



# 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.

The screenshot shows a data manipulation interface. On the left, there is a table with columns labeled Col1 through Col8. A context menu is open over the 6th row of this table, containing options: 'Import from SpreadSheet', 'Import from file', 'Export Spread Sheet Data', and 'Clear Spreadsheet'. Below this table is another table with columns labeled Col1 through Col8, showing data rows from 1 to 21. The first column is labeled 'User Header' and the second column is labeled 'User Row ID'. The interface includes a top navigation bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help.

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

## 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 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, with 'Data Manipulation' selected. A sub-menu 'Select Column(s)' is highlighted. To the right, a modal dialog titled 'Select Column(s)' displays a list of columns: Col2 -- age, Col3 -- sex, Col4 -- bmi, Col5 -- children, Col6 -- smoker, Col7 -- region, and Col8 -- charges. The 'Included Columns' section (blue border) contains Col2 -- age, Col3 -- sex, Col4 -- bmi, Col5 -- children, Col6 -- smoker, Col7 -- region, and Col8 -- charges. The 'Excluded Columns' section (orange border) is empty. Buttons for >>, >, <, and << are available for moving columns between the two lists. Below the dialog are Execute and Cancel buttons.

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".

The screenshot shows the Isalos Analytics Platform interface with two tabs: 'IMPORT' and 'TRAIN\_TEST\_SPLIT'. The 'IMPORT' tab contains the original dataset with columns: Col1, Col2 (I), Col3 (I), Col4 (D), Col5 (I), Col6 (I), Col7 (I), Col8 (D), and Col9. The 'TRAIN\_TEST\_SPLIT' tab is empty, showing only the column headers: Col1, Col2, Col3, Col4, Col5, Col6, Col7, and Col8.

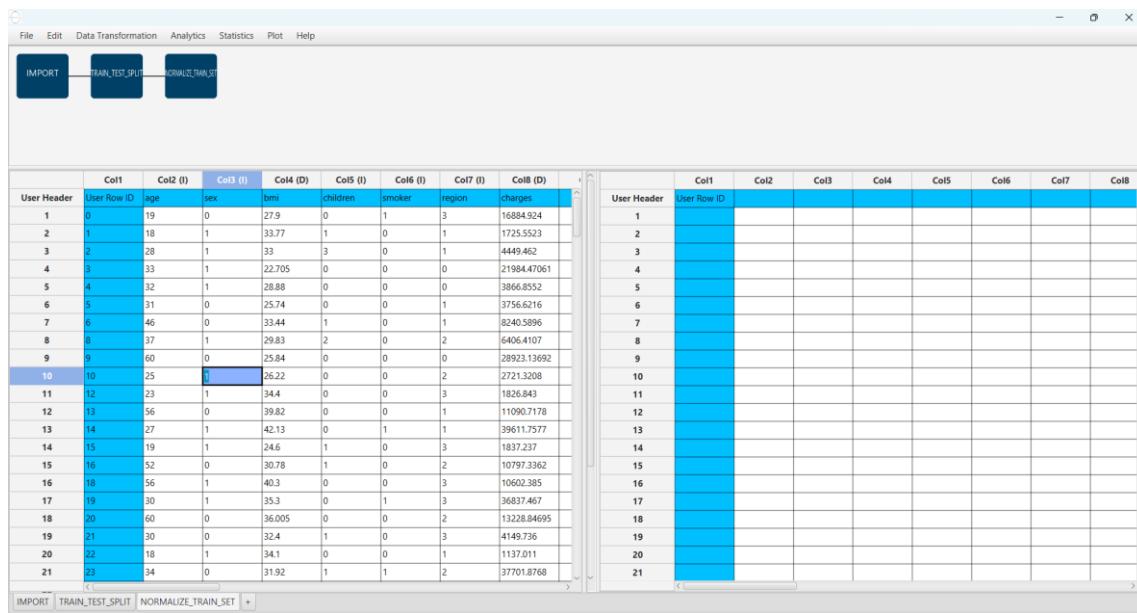
Split the dataset by choosing by browsing: "Data Transformation" → "Split" → "Random Partitioning". Then choose the "Training set percentage" and the column for the sampling as shown below:

The results will appear on the output spreadsheet.

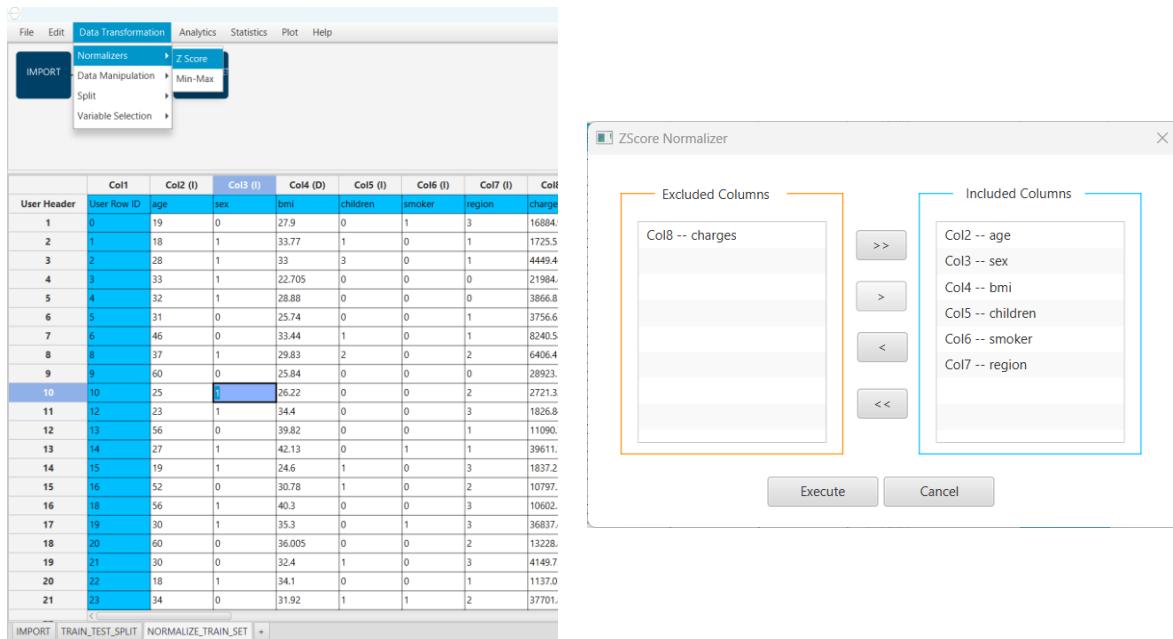
## Step 4: Normalize the training set

Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALIZE\_TRAIN\_SET".

Import data into the input spreadsheet of the "NORMALIZE\_TRAIN\_SET" tab the train set from the output of the "TRAIN\_TEST\_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet". From the available Select input tab options choose "TRAIN\_TEST\_SPLIT : Training Set"



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



The results will appear on the output spreadsheet.

The screenshot shows the Isalos Analytics Platform interface. At the top, there's a menu bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, Help, and a search bar. Below the menu is a toolbar with icons for IMPORT, TRAIN\_TEST\_SPLIT, and NORMALIZE\_TRAIN\_SET.

The main area contains a data flow diagram with three nodes: IMPORT, TRAIN\_TEST\_SPLIT, and NORMALIZE\_TRAIN\_SET. The TRAIN\_TEST\_SPLIT node has a connection from IMPORT and a connection to NORMALIZE\_TRAIN\_SET.

Below the diagram are two data tables:

	Col1	Col2 (I)	Col3 (I)	Col4 (D)	Col5 (I)	Col6 (I)	Col7 (I)	Col8 (D)
User Header	User Row ID	age	sex	bmi	children	smoker	region	charges
1	0	19	0	27.9	0	1	3	16684.924
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	1	29.83	2	0	2	6406.4107
9	8	60	0	25.84	0	0	0	28932.13692
10	9	25		26.22	0	0	2	2721.3206
11	10	23	1	34.4	0	0	3	1826.843
12	11	56	0	39.82	0	0	1	11090.7178
13	12	27	1	42.13	0	1	1	39611.7577
14	13	19	1	24.6	1	0	3	1837.237
15	14	52	0	30.78	1	0	2	10797.3362
16	15	56	1	40.3	0	0	3	10602.385
17	16	30	1	35.3	0	1	3	36837.467
18	17	60	0	36.005	0	0	2	13228.84695
19	18	30	0	32.4	1	0	3	4149.736
20	19	18	1	34.1	0	0	1	1137.011
21	20	34	0	31.92	1	1	2	37701.8768

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
User Header	User Row ID	age	sex	bmi	children	smoker	region	charges
1	0	-1.442052194	-1.013537686	-0.4480337111	-0.894080423	2.0358255719	1.3698195197	16684.924
2	1	-1.513797506	0.9856604221	0.49191124	0.403234	795726	160479	
3	2	-0.796344383	0.9856604221	0.3829151975	1.5720236018	-0.490711973	-0.438594415	1725.5523
4	3	-0.437617821	0.9856604221	-1.295073532	-0.894080423	-0.490711973	-1.342801383	
5	4	42380545	549418	7306575	2914	46850026	58298184	4449.462
6	5	-0.509363133	0.9856604221	-0.288060275	-0.894080423	-0.490711973	-1.342801383	21984.4706
7	6	7539002	549418	1963338	5403234	46850026	2324968	3866.8552
8	7	-0.581108446	0.9856604221	-0.800396912	-0.894080423	-0.490711973	-0.438594415	3756.6216
9	8	083995	61993	9782544	5403234	46850026	58298184	
10	9	674268	61993	610241	75050219	46850026	58298184	8240.5896
11	10	-0.150636572	0.9856604221	-0.133765159	0.7499862600	-0.490711973	0.4656125520	
12	11	1034263	549418	2379103	39311	46850026	6653307	6406.4107
13	12	1.4995056114	-1.013537686	-0.784097848	-0.894080423	-0.490711973	-1.342801383	28923.1369
14	13	887537	61993	0946271	5403234	46850026	2324968	
15	14	10	-1.013537686	0.9856604221	-0.894080423	-0.490711973	0.4656125520	2721.3208
16	15	0646337	61993	156847	5403234	46850026	6653307	
17	16	115670944	0.9856604221	0.0111021659	-0.894080423	-0.490711973	1.3698195197	1826.843
18	17	7247534	549418	436434	5403234	46850026	160479	
19	18	1.212543621	-1.013537686	1.4945114226	-0.894080423	-0.490711973	-0.438594415	11090.7178
20	19	683748	61993	36427	5403234	46850026	58298184	
21	20	4043742	549418	482124	5403234	795726	58298184	39611.7577
22	21	-1.442052194	0.9856604221	-0.896206252	-0.072051081	-0.490711973	1.3698195197	1837.237
23	22	0451325	549418	6516017	75050219	46850026	160479	
24	23	0.925543128	-1.013537686	0.0210759571	-0.072051081	-0.490711973	0.4656125520	10797.3362

## 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".

The screenshot shows the Isalos Analytics Platform interface. At the top, there's a menu bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, Help, and a search bar. Below the menu is a toolbar with icons for IMPORT, TRAIN\_TEST\_SPLIT, and NORMALIZE\_TRAIN\_SET.

The main area contains a data flow diagram with three nodes: IMPORT, TRAIN\_TEST\_SPLIT, and NORMALIZE\_TRAIN\_SET. The TRAIN\_TEST\_SPLIT node has a connection from IMPORT and a connection to NORMALIZE\_TRAIN\_SET. A new tab labeled "NORMALIZE\_TEST\_SET" is visible on the right side of the interface.

Below the diagram are two data tables:

	Col1	Col2 (I)	Col3 (I)	Col4 (D)	Col5 (I)	Col6 (I)	Col7 (I)	Col8 (D)	Col9
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	15158.18025	
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	Col3	Col4	Col5	Col6	Col7
User Header	User Row ID						
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							

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

The screenshot shows the Isalos Analytics Platform interface. On the left, a data pipeline is displayed with nodes: IMPORT, TRAIN\_TEST\_SPLIT, and NORMALIZE\_TRAIN\_SET. The TRAIN\_TEST\_SPLIT node has a dropdown menu open under the Analytics tab, showing options like Regression, Classification, Clustering, and Anomaly Detection. Below the pipeline is a data spreadsheet with columns Col1 through Col9. On the right, a modal dialog titled "Existing Model Execution" is open, showing a dropdown "Model" set to "(from Tab:) NORMALIZE\_TRAIN\_SET", a "Type" field containing "Z Score Normalizer Model", and a "Description" section with header mappings. At the bottom of the dialog are "Execute" and "Cancel" buttons.

The results will appear on the output spreadsheet.

The screenshot shows the output spreadsheet from the previous step. The data has been normalized, with values ranging from -0.58108446 to 1.013537686. The columns are labeled Col1 through Col9. A secondary view on the right shows the raw data from the "NORMALIZE\_TRAIN\_SET" tab, with columns User Header, User Row ID, age, sex, bmi, children, smoker, and region. This view includes row numbers 1 through 15 and detailed numerical values for each column.

## 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".

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col8
1	4.442052194	-1.013537686	-0.448337111	-0.894088423	2.0358255719	1.3698195197	160479	16884.924	
2	-1.513797506	0.9856604221	0.5084179971	-0.072051081	-0.490711973	0.438594415	795726	160479	
3	0.796344383	0.9856604221	0.38251975	1.5720236018	-0.490711973	0.438594415	46850026	58298184	1725.5523
4	-0.437617821	0.9856604221	-1.295073532	-0.894088423	-0.490711973	-1.342801383	29140761	46850026	449.462
5	42380545	549418	7306575	29140761	-0.894088423	-0.490711973	46850026	58298184	21984.47061
6	513377539002	0.9856604221	-0.288606275	-0.894088423	-0.490711973	-1.342801383	5403234	46850026	3866.8552
7	0.581108446	-1.013537686	-0.800396912	-0.894088423	-0.490711973	-0.438594415	61993	9782544	3756.6216
8	4.4950712388	-1.013537686	0.04084179971	0.448337111	0.438594415	0.438594415	174208	46850026	4606.4107
9	1.4995056114	-1.013537686	-0.784097848	-0.894088423	-0.490711973	-1.342801383	087537	61993	28923.13692
10	-1.011580320	0.9856604221	-0.722161401	-0.894088423	-0.490711973	0.4656125520	0645637	5368447	271.3208
11	1.155070944	0.9856604221	0.6111021059	-0.894088423	-0.490711973	1.3698195197	247534	549418	1826.843
12	7.2125243621	-1.013537686	-0.13765159	-0.894088423	-0.490711973	0.438594415	683748	5403234	11090.7178
13	0.680809695	0.9856604221	0.8770198214	-0.894088423	0.438594415	0.438594415	4043742	36427	39611.7577
14	4.442052194	0.9856604221	-0.986206252	-0.072051081	-0.490711973	1.3698195197	3752274	549418	1837.237
15	0451325	549418	6516017	75050219	-0.894088423	-0.490711973	46850026	160479	0.4656125520
16	0.9255431128	-1.013537686	0.0210759571	-0.072051081	-0.490711973	0.4656125520	4700EE	61993	10797.3362

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 ( $\alpha$ ) as 0.05 and include all columns.

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8	
1	0	4.442052194	-1.013537686	-0.448337111	-0.894088423	2.0358255719	1.3698195197	160479	
2	1	-1.513797506	0.9856604221	0.5084179971	-0.072051081	-0.490711973	0.438594415	795726	
3	2	-0.796344383	0.9856604221	0.38251975	1.5720236018	-0.490711973	0.438594415	46850026	
4	3	-0.437617821	0.9856604221	-1.295073532	-0.894088423	-0.490711973	-1.342801383	29140761	
5	4	42380545	549418	1963338	5403234	46850026	232496	58298184	
6	5	513377539002	0.9856604221	-0.288606275	-0.894088423	-0.490711973	5403234	46850026	3866.8552
7	6	0.581108446	-1.013537686	-0.800396912	-0.894088423	-0.490711973	61993	9782544	3756.6216
8	7	4.4950712388	-1.013537686	0.04084179971	0.448337111	0.438594415	174208	46850026	4606.4107
9	8	1.4995056114	-1.013537686	-0.784097848	-0.894088423	-0.490711973	087537	61993	28923.13692
10	9	-1.011580320	0.9856604221	-0.722161401	-0.894088423	-0.490711973	0645637	5368447	271.3208
11	10	1.155070944	0.9856604221	0.6111021059	-0.894088423	-0.490711973	247534	549418	1826.843
12	11	7.2125243621	-1.013537686	-0.13765159	-0.894088423	-0.490711973	683748	5403234	11090.7178
13	12	0.680809695	0.9856604221	0.8770198214	-0.894088423	0.438594415	4043742	36427	39611.7577
14	13	4.442052194	0.9856604221	-0.986206252	-0.072051081	-0.490711973	3752274	549418	1837.237
15	14	0451325	549418	6516017	75050219	-0.894088423	46850026	160479	0.4656125520
16	15	0.9255431128	-1.013537686	0.0210759571	-0.072051081	-0.490711973	0.4656125520	4700EE	61993

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)".

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)
1	-1.442052194	1.013537686	-0.448337111	children	smoker	region	
2	-1.513797506	0.9856604221	0.05084179971	-0.072051081	-0.490711973	-0.43859441!	
3	-0.796344380	0.9856604221	0.3829151975	0.5720326018	0.490711973	-0.43859441!	
4	-0.433787313	0.9856604221	0.05084179973	0.29180808423	-0.490711973	-0.43859441!	
5	0.509361313	0.9856604221	-0.288606275	-0.8940840423	-0.490711973	-1.42801381	
6	0.581108446	1.013537686	-0.800396912	-0.8940840423	-0.490711973	-0.43859441!	
7	0.4950712388	0.9856604221	0.45461310830	-0.072051081	-0.490711973	-0.43859441!	
8	-0.150635720	0.9856604221	-0.133765159	0.749866260	-0.490711973	0.465162552	
9	1.034263	0.949418	2379109	39319	46850026	5234968	
10	1.4995056114	1.013537686	-0.784097848	-0.8940840423	-0.490711973	-1.42801381	
11	0.838096955	0.9856604221	0.05084179973	0.40111021059	-0.490711973	-0.43859441!	
12	1.2125243621	1.013537686	1.4945114226	-0.8940840423	-0.490711973	-0.43859441!	
13	0.683748	0.9193	36427	5403234	46850026	5234968	
14	0.403742	0.949418	482108	5403234	795726	5234968	
15	-1.442052194	0.9856604221	-0.986206252	-0.072051081	-0.490711973	1.369819519	

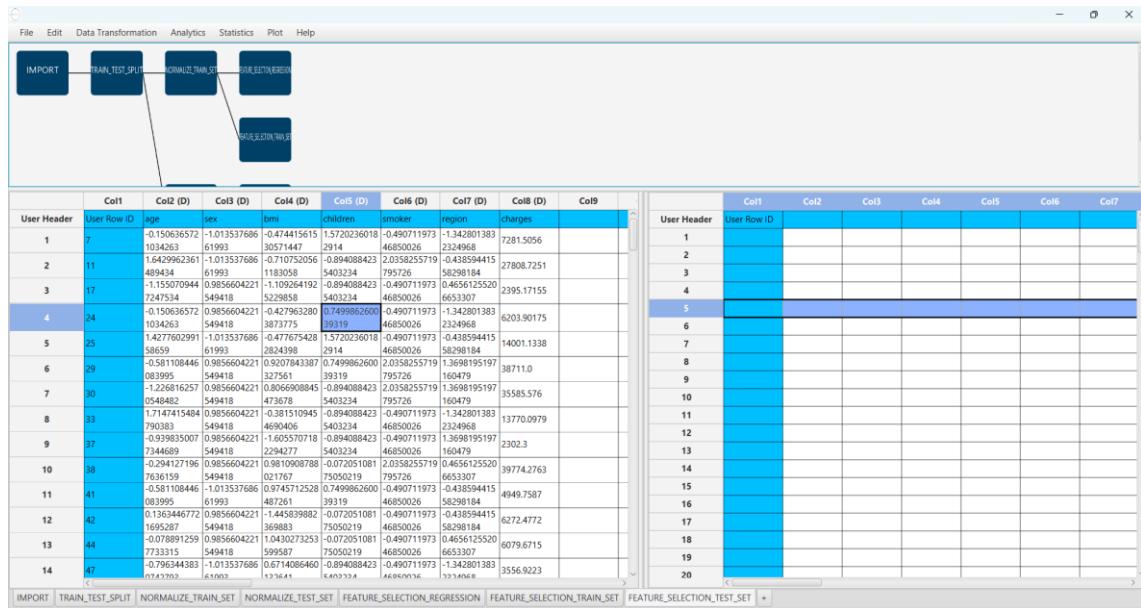
The results will appear on the output spreadsheet.

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
1	0	-1.442052194	0.9856604221	0.05084179971	children	smoker	region	charges
2	1	-1.513797506	0.9856604221	0.3829151975	0.5720326018	-0.490711973	-0.43859441!	
3	2	-0.796344380	0.9856604221	0.3829151975	0.5720326018	-0.490711973	-0.43859441!	
4	3	-0.433787313	0.9856604221	0.05084179973	0.29180808423	-0.490711973	-0.43859441!	
5	4	0.509361313	0.9856604221	-0.288606275	-0.8940840423	-0.490711973	-1.42801381	
6	5	0.581108446	1.013537686	-0.800396912	-0.8940840423	-0.490711973	-0.43859441!	
7	6	0.4950712388	0.9856604221	0.45461310830	-0.072051081	-0.490711973	-0.43859441!	
8	7	0.838096955	0.9856604221	-0.133765159	0.749866260	-0.490711973	-0.43859441!	
9	8	1.034263	0.949418	2379109	39319	46850026	5234968	
10	9	1.4995056114	1.013537686	-0.784097848	-0.8940840423	-0.490711973	-1.42801381	
11	10	0.683748	0.9193	36427	5403234	46850026	5234968	
12	11	1.2125243621	1.013537686	1.4945114226	-0.8940840423	-0.490711973	-0.43859441!	
13	12	0.403742	0.949418	482108	5403234	795726	5234968	
14	13	0.868089695	0.9856604221	1.8710198214	-0.8940840423	0.05084179971	-0.43859441!	
15	14	0.442052194	0.9856604221	-0.986206252	-0.072051081	-0.490711973	1.369819519	

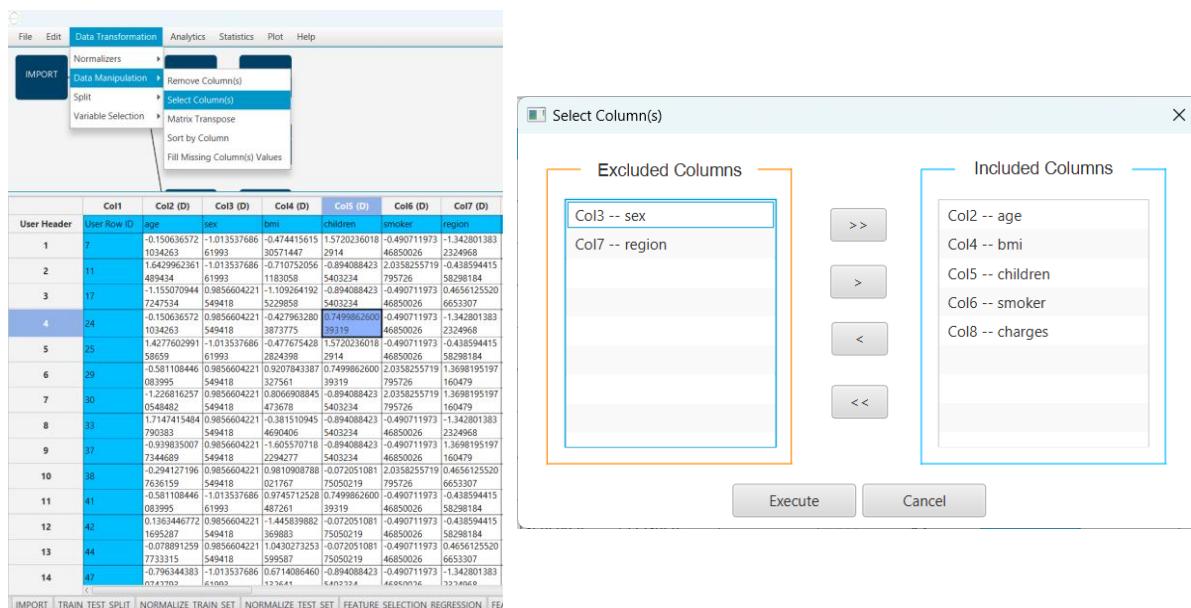
## Step 8: Feature selection: test set

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.

User Header	User Row ID	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9
1	7	-0.150636572	-1.013537686	-0.474415615	0.5720236018	-0.490711973	-1.342801383			7281.5056
2	11	1.6429962361	-1.013537686	-0.710752056	-0.894088423	2.0358255719	-0.438594415			
3	17	1.155070944	0.9856604221	-1.109264192	-0.894088423	-0.490711973	0.4656125520			2395.17155
4	24	-0.150636572	0.9856604221	-0.427963280	0.749986269	-0.490711973	-1.342801383			6203.90175
5	25	1.4277092991	-1.013537686	-0.477675428	1.5720236018	-0.490711973	-0.438594415			14001.1338
6	29	0.581108446	0.9856604221	-0.381510945	0.749986260	-0.490711973	0.569195184			
7	30	0.80995	1.69193	0.49419	0.37764	0.39119	0.795726	160471		38711.0
8	33	1.7147415484	0.9856604221	-0.381510945	-0.894088423	2.0358255719	-1.342801383			13770.0979
9	37	0.939835007	0.9856604221	-1.605570718	-0.894088423	-0.490711973	1.3698195197			2302.3
10	38	7.344689	0.9856604221	0.294277196	0.403234	0.4685026	0.160479			
11	41	0.636192159	-0.1913537686	0.7474712520	0.749986260	-0.072051081	2.0358255719	0.4656125520		39774.2763
12	42	0.525108446	-0.1913537686	0.7474712520	0.749986260	-0.072051081	0.49494415	4949.7587		
13	44	1.136344672	0.9856604221	-1.445839882	-0.894088423	-0.490711973	-0.438594415			
14	47	0.747534	0.9856604221	1.0430273253	0.072051081	-0.490711973	0.4656125520			

User Header	User Row ID	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7
1	7	0.150636572	-0.474415615	1.5720236018	-0.490711973			
2	11	1.6429962361	-0.710752056	2.0358255719	-0.438594415			
3	17	1.155070944	-1.109264192	-0.894088423	-0.490711973			
4	24	-0.150636572	0.749986260	-0.490711973	-1.342801383			
5	25	1.4277092991	-0.477675428	1.5720236018	-0.490711973			
6	29	0.581108446	0.749986260	-0.490711973	0.569195184			
7	30	0.80995	1.69193	0.49419	0.37764	0.39119	0.795726	160471
8	33	1.7147415484	-0.381510945	-0.894088423	2.0358255719	-1.342801383		
9	37	0.939835007	-1.605570718	-0.894088423	-0.490711973			
10	38	7.344689	0.294277196	0.403234	0.4685026	0.160479		
11	41	0.636192159	0.7474712520	0.749986260	-0.072051081	2.0358255719		
12	42	0.525108446	-0.1913537686	0.7474712520	0.749986260	-0.072051081	0.49494415	4949.7587
13	44	1.136344672	0.9856604221	-1.445839882	-0.894088423	-0.490711973		
14	47	0.747534	0.9856604221	1.0430273253	0.072051081	-0.490711973		

## 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".

User Header	User Row ID	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7	Col8
1	4	1.442052194	-0.448337111	-0.894088423	2.0358255719	1684.924			
2	1	-1.513797506	0.5084179971	-0.072051081	-0.490711973	1725.5523			
3	2	3752274	0.769941	0.49191124	0.503234	795726			
4	3	0.796344383	0.3829151975	1.5720236018	-0.490711973	4449.462			
5	4	0.42637617821	0.985315332	0.493134	0.490711973	21984.47061			
6	5	0.509303133	-0.2808060275	-0.894088423	-0.490711973	3866.8552			
7	6	0.83995	0.972544	0.503234	0.4685026	3756.6216			
8	7	0.4950712388	0.4546310830	-0.072051081	-0.490711973	0.8240.5896			
9	8	0.74268	0.610241	0.75050219	0.4685026	6272.4772			
10	9	1.034261	0.2379103	0.39319	0.4685026	599587			
11	10	0.747534	0.49494415	0.671406460	-0.894088423	-0.490711973	2721.3208		
12	11	1.155070944	0.6111021059	-0.894088423	-0.490711973	1826.843			
13	12	7.247534	0.49494415	0.671406460	-0.894088423	-0.490711973	28923.13692		

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

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 screenshot shows the Isalos Analytics Platform's interface. At the top, there's a navigation bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help. The Analytics tab is active. Below the navigation bar is a dropdown menu for 'Regression' which includes k Nearest Neighbors (kNN), Multiple Layer Perception (MLP), Radial Basis Function (RBF), Linear SGD, XGBoost, and Random Forest. The main workspace contains a spreadsheet with columns labeled Col1 through Col8. Below the spreadsheet is a toolbar with buttons for IMPORT, TRAIN\_TEST\_SPLIT, NORMALIZE\_TRAIN\_SET, NORMALIZE\_TEST\_SET, FEATURE\_SELECTION\_REGRESSION, and FEATURE\_SELECTION. A modal window titled 'kNN Regression Model' is displayed, asking for the target column ('Col6 -- charges') and the number of neighbors ('3').

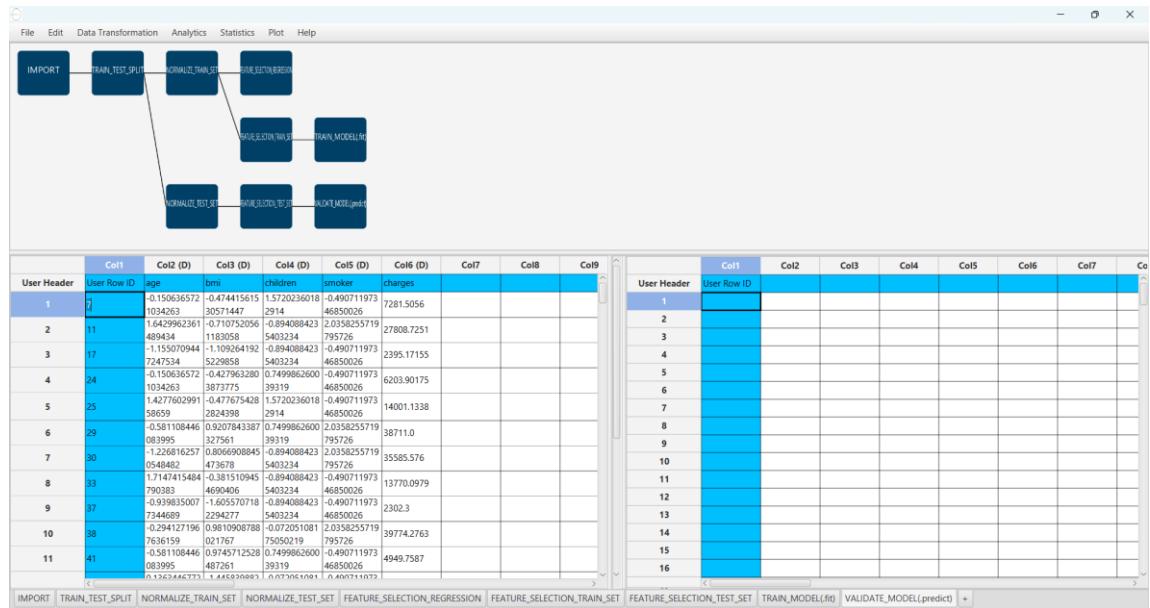
The predictions will appear on the output spreadsheet.

This screenshot shows the completed workflow from the previous one. The 'Analytics' tab remains active. The main workspace now displays a spreadsheet with columns Col1 through Col9. A modal window titled 'kNN Regression Model' is still present, showing the same settings as before. The bottom of the screen features a toolbar with various buttons for data management and model training, including 'TRAIN\_MODEL(.predict)' which is highlighted.

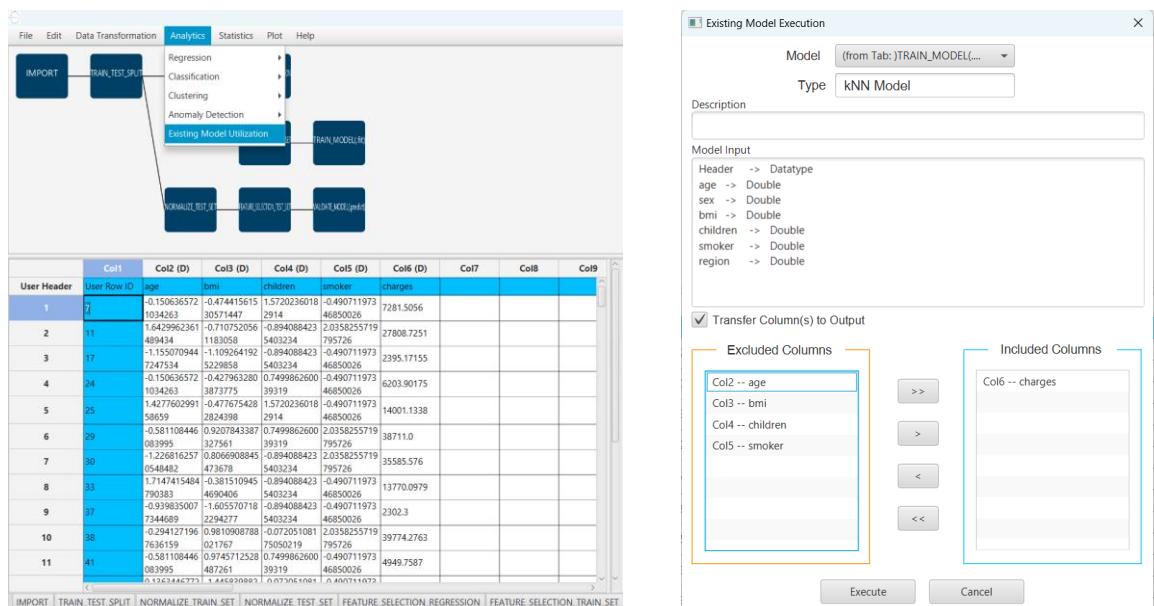
## Step 10: Validate the model

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

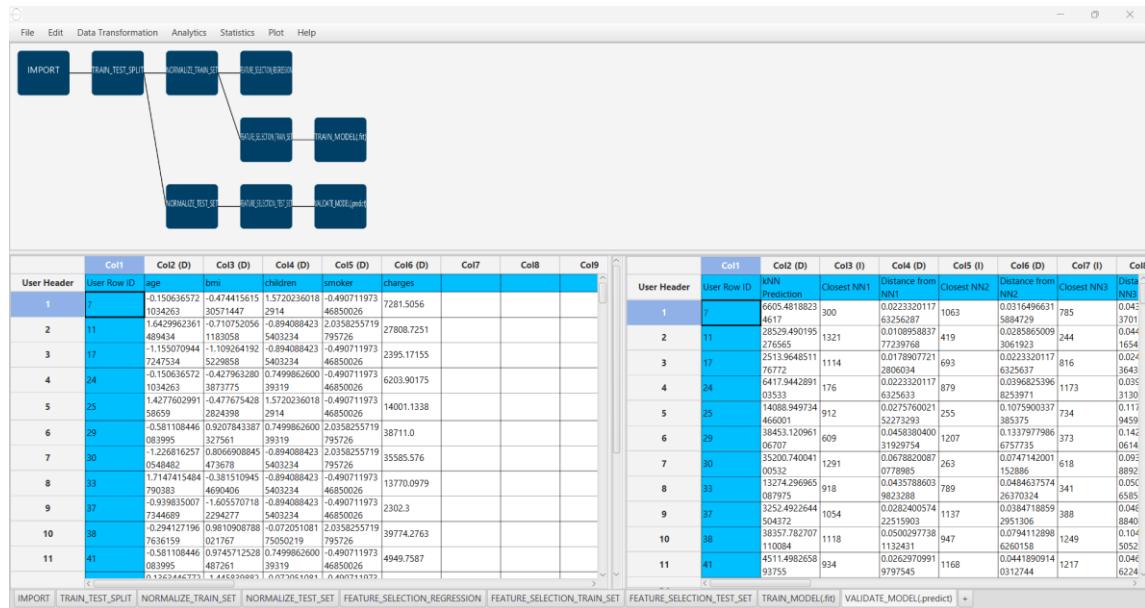
Import data into the input spreadsheet of the "VALIDATE\_MODEL(.predict)" tab from the output of the "FEATURE\_SELECTION\_TEST\_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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.



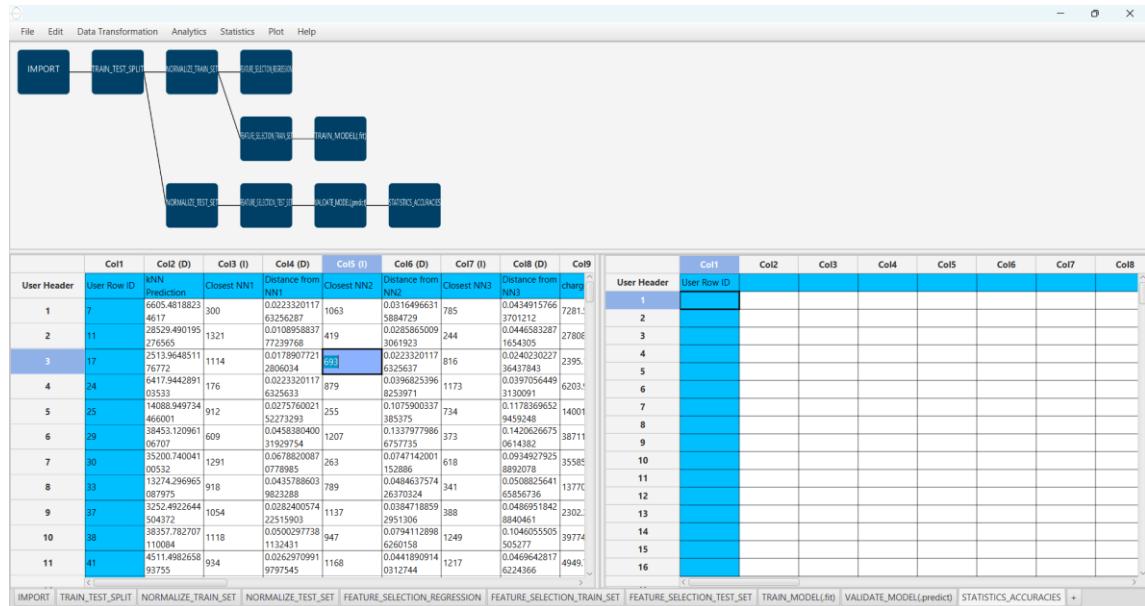
The predictions will appear on the output spreadsheet.



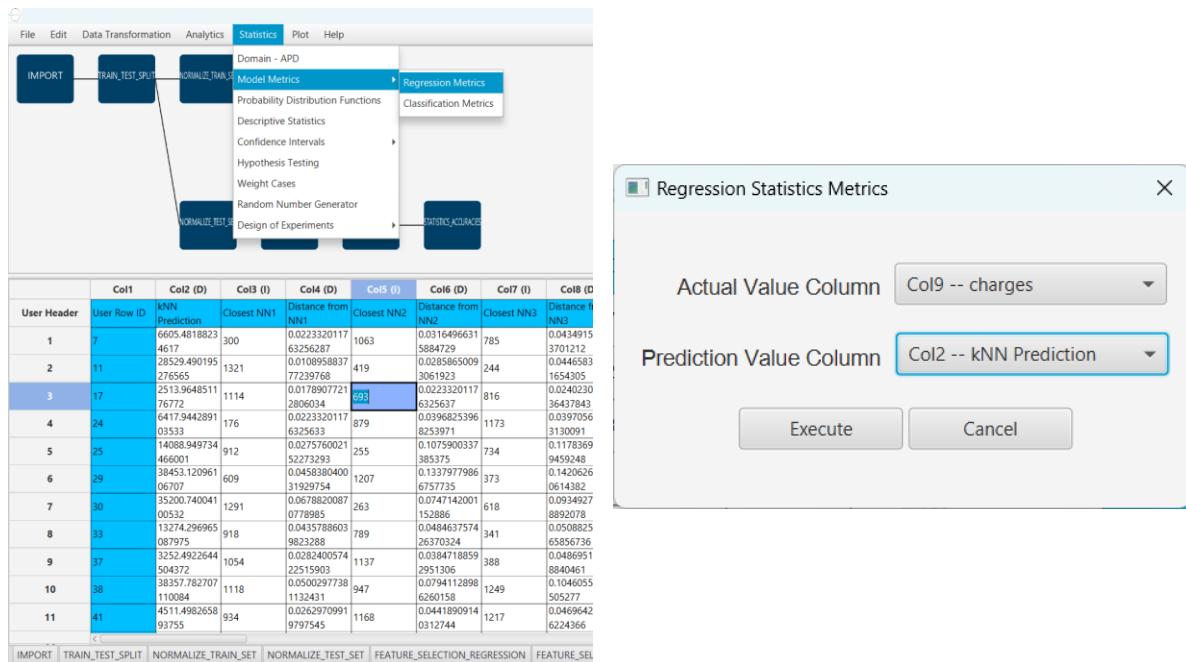
## Step 11: Statistics calculation

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

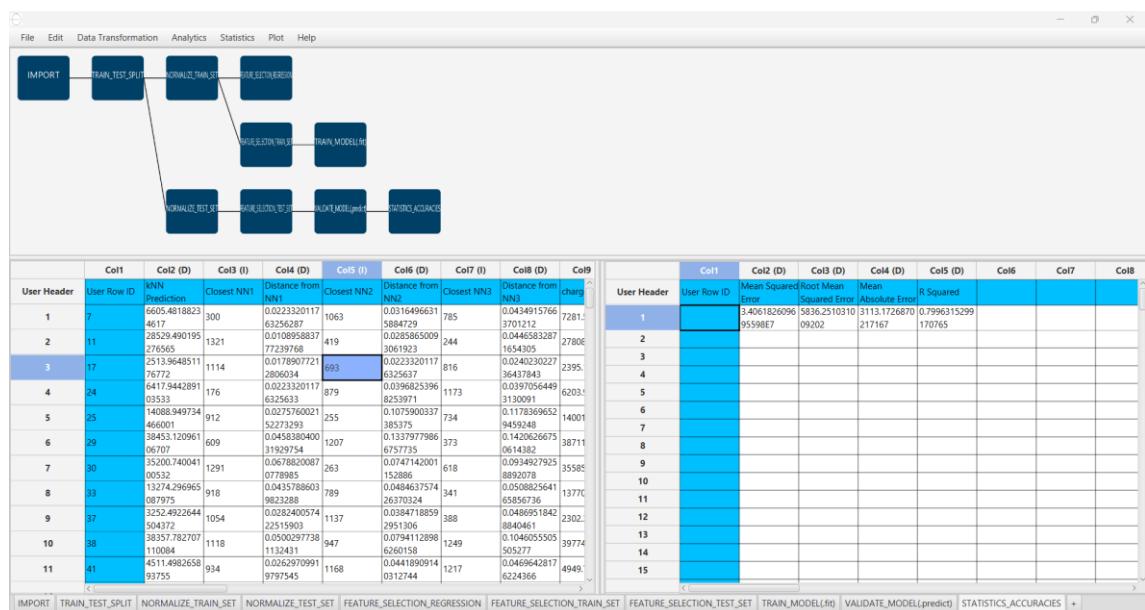
Import data into the input spreadsheet of the "STATISTICS\_ACCURACIES" tab from the output of the "VALIDATE\_MODEL(.predict)" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



The results will appear on the output spreadsheet.

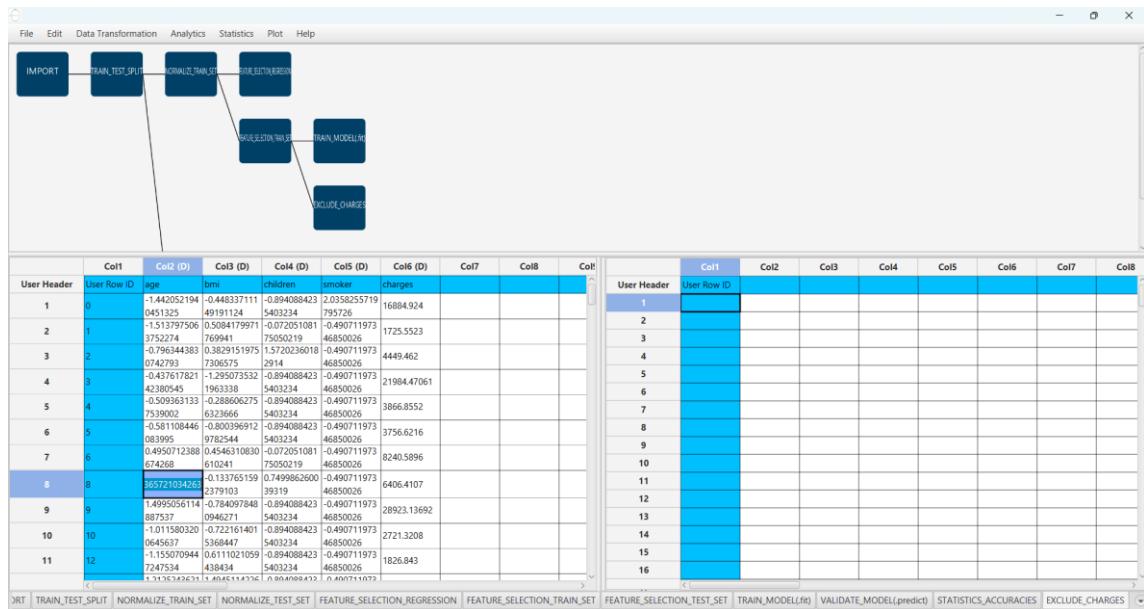


## Step 12: Reliability check of each record of the test set

### 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".



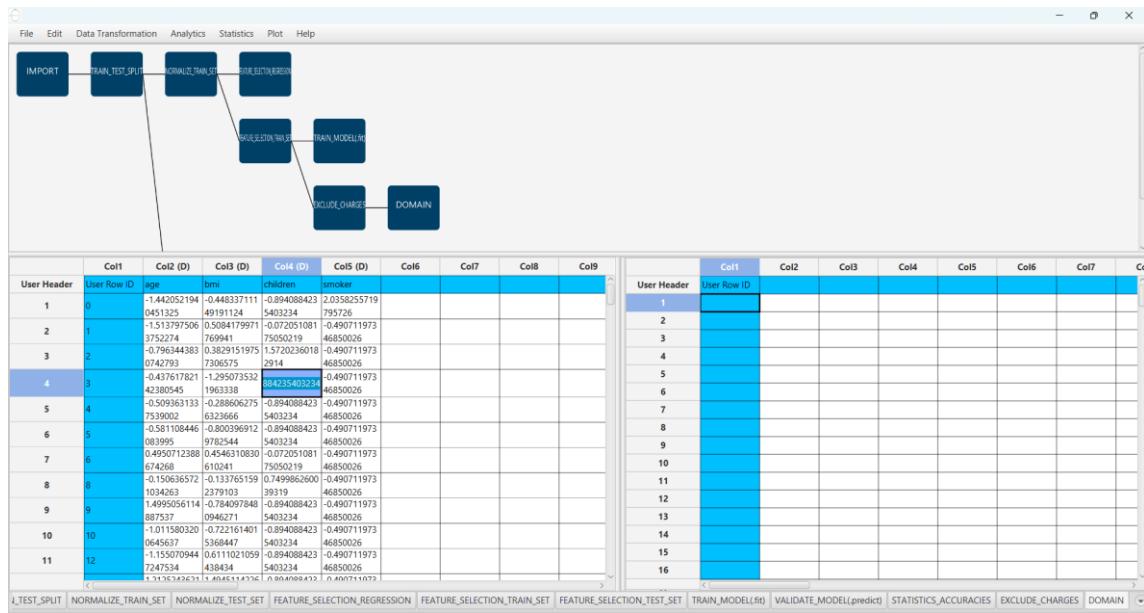
Manipulate the data to exclude the column that corresponds to the "charges" by browsing: "Data Transformation" → "Data Manipulation" → "Select Column(s)". Then select all the columns except the "charges".

This screenshot shows the 'Data Transformation' menu open, with 'Data Manipulation' selected. Under 'Data Manipulation', 'Select Column(s)' is highlighted. To the right, a 'Select Column(s)' dialog box is displayed. It has two main sections: 'Excluded Columns' (highlighted in orange) containing 'Col6 -- charges', and 'Included Columns' (highlighted in blue) containing 'Col2 -- age', 'Col3 -- bmi', 'Col4 -- children', and 'Col5 -- smoker'. At the bottom of the dialog are 'Execute' and 'Cancel' buttons.

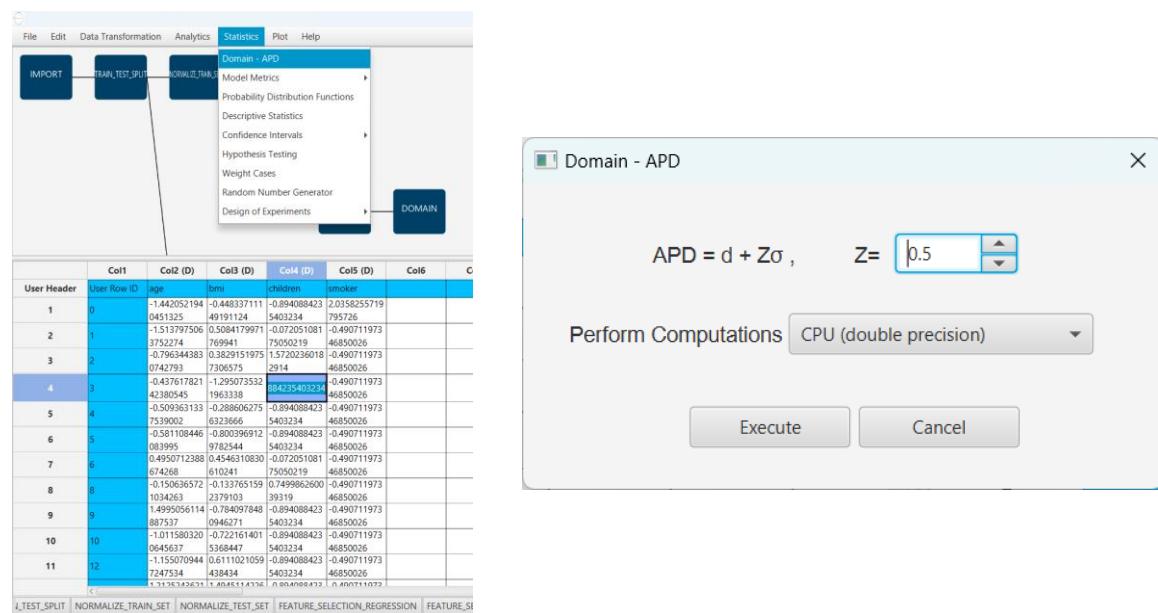
The results will appear on the output spreadsheet.

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

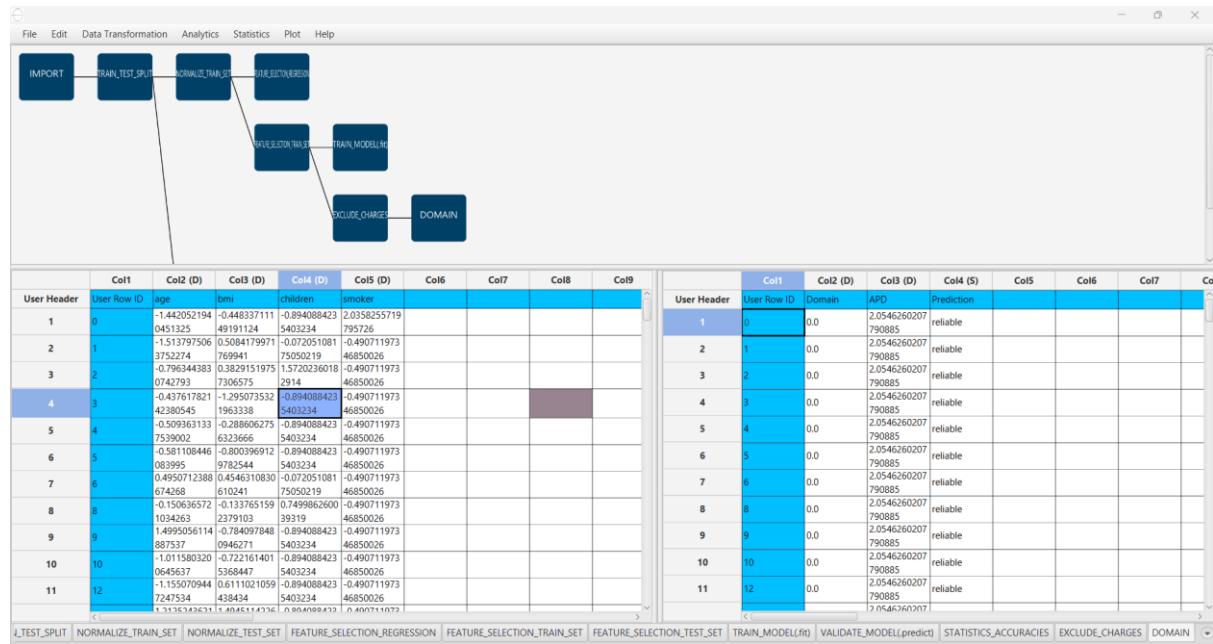
Import data into the input spreadsheet of the "DOMAIN" tab from the output of the "EXCLUDE\_CHARGES" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



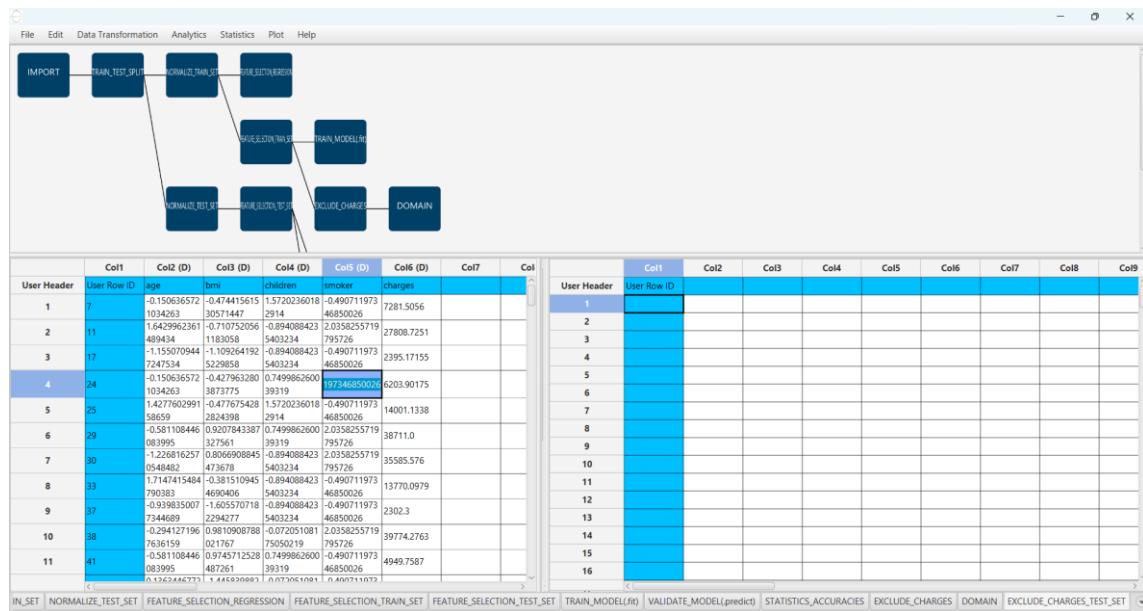
The results will appear on the output spreadsheet.



## 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 screenshot shows the Isalos Analytics Platform's Data Transformation interface. A context menu is open under the 'Data Manipulation' option, with 'Select Column(s)' highlighted. To the right, a 'Select Column(s)' dialog box is displayed, containing two main sections: 'Excluded Columns' (which lists 'Col6 -- charges') and 'Included Columns' (which lists 'Col2 -- age', 'Col3 -- bmi', 'Col4 -- children', and 'Col5 -- smoker'). Below the dialog are two buttons: 'Execute' and 'Cancel'.

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".

The screenshot shows the Isalos Analytics Platform's interface with the 'RELIABILITY' tab active. The input spreadsheet on the left contains data with columns such as 'User Row ID', 'age', 'bmi', 'children', 'smoker', 'Col5 (D)', 'Col6', 'Col7', and 'Col8'. The output spreadsheet on the right is currently empty, showing only the column headers 'Col1' through 'Col8'.

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

The screenshot shows the Isalos Analytics Platform interface. On the left, the Analytics menu is open, showing options like Regression, Classification, Clustering, Anomaly Detection, and Existing Model Utilization. The Existing Model Utilization option is selected, which triggers a flowchart diagram. This flowchart starts with 'NORMALIZE\_TEST\_SET' (highlighted in blue), followed by 'VALIDATE\_MODEL(predict)', then branches into 'EXCLUDE\_CHARGES' and 'DOMAIN'. From 'EXCLUDE\_CHARGES', it leads to 'VALIDATE\_MODEL(predict)' and then to 'STATISTICS\_ACCURACIES'. From 'DOMAIN', it leads to 'VALIDATE\_MODEL(predict)' and then to 'RELIABILITY'. Below the flowchart is a table with columns Col1 through Col7. A row in this table is highlighted in blue, corresponding to the row in the flowchart. To the right of the table is a 'Existing Model Execution' dialog box. It shows the 'Model' dropdown set to '(from Tab:)DOMAIN' and the 'Type' dropdown set to 'APD Model'. Under 'Model Input', there is a list of variables and their types: age -> Double, sex -> Double, bmi -> Double, children -> Double, smoker -> Double, and region -> Double. Below this is a checked checkbox labeled 'Transfer Column(s) to Output'. On the right side of the dialog, there are two sections: 'Excluded Columns' (empty) and 'Included Columns' containing 'Col2 -- age', 'Col3 -- bmi', 'Col4 -- children', and 'Col5 -- smoker'. At the bottom are 'Execute' and 'Cancel' buttons.

The results will appear on the output spreadsheet.

This screenshot shows the Isalos Analytics Platform interface with the output spreadsheet. The table has columns Col1 through Col8. The first few rows of data are identical to the input table. However, starting from row 11, the data is processed according to the flowchart. Row 11 shows the result of 'EXCLUDE\_CHARGES', where the 'smoker' column is removed. Rows 12 and 13 show the result of 'DOMAIN', where the 'age' column is removed. The remaining columns are 'User Row ID', 'User Row ID', 'Domain', 'APD', 'Prediction', 'age', 'bmi', 'children', and 'smoker'. The 'age' column is present in the first three rows but absent in the subsequent ones. The 'smoker' column is also absent in the later rows. The 'APD' and 'Prediction' columns are populated with values from the model execution dialog. The 'User Row ID' column contains IDs 7, 11, 17, 24, 25, 29, 30, 33, 37, 38, and 41. The 'Domain' column contains 'TEST\_SET'. The 'APD' column contains values like 0.050636572, 0.16429962361, etc. The 'Prediction' column contains values like reliable, 0.155070944, etc. The 'age', 'bmi', 'children', and 'smoker' columns contain numerical values corresponding to the rows.

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:

