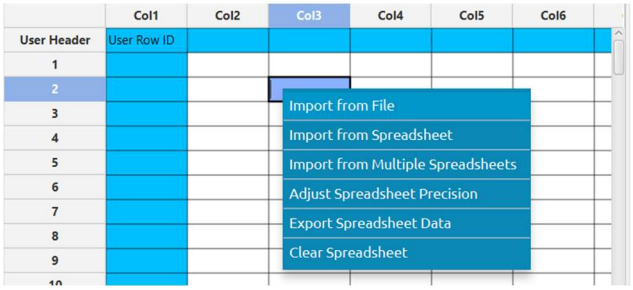
Tips Dataset

This dataset, which can be found in https://www.kaggle.com/datasets/sakshisatre/tips-dataset, contains information about tipping behavior from customers. It includes various attributes like the total bill amount, tip amount, gender, whether the customer smokes or not, the day of the week, time of day, and the size of the party. It contains 7 features and 244 samples and it is often used for demonstration and practice in data analysis and visualization.

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 vehicle data.

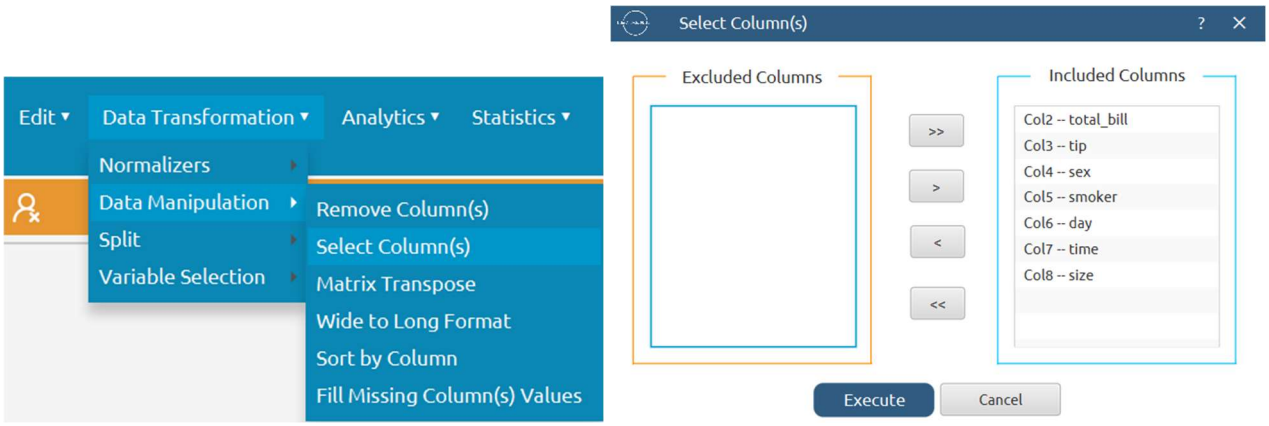


The data will appear on the left spreadsheet.

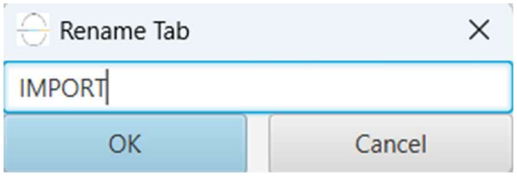
	Col1	Col2 (D)	Col3 (D)	Col4 (S)	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (I)
User Header	User Row ID	total_bill	tip	sex	smoker	day	time	size
1		16.99	1.01	Female	No	Sun	Dinner	2
2		10.34	1.66	Male	No	Sun	Dinner	3
3		21.01	3.5	Male	No	Sun	Dinner	3
4		23.68	3.31	Male	No	Sun	Dinner	2
5		24.59	3.61	Female	No	Sun	Dinner	4
6		25.29	4.71	Male	No	Sun	Dinner	4
7		8.77	2.0	Male	No	Sun	Dinner	2
8		26.88	3.12	Male	No	Sun	Dinner	4
9		15.04	1.96	Male	No	Sun	Dinner	2
10		14.78	3.23	Male	No	Sun	Dinner	2
11		10.27	1.71	Male	No	Sun	Dinner	2
12		35.26	5.0	Female	No	Sun	Dinner	4
13		15.42	1.57	Male	No	Sun	Dinner	2
14		18.43	3.0	Male	No	Sun	Dinner	4
15		14.83	3.02	Female	No	Sun	Dinner	2

Step 2: Manipulate data

In this dataset there are not any empty values, 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.



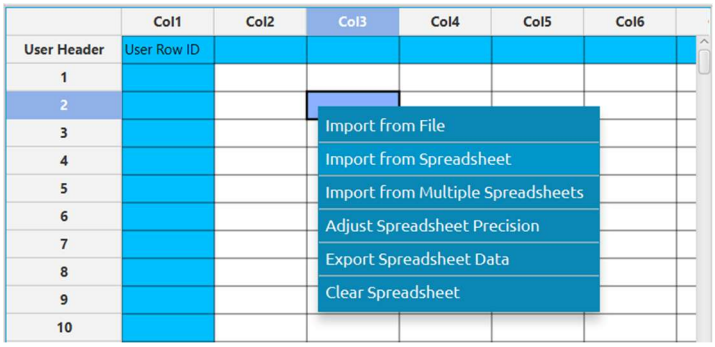
All of the data will appear in the output (right) spreadsheet. This tab can be renamed "IMPORT" by right-clicking on it and choosing the "Rename" option.



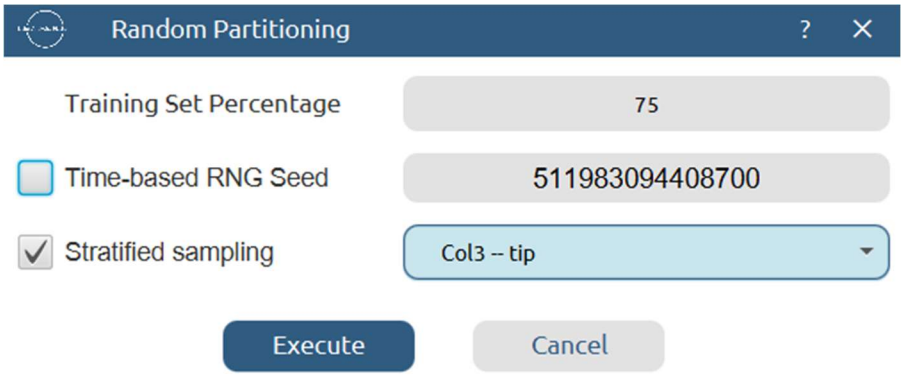
Step 3: Split data

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:



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.

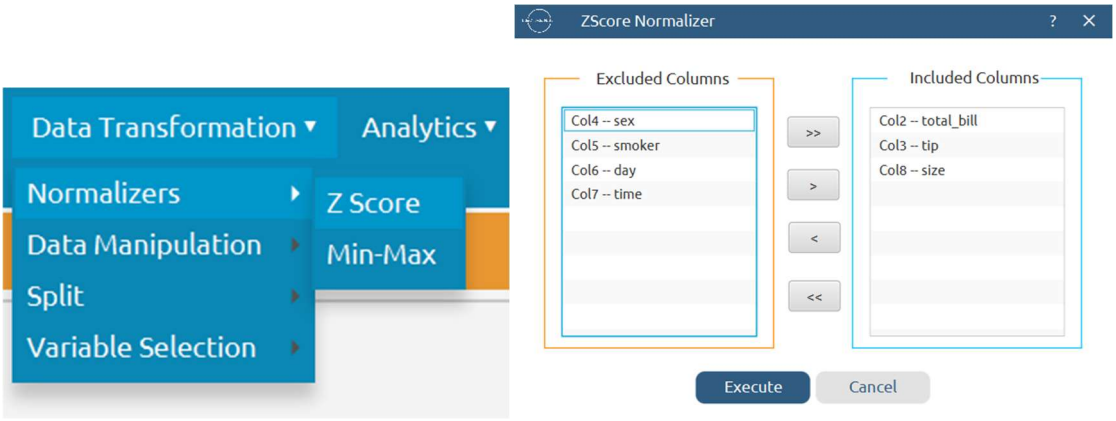
Step 4: Normalize the training set

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 (D)	Col3 (D)	Col4 (S)	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (I)
User Header	User Row ID	total_bill	tip	sex	smoker	day	time	size
1		16.99	1.01	Female	No	Sun	Dinner	2
2		10.34	1.66	Male	No	Sun	Dinner	3
3		23.68	3.31	Male	No	Sun	Dinner	2
4		24.59	3.61	Female	No	Sun	Dinner	4
5		25.29	4.71	Male	No	Sun	Dinner	4
6		15.04	1.96	Male	No	Sun	Dinner	2
7		14.78	3.23	Male	No	Sun	Dinner	2
8		10.27	1.71	Male	No	Sun	Dinner	2
9		15.42	1.57	Male	No	Sun	Dinner	2
10		14.83	3.02	Female	No	Sun	Dinner	2
11		10.33	1.67	Female	No	Sun	Dinner	3
12		16.97	3.5	Female	No	Sun	Dinner	3
13		17.92	4.08	Male	No	Sat	Dinner	2
14		20.29	2.75	Female	No	Sat	Dinner	2
15		19.82	3.18	Male	No	Sat	Dinner	2

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



The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (S)	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (D)
User Header	User Row ID	total_bill	tip	sex	smoker	day	time	size
1		-0.2992092	-1.4705782	Female	No	Sun	Dinner	-0.5737182
2		-1.0280668	-0.9724001	Male	No	Sun	Dinner	0.5674822
3		0.4340325	0.2922059	Male	No	Sun	Dinner	-0.5737182
4		0.5337709	0.5221342	Female	No	Sun	Dinner	1.7086825
5		0.6104927	1.3652049	Male	No	Sun	Dinner	1.7086825
6		-0.5129344	-0.7424718	Male	No	Sun	Dinner	-0.5737182
7		-0.5414311	0.2308916	Male	No	Sun	Dinner	-0.5737182
8		-1.0357390	-0.9340787	Male	No	Sun	Dinner	-0.5737182
9		-0.4712854	-1.0413786	Male	No	Sun	Dinner	-0.5737182
10		-0.5359509	0.0699418	Female	No	Sun	Dinner	-0.5737182
11		-1.0291628	-0.9647358	Female	No	Sun	Dinner	0.5674822
12		-0.3014013	0.4378272	Female	No	Sun	Dinner	0.5674822
13		-0.1972788	0.8823553	Male	No	Sat	Dinner	-0.5737182
14		0.0624795	-0.1369937	Female	No	Sat	Dinner	-0.5737182
15		0.0109663	0.1925702	Male	No	Sat	Dinner	-0.5737182

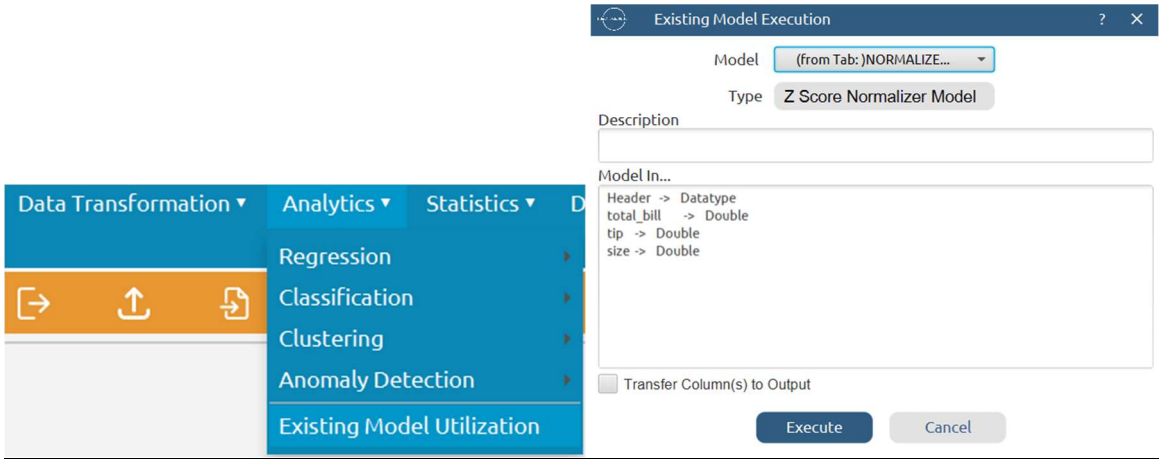
Step 5: Normalize the test set

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 (D)	Col3 (D)	Col4 (S)	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (I)
User Header	User Row ID	total_bill	tip	sex	smoker	day	time	size
1		21.01	3.5	Male	No	Sun	Dinner	3
2		8.77	2.0	Male	No	Sun	Dinner	2
3		26.88	3.12	Male	No	Sun	Dinner	4
4		35.26	5.0	Female	No	Sun	Dinner	4
5		18.43	3.0	Male	No	Sun	Dinner	4
6		21.58	3.92	Male	No	Sun	Dinner	2
7		16.29	3.71	Male	No	Sun	Dinner	3
8		20.65	3.35	Male	No	Sat	Dinner	3
9		15.77	2.23	Female	No	Sat	Dinner	2
10		39.42	7.58	Male	No	Sat	Dinner	4
11		17.46	2.54	Male	No	Sun	Dinner	2
12		9.68	1.32	Male	No	Sun	Dinner	2
13		32.4	6.0	Male	No	Sun	Dinner	4
14		18.04	3.0	Male	No	Sun	Dinner	2
15		25.56	4.34	Male	No	Sun	Dinner	4

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



The results will appear on the output spreadsheet.

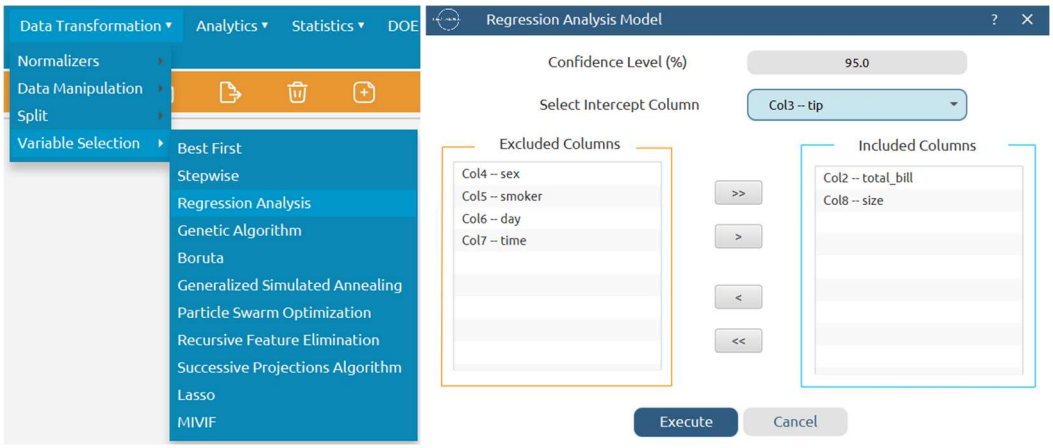
	Col1	Col2 (D)	Col3 (D)	Col4 (S)	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (D)
User Header	User Row ID	total_bill	tip	sex	smoker	day	time	size
1		0.1330918	0.3863679	Male	No	Sun	Dinner	0.4064131
2		-1.2126172	-0.7208248	Male	No	Sun	Dinner	-0.6560669
3		0.7784604	0.1058791	Male	No	Sun	Dinner	1.4688932
4		1.6997873	1.4935606	Female	No	Sun	Dinner	1.4688932
5		-0.1505626	0.0173037	Male	No	Sun	Dinner	1.4688932
6		0.1957596	0.6963819	Male	No	Sun	Dinner	-0.6560669
7		-0.3858418	0.5413749	Male	No	Sun	Dinner	0.4064131
8		0.0935121	0.2756486	Male	No	Sat	Dinner	0.4064131
9		-0.4430124	-0.5510552	Female	No	Sat	Dinner	-0.6560669
10		2.1571524	3.3979320	Male	No	Sat	Dinner	1.4688932
11		-0.2572078	-0.3222354	Male	No	Sun	Dinner	-0.6560669
12		-1.1125686	-1.2227521	Male	No	Sun	Dinner	-0.6560669
13		1.3853487	2.2316890	Male	No	Sun	Dinner	1.4688932
14		-0.1934406	0.0173037	Male	No	Sun	Dinner	-0.6560669
15		0.6333349	1.0063958	Male	No	Sun	Dinner	1.4688932

Step 6: Regression analysis

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

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

Perform regression analysis by choosing *Data Transformation → Variable Selection → Regression Analysis* and selecting "tip" as the intercept column.



The results will appear on the output spreadsheet.

	Col1	Col2 (S)	Col3 (S)	Col4 (S)	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (S)
User Header	User Row ID							
1		Regression Statistics						
2		Multiple R	0.6238992					
3		R Square	0.3892503					
4		Adjusted R Square	0.3824642					
5		Standard Error	0.7858345					
6		Observations	183					
7								
8			Degrees of Freedom	Sum of Squares	Mean Square	F-statistic	Significance F	
9		Regression	2	70.8435488	35.4217744	57.3598683	0E-7	
10		Residual	180	111.1564512	0.6175358			
11		Total	182	182.0000000				
12			Coefficients	Standard Error	t-statistic	P-value	Lower 95.0%	Upper 95.0%
13		tip	0E-7	0.0580906	0E-7	1.0000000	-0.1146261	0.1146261
14		total_bill	0.5773858	0.0708695	8.1471698	0E-7	0.4375439	0.7172277
15		size	0.0761378	0.0708695	1.0743381	0.2841092	-0.0637041	0.2159797

Step 7: 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 k-Nearest Neighbours (kNN) method to train and fit the model: *Analytics → Regression → k-Nearest Neighbors (kNN)*

The predictions will appear on the output spreadsheet.

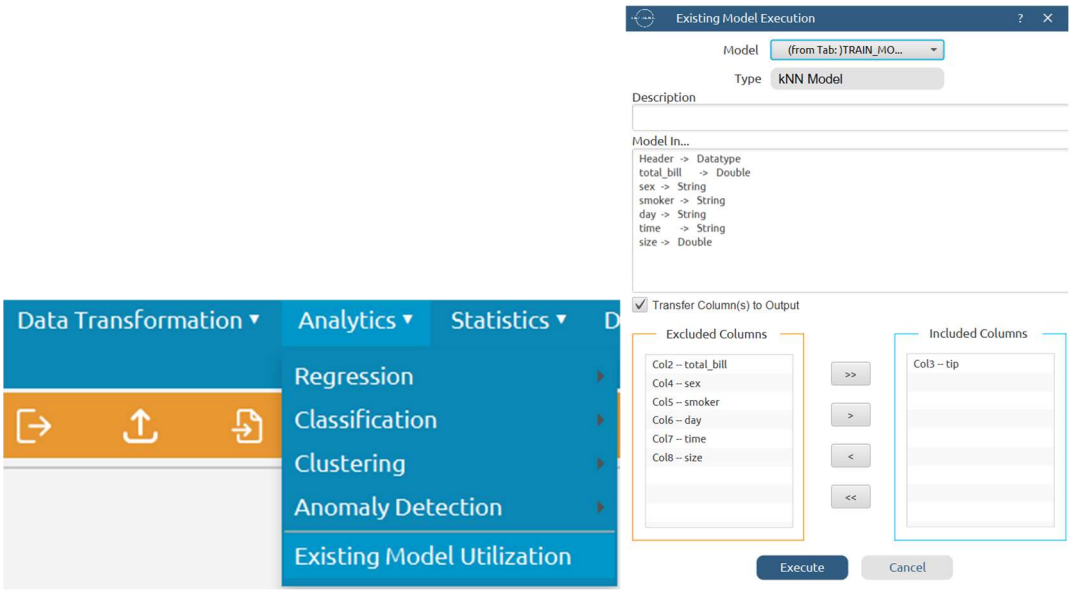
	Col1	Col2 (D)	Col3 (D)	Col4 (S)	Col5 (D)	Col6 (S)	Col7 (D)	Col8 (S)	Col9 (D)
User Header	User Row ID	tip	kNN Prediction	Closest NN1	Distance from NN1	Closest NN2	Distance from NN2	Closest NN3	Distance from NN3
1		-1.4705782	-1.2264836	Entry 1	0.0	Entry 87	0.0067030	Entry 10	0.0452451
2		-0.9724001	-0.9725886	Entry 2	0.0	Entry 8	0.2000054	Entry 40	0.2001754
3		0.2922059	0.2530970	Entry 3	0.0	Entry 86	0.0056556	Entry 35	0.0303729
4		0.5221342	0.5373350	Entry 4	0.0	Entry 114	0.0085882	Entry 39	0.2140762
5		1.3652049	1.3741545	Entry 5	0.0	Entry 88	0.0971931	Entry 33	0.1070381
6		-0.7424718	-0.6496052	Entry 6	0.0	Entry 7	0.0054462	Entry 9	0.0079598
7		0.2308916	0.0354606	Entry 7	0.0	Entry 6	0.0054462	Entry 9	0.0134059
8		-0.9340787	-0.9358731	Entry 8	0.0	Entry 40	0.0069124	Entry 37	0.0475492
9		-1.0413786	-0.9383524	Entry 9	0.0	Entry 6	0.0079598	Entry 7	0.0134059
10		0.0699418	0.0446042	Entry 10	0.0	Entry 1	0.0452451	Entry 87	0.0519481
11		-0.9647358	-0.9528671	Entry 11	0.0	Entry 117	0.1231672	Entry 12	0.1390867
12		0.4378272	0.3647109	Entry 12	0.0	Entry 117	0.0159196	Entry 11	0.1390867
13		0.8823553	0.4164619	Entry 13	0.0	Entry 182	0.0020947	Entry 25	0.0029326
14		-0.1369937	-0.1271339	Entry 14	0.0	Entry 20	0.0134059	Entry 50	0.0804357
15		0.1925702	0.2251059	Entry 15	0.0	Entry 82	0.0330959	Entry 19	0.0393800

Step 8: 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 "tip" to be transferred to the output spreadsheet.



The predictions will appear on the output spreadsheet.

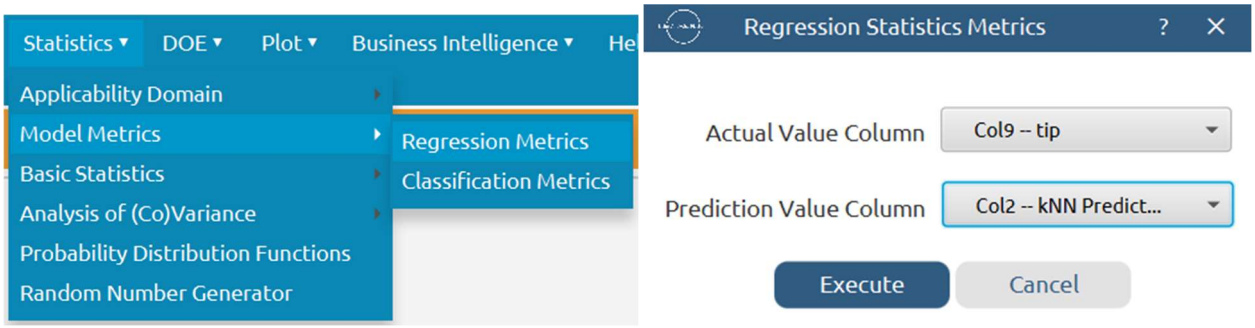
	Col1	Col2 (D)	Col3 (S)	Col4 (D)	Col5 (S)	Col6 (D)	Col7 (S)	Col8 (D)	Col9 (D)
User Header	User Row ID	kNN Prediction	Closest NN1	Distance from NN1	Closest NN2	Distance from NN2	Closest NN3	Distance from NN3	tip
1		0.3796556	Entry 3	0.0	Entry 124	0.0775519	Entry 115	0.0828546	0.3863679
2		-0.7576050	Entry 9	0.0331418	Entry 42	0.0832965	Entry 114	0.0963323	-0.7208248
3		0.1192800	Entry 6	0.0	Entry 44	0.0291648	Entry 116	0.0514803	0.1058791
4		1.4899810	Entry 10	0.0	Entry 120	0.2266902	Entry 118	0.2329974	1.4935606
5		-0.0214889	Entry 12	0.0	Entry 117	0.0296067	Entry 122	0.0428635	0.0173037
6		0.6847745	Entry 14	0.0	Entry 39	0.0143615	Entry 125	0.0181175	0.6963819
7		0.5096973	Entry 16	0.0	Entry 115	0.0214317	Entry 3	0.1042863	0.5413749
8		0.2499345	Entry 17	0.0	Entry 53	0.0125939	Entry 33	0.0433053	0.2756486
9		-0.5108578	Entry 19	0.0	Entry 54	0.0150243	Entry 28	0.0156871	-0.5510552
10		3.4000159	Entry 20	0.0	Entry 49	0.1955369	Entry 160	0.1968626	3.3979320
11		-0.3221637	Entry 36	0.0	Entry 45	0.0448520	Entry 11	0.0450729	-0.3222354
12		-0.8160529	Entry 9	0.0130358	Entry 42	0.0631905	Entry 114	0.0762262	-1.2227521
13		2.2135316	Entry 40	0.0	Entry 38	0.0441891	Entry 88	0.0545736	2.2316890
14		-0.2427139	Entry 36	0.0128148	Entry 45	0.0320371	Entry 11	0.0578878	0.0173037
15		0.9098070	Entry 44	0.0	Entry 116	0.0223155	Entry 6	0.0291648	1.0063958

Step 9: 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 regression: *Statistics → Model Metrics → Regression Metrics*



The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)
User Header	User Row ID	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R Squared
1		0.1459887	0.3820847	0.1992922	0.8969911

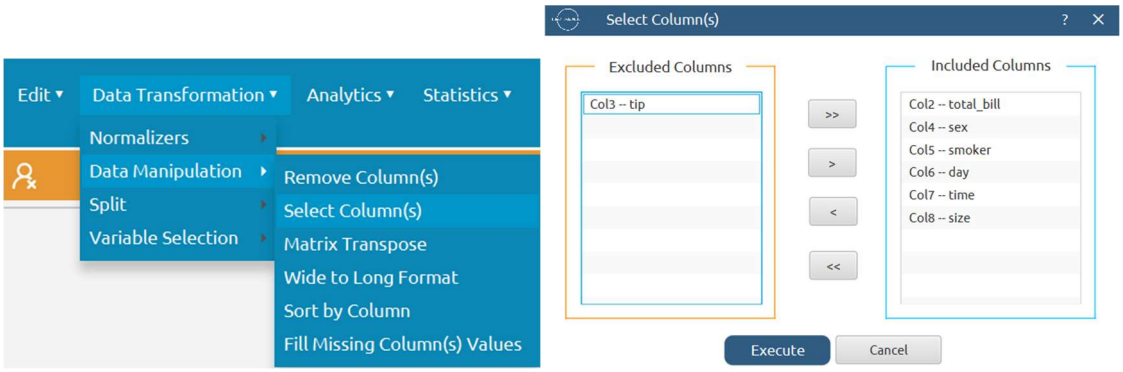
Step 10: Reliability check for each record of the test set

Step 11.a: Create the domain

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

Import data into the input spreadsheet of the "EXCLUDE_TIP" 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 "tip": *Data Transformation → Data Manipulation → Select Column(s)*

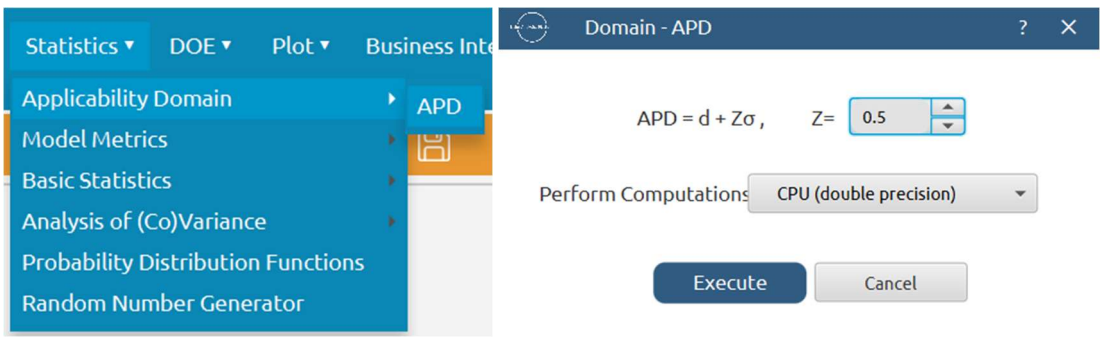


The results will appear on the output spreadsheet.

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_TIP" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

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



The results will appear on the output spreadsheet.

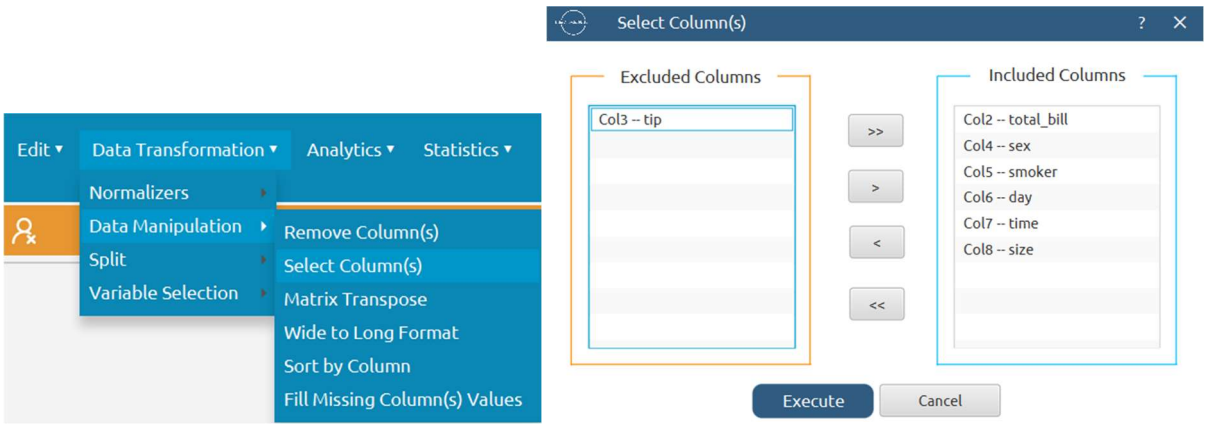
	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1		0.0	1.0305824	reliable
2		0.0	1.0305824	reliable
3		0.0	1.0305824	reliable
4		0.0	1.0305824	reliable
5		0.0	1.0305824	reliable
6		0.0	1.0305824	reliable
7		0.0	1.0305824	reliable
8		0.0	1.0305824	reliable
9		0.0	1.0305824	reliable
10		0.0	1.0305824	reliable
11		0.0	1.0305824	reliable
12		0.0	1.0305824	reliable
13		0.0	1.0305824	reliable
14		0.0	1.0305824	reliable
15		0.0	1.0305824	reliable

Step 11.b: Check the test set reliability

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

Import data into the input spreadsheet of the "EXCLUDE_TIP_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 "tip": *Data Transformation → Data Manipulation → Select Column(s)*

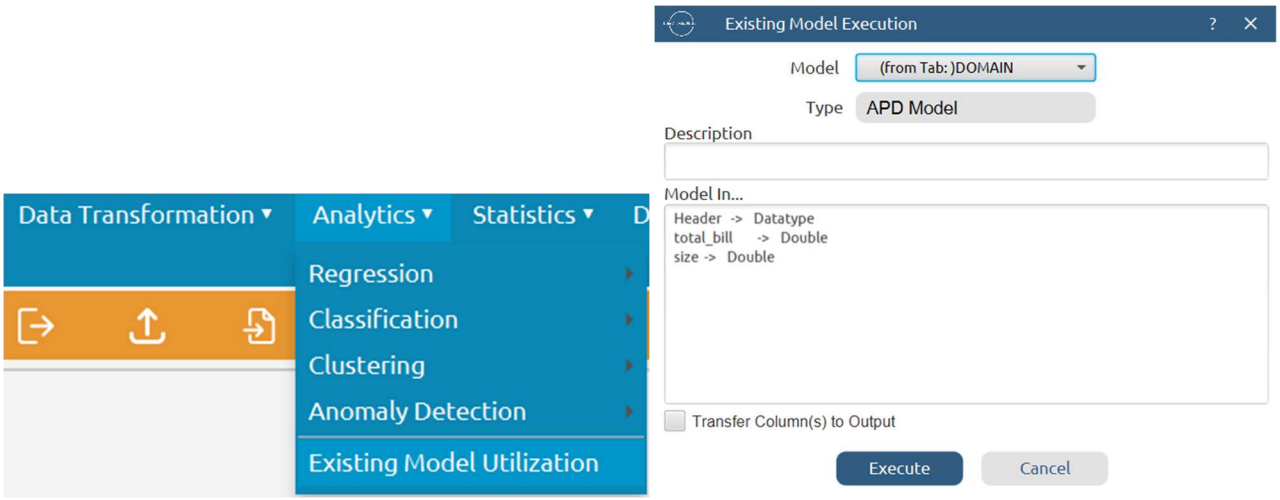


The results will appear on the output spreadsheet.

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_TIP_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*



The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1		0.1611128	1.0305824	reliable
2		0.0838929	1.0305824	reliable
3		0.2433325	1.0305824	reliable
4		0.2437362	1.0305824	reliable
5		0.2397897	1.0305824	reliable
6		0.0850487	1.0305824	reliable
7		0.1610731	1.0305824	reliable
8		0.1650050	1.0305824	reliable
9		0.0823593	1.0305824	reliable
10		0.2508305	1.0305824	reliable
11		0.0826397	1.0305824	reliable
12		0.0823751	1.0305824	reliable
13		0.3219195	1.0305824	reliable
14		0.0824381	1.0305824	reliable
15		0.2408748	1.0305824	reliable

Final Isalos Workflow

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