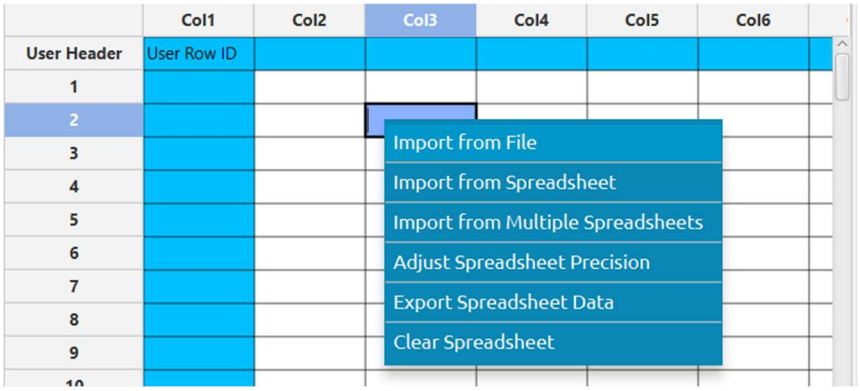
Wine Quality Classification Dataset

The dataset, which can be found in https://archive.ics.uci.edu/dataset/186/wine+quality, includes red and white variants of the Portuguese Vinho Verde wine. It contains 11 physicochemical attributes (such as acidity, sugar, pH) as inputs, and a sensory quality score as the output. The data does not include private or brand-specific information (e.g., grape type, price, or brand name). The main task is a binary classification problem, aiming to categorize each wine as good (quality ≥ 7) = G or bad (quality < 7) = B. The original dataset includes a numerical quality score which was transformed into a categorical score in Microsoft Excel, with the addition of the column "binary_quality". The dataset is imbalanced, with significantly more "average" wines than excellent or poor ones.

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 wine quality data.

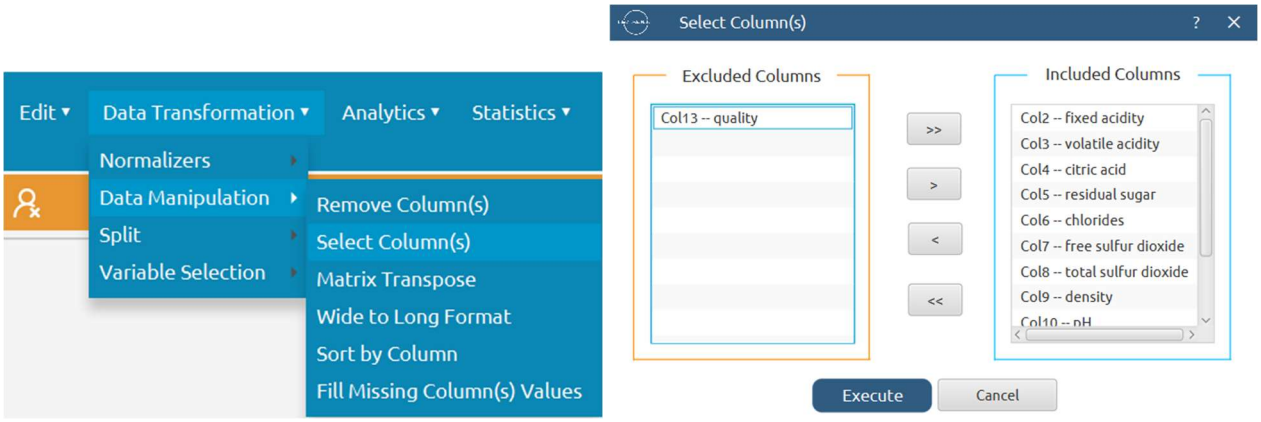


The data will appear on the left spreadsheet.

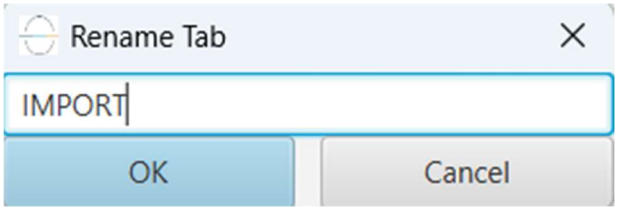
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)	Col13 (S)
User Header	User Row ID	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	binary_quality
1		7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	B
2		6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	B
3		8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	B
4		7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	B
5		7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	B
6		8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	B
7		6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	B
8		7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	B
9		6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	B
10		8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	B
11		8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	B
12		8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	B
13		7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	B
14		6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	G
15		8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	B
16		6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	G
17		6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24	0.36	9.6	B
18		6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	G
19		7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12	0.53	11.3	B
20		6.5	0.31	0.14	7.5	0.044	34	133	0.9955	3.22	0.5	9.5	B

Step 2: Manipulate data

In our 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, except the column “quality”, which is the numerical quality score.



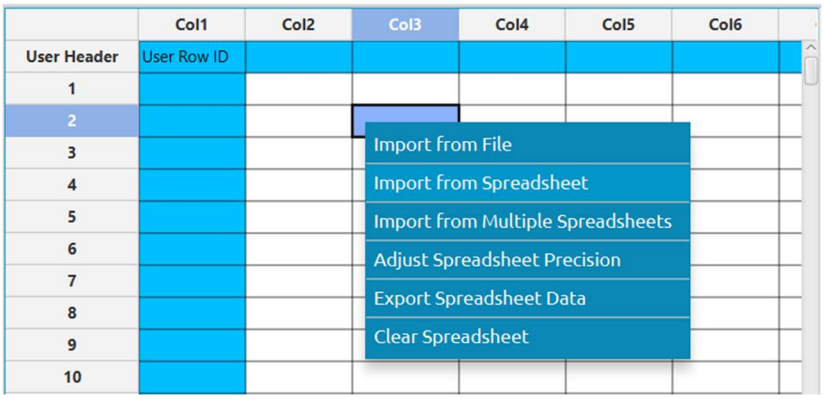
All of the data will appear in the output (right) spreadsheet. This tab can be renamed “IMPORT” by right-clicking on it and choosing the “Rename” option.



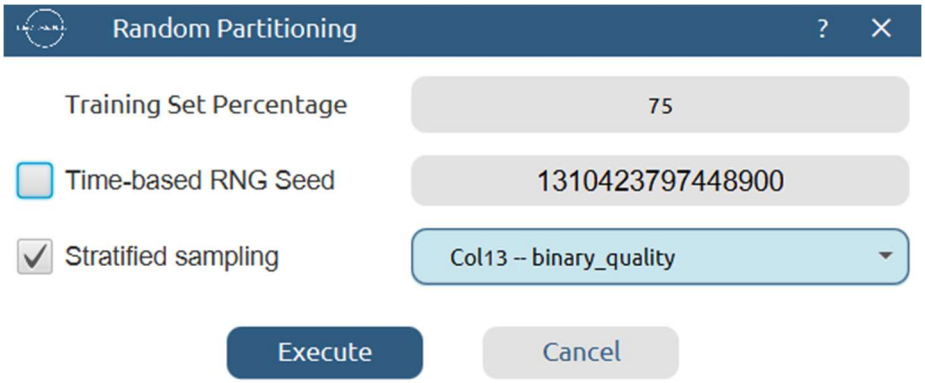
Step 3: Split data

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.

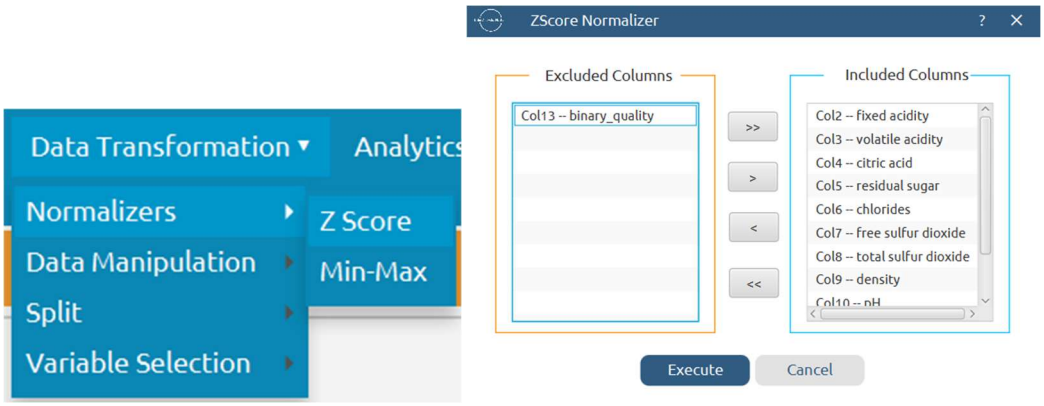
Step 4: Normalize the training set

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 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)	Col13 (S)
User Header	User Row ID	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	binary_quality
1		7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	B
2		6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	B
3		8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	B
4		7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	B
5		7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	B
6		8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	B
7		7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	B
8		6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	B
9		8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	B
10		7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	B
11		8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	B
12		6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	G
13		6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24	0.36	9.6	B
14		6.5	0.31	0.14	7.5	0.044	34	133	0.9955	3.22	0.5	9.5	B
15		6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	G

Normalize the data using Z-score: *Data Transformation → Normalizers → Z Score* and select all columns.



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 (D)	Col11 (D)	Col12 (D)	Col13 (S)
User Header	User Row ID	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	binary_quality
1		0.1967693	-0.0844381	0.2120661	2.7883490	-0.0358488	0.5706472	0.7579506	2.3056169	-1.2252979	-0.3506010	-1.4070922	B
2		-0.6403018	0.2116728	0.0471532	-0.9413659	0.1451318	-1.2633011	-0.1382244	0.0000359	0.7519844	-0.0005241	-0.8381853	B
3		1.5121669	0.0142655	0.5418918	0.0935812	0.1903769	-0.3167471	-0.9636487	0.3623415	0.4883467	-0.4381203	-0.3505509	B
4		0.4359325	-0.4792525	-0.1177596	0.4060180	0.5523382	0.6889664	1.1352874	0.5270258	0.0269809	-0.7881972	-0.5130957	B
5		0.4359325	-0.4792525	-0.1177596	0.4060180	0.5523382	0.6889664	1.1352874	0.5270258	0.0269809	-0.7881972	-0.5130957	B
6		1.5121669	0.0142655	0.5418918	0.0935812	0.1903769	-0.3167471	-0.9636487	0.3623415	0.4883467	-0.4381203	-0.3505509	B
7		0.1967693	-0.0844381	0.2120661	2.7883490	-0.0358488	0.5706472	0.7579506	2.3056169	-1.2252979	-0.3506010	-1.4070922	B
8		-0.6403018	0.2116728	0.0471532	-0.9413659	0.1451318	-1.2633011	-0.1382244	0.0000359	0.7519844	-0.0005241	-0.8381853	B
9		2.1100749	-0.4792525	0.5418918	-0.4336560	-0.4883004	-1.0858222	-0.6806461	0.2305940	-0.3025662	0.3495529	-0.6756405	B
10		1.2730037	-0.9727705	0.2945225	-1.0194751	-0.2620746	-1.1449818	-1.4824869	-0.6587016	-0.0389285	1.2247453	0.2183560	B
11		1.7513301	1.3961160	2.3559332	2.5052031	-0.2620746	0.3340087	0.8051177	2.0421219	-1.3571167	1.5748223	-0.6756405	B
12		-0.2815570	-1.0714741	0.3769789	-0.9608932	-0.6240358	-0.4350664	-0.6098954	-0.8563228	0.4224373	0.5245914	0.7059905	G
13		-0.6403018	1.9883377	-2.4265396	-1.0390024	0.0093963	-0.3167471	-0.9164816	-0.3952066	0.3565279	-1.1382742	-0.7569129	B
14		-0.4011386	0.3103764	-1.6019753	0.2107450	-0.0810940	-0.0801086	-0.1146408	0.4940889	0.2247091	0.0869952	-0.8381853	B
15		-0.7598834	3.7650026	1.2015432	-1.0194751	-0.7597713	-0.3759067	-1.4824869	-1.5809340	0.9497126	-0.8757165	1.8438042	G

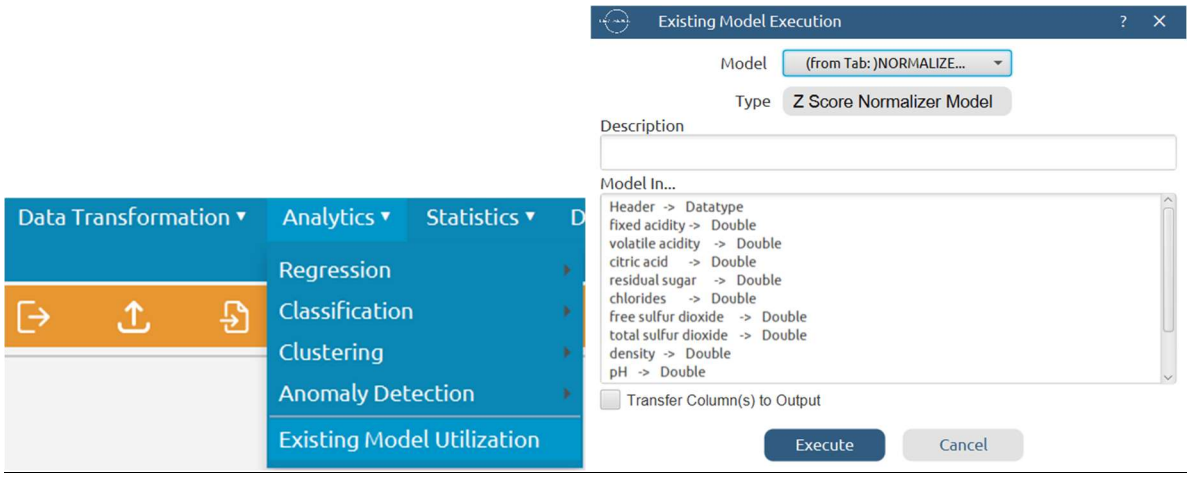
Step 5: Normalize the test set

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 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)	Col13 (S)
User Header	User Row ID	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	binary_quality
1		6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	B
2		8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	B
3		8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	B
4		6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	G
5		6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	G
6		7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12	0.53	11.3	B
7		6.9	0.24	0.35	1	0.052	35	146	0.993	3.45	0.44	10	B
8		7.4	0.25	0.36	2.05	0.05	31	100	0.992	3.19	0.44	10.8	B
9		6.2	0.12	0.34	1.5	0.045	43	117	0.9939	3.42	0.51	9	B
10		6.5	0.39	0.23	5.4	0.051	25	149	0.9934	3.24	0.35	10	B
11		7.3	0.24	0.39	17.95	0.057	45	149	0.9999	3.21	0.36	8.6	B
12		7.3	0.24	0.39	17.95	0.057	45	149	0.9999	3.21	0.36	8.6	B
13		7.2	0.19	0.31	1.6	0.062	31	173	0.9917	3.35	0.44	11.7	B
14		6	0.19	0.26	12.4	0.048	50	147	0.9972	3.3	0.36	8.9	B
15		6.6	0.38	0.15	4.6	0.044	25	78	0.9931	3.11	0.38	10.2	B

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.

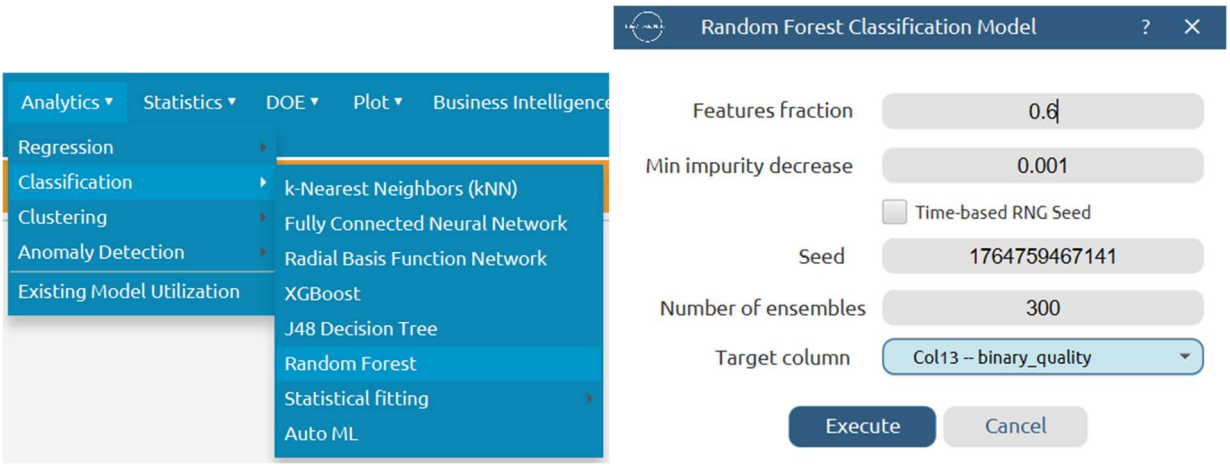
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)	Col13 (S)
User Header	User Row ID	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	binary_quality
1		-0.7598834	0.4090800	-1.4370625	0.1131085	-0.0358488	-0.3167471	-0.0438902	0.2964677	-0.0389285	-0.1755625	-0.7569129	B
2		1.5121669	-0.5779561	0.7892611	-0.9608932	-0.0810940	-0.4350664	-0.2089751	-0.0658379	0.2247091	-0.3506010	0.3809008	B
3		1.5121669	-0.0844381	0.6243482	-0.9706569	-0.5787907	-1.4407799	-1.7654895	-1.0539441	-1.2912073	0.6121106	1.1936249	B
4		-0.2815570	-1.1701777	0.5418918	-0.9608932	-0.0810940	0.7481261	0.1211947	-0.9221966	2.3338102	0.2620337	1.5187146	G
5		-0.7598834	3.7650026	1.2015432	-1.0194751	-0.7597713	-0.3759067	-1.4824869	-1.5809340	0.9497126	-0.8757165	1.8438042	G
6		0.6750957	0.6064872	0.7068047	-1.0390024	-0.5787907	-1.0858222	0.7815341	-0.7575122	-0.4343850	0.3495529	0.6247181	B
7		0.0771877	-0.3805489	0.1296097	-1.0585297	0.2808672	-0.0209490	0.1919453	-0.3293329	1.7406255	-0.4381203	-0.4318233	B
8		0.6750957	-0.2818453	0.2120661	-0.8534930	0.1903769	-0.2575875	-0.8928981	-0.6587016	0.0269809	-0.4381203	0.2183560	B
9		-0.7598834	-1.5649921	0.0471532	-0.9608932	-0.0358488	0.4523280	-0.4919777	-0.0329010	1.5428973	0.1745144	-1.2445474	B
10		-0.4011386	1.1000052	-0.8598675	-0.1993284	0.2356221	-0.6125452	0.2626960	-0.1975854	0.3565279	-1.2257934	-0.4318233	B
11		0.5555141	-0.3805489	0.4594354	2.2513482	0.5070930	0.5706472	0.2626960	1.9433113	0.1587997	-1.1382742	-1.5696370	B
12		0.5555141	-0.3805489	0.4594354	2.2513482	0.5070930	0.5706472	0.2626960	1.9433113	0.1587997	-1.1382742	-1.5696370	B
13		0.4359325	-0.8740669	-0.2002161	-0.9413659	0.7333188	-0.2575875	0.8287013	-0.7575122	1.0815314	-0.4381203	0.9498077	B
14		-0.9990466	-0.8740669	-0.6124982	1.1675829	0.0998866	0.8664453	0.2155289	1.0540158	0.7519844	-1.1382742	-1.3258198	B
15		-0.2815570	1.0013016	-1.5195189	-0.3555468	-0.0810940	-0.6125452	-1.4117362	-0.2963960	-0.5002944	-0.9632357	-0.2692785	B

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



The predictions will appear on the output spreadsheet.

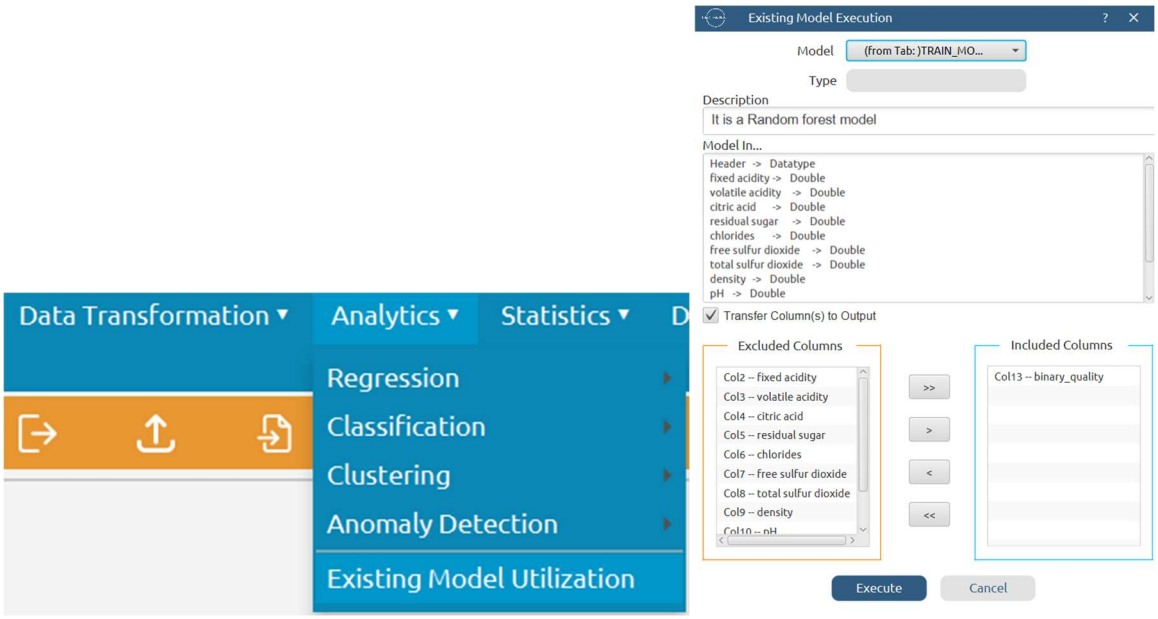
	Col1	Col2 (S)	Col3 (S)
User Header	User Row ID	binary_quality	Prediction
1		B	B
2		B	B
3		B	B
4		B	B
5		B	B
6		B	B
7		B	B
8		B	B
9		B	B
10		B	B
11		B	B
12		G	B
13		B	B
14		B	B
15		G	G

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 "binary_quality" to be transferred to the output spreadsheet.



The predictions will appear on the output spreadsheet.

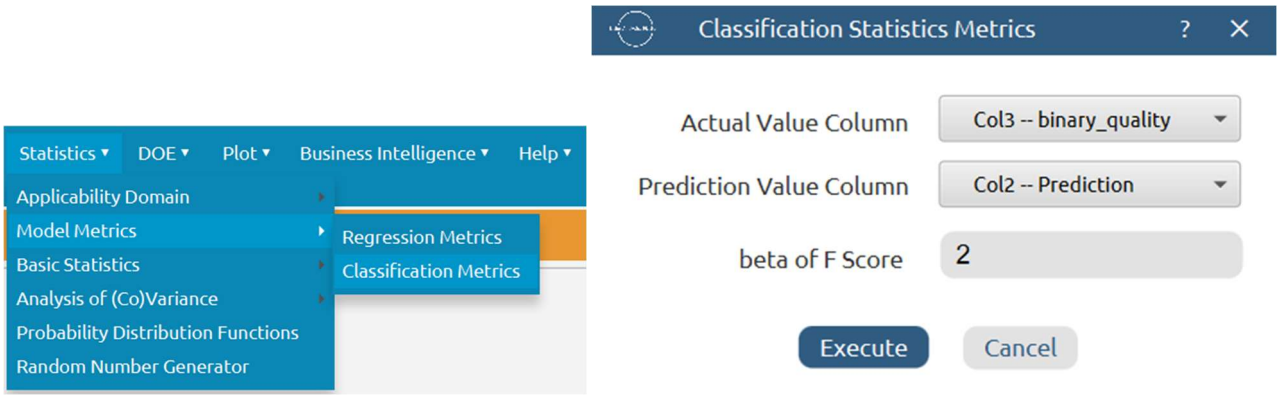
	Col1	Col2 (S)	Col3 (S)
User Header	User Row ID	Prediction	binary_quality
1		B	B
2		B	B
3		B	B
4		G	G
5		G	G
6		B	B
7		B	B
8		B	B
9		B	B
10		B	B
11		B	B
12		B	B
13		B	B
14		B	B
15		B	B

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*



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			B	G
3	Actual Class	B	930	29
4	Actual Class	G	170	95
5				
6				
7	Classification Accuracy	0.8374183		
8				
9	Precision		0.8454545	0.7661290
10				
11	Recall/Sensitivity		0.9697602	0.3584906
12				
13	Specificity		0.3584906	0.9697602
14				
15	F1 Score		0.9033511	0.4884319
16				
17	F (beta=2)		0.9420583	0.4011824
18				
19	MCC	0.4480544		

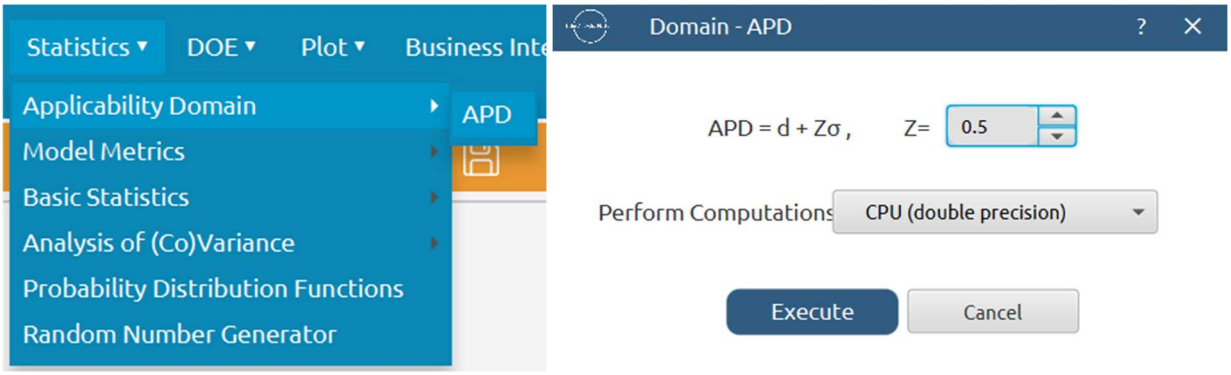
Step 9: Reliability check for each record of the test set

Step 9.a: Create the domain

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

Import data into the input spreadsheet of the "DOMAIN" tab from the output of the "NORMALIZE_TRAIN_SET" 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.

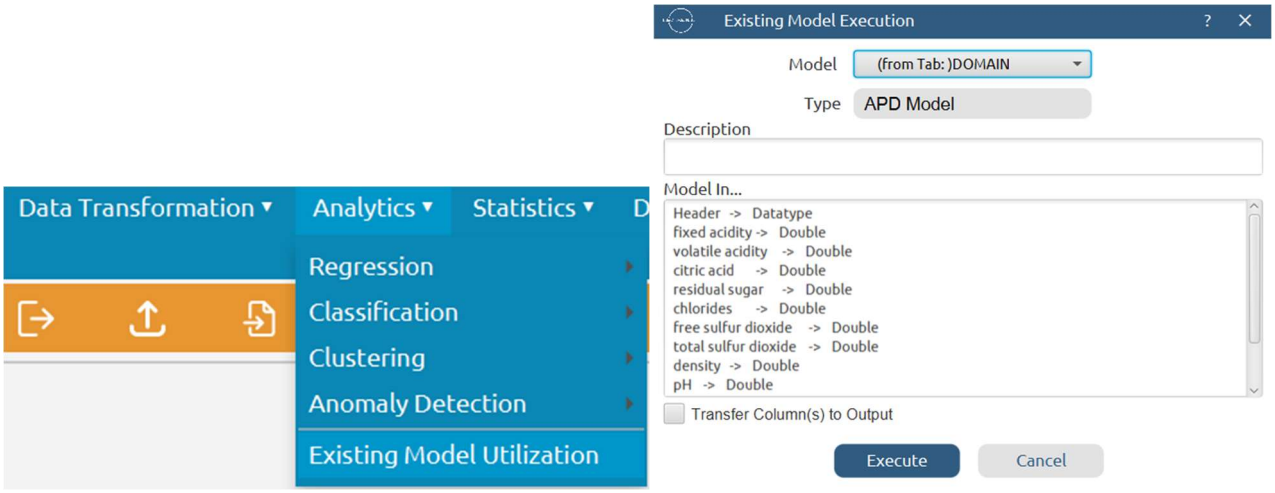
	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1		0.0	3.7344623	reliable
2		0.0	3.7344623	reliable
3		0.0	3.7344623	reliable
4		0.0	3.7344623	reliable
5		0.0	3.7344623	reliable
6		0.0	3.7344623	reliable
7		0.0	3.7344623	reliable
8		0.0	3.7344623	reliable
9		0.0	3.7344623	reliable
10		0.0	3.7344623	reliable
11		0.0	3.7344623	reliable
12		0.0	3.7344623	reliable
13		0.0	3.7344623	reliable
14		0.0	3.7344623	reliable
15		0.0	3.7344623	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 "RELIABILITY".

Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "NORMALIZE_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		0.6498620	3.7344623	reliable
2		1.2406787	3.7344623	reliable
3		1.4503034	3.7344623	reliable
4		1.3860351	3.7344623	reliable
5		0.0	3.7344623	reliable
6		1.2702076	3.7344623	reliable
7		0.8537607	3.7344623	reliable
8		0.6634052	3.7344623	reliable
9		1.3150571	3.7344623	reliable
10		1.3691783	3.7344623	reliable
11		0.9852532	3.7344623	reliable
12		0.9852532	3.7344623	reliable
13		1.3561545	3.7344623	reliable
14		0.0	3.7344623	reliable
15		0.0	3.7344623	reliable

Final Isalos Workflow

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