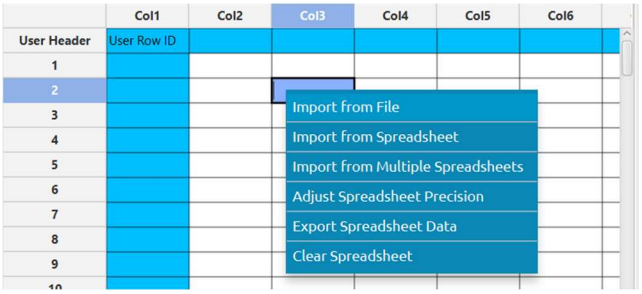
Bank Customer Churn Prediction

This dataset, which can be found in https://www.kaggle.com/datasets/shubhammeshram579/bank-customer-churn-prediction, contains information about customer churn in the banking industry. It contains 12 features and 10002 samples, aiming to predict whether a customer will leave the bank or not. The target column is called "Exited" where a value of 1 indicates that the customer has left the bank and a value of 0 indicates they haven't.

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 bank customer churn data.

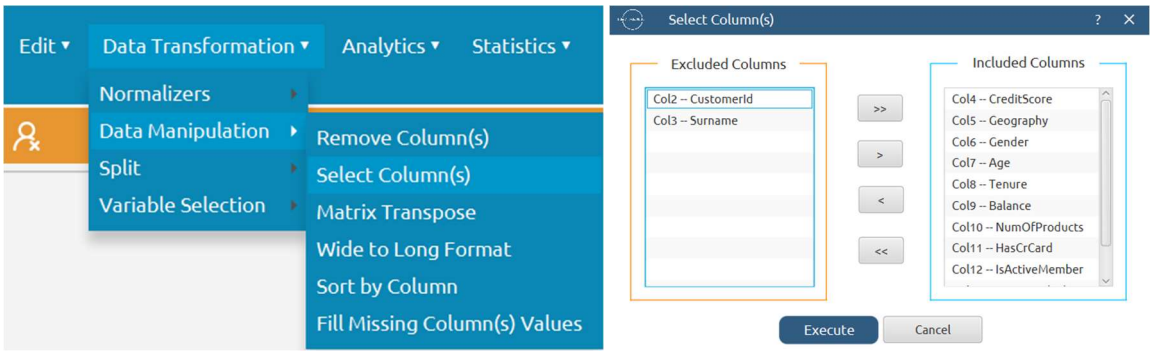


The data will appear on the left spreadsheet.

	Col1	Col2 (I)	Col3 (S)	Col4 (I)	Col5 (S)	Col6 (S)	Col7 (D)	Col8 (I)	Col9 (D)	Col10 (I)	Col11 (I)	Col12 (I)	Col13 (D)	Col14 (I)
User Header	User Row ID	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
2	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1		1	79084.1	0
6	6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	7	15592531	Bartlett	822		Male	50	7	0	2	1	1	10062.8	0
8	8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
9	9	15792365	He	501	France	Male	44	4	142051.07	2	0		74940.5	0
10	10	15592389	H?	684	France	Male		2	134603.88	1	1	1	71725.73	0
11	11	15767821	Bearce	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
12	12	15737173	Andrews	497	Spain	Male	24	3	0	2	1	0	76390.01	0
13	13	15632264	Kay	476	France	Female	34	10	0	2	1	0	26260.98	0
14	14	15691483	Chin	549	France	Female	25	5	0	2	0	0	190857.79	0
15	15	15600882	Scott	635	Spain	Female	35	7	0	2	1	1	65951.65	0

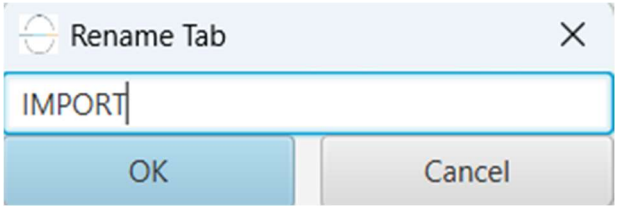
Step 2: Manipulate data

From the available columns we will select the ones that will be useful features for predicting customer churn. Customer IDs and surnames do not provide any useful information about customer behaviour, so they will be left out. On the menu click on *Data Transformation → Data Manipulation → Select Column(s)* and select the rest of the columns.



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

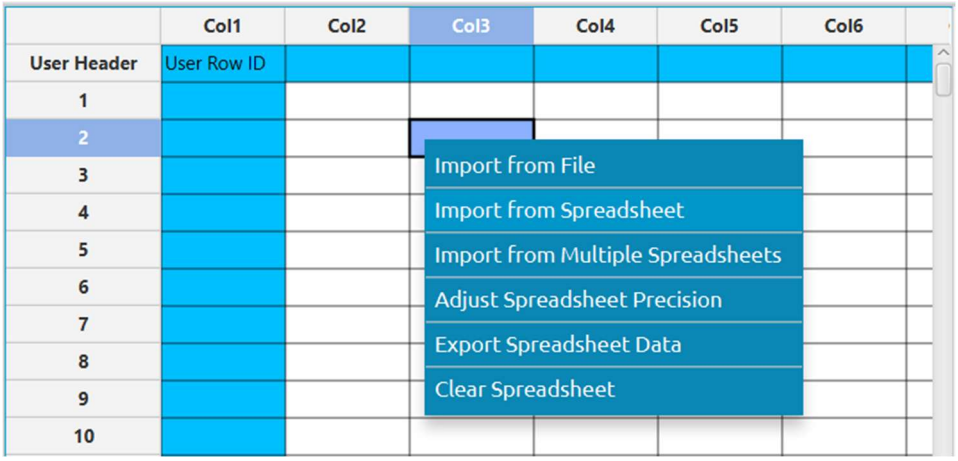
	Col1	Col2 (I)	Col3 (S)	Col4 (S)	Col5 (D)	Col6 (I)	Col7 (D)	Col8 (I)	Col9 (I)	Col10 (I)	Col11 (D)	Col12 (I)
User Header	User Row ID	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	1	619	France	Female	42	2	0	1	1	1	101348.88	1
2	2	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	3	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	4	699	France	Female	39	1	0	2	0	0	93826.63	0
5	5	850	Spain	Female	43	2	125510.82	1		1	79084.1	0
6	6	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	7	822		Male	50	7	0	2	1	1	10062.8	0
8	8	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
9	9	501	France	Male	44	4	142051.07	2	0		74940.5	0
10	10	684	France	Male		2	134603.88	1	1	1	71725.73	0
11	11	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
12	12	497	Spain	Male	24	3	0	2	1	0	76390.01	0
13	13	476	France	Female	34	10	0	2	1	0	26260.98	0
14	14	549	France	Female	25	5	0	2	0	0	190857.79	0
15	15	635	Spain	Female	35	7	0	2	1	1	65951.65	0



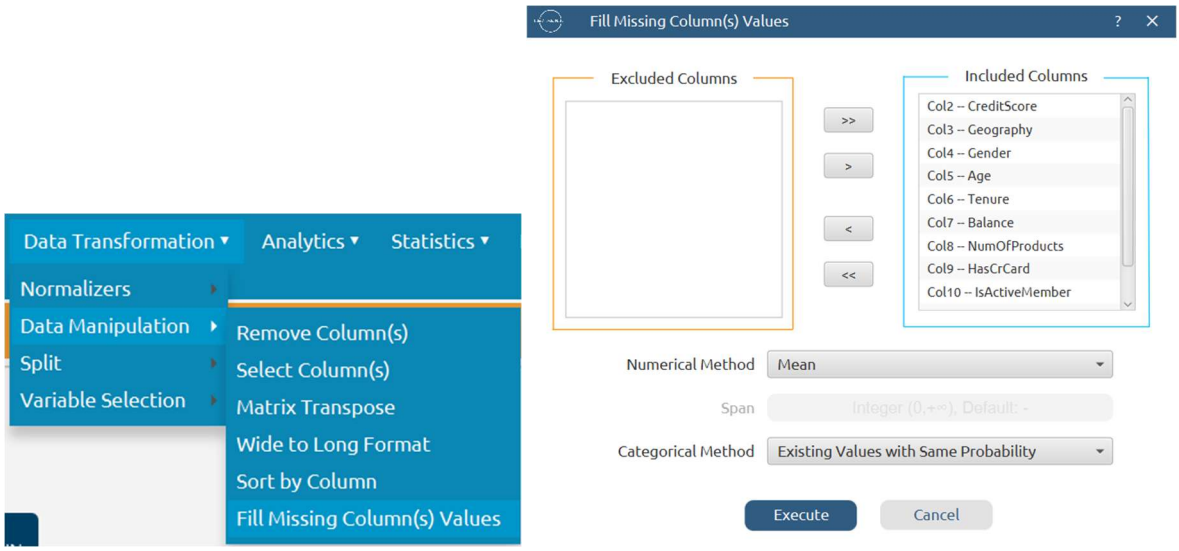
Step 3: Fill missing data

This dataset includes some missing values therefore we need to fill them appropriately. Create a new tab by pressing the "+" button on the bottom of the page with the name "FILL_MISSING" which we will use for filling the missing values.

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



To fill the missing data, choose: *Data Transformation → Data Manipulation → Fill Missing Column(s) Values*



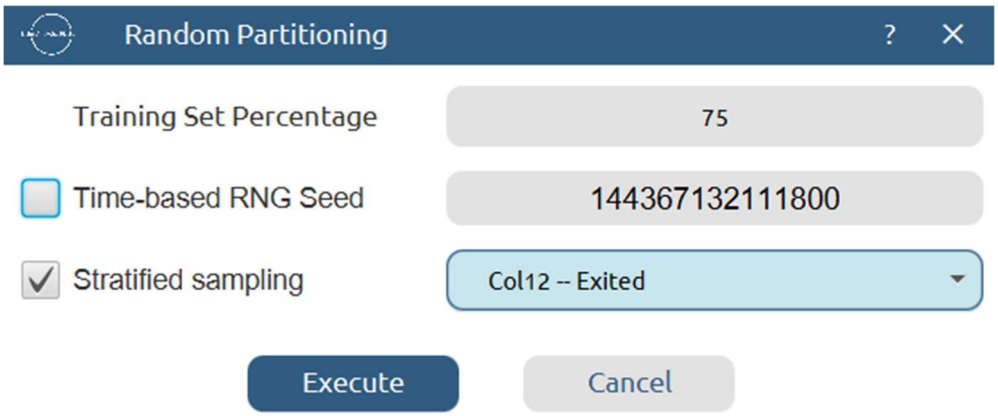
The results will appear on the output spreadsheet.

Step 4: 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 "FILL_MISSING" 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.

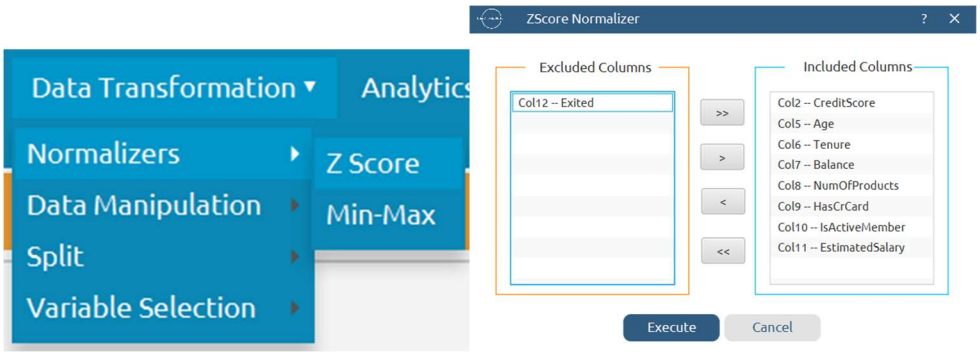
Step 5: Normalize the training set

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 (S)	Col4 (S)	Col5 (D)	Col6 (I)	Col7 (D)	Col8 (I)	Col9 (I)	Col10 (I)	Col11 (D)	Col12 (I)
User Header	User Row ID	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	1	619	France	Female	42	2	0	1	1	1	101348.88	1
2	2	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	3	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	4	699	France	Female	39	1	0	2	0	0	93826.63	0
5	6	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
6	7	822	Spain	Male	50	7	0	2	1	1	10062.8	0
7	9	501	France	Male	44	4	142051.07	2	0	1	74940.5	0
8	10	684	France	Male	38.9223	2	134603.88	1	1	1	71725.73	0
9	13	476	France	Female	34	10	0	2	1	0	26260.98	0
10	14	549	France	Female	25	5	0	2	0	0	190857.79	0
11	15	635	Spain	Female	35	7	0	2	1	1	65951.65	0
12	16	616	Germany	Male	45	3	143129.41	2	0	1	64327.26	0
13	18	549	Spain	Female	24	9	0	2	1	1	14406.41	0
14	19	587	Spain	Male	45.25	6	0	1	0	0	158684.81	0
15	20	726	France	Female	24	6	0	2	1	1	54724.03	0

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



The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (S)	Col4 (S)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)
User Header	User Row ID	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	1	-0.3304575	France	Female	0.2980531	-1.0461880	-1.2251290	-0.9064502	0.6478440	0.9736216	0.0257231	1.0
2	2	-0.4450013	Spain	Female	0.2024135	-1.3922724	0.1139903	-0.9064502	-1.5433758	0.9736216	0.2200306	0.0
3	3	-1.5487870	France	Female	0.2980531	1.0303185	1.3260023	2.5332006	0.6478440	-1.0269561	0.2441416	1.0
4	4	0.5025883	France	Female	0.0111344	-1.3922724	-1.2251290	0.8133752	-1.5433758	-1.0269561	-0.1048530	0.0
5	6	-0.0597176	Spain	Male	0.4893322	1.0303185	0.5925115	0.8133752	0.6478440	-1.0269561	0.8660178	1.0
6	7	1.7833962	Spain	Male	1.0631696	0.6842341	-1.2251290	0.8133752	0.6478440	0.9736216	-1.5588803	0.0
7	9	-1.5592000	France	Male	0.4893322	-0.3540192	1.0446262	0.8133752	-1.5433758	0.9736216	-0.4326908	0.0
8	10	0.3463922	France	Male	0.0037032	-1.0461880	0.9256317	-0.9064502	0.6478440	0.9736216	-0.4884949	0.0
9	13	-1.8195268	France	Female	-0.4670634	1.7224874	-1.2251290	0.8133752	0.6478440	-1.0269561	-1.2777017	0.0
10	14	-1.0593726	France	Female	-1.3278194	-0.0079348	-1.2251290	0.8133752	-1.5433758	-1.0269561	1.5794772	0.0
11	15	-0.1638483	Spain	Female	-0.3714238	0.6842341	-1.2251290	0.8133752	0.6478440	0.9736216	-0.5887251	0.0
12	16	-0.3616967	Germany	Male	0.5849718	-0.7001036	1.0618564	0.8133752	-1.5433758	0.9736216	-0.6169224	0.0
13	18	-1.0593726	Spain	Female	-1.4234590	1.3764029	-1.2251290	0.8133752	0.6478440	0.9736216	-1.4834810	0.0
14	19	-0.6636758	Spain	Male	0.6088817	0.3381497	-1.2251290	-0.9064502	-1.5433758	-1.0269561	1.0209976	0.0
15	20	0.7837412	France	Female	-1.4234590	0.3381497	-1.2251290	0.8133752	0.6478440	0.9736216	-0.7836215	0.0

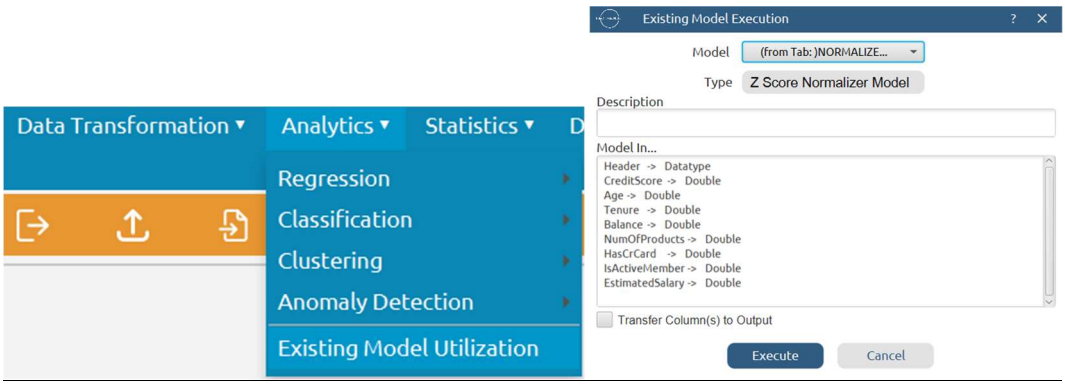
Step 6: Normalize the test set

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 (S)	Col4 (S)	Col5 (D)	Col6 (I)	Col7 (D)	Col8 (I)	Col9 (I)	Col10 (I)	Col11 (D)	Col12 (I)
User Header	User Row ID	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	5	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0
2	8	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
3	11	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
4	12	497	Spain	Male	24	3	0	2	1	0	76390.01	0
5	17	653	Germany	Male	58	1	132602.88	1	1	0	5097.67	1
6	21	732	France	Male	41	8	0	2	1	1	170886.17	0
7	22	636	Spain	Female	32.34	8	0	2	1	0	138555.46	0
8	23	510	Spain	Female	38	4	0	1	1	0	118913.53	1
9	26	577	France	Male	25	3	0	2	0	1	124508.29	0
10	28	571	France	Male	44	9	0	2	0	0	38433.35	0
11	29	574	Germany	Female	43	3	141349.43	1	1	1	100187.43	0
12	33	553	Germany	Male	41	9	110112.54	2	0	0	81898.81	0
13	36	475	France	Female	45	0	134264.04	1	1	0	27822.99	1
14	38	804	Spain	Male	33	7	76548.6	1	0	1	98453.45	0
15	40	582	Germany	Male	41	6	70349.48	2	0	1	178074.04	0

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.

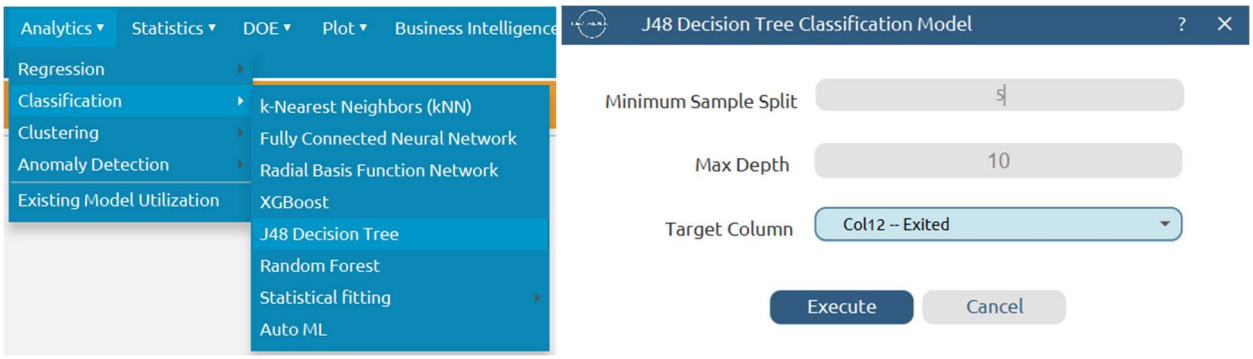
	Col1	Col2 (D)	Col3 (S)	Col4 (S)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)
User Header	User Row ID	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	5	2.0749622	Spain	Female	0.3936927	-1.0461880	0.7803387	-0.9064502	0.6478440	0.9736216	-0.3607635	0.0
2	8	-2.8608341	Germany	Female	-0.9452612	-0.3540192	0.6131390	4.2530260	0.6478440	-1.0269561	0.3381441	1.0
3	11	-1.2780471	France	Male	-0.7539821	0.3381497	0.4049395	0.8133752	-1.5433758	-1.0269561	-0.3417207	0.0
4	12	-1.6008523	Spain	Male	-1.4234590	-0.7001036	-1.2251290	0.8133752	0.6478440	-1.0269561	-0.4075293	0.0
5	17	0.0235870	Germany	Male	1.8282861	-1.3922724	0.8936588	-0.9064502	0.6478440	-1.0269561	-1.6450682	1.0
6	21	0.8462197	France	Male	0.2024135	1.0303185	-1.2251290	0.8133752	0.6478440	0.9736216	1.2327967	0.0
7	22	-0.1534353	Spain	Female	-0.6258251	1.0303185	-1.2251290	0.8133752	0.6478440	-1.0269561	0.6715792	0.0
8	23	-1.4654824	Spain	Female	-0.0845051	-0.3540192	-1.2251290	-0.9064502	0.6478440	-1.0269561	0.3306217	1.0
9	26	-0.7678065	France	Male	-1.3278194	-0.7001036	-1.2251290	0.8133752	-1.5433758	0.9736216	0.4277392	0.0
10	28	-0.8302850	France	Male	0.4893322	1.3764029	-1.2251290	0.8133752	-1.5433758	-1.0269561	-1.0664057	0.0
11	29	-0.7990458	Germany	Female	0.3936927	-0.7001036	1.0334151	-0.9064502	0.6478440	0.9736216	0.0055619	0.0
12	33	-1.0177203	Germany	Male	0.2024135	1.3764029	0.5342982	0.8133752	-1.5433758	-1.0269561	-0.3119039	0.0
13	36	-1.8299399	France	Female	0.5849718	-1.7383569	0.9202016	-0.9064502	0.6478440	-1.0269561	-1.2505873	1.0
14	38	1.5959609	Spain	Male	-0.5627030	0.6842341	-0.0020014	-0.9064502	-1.5433758	0.9736216	-0.0245377	0.0
15	40	-0.7157412	Germany	Male	0.2024135	0.3381497	-0.1010537	0.8133752	-1.5433758	0.9736216	1.3575685	0.0

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



The predictions will appear on the output spreadsheet.

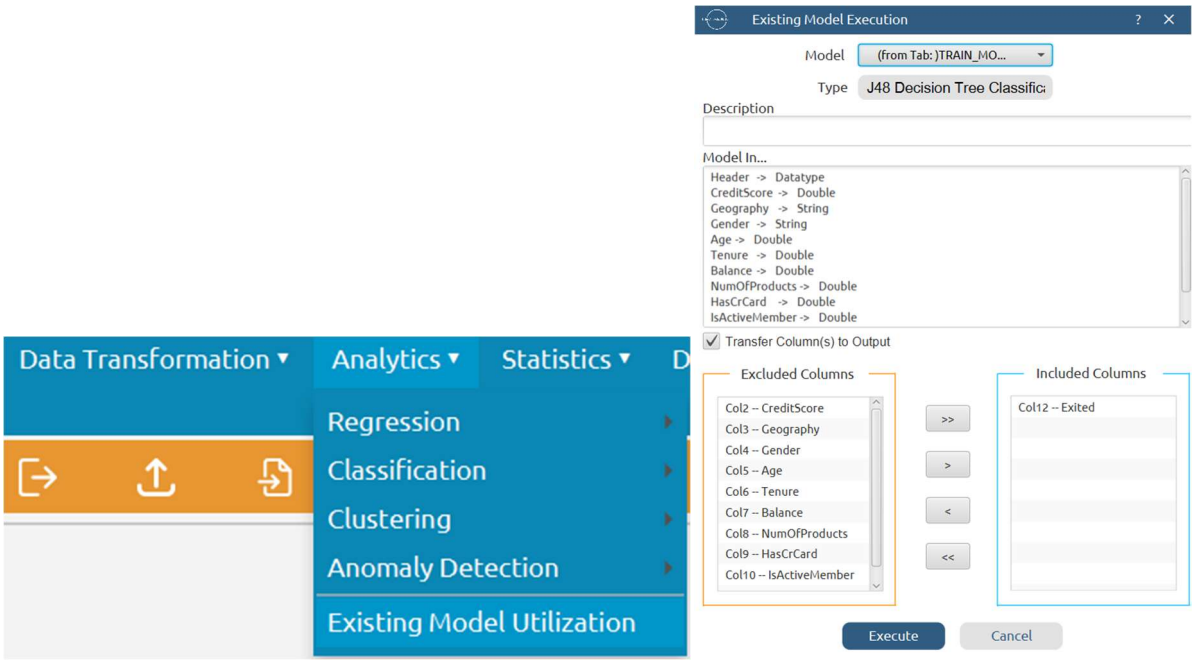
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Exited	Prediction
1	1	1.0	1.0
2	2	0.0	0.0
3	3	1.0	1.0
4	4	0.0	0.0
5	6	1.0	0.0
6	7	0.0	0.0
7	9	0.0	0.0
8	10	0.0	0.0
9	13	0.0	0.0
10	14	0.0	0.0
11	15	0.0	0.0
12	16	0.0	0.0
13	18	0.0	0.0
14	19	0.0	1.0
15	20	0.0	0.0

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 "Exited" to be transferred to the output spreadsheet.



The predictions will appear on the output spreadsheet.

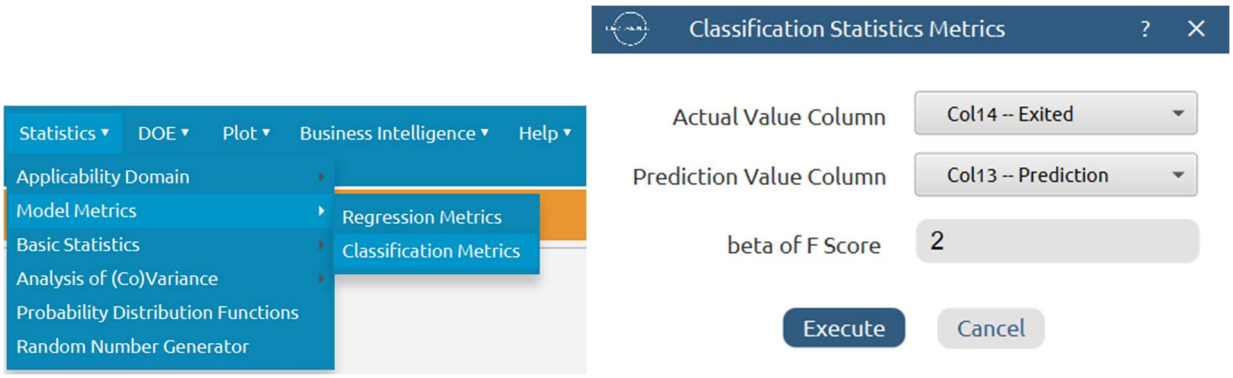
	Col13 (D)	Col14 (D)
User Header	Prediction	Exited
1	0.0	0.0
2	1.0	1.0
3	0.0	0.0
4	0.0	0.0
5	1.0	1.0
6	0.0	0.0
7	0.0	0.0
8	0.0	1.0
9	0.0	0.0
10	0.0	0.0
11	0.0	0.0
12	0.0	0.0
13	0.0	1.0
14	0.0	0.0
15	0.0	0.0

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 classification: *Statistics → Model Metrics → Classification Metrics*



The results will appear on the output spreadsheet.

	Col13 (D)	Col14 (D)
User Header	Prediction	Exited
1	0.0	0.0
2	1.0	1.0
3	0.0	0.0
4	0.0	0.0
5	1.0	1.0
6	0.0	0.0
7	0.0	0.0
8	0.0	1.0
9	0.0	0.0
10	0.0	0.0
11	0.0	0.0
12	0.0	0.0
13	0.0	1.0
14	0.0	0.0
15	0.0	0.0

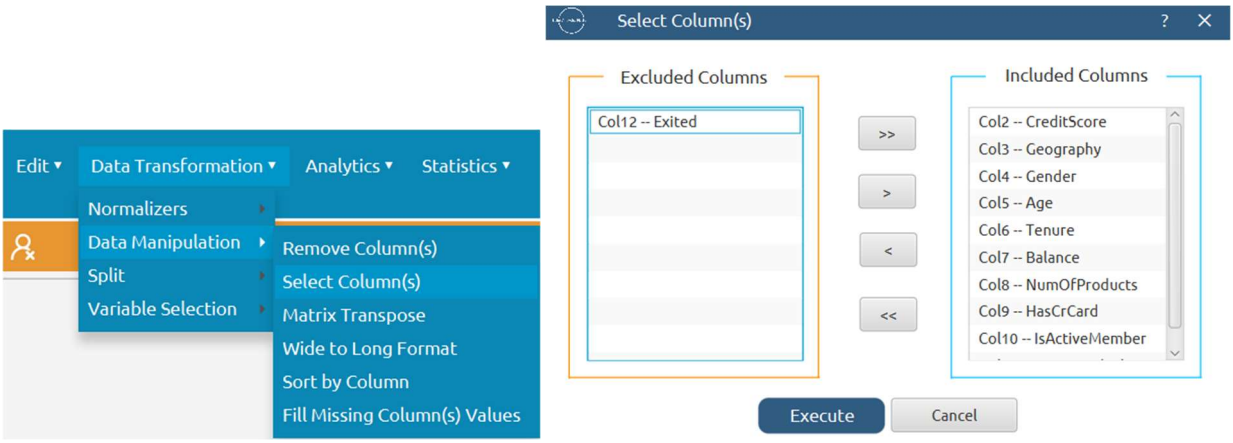
Step 10: Reliability check for each record of the test set

Step 10.a: Create the domain

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

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

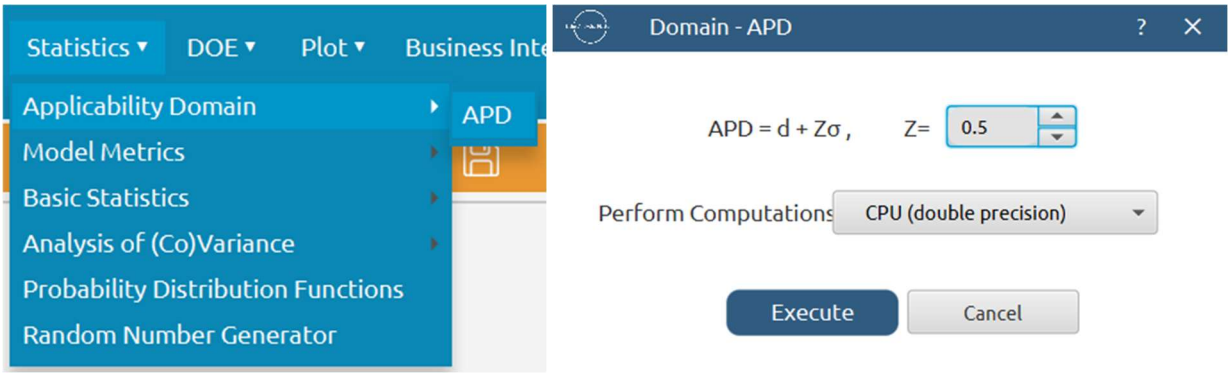


The results will appear on the output spreadsheet.

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_EXIT" 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.

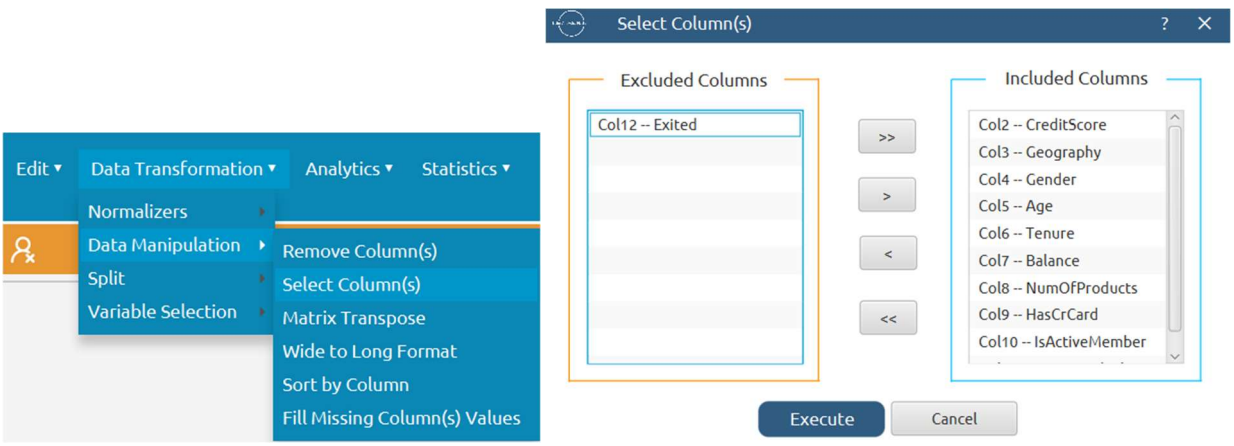
	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1	1	0.0	3.4497162	reliable
2	2	0.0	3.4497162	reliable
3	3	0.0	3.4497162	reliable
4	4	0.0	3.4497162	reliable
5	6	0.0	3.4497162	reliable
6	7	0.0	3.4497162	reliable
7	9	0.0	3.4497162	reliable
8	10	0.0	3.4497162	reliable
9	13	0.0	3.4497162	reliable
10	14	0.0	3.4497162	reliable
11	15	0.0	3.4497162	reliable
12	16	0.0	3.4497162	reliable
13	18	0.0	3.4497162	reliable
14	19	0.0	3.4497162	reliable
15	20	0.0	3.4497162	reliable

Step 10.b: Check the test set reliability

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

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

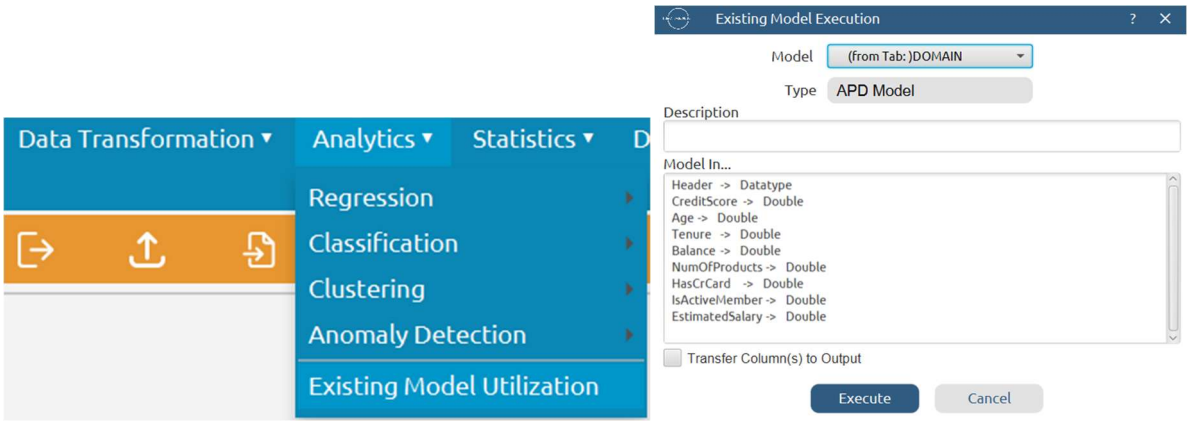


The results will appear on the output spreadsheet.

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_EXIT_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	5	0.4974462	3.4497162	reliable
2	8	1.9535274	3.4497162	reliable
3	11	0.5925998	3.4497162	reliable
4	12	0.6388762	3.4497162	reliable
5	17	0.5839557	3.4497162	reliable
6	21	0.3828181	3.4497162	reliable
7	22	0.1217719	3.4497162	reliable
8	23	0.4540411	3.4497162	reliable
9	26	0.3072059	3.4497162	reliable
10	28	0.4855175	3.4497162	reliable
11	29	0.4021592	3.4497162	reliable
12	33	0.5581759	3.4497162	reliable
13	36	1.0419899	3.4497162	reliable
14	38	0.6518536	3.4497162	reliable
15	40	0.7596882	3.4497162	reliable

Final Isalos Workflow

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