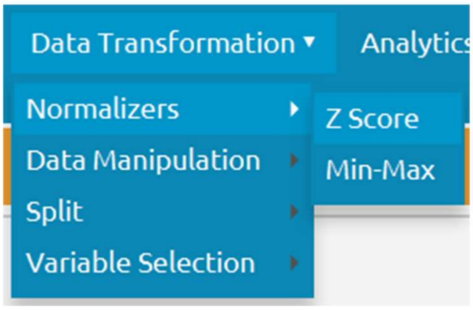
Create a new tab by pressing the “+” button on the bottom of the page with the name “NORMALIZE_TRAIN_SET”.

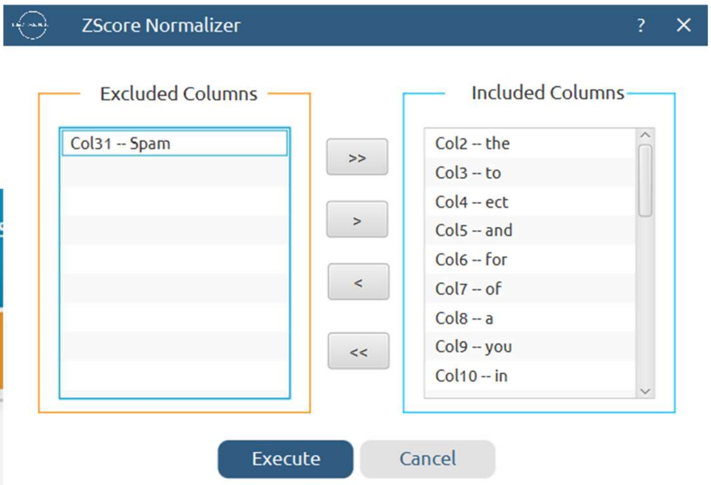
Import into the input spreadsheet of the “NORMALIZE_TRAIN_SET” tab the train set from the output of the “TRAIN_TEST_SPLIT” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”. From the available Select input tab options choose “TRAIN_TEST_SPLIT: Training Set”.



	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (I)	Col6 (I)	Col7 (I)	Col8 (I)
User Header	User Row ID	the	to	ect	and	for	of	a
1	Email 1	0	0	1	0	0	0	2
2	Email 2	8	13	24	6	6	2	102
3	Email 4	0	5	22	0	5	1	51
4	Email 5	7	6	17	1	5	2	57
5	Email 6	4	5	1	4	2	3	45
6	Email 7	5	3	1	3	2	1	37
7	Email 10	4	4	35	0	1	0	49
8	Email 11	22	14	2	9	2	2	104
9	Email 13	27	17	3	7	5	8	106
10	Email 14	4	5	7	1	5	1	37
11	Email 15	2	4	6	0	3	1	16
12	Email 17	3	1	2	2	0	1	17
13	Email 19	1	3	1	0	2	0	14
14	Email 20	3	4	11	0	4	2	32
15	Email 21	0	0	1	1	0	0	15
16	Email 22	5	1	13	2	3	1	36
17	Email 24	4	0	1	0	2	1	15
18	Email 25	0	0	1	0	4	0	10
19	Email 26	12	53	2	14	18	14	287
20	Email 27	5	4	1	1	4	4	51

Normalize the data using Z-score: *Data Transformation → Normalizers → Z Score* and select all columns except the “Spam” target column.





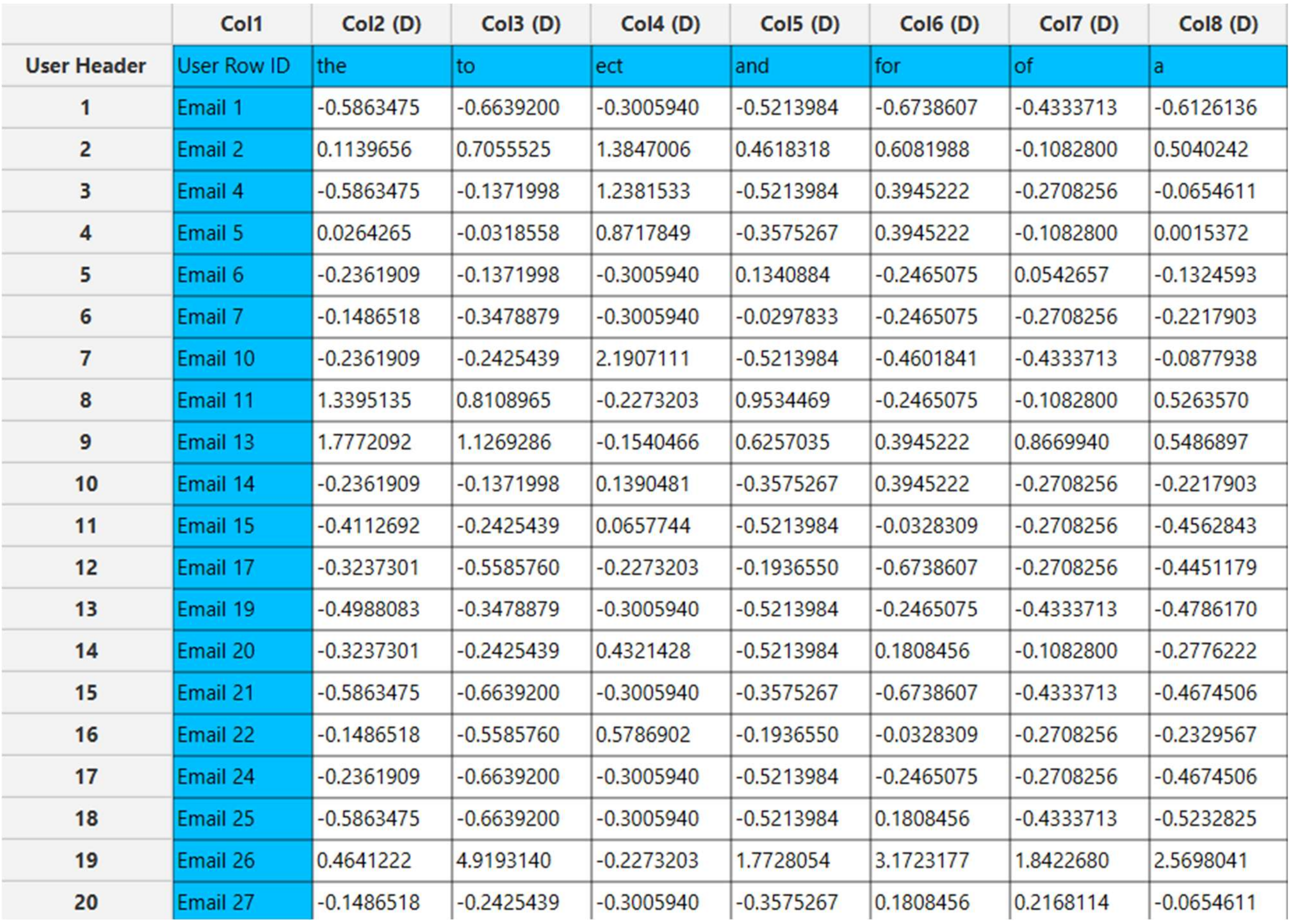
Excluded Columns

Col31 -- Spam

Included Columns

Col2 -- the
Col3 -- to
Col4 -- ect
Col5 -- and
Col6 -- for
Col7 -- of
Col8 -- a
Col9 -- you
Col10 -- in

The results will appear on the output spreadsheet.

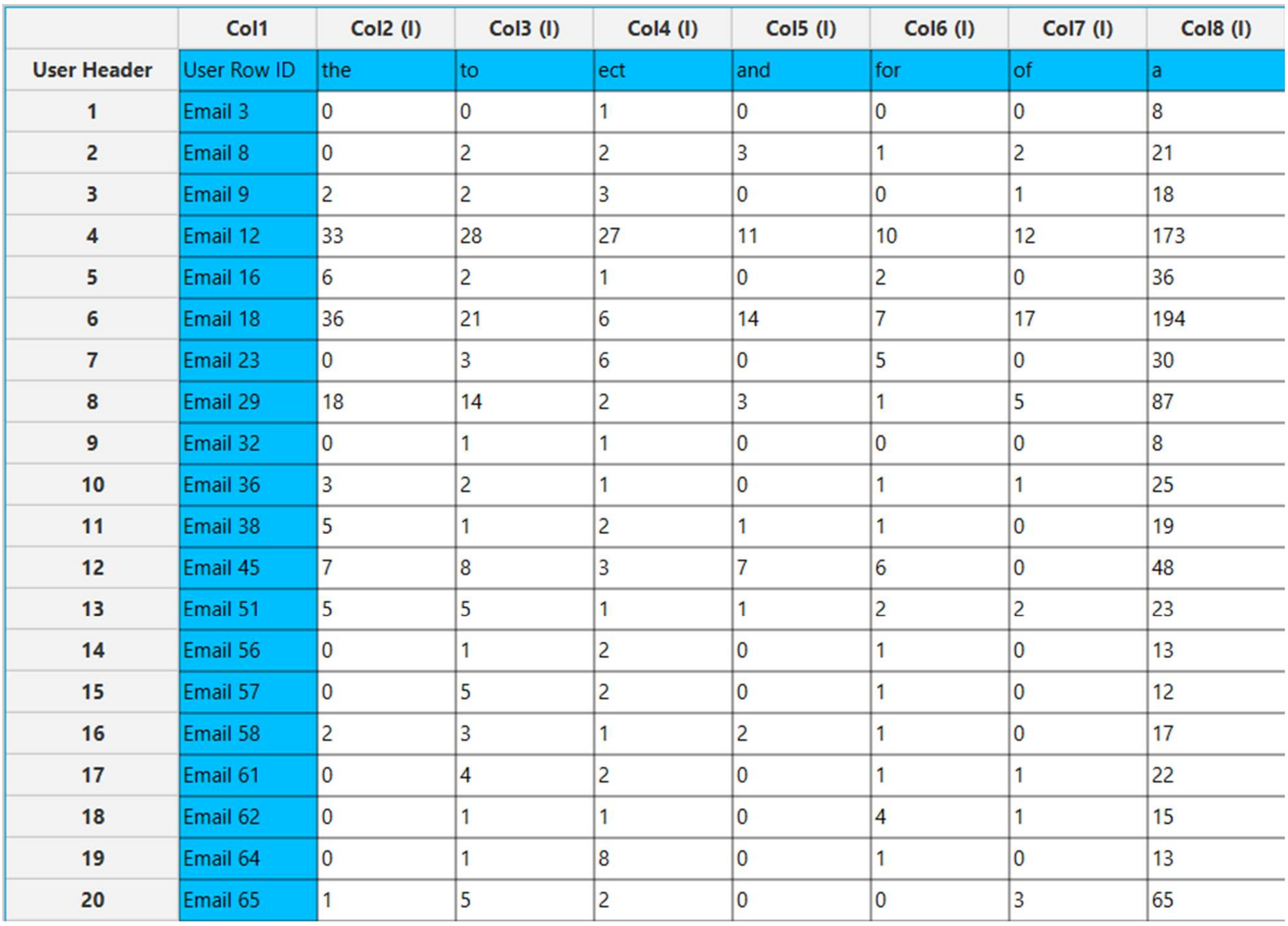


	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
User Header	User Row ID	the	to	ect	and	for	of	a
1	Email 1	-0.5863475	-0.6639200	-0.3005940	-0.5213984	-0.6738607	-0.4333713	-0.6126136
2	Email 2	0.1139656	0.7055525	1.3847006	0.4618318	0.6081988	-0.1082800	0.5040242
3	Email 4	-0.5863475	-0.1371998	1.2381533	-0.5213984	0.3945222	-0.2708256	-0.0654611
4	Email 5	0.0264265	-0.0318558	0.8717849	-0.3575267	0.3945222	-0.1082800	0.0015372
5	Email 6	-0.2361909	-0.1371998	-0.3005940	0.1340884	-0.2465075	0.0542657	-0.1324593
6	Email 7	-0.1486518	-0.3478879	-0.3005940	-0.0297833	-0.2465075	-0.2708256	-0.2217903
7	Email 10	-0.2361909	-0.2425439	2.1907111	-0.5213984	-0.4601841	-0.4333713	-0.0877938
8	Email 11	1.3395135	0.8108965	-0.2273203	0.9534469	-0.2465075	-0.1082800	0.5263570
9	Email 13	1.7772092	1.1269286	-0.1540466	0.6257035	0.3945222	0.8669940	0.5486897
10	Email 14	-0.2361909	-0.1371998	0.1390481	-0.3575267	0.3945222	-0.2708256	-0.2217903
11	Email 15	-0.4112692	-0.2425439	0.0657744	-0.5213984	-0.0328309	-0.2708256	-0.4562843
12	Email 17	-0.3237301	-0.5585760	-0.2273203	-0.1936550	-0.6738607	-0.2708256	-0.4451179
13	Email 19	-0.4988083	-0.3478879	-0.3005940	-0.5213984	-0.2465075	-0.4333713	-0.4786170
14	Email 20	-0.3237301	-0.2425439	0.4321428	-0.5213984	0.1808456	-0.1082800	-0.2776222
15	Email 21	-0.5863475	-0.6639200	-0.3005940	-0.3575267	-0.6738607	-0.4333713	-0.4674506
16	Email 22	-0.1486518	-0.5585760	0.5786902	-0.1936550	-0.0328309	-0.2708256	-0.2329567
17	Email 24	-0.2361909	-0.6639200	-0.3005940	-0.5213984	-0.2465075	-0.2708256	-0.4674506
18	Email 25	-0.5863475	-0.6639200	-0.3005940	-0.5213984	0.1808456	-0.4333713	-0.5232825
19	Email 26	0.4641222	4.9193140	-0.2273203	1.7728054	3.1723177	1.8422680	2.5698041
20	Email 27	-0.1486518	-0.2425439	-0.3005940	-0.3575267	0.1808456	0.2168114	-0.0654611

Step 5: Normalize the test set

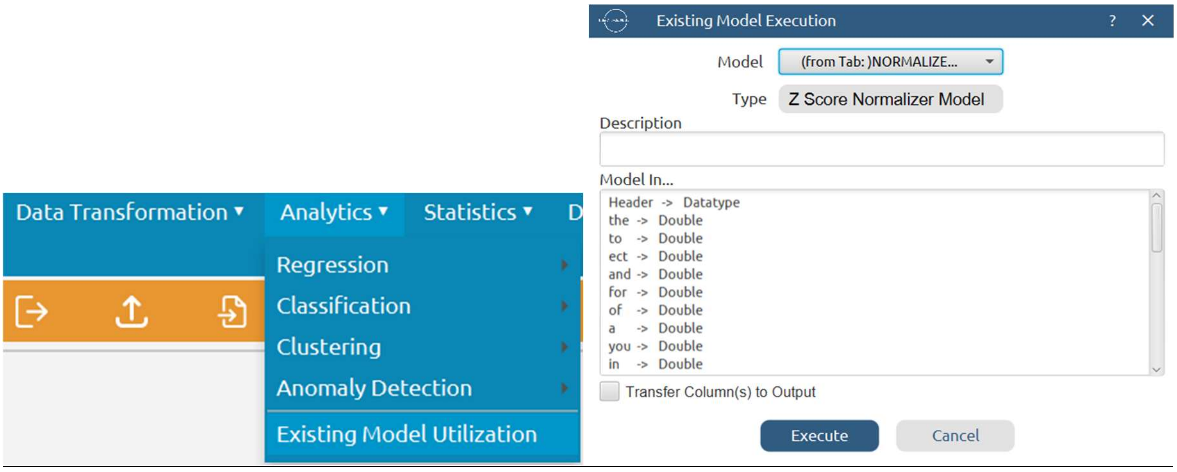
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALIZE_TEST_SET".

Import into the input spreadsheet of the "NORMALIZE_TEST_SET" tab the test set from the output of the "TRAIN_TEST_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet". From the available Select input tab options choose "TRAIN_TEST_SPLIT: Test Set".



	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (I)	Col6 (I)	Col7 (I)	Col8 (I)
User Header	User Row ID	the	to	ect	and	for	of	a
1	Email 3	0	0	1	0	0	0	8
2	Email 8	0	2	2	3	1	2	21
3	Email 9	2	2	3	0	0	1	18
4	Email 12	33	28	27	11	10	12	173
5	Email 16	6	2	1	0	2	0	36
6	Email 18	36	21	6	14	7	17	194
7	Email 23	0	3	6	0	5	0	30
8	Email 29	18	14	2	3	1	5	87
9	Email 32	0	1	1	0	0	0	8
10	Email 36	3	2	1	0	1	1	25
11	Email 38	5	1	2	1	1	0	19
12	Email 45	7	8	3	7	6	0	48
13	Email 51	5	5	1	1	2	2	23
14	Email 56	0	1	2	0	1	0	13
15	Email 57	0	5	2	0	1	0	12
16	Email 58	2	3	1	2	1	0	17
17	Email 61	0	4	2	0	1	1	22
18	Email 62	0	1	1	0	4	1	15
19	Email 64	0	1	8	0	1	0	13
20	Email 65	1	5	2	0	0	3	65

Normalize the test set using the existing normalizer of the training set: *Analytics → Existing Model Utilization → Model (from Tab:) NORMALIZE_TRAIN_SET*



Model (from Tab:)NORMALIZE...

Type Z Score Normalizer Model

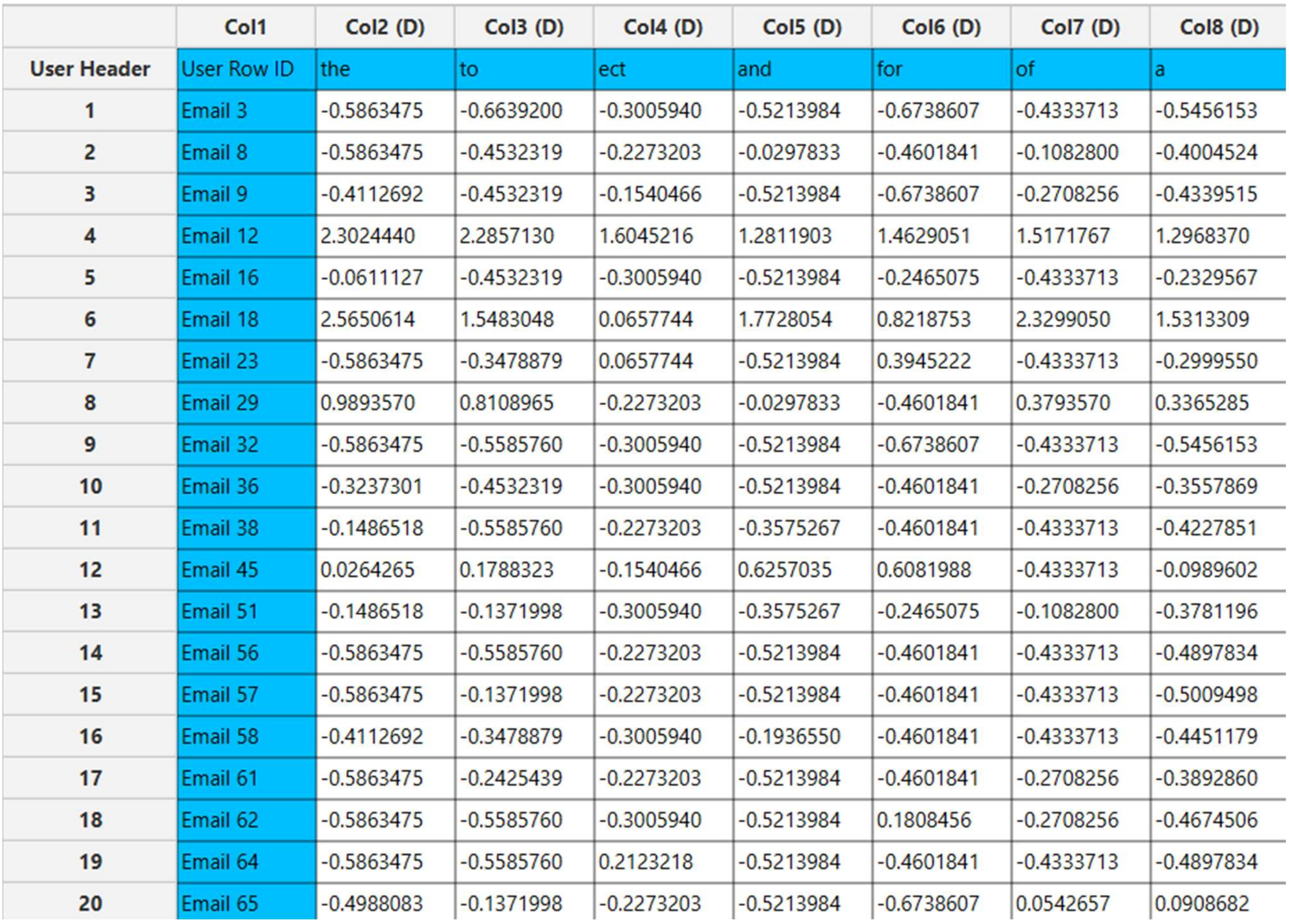
Description

Model In...

```
Header -> Datatype
the  -> Double
to   -> Double
ect  -> Double
and  -> Double
for  -> Double
of   -> Double
a    -> Double
you  -> Double
in   -> Double
```

☐ Transfer Column(s) to Output

The results will appear on the output spreadsheet.



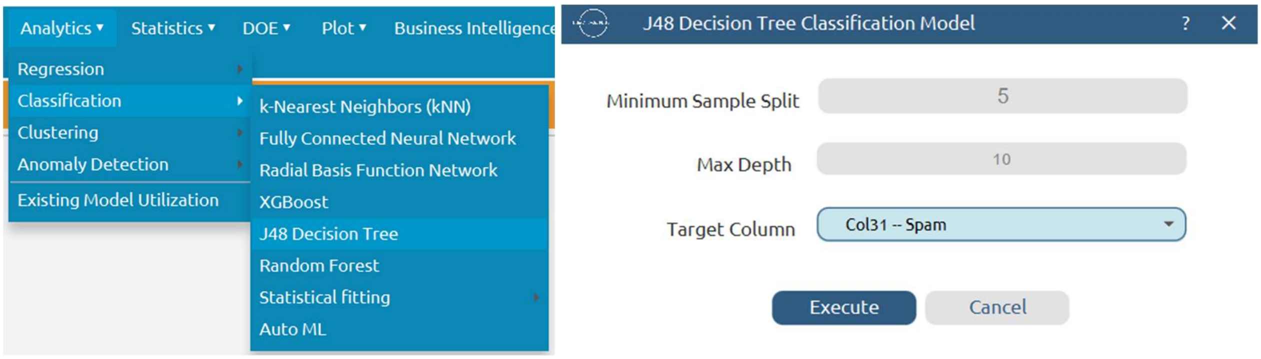
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
User Header	User Row ID	the	to	ect	and	for	of	a
1	Email 3	-0.5863475	-0.6639200	-0.3005940	-0.5213984	-0.6738607	-0.4333713	-0.5456153
2	Email 8	-0.5863475	-0.4532319	-0.2273203	-0.0297833	-0.4601841	-0.1082800	-0.4004524
3	Email 9	-0.4112692	-0.4532319	-0.1540466	-0.5213984	-0.6738607	-0.2708256	-0.4339515
4	Email 12	2.3024440	2.2857130	1.6045216	1.2811903	1.4629051	1.5171767	1.2968370
5	Email 16	-0.0611127	-0.4532319	-0.3005940	-0.5213984	-0.2465075	-0.4333713	-0.2329567
6	Email 18	2.5650614	1.5483048	0.0657744	1.7728054	0.8218753	2.3299050	1.5313309
7	Email 23	-0.5863475	-0.3478879	0.0657744	-0.5213984	0.3945222	-0.4333713	-0.2999550
8	Email 29	0.9893570	0.8108965	-0.2273203	-0.0297833	-0.4601841	0.3793570	0.3365285
9	Email 32	-0.5863475	-0.5585760	-0.3005940	-0.5213984	-0.6738607	-0.4333713	-0.5456153
10	Email 36	-0.3237301	-0.4532319	-0.3005940	-0.5213984	-0.4601841	-0.2708256	-0.3557869
11	Email 38	-0.1486518	-0.5585760	-0.2273203	-0.3575267	-0.4601841	-0.4333713	-0.4227851
12	Email 45	0.0264265	0.1788323	-0.1540466	0.6257035	0.6081988	-0.4333713	-0.0989602
13	Email 51	-0.1486518	-0.1371998	-0.3005940	-0.3575267	-0.2465075	-0.1082800	-0.3781196
14	Email 56	-0.5863475	-0.5585760	-0.2273203	-0.5213984	-0.4601841	-0.4333713	-0.4897834
15	Email 57	-0.5863475	-0.1371998	-0.2273203	-0.5213984	-0.4601841	-0.4333713	-0.5009498
16	Email 58	-0.4112692	-0.3478879	-0.3005940	-0.1936550	-0.4601841	-0.4333713	-0.4451179
17	Email 61	-0.5863475	-0.2425439	-0.2273203	-0.5213984	-0.4601841	-0.2708256	-0.3892860
18	Email 62	-0.5863475	-0.5585760	-0.3005940	-0.5213984	0.1808456	-0.2708256	-0.4674506
19	Email 64	-0.5863475	-0.5585760	0.2123218	-0.5213984	-0.4601841	-0.4333713	-0.4897834
20	Email 65	-0.4988083	-0.1371998	-0.2273203	-0.5213984	-0.6738607	0.0542657	0.0908682

Step 6: Train the model

Create a new tab by pressing the "+" button on the bottom of the page with the name "TRAIN_MODEL(.fit)".

Import data into the input spreadsheet of the "TRAIN_MODEL(.fit)" tab from the output of the "NORMALIZE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

Use the J48 Decision Tree Method to train and fit the model: *Analytics → Classification → J48 Decision Tree*



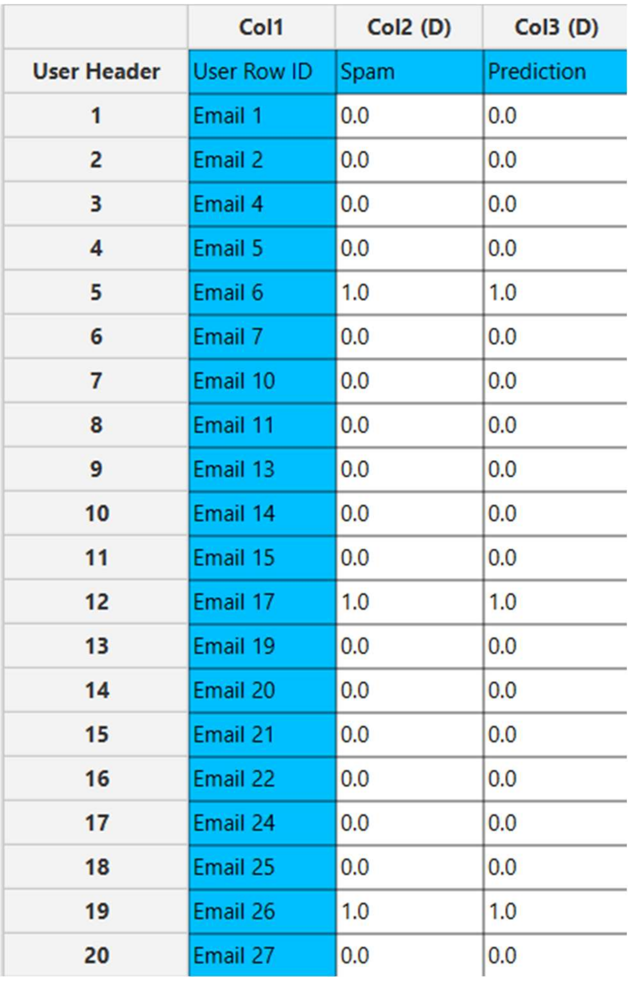
J48 Decision Tree Classification Model

Minimum Sample Split: 5

Max Depth: 10

Target Column: Col31 -- Spam

The predictions will appear on the output spreadsheet.



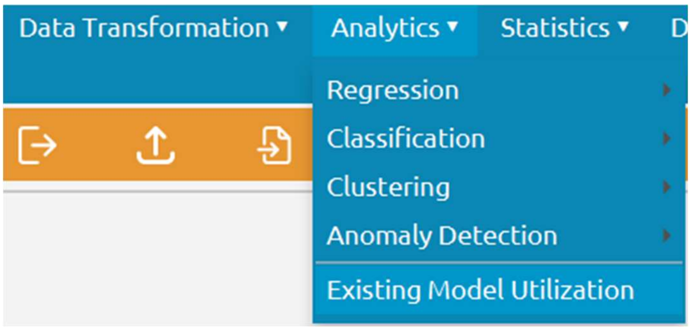
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Spam	Prediction
1	Email 1	0.0	0.0
2	Email 2	0.0	0.0
3	Email 4	0.0	0.0
4	Email 5	0.0	0.0
5	Email 6	1.0	1.0
6	Email 7	0.0	0.0
7	Email 10	0.0	0.0
8	Email 11	0.0	0.0
9	Email 13	0.0	0.0
10	Email 14	0.0	0.0
11	Email 15	0.0	0.0
12	Email 17	1.0	1.0
13	Email 19	0.0	0.0
14	Email 20	0.0	0.0
15	Email 21	0.0	0.0
16	Email 22	0.0	0.0
17	Email 24	0.0	0.0
18	Email 25	0.0	0.0
19	Email 26	1.0	1.0
20	Email 27	0.0	0.0

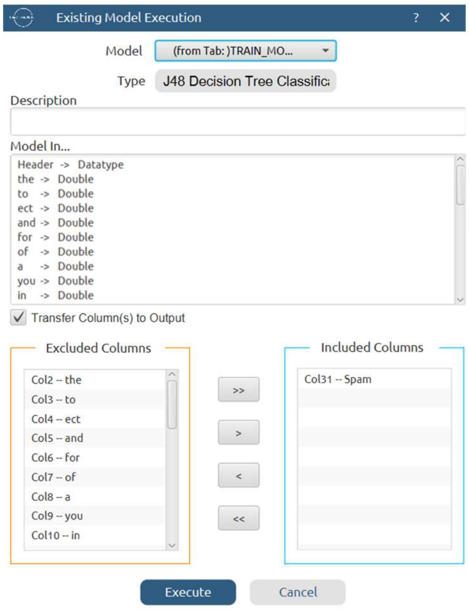
Step 7: Validate the model

Create a new tab by pressing the "+" button on the bottom of the page with the name "VALIDATE_MODEL(.predict)".

Import data into the input spreadsheet of the "VALIDATE_MODEL(.predict)" tab from the output of the "NORMALIZE_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

To validate the model: *Analytics → Existing Model Utilization→ Model (from Tab:) TRAIN_MODEL(.fit)*. Choose the column "Spam" to be transferred to the output spreadsheet.





Model (from Tab:)TRAIN_MO...

Type J48 Decision Tree Classifica

Description

Model In...

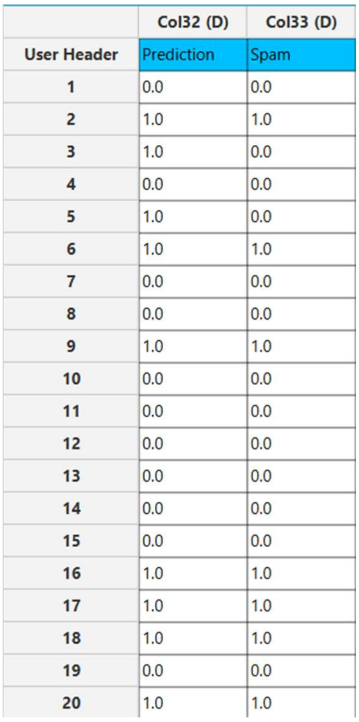
```
Header -> Datatype
the -> Double
to -> Double
ect -> Double
and -> Double
for -> Double
of -> Double
a -> Double
you -> Double
in -> Double
```

☑ Transfer Column(s) to Output

Excluded Columns: Col2 -- the, Col3 -- to, Col4 -- ect, Col5 -- and, Col6 -- for, Col7 -- of, Col8 -- a, Col9 -- you, Col10 -- in

Included Columns: Col31 -- Spam

The predictions will appear on the output spreadsheet.



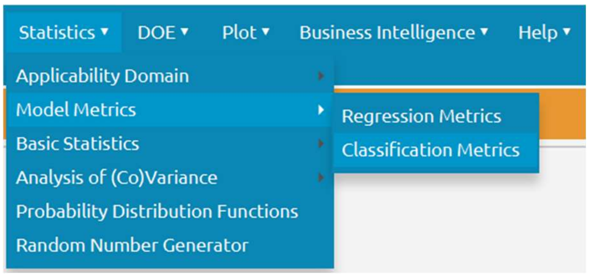
	Col32 (D)	Col33 (D)
User Header	Prediction	Spam
1	0.0	0.0
2	1.0	1.0
3	1.0	0.0
4	0.0	0.0
5	1.0	0.0
6	1.0	1.0
7	0.0	0.0
8	0.0	0.0
9	1.0	1.0
10	0.0	0.0
11	0.0	0.0
12	0.0	0.0
13	0.0	0.0
14	0.0	0.0
15	0.0	0.0
16	1.0	1.0
17	1.0	1.0
18	1.0	1.0
19	0.0	0.0
20	1.0	1.0

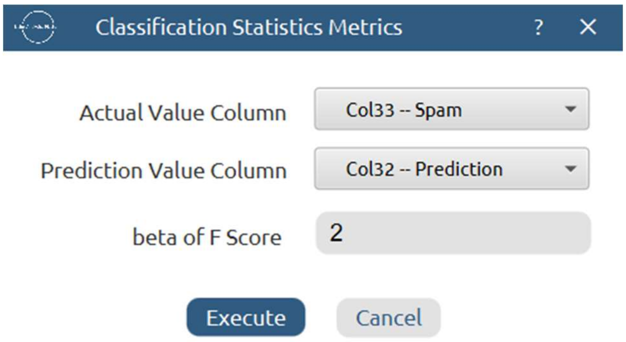
Step 8: Statistics calculation

Create a new tab by pressing the “+” button on the bottom of the page with the name “STATISTICS_ACCURACIES”.

Import data into the input spreadsheet of the “STATISTICS_ACCURACIES” tab from the output of the “VALIDATE_MODEL(.predict)” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Calculate the statistical metrics for the classification: *Statistics → Model Metrics → Classification Metrics*



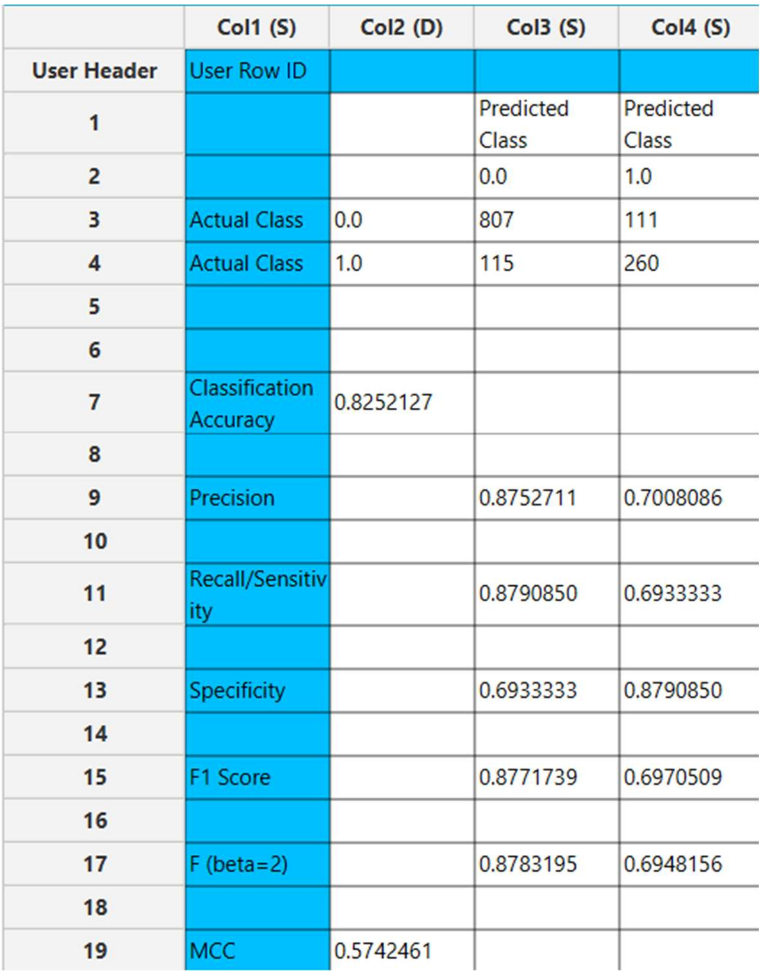


Actual Value Column: Col33 -- Spam

Prediction Value Column: Col32 -- Prediction

beta of F Score: 2

The results will appear on the output spreadsheet.



	Col1 (S)	Col2 (D)	Col3 (S)	Col4 (S)
User Header	User Row ID			
1			Predicted Class	Predicted Class
2			0.0	1.0
3	Actual Class	0.0	807	111
4	Actual Class	1.0	115	260
5				
6				
7	Classification Accuracy	0.8252127		
8				
9	Precision		0.8752711	0.7008086
10				
11	Recall/Sensitivity		0.8790850	0.6933333
12				
13	Specificity		0.6933333	0.8790850
14				
15	F1 Score		0.8771739	0.6970509
16				
17	F (beta=2)		0.8783195	0.6948156
18				
19	MCC	0.5742461		

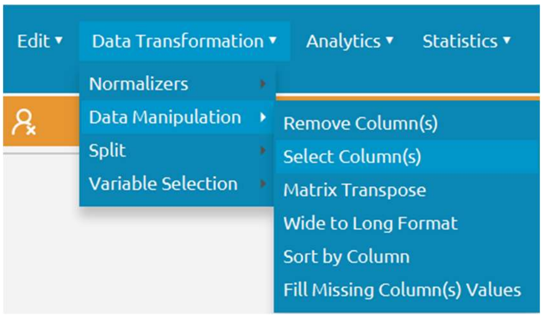
Step 9: Reliability check for each record of the test set

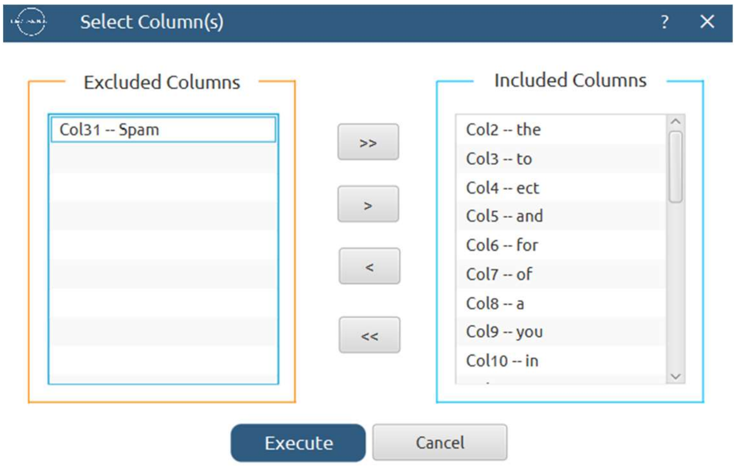
Step 9.a: Create the domain

Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_SPAM".

Import data into the input spreadsheet of the "EXCLUDE_SPAM" tab from the output of the "NORMALIZE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

Manipulate the data to exclude the target column "Spam": *Data Transformation → Data Manipulation → Select Column(s)*





Excluded Columns: Col31 -- Spam

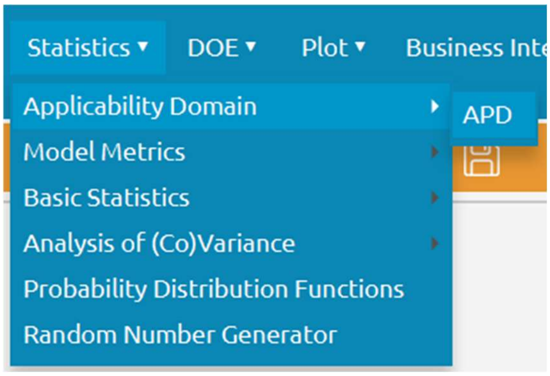
Included Columns: Col2 -- the, Col3 -- to, Col4 -- ect, Col5 -- and, Col6 -- for, Col7 -- of, Col8 -- a, Col9 -- you, Col10 -- in

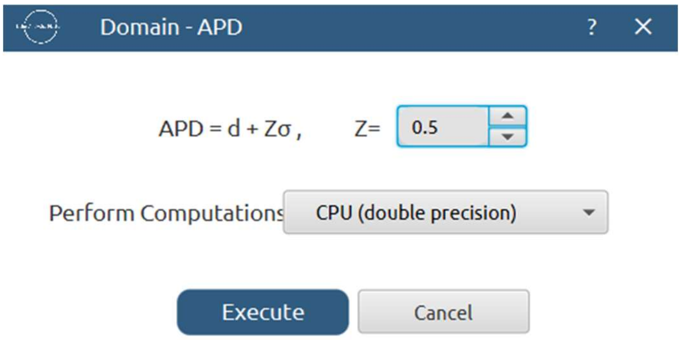
The results will appear on the output spreadsheet.

Create a new tab by pressing the "+" button on the bottom of the page with the name "DOMAIN".

Import data into the input spreadsheet of the "DOMAIN" tab from the output of the "EXCLUDE_SPAM" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

Create the domain: *Statistics → Applicability Domain → APD*

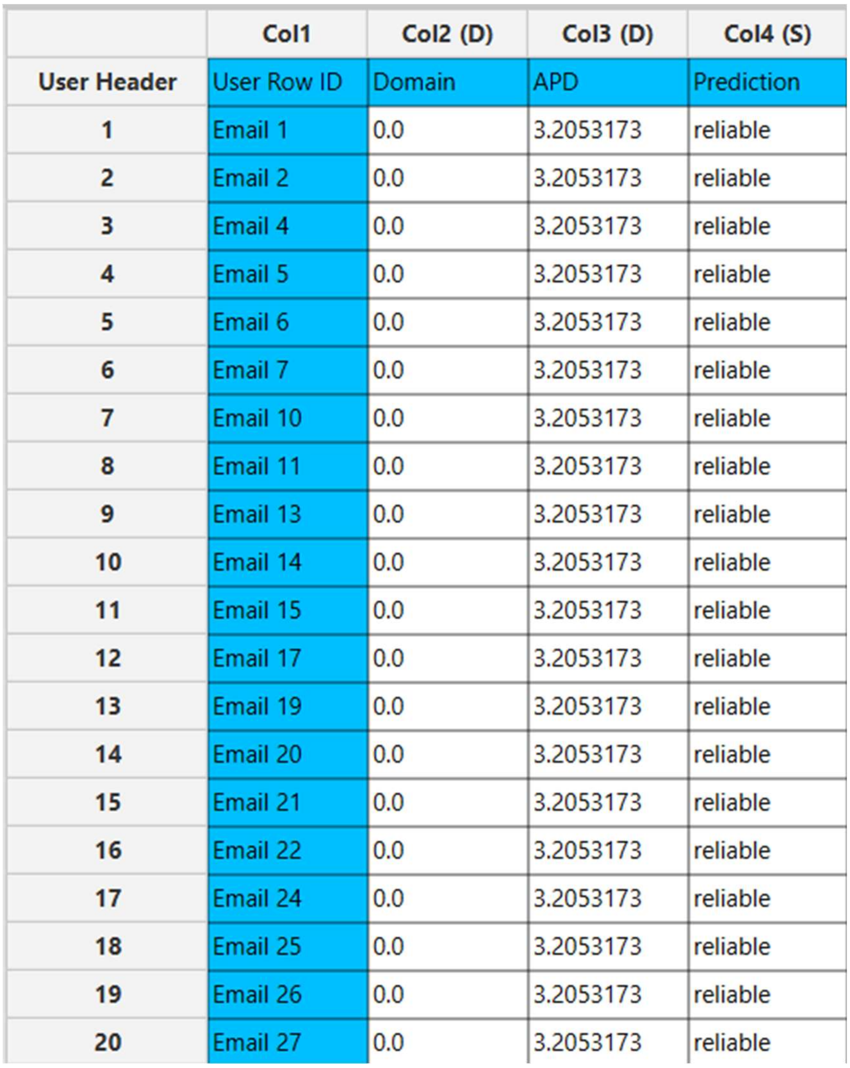




APD = d + Zσ , Z= 0.5

Perform Computations: CPU (double precision)

The results will appear on the output spreadsheet.



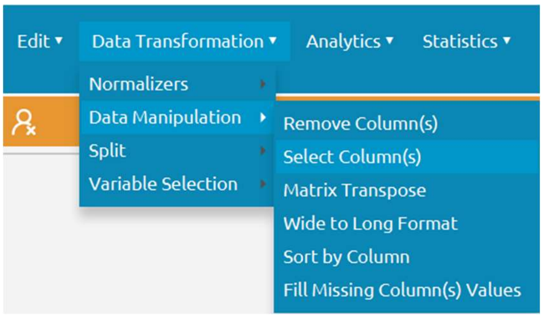
	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1	Email 1	0.0	3.2053173	reliable
2	Email 2	0.0	3.2053173	reliable
3	Email 4	0.0	3.2053173	reliable
4	Email 5	0.0	3.2053173	reliable
5	Email 6	0.0	3.2053173	reliable
6	Email 7	0.0	3.2053173	reliable
7	Email 10	0.0	3.2053173	reliable
8	Email 11	0.0	3.2053173	reliable
9	Email 13	0.0	3.2053173	reliable
10	Email 14	0.0	3.2053173	reliable
11	Email 15	0.0	3.2053173	reliable
12	Email 17	0.0	3.2053173	reliable
13	Email 19	0.0	3.2053173	reliable
14	Email 20	0.0	3.2053173	reliable
15	Email 21	0.0	3.2053173	reliable
16	Email 22	0.0	3.2053173	reliable
17	Email 24	0.0	3.2053173	reliable
18	Email 25	0.0	3.2053173	reliable
19	Email 26	0.0	3.2053173	reliable
20	Email 27	0.0	3.2053173	reliable

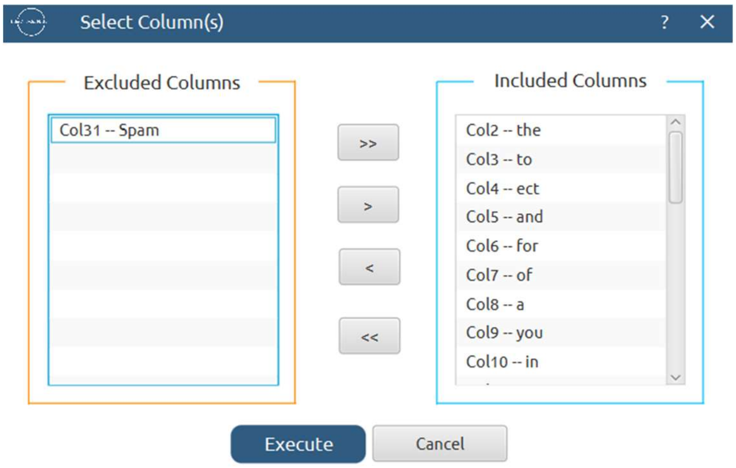
Step 9.b: Check the test set reliability

Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_SPAM_TEST_SET".

Import data into the input spreadsheet of the "EXCLUDE_SPAM_TEST_SET" tab from the output of the "NORMALIZE_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

Manipulate the data to exclude the target column "Spam": *Data Transformation → Data Manipulation → Select Column(s)*




Select Column(s)

Excluded Columns

Col31 -- Spam

Included Columns

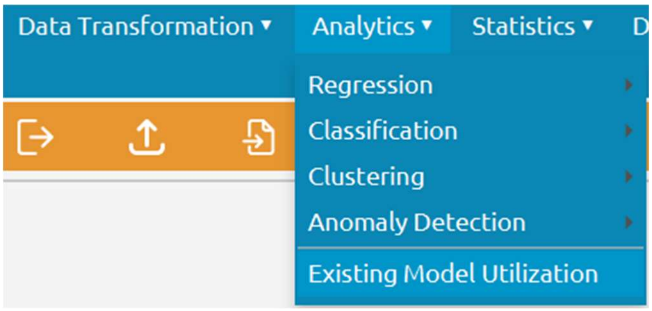
Col2 -- the
Col3 -- to
Col4 -- ect
Col5 -- and
Col6 -- for
Col7 -- of
Col8 -- a
Col9 -- you
Col10 -- in

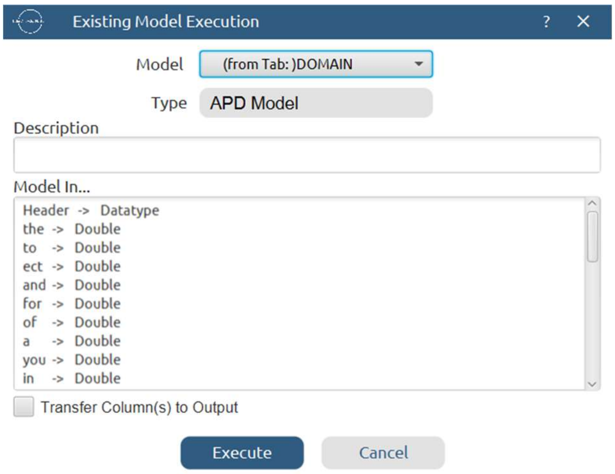
The results will appear on the output spreadsheet.

Create a new tab by pressing the "+" button on the bottom of the page with the name "RELIABILITY".

Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "EXCLUDE_SPAM_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

Check the Reliability: *Analytics → Existing Model Utilization → Model (from Tab:) DOMAIN*




Existing Model Execution

Model (from Tab:)DOMAIN

Type APD Model

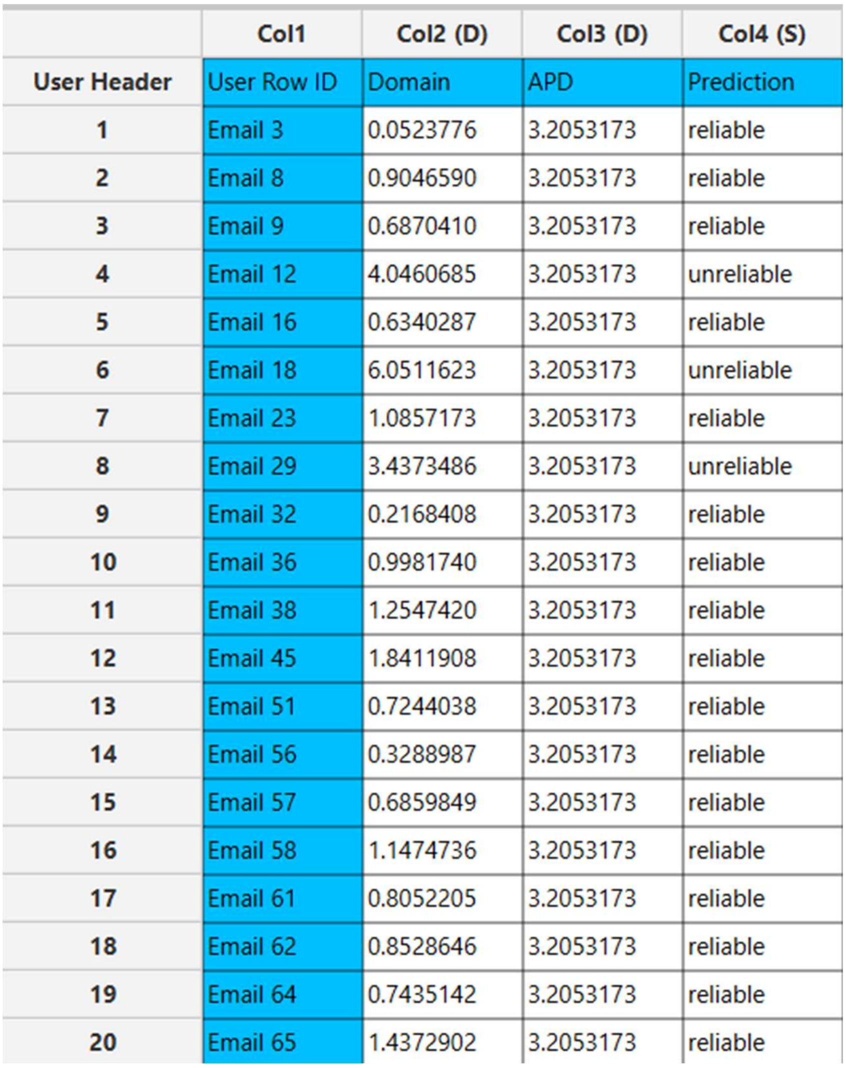
Description

Model In...

```
Header -> Datatype
the  -> Double
to   -> Double
ect  -> Double
and  -> Double
for  -> Double
of   -> Double
a    -> Double
you  -> Double
in   -> Double
```

□ Transfer Column(s) to Output

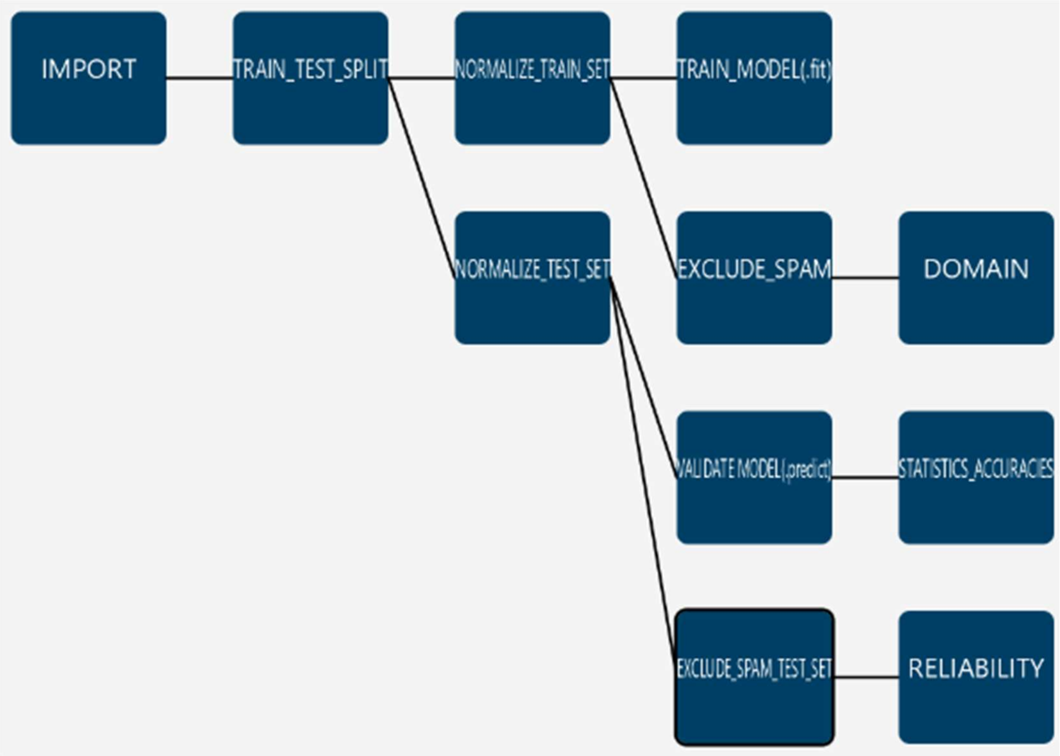
The results will appear on the output spreadsheet.



	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1	Email 3	0.0523776	3.2053173	reliable
2	Email 8	0.9046590	3.2053173	reliable
3	Email 9	0.6870410	3.2053173	reliable
4	Email 12	4.0460685	3.2053173	unreliable
5	Email 16	0.6340287	3.2053173	reliable
6	Email 18	6.0511623	3.2053173	unreliable
7	Email 23	1.0857173	3.2053173	reliable
8	Email 29	3.4373486	3.2053173	unreliable
9	Email 32	0.2168408	3.2053173	reliable
10	Email 36	0.9981740	3.2053173	reliable
11	Email 38	1.2547420	3.2053173	reliable
12	Email 45	1.8411908	3.2053173	reliable
13	Email 51	0.7244038	3.2053173	reliable
14	Email 56	0.3288987	3.2053173	reliable
15	Email 57	0.6859849	3.2053173	reliable
16	Email 58	1.1474736	3.2053173	reliable
17	Email 61	0.8052205	3.2053173	reliable
18	Email 62	0.8528646	3.2053173	reliable
19	Email 64	0.7435142	3.2053173	reliable
20	Email 65	1.4372902	3.2053173	reliable

Final Isalos Workflow

Following the above-described steps, the final workflow on Isalos will look like this:


IMPORT
TRAIN_TEST_SPLIT
NORMALIZE_TRAIN_SET
TRAIN_MODEL(.fit)
NORMALIZE_TEST_SET
EXCLUDE_SPAM
DOMAIN
VALIDATE MODEL(.predict)
STATISTICS_ACCURACIES
EXCLUDE_SPAM_TEST_SET
RELIABILITY