Insurance Charges (Regression)

The goal of this study is to train a model in order to predict insurance charges. The dataset used in this case study is found in https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender/data and has 8 features and 1338 labelled samples. This dataset contains detailed information about insurance customers, including their age, sex, body mass index (BMI), number of children, smoking status and region. Having access to such valuable insights allows analysts to get a better view into customer behaviour and the factors that contribute to their insurance charges.

The dataset contains no missing values and includes several categorical features. Categorical features contain multiple levels, and the data was transformed to corresponding numeric codes, as detailed below:

Gender:

- Female (0)
- Male (1)

Smoker:

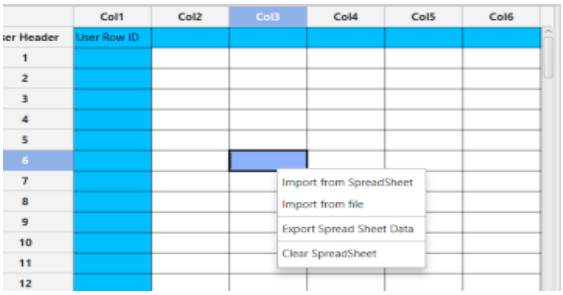
- No (0)
- Yes (1)

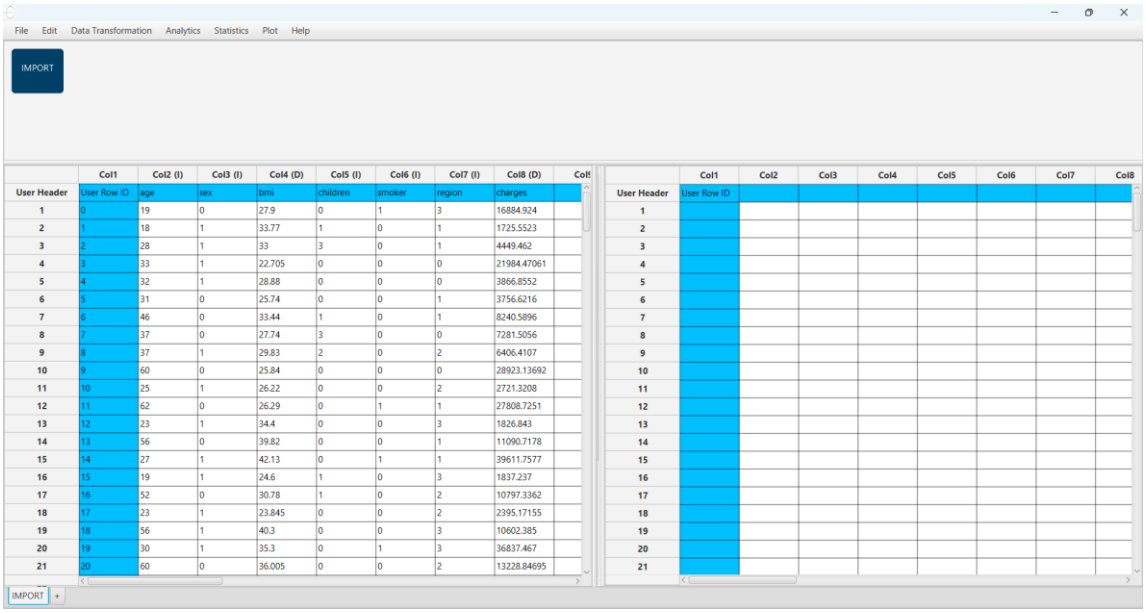
Region:

- Northwest (0)
- Southeast (1)
- Northeast (2)
- Southwest (3)

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 salary data.

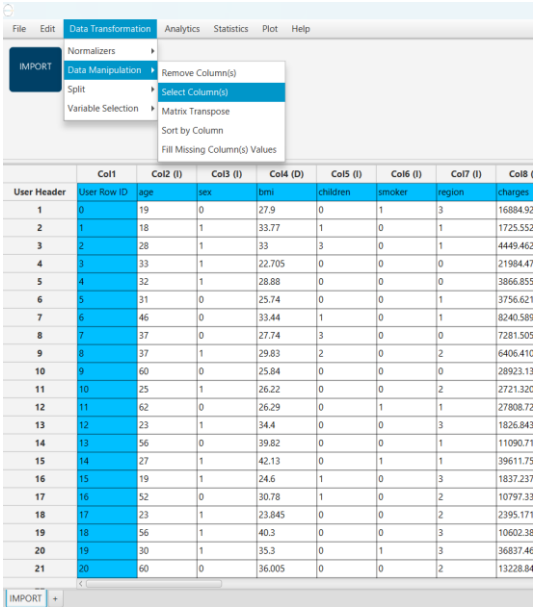


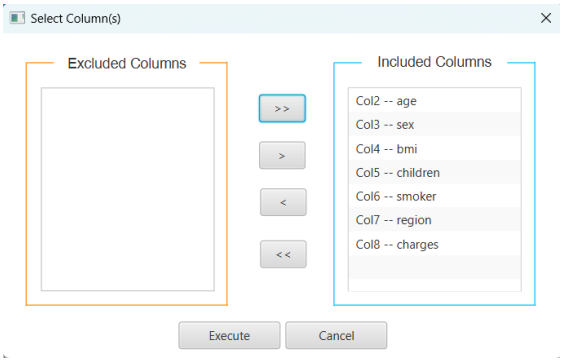


Step 2: Manipulate data

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.



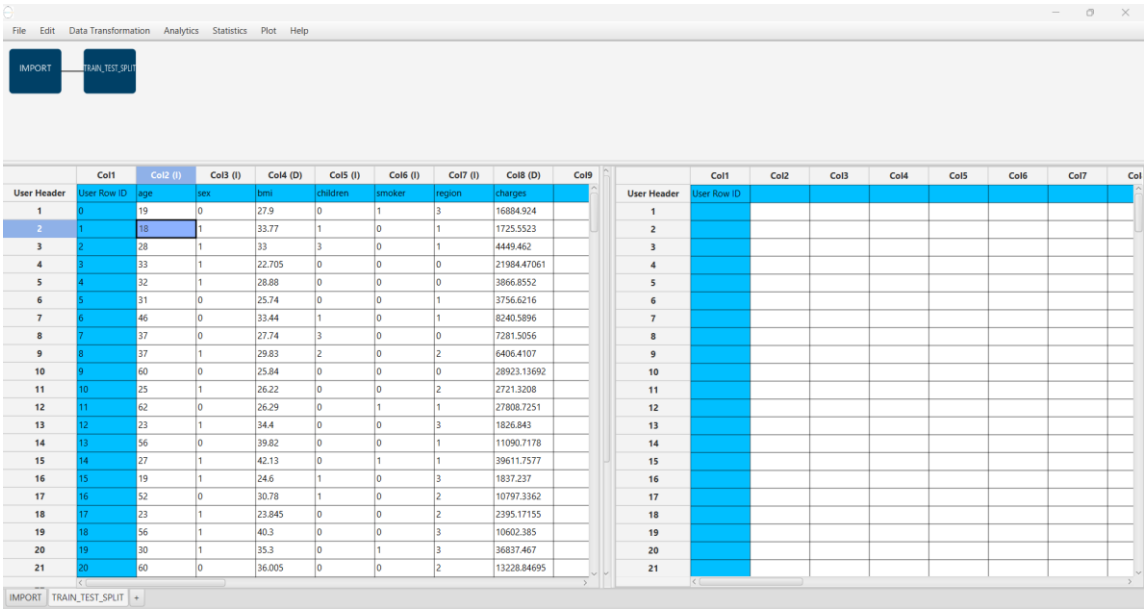


The data will appear in the output spreadsheet.

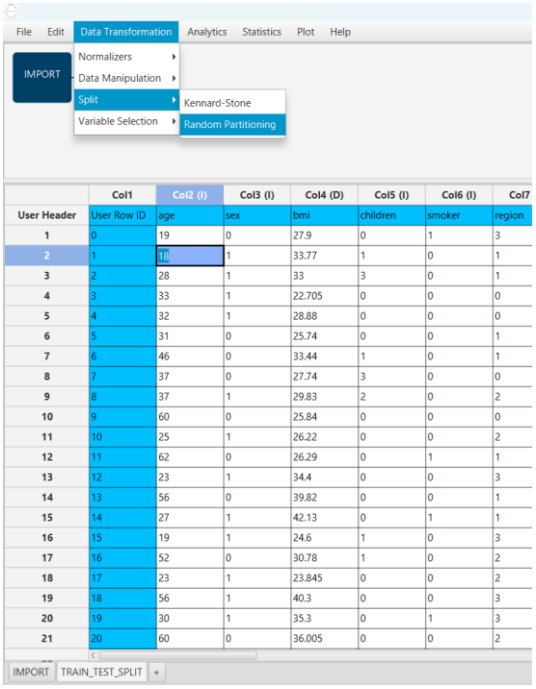
Step 3: Split data

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.

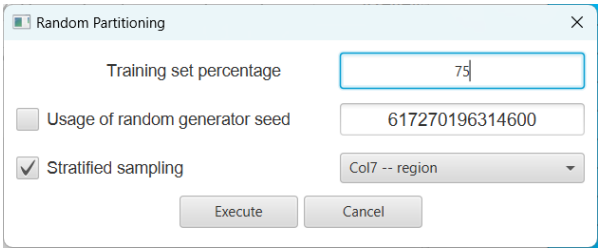
Import data into the input spreadsheet of the "TRAIN_TEST_SPLIT" tab from the output of the "IMPORT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



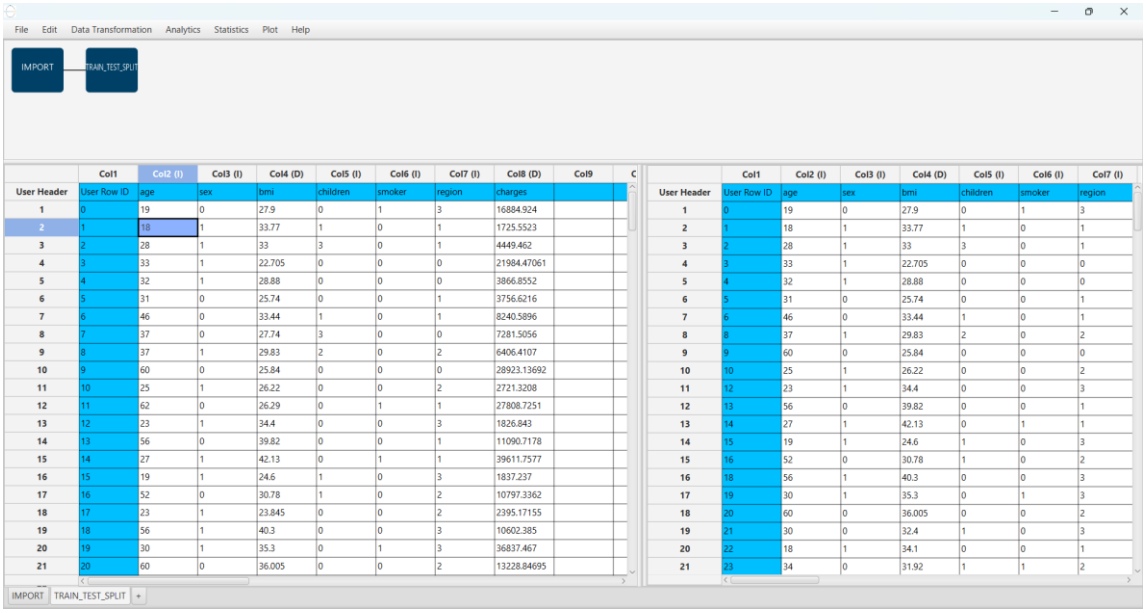
Split the dataset by choosing by browsing: "Data Transformation" → "Split" → "Random Partitioning". Then choose the "Training set percentage" and the column for the sampling as shown below:



	Col1	Col2 (I)	Col3 (I)	Col4 (D)	Col5 (I)	Col6 (I)	Col7
User Header	User Row ID	age	sex	bmi	children	smoker	region
1	0	19	0	27.9	0	1	3
2	1	18	1	33.77	1	0	1
3	2	28	1	33	3	0	1
4	3	33	1	22.705	0	0	0
5	4	32	1	28.88	0	0	0
6	5	31	0	25.74	0	0	1
7	6	46	0	33.44	1	0	1
8	7	37	0	27.74	3	0	0
9	8	37	1	29.83	2	0	2
10	9	60	0	25.84	0	0	0
11	10	25	1	26.22	0	0	2
12	11	62	0	26.29	0	1	1
13	12	23	1	34.4	0	0	3
14	13	56	0	39.82	0	0	1
15	14	27	1	42.13	0	1	1
16	15	19	1	24.6	1	0	3
17	16	52	0	30.78	1	0	2
18	17	23	1	23.845	0	0	2
19	18	56	1	40.3	0	0	3
20	19	30	1	35.3	0	1	3
21	20	60	0	36.005	0	0	2



The results will appear on the output spreadsheet.



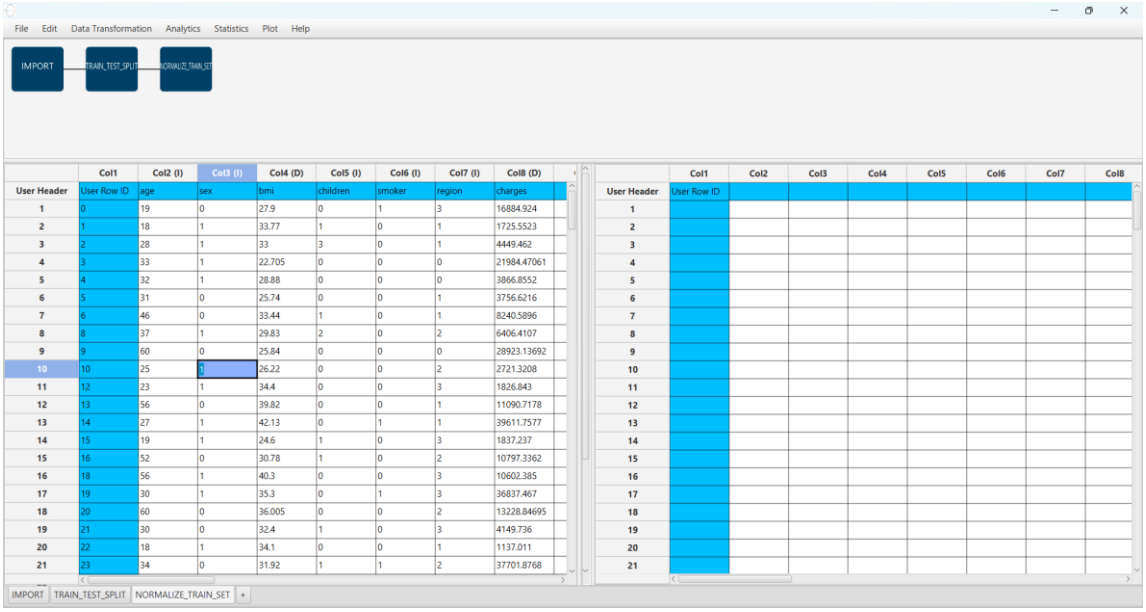
	Col1	Col2 (I)	Col3 (I)	Col4 (D)	Col5 (I)	Col6 (I)	Col7 (I)	Col8 (D)	Col9
User Header	User Row ID	age	sex	bmi	children	smoker	region	charges	
1	0	19	0	27.9	0	1	3	16884.924	
2	1	18	1	33.77	1	0	1	1725.5523	
3	2	28	1	33	3	0	1	4449.462	
4	3	33	1	22.705	0	0	0	21984.47061	
5	4	32	1	28.88	0	0	0	3866.8552	
6	5	31	0	25.74	0	0	1	3756.6216	
7	6	46	0	33.44	1	0	1	8240.5896	
8	7	37	0	27.74	3	0	0	7281.5056	
9	8	37	1	29.83	2	0	2	6406.4107	
10	9	60	0	25.84	0	0	0	28923.13692	
11	10	25	1	26.22	0	0	2	2721.3208	
12	11	62	0	26.29	0	1	1	27808.7251	
13	12	23	1	34.4	0	0	3	1826.843	
14	13	56	0	39.82	0	0	1	11090.7178	
15	14	27	1	42.13	0	1	1	39611.7577	
16	15	19	1	24.6	1	0	3	1837.237	
17	16	52	0	30.78	1	0	2	10797.3362	
18	17	23	1	23.845	0	0	2	2395.17155	
19	18	56	1	40.3	0	0	3	10602.385	
20	19	30	1	35.3	0	1	3	36837.467	
21	20	60	0	36.005	0	0	2	13228.84695	

	Col1	Col2 (I)	Col3 (I)	Col4 (D)	Col5 (I)	Col6 (I)	Col7 (I)
User Header	User Row ID	age	sex	bmi	children	smoker	region
1	0	19	0	27.9	0	1	3
2	1	18	1	33.77	1	0	1
3	2	28	1	33	3	0	1
4	3	33	1	22.705	0	0	0
5	4	32	1	28.88	0	0	0
6	5	31	0	25.74	0	0	1
7	6	46	0	33.44	1	0	1
8	8	37	1	29.83	2	0	2
9	9	60	0	25.84	0	0	0
10	10	25	1	26.22	0	0	2
11	12	23	1	34.4	0	0	3
12	13	56	0	39.82	0	0	1
13	14	27	1	42.13	0	1	1
14	15	19	1	24.6	1	0	3
15	16	52	0	30.78	1	0	2
16	18	56	1	40.3	0	0	3
17	19	30	1	35.3	0	1	3
18	20	60	0	36.005	0	0	2
19	21	30	0	32.4	1	0	3
20	22	18	1	34.1	0	0	1
21	23	34	0	31.92	1	1	2

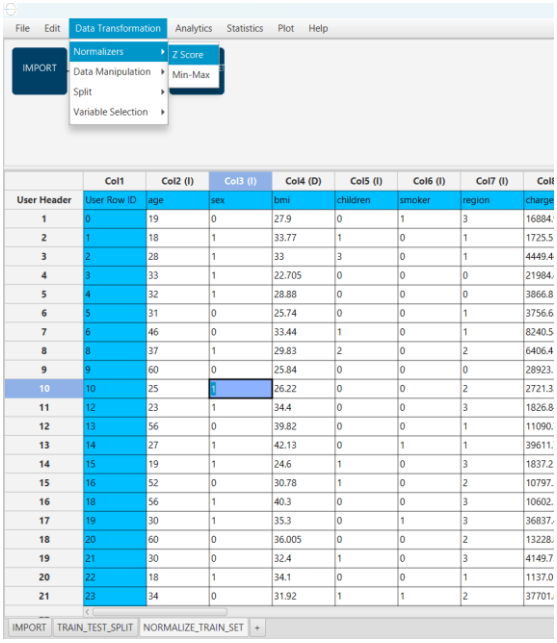
Step 4: Normalize the training set

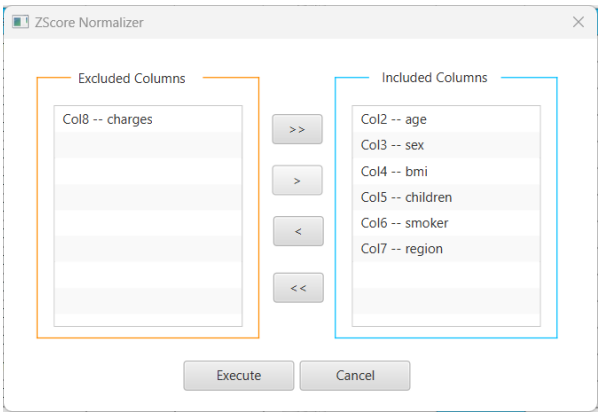
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALIZE_TRAIN_SET".

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"

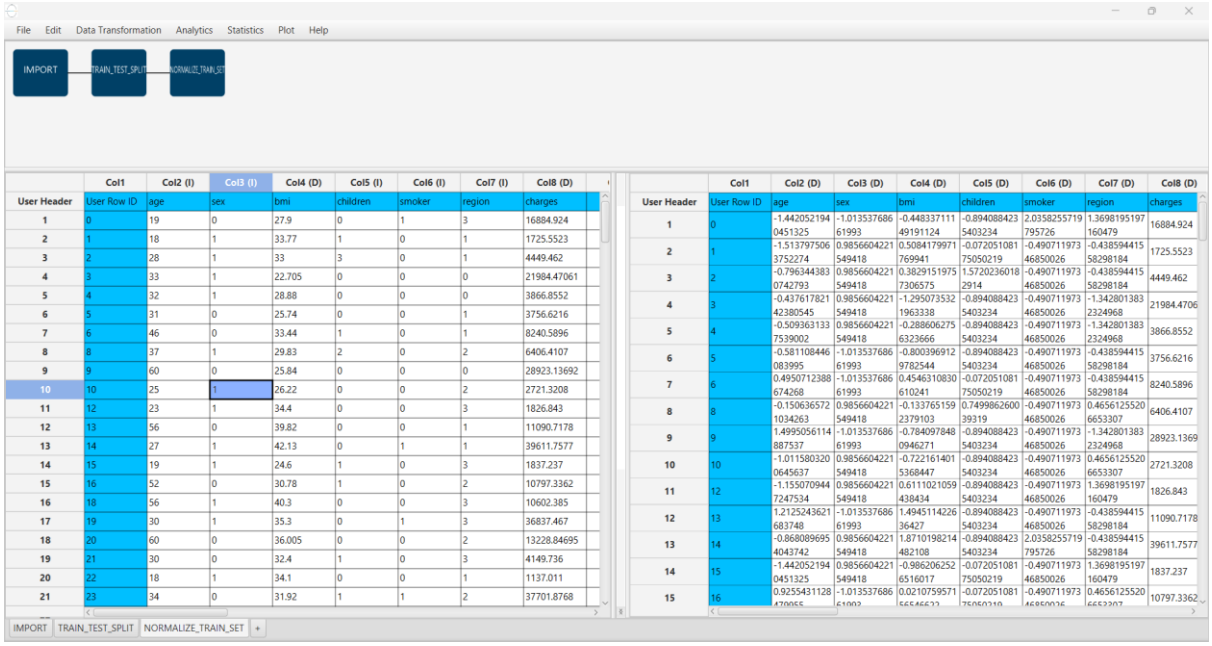


Normalize the data using Z-score by browsing: "Data Transformation" → "Normalizers" → "Z-Score". Then select all columns except "charges" and click "Execute".





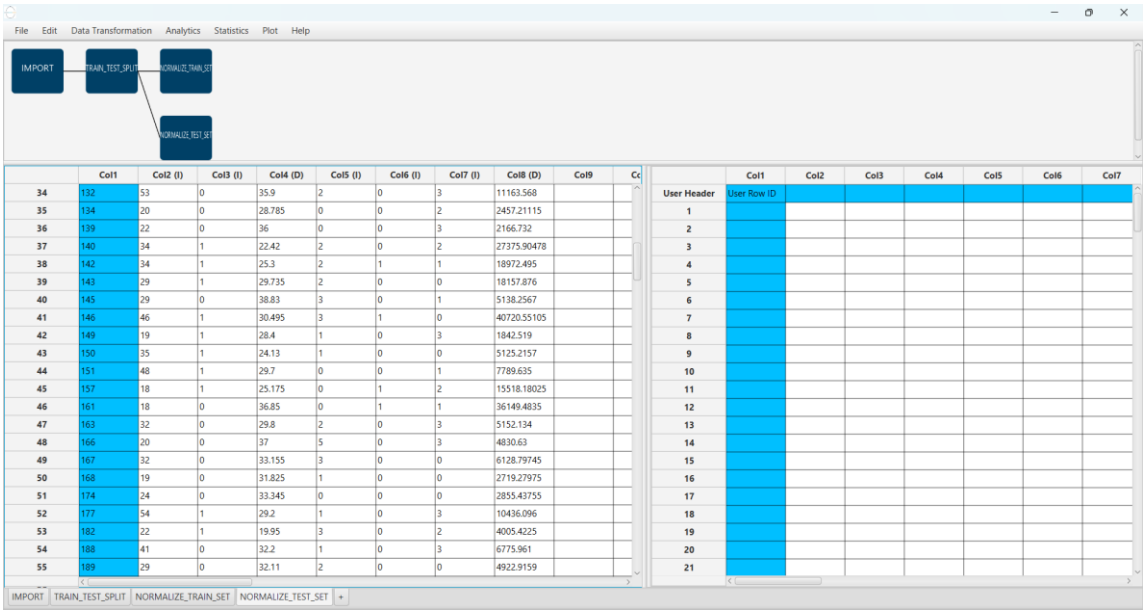
The results will appear on the output spreadsheet.



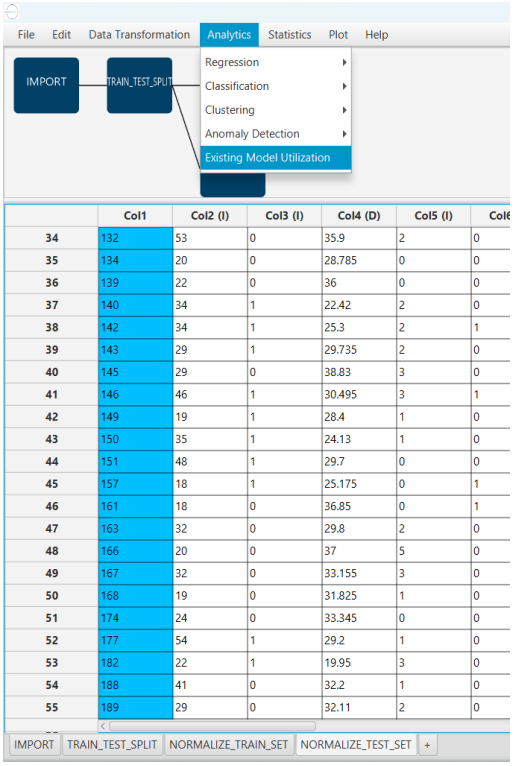
Step 5: Normalize the test set

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

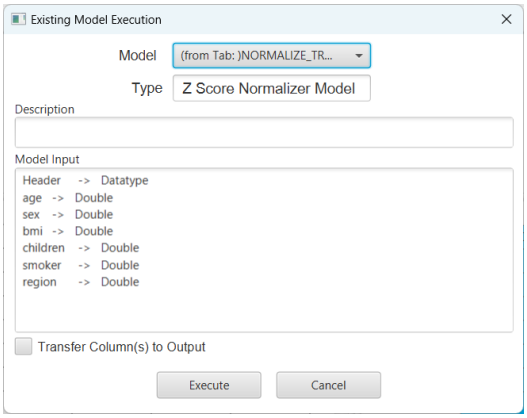
Import data into the input spreadsheet of the "NORMALIZE_TEST_SET" tab the test set from the output of the "TRAIN_TEST_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet". From the available Select input tab options choose "TRAIN_TEST_SPLIT: Test Set".



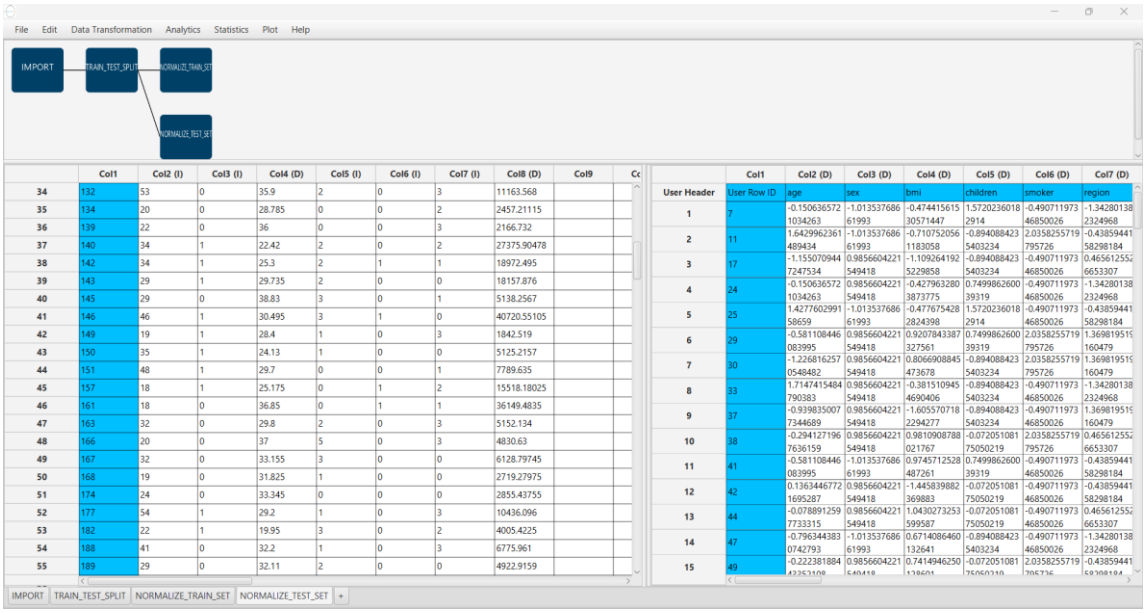
Normalize the test set using the existing normalizer of the training set by browsing: "Analytics" → "Existing Model Utilization" → "Model (from Tab:) NORMALIZE_TRAIN_SET".



	Col1	Col2 (I)	Col3 (I)	Col4 (D)	Col5 (I)	Col6
34	132	53	0	35.9	2	0
35	134	20	0	28.785	0	0
36	139	22	0	36	0	0
37	140	34	1	22.42	2	0
38	142	34	1	25.3	2	1
39	143	29	1	29.735	2	0
40	145	29	0	38.83	3	0
41	146	46	1	30.495	3	1
42	149	19	1	28.4	1	0
43	150	35	1	24.13	1	0
44	151	48	1	29.7	0	0
45	157	18	1	25.175	0	1
46	161	18	0	36.85	0	1
47	163	32	0	29.8	2	0
48	166	20	0	37	5	0
49	167	32	0	33.155	3	0
50	168	19	0	31.825	1	0
51	174	24	0	33.345	0	0
52	177	54	1	29.2	1	0
53	182	22	1	19.95	3	0
54	188	41	0	32.2	1	0
55	189	29	0	32.11	2	0



The results will appear on the output spreadsheet.



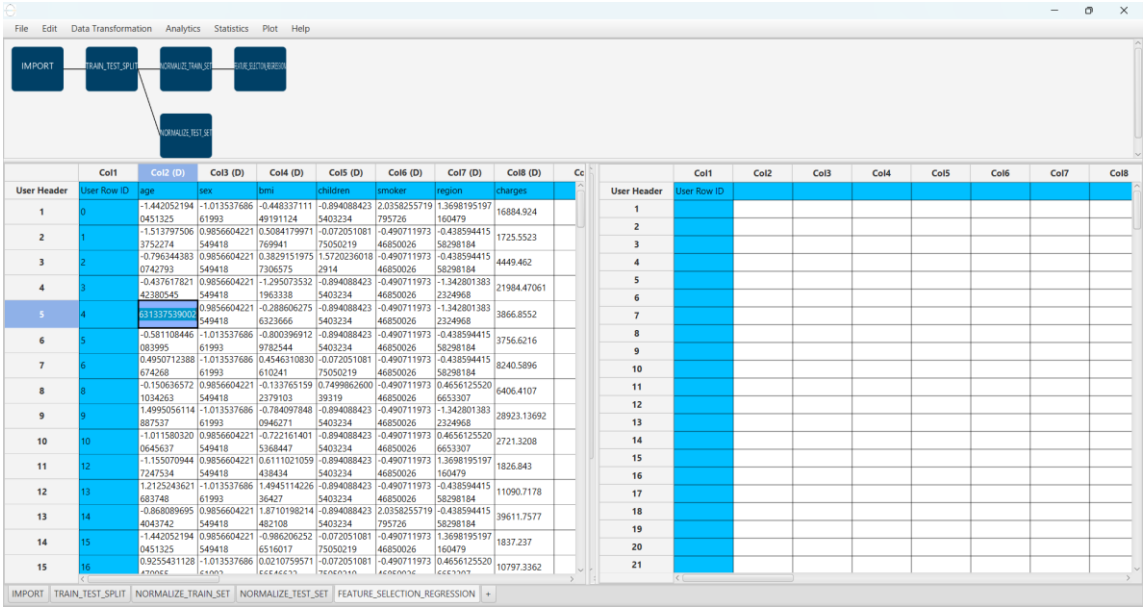
	Col1	Col2 (I)	Col3 (I)	Col4 (D)	Col5 (I)	Col6 (I)	Col7 (I)	Col8 (D)	Col9	Cc
34	132	53	0	35.9	2	0	3	11163.568		
35	134	20	0	28.785	0	0	2	2457.21115		
36	139	22	0	36	0	0	3	2166.732		
37	140	34	1	22.42	2	0	2	27375.90478		
38	142	34	1	25.3	2	1	1	18972.495		
39	143	29	1	29.735	2	0	0	18157.876		
40	145	29	0	38.83	3	0	1	5138.2567		
41	146	46	1	30.495	3	1	0	40720.55105		
42	149	19	1	28.4	1	0	3	1842.519		
43	150	35	1	24.13	1	0	0	5125.2157		
44	151	48	1	29.7	0	0	1	7789.635		
45	157	18	1	25.175	0	1	2	15518.18025		
46	161	18	0	36.85	0	1	1	36149.4835		
47	163	32	0	29.8	2	0	3	5152.134		
48	166	20	0	37	5	0	3	4830.63		
49	167	32	0	33.155	3	0	0	6128.79745		
50	168	19	0	31.825	1	0	0	2719.27975		
51	174	24	0	33.345	0	0	0	2855.43755		
52	177	54	1	29.2	1	0	3	10436.096		
53	182	22	1	19.95	3	0	2	4005.4225		
54	188	41	0	32.2	1	0	3	6775.961		
55	189	29	0	32.11	2	0	0	4922.9159		

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)
User Header	User Row ID	age	sex	bmi	children	smoker	region
1	7	-0.150636572 1034263	-1.013537686 61993	-0.474415615 30571447	1.5720236018 2914	-0.490711973 46850026	-1.34280138 2324968
2	11	1.6429962361 489434	-1.013537686 61993	-0.710752056 1183058	-0.894088423 5403234	2.0358255719 795726	-0.43859441 58298184
3	17	-1.155070944 7247534	0.9856604221 549418	-1.109264192 5229858	-0.894088423 5403234	-0.490711973 46850026	0.465612552 6653307
4	24	-0.150636572 1034263	0.9856604221 549418	-0.427963280 3873775	0.7499862600 39319	-0.490711973 46850026	-1.34280138 2324968
5	25	1.4277602991 58659	-1.013537686 61993	-0.477675428 2824398	1.5720236018 2914	-0.490711973 46850026	-0.43859441 58298184
6	29	-0.581108446 083995	0.9856604221 549418	0.9207843387 327561	0.7499862600 39319	2.0358255719 795726	1.369819519 160479
7	30	-1.226816257 0548482	0.9856604221 549418	0.8066908845 473678	-0.894088423 5403234	2.0358255719 795726	1.369819519 160479
8	33	1.7147415484 790383	0.9856604221 549418	-0.381510945 4690406	-0.894088423 5403234	-0.490711973 46850026	-1.34280138 2324968
9	37	-0.939835007 7344689	0.9856604221 549418	-1.605570718 2294277	-0.894088423 5403234	-0.490711973 46850026	1.369819519 160479
10	38	-0.294127196 7636159	0.9856604221 549418	0.9810908788 021767	-0.072051081 75050219	2.0358255719 795726	0.465612552 6653307
11	41	-0.581108446 083995	-1.013537686 61993	0.9745712528 487261	0.7499862600 39319	-0.490711973 46850026	-0.43859441 58298184
12	42	0.1363446772 1695287	0.9856604221 549418	-1.445839882 369883	-0.072051081 75050219	-0.490711973 46850026	-0.43859441 58298184
13	44	-0.078891259 7733315	0.9856604221 549418	1.0430273253 599587	-0.072051081 75050219	-0.490711973 46850026	0.465612552 6653307
14	47	-0.796344383 0742793	-1.013537686 61993	0.6714086460 132641	-0.894088423 5403234	-0.490711973 46850026	-1.34280138 2324968
15	49	-0.222381884 4335310	0.9856604221 549418	0.7414946250 139601	-0.072051081 75050219	2.0358255719 795726	-0.43859441 58298184

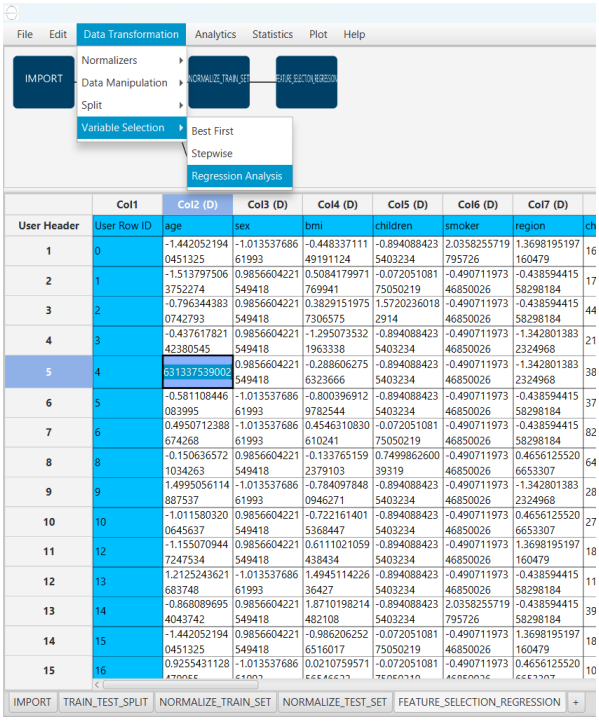
Step 6: Feature selection

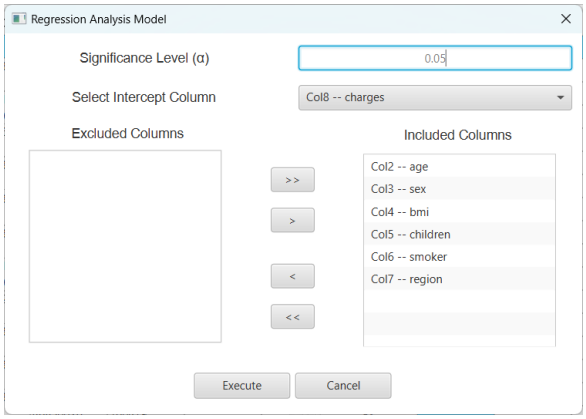
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_REGRESSION".

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".

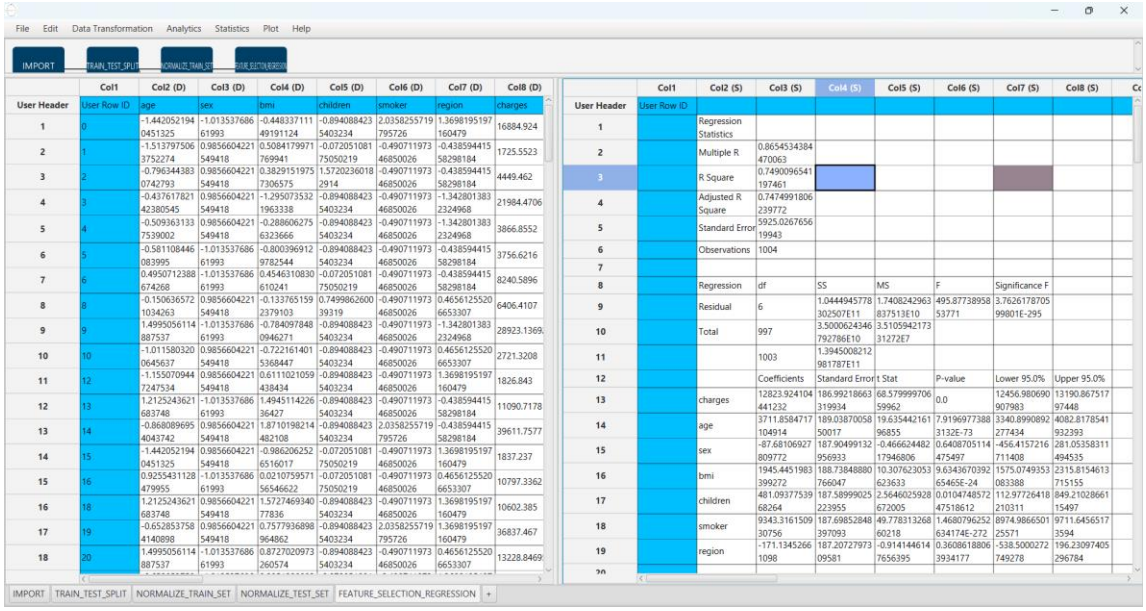


Choose the most important features using the Regression Analysis by browsing: "Data Transformation" → "Variable Selection" → "Regression Analysis". Then choose the "charges" column as the intercept column, the Significance level (α) as 0.05 and include all columns.





The results will appear on the output spreadsheet.



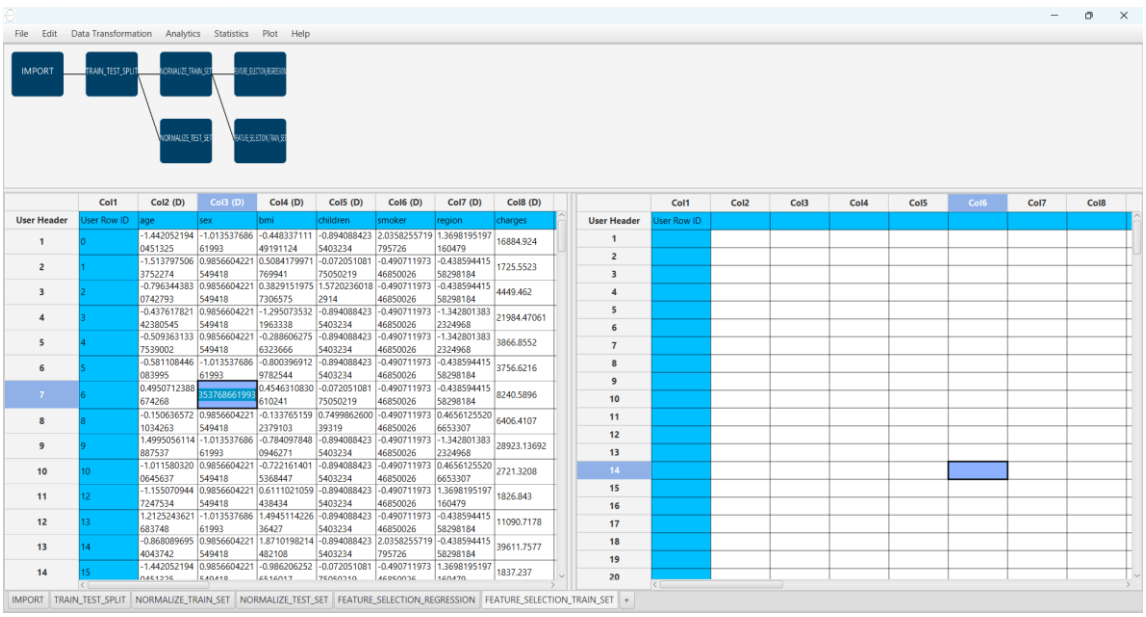
The significant features according to the p-value are the following:

- charges (p-value = 0.0)
- age (p-value = 7.91969773883132E-73)
- bmi (p-value = 9.634367039265465E-24)
- children (p-value = 0.010474857247518612)
- smoker (p-value = 1.4680796252634174E-272)

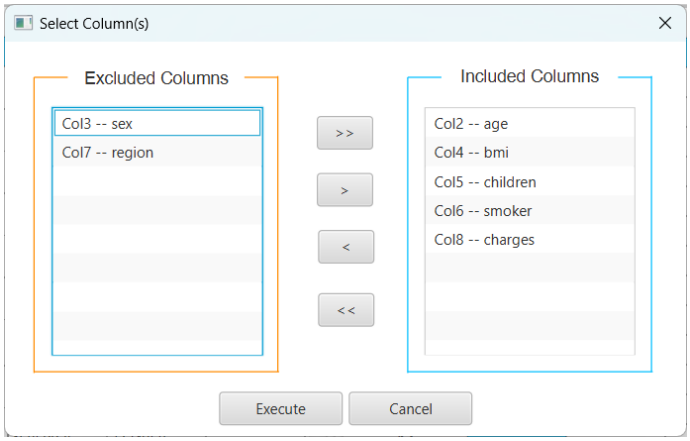
Step 7: Feature selection: train set

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

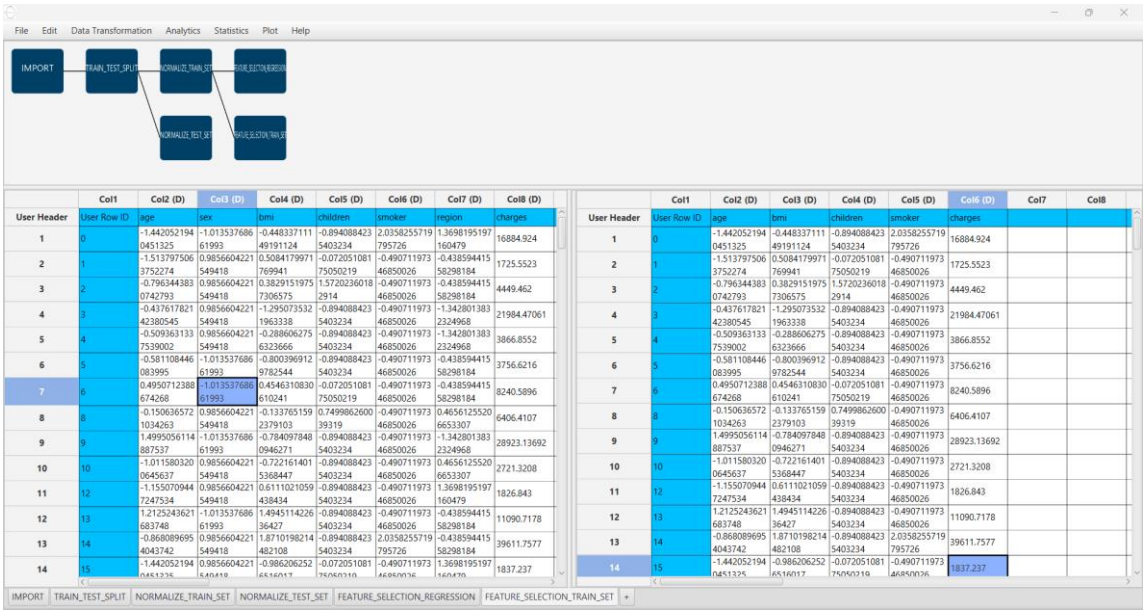
Import data into the input spreadsheet of the "FEATURE_SELECTION_TRAIN_SET" tab from the output of the "NORMALIZE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



Manipulate the data by choosing the columns that correspond to the significant features (from the previous step) by browsing: "Data Transformation" → "Data Manipulation" → "Select Column(s)".



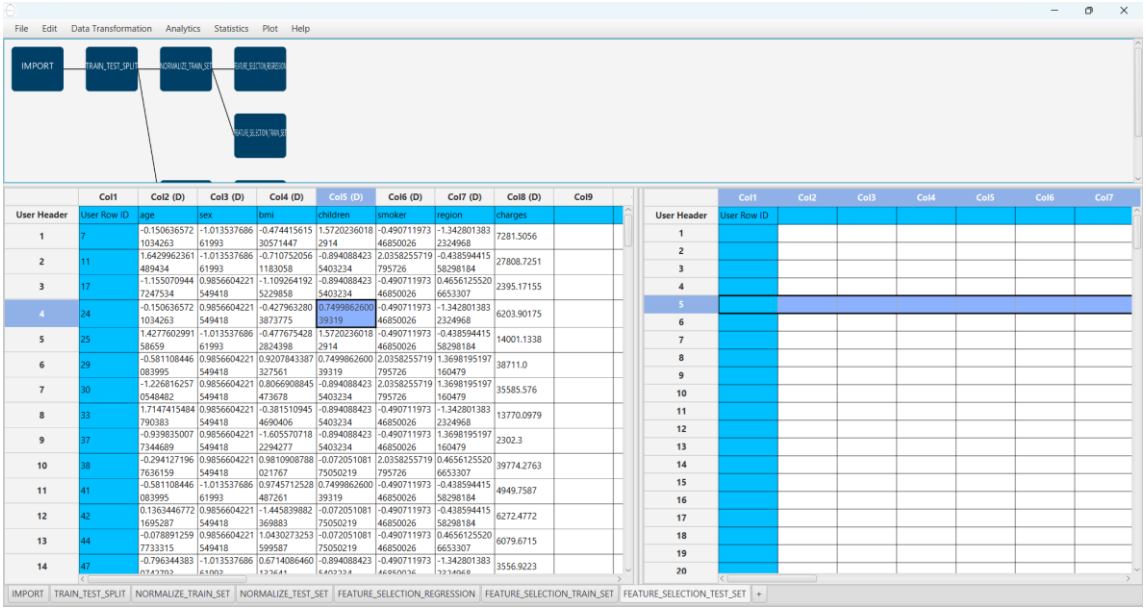
The results will appear on the output spreadsheet.



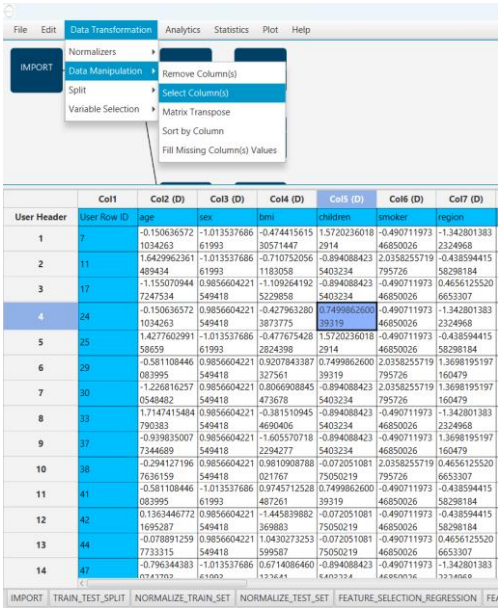
Step 8: Feature selection: test set

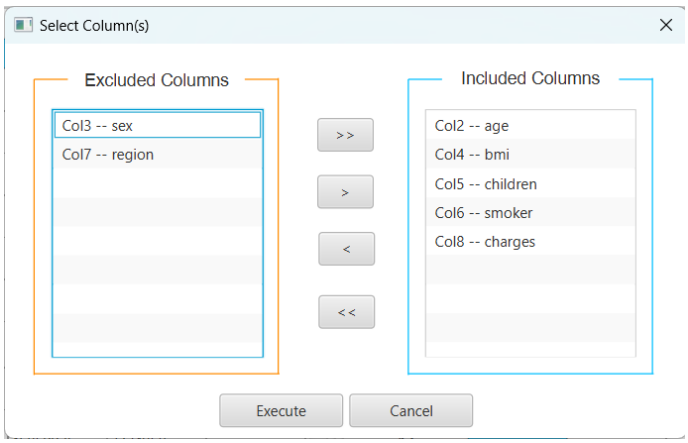
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_TEST_SET".

Import data into the input spreadsheet of the "FEATURE_SELECTION_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".

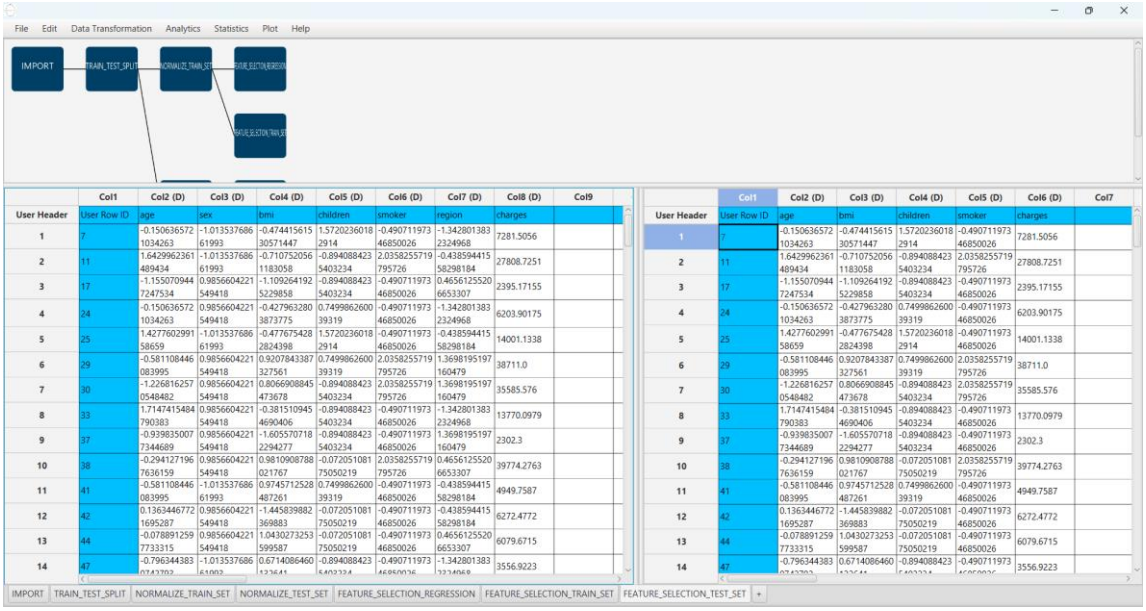


Manipulate the data by choosing the columns that correspond to the significant features (from the step 6) by browsing: "Data Transformation" → "Data Manipulation" → "Select Column(s)".





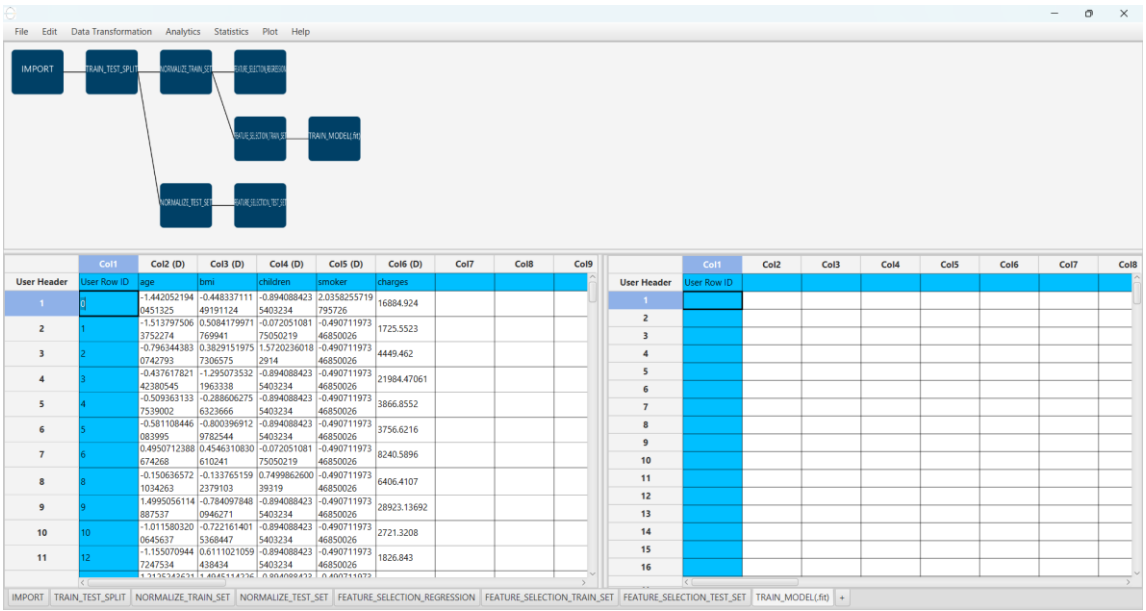
The results will appear on the output spreadsheet.



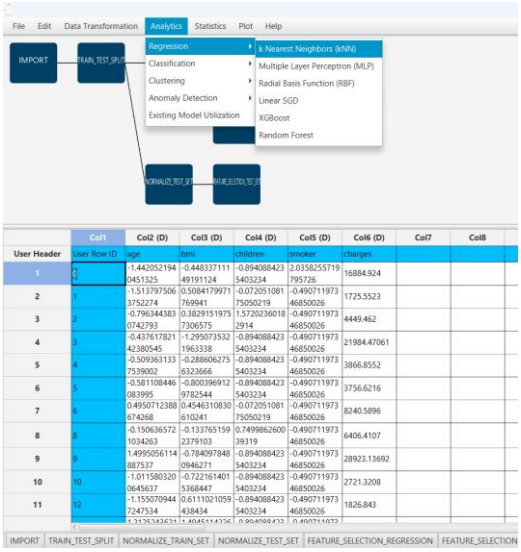
Step 9: Train the model

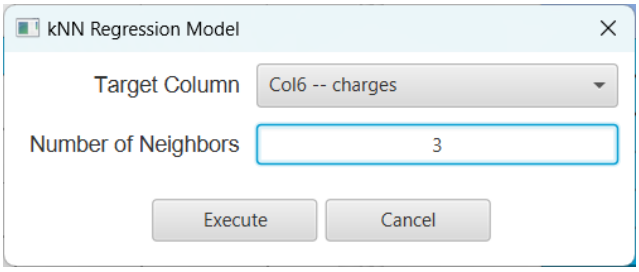
Create a new tab by pressing the "+" button on the bottom of the page with the name "TRAIN_MODEL(.fit)".

Import data into the input spreadsheet of the "TRAIN_MODEL(.fit)" tab from the output of the "FEATURE_SELECTION_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".

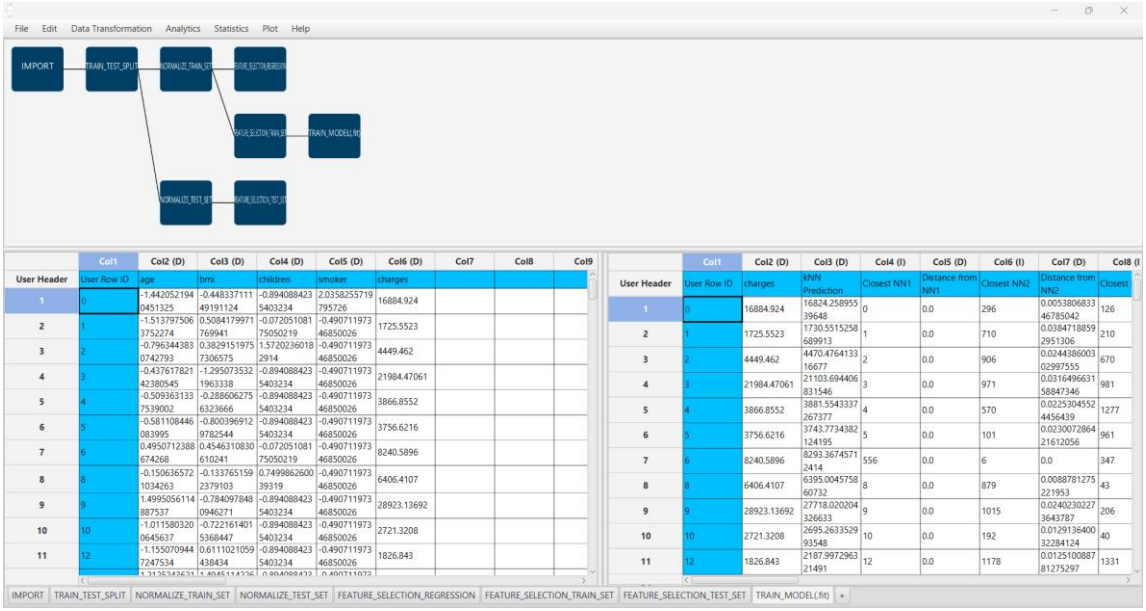


Use the k Nearest Neighbors (kNN) method to train and fit the model by browsing: "Analytics" → "Regression" → "k Nearest Neighbors (kNN)" and set the "Target Column" as the column corresponding to "charges" and the "Number of Neighbors" to 3.





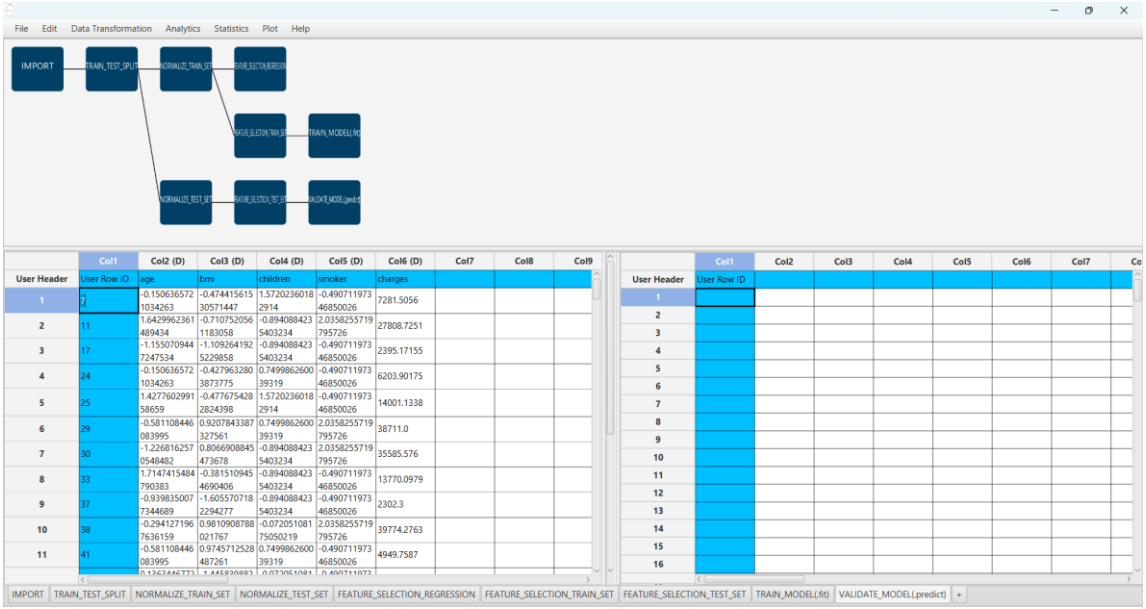
The predictions will appear on the output spreadsheet.



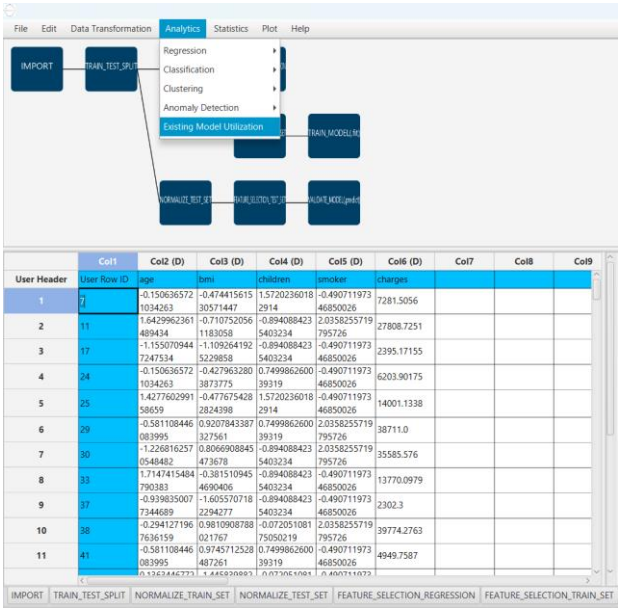
Step 10: Validate the model

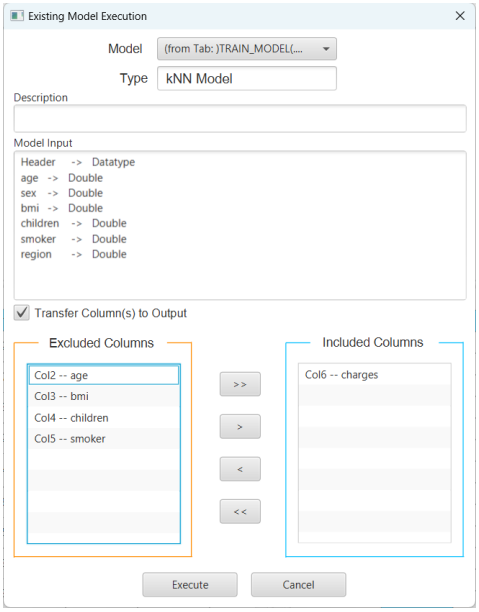
Create a new tab by pressing the "+" button on the bottom of the page with the name "VALIDATE_MODEL(.predict)".

Import data into the input spreadsheet of the "VALIDATE_MODEL(.predict)" tab from the output of the "FEATURE_SELECTION_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".

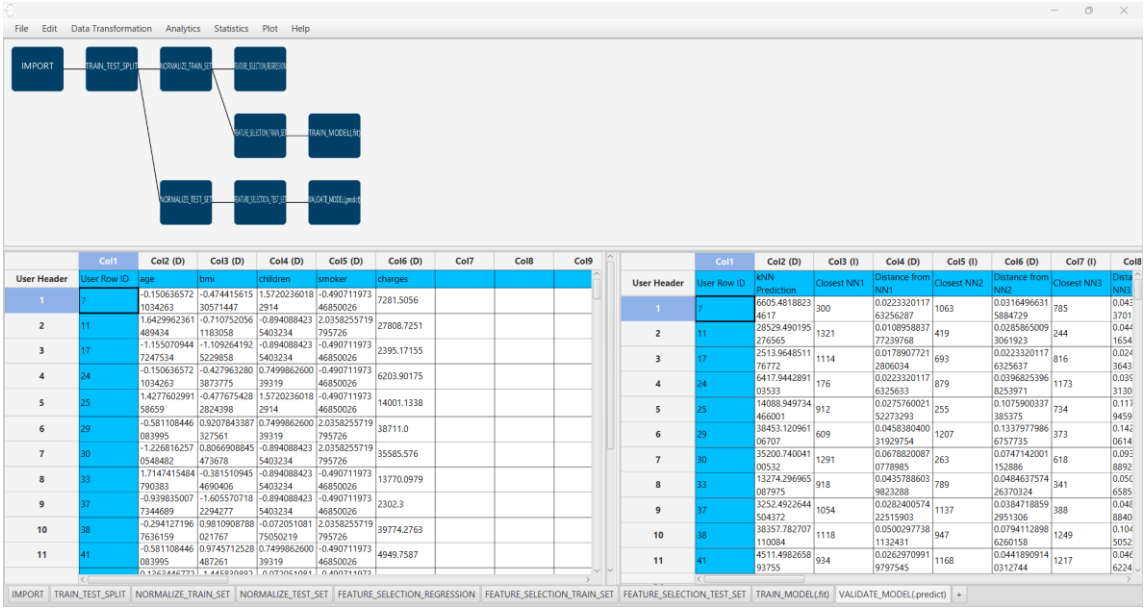


To validate the model browse: "Analytics" → "Existing Model Utilization". Then choose Model "(from Tab:) TRAIN_MODEL (.fit)" and transfer the "charges" column in the output.





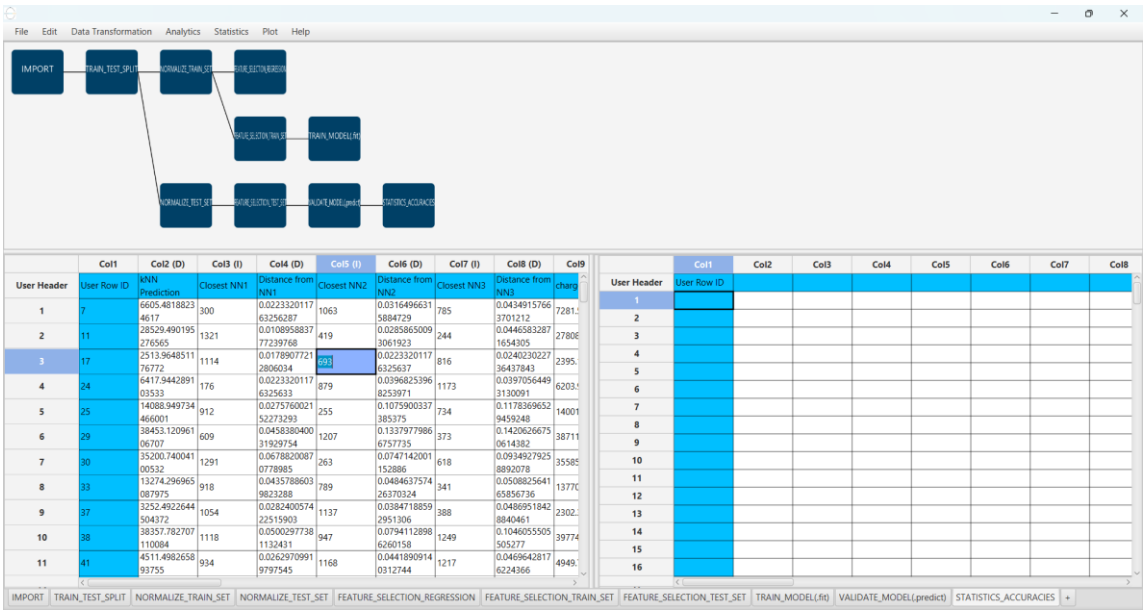
The predictions will appear on the output spreadsheet.



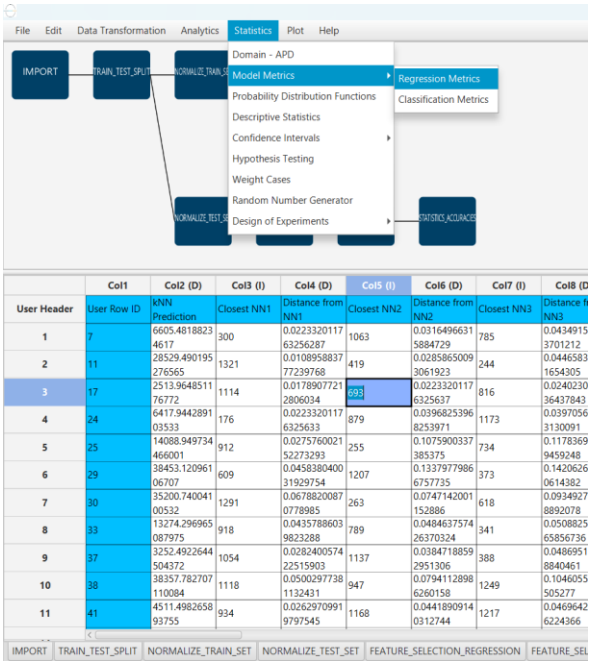
Step 11: Statistics calculation

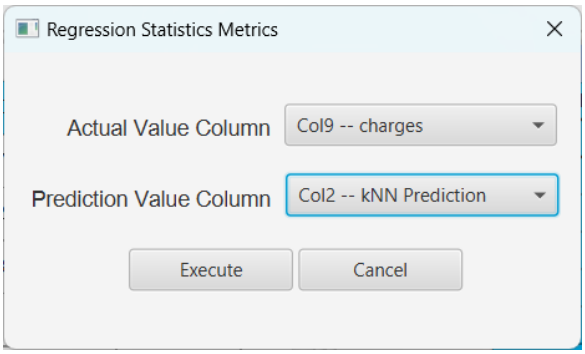
Create a new tab by pressing the "+" button on the bottom of the page with the name "STATISTICS_ACCURACIES".

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".

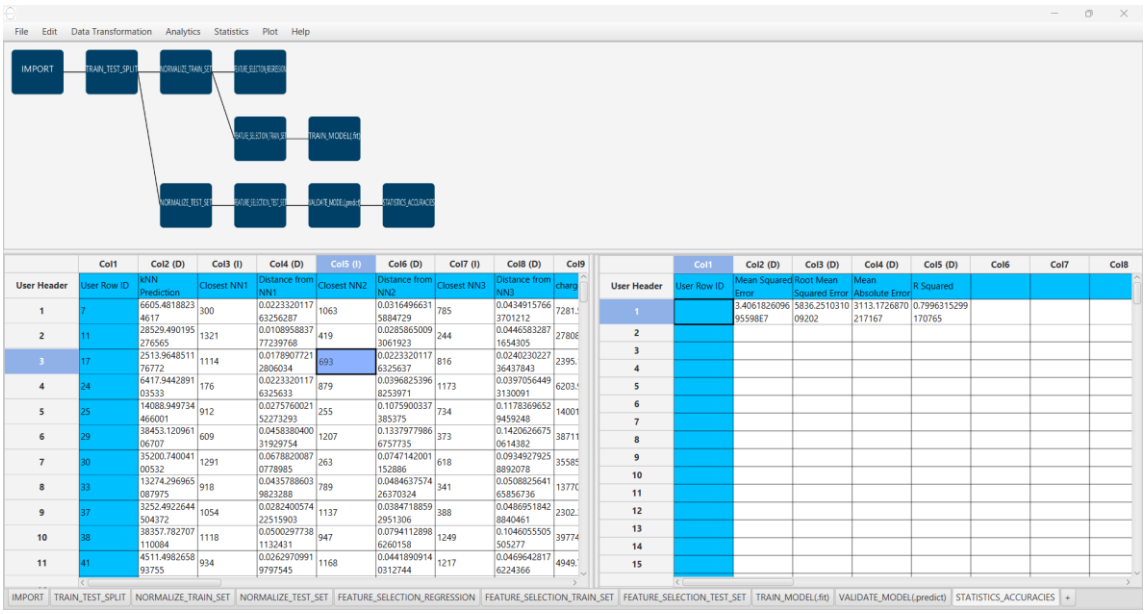


Calculate the statistical metrics for the regression by browsing: "Statistics" → "Model Metrics" → "Regression Metrics".





The results will appear on the output spreadsheet.

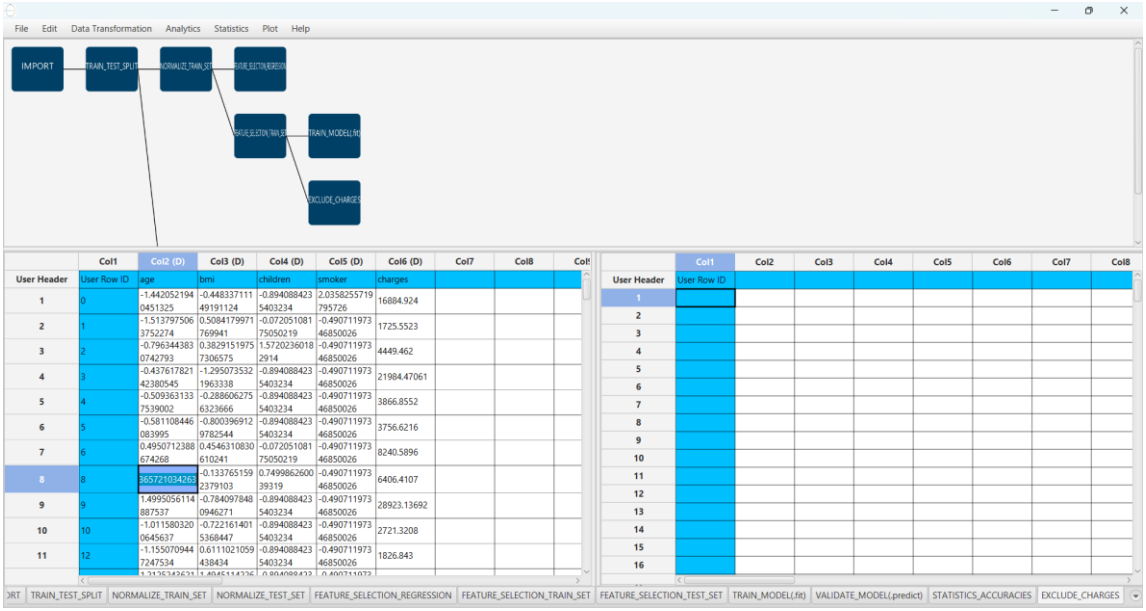


Step 12: Reliability check of each record of the test set

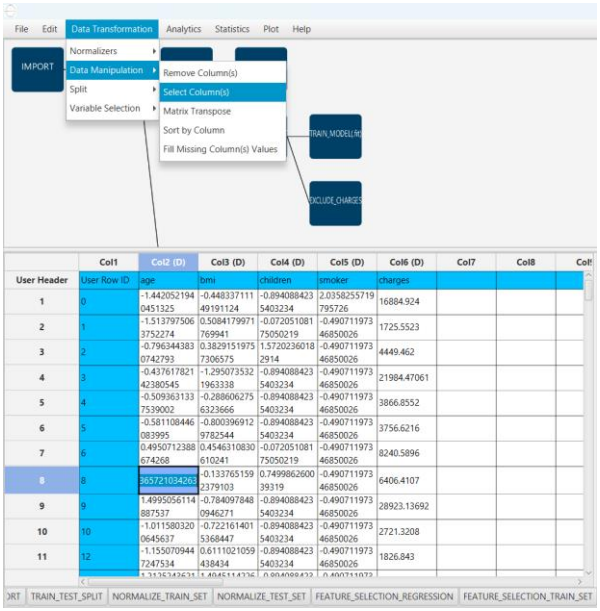
Step 12.a: Create the domain

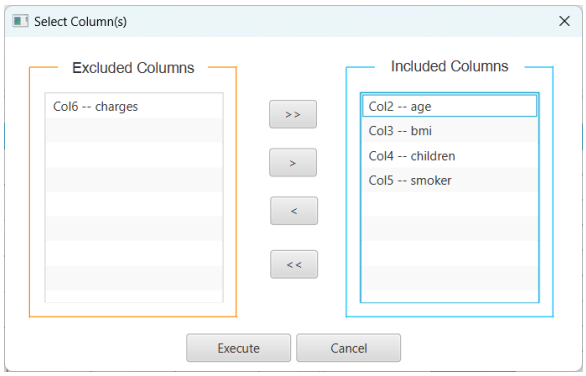
Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_CHARGES".

Import data into the input spreadsheet of the "EXCLUDE_CHARGES" tab from the output of the "FEATURE_SELECTION_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 "charges" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except the "charges".

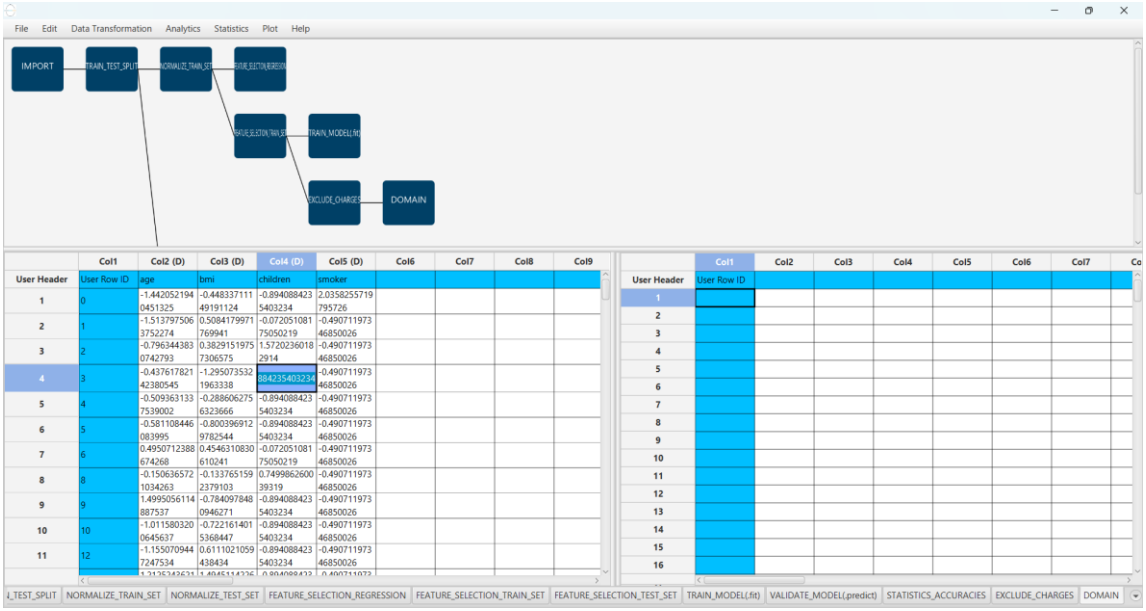




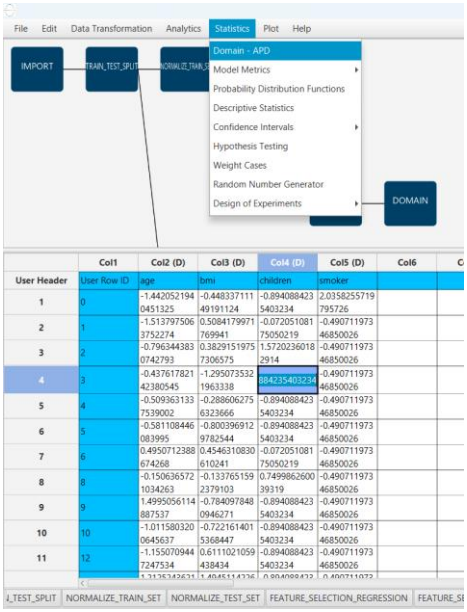
The results will appear on the output spreadsheet.

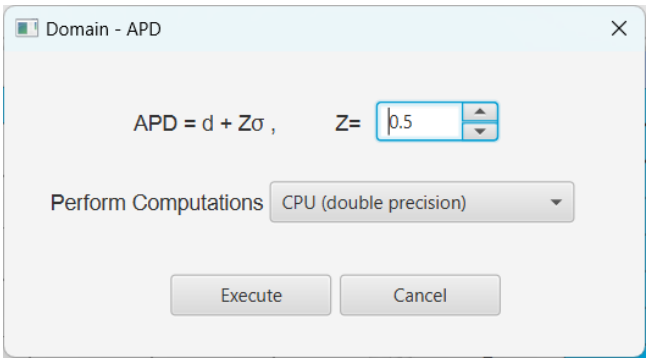
Create a new tab by pressing the "+" button on the bottom of the page with the name "DOMAIN".

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

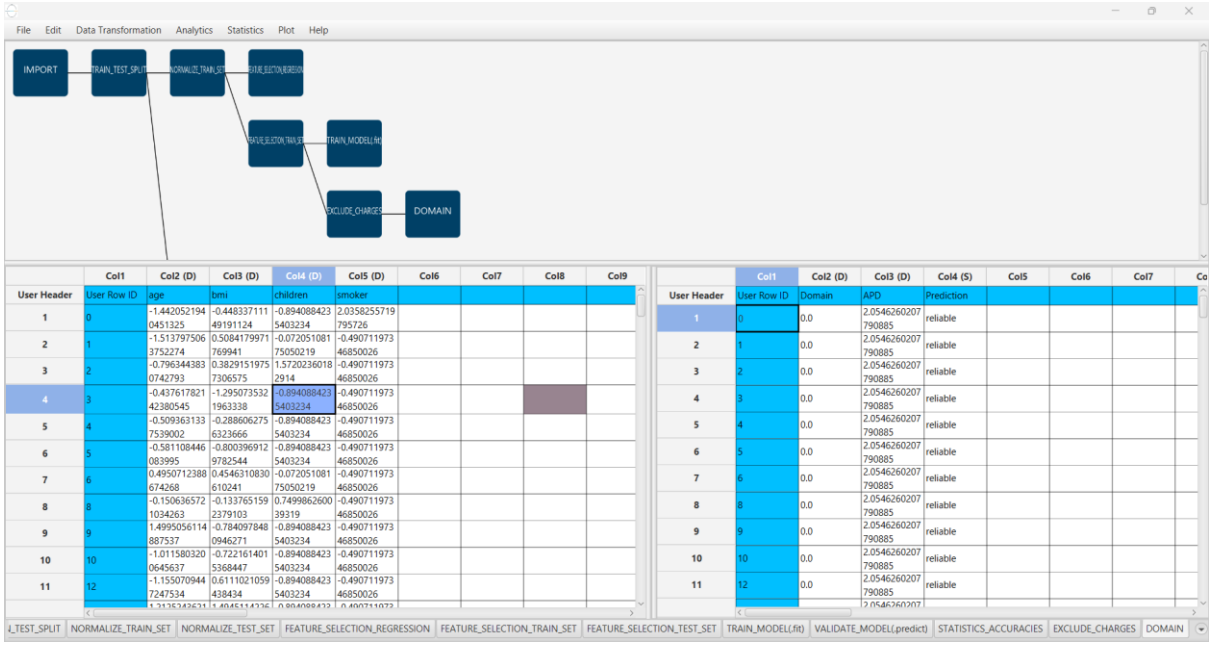


Create the domain by browsing: "Statistics" → "Domain APD".





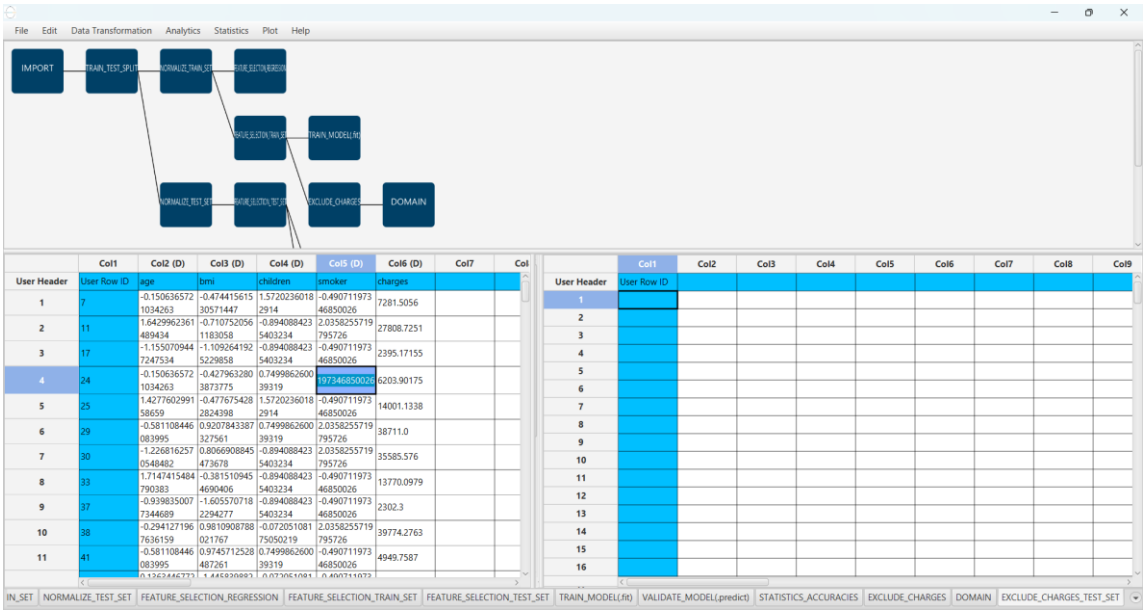
The results will appear on the output spreadsheet.



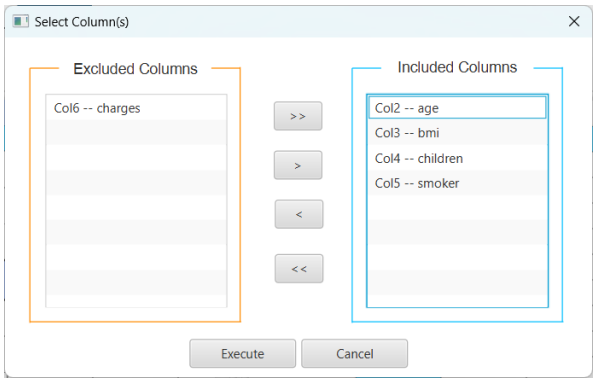
Step 12.b: Check the test set reliability

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

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



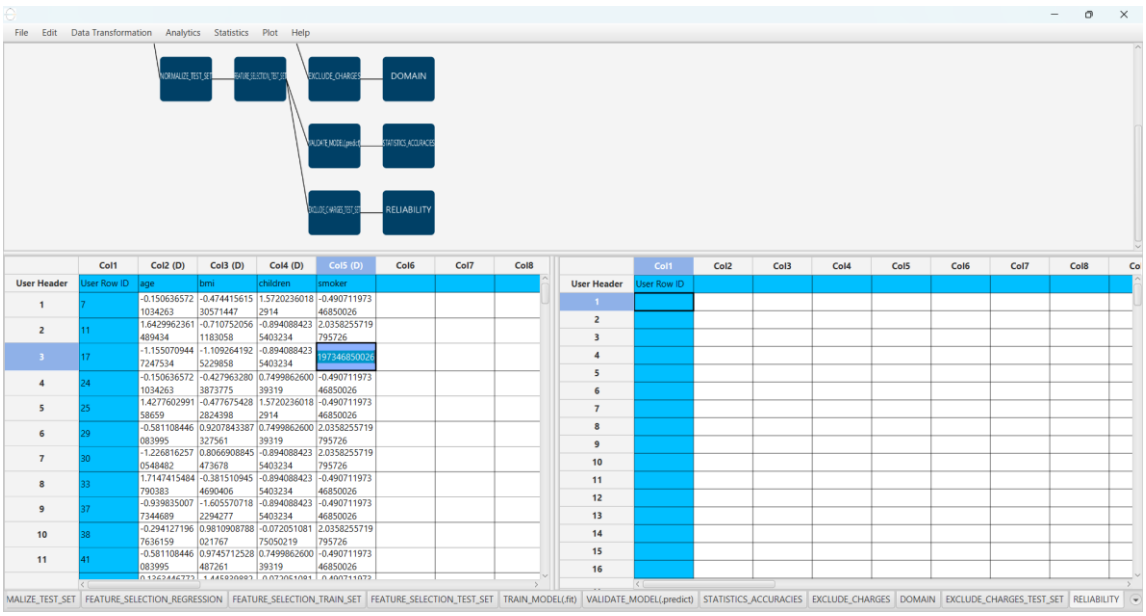
Filter the data to exclude the column that corresponds to the "charges" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except "charges".



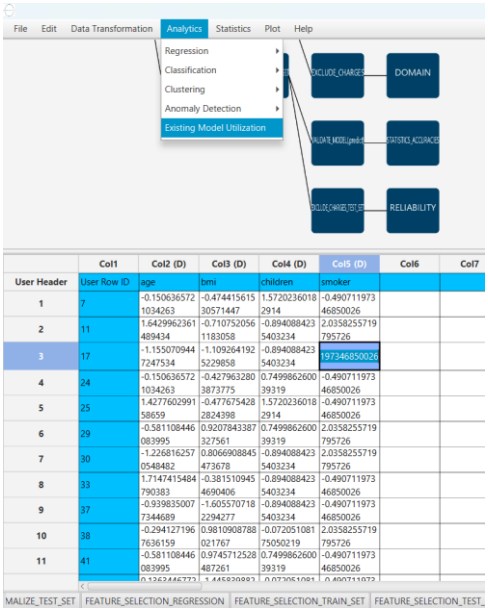
The results will appear on the output spreadsheet.

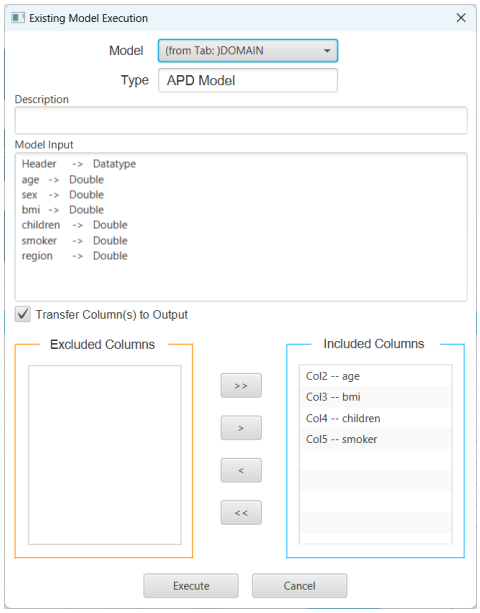
Create a new tab by pressing the "+" button on the bottom of the page with the name "RELIABILITY".

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

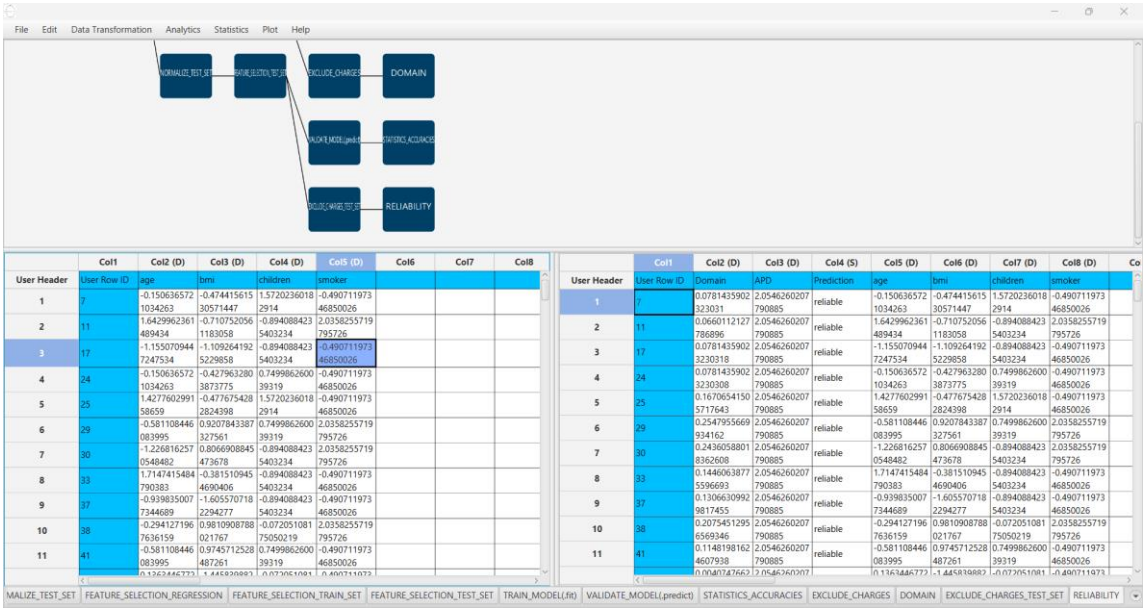


Check the Reliability by browsing: "Analytics" → "Existing Model Utilization". Then select as Model "(from Tab:) DOMAIN".





The results will appear on the output spreadsheet.



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: