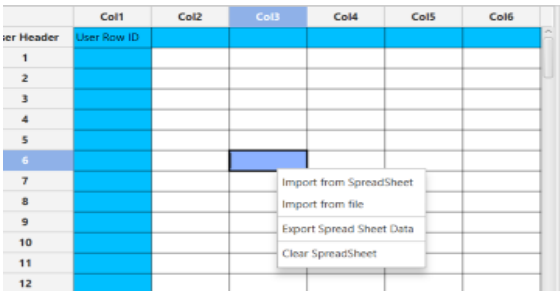
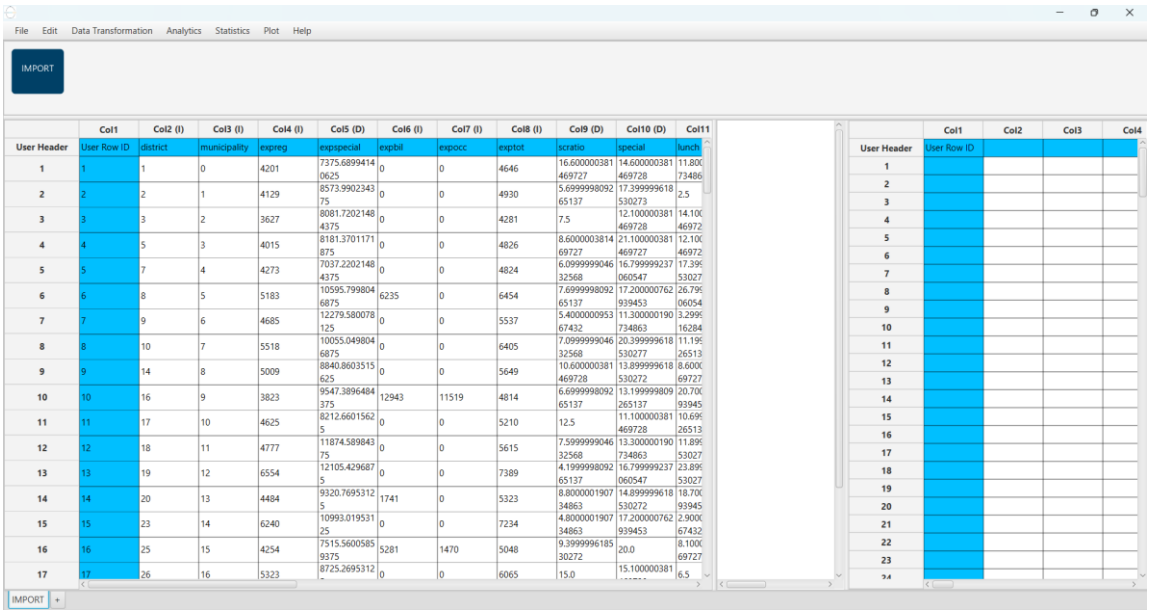
MA score Data (Regression)

The goal of this study is to train a model in order to predict the test score for school districts in Massachusetts. The dataset used in this case study is found in [R: Massachusetts Test Score Data (vincentarelbundock.github.io)](#) and has 16 features and 220 labelled samples. The dataset contains data on test performance, school characteristics and student demographic backgrounds for school districts in Massachusetts. The data analysed here are the overall total score, which is the sum of the scores on the English, Math, and Science portions of the test.

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 MA score data.

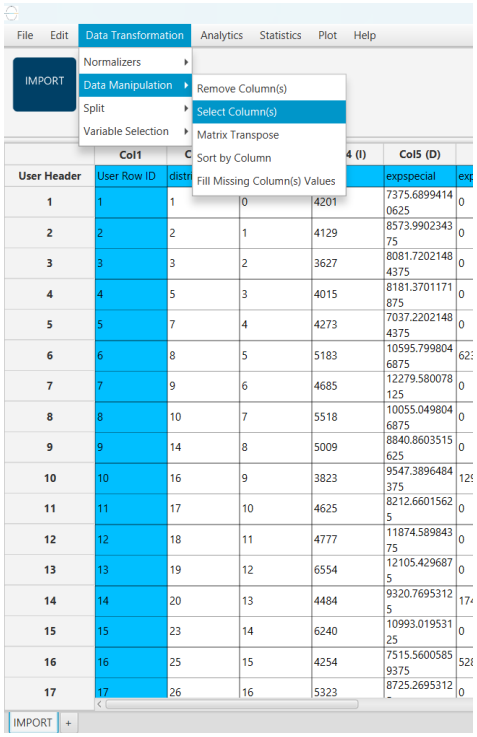


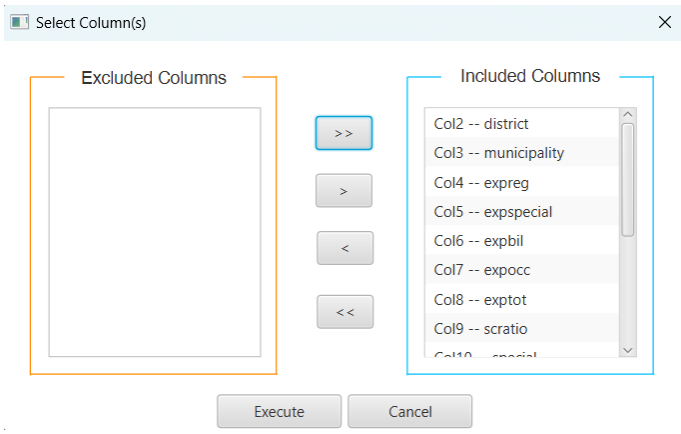


Step 2: Manipulate data

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.



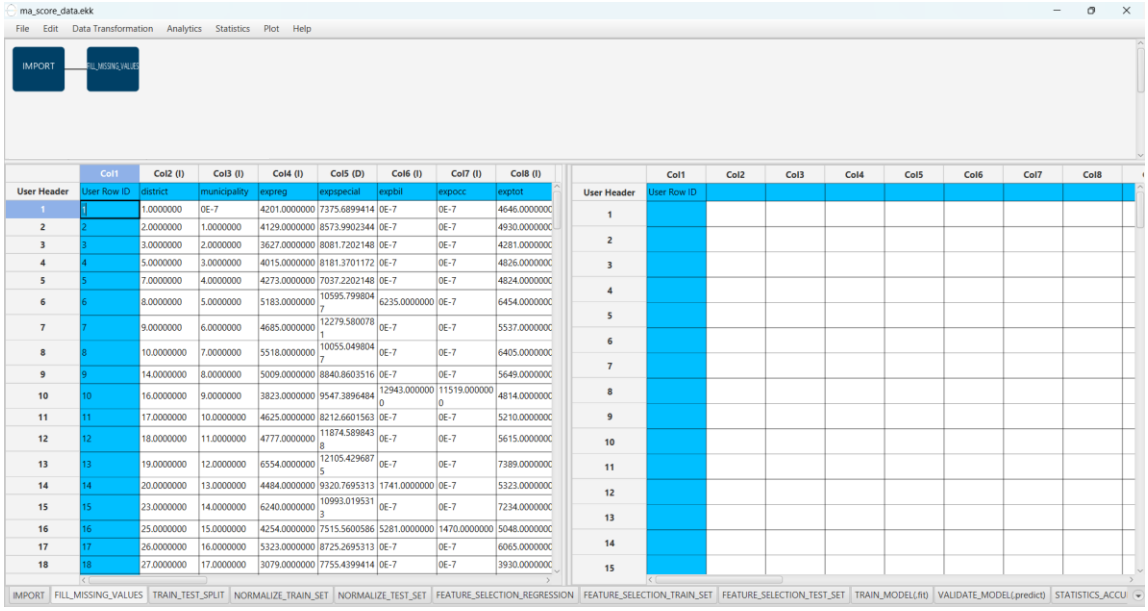


The data will appear in the output spreadsheet.

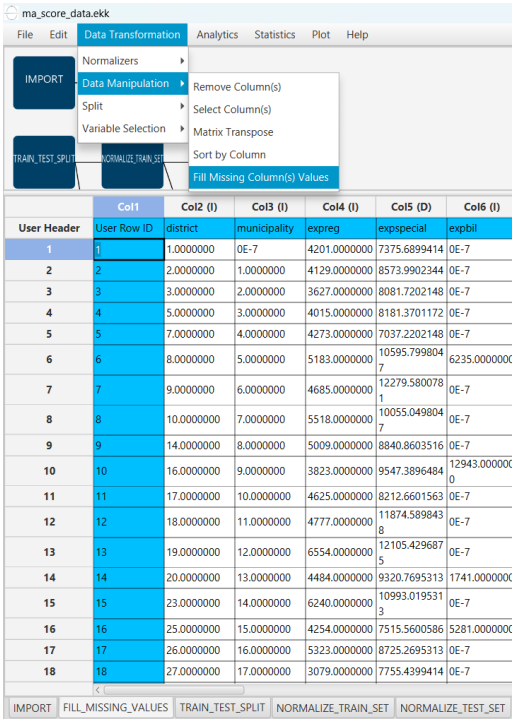
Step 3: Fill missing values

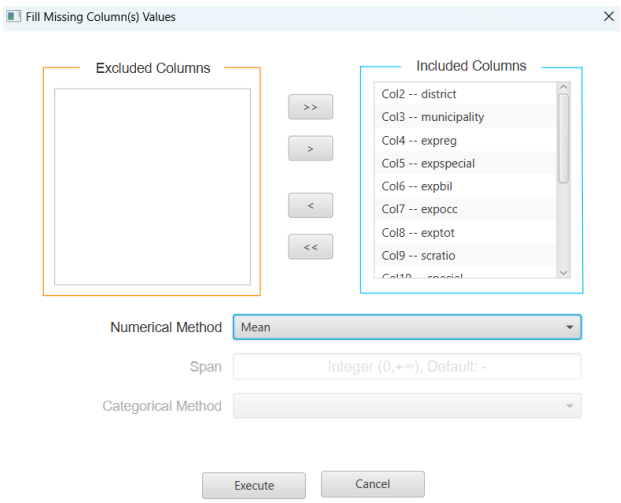
Create a new tab by pressing the "+" button on the bottom of the page with the name "FILL_MISSING_VALUES" which we will use for filling the missing values of the dataset

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

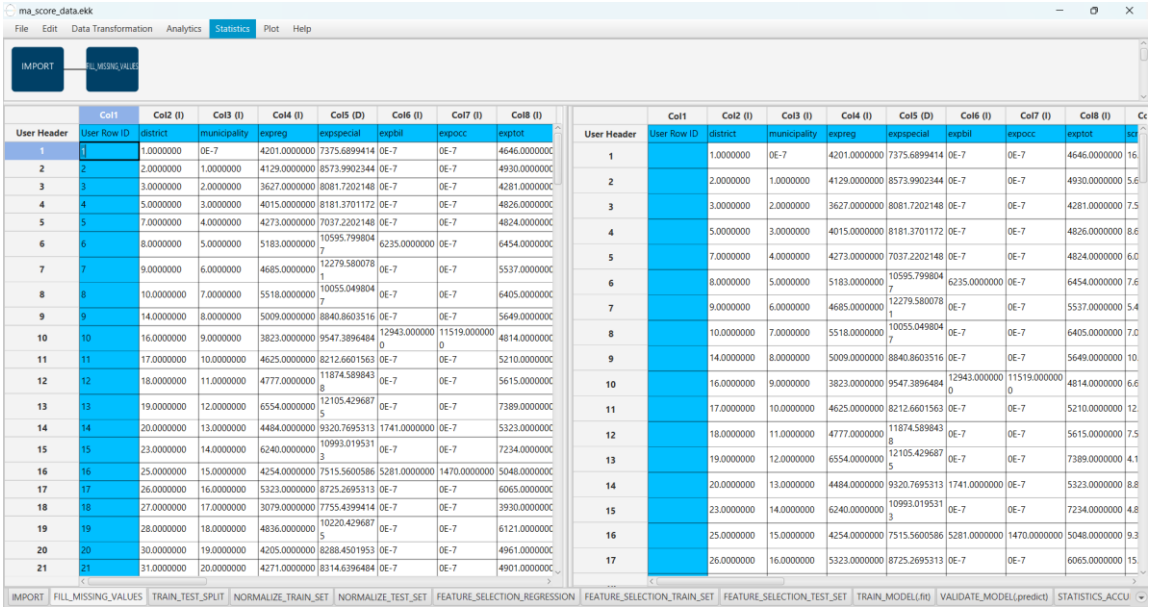


Fill the missing values in the dataset by browsing: "Data Transformation" → "Data Manipulation" → "Fill Missing Column Values". Then include all columns and select the "Mean" as the "Numerical Method" to fill the missing values.





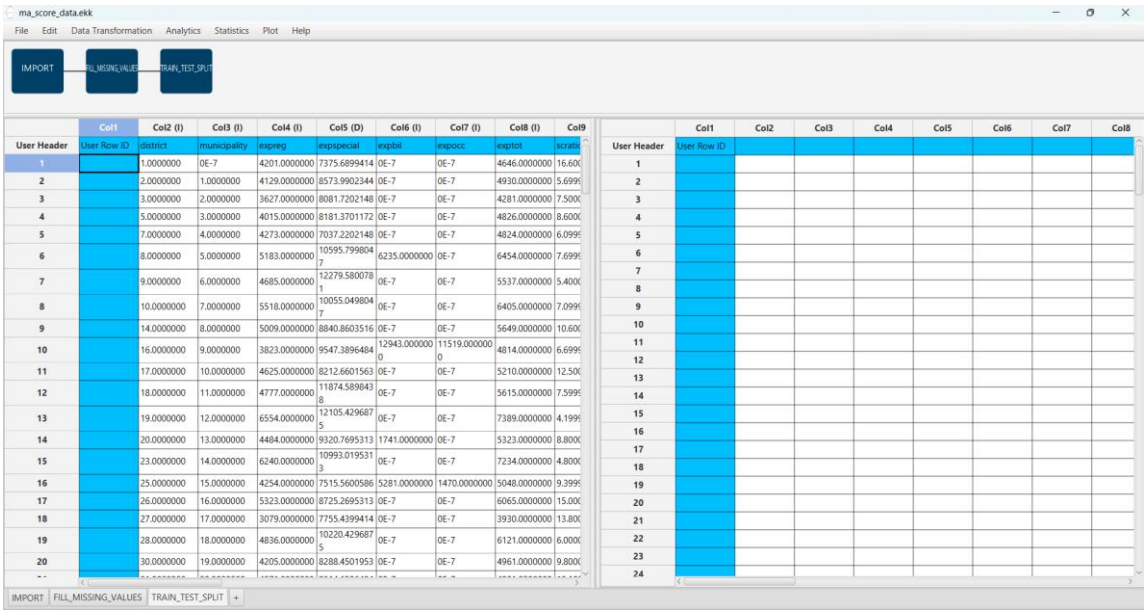
The data will appear in the output spreadsheet.



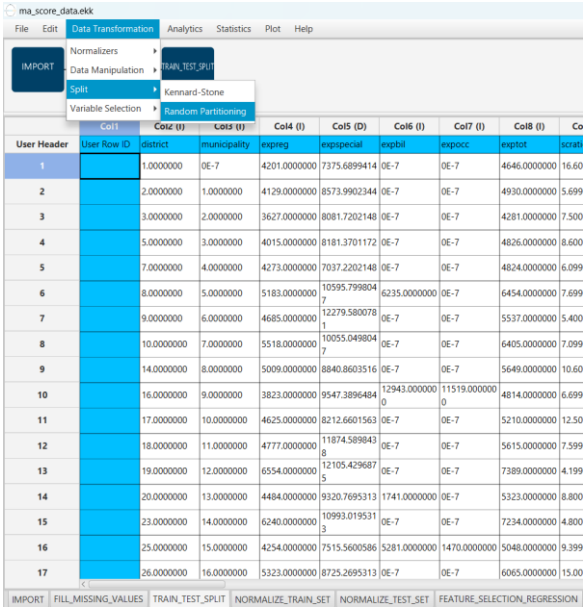
Step 4: Split data

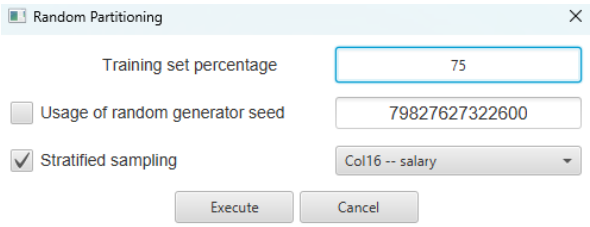
Create a new tab by pressing the "+" button on the bottom of the page with the name "TRAIN_TEST_SPLIT" which we will use for splitting to create the train and test set.

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

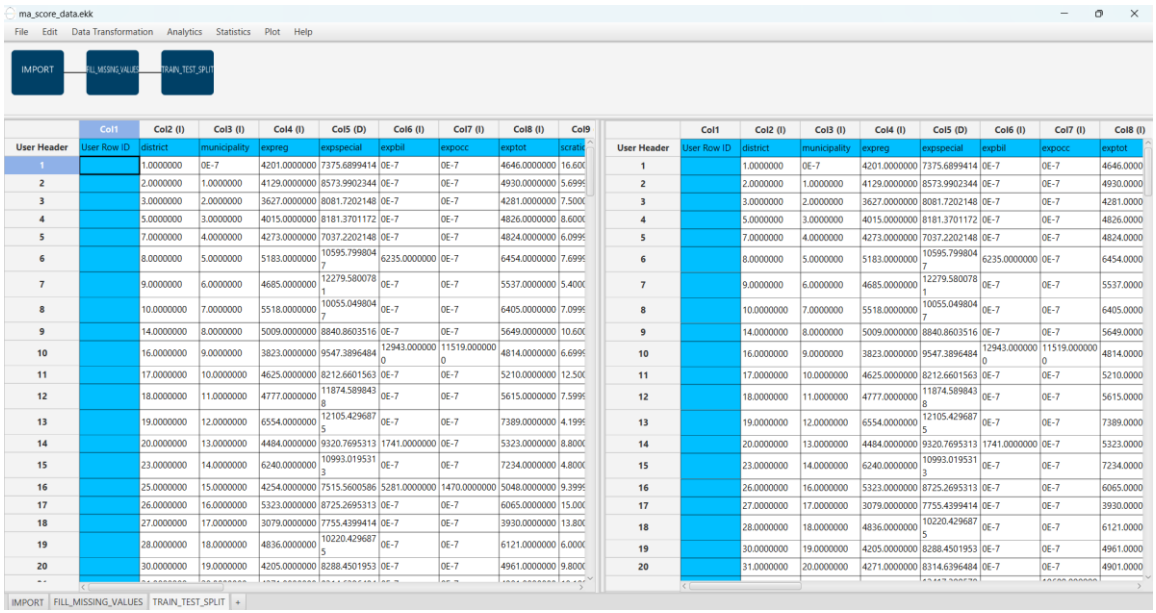


Split the dataset by browsing: "Data Transformation" → "Split" → "Random Partitioning". Then choose the "Training set percentage" and the column for the sampling as shown below:





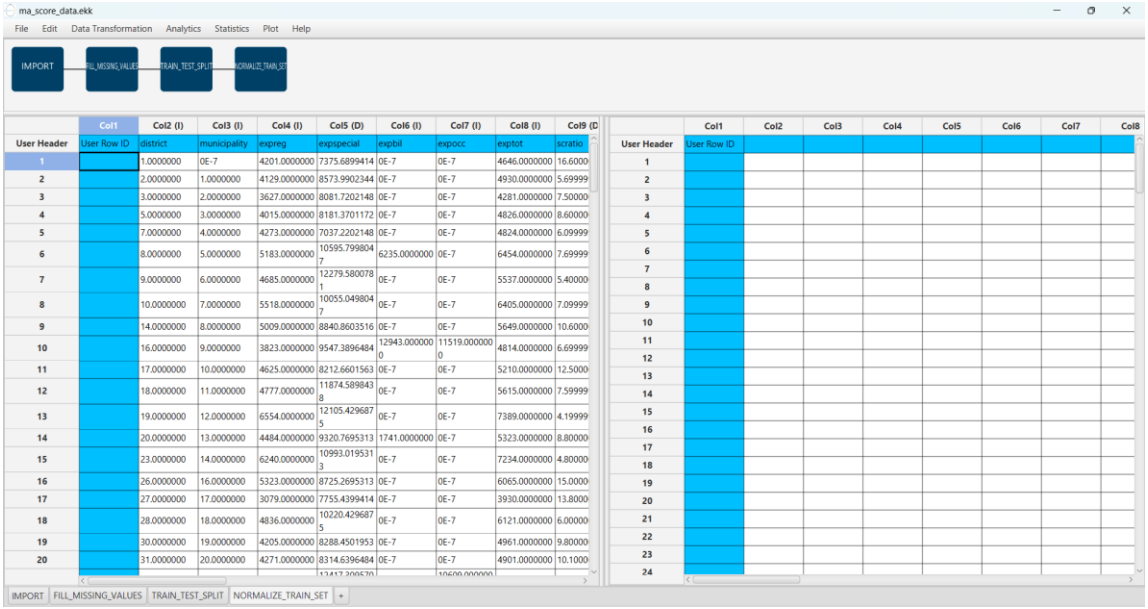
The results will appear on the output spreadsheet.



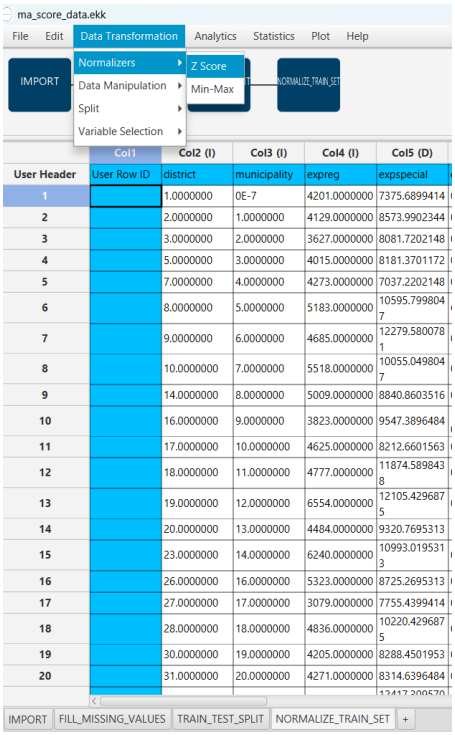
Step 5: Normalize the training set

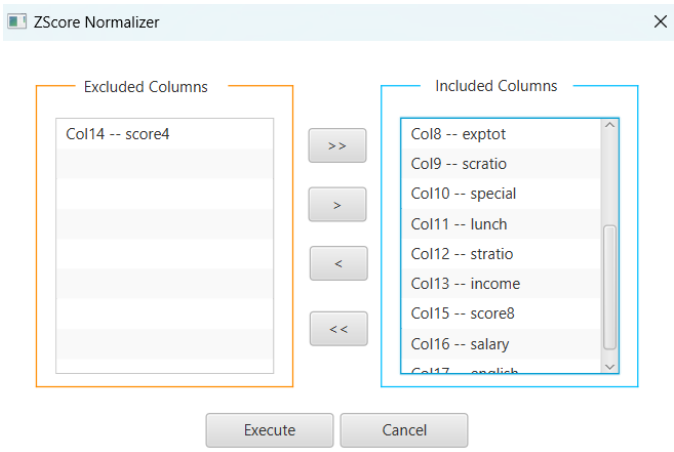
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALIZE_TRAIN_SET".

Import 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"

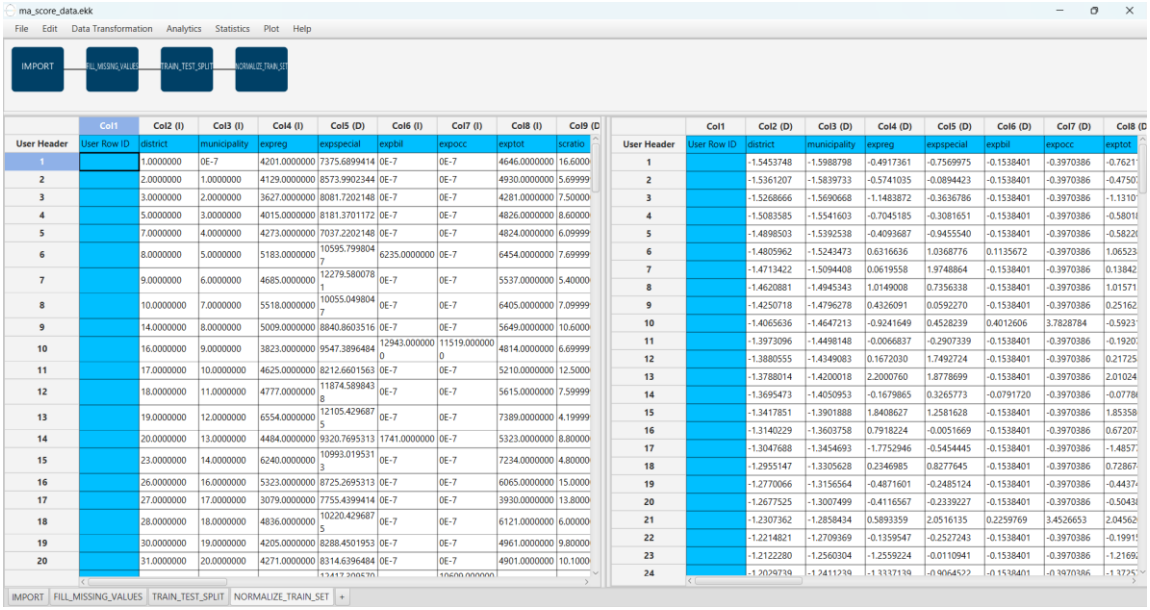


Normalize the data using Z-score by browsing: "Data Transformation" → "Normalizers" → "Z-Score". Then select all columns except "score4" and click "Execute".





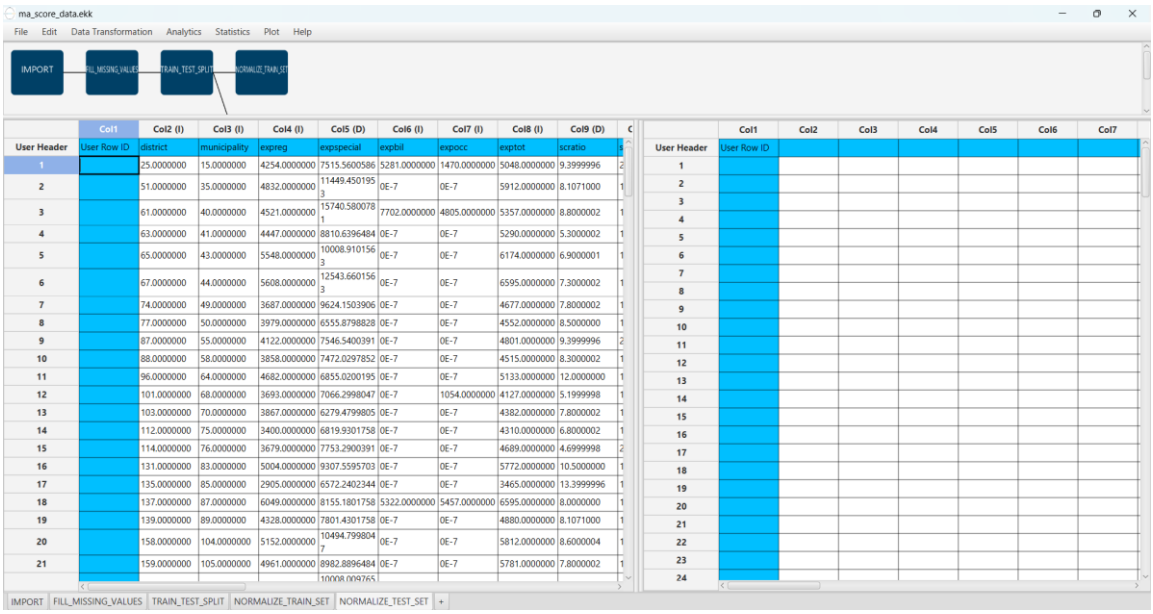
The results will appear on the output spreadsheet.



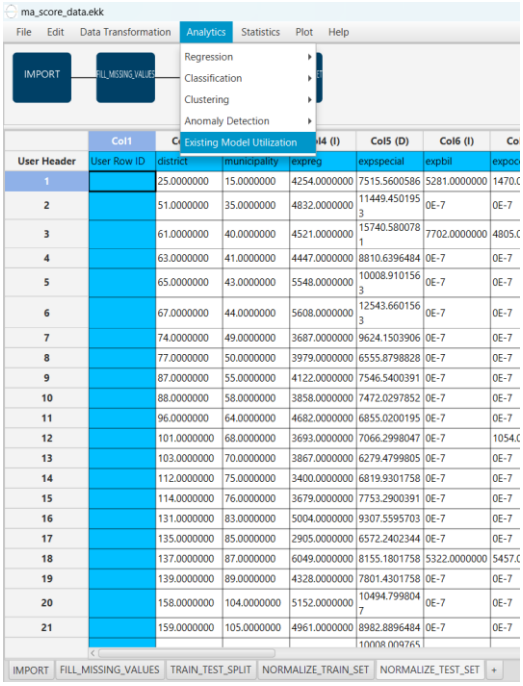
Step 6: Normalize the test set

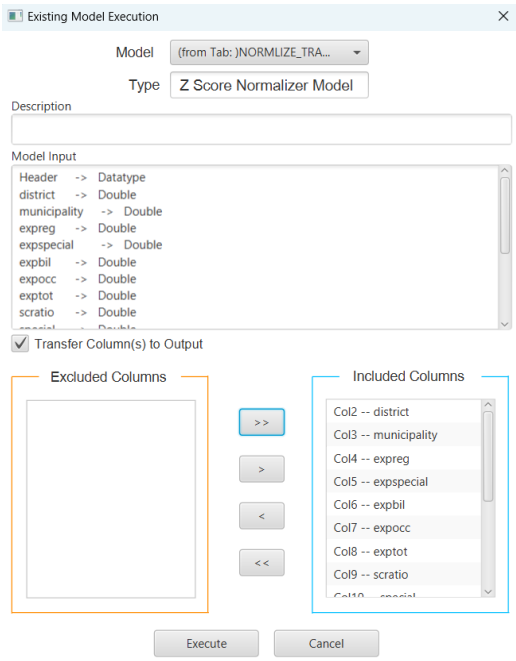
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALIZE_TEST_SET".

Import 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".

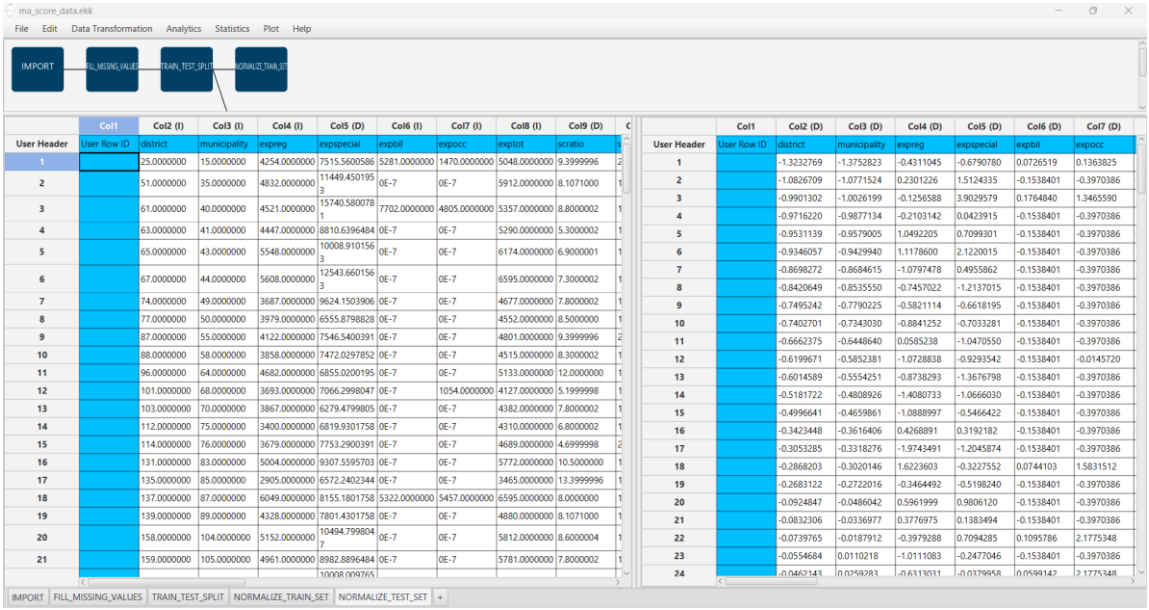


Normalize the test set using the existing normalizer of the training set by browsing: "Analytics" → "Existing Model Utilization" → "Model (from Tab:) NORMALIZE_TRAIN_SET".





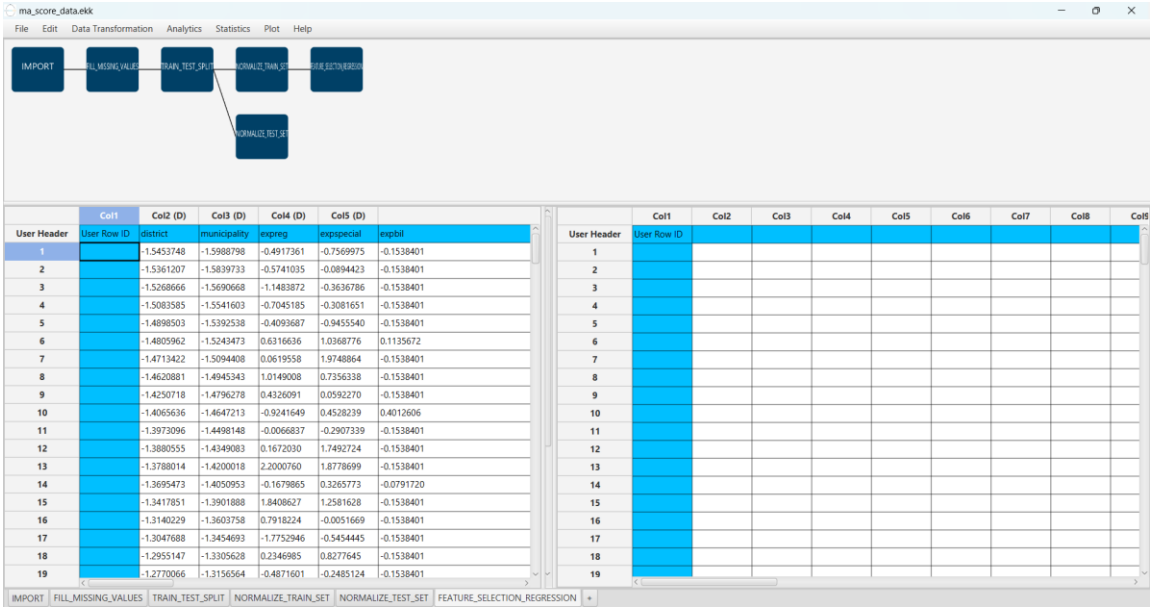
The results will appear on the output spreadsheet.



Step 7: 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".

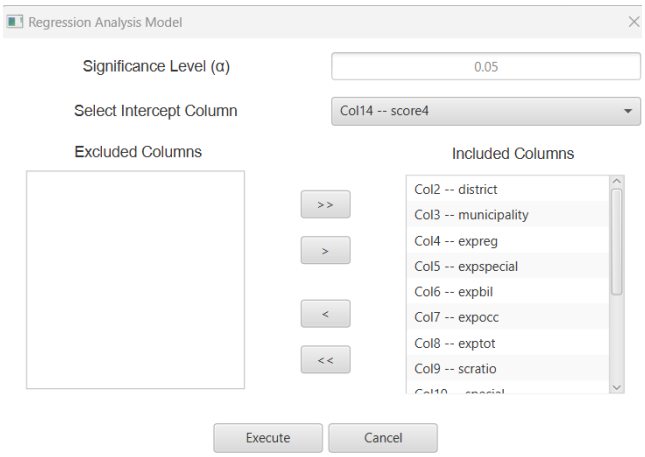


	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	
User Header	User Row ID	district	municipality	expreg	expspecial	expbil
1		-1.5453748	-1.5988798	-0.4917361	-0.7569975	-0.1538401
2		-1.5361207	-1.5839733	-0.5741035	-0.0894423	-0.1538401
3		-1.5268666	-1.5690668	-1.1483872	-0.3636786	-0.1538401
4		-1.5083585	-1.5541603	-0.7045185	-0.3081651	-0.1538401
5		-1.4898503	-1.5392538	-0.4093687	-0.9455540	-0.1538401
6		-1.4805962	-1.5243473	0.6316636	1.0368776	0.1135672
7		-1.4713422	-1.5094408	0.0619558	1.9748864	-0.1538401
8		-1.4620881	-1.4945343	1.0149008	0.7356338	-0.1538401
9		-1.4250718	-1.4796278	0.4326091	0.0592270	-0.1538401
10		-1.4065636	-1.4647213	-0.9241649	0.4528239	0.4012606
11		-1.3973096	-1.4498148	-0.0066837	-0.2907339	-0.1538401
12		-1.3880555	-1.4349083	0.1672030	1.7492724	-0.1538401
13		-1.3788014	-1.4200018	2.2000760	1.8778699	-0.1538401
14		-1.3695473	-1.4050953	-0.1679865	0.3265773	-0.0791720
15		-1.3417851	-1.3901888	1.8408627	1.2581628	-0.1538401
16		-1.3140229	-1.3603758	0.7918224	-0.0051669	-0.1538401
17		-1.3047688	-1.3454693	-1.7752946	-0.5454445	-0.1538401
18		-1.2955147	-1.3305628	0.2346985	0.8277645	-0.1538401
19		-1.2770066	-1.3156564	-0.4871601	-0.2485124	-0.1538401

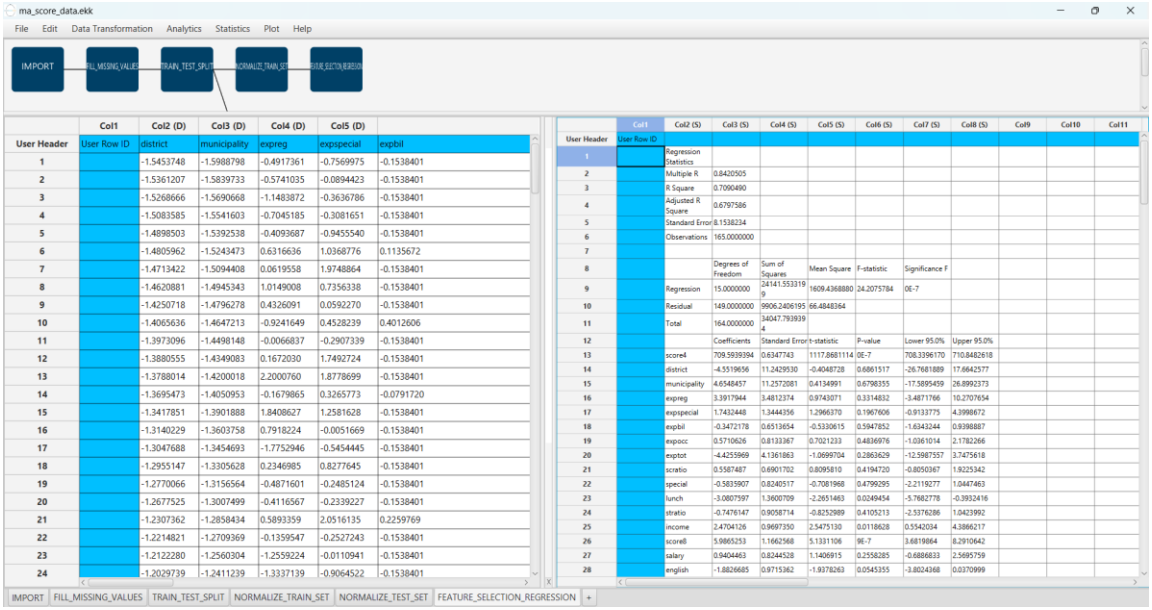
	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9
User Header	User Row ID								
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									

Choose the most important features using the Regression Analysis by browsing: "Data Transformation" → "Variable Selection" → "Regression Analysis". Then choose the "score4" column as the intercept column, the Significance level (α) as 0.05 and include all columns.

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	
User Header	User Row ID	district	municipality	expreg	expspecial	expbil
1		-1.5453748	-1.5988798	-0.4917361	-0.7569975	-0.1538401
2		-1.5361207	-1.5839733	-0.5741035	-0.0894423	-0.1538401
3		-1.5268666	-1.5690668	-1.1483872	-0.3636786	-0.1538401
4		-1.5083585	-1.5541603	-0.7045185	-0.3081651	-0.1538401
5		-1.4898503	-1.5392538	-0.4093687	-0.9455540	-0.1538401
6		-1.4805962	-1.5243473	0.6316636	1.0368776	0.1135672
7		-1.4713422	-1.5094408	0.0619558	1.9748864	-0.1538401
8		-1.4620881	-1.4945343	1.0149008	0.7356338	-0.1538401
9		-1.4250718	-1.4796278	0.4326091	0.0592270	-0.1538401
10		-1.4065636	-1.4647213	-0.9241649	0.4528239	0.4012606
11		-1.3973096	-1.4498148	-0.0066837	-0.2907339	-0.1538401
12		-1.3880555	-1.4349083	0.1672030	1.7492724	-0.1538401
13		-1.3788014	-1.4200018	2.2000760	1.8778699	-0.1538401
14		-1.3695473	-1.4050953	-0.1679865	0.3265773	-0.0791720
15		-1.3417851	-1.3901888	1.8408627	1.2581628	-0.1538401
16		-1.3140229	-1.3603758	0.7918224	-0.0051669	-0.1538401
17		-1.3047688	-1.3454693	-1.7752946	-0.5454445	-0.1538401
18		-1.2955147	-1.3305628	0.2346985	0.8277645	-0.1538401
19		-1.2770066	-1.3156564	-0.4871601	-0.2485124	-0.1538401



The results will appear on the output spreadsheet.



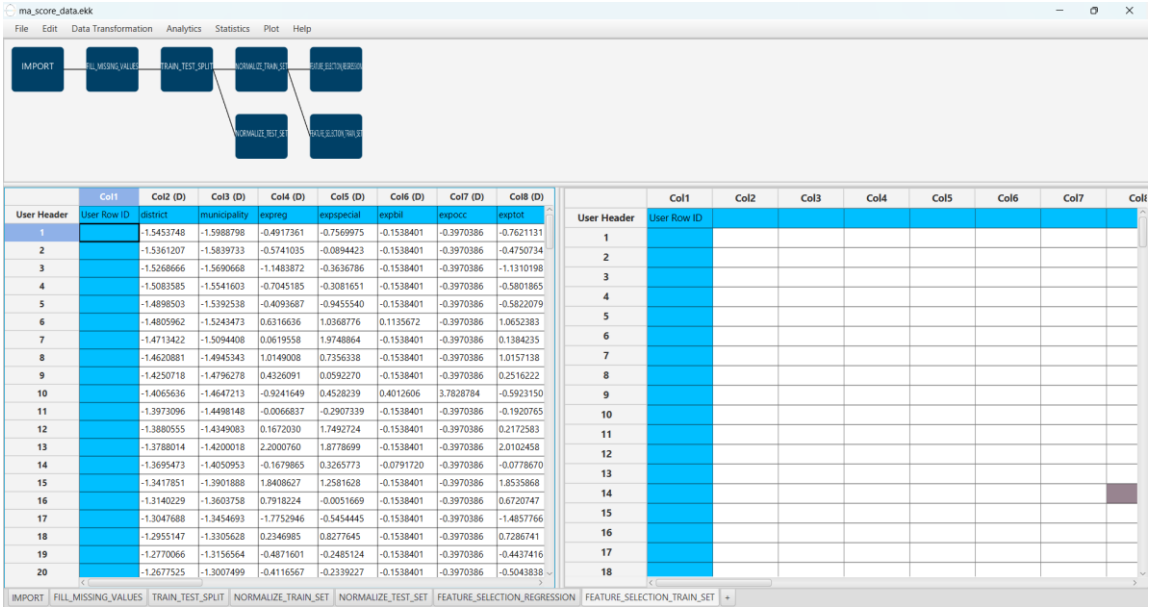
The significant features according to the p-value are the following:

- score4 (p-value = 0.0)
- lunch (p-value = 0.024945423143072513)
- score8 (p-value = 8.766402775289995E-7)
- income (p-value = 0.011862821687884247)
- english (p-value = 0.05453546054604336)

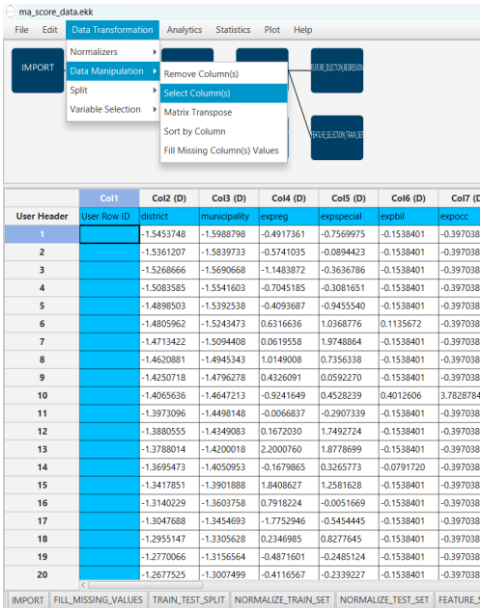
Step 8: Feature selection: train set

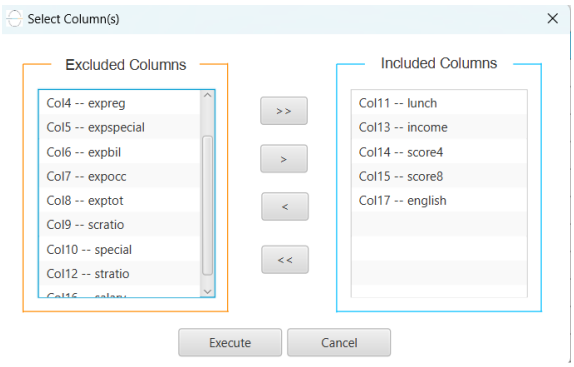
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_TRAIN_SET".

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".

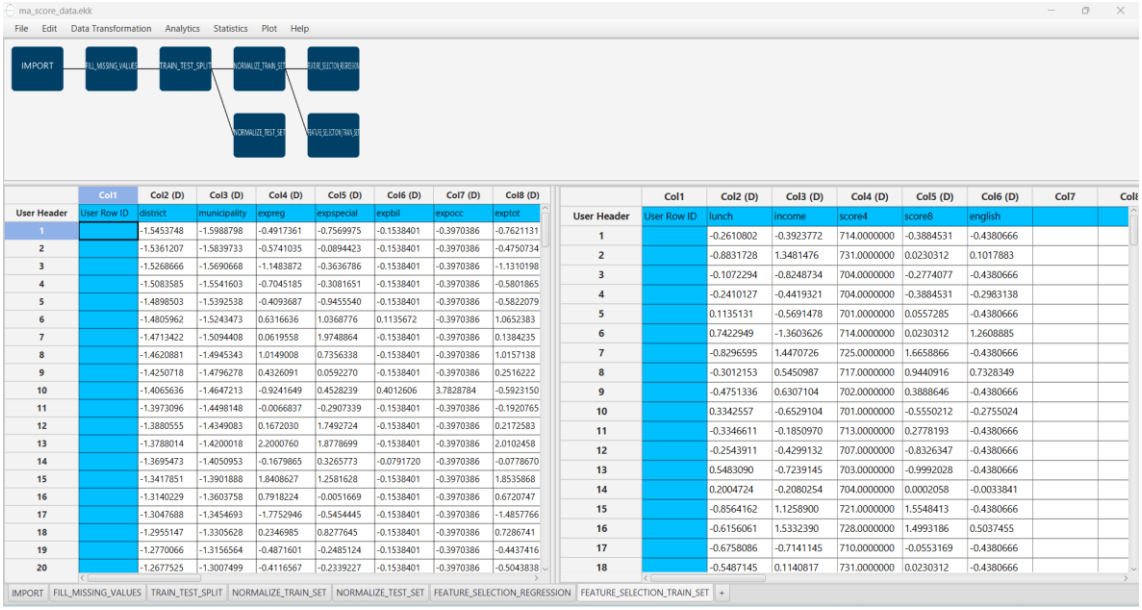


Manipulate the data by choosing the columns that correspond to the significant features (from the previous step): "Data Transformation" → "Data Manipulation" → "Select Column(s)".





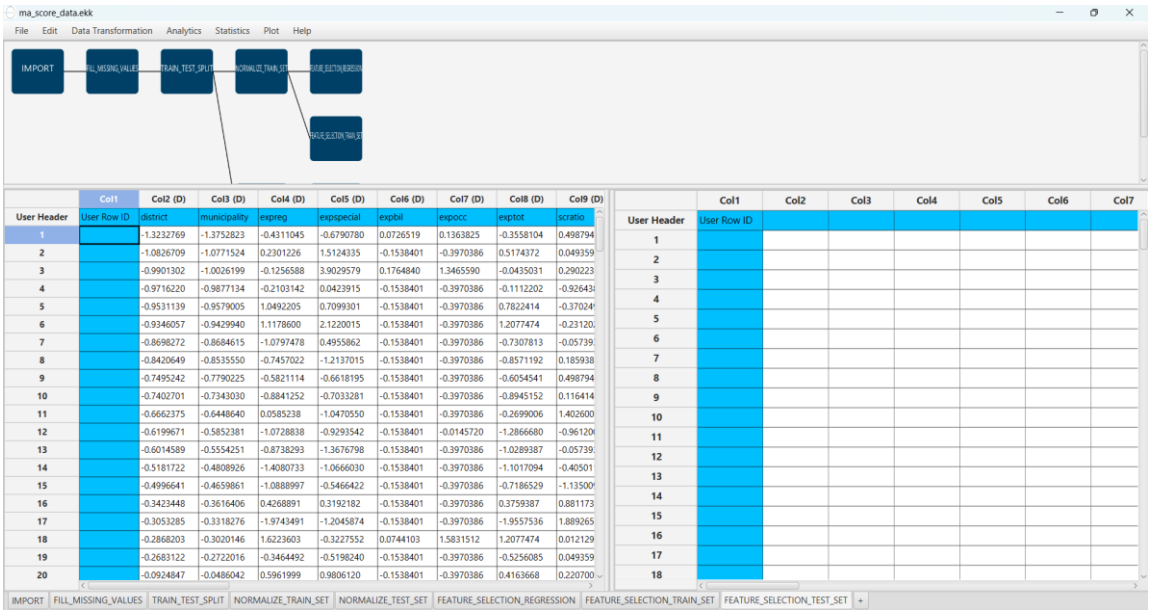
The results will appear on the output spreadsheet.



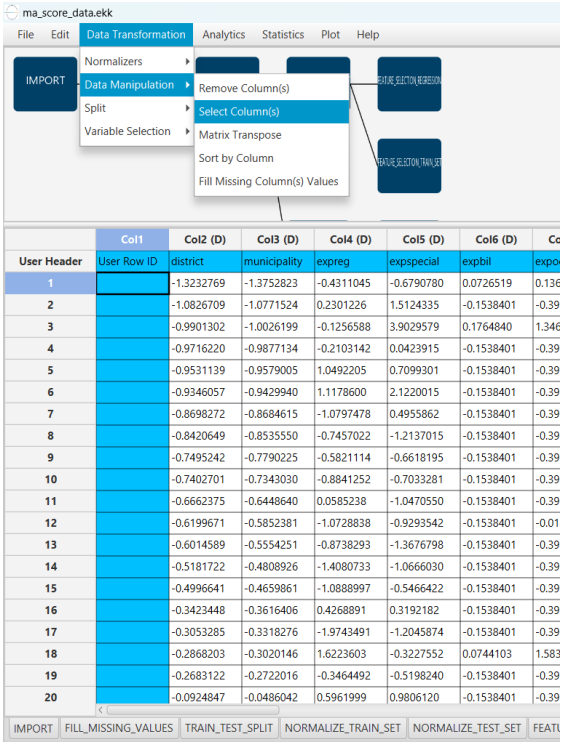
Step 9: Feature selection: test set

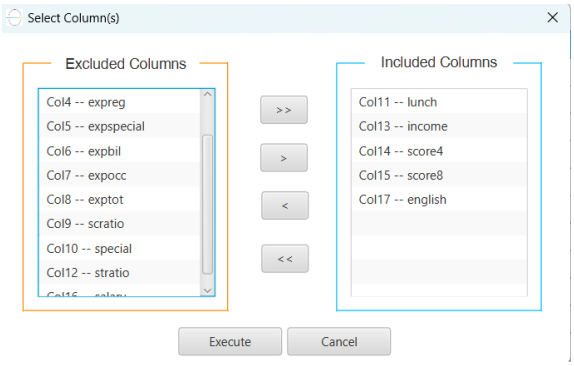
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_TEST_SET".

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".

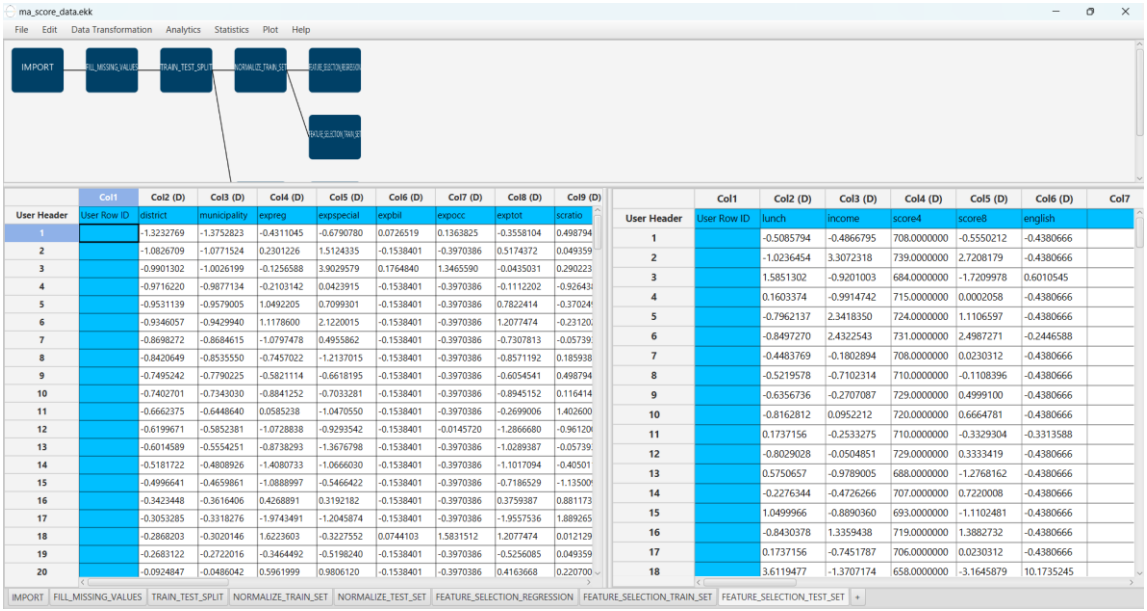


Manipulate the data by choosing the columns that correspond to the significant features (from the step 7): "Data Transformation" → "Data Manipulation" → "Select Column(s)".





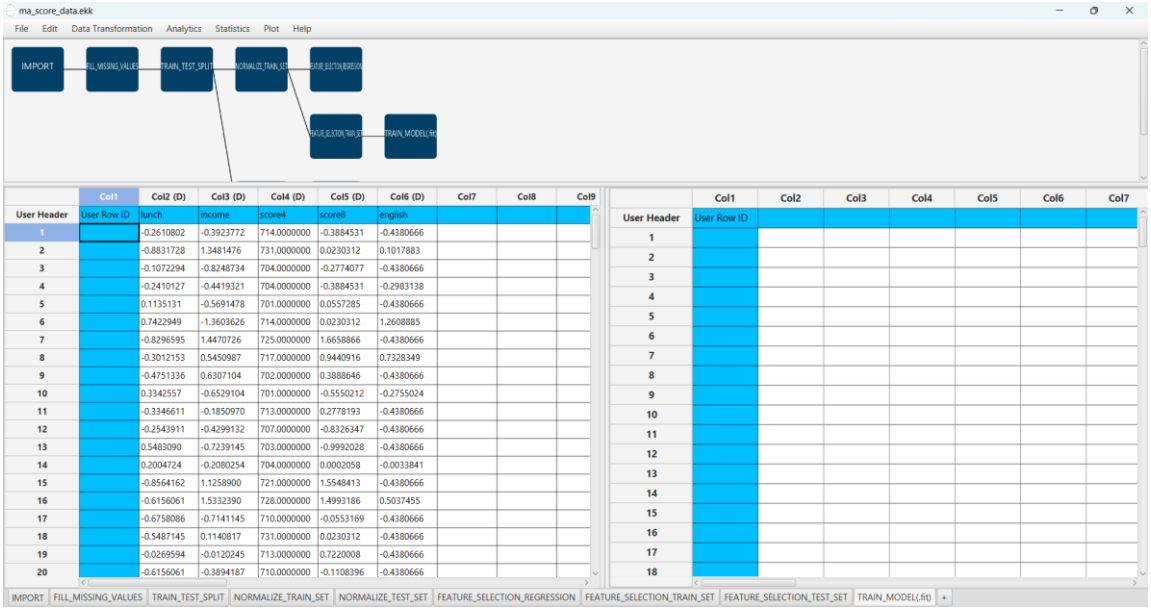
The results will appear on the output spreadsheet.



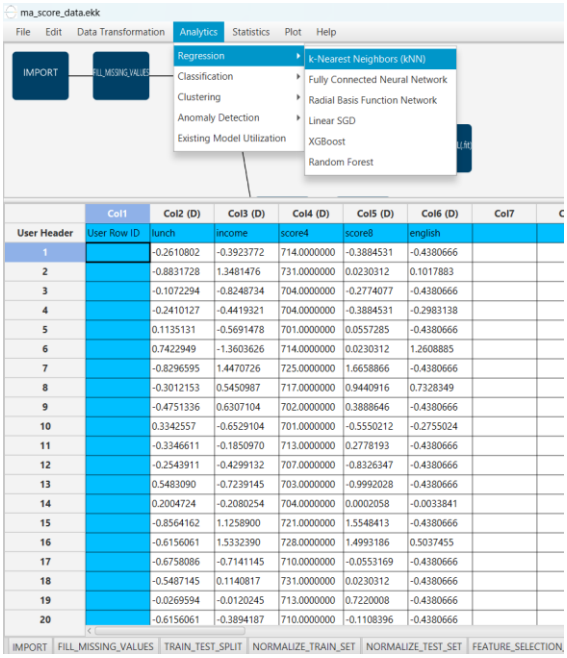
Step 10: Train the model

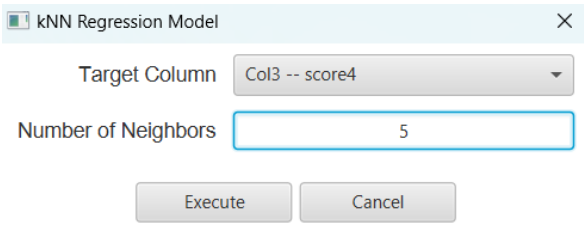
Create a new tab by pressing the "+" button on the bottom of the page with the name "TRAIN_MODEL(.fit)".

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".

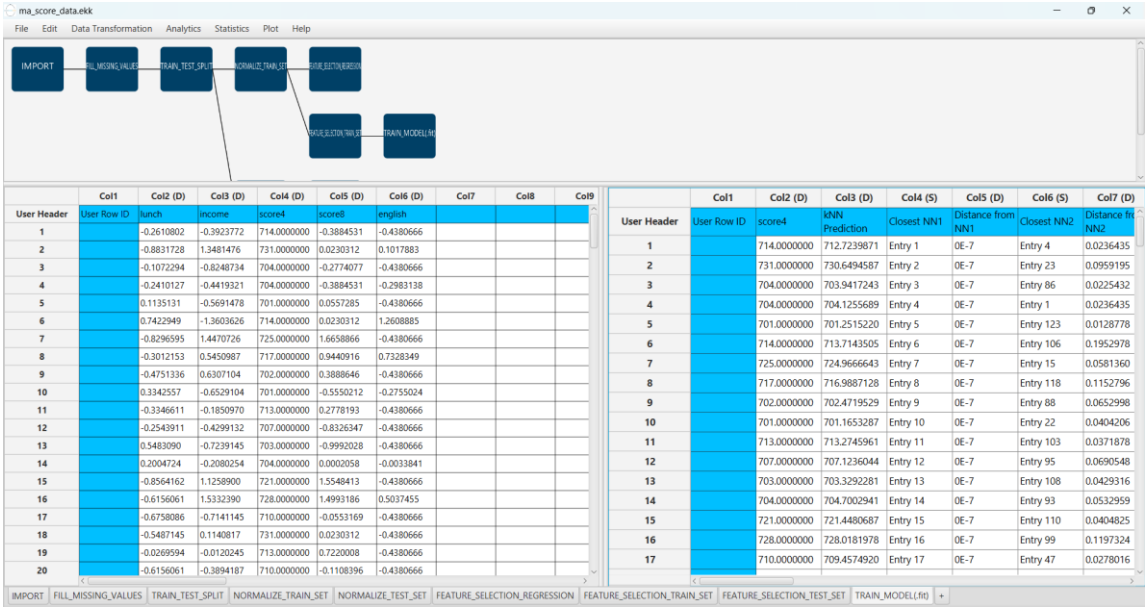


Use the k Nearest Neighbors (kNN) method to train and fit the model by browsing: "Analytics" → "Regression" → "k Nearest Neighbors (kNN)" and set the "Target Column" as the column corresponding to "score4" and the "Number of Neighbors" to 5.





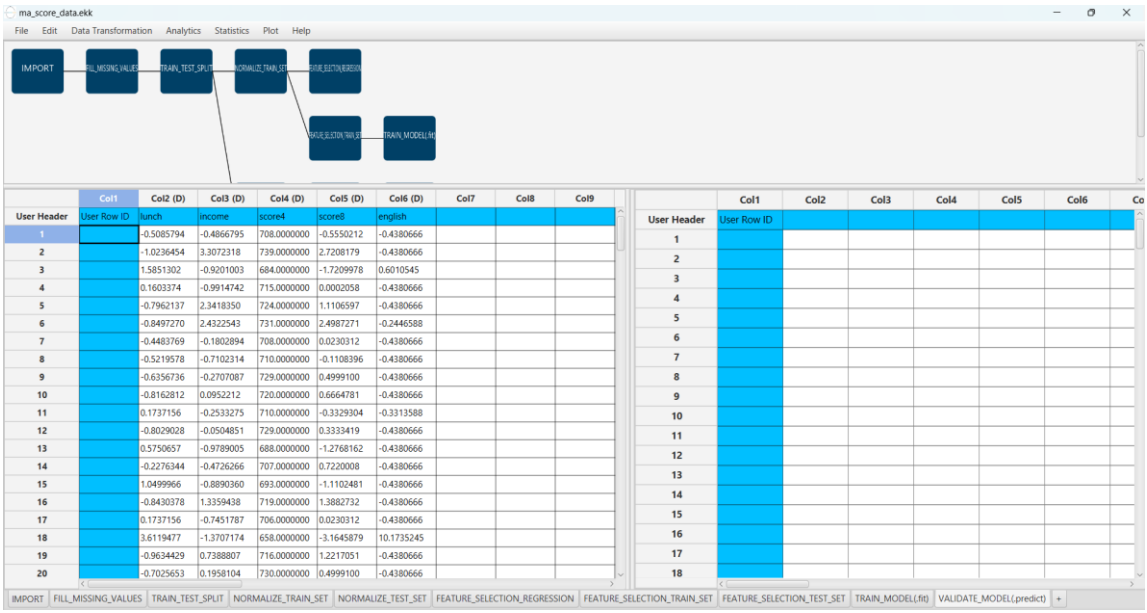
The predictions will appear on the output spreadsheet.



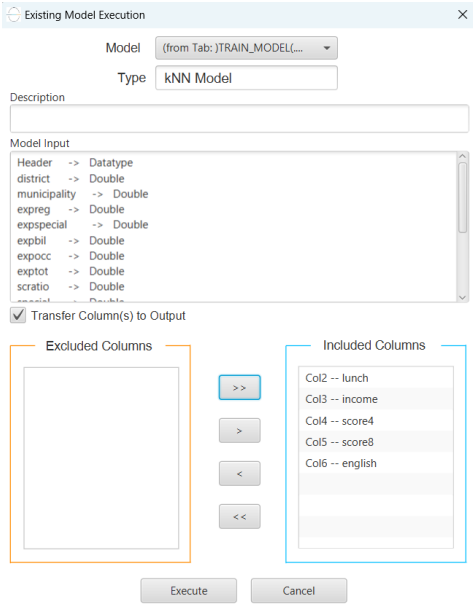
Step 11: Validate the model

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

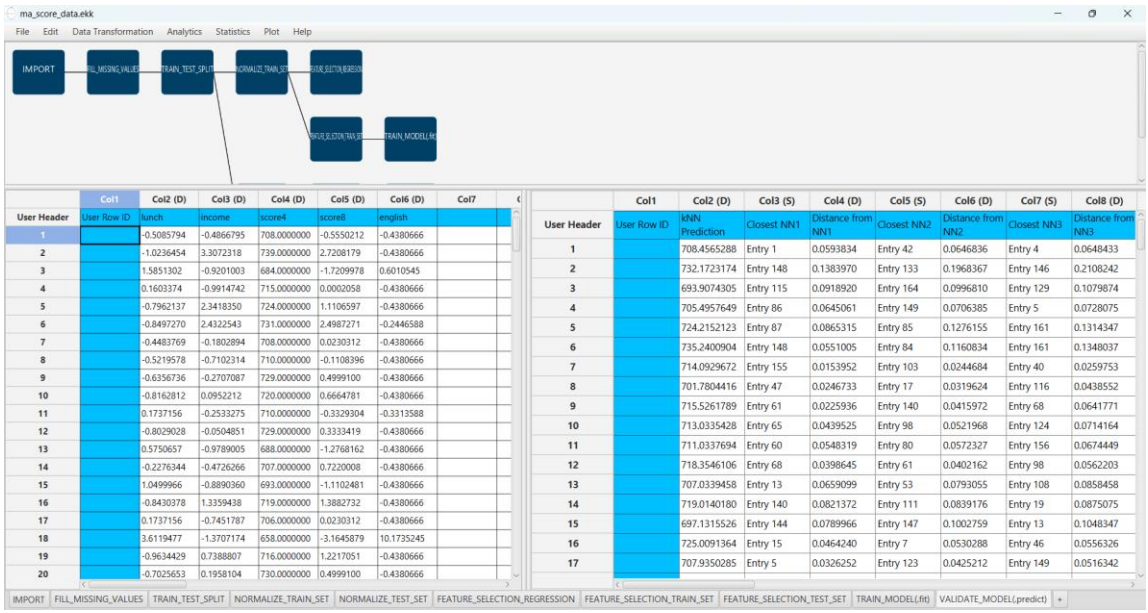
Import data into the input spreadsheet of the "VALIDATE_MODEL(.predict)" tab from the output of the "FEATURE_SELECTION_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



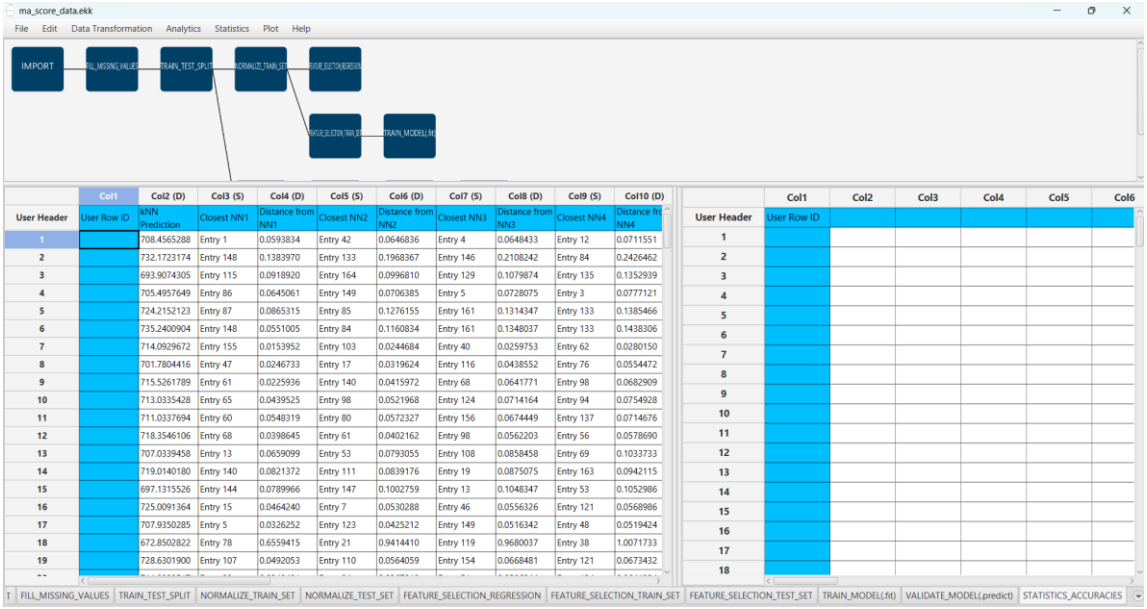
The predictions will appear on the output spreadsheet.



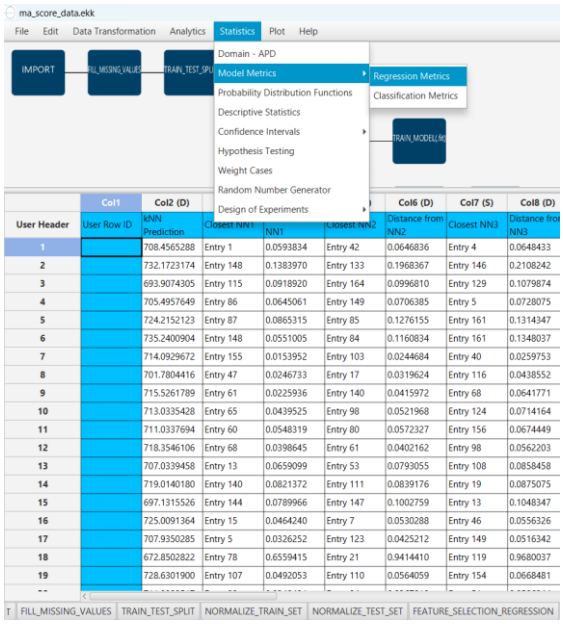
Step 12: Statistics calculation

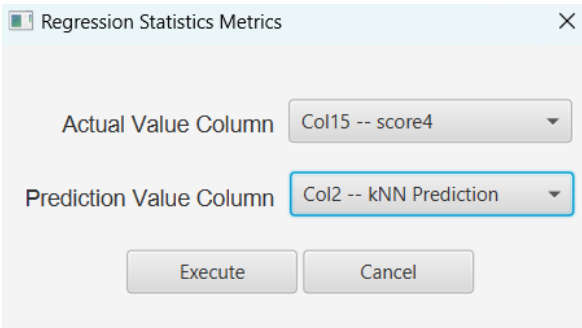
Create a new tab by pressing the "+" button on the bottom of the page with the name "STATISTICS_ACCURACIES".

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".

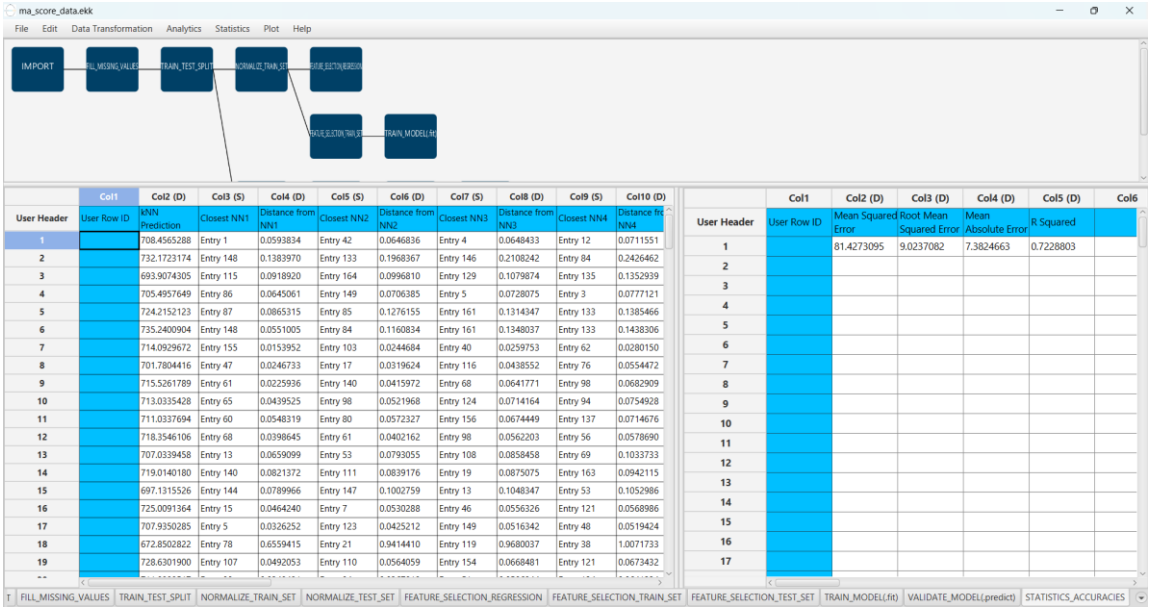


Calculate the statistical metrics for the regression by browsing: "Statistics" → "Model Metrics" → "Regression Metrics".





The results will appear on the output spreadsheet.

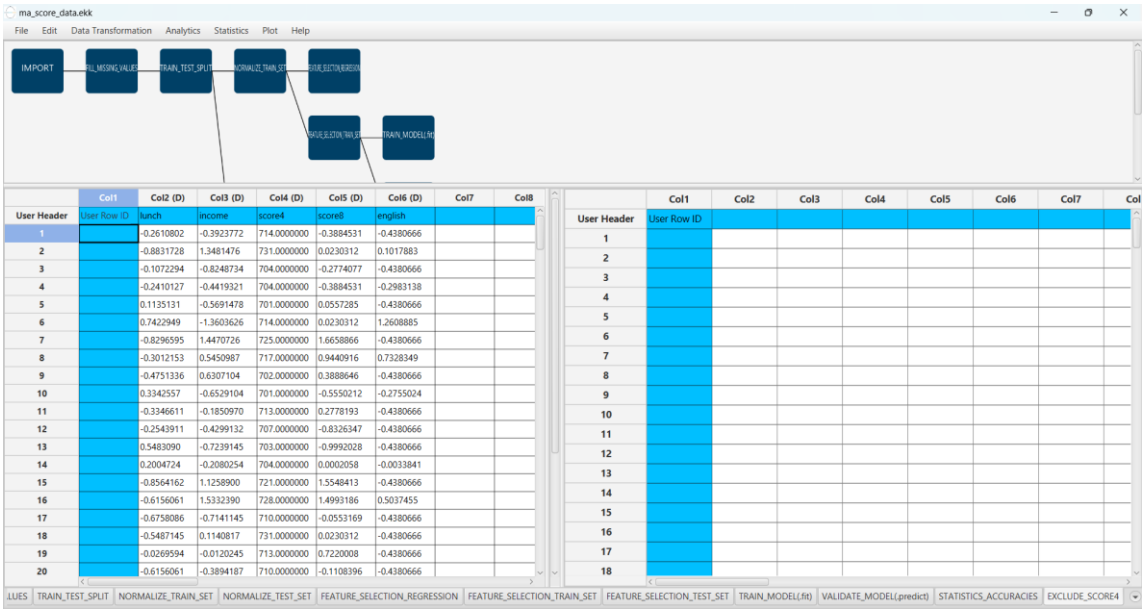


Step 13: Reliability check of each record of the test set

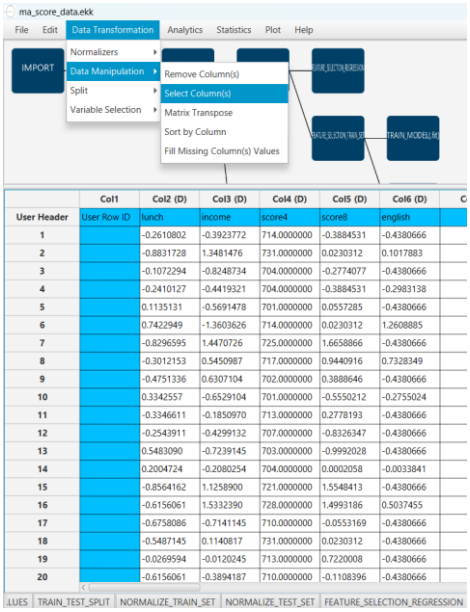
Step 13.a: Create the domain

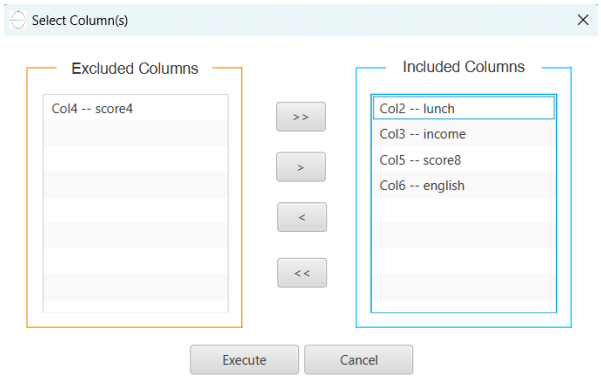
Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_SCORE4".

Import data into the input spreadsheet of the "EXCLUDE_SCORE4" 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 "score4" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except the "score4".

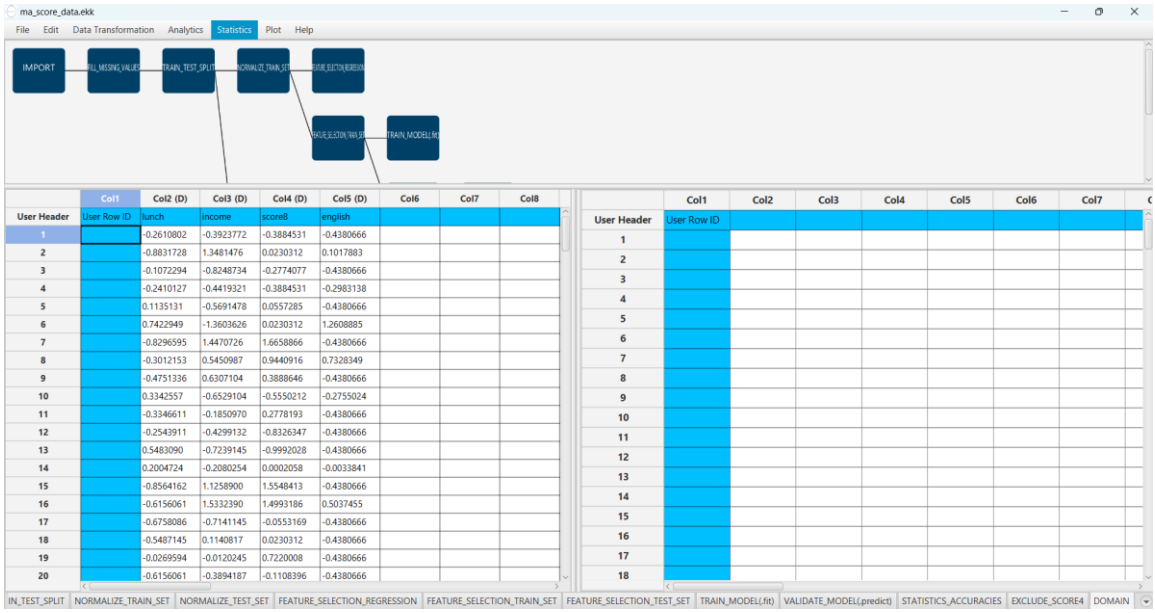




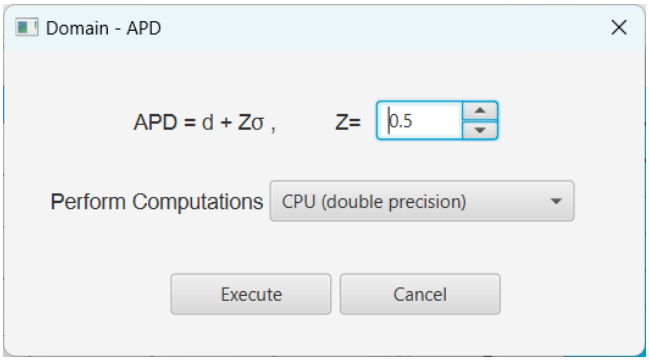
The results will appear on the output spreadsheet.

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

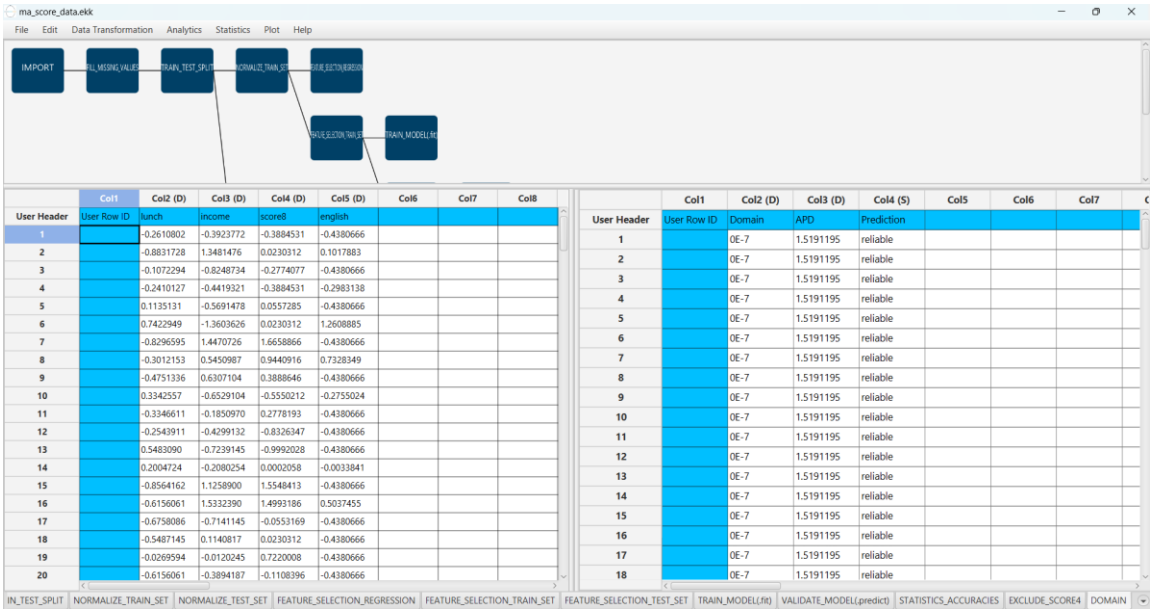
Import data into the input spreadsheet of the "DOMAIN" tab from the output of the "EXCLUDE_SCORE4" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



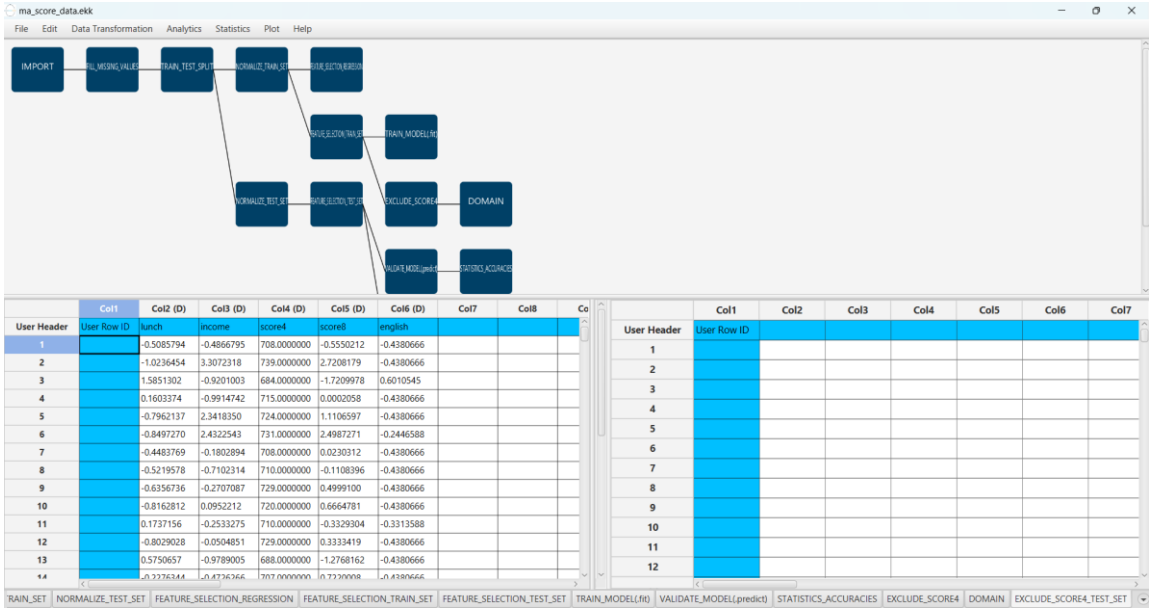
The results will appear on the output spreadsheet.



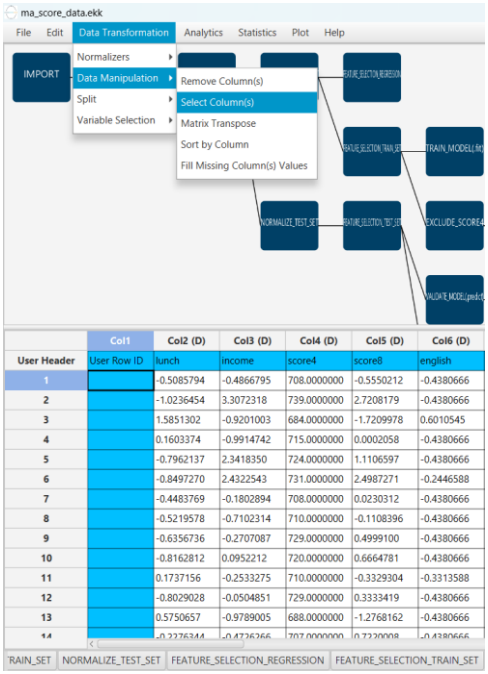
Step 12.b: Check the test set reliability

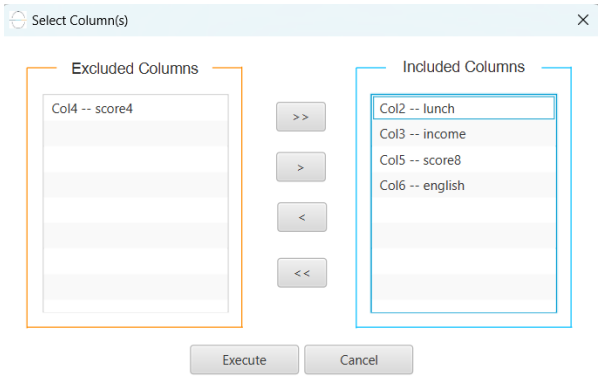
Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_SCORE4_TEST_SET".

Import data into the input spreadsheet of the "EXCLUDE_SCORE4_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 "score4" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except "score4".

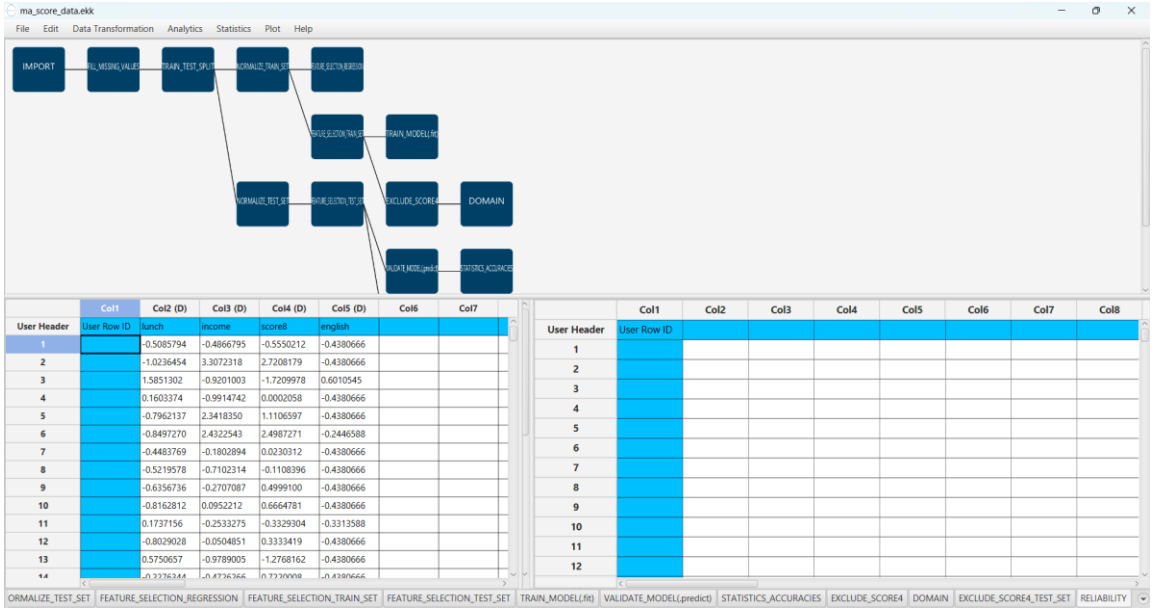




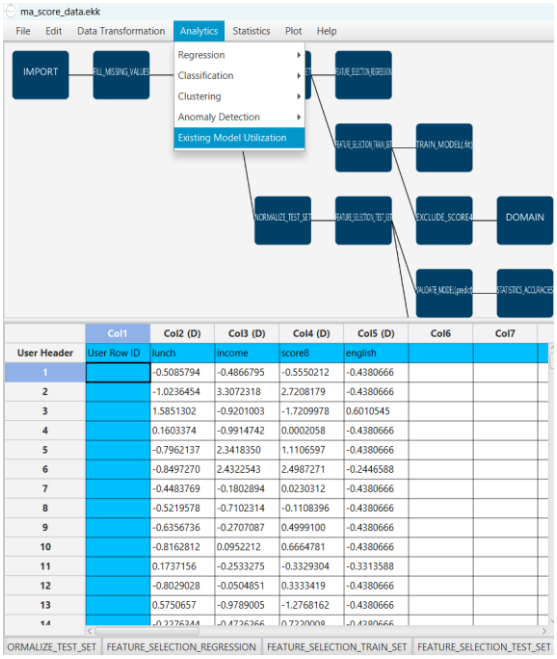
The results will appear on the output spreadsheet.

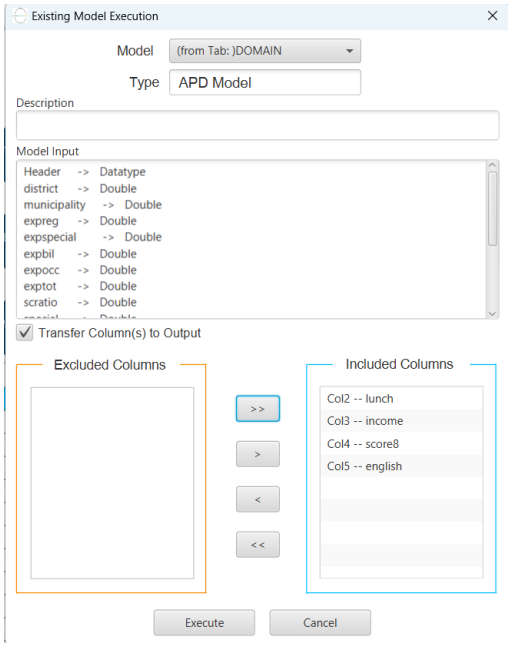
Create a new tab by pressing the "+" button on the bottom of the page with the name "RELIABILITY".

Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "EXCLUDE_SCORE4_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".

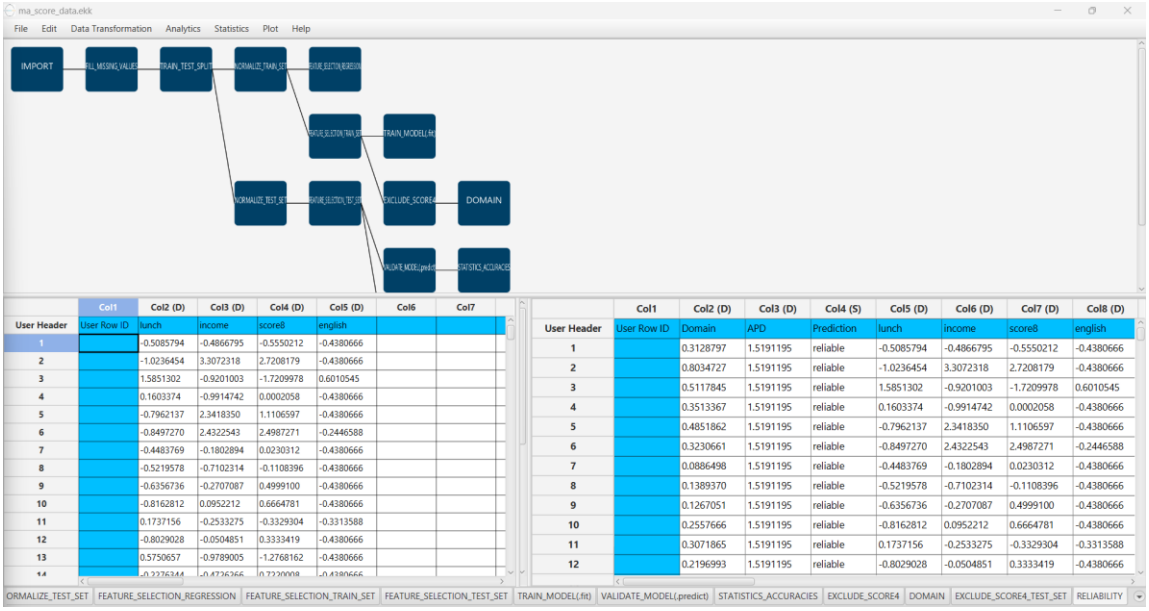


Check the Reliability by browsing: "Analytics" → "Existing Model Utilization". Then select as Model "(from Tab:) DOMAIN".





The results will appear on the output spreadsheet.



User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6	Col7
	User Row ID	lunch	income	score8	english		
1		-0.5085794	-0.4866795	-0.5550212	-0.4380666		
2		-1.0236454	3.3072318	2.7208179	-0.4380666		
3		1.5851302	-0.9201003	-1.7209978	0.6010545		
4		0.1603374	-0.9914742	0.0002058	-0.4380666		
5		-0.7962137	2.3418350	1.1106597	-0.4380666		
6		-0.8497270	2.4322543	2.4987271	-0.2446588		
7		-0.4483769	-0.1802894	0.0230312	-0.4380666		
8		-0.5219578	-0.7102314	-0.1108396	-0.4380666		
9		-0.6356736	-0.2707087	0.4999100	-0.4380666		
10		-0.8162812	0.0952212	0.6664781	-0.4380666		
11		0.1737156	-0.2533275	-0.3329304	-0.3313588		
12		-0.8029028	-0.0504851	0.3333419	-0.4380666		
13		0.5750657	-0.9789005	-1.2768162	-0.4380666		

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (S)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
	User Row ID	Domain	APD	Prediction	lunch	income	score8	english
1		0.3128797	1.5191195	reliable	-0.5085794	-0.4866795	-0.5550212	-0.4380666
2		0.8034727	1.5191195	reliable	-1.0236454	3.3072318	2.7208179	-0.4380666
3		0.5117845	1.5191195	reliable	1.5851302	-0.9201003	-1.7209978	0.6010545
4		0.3513367	1.5191195	reliable	0.1603374	-0.9914742	0.0002058	-0.4380666
5		0.4851862	1.5191195	reliable	-0.7962137	2.3418350	1.1106597	-0.4380666
6		0.3230661	1.5191195	reliable	-0.8497270	2.4322543	2.4987271	-0.2446588
7		0.0886498	1.5191195	reliable	-0.4483769	-0.1802894	0.0230312	-0.4380666
8		0.1389370	1.5191195	reliable	-0.5219578	-0.7102314	-0.1108396	-0.4380666
9		0.1267051	1.5191195	reliable	-0.6356736	-0.2707087	0.4999100	-0.4380666
10		0.2557666	1.5191195	reliable	-0.8162812	0.0952212	0.6664781	-0.4380666
11		0.3071865	1.5191195	reliable	0.1737156	-0.2533275	-0.3329304	-0.3313588
12		0.2196993	1.5191195	reliable	-0.8029028	-0.0504851	0.3333419	-0.4380666

There are 3 unreliable samples in the test set.

Final Isalos Workflow

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