



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.

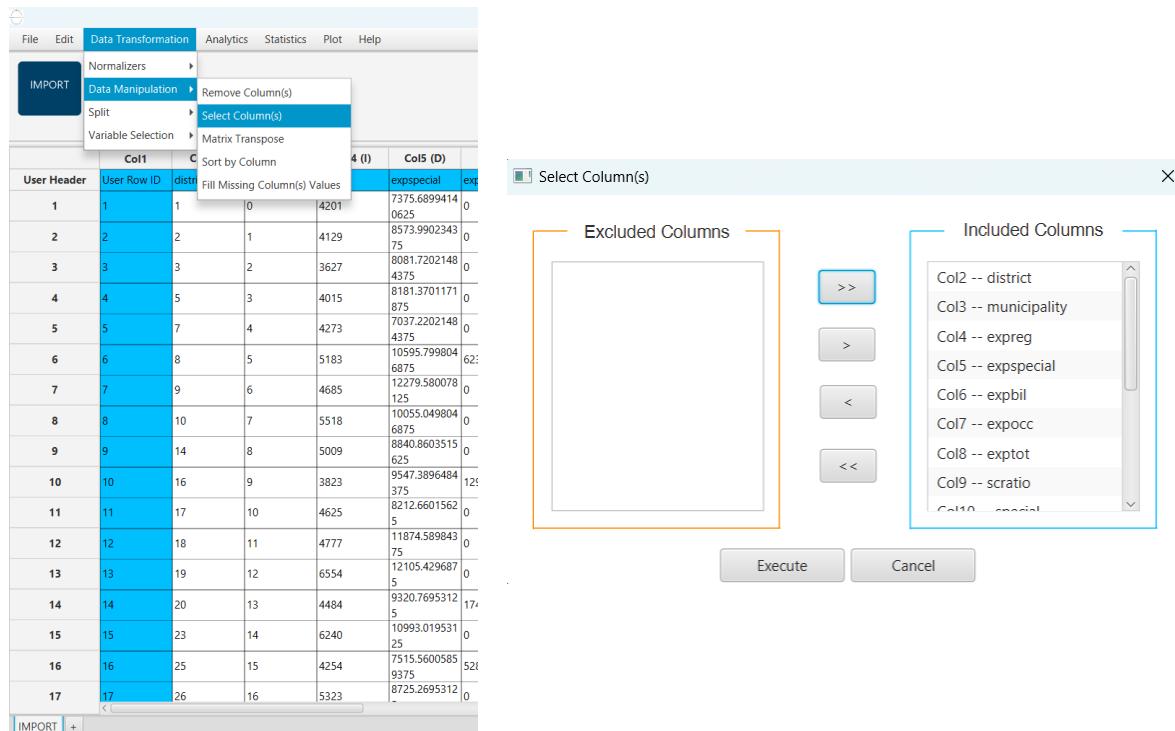
 A screenshot of the Isalos Analytics Platform interface. At the top, there's a menu bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help. Below the menu is a toolbar with an 'IMPORT' button. The main area shows a spreadsheet with 12 rows and 11 columns. Row 6 is highlighted in blue. A context menu is open over row 6, with options: Import from Spreadsheet, Import from file, Export Spread Sheet Data, and Clear SpreadSheet. The right side of the screen shows two vertical panes of the same spreadsheet data.

User Header	Col1	Col2	Col3	Col4	Col5	Col6
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						

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".

The screenshot shows the 'ma_score_data.ekk' spreadsheet interface. The 'IMPORT' and 'FILL_MISSING_VALUES' buttons are highlighted in blue. The spreadsheet itself contains data with columns labeled Col1 through Col8. The first few rows show values like:

User Header	User Row ID	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (D)	Col6 (I)	Col7 (I)	Col8 (I)
1	1	1.0000000	0E-7	4201.0000000	7375.6899414	0E-7	4646.0000000		
2	2	2.0000000	1.0000000	4129.0000000	8573.9902344	0E-7	4930.0000000		
3	3	3.0000000	2.0000000	3627.0000000	8081.7202148	0E-7	4281.0000000		
4	4	5.0000000	3.0000000	4015.0000000	8181.3701172	0E-7	4826.0000000		
5	5	7.0000000	4.0000000	4273.0000000	7037.2202148	0E-7	4824.0000000		
6	6	8.0000000	5.0000000	5183.0000000	10595.7998047	6235.0000000	6454.0000000		
7	7	9.0000000	6.0000000	4685.0000000	12279.5800781	0E-7	5537.0000000		
8	8	10.0000000	7.0000000	5518.0000000	10055.0498077	0E-7	6405.0000000		
9	9	14.0000000	8.0000000	5009.0000000	8840.8603516	0E-7	5649.0000000		
10	10	16.0000000	9.0000000	3823.0000000	9547.3896484	12943.0000000	11519.0000000		
11	11	17.0000000	10.0000000	4625.0000000	8212.6601563	0E-7	5210.0000000		
12	12	18.0000000	11.0000000	4777.0000000	11874.5898438	0E-7	5615.0000000		
13	13	19.0000000	12.0000000	6554.0000000	12105.4296875	0E-7	7389.0000000		
14	14	20.0000000	13.0000000	4484.0000000	9320.7695313	1741.0000000	5233.0000000		
15	15	23.0000000	14.0000000	6240.0000000	10993.0195313	0E-7	7234.0000000		
16	16	25.0000000	15.0000000	4254.0000000	7515.5600586	5281.0000000	1470.0000000	5048.0000000	
17	17	26.0000000	16.0000000	5323.0000000	8725.2695313	0E-7	6065.0000000		
18	18	27.0000000	17.0000000	3079.0000000	7755.4399414	0E-7	3930.0000000		

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 screenshot shows the 'Data Transformation' menu open, with 'Data Manipulation' selected. A sub-menu 'Fill Missing Column(s) Values' is highlighted. To the right, a dialog box titled 'Fill Missing Column(s) Values' is open, showing the following settings:

- Excluded Columns:** Col2 -- district, Col3 -- municipality, Col4 -- expreg, Col5 -- expspecial, Col6 -- expbil, Col7 -- expocc, Col8 -- exptot, Col9 -- scratio, Col10 -- special
- Included Columns:** All other columns (Col1, Col2, Col3, Col4, Col5, Col6, Col7, Col8, Col9, Col10)
- Numerical Method:** Mean
- Span:** Integer (0,+∞), Default: -
- Categorical Method:** Categorical Method

The data will appear in the output spreadsheet.

The screenshot shows two tabs side-by-side. The left tab is titled 'ma_score_data.ekk' and contains raw data with columns labeled Col1 through Col8. The right tab is also titled 'ma_score_data.ekk' and contains the same data with missing values filled in, indicated by the '(1)' suffix. This tab includes an additional column labeled 'label'.

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".

The screenshot shows the 'TRAIN_TEST_SPLIT' tab, which is the active tab. It contains the same data as the 'FILL_MISSING_VALUES' tab, including columns Col1 through Col8 and an additional 'label' column. The tab has a blue header and is the active tab.

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 screenshot shows the Isalos Analytics Platform interface. In the top menu, 'Data Transformation' is selected. Under 'Data Manipulation', 'Split' is chosen, which then leads to 'Random Partitioning'. A sub-dialog titled 'Random Partitioning' is open, showing a 'Training set percentage' input field set to 75, a checked 'Usage of random generator seed' field with value 79827627322600, and a checked 'Stratified sampling' checkbox with dropdown 'Col16 -- salary'. Below these are 'Execute' and 'Cancel' buttons.

User Header	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (D)	Col6 (I)	Col7 (I)	Col8 (I)	Co
1	1.0000000	OE-7	4201.0000000	7375.6899414	OE-7	OE-7	4646.0000000	16.60	
2	2.0000000	1.0000000	4129.0000000	8573.9902344	OE-7	OE-7	4930.0000000	5.699	
3	3.0000000	2.0000000	3627.0000000	8081.7202148	OE-7	OE-7	4281.0000000	7.500	
4	5.0000000	3.0000000	4015.0000000	8181.3701172	OE-7	OE-7	4826.0000000	8.600	
5	7.0000000	4.0000000	4273.0000000	7037.2202148	OE-7	OE-7	4824.0000000	6.099	
6	8.0000000	5.0000000	5183.0000000	10595.799804	6235.0000000	OE-7	6454.0000000	7.699	
7	9.0000000	6.0000000	4685.0000000	12279.580078	1	OE-7	5537.0000000	5.400	
8	10.0000000	7.0000000	10055.049804	7	OE-7	OE-7	6405.0000000	7.099	
9	14.0000000	8.0000000	5009.0000000	8940.8603516	OE-7	OE-7	5649.0000000	10.60	
10	16.0000000	9.0000000	3823.0000000	9547.3896484	0	12943.0000000	11519.0000000	OE-7	4814.0000000
11	17.0000000	10.0000000	4625.0000000	8212.6601563	OE-7	OE-7	5210.0000000	12.50	
12	18.0000000	11.0000000	4777.0000000	11874.589843	8	OE-7	5615.0000000	7.599	
13	19.0000000	12.0000000	6554.0000000	12105.429687	5	OE-7	7389.0000000	4.199	
14	20.0000000	13.0000000	4484.0000000	9320.7695313	1741.0000000	OE-7	5323.0000000	8.800	
15	23.0000000	14.0000000	6240.0000000	10993.019531	3	OE-7	7234.0000000	4.800	
16	25.0000000	15.0000000	4254.0000000	7515.5600586	5281.0000000	1470.0000000	5048.0000000	9.399	
17	26.0000000	17.0000000	3079.0000000	7755.4399414	OE-7	OE-7	3930.0000000	13.800	
18	28.0000000	18.0000000	4836.0000000	10228.429687	5	OE-7	6121.0000000	6.000	
19	30.0000000	19.0000000	4205.0000000	8288.4501953	OE-7	OE-7	4961.0000000	9.800	
20	31.0000000	20.0000000	4205.0000000	8288.4501953	OE-7	OE-7	4901.0000000		

The results will appear on the output spreadsheet.

The screenshot shows the Isalos Analytics Platform interface with two spreadsheets side-by-side. The left spreadsheet has tabs at the bottom: 'IMPORT', 'FILL_MISSING_VALUES', and 'TRAIN_TEST_SPLIT'. The right spreadsheet also has tabs at the bottom: 'IMPORT', 'FILL_MISSING_VALUES', 'TRAIN_TEST_SPLIT', and 'NORMALIZE_TRAIN_SET'. Both spreadsheets contain identical data, showing columns: Col1, Col2 (I), Col3 (I), Col4 (I), Col5 (D), Col6 (I), Col7 (I), Col8 (I), and Col9.

User Header	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (D)	Col6 (I)	Col7 (I)	Col8 (I)	Col9
1	1.0000000	OE-7	4201.0000000	7375.6899414	OE-7	OE-7	4646.0000000	16.60	
2	2.0000000	1.0000000	4129.0000000	8573.9902344	OE-7	OE-7	4930.0000000	5.699	
3	3.0000000	2.0000000	3627.0000000	8081.7202148	OE-7	OE-7	4281.0000000	7.500	
4	5.0000000	3.0000000	4015.0000000	8181.3701172	OE-7	OE-7	4826.0000000	8.600	
5	7.0000000	4.0000000	4273.0000000	7037.2202148	OE-7	OE-7	4824.0000000	6.099	
6	8.0000000	5.0000000	5183.0000000	10595.799804	6235.0000000	OE-7	6454.0000000	7.699	
7	9.0000000	6.0000000	4685.0000000	12279.580078	1	OE-7	5537.0000000	5.400	
8	10.0000000	7.0000000	10055.049804	7	OE-7	OE-7	6405.0000000	7.099	
9	14.0000000	8.0000000	5009.0000000	8940.8603516	OE-7	OE-7	5649.0000000	10.60	
10	16.0000000	9.0000000	3823.0000000	9547.3896484	0	12943.0000000	11519.0000000	OE-7	4814.0000000
11	17.0000000	10.0000000	4625.0000000	8212.6601563	OE-7	OE-7	5210.0000000	12.50	
12	18.0000000	11.0000000	4777.0000000	11874.589843	8	OE-7	5615.0000000	7.599	
13	19.0000000	12.0000000	6554.0000000	12105.429687	5	OE-7	7389.0000000	4.199	
14	20.0000000	13.0000000	4484.0000000	9320.7695313	1741.0000000	OE-7	5323.0000000	8.800	
15	23.0000000	14.0000000	6240.0000000	10993.019531	3	OE-7	7234.0000000	4.800	
16	25.0000000	15.0000000	4254.0000000	7515.5600586	5281.0000000	1470.0000000	5048.0000000	9.399	
17	26.0000000	16.0000000	5323.0000000	8725.2695313	OE-7	OE-7	6065.0000000	15.00	
18	27.0000000	17.0000000	3079.0000000	7755.4399414	OE-7	OE-7	3930.0000000	13.800	
19	28.0000000	18.0000000	4836.0000000	10228.429687	5	OE-7	6121.0000000	6.000	
20	30.0000000	19.0000000	4205.0000000	8288.4501953	OE-7	OE-7	4961.0000000	9.800	

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"

The screenshot shows a data flow interface with the following steps:

- IMPORT**
- FILL_MISSING_VALUES**
- TRAIN_TEST_SPLIT**
- NORMALIZE_TRAIN_SET**

The data table has the following structure:

User Header	User Row ID	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (D)	Col6 (I)	Col7 (I)	Col8 (I)	Col9 (D)
1	1.000000	0E-7	4201.000000	7375.689941	0E-7	0E-7	4646.000000	16.6000		
2	2.000000	1.000000	4129.000000	8573.990234	0E-7	0E-7	4930.000000	5.6999		
3	3.000000	2.000000	3627.000000	8081.720214	0E-7	0E-7	4281.000000	7.5000		
4	5.000000	3.000000	4015.000000	8181.370117	0E-7	0E-7	4826.000000	8.6000		
5	7.000000	4.000000	4273.000000	7037.220214	0E-7	0E-7	4824.000000	6.0999		
6	8.000000	5.000000	5183.000000	10595.799804	6235.000000	0E-7	6454.000000	7.6999		
7	9.000000	6.000000	4685.000000	12279.580078	0E-7	0E-7	5537.000000	5.4000		
8	10.000000	7.000000	5518.000000	10055.049804	7	0E-7	6405.000000	7.0999		
9	14.000000	8.000000	5009.000000	8840.860351	6	0E-7	5649.000000	10.6000		
10	16.000000	9.000000	3823.000000	9547.389648	4	12943.000000	11519.000000	4814.000000	5.6999	
11	17.000000	10.000000	4625.000000	8212.660156	3	0E-7	5210.000000	12.5000		
12	18.000000	11.000000	4777.000000	11874.589843	8	0E-7	5615.000000	7.5999		
13	19.000000	12.000000	6554.000000	12105.429687	5	0E-7	7389.000000	4.1999		
14	20.000000	13.000000	4484.000000	9320.769531	17	1471.000000	5323.000000	8.8000		
15	23.000000	14.000000	6240.000000	10993.019531	3	0E-7	7234.000000	4.8000		
16	26.000000	16.000000	5323.000000	8725.269531	7	0E-7	6065.000000	15.0000		
17	27.000000	17.000000	3079.000000	7755.439941	1	0E-7	3930.000000	13.8000		
18	28.000000	18.000000	4836.000000	10220.429687	5	0E-7	6121.000000	6.0000		
19	30.000000	19.000000	4205.000000	8288.450195	3	0E-7	4961.000000	9.8000		
20	31.000000	20.000000	4271.000000	8314.639648	4	0E-7	4901.000000	10.1000		

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

The screenshot shows a data flow interface with the following steps:

- IMPORT**
- Data Transformation** (with **Normalizers** selected)
- Min-Max**
- NORMALIZE_TRAIN_SET**

The dialog box for **ZScore Normalizer** has the following settings:

- Excluded Columns:** Col14 -- score4
- Included Columns:** Col8 -- exptot, Col9 -- scratio, Col10 -- special, Col11 -- lunch, Col12 -- stratio, Col13 -- income, Col15 -- score8, Col16 -- salary, Col17 -- english

The results will appear on the output spreadsheet.

ma_score_data.ekk

User Header	User Row ID	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (D)	Col6 (I)	Col7 (I)	Col8 (I)	Col9 (D)	Col10
1	1.0000000	0E-7	4201.0000000	7375.6899414	0E-7	0E-7	4646.0000000	16.60000			
2	2.0000000	1.0000000	4129.0000000	8573.9902344	0E-7	0E-7	4930.0000000	5.69999			
3	3.0000000	2.0000000	3627.0000000	8081.7202148	0E-7	0E-7	4281.0000000	7.50000			
4	5.0000000	3.0000000	4015.0000000	8181.3701172	0E-7	0E-7	4826.0000000	8.60000			
5	7.0000000	4.0000000	4273.0000000	7037.2202148	0E-7	0E-7	4824.0000000	6.09999			
6	8.0000000	5.0000000	5183.0000000	10595.799804	6235.0000000	0E-7	6454.0000000	7.69999			
7	9.0000000	6.0000000	4685.0000000	12279.580078	0E-7	0E-7	5537.0000000	5.40000			
8	10.0000000	7.0000000	5518.0000000	10055.049804	0E-7	0E-7	6405.0000000	7.09999			
9	14.0000000	8.0000000	5009.0000000	8840.8603516	0E-7	0E-7	5649.0000000	10.60000			
10	16.0000000	9.0000000	3823.0000000	9547.3896484	12943.0000000	11519.0000000	0E-7	4814.0000000	6.69999		
11	17.0000000	10.0000000	4625.0000000	8212.6601563	0E-7	0E-7	5210.0000000	12.50000			
12	18.0000000	11.0000000	4777.0000000	11874.589843	0E-7	0E-7	5615.0000000	7.59999			
13	19.0000000	12.0000000	6554.0000000	12105.429687	0E-7	0E-7	7389.0000000	4.19999			
14	20.0000000	13.0000000	4484.0000000	9320.7695313	1741.0000000	0E-7	5323.0000000	8.80000			
15	23.0000000	14.0000000	10993.019531	3	0E-7	0E-7	7234.0000000	4.80000			
16	26.0000000	16.0000000	5233.0000000	8725.2695313	0E-7	0E-7	6065.0000000	15.00000			
17	27.0000000	17.0000000	3079.0000000	7755.4399414	0E-7	0E-7	3930.0000000	13.80000			
18	28.0000000	18.0000000	4836.0000000	10220.429687	0E-7	0E-7	6121.0000000	6.00000			
19	30.0000000	19.0000000	4205.0000000	8288.4501953	0E-7	0E-7	4961.0000000	9.80000			
20	31.0000000	20.0000000	4271.0000000	8314.6396484	0E-7	0E-7	4901.0000000	10.10000			

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 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".

ma_score_data.ekk

User Header	User Row ID	Col1	Col2	Col3	Col4	Col5	Col6	Col7
1	25.0000000	15.0000000	4254.0000000	7515.560056	5281.0000000	1470.0000000	5048.0000000	9.399996
2	51.0000000	35.0000000	4832.0000000	11449.450195	0E-7	0E-7	5912.0000000	8.107100
3	61.0000000	40.0000000	4521.0000000	15740.580078	7702.0000000	4805.0000000	5357.0000000	8.800002
4	63.0000000	41.0000000	4447.0000000	8810.6396484	0E-7	0E-7	5290.0000000	5.300002
5	65.0000000	43.0000000	10008.910156	3	0E-7	0E-7	6174.0000000	6.900001
6	67.0000000	44.0000000	5608.0000000	12543.660156	0E-7	0E-7	6595.0000000	7.300002
7	74.0000000	49.0000000	3687.0000000	9624.1503906	0E-7	0E-7	4677.0000000	7.800002
8	77.0000000	50.0000000	3979.0000000	6555.8798828	0E-7	0E-7	4552.0000000	8.500000
9	87.0000000	55.0000000	4122.0000000	7546.5400391	0E-7	0E-7	4801.0000000	9.399996
10	88.0000000	58.0000000	3858.0000000	7472.0297852	0E-7	0E-7	4515.0000000	8.300002
11	96.0000000	64.0000000	4682.0000000	6855.020195	0E-7	0E-7	5133.0000000	12.000000
12	101.0000000	68.0000000	3693.0000000	7066.2998047	0E-7	1054.0000000	4127.0000000	5.199998
13	103.0000000	70.0000000	3867.0000000	6279.4799805	0E-7	0E-7	4382.0000000	7.800002
14	112.0000000	75.0000000	3400.0000000	6819.9301758	0E-7	0E-7	4310.0000000	6.800002
15	114.0000000	76.0000000	3679.0000000	7753.2900391	0E-7	0E-7	4689.0000000	4.699998
16	131.0000000	83.0000000	5004.0000000	9307.5595703	0E-7	0E-7	5772.0000000	10.500000
17	135.0000000	85.0000000	2905.0000000	6572.2402344	0E-7	0E-7	3465.0000000	13.399996
18	137.0000000	87.0000000	6049.0000000	8155.1801758	5322.0000000	5457.0000000	6595.0000000	8.000000
19	139.0000000	89.0000000	4328.0000000	7801.4301758	0E-7	0E-7	4880.0000000	8.107100
20	158.0000000	104.0000000	5152.0000000	10494.799804	7	0E-7	5812.0000000	8.600004
21	159.0000000	105.0000000	4961.0000000	8982.8896484	0E-7	0E-7	5781.0000000	7.800002

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 screenshot shows the Isalos Analytics Platform interface. On the left, the 'ma_score_data.ekk' window displays a data flow starting with an 'IMPORT' node, followed by a 'FILL_MISSING_VALUES' node, and then a table with columns labeled Col1 through Col9. The table has 21 rows of data. On the right, the 'Existing Model Execution' window shows a configuration for a 'Z Score Normalizer Model'. It includes a 'Model' dropdown set to '(from Tab:) NORMALIZE_TRAIN_SET...', a 'Type' dropdown set to 'Z Score Normalizer Model', and a 'Description' field. Below these are sections for 'Model Input' (listing columns and their dtypes) and 'Transfer Column(s) to Output' (checkbox checked). To the right are 'Excluded Columns' (empty) and 'Included Columns' (listing columns Col2 through Col9). At the bottom are 'Execute' and 'Cancel' buttons.

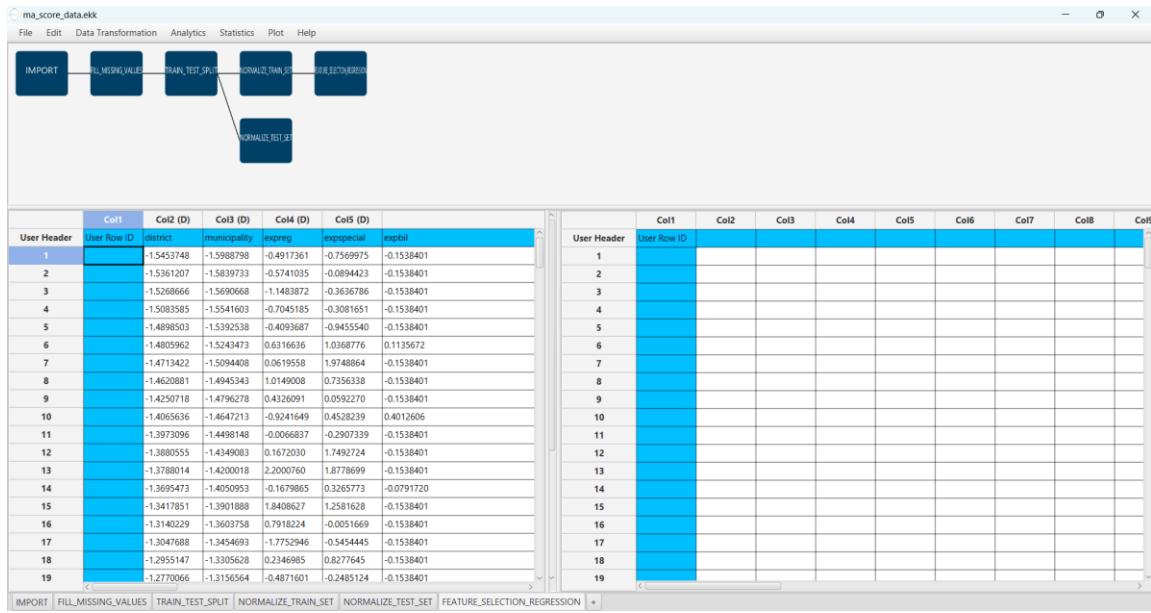
The results will appear on the output spreadsheet.

The screenshot shows the Isalos Analytics Platform interface again. The 'ma_score_data.ekk' window is open, showing a data flow with 'IMPORT' and 'FILL_MISSING_VALUES' nodes, followed by two tabs: 'TRAIN_TEST_SPLIT' and 'NORMALIZE_TRAIN_SET'. The 'NORMALIZE_TRAIN_SET' tab is active, displaying a normalized dataset with columns Col1 through Col9. The 'ma_score_data.ekk' tab shows the original dataset. Both tabs have 21 rows of data.

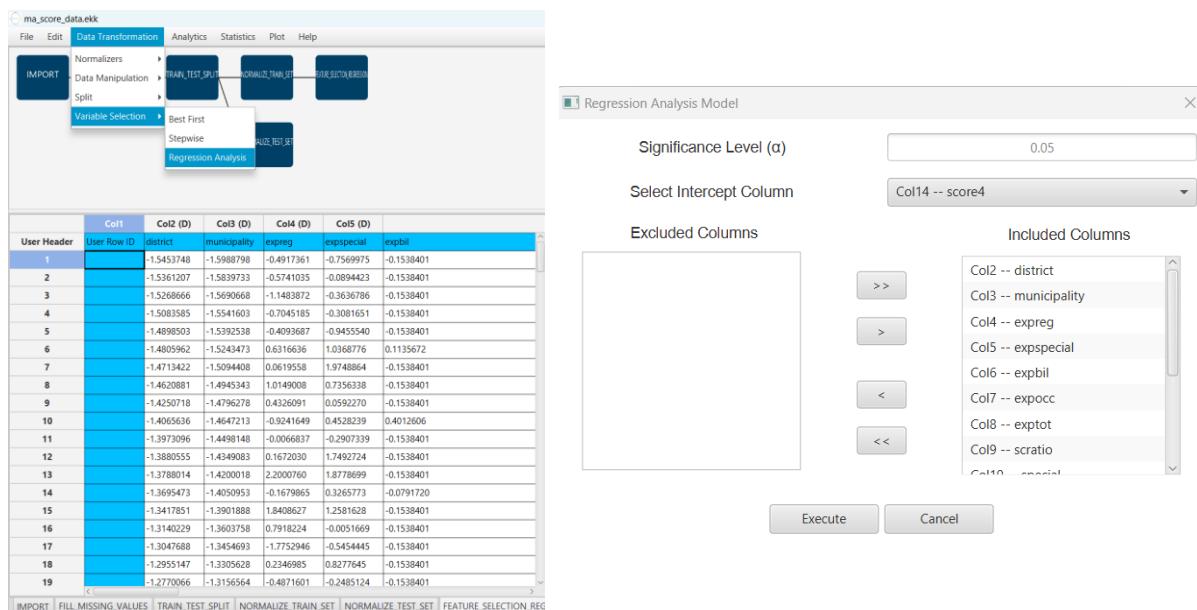
Step 7: Feature selection

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

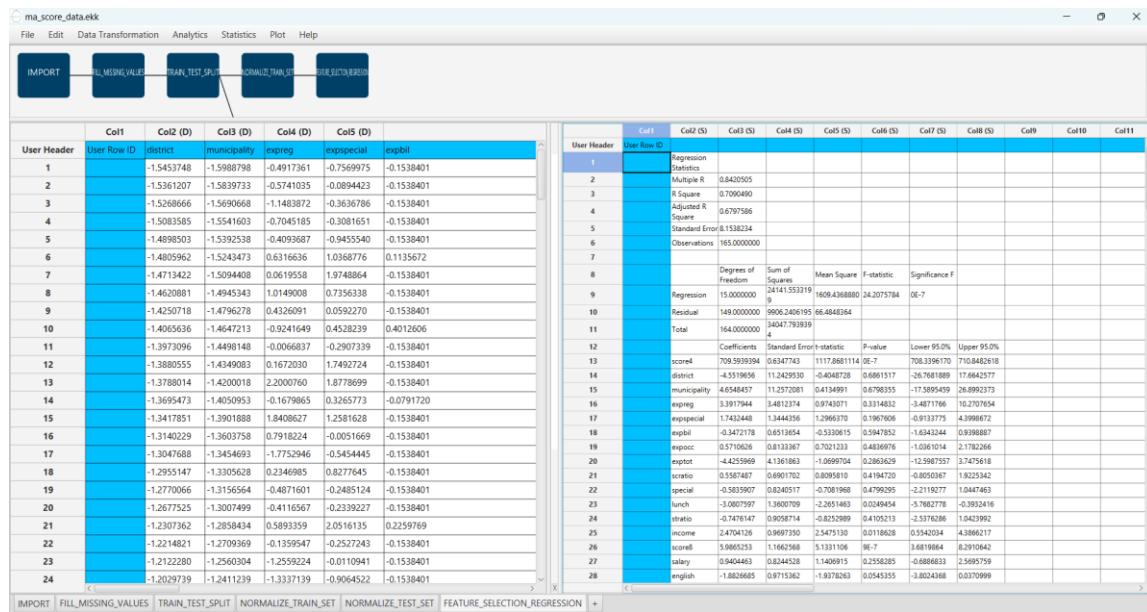
Import data into the input spreadsheet of the "FEATURE_SELECTION_REGRESSION" tab from the output of the "NORMALIZE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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.



The results will appear on the output spreadsheet.



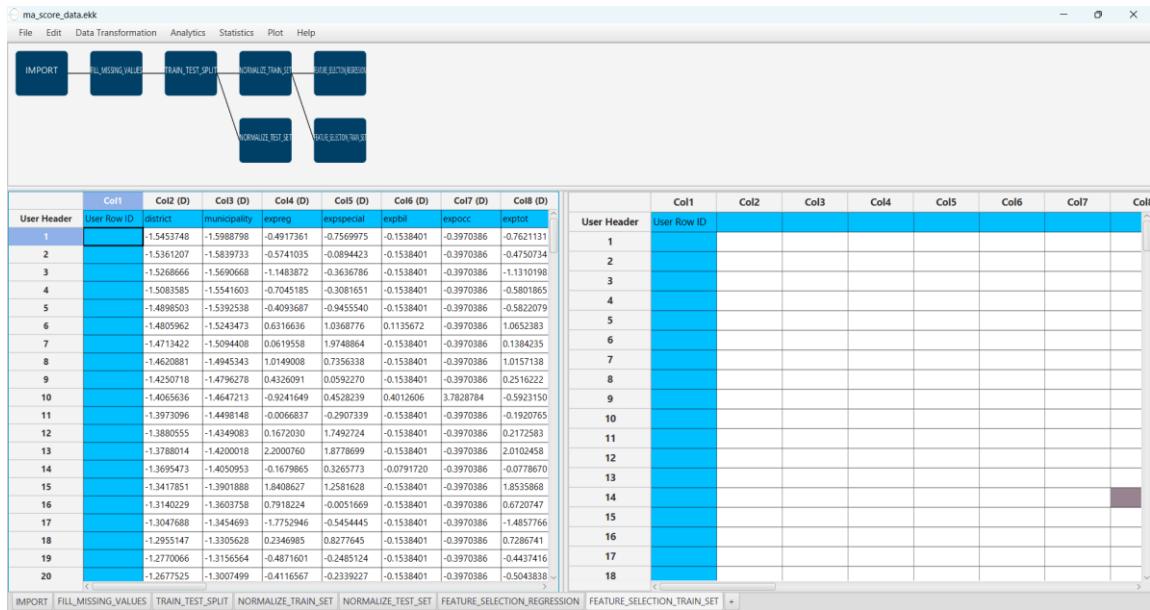
The significant features according to the p-value are the following:

- 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)

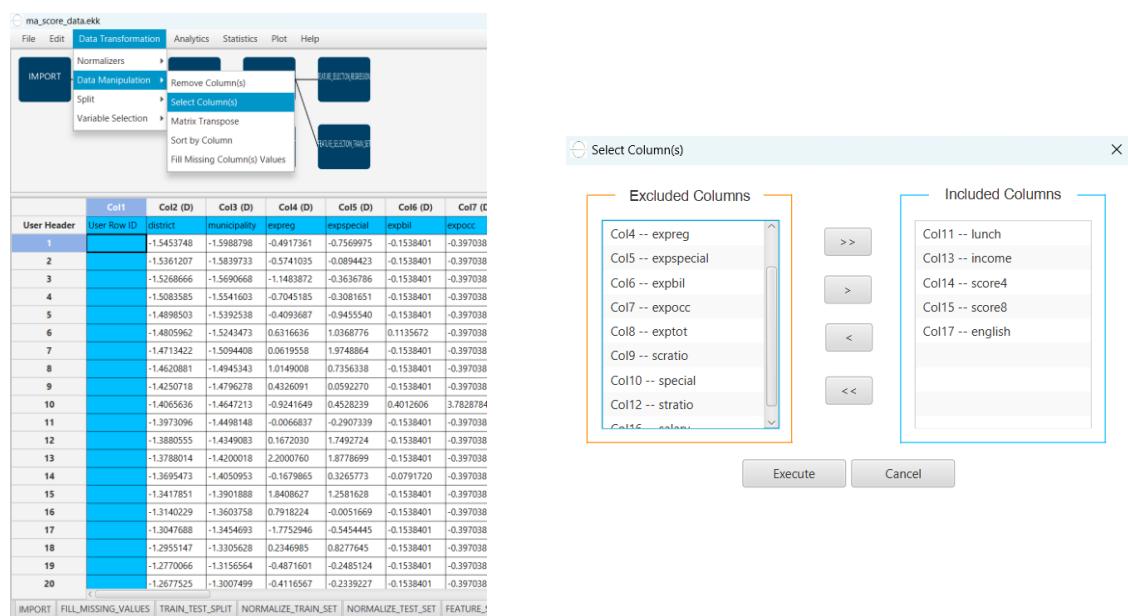
Step 8: Feature selection: train set

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.

The screenshot shows the Isalos Analytics Platform interface. At the top, there is a menu bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, Help, and a logo. Below the menu is a toolbar with buttons for IMPORT, FILL_MISSING_VALUES, TRAIN_TEST_SPLIT, NORMALIZE_TRAIN_SET, NORMALIZE_TEST_SET, FEATURE_SELECTION_REGRESSION, and FEATURE_SELECTION_TRAIN_SET.

The main area contains a data flow diagram with nodes: IMPORT, FILL_MISSING_VALUES, TRAIN_TEST_SPLIT, NORMALIZE_TRAIN_SET, and FEATURE_SELECTION_REGRESSION. The TRAIN_TEST_SPLIT node has two outputs: one to NORMALIZE_TRAIN_SET and one to NORMALIZE_TEST_SET. The NORMALIZE_TRAIN_SET node has two outputs: one to FEATURE_SELECTION_REGRESSION and one to FEATURE_SELECTION_TRAIN_SET.

Below the data flow diagram are two data tables:

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
1	-1.5453748	-1.5980796	-0.4917361	-0.7569975	-0.1538401	-0.3970386	-0.7621131	
2	-1.5361207	-1.5839733	-0.5741035	-0.0894423	-0.1538401	-0.3970386	-0.4750734	
3	-1.5266666	-1.5690668	-1.1483872	-0.3636786	-0.1538401	-0.3970386	-1.1310198	
4	-1.5083585	-1.5541603	-0.7045185	-0.3081651	-0.1538401	-0.3970386	-0.5801865	
5	-1.4898503	-1.5392538	-0.4093687	-0.9455540	-0.1538401	-0.3970386	-0.5822079	
6	-1.4805962	-1.5249473	0.6316636	1.0368776	-0.1538401	-0.3970386	1.0652383	
7	-1.4713422	-1.5094400	0.0619558	1.9748864	-0.1538401	-0.3970386	0.1384253	
8	-1.4620881	-1.4945343	1.0149008	0.7356338	-0.1538401	-0.3970386	1.0157138	
9	-1.4250718	-1.4796278	0.4326091	0.0592270	-0.1538401	-0.3970386	0.2516222	
10	-1.4065636	-1.4647213	-0.9241649	0.4528239	0.4012606	3.7828784	-0.5923150	
11	-1.3973096	-1.4498148	-0.0066837	-0.2970339	-0.1538401	-0.3970386	-0.1920765	
12	-1.3880555	-1.4349083	0.1672030	1.7492724	-0.1538401	-0.3970386	0.2172583	
13	-1.3788014	-1.4200018	2.2000760	1.8778999	-0.1538401	-0.3970386	2.0102458	
14	-1.3695473	-1.4050953	-0.1679865	0.3265773	-0.0791720	-0.3970386	-0.0778670	
15	-1.3417851	-1.3901888	1.8408627	1.2581628	-0.1538401	-0.3970386	1.8535868	
16	-1.3142029	-1.3603758	0.7918224	-0.0051669	-0.1538401	-0.3970386	0.6720747	
17	-1.3047688	-1.3454693	-0.7752946	-0.5454445	-0.1538401	-0.3970386	-1.4857766	
18	-1.2955147	-1.3305626	0.2346986	0.8277645	-0.1538401	-0.3970386	0.7286741	
19	-1.2770066	-1.3156564	-0.4871601	0.4871601	-0.1538401	-0.3970386	-0.4437416	
20	-1.2677525	-1.3007499	-0.4116567	-0.2339227	-0.1538401	-0.3970386	-0.5403838	

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
1	0.2610802	-0.3923772	714.0000000	-0.3884531	-0.4380666			
2	-0.8831728	1.3481476	731.0000000	0.0230312	0.1017883			
3	-0.1072294	-0.8248734	704.0000000	-0.2774707	-0.4380666			
4	-0.2410127	-0.4419321	704.0000000	-0.3884531	-0.2983138			
5	0.1135131	-0.5691478	701.0000000	0.0557285	-0.4380666			
6	0.7422949	-1.3603626	714.0000000	0.0230312	1.2608885			
7	-0.8296595	1.4470726	725.0000000	1.6658866	-0.4380666			
8	0.3012153	0.5450987	717.0000000	0.9440916	0.7328349			
9	0.4751336	0.6307104	702.0000000	0.3888646	-0.4380666			
10	0.3342557	-0.6529104	701.0000000	-0.5550212	-0.2755024			
11	-0.3466111	-0.1850970	713.0000000	0.2778193	-0.4380666			
12	0.2543911	-0.4299132	707.0000000	-0.8326347	-0.4380666			
13	0.5483090	-0.7239145	703.0000000	-0.9992028	-0.4380666			
14	0.2004724	-0.2080254	704.0000000	0.0002058	-0.0033341			
15	0.8564162	1.1258900	721.0000000	1.5548413	-0.4380666			
16	0.6156061	1.5332390	728.0000000	1.4993186	0.5037455			
17	-0.6758086	-0.7141145	710.0000000	-0.0553169	-0.4380666			
18	0.5487145	0.1140817	731.0000000	0.0230312	-0.4380666			

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".

The screenshot shows the Isalos Analytics Platform interface. At the top, there is a menu bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, Help, and a logo. Below the menu is a toolbar with buttons for IMPORT, FILL_MISSING_VALUES, TRAIN_TEST_SPLIT, NORMALIZE_TRAIN_SET, NORMALIZE_TEST_SET, FEATURE_SELECTION_REGRESSION, and FEATURE_SELECTION_TRAIN_SET.

The main area contains a data flow diagram with nodes: IMPORT, FILL_MISSING_VALUES, TRAIN_TEST_SPLIT, NORMALIZE_TRAIN_SET, and FEATURE_SELECTION_REGRESSION. The TRAIN_TEST_SPLIT node has two outputs: one to NORMALIZE_TRAIN_SET and one to NORMALIZE_TEST_SET. The NORMALIZE_TRAIN_SET node has two outputs: one to FEATURE_SELECTION_REGRESSION and one to FEATURE_SELECTION_TRAIN_SET.

Below the data flow diagram are two data tables:

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
1	-1.3232769	-1.3752823	-0.4311045	-0.6790780	0.0726519	0.1363825	-0.3558104	0.498794
2	-1.0826709	-1.0771524	0.2301226	1.5124335	-0.1538401	-0.3970386	0.5174372	0.049359
3	-0.9901302	-1.0026199	-0.1205688	3.9029579	0.1764840	1.3465590	-0.0435031	0.290223
4	-0.9716220	-0.9877134	-0.2103142	0.0423915	-0.1538401	-0.3970386	-0.1112202	-0.926431
5	-0.9531139	-0.9579009	1.0492205	0.7099301	-0.1538401	-0.3970386	0.7822414	-0.37024
6	-0.9346057	-0.9429940	1.1786600	2.1220015	-0.1538401	-0.3970386	1.2077474	-0.23120
7	-0.8698272	-0.8684615	-1.0797478	0.4955862	-0.1538401	-0.3970386	-0.7307813	-0.05739
8	-0.8420649	-0.8535550	-0.7457022	-1.2137015	-0.1538401	-0.3970386	-0.8571192	0.185938
9	-0.7495242	-0.7790225	-0.5821111	-0.6618195	-0.1538401	-0.3970386	0.6054541	0.498794
10	-0.7402701	-0.7343030	-0.8841252	-0.7033281	-0.1538401	-0.3970386	-0.8945152	0.116414
11	0.6662375	0.6448640	0.0585238	1.0470550	-0.1538401	-0.3970386	0.2699006	1.402600
12	-0.6199671	-0.5852381	-1.0728838	-0.9295542	-0.1538401	-0.0415720	-1.2866680	-0.96120
13	-0.6014589	-0.5554251	-0.8738293	-1.3676798	-0.1538401	-0.3970386	-1.0289387	-0.05739
14	-0.5181722	-0.4808926	-1.4080733	-1.0666030	-0.1538401	-0.3970386	-1.1017094	-0.49501
15	-0.4996641	-0.4659861	-1.0888997	-0.5466422	-0.1538401	-0.3970386	-0.7186529	-1.13500
16	-0.3423448	-0.3616406	0.4268891	0.3192182	-0.1538401	-0.3970386	0.3793987	0.881173
17	-0.3052825	-0.3318276	-1.9743491	-1.2045874	-0.1538401	-0.3970386	-1.9557536	1.889265
18	-0.2868203	-0.3020146	1.6223603	-0.3227552	0.0744103	1.5831512	1.2077474	0.012129
19	-0.2683122	-0.2722016	-0.3464492	-0.5198240	-0.1538401	-0.3970386	-0.5256085	0.049359
20	-0.0924847	-0.0486042	0.5961999	0.9806120	-0.1538401	-0.3970386	0.4163668	0.220700

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)
1	0.2301250	-0.7239145	731.0000000	0.0230312	-0.4380666		
2	0.8831728	1.3481476	731.0000000	0.0230312	0.1017883		
3	-0.1072294	-0.8248734	704.0000000	-0.2774707	-0.4380666		
4	-0.2410127	-0.4419321	704.0000000	-0.3884531	-0.2983138		
5	0.1135131	-0.5691478	701.0000000	0.0557285	-0.4380666		
6	0.7422949	-1.3603626	714.0000000	0.0230312	1.2608885		
7	-0.8296595	1.4470726	725.0000000	1.6658866	-0.4380666		
8	0.3012153	0.5450987	717.0000000	0.9440916	0.7328349		
9	0.4751336	0.6307104	702.0000000	0.3888646	-0.4380666		
10	0.3342557	-0.6529104	701.0000000	-0.5550212	-0.2755024		
11	-0.3466111	-0.1850970	713.0000000	0.2778193	-0.4380666		
12	0.2543911	-0.4299132	707.0000000	-0.8326347	-0.4380666		
13	0.5483090	-0.7239145	703.0000000	-0.9992028	-0.4380666		
14	0.2004724	-0.2080254	704.0000000	0.0002058	-0.0033341		
15	0.8564162	1.1258900	721.0000000	1.5548413	-0.4380666		
16	0.6156061	1.5332390	728.0000000	1.4993186	0.5037455		
17	-0.6758086	-0.7141145	710.0000000	-0.0553169	-0.4380666		
18	0.5487145	0.1140817	731.0000000	0.0230312	-0.4380666		

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 screenshot shows the Isalos Analytics Platform interface. The top navigation bar includes File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help. The Data Transformation menu is open, showing sub-options like Import, Data Manipulation, Split, and Variable Selection. The 'Data Manipulation' option is selected, and its submenu includes Remove Column(s), Select Column(s), Matrix Transpose, Sort by Column, and Fill Missing Column(s) Values. A callout box highlights the 'FEATURE_SELECTION_REGRESSION' node in the process flow below the menu.

The main workspace displays a spreadsheet with columns: Col1, Col2 (D), Col3 (D), Col4 (D), Col5 (D), Col6 (D), and Col7 (D). The rows are numbered 1 to 20, with various numerical values. Below the spreadsheet are several buttons: IMPORT, FILL_MISSING_VALUES, TRAIN_TEST_SPLIT, NORMALIZE_TRAIN_SET, NORMALIZE_TEST_SET, and FEATU.

A separate window titled 'Select Column(s)' is shown, containing two lists: 'Excluded Columns' (which includes Col4--expreg, Col5--expspecial, Col6--expbil, Col7--expocc, Col8--exptot, Col9--scratio, Col10--special, Col12--stratio, and Col13--income) and 'Included Columns' (which includes Col11--lunch, Col13--income, Col14--score4, Col15--score8, and Col17--english). At the bottom are Execute and Cancel buttons.

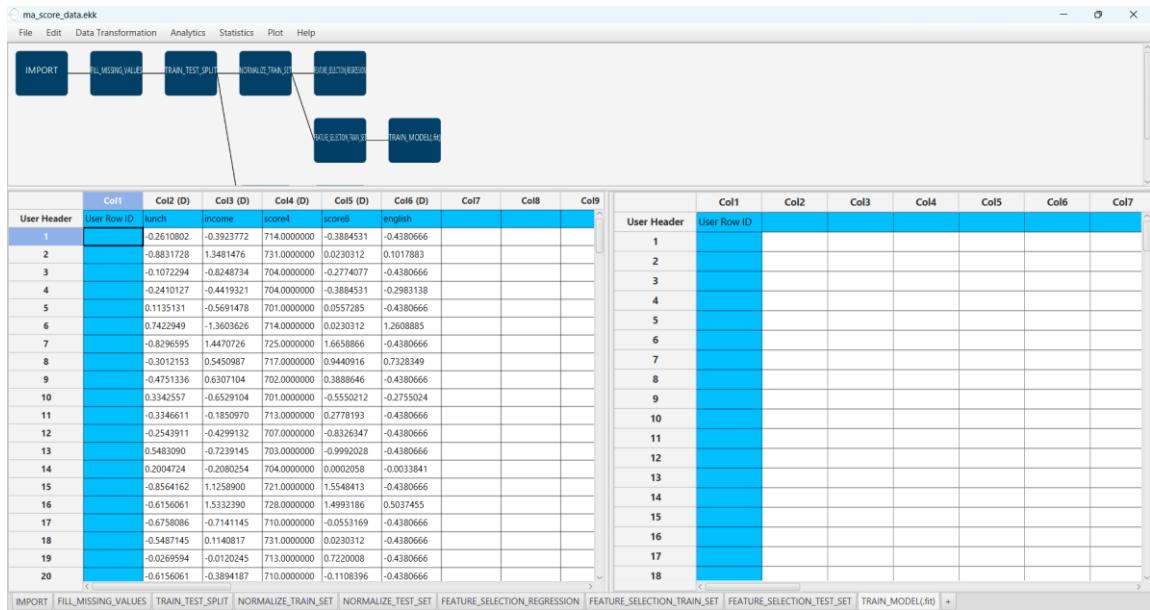
The results will appear on the output spreadsheet.

This screenshot shows the Isalos Analytics Platform interface after the 'FEATURE_SELECTION_REGRESSION' step has been executed. The top navigation bar and Data Transformation menu are identical to the previous screenshot. The process flow now includes the 'FILL_MISSING_VALUES' node, which is connected to the 'FEATURE_SELECTION_REGRESSION' node. The resulting spreadsheet is identical to the one in the previous screenshot, with columns: Col1, Col2 (D), Col3 (D), Col4 (D), Col5 (D), Col6 (D), Col7 (D), Col8 (D), and Col9 (D). The rows are numbered 1 to 20, with values corresponding to the input data.

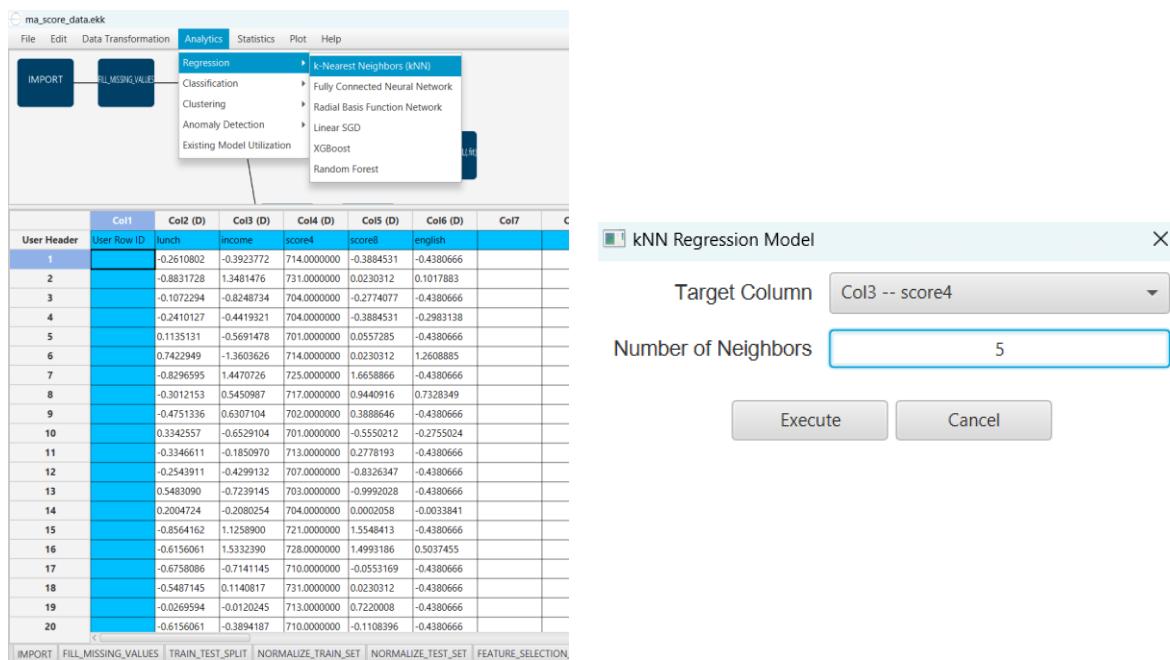
Step 10: Train the model

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

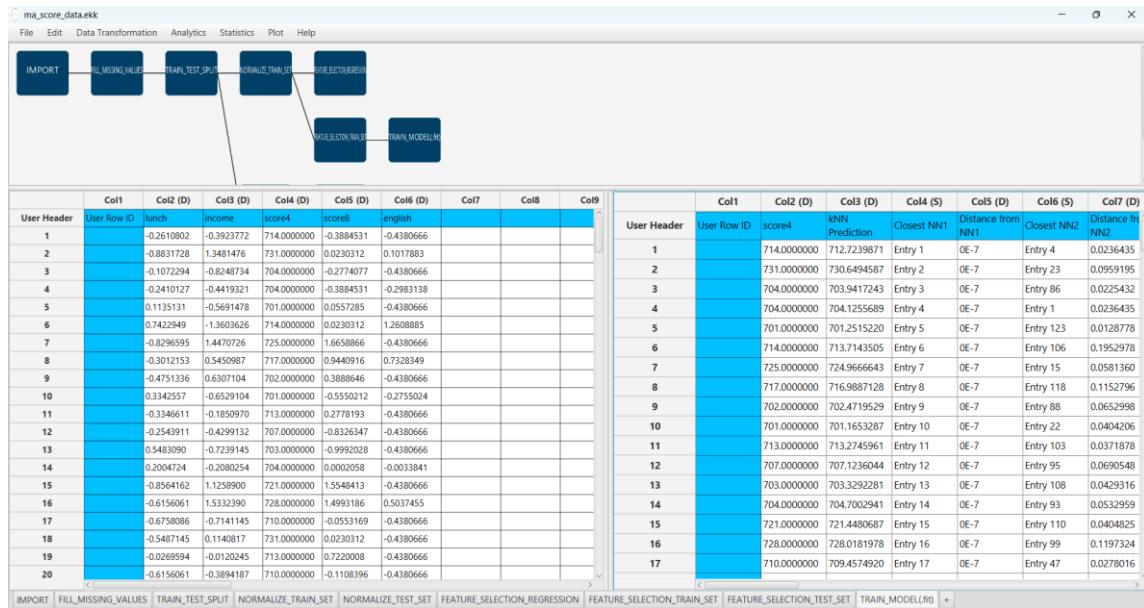
Import data into the input spreadsheet of the "TRAIN_MODEL(.fit)" tab from the output of the "FEATURE_SELECTION_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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.



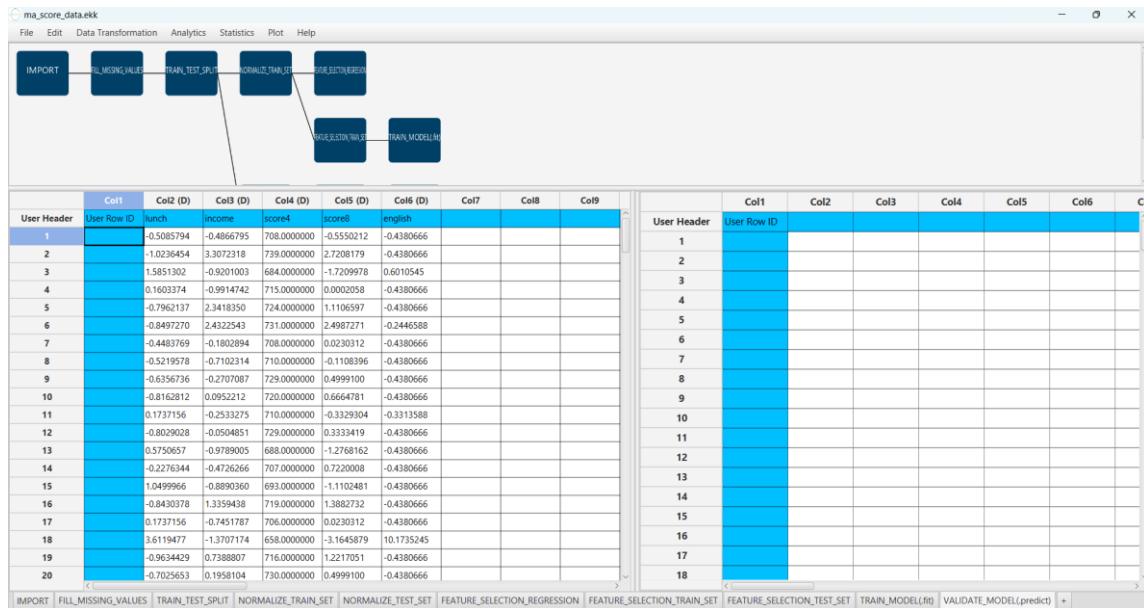
The predictions will appear on the output spreadsheet.



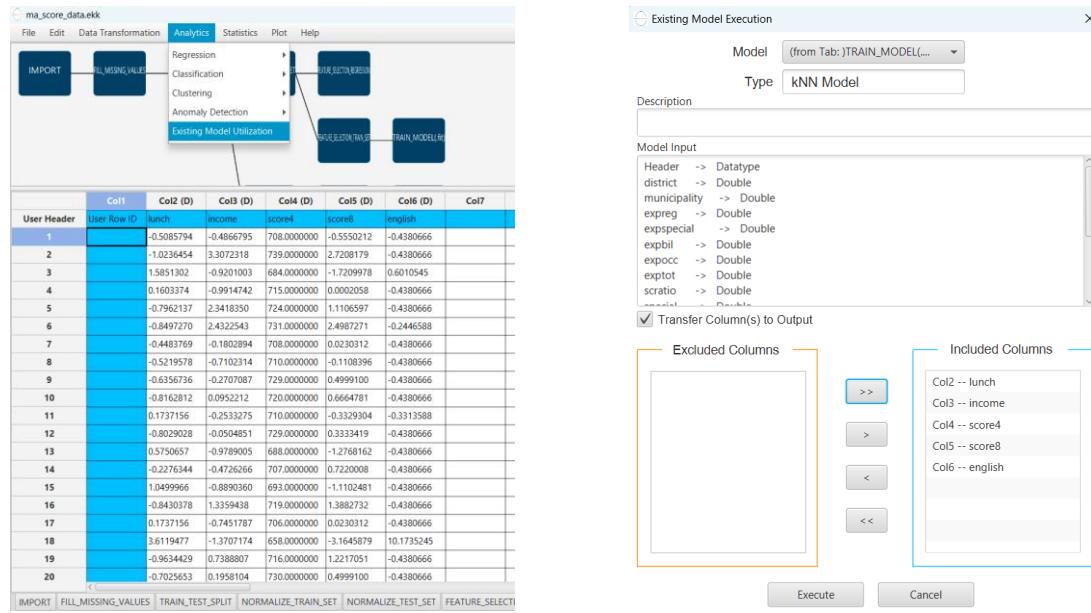
Step 11: Validate the model

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

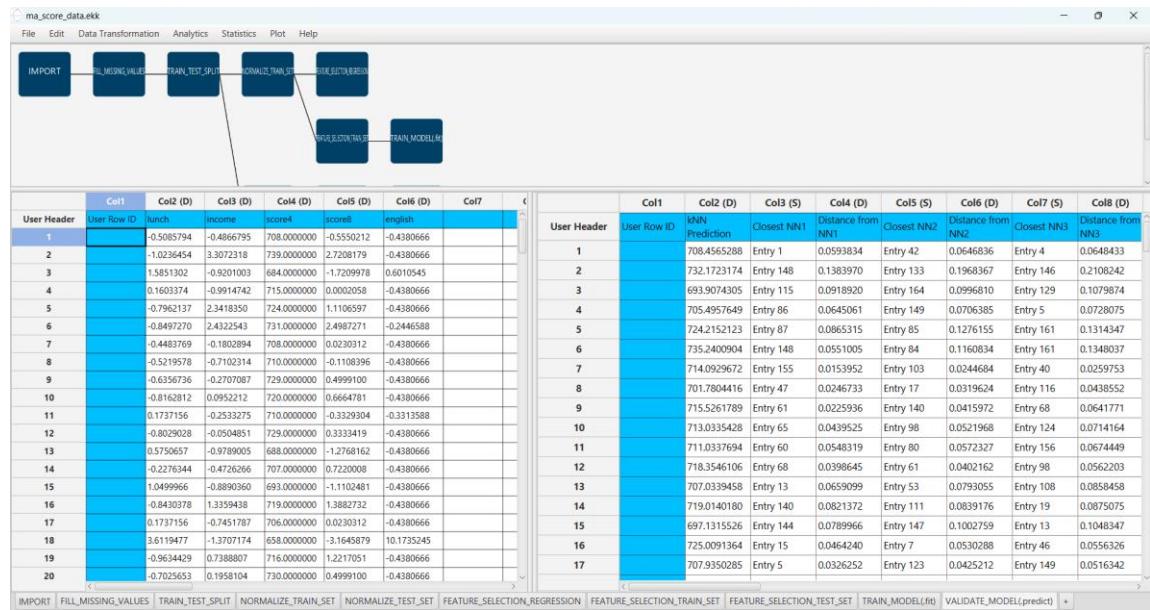
Import data into the input spreadsheet of the "VALIDATE_MODEL(.predict)" tab from the output of the "FEATURE_SELECTION_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



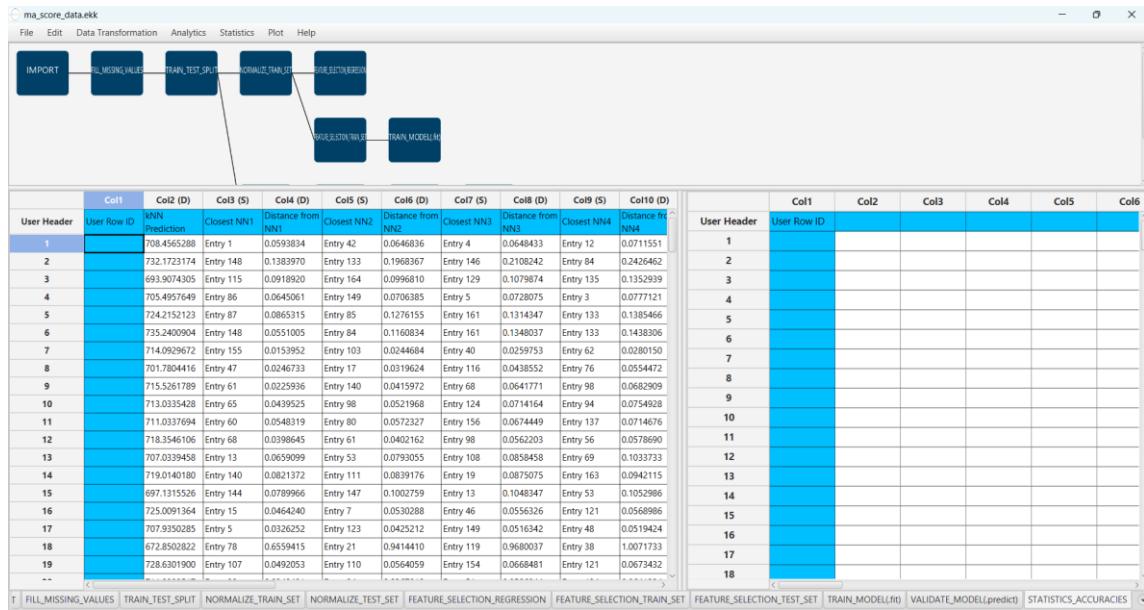
The predictions will appear on the output spreadsheet.



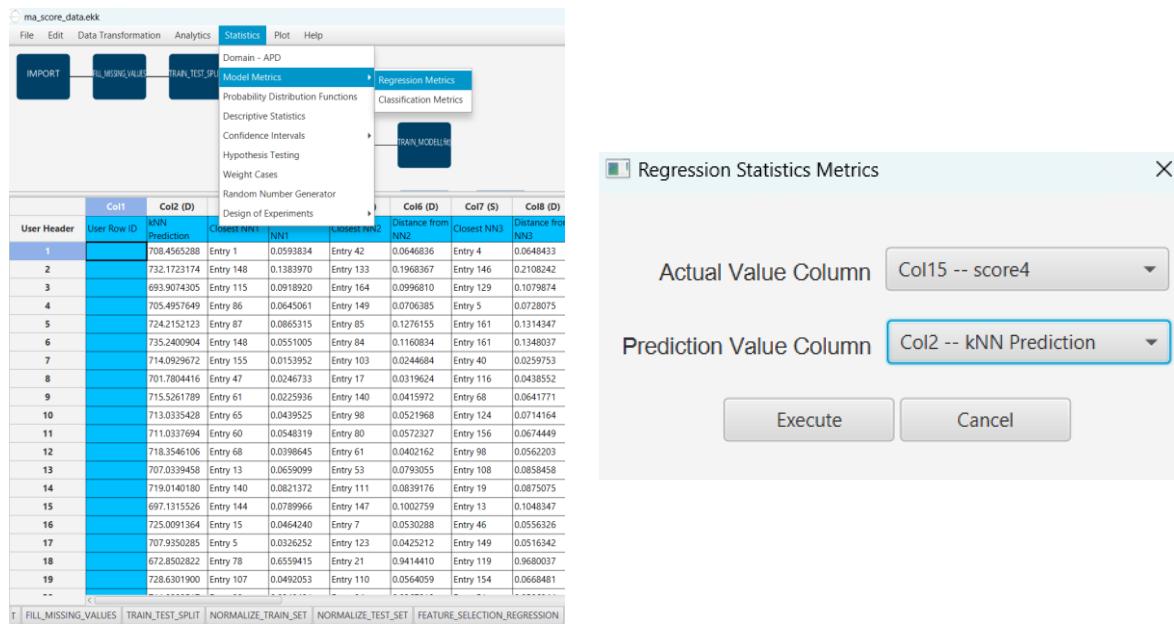
Step 12: Statistics calculation

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

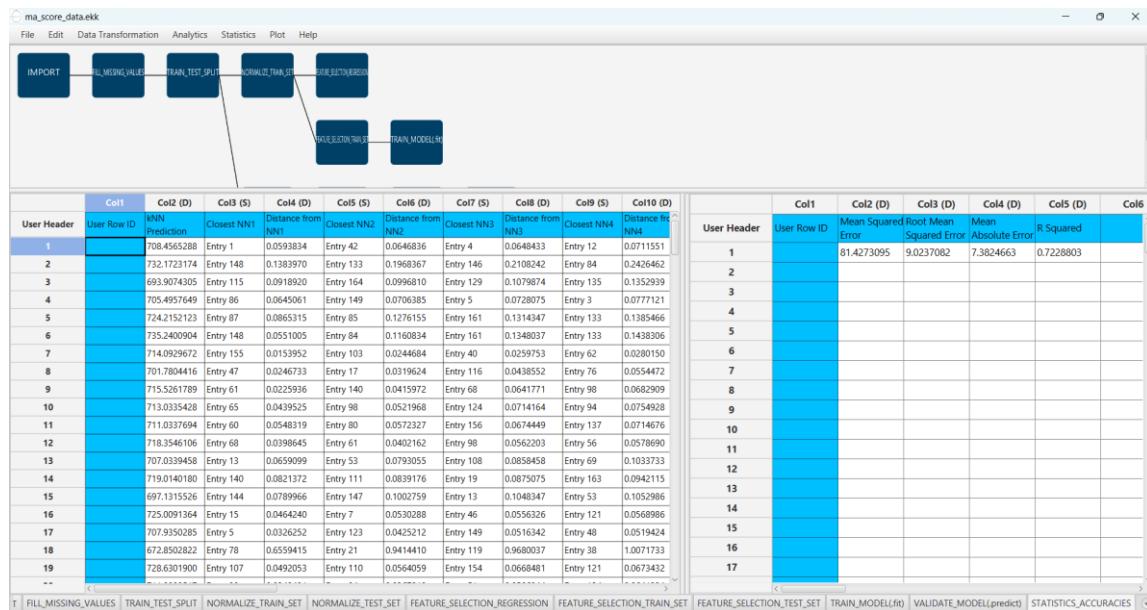
Import data into the input spreadsheet of the "STATISTICS_ACCURACIES" tab from the output of the "VALIDATE_MODEL(.predict)" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



The results will appear on the output spreadsheet.

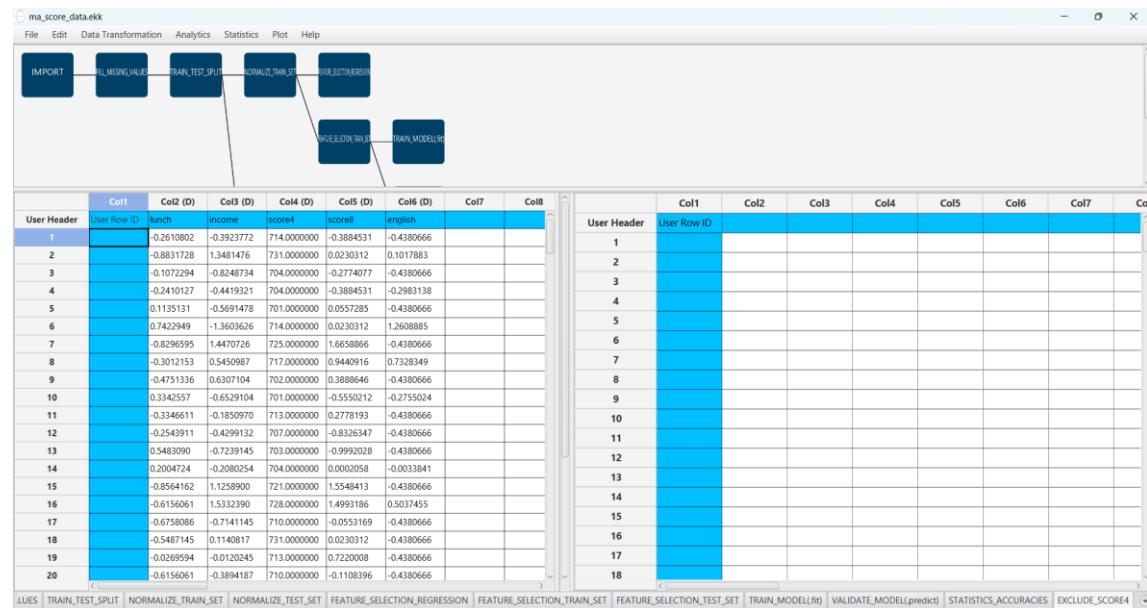


Step 13: Reliability check of each record of the test set

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 Column(s)". Then select all the columns except the "score4".

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7	Col8
1		lunch	income	score4	score8	english		
2	-0.2610802	-0.3923772	714.0000000	0.3884531	-0.4380666			
3	-0.8831728	1.3481476	731.0000000	0.0230312	0.1017883			
4	-0.1072294	-0.8248734	704.0000000	-0.2774077	-0.4380666			
5	-0.2410127	-0.4419321	704.0000000	-0.3884531	-0.2983138			
6	0.1135131	-0.5691478	701.0000000	0.0557285	-0.4380666			
7	0.7422949	-1.3603626	714.0000000	0.0230312	1.260885			
8	-0.8296595	1.4470726	725.0000000	1.6658866	-0.4380666			
9	-0.3012153	0.5450987	717.0000000	0.9440916	0.7328349			
10	-0.4751336	0.6307104	702.0000000	0.3888646	-0.4380666			
11	0.3342557	-0.6529108	701.0000000	-0.5550212	-0.2755024			
12	-0.3466611	-0.1850970	713.0000000	0.2778193	-0.4380666			
13	-0.2543911	-0.4299132	707.0000000	-0.8326347	-0.4380666			
14	0.5483090	-0.7239145	703.0000000	-0.9992028	-0.4380666			
15	0.2004724	-0.2080254	704.0000000	0.0002058	-0.0033841			
16	-0.8564162	1.1258900	721.0000000	1.5540413	-0.4380666			
17	-0.6156061	1.5332390	728.0000000	1.4993186	0.5037455			
18	-0.6758086	-0.7141145	710.0000000	-0.0553169	-0.4380666			
19	-0.5487145	0.1140817	731.0000000	0.0230312	-0.4380666			
20	-0.0269594	-0.0120245	713.0000000	0.7220008	-0.4380666			

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".

User Header	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
1		lunch	income	score4	score8	english		
2	-0.2610802	-0.3923772	-0.3884531	-0.4380666				
3	-0.8831728	1.3481476	0.0230312	0.1017883				
4	-0.1072294	-0.8248734	-0.2774077	-0.4380666				
5	-0.2410127	-0.4419321	-0.3884531	-0.2983138				
6	0.1135131	-0.5691478	0.0557285	-0.4380666				
7	0.7422949	-1.3603626	0.0230312	1.260885				
8	-0.8296595	1.4470726	1.6658866	-0.4380666				
9	-0.3012153	0.5450987	0.9440916	0.7328349				
10	-0.4751336	0.6307104	0.3888646	-0.4380666				
11	0.3342557	-0.6529104	-0.5550212	-0.2755024				
12	-0.3466611	-0.1850970	0.2778193	-0.4380666				
13	-0.2543911	-0.4299132	-0.8326347	-0.4380666				
14	0.5483090	-0.7239145	-0.9992028	-0.4380666				
15	0.2004724	-0.2080254	0.0002058	-0.0033841				
16	-0.8564162	1.1258900	1.5540413	-0.4380666				
17	-0.6156061	1.5332390	1.4993186	0.5037455				
18	-0.6758086	-0.7141145	-0.0553169	-0.4380666				
19	-0.5487145	0.1140817	0.0230312	-0.4380666				
20	-0.0269594	-0.0120245	0.7220008	-0.4380666				

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

The screenshot shows the Isalos Analytics Platform interface. On the left, a data flow diagram for 'ma_score_data.ekk' is displayed, featuring nodes like 'IMPORT', 'ML_MISSING_VALUES', 'TRAIN_TEST_SPLIT', 'NORMALIZE_TRAIN_SET', 'FEATURE_SELECTION_REGRESSION', and 'TRAIN_MODEL'. A context menu is open over the 'TRAIN_MODEL' node, showing options such as 'Domain - APD', 'Model Metrics', 'Probability Distribution Functions', 'Descriptive Statistics', 'Confidence Intervals', 'Hypothesis Testing', 'Weight Cases', and 'Random Number Generator'. On the right, a detailed dialog box titled 'Domain - APD' is shown, containing the formula $APD = d + Z\sigma$ and a slider set to 0.5. It also includes a 'Perform Computations' button, a 'CPU (double precision)' dropdown, and 'Execute' and 'Cancel' buttons.

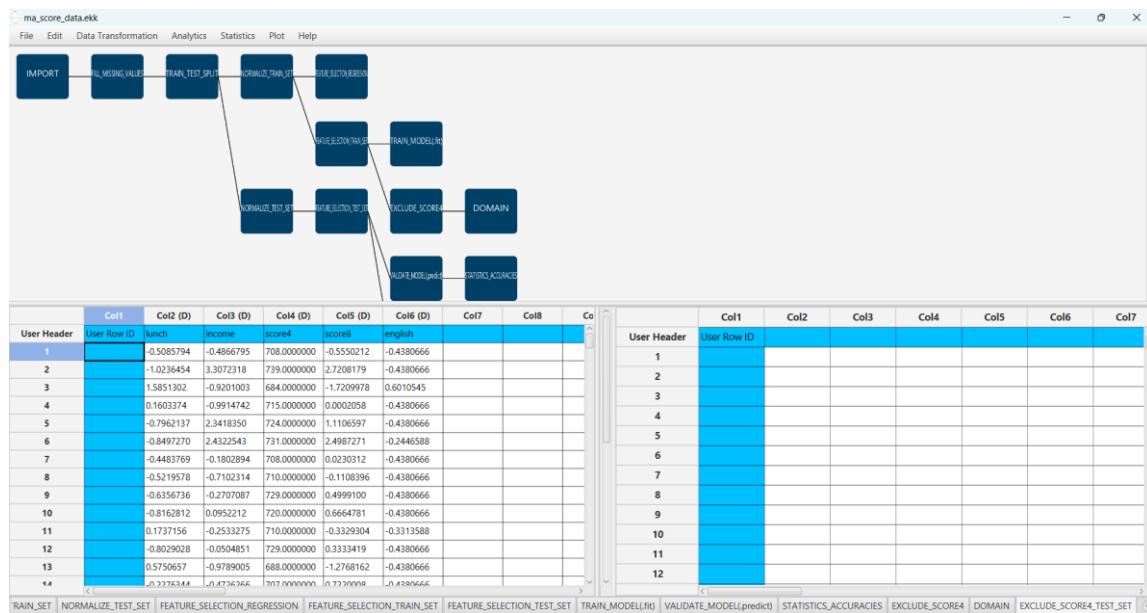
The results will appear on the output spreadsheet.

This screenshot shows the Isalos Analytics Platform interface again. The data flow diagram on the left is identical to the previous one. The output spreadsheet on the right has two tabs: 'FEATURE_SELECTION_TRAIN_SET' and 'FEATURE_SELECTION_TEST_SET'. The 'FEATURE_SELECTION_TRAIN_SET' tab contains columns for User Header, User Row ID, Domain, APD, and Prediction. The 'FEATURE_SELECTION_TEST_SET' tab contains columns for User Header, User Row ID, Col1 through Col8. Both tabs show the same data rows as the input spreadsheet, with the addition of the 'APD' and 'Prediction' columns in the first tab.

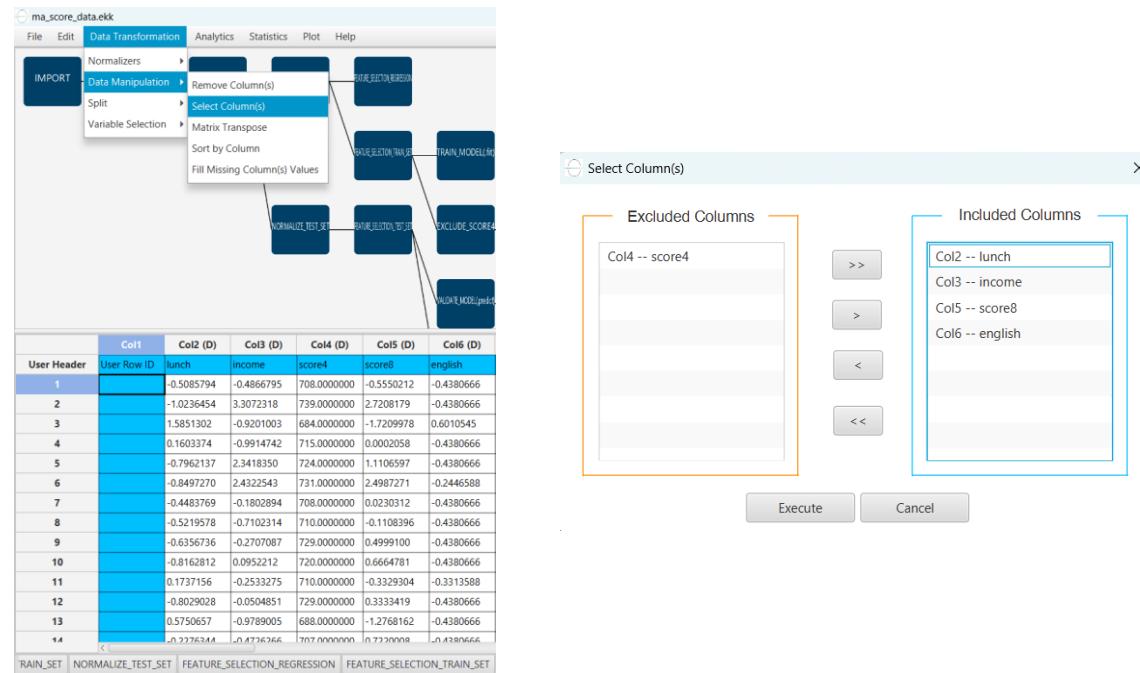
Step 12.b: Check the test set reliability

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".



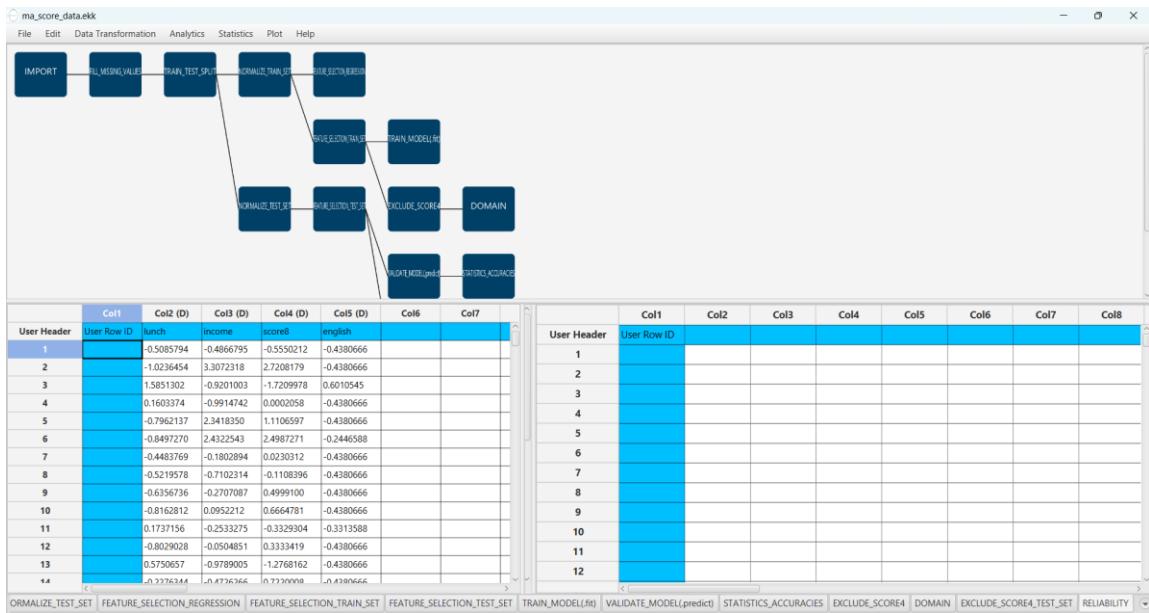
Filter the data to exclude the column that corresponds to the "score4" by browsing: "Data Transformation" → "Data Manipulation" → "Select Column(s)". Then select all the columns except "score4".



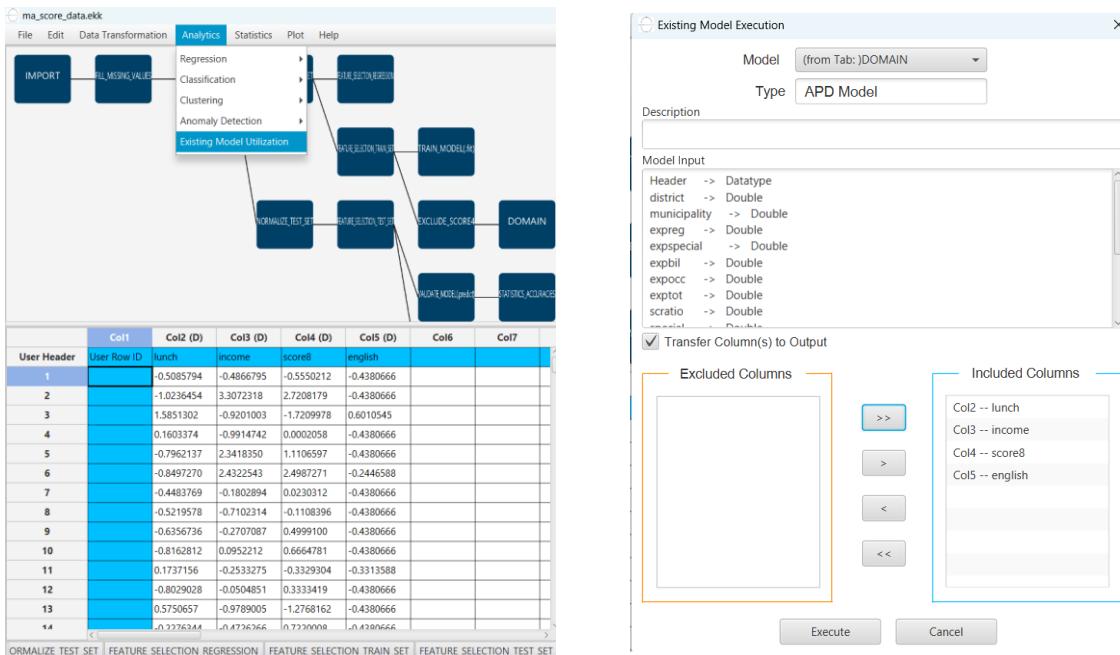
The results will appear on the output spreadsheet.

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

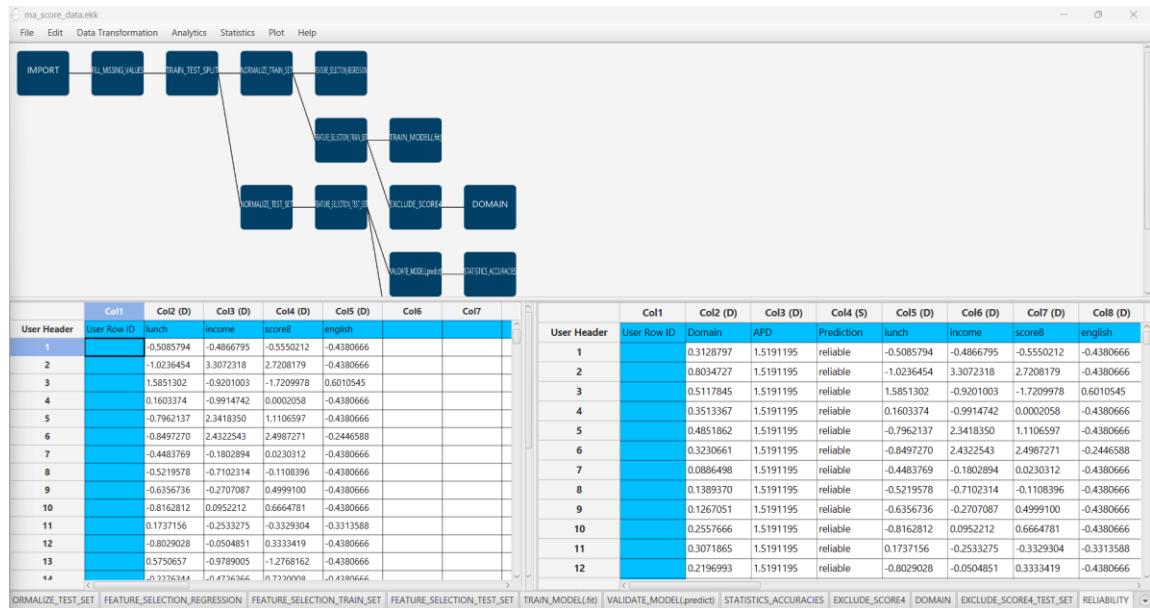
Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "EXCLUDE_SCORE4_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



The results will appear on the output spreadsheet.



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:

