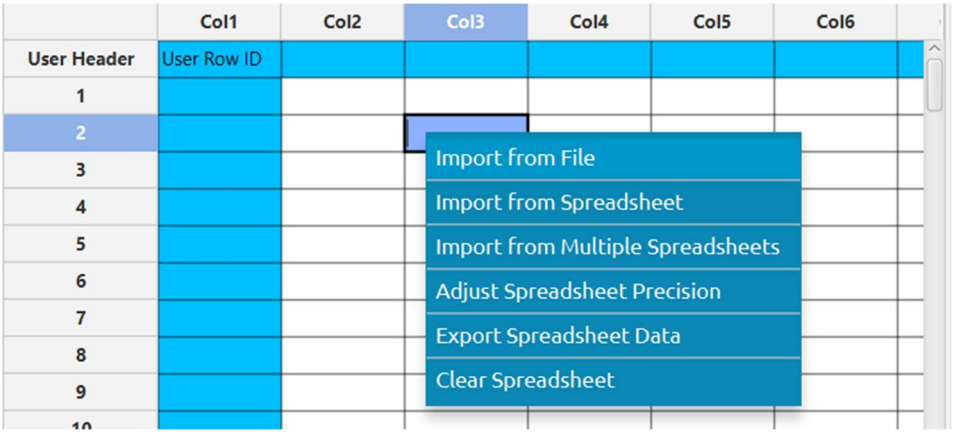
Breast Cancer Wisconsin (Diagnostic) Data set

This dataset contains, which can be found in https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data 31 features and 569 samples and it is used to predict whether a breast cancer diagnosis is malignant (M) or benign (B). The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass and they describe characteristics of the cell nuclei present in the image.

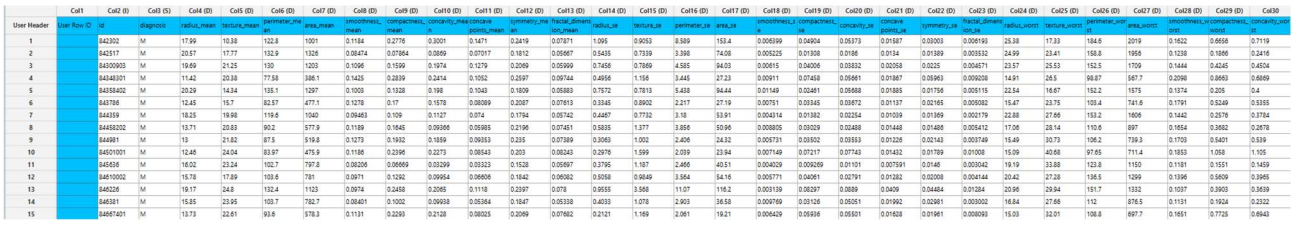
Isalos version used: 2.0.6

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 breast cancer data.

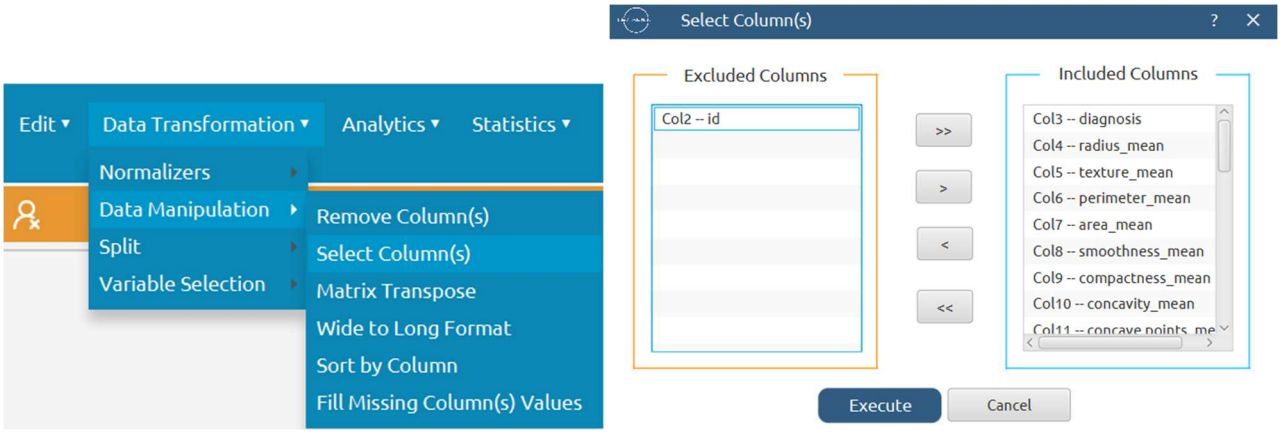


The data will appear on the left spreadsheet.

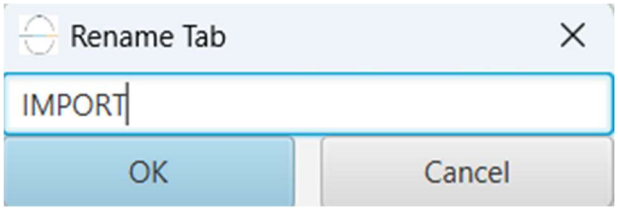


Step 2: Manipulate data

In our dataset there are not any empty values, so we can select all the columns to be used. However, since the column "id" does not offer any significant information about the breast cancer diagnosis we will exclude it. On the menu click on *Data Transformation → Data Manipulation → Select Column(s)* and select all columns except "id".



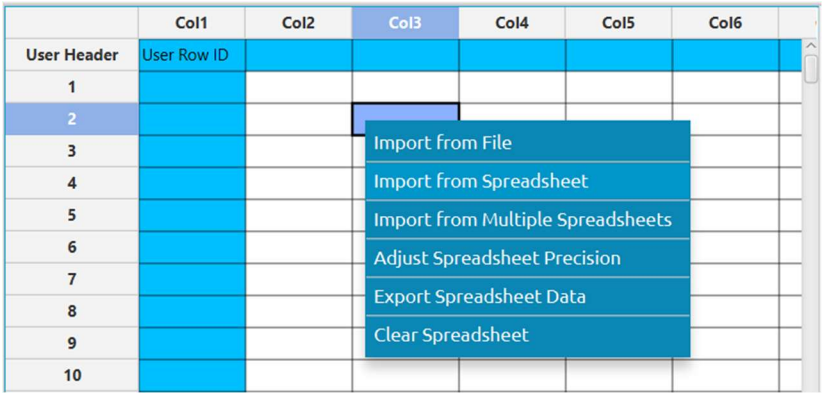
All of the data will appear in the output (right) spreadsheet. This tab can be renamed "IMPORT" by right-clicking on it and choosing the "Rename" option.



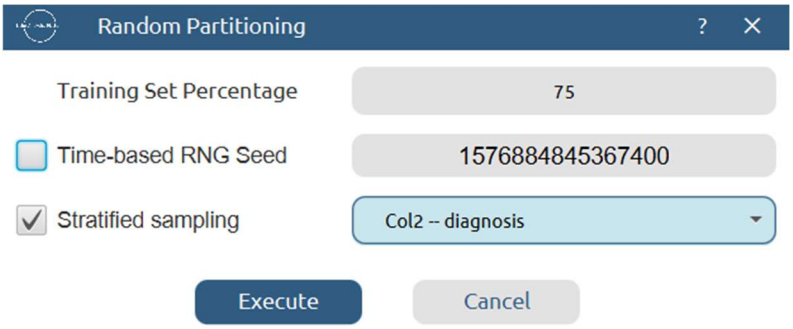
Step 3: Split data

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:

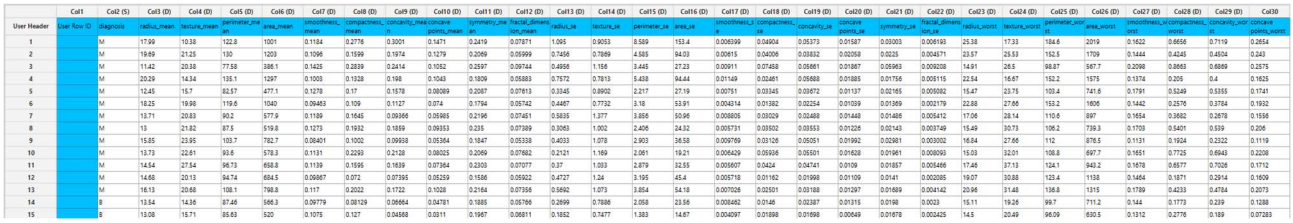


The results will be two separate spreadsheets, "TRAIN_TEST_SPLIT: Training Set" and "TRAIN_TEST_SPLIT: Test Set", which will be available to import into the next tabs.

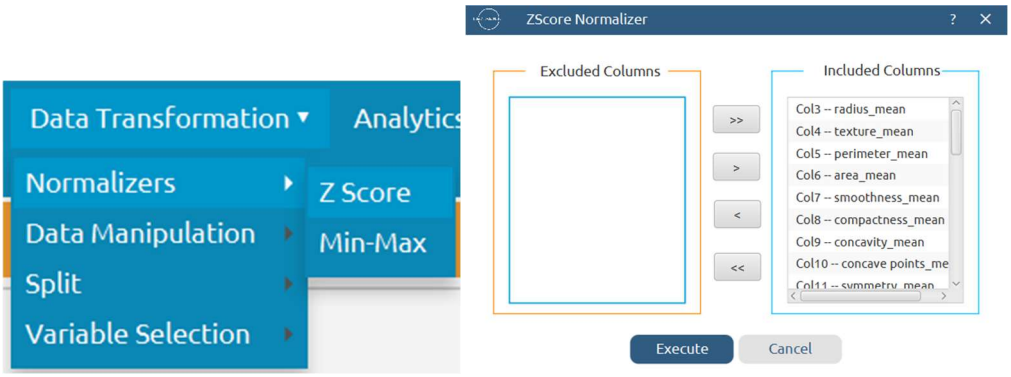
Step 4: Normalize the training set

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

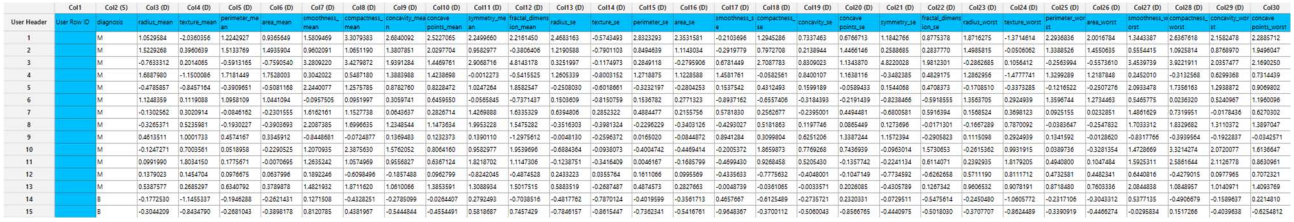
Import into the input spreadsheet of the "NORMALIZE_TRAIN_SET" tab the train set from the output of the "TRAIN_TEST_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet". From the available Select input tab options choose "TRAIN_TEST_SPLIT: Training Set".



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



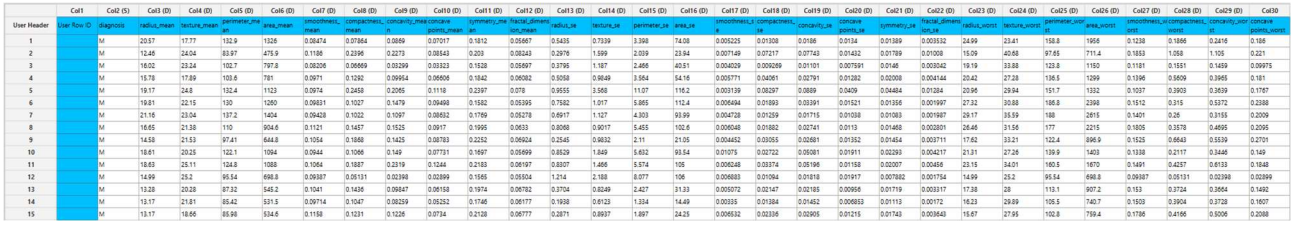
The results will appear on the output spreadsheet.



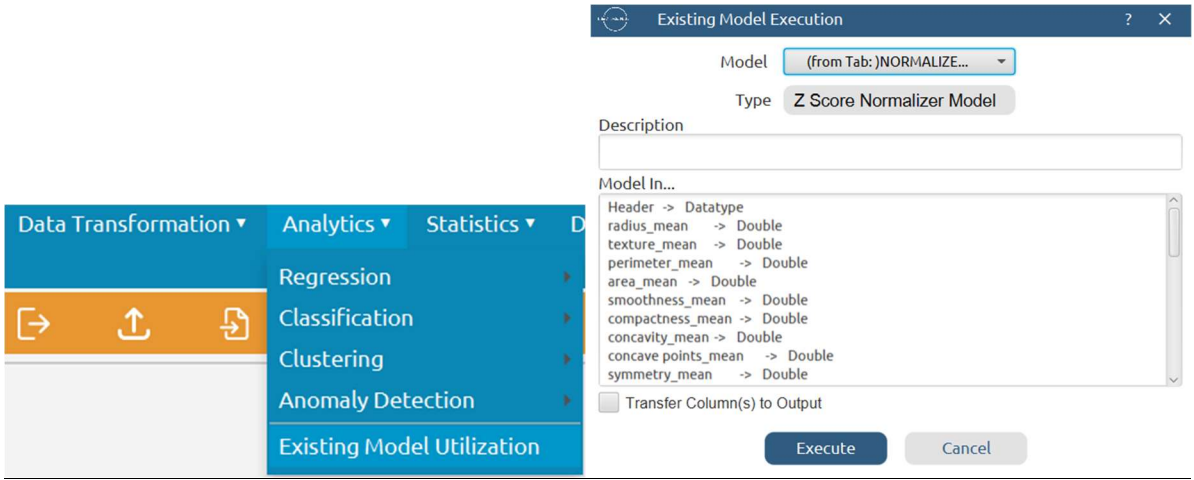
Step 5: Normalize the test set

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

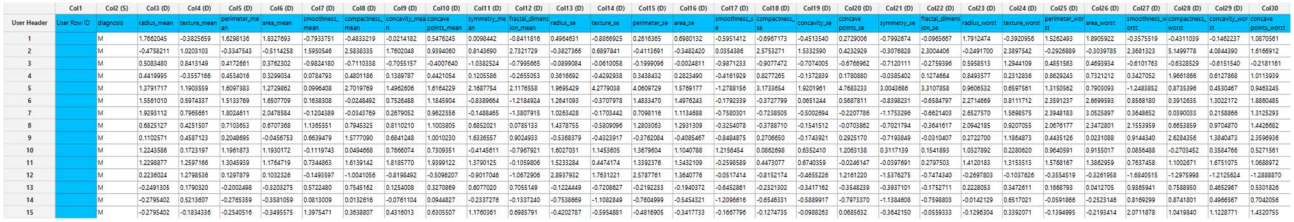
Import into the input spreadsheet of the "NORMALIZE_TEST_SET" tab the test set from the output of the "TRAIN_TEST_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet". From the available Select input tab options choose "TRAIN_TEST_SPLIT: Test Set".



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.

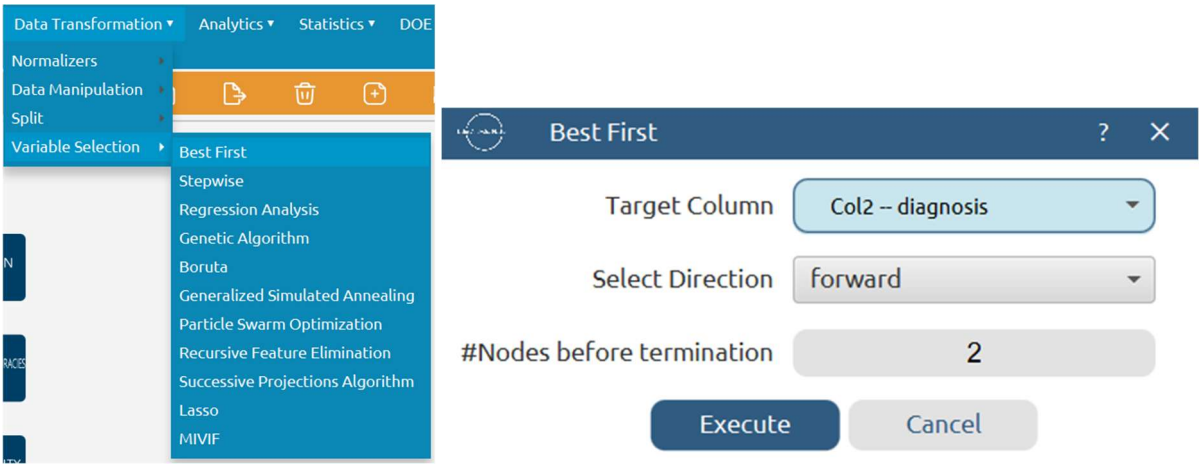


Step 6: Best First Algorithm

We want to choose the features that will be the most useful for predicting the diabetes outcome. Create a new tab by pressing the "+" button on the bottom of the page with the name "BEST_FIRST".

Import data into the input spreadsheet of the "BEST_FIRST" 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 best first algorithm by choosing: *Data Transformation → Variable Selection → Best First*



The results will appear on the output spreadsheet.

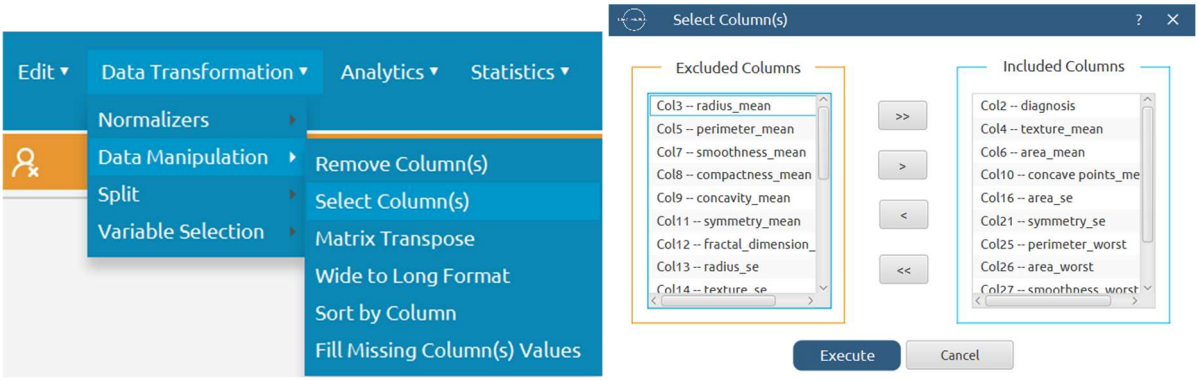
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (S)
User Header	User Row ID	texture_mean	area_mean	concave points_mean	area_se	symmetry_se	perimeter_worst	area_worst	smoothness_worst	concavity_worst	concave points_worst	diagnosis
1		-2.0360356	0.9365649	2.5227065	2.3531581	1.1842766	2.2936836	2.0016784	1.3443387	2.1582478	2.2885712	M
2		0.3960639	1.4935904	2.0297704	1.1143034	0.2588685	1.3388526	1.4550635	0.5554415	0.8768970	1.9496047	M
3		0.2014065	-0.7590540	1.4469761	-0.2795906	4.8220028	-0.2563994	-0.5573610	3.4539739	2.0357477	2.1690250	M
4		-1.1500086	1.7528003	1.4238698	1.1228588	-0.3482385	1.3299289	1.2187848	0.2452010	0.6299368	0.7314439	M
5		-0.8457164	-0.5081168	0.8228472	-0.2804253	0.1544068	-0.1216522	-0.2507276	2.0933478	1.2938872	0.9069802	M
6		0.1119088	1.0441094	0.6459550	0.2771323	-0.8238466	1.3596744	1.2734463	0.5465775	0.5240967	1.1960096	M
7		0.3020914	-0.2301555	0.2826714	0.2155756	-0.6800581	0.0925155	0.0232851	1.4861629	-0.0178436	0.6270302	M
8		0.5235981	-0.3903693	1.1473634	-0.3403126	0.1273696	-0.0383647	-0.2547832	1.7033312	1.3110372	1.3897047	M
9		1.0001733	0.3345912	0.1232373	-0.0844872	1.1572394	0.1341592	-0.0128620	-0.8317766	-0.1922837	-0.0342571	M
10		0.7003561	-0.2290525	0.8064160	-0.4469414	-0.0963014	0.0389736	-0.3281354	1.4728669	2.0720077	1.6136647	M
11		1.8034150	-0.0070695	0.6367124	-0.1685799	-0.2241134	0.4940800	0.1047484	1.5925311	2.1126778	0.8630961	M
12		0.1454704	0.0637996	0.0962799	0.0995569	-0.7734592	0.4732581	0.4482341	0.6440816	0.0977965	0.7072321	M
13		0.2685297	0.3789878	1.3853591	0.2827663	-0.4305789	0.8718480	0.7603336	2.0844838	1.0140971	1.4093769	M
14		-1.1455337	-0.2621431	-0.0264407	-0.3561713	-0.0729511	-0.2317106	-0.3043312	0.5377135	-0.1589637	0.2214810	B
15		-0.8434790	-0.3898178	-0.4554491	-0.5416761	-0.4440975	-0.3390919	-0.4466274	-0.0295834	-0.4039638	-0.6254812	B

Step 7: Feature Selection: Test set

We need to select the features of the test set that the best first algorithm indicated. Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION".

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

Select the columns that correspond to the important features: *Data Transformation → Data Manipulation → Select Column(s)*



The results will appear on the output spreadsheet.

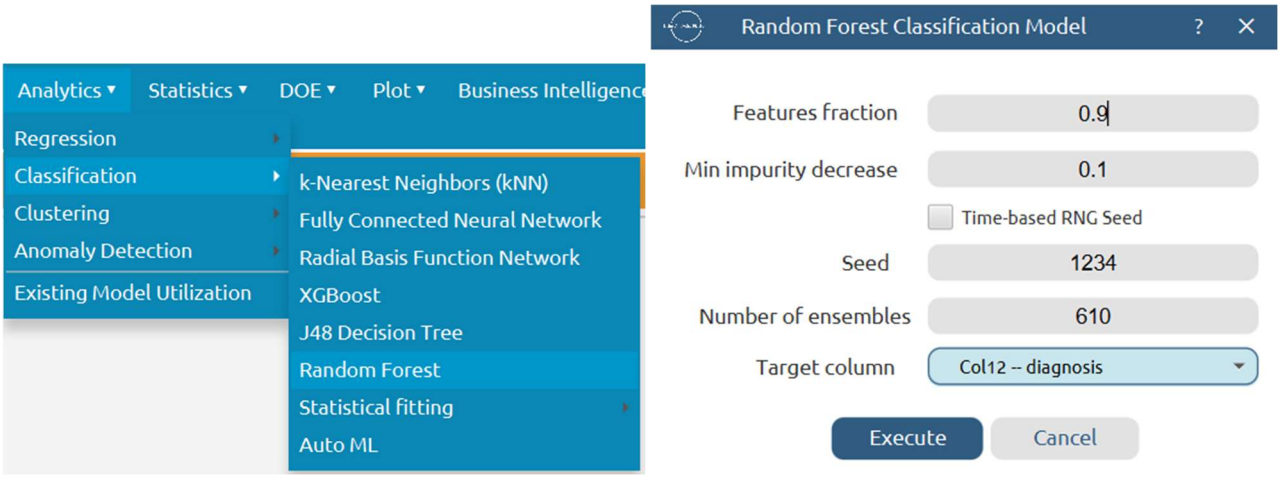
	Col1	Col2 (S)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)
User Header	User Row ID	diagnosis	texture_mean	area_mean	concave points_mean	area_se	symmetry_se	perimeter_worst	area_worst	smoothness_worst	concavity_worst	concave points_worst
1		M	-0.3825659	1.8327693	0.5476245	0.6980132	-0.7992674	1.5262493	1.8905922	-0.3575519	-0.1462237	1.0870561
2		M	1.0203103	-0.5114258	0.9394060	-0.3482420	-0.3076828	-0.2926889	-0.3039785	2.3681323	4.0844390	1.6166912
3		M	0.8413149	0.3762302	-0.4007640	-0.0024811	-0.7120111	0.4851563	0.4693934	-0.6101763	-0.6151540	-0.2181161
4		M	-0.3557166	0.3299034	0.4421054	0.2823490	-0.0385402	0.8629243	0.7321212	0.3427052	0.6127868	1.0113939
5		M	1.1903559	1.2729862	1.6164229	1.5769177	3.0043686	1.3150562	0.7903093	-1.2483852	0.4530467	0.9463245
6		M	0.5974337	1.6507709	1.1845904	1.4976243	-0.8398231	2.3591237	2.6699593	0.8568180	1.3022172	1.8860485
7		M	0.7965661	2.0478584	0.9622556	1.1134688	-1.1753296	2.3948183	3.0525897	0.3648652	0.2158866	1.3125293
8		M	0.4251507	0.6707368	1.1003805	1.2931309	-0.7021794	2.0676177	2.3472801	2.1553959	0.9704870	1.4426682
9		M	0.4587123	-0.0456753	1.0010230	-0.4085467	-0.7193849	0.4435126	0.0231088	0.9144340	1.3840473	2.3596936
10		M	0.1723197	1.1930172	0.7309351	1.1040788	0.3117139	0.9640591	0.9155017	0.0856488	0.3584766	0.5271561
11		M	1.2597166	1.1764719	1.9399122	1.3432109	-0.0397691	1.5768167	1.3862959	0.7637458	1.6751075	1.0688972
12		M	1.2798536	0.1032326	-0.5096207	1.3640776	-1.5376275	-0.3554519	-0.3261958	-1.6840515	-1.2125624	-1.2888870
13		M	0.1790320	-0.3203275	0.3270869	-0.1940372	-0.3937101	0.1668793	0.0412705	0.9365941	0.4652967	0.5301826
14		M	0.5213607	-0.3581059	0.0944827	-0.5454321	-1.1384608	-0.0591866	-0.2523146	0.8169299	0.4966567	0.7042056
15		M	-0.1834336	-0.3495575	0.6305507	-0.3417733	-0.3642150	-0.1394995	-0.2193414	2.0711878	1.1228771	1.4320755

Step 8: 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 "BEST_FIRST" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

Use the Random Forest method to train and fit the model: *Analytics → Classification → Random Forest*



The predictions will appear on the output spreadsheet.

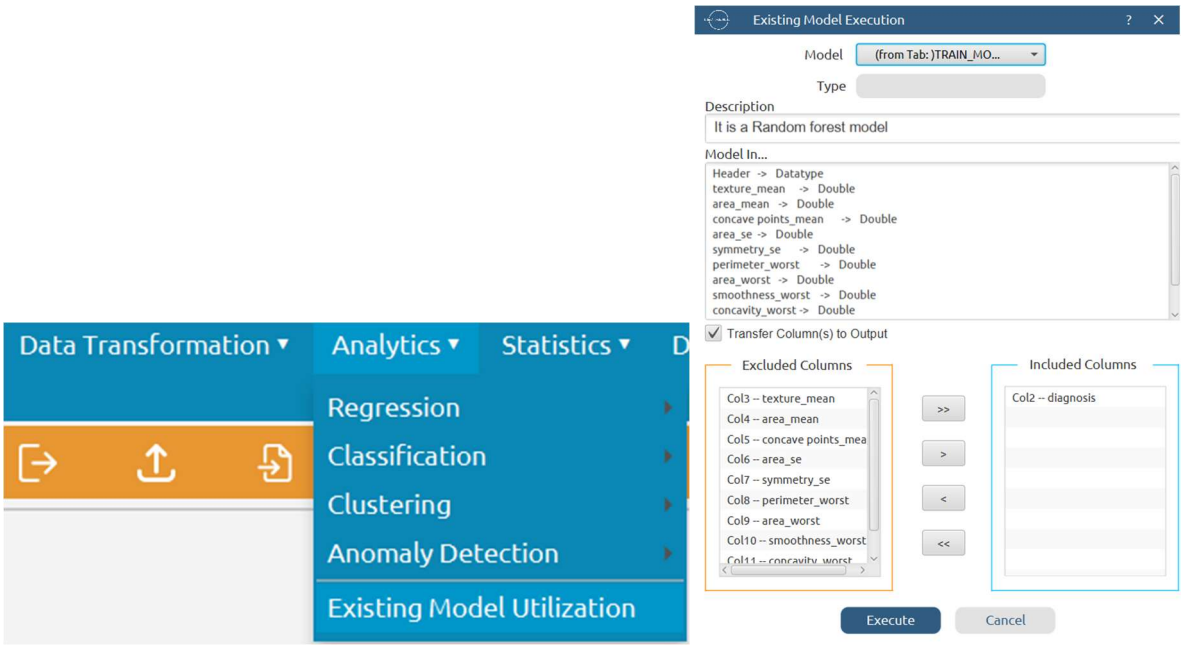
	Col1	Col2 (S)	Col3 (S)
User Header	User Row ID	diagnosis	Prediction
1		M	M
2		M	M
3		M	B
4		M	M
5		M	B
6		M	M
7		M	M
8		M	B
9		M	M
10		M	B
11		M	M
12		M	M
13		M	M
14		B	B
15		B	B

Step 9: 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 "FEATURE_SELECTION" 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 "diagnosis" to be transferred to the output spreadsheet.



The predictions will appear on the output spreadsheet.

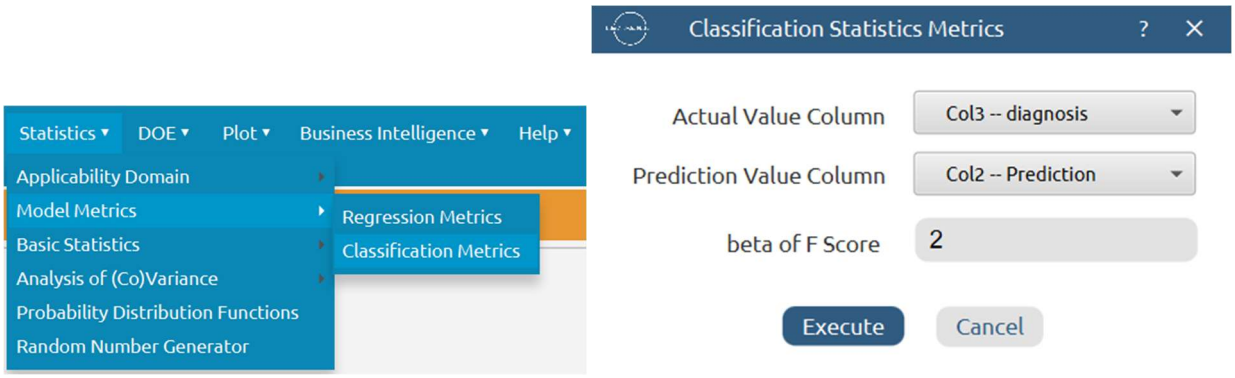
	Col1	Col2 (S)	Col3 (S)
User Header	User Row ID	Prediction	diagnosis
1		M	M
2		B	M
3		M	M
4		M	M
5		M	M
6		M	M
7		M	M
8		M	M
9		M	M
10		M	M
11		M	M
12		B	M
13		M	M
14		B	M
15		B	M

Step 10: 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*



The results will appear on the output spreadsheet.

	Col1 (S)	Col2 (S)	Col3 (S)	Col4 (S)
User Header	User Row ID			
1			Predicted Class	Predicted Class
2			M	B
3	Actual Class	M	42	11
4	Actual Class	B	4	85
5				
6				
7	Classification Accuracy	0.8943662		
8				
9	Precision		0.9130435	0.8854167
10				
11	Recall/Sensitivity		0.7924528	0.9550562
12				
13	Specificity		0.9550562	0.7924528
14				
15	F1 Score		0.8484848	0.9189189
16				
17	F (beta=2)		0.8139535	0.9402655
18				
19	MCC	0.7725647		

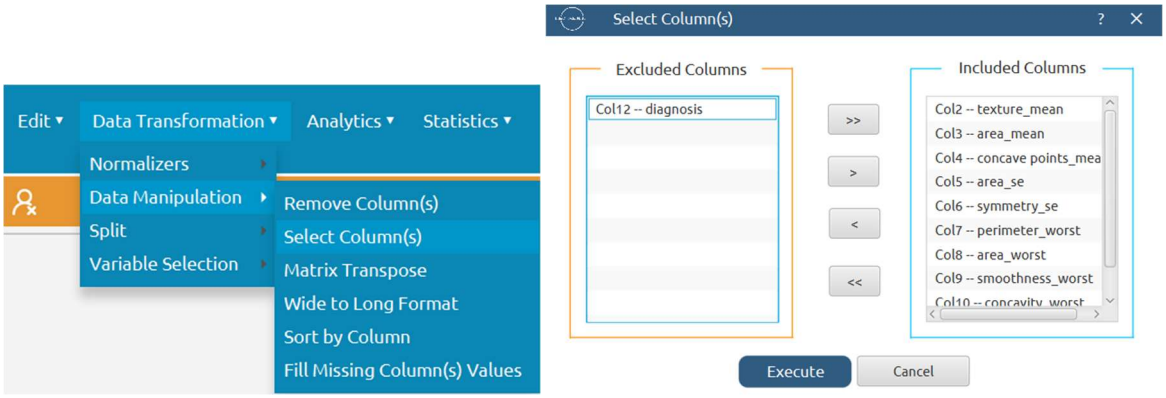
Step 11: Reliability check for each record of the test set

Step 11.a: Create the domain

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

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

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

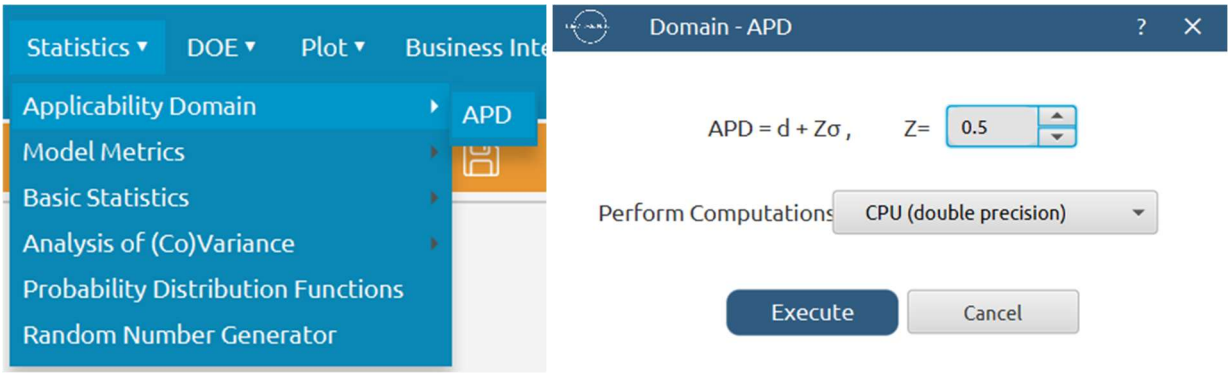


The results will appear on the output spreadsheet.

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_DIAGNOSIS" 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.

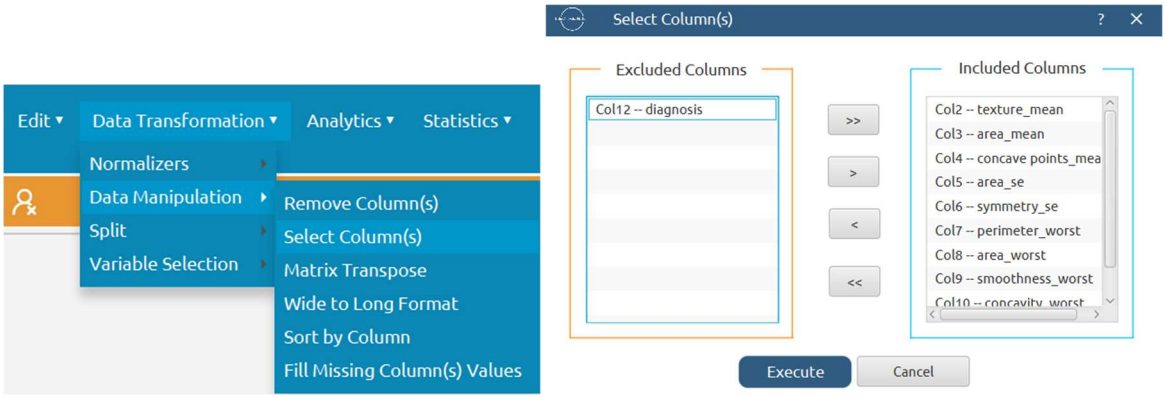
	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1		0.0	2.9702124	reliable
2		0.0	2.9702124	reliable
3		0.0	2.9702124	reliable
4		0.0	2.9702124	reliable
5		0.0	2.9702124	reliable
6		0.0	2.9702124	reliable
7		0.0	2.9702124	reliable
8		0.0	2.9702124	reliable
9		0.0	2.9702124	reliable
10		0.0	2.9702124	reliable
11		0.0	2.9702124	reliable
12		0.0	2.9702124	reliable
13		0.0	2.9702124	reliable
14		0.0	2.9702124	reliable
15		0.0	2.9702124	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_DIAGNOSIS_TEST_SET".

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

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

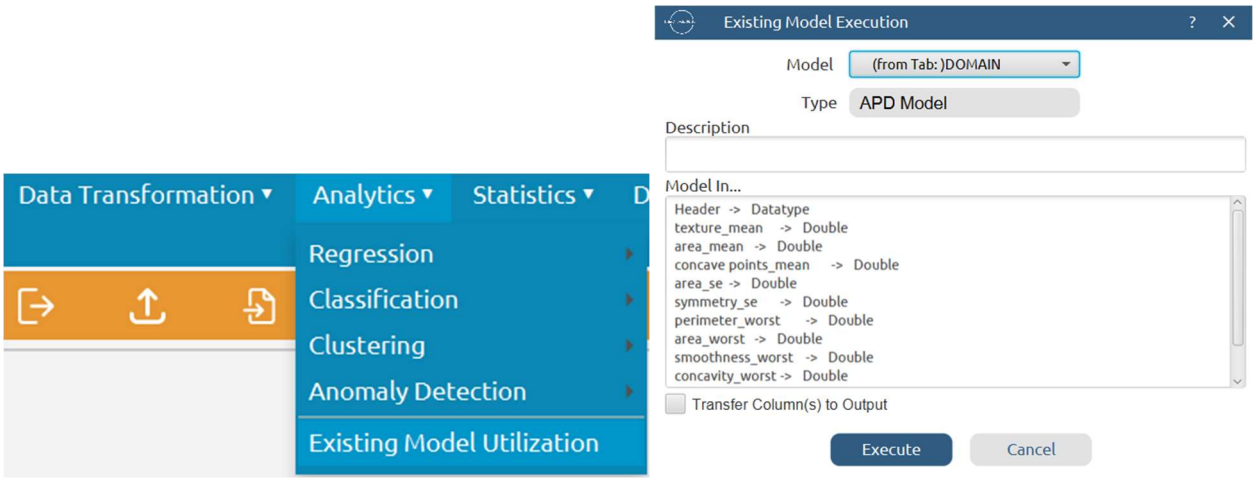


The results will appear on the output spreadsheet.

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_DIAGNOSIS_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		1.0748767	2.9702124	reliable
2		2.2838975	2.9702124	reliable
3		0.5921315	2.9702124	reliable
4		0.9078477	2.9702124	reliable
5		1.8209589	2.9702124	reliable
6		1.5301520	2.9702124	reliable
7		1.7180436	2.9702124	reliable
8		1.3950809	2.9702124	reliable
9		1.1623175	2.9702124	reliable
10		0.9187315	2.9702124	reliable
11		1.4637847	2.9702124	reliable
12		2.1716145	2.9702124	reliable
13		0.6998135	2.9702124	reliable
14		0.8731378	2.9702124	reliable
15		1.0393431	2.9702124	reliable

Final Isalos Workflow

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