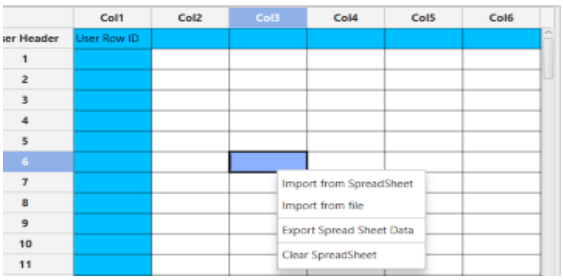
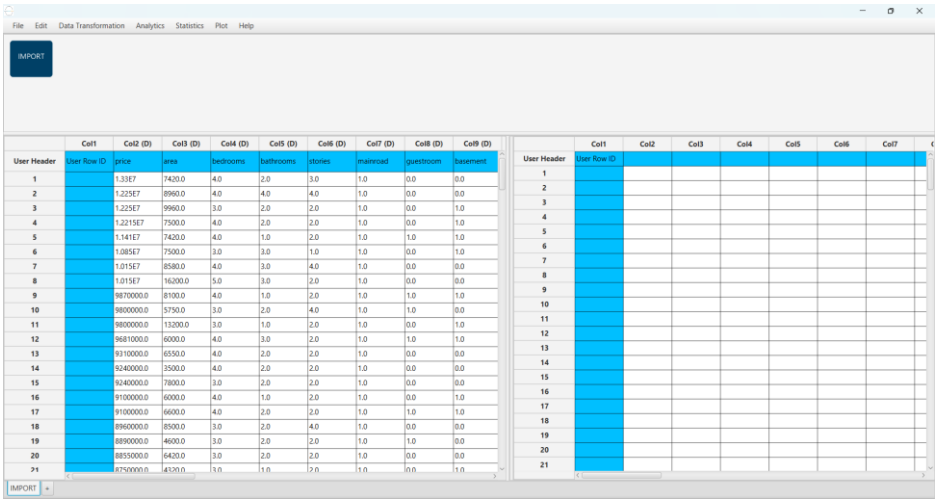
Bodyfat Percentage (Regression)

The goal of this study is to train a model in order to predict the percentage of bodyfat in men. The dataset used in this case study is found in https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset and has 14 features and 251 samples. This dataset contains information on certain factors like Density determined from underwater weighing, age, weight, height, neck circumference, Chest circumference, Abdomen 2 circumference, Hip circumference, Thigh circumference, Knee circumference, Ankle circumference, Biceps (extended) circumference, Forearm circumference, Wrist circumference. The dataset contains no missing values and also all features are numerical, therefore there is no need to for transforming any of them.

Step 1: Import data from file

Right click on the input spreadsheet and choose the option "Import from file". Then navigate through your files to load the one with the housing price data.

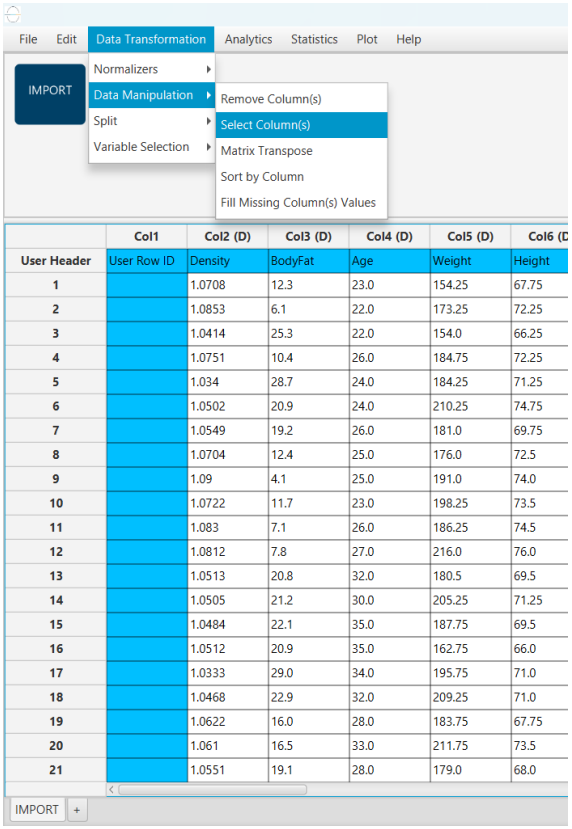




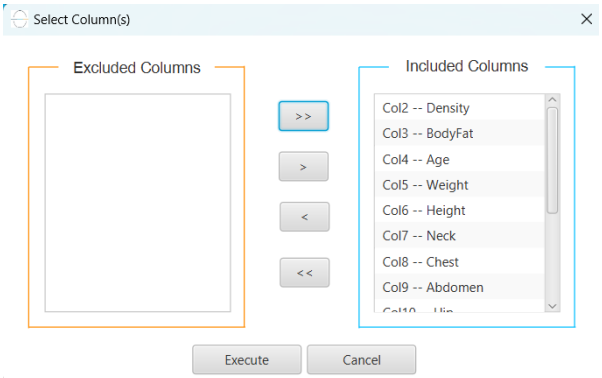
Step 2: Manipulate data

In order to use the data for training we have to exclude any columns that do not contain features. In our dataset there are no such columns. Therefore, we will include all columns in the training. We follow these steps to execute this:

- On the menu click on "Data Transformation" → "Data Manipulation" → "Select Column(s)"
- Select all columns.



	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D
User Header	User Row ID	Density	BodyFat	Age	Weight	Height
1		1.0708	12.3	23.0	154.25	67.75
2		1.0853	6.1	22.0	173.25	72.25
3		1.0414	25.3	22.0	154.0	66.25
4		1.0751	10.4	26.0	184.75	72.25
5		1.034	28.7	24.0	184.25	71.25
6		1.0502	20.9	24.0	210.25	74.75
7		1.0549	19.2	26.0	181.0	69.75
8		1.0704	12.4	25.0	176.0	72.5
9		1.09	4.1	25.0	191.0	74.0
10		1.0722	11.7	23.0	198.25	73.5
11		1.083	7.1	26.0	186.25	74.5
12		1.0812	7.8	27.0	216.0	76.0
13		1.0513	20.8	32.0	180.5	69.5
14		1.0505	21.2	30.0	205.25	71.25
15		1.0484	22.1	35.0	187.75	69.5
16		1.0512	20.9	35.0	162.75	66.0
17		1.0333	29.0	34.0	195.75	71.0
18		1.0468	22.9	32.0	209.25	71.0
19		1.0622	16.0	28.0	183.75	67.75
20		1.061	16.5	33.0	211.75	73.5
21		1.0551	19.1	28.0	179.0	68.0

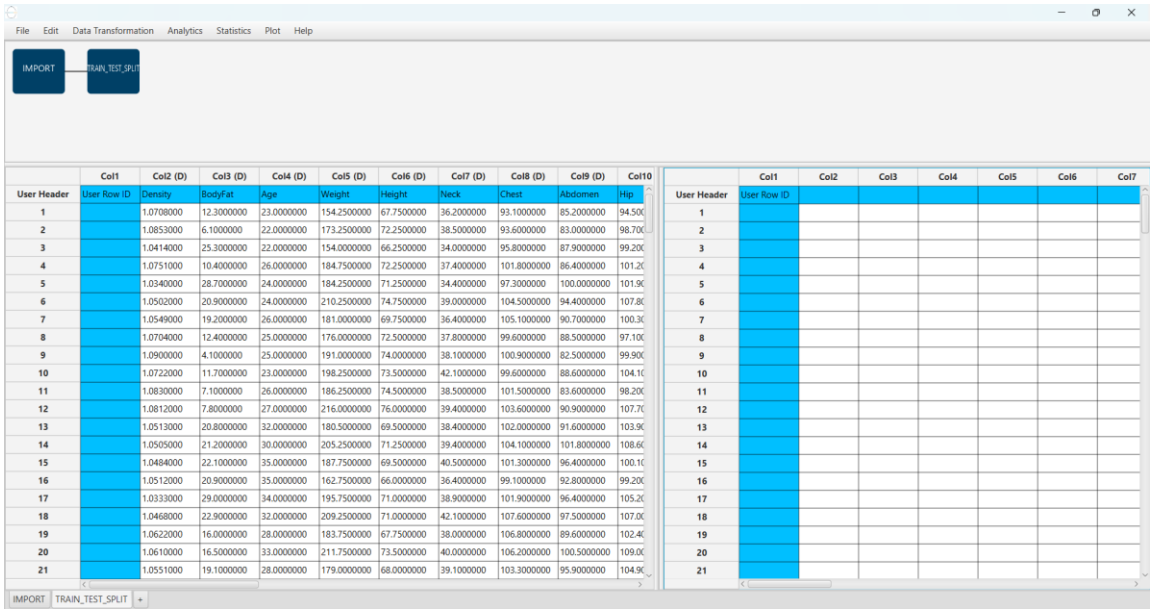


The data will appear in the output spreadsheet.

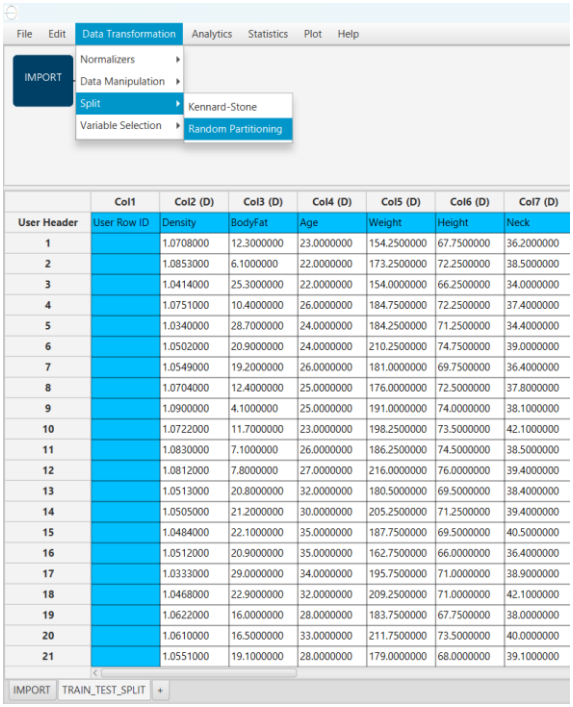
Step 3: Split data

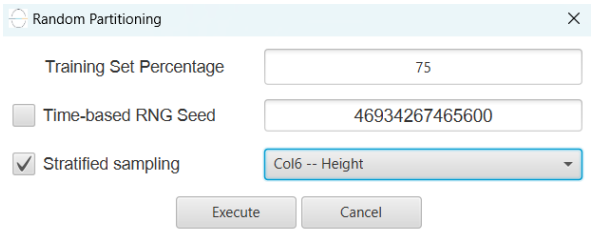
Create a new tab by pressing the "+" button on the bottom of the page with the name "TRAIN_TEST_SPLIT" which we will use for splitting to create the train and test set.

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".

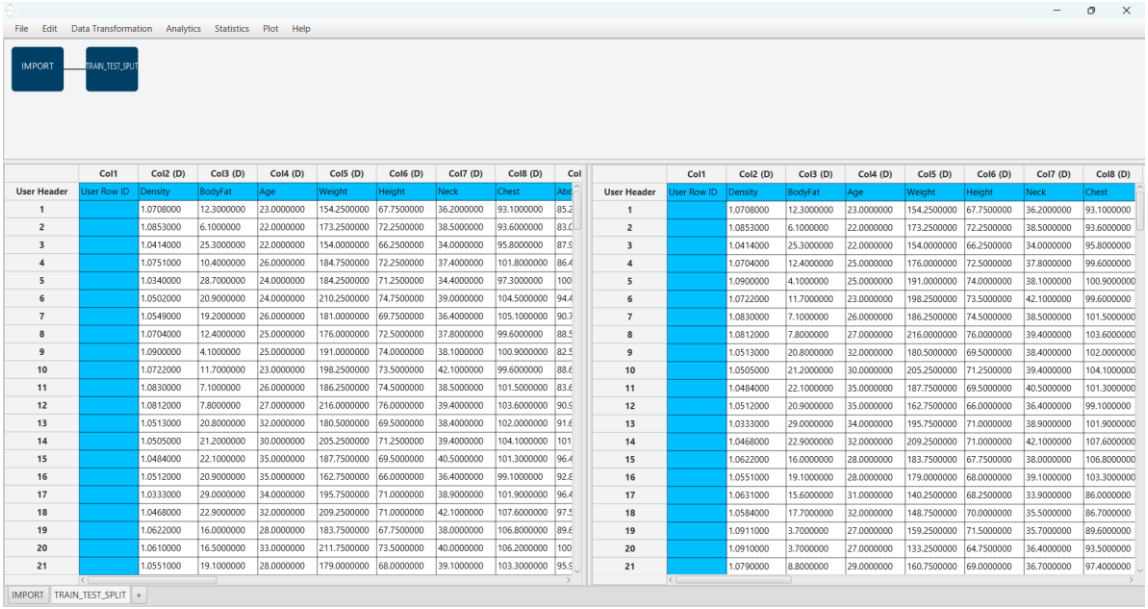


Split the dataset by choosing from the top ribbon: "Data Transformation" → "Split" → "Random Partitioning". Then choose the "Training set percentage" and the column for the sampling as shown below:





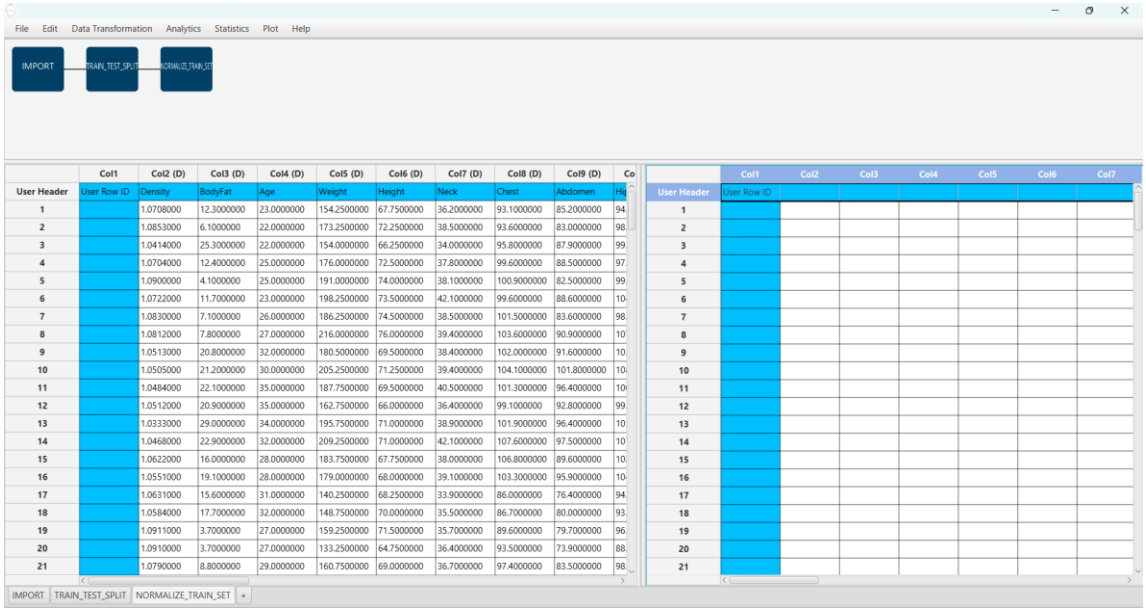
The results will appear on the output spreadsheet.



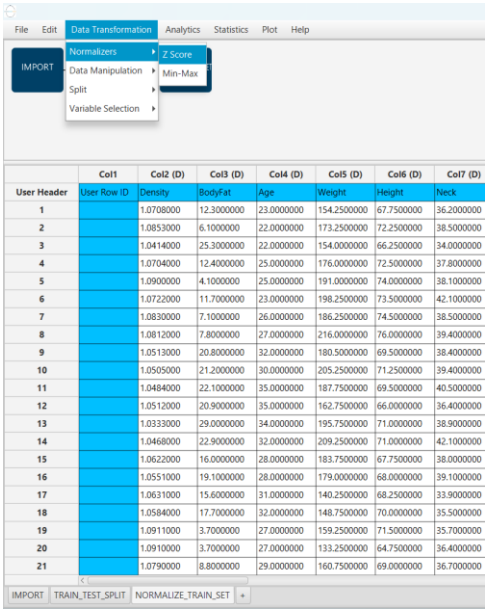
Step 4: Normalize the training set

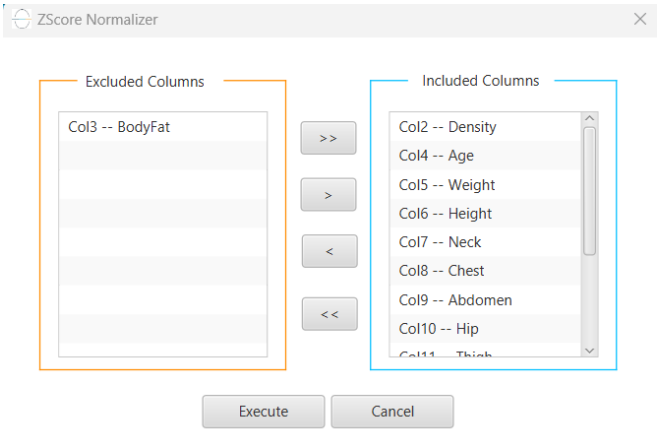
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALIZE_TRAIN_SET".

Import data 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".

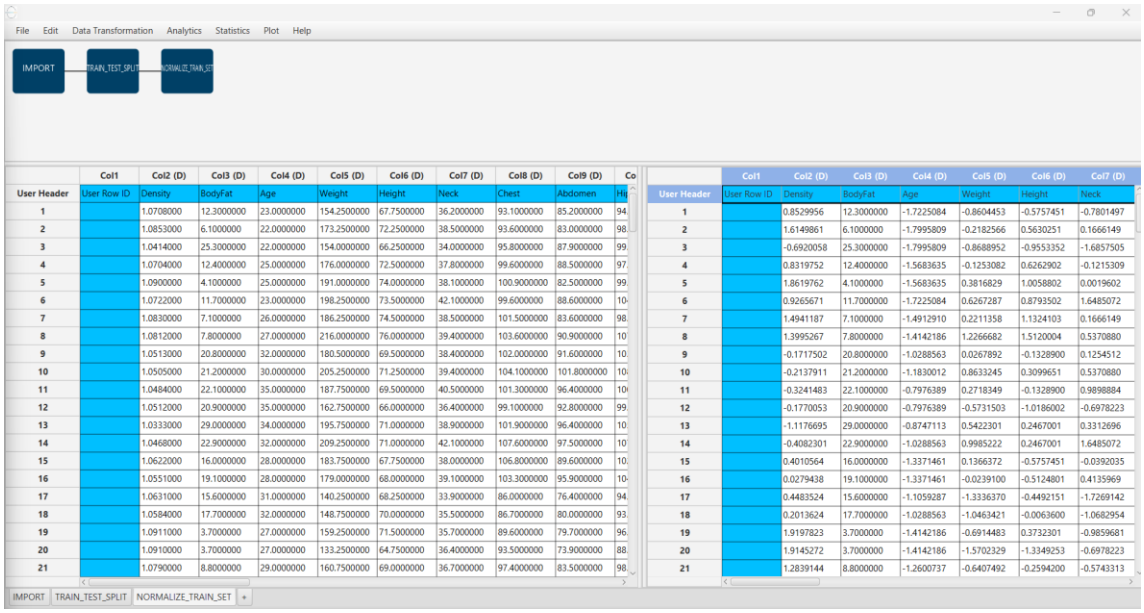


Normalize the data using Z-score by browsing: "Data Transformation" → "Normalizers" → "Z-Score". Then select all columns except "BodyFat" and click "Execute".





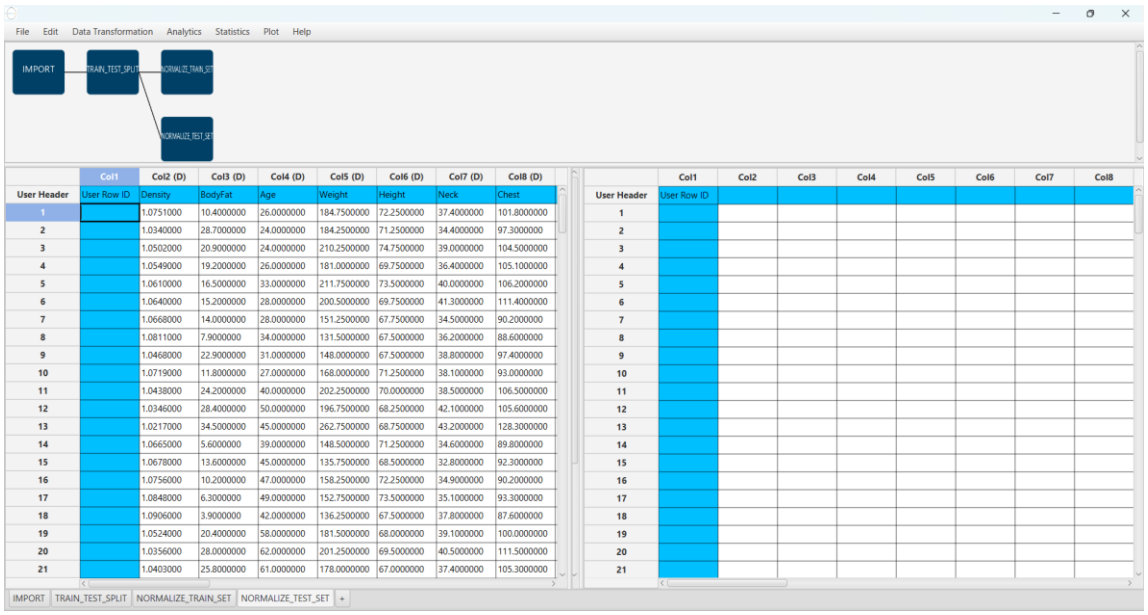
The results will appear on the output spreadsheet.



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".

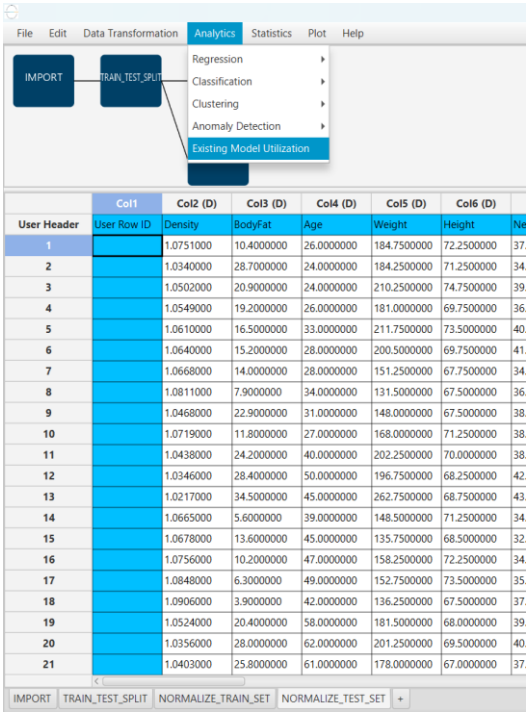
Import data into the input spreadsheet of the "NORMALIZE_TEST_SET" tab the test set from the output of the "TRAIN_TEST_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet". From the available Select input tab options choose "TRAIN_TEST_SPLIT: Test Set".



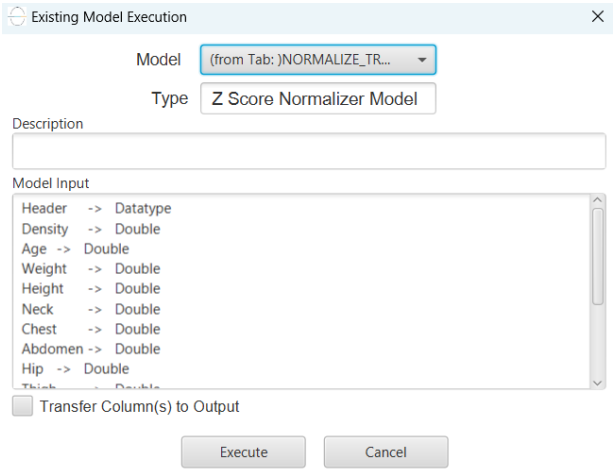
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
User Header	User Row ID	Density	BodyFat	Age	Weight	Height	Neck	Chest
1		1.0751000	10.4000000	26.0000000	184.7500000	72.2500000	37.4000000	101.8000000
2		1.0340000	28.7000000	24.0000000	184.2500000	71.2500000	34.4000000	97.3000000
3		1.0502000	20.9000000	24.0000000	210.2500000	74.7500000	39.0000000	104.5000000
4		1.0549000	19.2000000	26.0000000	181.0000000	69.7500000	36.4000000	105.1000000
5		1.0610000	16.5000000	33.0000000	211.7500000	73.5000000	40.0000000	106.2000000
6		1.0640000	15.2000000	28.0000000	200.5000000	69.7500000	41.3000000	111.4000000
7		1.0668000	14.0000000	28.0000000	151.2500000	67.7500000	34.5000000	90.2000000
8		1.0811000	7.9000000	34.0000000	131.5000000	67.5000000	36.2000000	88.6000000
9		1.0468000	22.9000000	31.0000000	148.0000000	67.5000000	38.8000000	97.4000000
10		1.0719000	11.8000000	27.0000000	168.0000000	71.2500000	38.1000000	93.0000000
11		1.0438000	24.2000000	40.0000000	202.2500000	70.0000000	38.5000000	106.5000000
12		1.0346000	28.4000000	50.0000000	196.7500000	68.2500000	42.1000000	105.6000000
13		1.0217000	34.5000000	45.0000000	262.7500000	68.7500000	43.2000000	128.3000000
14		1.0665000	5.6000000	39.0000000	148.5000000	71.2500000	34.6000000	89.8000000
15		1.0678000	13.6000000	45.0000000	135.7500000	68.5000000	32.8000000	92.3000000
16		1.0756000	10.2000000	47.0000000	158.2500000	72.2500000	34.9000000	90.2000000
17		1.0848000	6.3000000	49.0000000	152.7500000	73.5000000	35.1000000	93.3000000
18		1.0906000	3.9000000	42.0000000	136.2500000	67.5000000	37.8000000	87.6000000
19		1.0524000	20.4000000	58.0000000	181.5000000	68.0000000	39.1000000	100.0000000
20		1.0356000	28.0000000	62.0000000	201.2500000	69.5000000	40.5000000	111.5000000
21		1.0403000	25.8000000	61.0000000	178.0000000	67.0000000	37.4000000	105.3000000

	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
User Header	User Row ID							
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

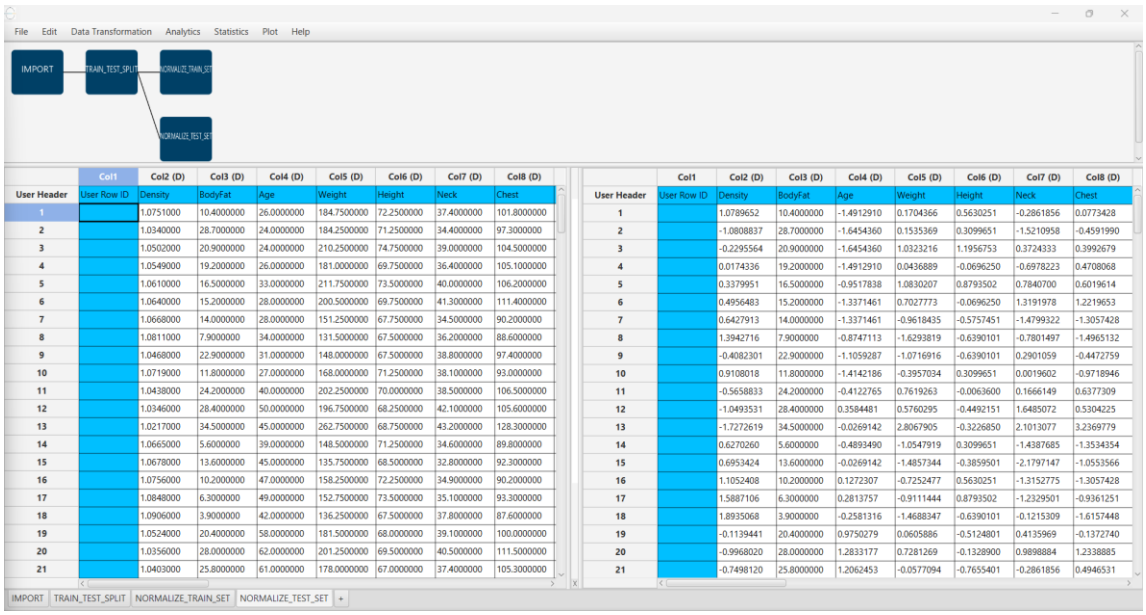
Normalize the test set using the existing normalizer of the training set by browsing: "Analytics" → "Existing Model Utilization" → "Model (from Tab:) NORMALIZE_TRAIN_SET".



	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	
User Header	User Row ID	Density	BodyFat	Age	Weight	Height	Ne
1		1.0751000	10.4000000	26.0000000	184.7500000	72.2500000	37.
2		1.0340000	28.7000000	24.0000000	184.2500000	71.2500000	34.
3		1.0502000	20.9000000	24.0000000	210.2500000	74.7500000	39.
4		1.0549000	19.2000000	26.0000000	181.0000000	69.7500000	36.
5		1.0610000	16.5000000	33.0000000	211.7500000	73.5000000	40.
6		1.0640000	15.2000000	28.0000000	200.5000000	69.7500000	41.
7		1.0668000	14.0000000	28.0000000	151.2500000	67.7500000	34.
8		1.0811000	7.9000000	34.0000000	131.5000000	67.5000000	36.
9		1.0468000	22.9000000	31.0000000	148.0000000	67.5000000	38.
10		1.0719000	11.8000000	27.0000000	168.0000000	71.2500000	38.
11		1.0438000	24.2000000	40.0000000	202.2500000	70.0000000	38.
12		1.0346000	28.4000000	50.0000000	196.7500000	68.2500000	42.
13		1.0217000	34.5000000	45.0000000	262.7500000	68.7500000	43.
14		1.0665000	5.6000000	39.0000000	148.5000000	71.2500000	34.
15		1.0678000	13.6000000	45.0000000	135.7500000	68.5000000	32.
16		1.0756000	10.2000000	47.0000000	158.2500000	72.2500000	34.
17		1.0848000	6.3000000	49.0000000	152.7500000	73.5000000	35.
18		1.0906000	3.9000000	42.0000000	136.2500000	67.5000000	37.
19		1.0524000	20.4000000	58.0000000	181.5000000	68.0000000	39.
20		1.0356000	28.0000000	62.0000000	201.2500000	69.5000000	40.
21		1.0403000	25.8000000	61.0000000	178.0000000	67.0000000	37.

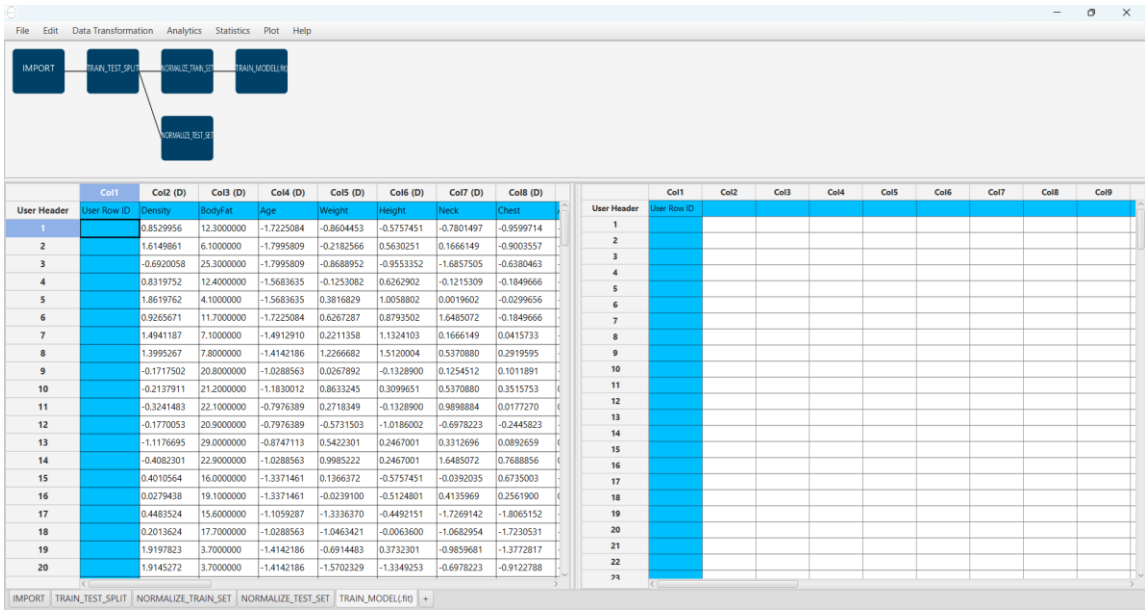


The results will appear on the output spreadsheet.



Step 6: Train the model

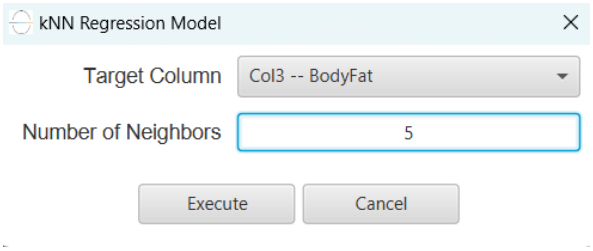
Create a new tab by pressing the "+" button on the bottom of the page with the name "TRAIN_MODEL(.fit)". Import data into the input spreadsheet of the "TRAIN_MODEL(.fit)" tab from the output of the "NORMALIZE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



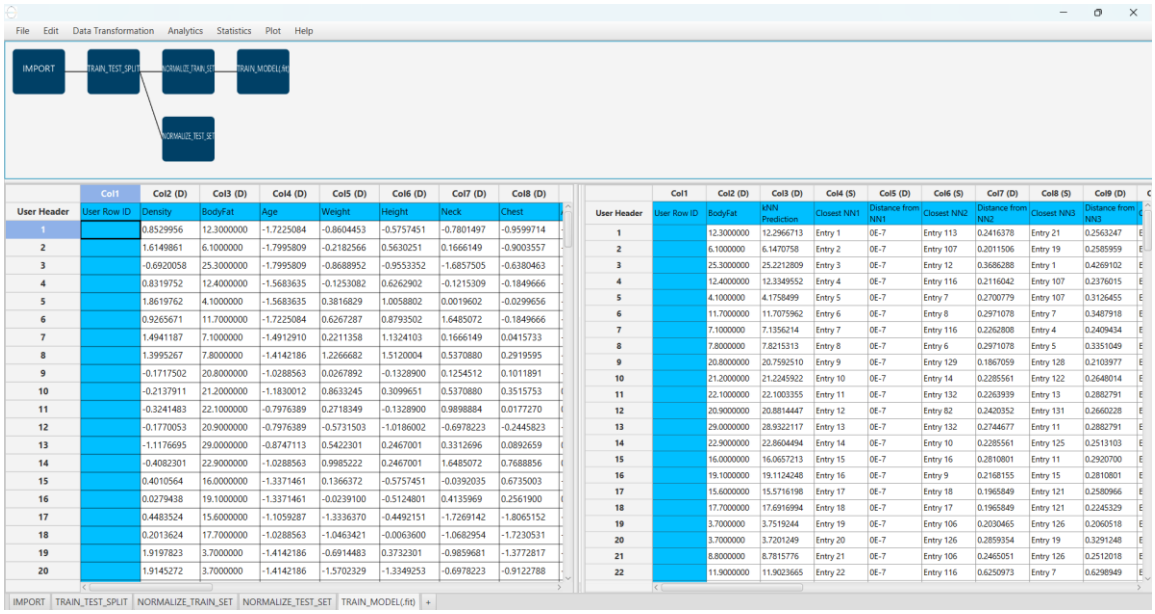
Use the k Nearest Neighbors (kNN) method to train and fit the model by browsing: "Analytics" → "Regression" → "k Nearest Neighbors (kNN)" and set the "Target Column" as the column corresponding to "BodyFat" and the "Number of Neighbors" to 5.

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)
User Header	User Row ID	Density	BodyFat	Age	Weight	Height	Neck
1		0.8529956	12.3000000	-1.7225084	-0.8604453	-0.5757451	-0.7801497
2		1.6149861	6.1000000	-1.7995809	-0.2182566	0.5630251	0.1666149
3		-0.6920058	25.3000000	-1.7995809	-0.8688952	-0.9553352	-1.6857505
4		0.8319752	12.4000000	-1.5683635	-0.1253082	0.6262902	-0.1215309
5		1.8619762	4.1000000	-1.5683635	0.3816829	1.0058802	0.0019602
6		0.9265671	11.7000000	-1.7225084	0.6267287	0.8793502	1.6485072
7		1.4941187	7.1000000	-1.4912910	0.2211358	1.1324103	0.1666149
8		1.3995267	7.8000000	-1.4142186	1.2266682	1.5120004	0.5370880
9		-0.1717502	20.8000000	-1.0288563	0.0267892	-0.1328900	0.1254512
10		-0.2137911	21.2000000	-1.1830012	0.8633245	0.3099651	0.5370880
11		-0.3241483	22.1000000	-0.7976389	0.2718349	-0.1328900	0.9898884
12		-0.1770053	20.9000000	-0.7976389	-0.5731503	-1.0186002	-0.6978223
13		-1.1176695	29.0000000	-0.8747113	0.5422301	0.2467001	0.3312696
14		-0.4082301	22.9000000	-1.0288563	0.9985222	0.2467001	1.6485072
15		0.4010564	16.0000000	-1.3371461	0.1366372	-0.5757451	-0.0392035
16		0.0279438	19.1000000	-1.3371461	-0.0239100	-0.5124801	0.4135969
17		0.4483524	15.6000000	-1.1059287	-1.3336370	-0.4492151	-1.7269142
18		0.2013624	17.7000000	-1.0288563	-1.0463421	-0.0063600	-1.0682954
19		1.9197823	3.7000000	-1.4142186	-0.6914483	0.3732301	-0.9859681
20		1.9145272	3.7000000	-1.4142186	-1.5702329	-1.3349253	-0.6978223

IMPORT | TRAIN_TEST_SPLIT | NORMALIZE_TRAIN_SET | NORMALIZE_TEST_SET | TRAIN_MODEL(.fit) | +



The predictions 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	Density	BodyFat	Age	Weight	Height	Neck	Chest
1		0.8529956	12.3000000	-1.7225084	-0.8604453	-0.5757451	-0.7801497	-0.9599714
2		1.6149861	6.1000000	-1.7995809	-0.2182566	0.5630251	0.1666149	-0.9003557
3		-0.6920058	25.3000000	-1.7995809	-0.8688952	-0.9553352	-1.6857505	-0.6380463
4		0.8319752	12.4000000	-1.5683635	-0.1253082	0.6262902	-0.1215309	-0.1849666
5		1.8619762	4.1000000	-1.5683635	0.3816829	1.0058802	0.0019602	-0.0299656
6		0.9265671	11.7000000	-1.7225084	0.6267287	0.8793502	1.6485072	-0.1849666
7		1.4941187	7.1000000	-1.4912910	0.2211358	1.1324103	0.1666149	0.0415733
8		1.3995267	7.8000000	-1.4142186	1.2266682	1.5120004	0.5370880	0.2919595
9		-0.1717502	20.8000000	-1.0288563	0.0267892	-0.1328900	0.1254512	0.1011891
10		-0.2137911	21.2000000	-1.1830012	0.8633245	0.3099651	0.5370880	0.3515753
11		-0.3241483	22.1000000	-0.7976389	0.2718349	-0.1328900	0.9898884	0.0177270
12		-0.1770053	20.9000000	-0.7976389	-0.5731503	-1.0186002	-0.6978223	-0.2445823
13		-1.1176695	29.0000000	-0.8747113	0.5422301	0.2467001	0.3312696	0.0892659
14		-0.4082301	22.9000000	-1.0288563	0.9985222	0.2467001	1.6485072	0.7688856
15		0.4010564	16.0000000	-1.3371461	0.1366372	-0.5757451	-0.0392035	0.6735003
16		0.0279438	19.1000000	-1.3371461	-0.0239100	-0.5124801	0.4135969	0.2561900
17		0.4483524	15.6000000	-1.1059287	-1.3336370	-0.4492151	-1.7269142	-1.8065152
18		0.2013624	17.7000000	-1.0288563	-1.0463421	-0.0063600	-1.0682954	-1.7230531
19		1.9197823	3.7000000	-1.4142186	-0.6914483	0.3732301	-0.9859681	-1.3772817
20		1.9145272	3.7000000	-1.4142186	-1.5702329	-1.3349253	-0.6978223	-0.9122788

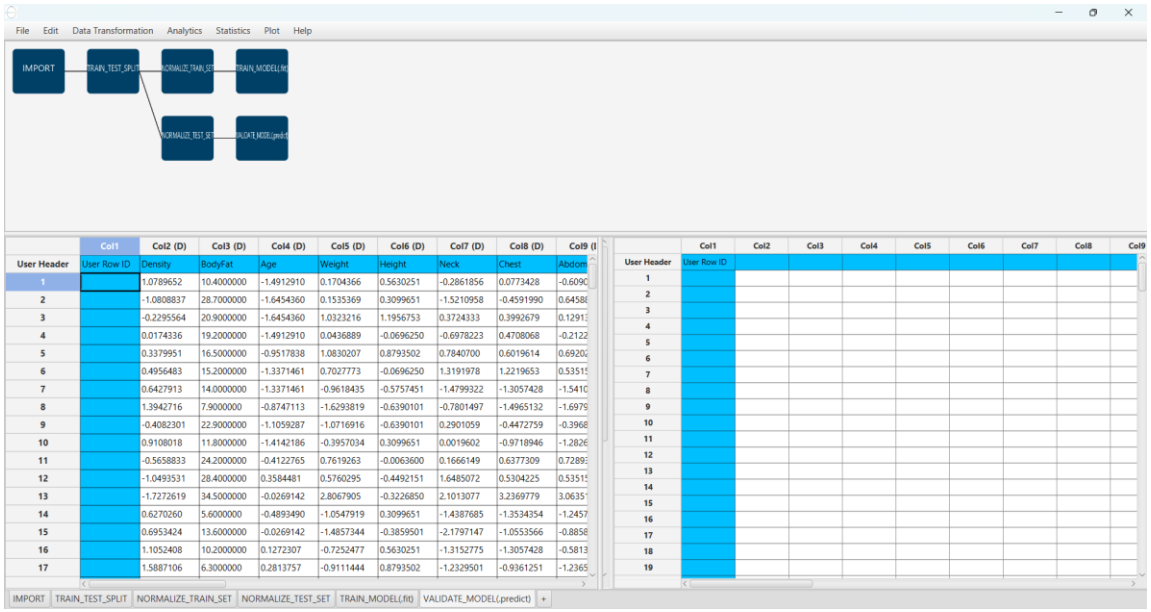
	Col1	Col2 (D)	Col3 (D)	Col4 (S)	Col5 (D)	Col6 (S)	Col7 (D)	Col8 (S)	Col9 (D)
User Header	User Row ID	BodyFat	kNN Prediction	Closest NN1	Distance from NN1	Closest NN2	Distance from NN2	Closest NN3	Distance from NN3
1		12.3000000	12.2966713	Entry 1	0E-7	Entry 113	0.2416378	Entry 21	0.2563247
2		6.1000000	6.1470758	Entry 2	0E-7	Entry 107	0.2011506	Entry 19	0.2585959
3		25.3000000	25.2212809	Entry 3	0E-7	Entry 12	0.3686288	Entry 1	0.4269102
4		12.4000000	12.3349552	Entry 4	0E-7	Entry 116	0.2116042	Entry 107	0.2376015
5		4.1000000	4.1758499	Entry 5	0E-7	Entry 7	0.2700779	Entry 107	0.3126455
6		11.7000000	11.7075962	Entry 6	0E-7	Entry 8	0.2971078	Entry 7	0.3487918
7		7.1000000	7.1356214	Entry 7	0E-7	Entry 116	0.2262808	Entry 4	0.2409434
8		7.8000000	7.8215313	Entry 8	0E-7	Entry 6	0.2971078	Entry 5	0.3351049
9		20.8000000	20.7592510	Entry 9	0E-7	Entry 129	0.1867059	Entry 128	0.2103977
10		21.2000000	21.2245922	Entry 10	0E-7	Entry 14	0.2285561	Entry 122	0.2648014
11		22.1000000	22.1003355	Entry 11	0E-7	Entry 132	0.2263939	Entry 13	0.2882791
12		20.9000000	20.8814447	Entry 12	0E-7	Entry 82	0.2420352	Entry 131	0.2660228
13		29.0000000	28.9322117	Entry 13	0E-7	Entry 132	0.2744677	Entry 11	0.2882791
14		22.9000000	22.8604494	Entry 14	0E-7	Entry 10	0.2285561	Entry 125	0.2513103
15		16.0000000	16.0657213	Entry 15	0E-7	Entry 16	0.2810801	Entry 11	0.2920700
16		19.1000000	19.1124248	Entry 16	0E-7	Entry 9	0.2168155	Entry 15	0.2810801
17		15.6000000	15.5716198	Entry 17	0E-7	Entry 18	0.1965849	Entry 121	0.2580966
18		17.7000000	17.6916994	Entry 18	0E-7	Entry 17	0.1965849	Entry 121	0.2245329
19		3.7000000	3.7519244	Entry 19	0E-7	Entry 106	0.2030465	Entry 126	0.2060518
20		3.7000000	3.7201249	Entry 20	0E-7	Entry 126	0.2859354	Entry 19	0.3291248
21		8.8000000	8.7815776	Entry 21	0E-7	Entry 106	0.2465051	Entry 126	0.2512018
22		11.9000000	11.9023665	Entry 22	0E-7	Entry 116	0.6250973	Entry 7	0.6298949

IMPORT | TRAIN_TEST_SPLIT | NORMALIZE_TRAIN_SET | NORMALIZE_TEST_SET | TRAIN_MODEL(.fit) | +

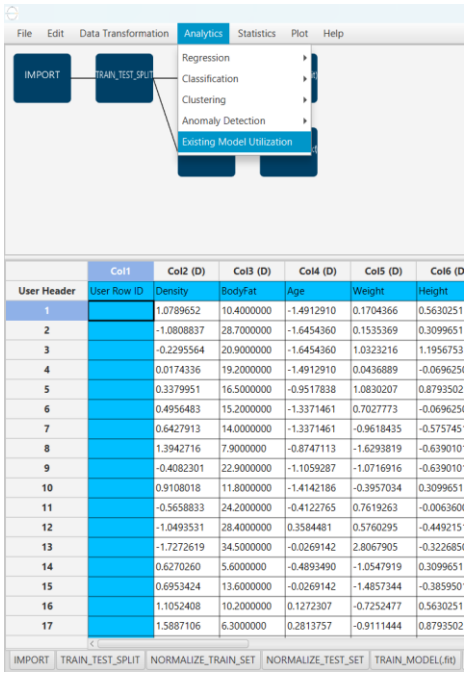
Step 7: Validate the model

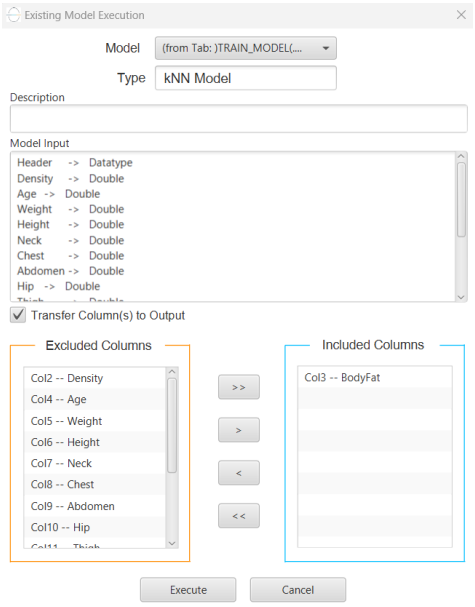
Create a new tab by pressing the "+" button on the bottom of the page with the name "VALIDATE_MODEL(.predict)".

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".

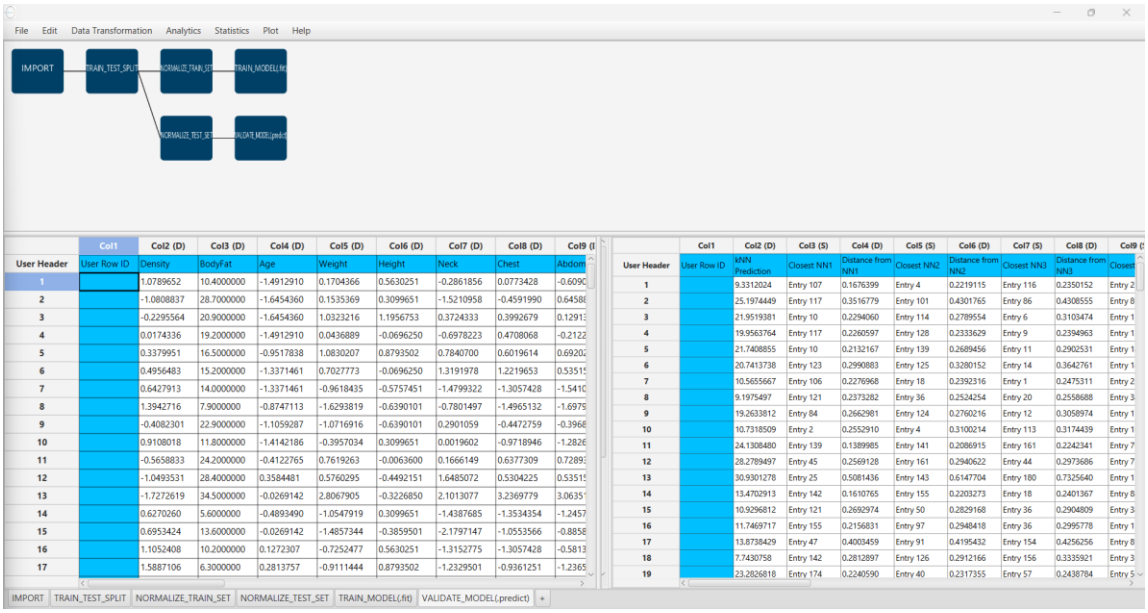


To validate the model browse: "Analytics" → "Existing Model Utilization". Then choose Model "(from Tab:) TRAIN_MODEL (.fit)". and transfer the "BodyFat" column in the output.





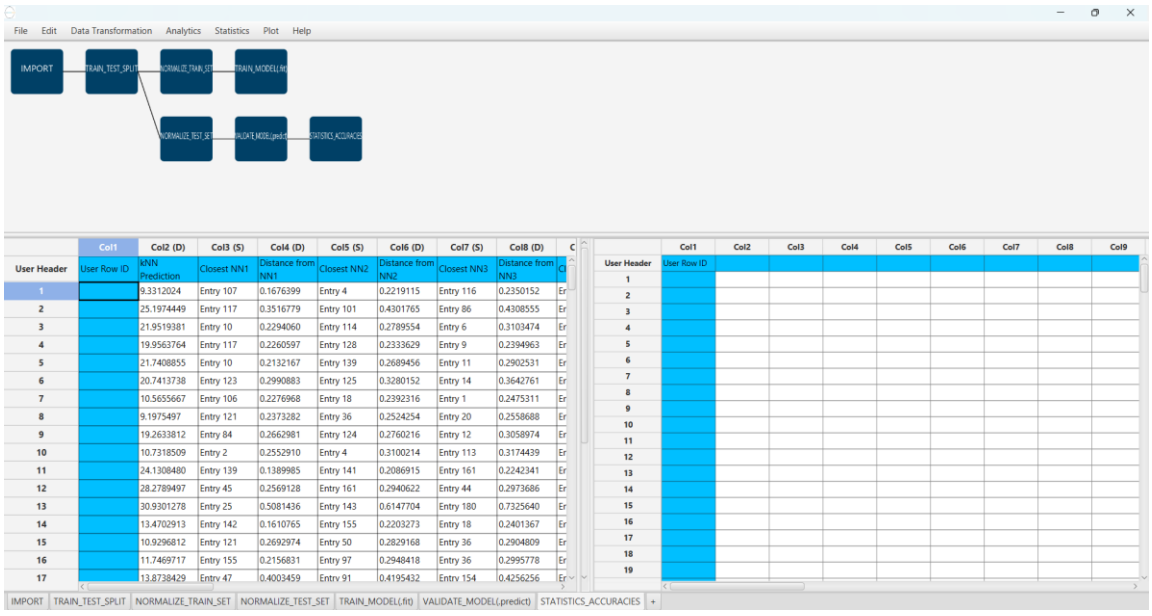
The predictions will appear on the output spreadsheet.



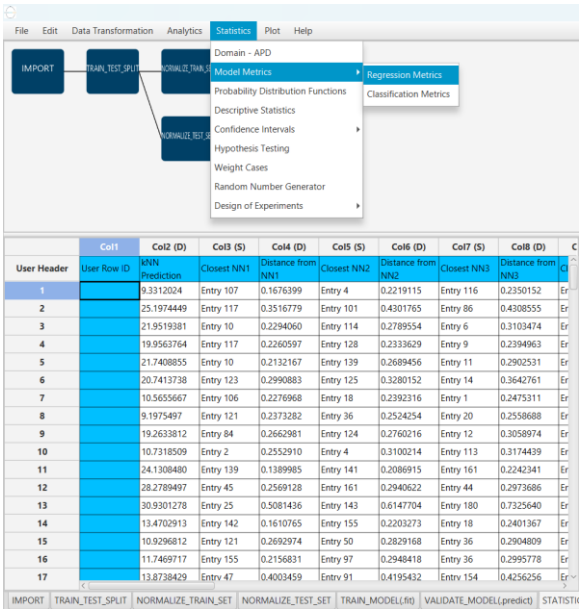
Step 8: Statistics calculation

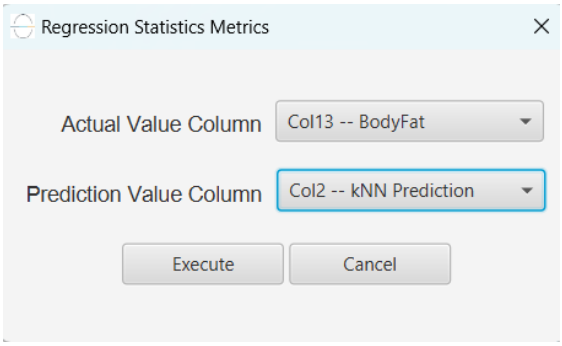
Create a new tab by pressing the "+" button on the bottom of the page with the name "STATISTICS_ACCURACIES".

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".

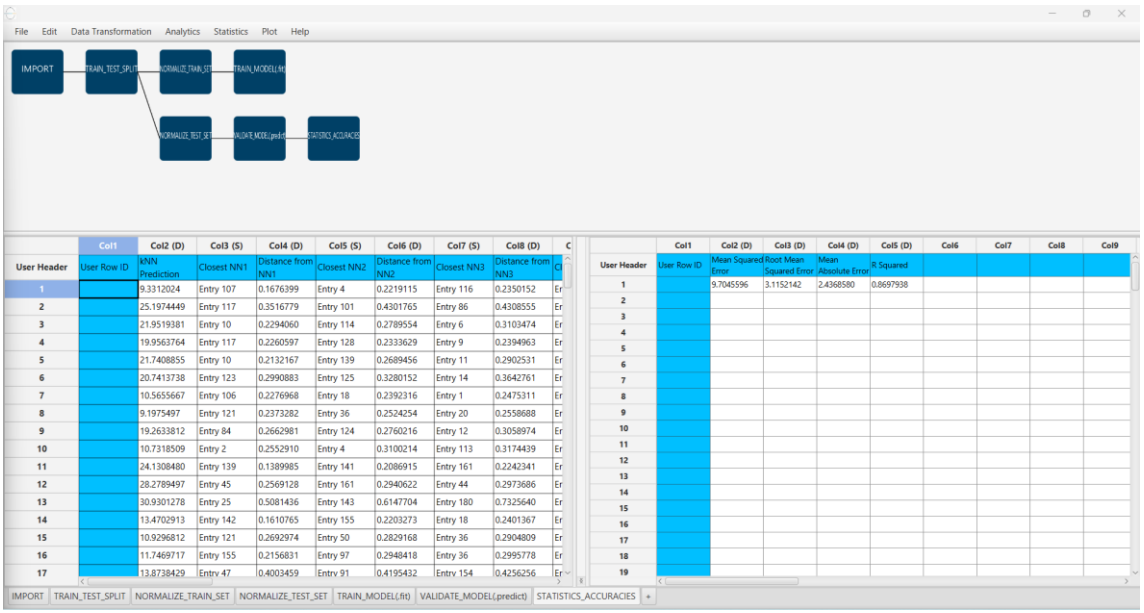


Calculate the statistical metrics for the regression by browsing: "Statistics" → "Model Metrics" → "Regression Metrics".





The results will appear on the output spreadsheet.

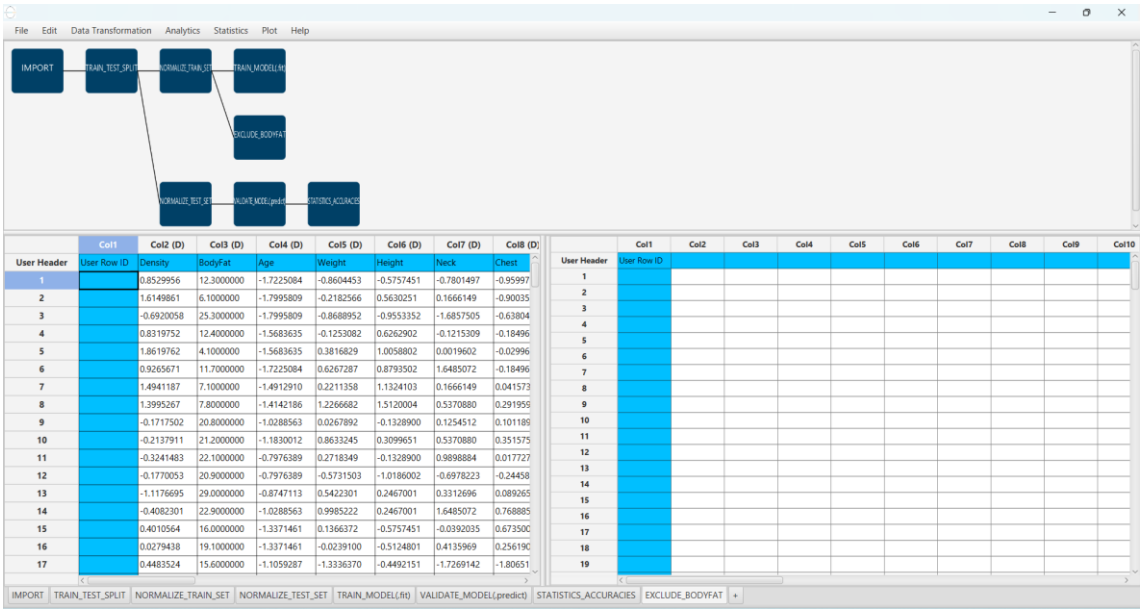


Step 9: Reliability check of each record of the test set

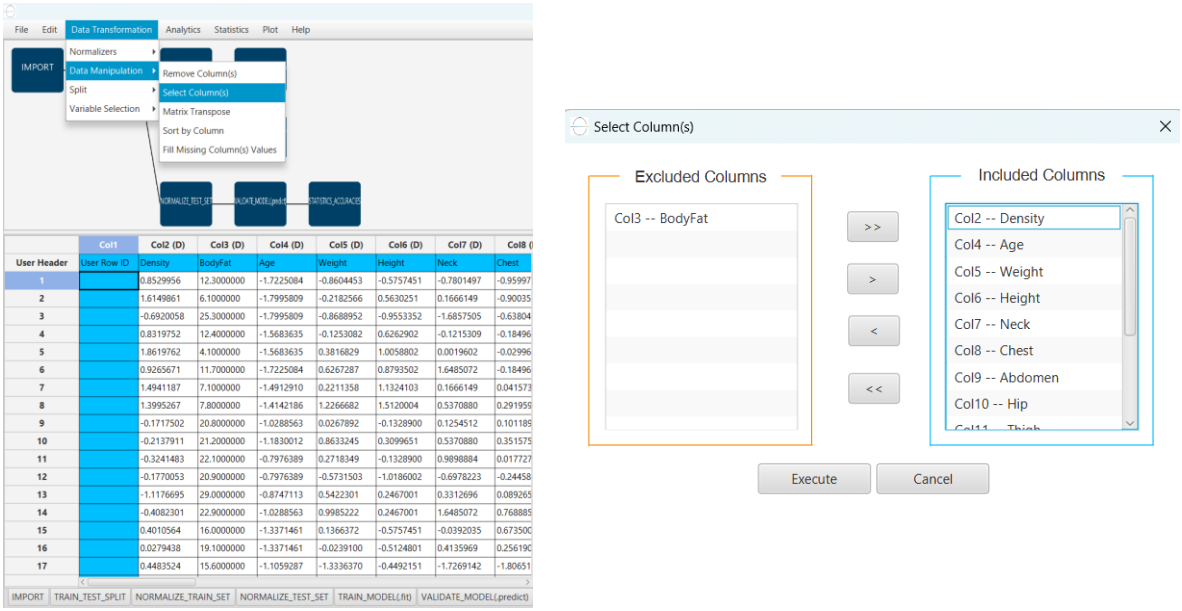
Step 9.a: Create the domain

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

Import data into the input spreadsheet of the "EXCLUDE_ BODYFAT" tab from the output of the "NORMALIZE _TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".

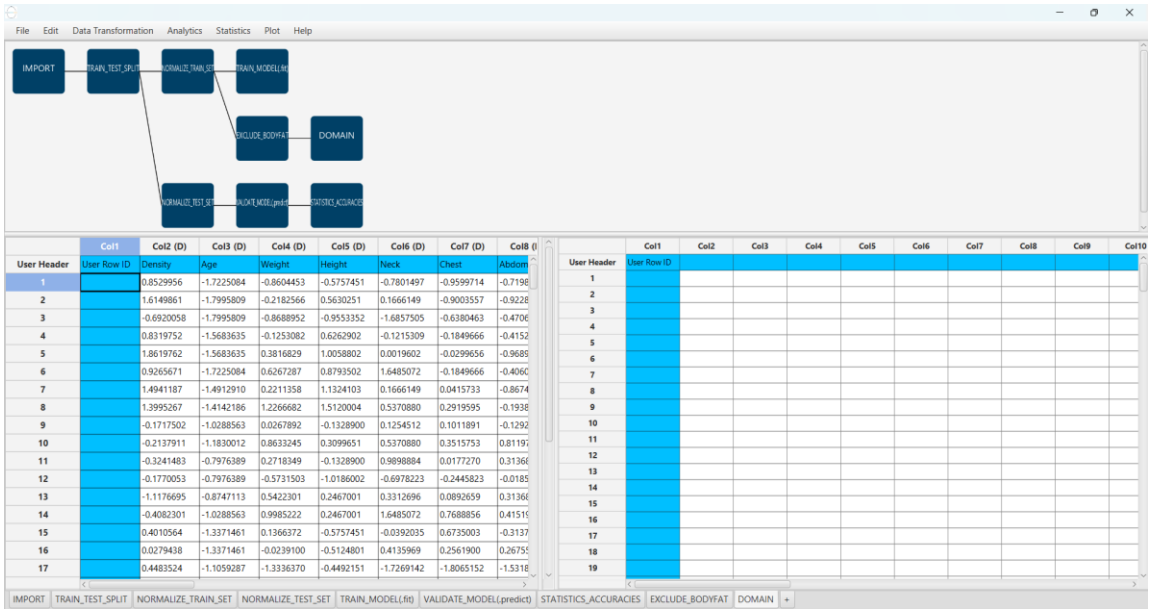


Manipulate the data to exclude the column that corresponds to the "BodyFat" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except the "BodyFat". The filtered table will appear on the output spreadsheet.

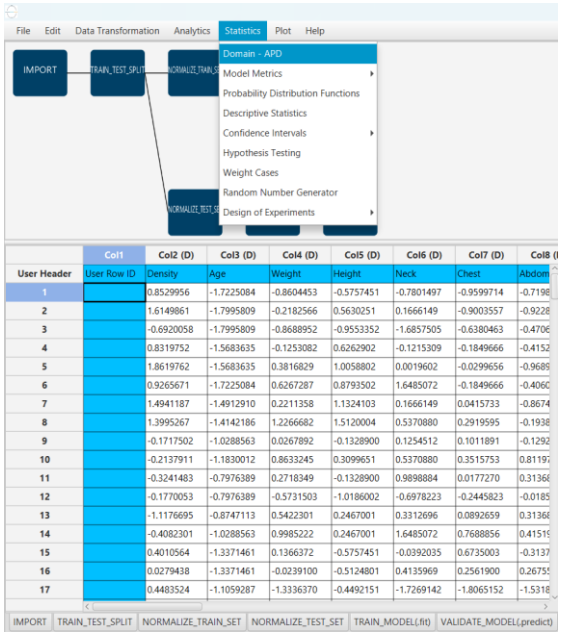


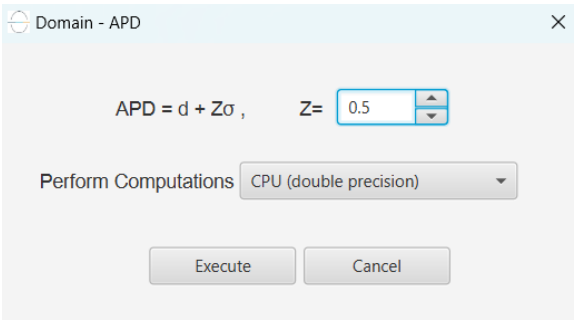
Create a new tab by pressing the "+" button on the bottom of the page with the name "DOMAIN".

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

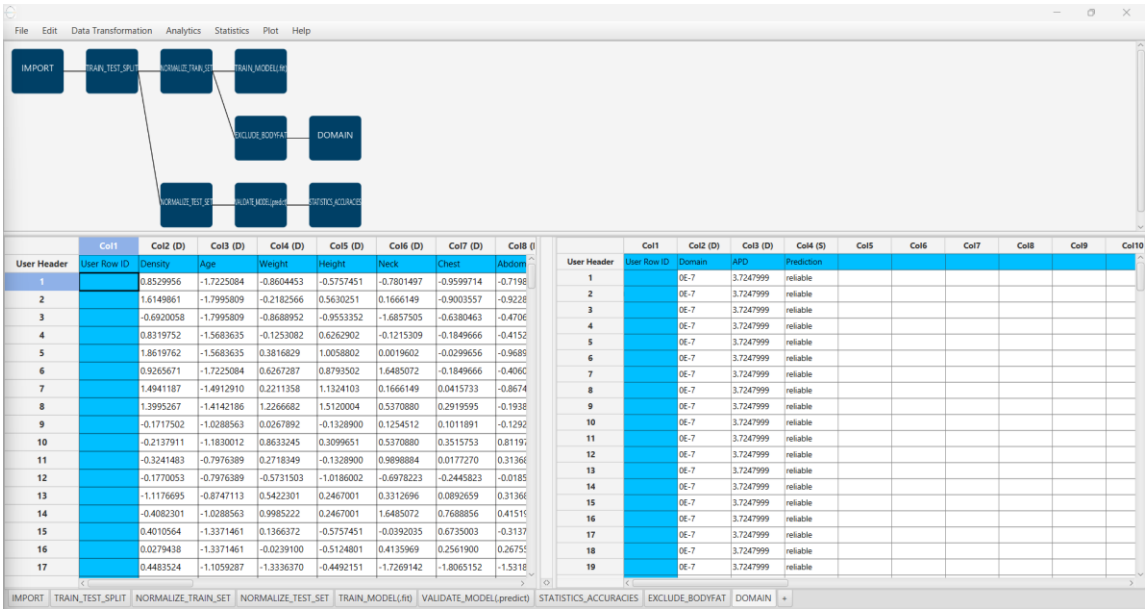


Create the domain of applicability by browsing: "Statistics" → "Domain APD".





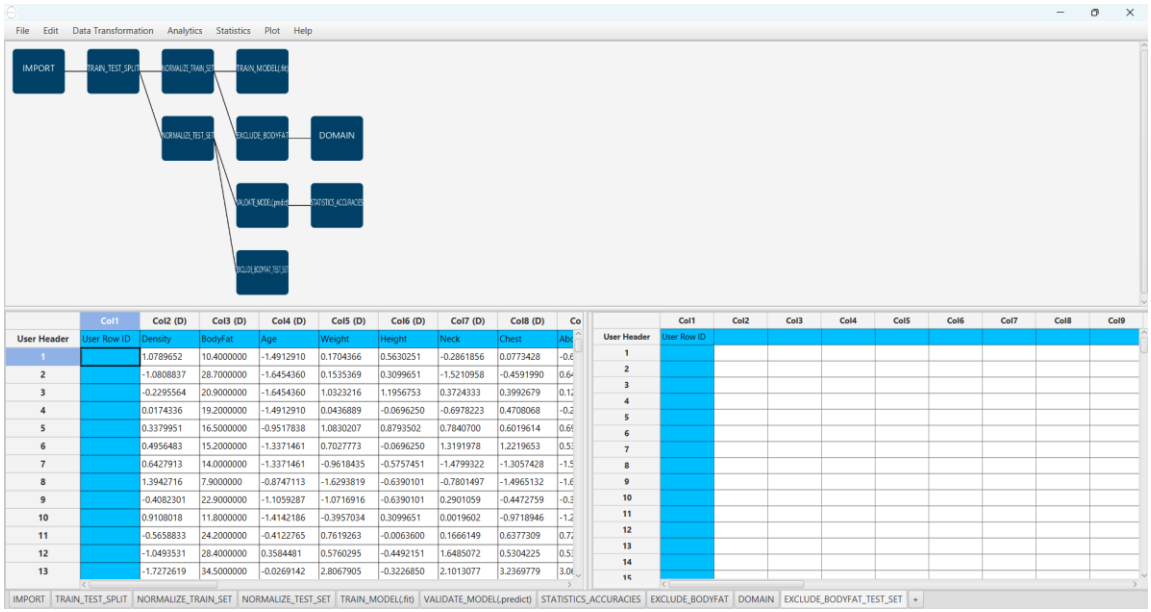
The results will appear on the output spreadsheet.



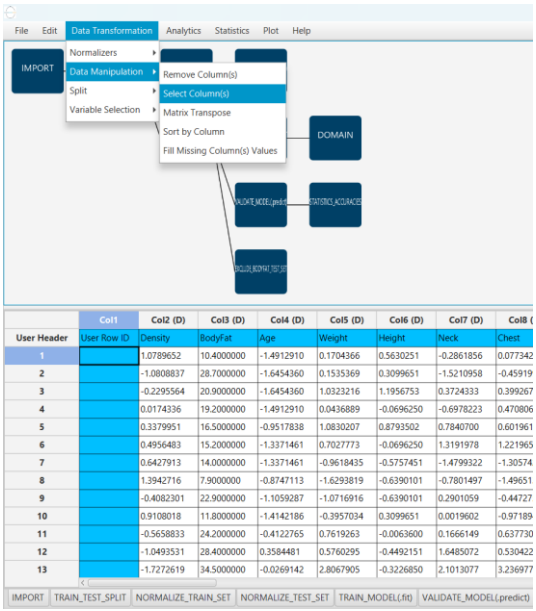
Step 9.b: Check the test set reliability

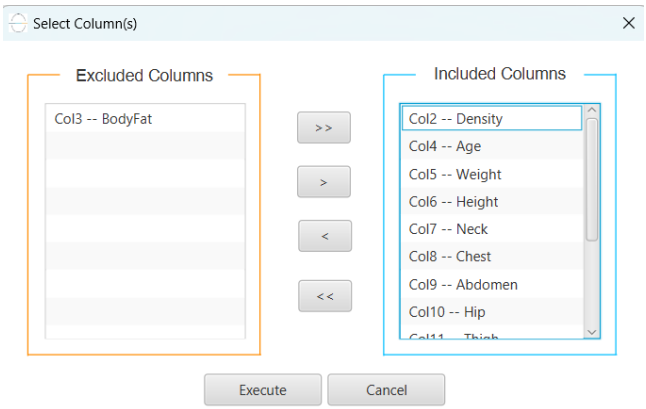
Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_BODYFAT_TEST_SET".

Import data into the input spreadsheet of the "EXCLUDE_ BODYFAT _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".



Filter the data to exclude the column that corresponds to the "BodyFat" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except "BodyFat".

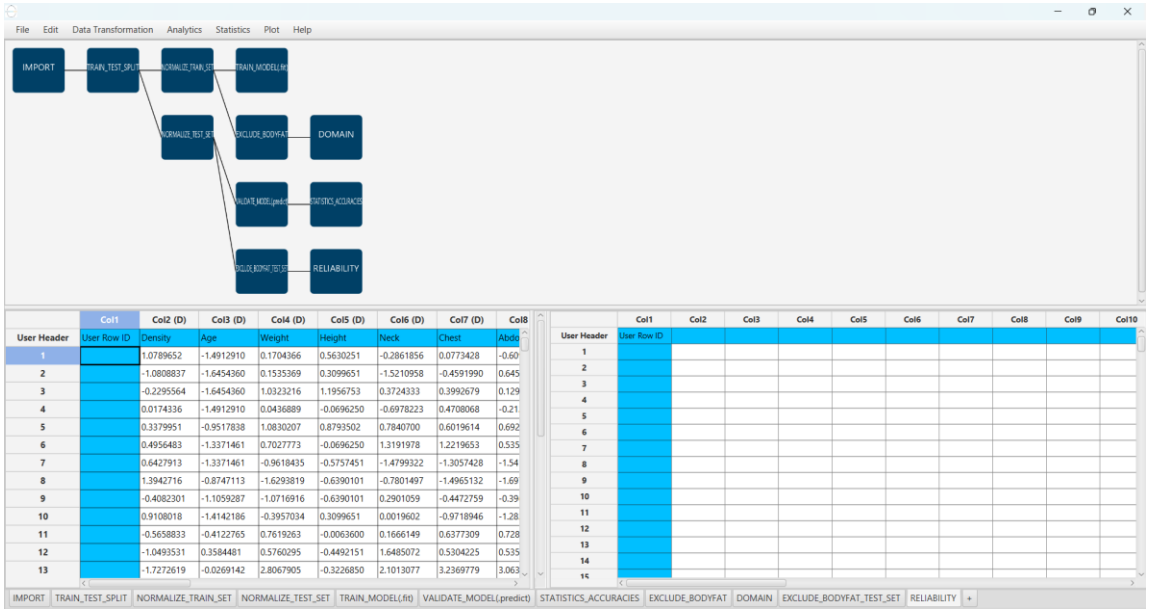




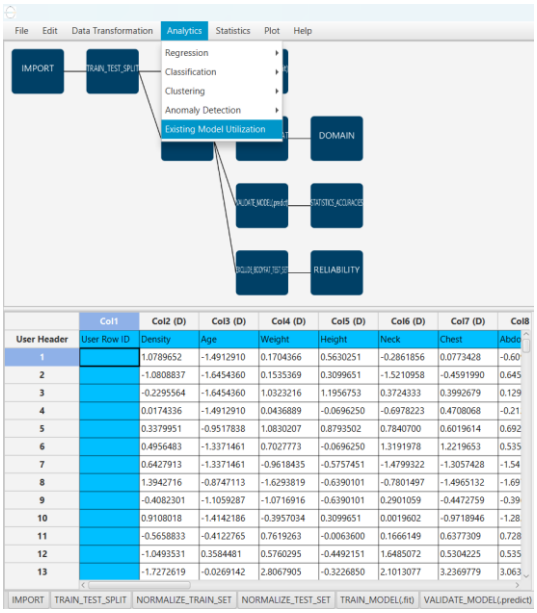
The results will appear on the output spreadsheet.

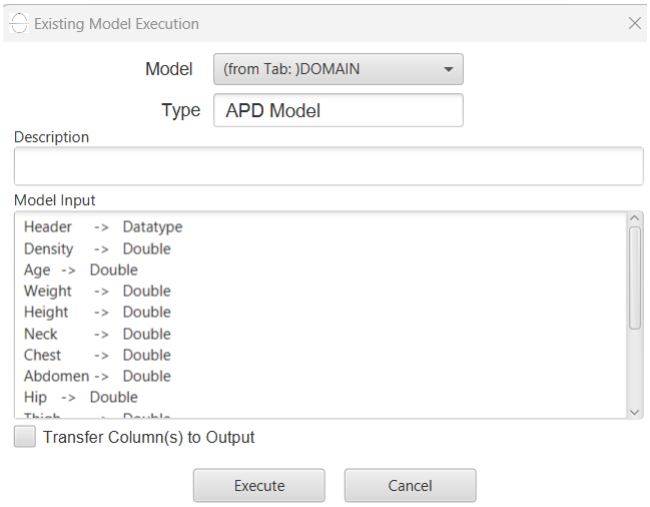
Create a new tab by pressing the "+" button on the bottom of the page with the name "RELIABILITY".

Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "EXCLUDE_ BODYFAT _TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".

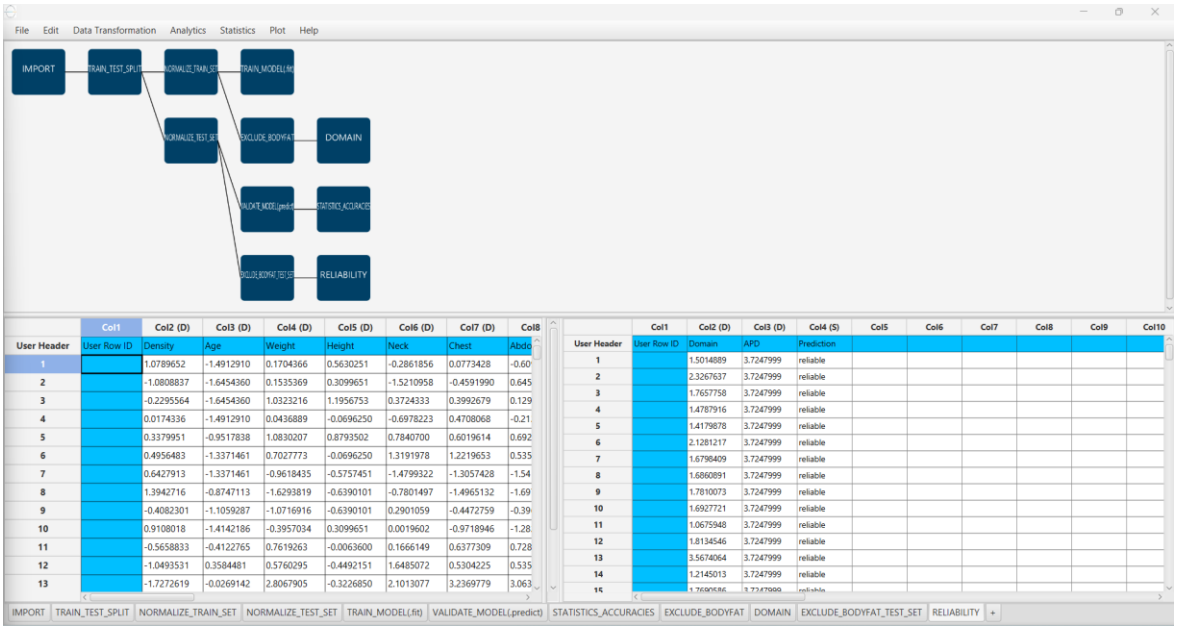


Check the predictions' reliability by browsing: "Analytics" → "Existing Model Utilization". Then select as Model "(from Tab:) DOMAIN".





The results will appear on the output spreadsheet. We can observe that there are no unreliable samples in the test set.



Final Isalos Workflow

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