



# Breast Cancer (Binary Classification)

The goal of this study is to train a model in order to predict whether the cancer is benign (B) or malignant (M). The dataset used in this case study is found in <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data> and has 32 features and 569 labelled samples. The features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

## 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 breast cancer data.

The screenshot shows the Isalos Analytics Platform interface. At the top, there is a header row with columns labeled "User Header" and "User Row ID". Below this, rows 1 through 12 are listed. Row 6 is highlighted in blue. A context menu is open over row 6, containing options: "Import from SpreadSheet", "Import from file", "Export Spread Sheet Data", and "Clear SpreadSheet". Below the table, there is a large text area with the heading "IMPORT" and a table of data. This table has a header row with columns "Col1" through "Col12" and a "User Header" row with columns "User Row ID" and "id". The data rows (1 through 21) contain various numerical values and labels like "M" and "B".

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

Col1	Col2 (I)	Col3 (S)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)
User Header	User Row ID	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	symmetry_mean
1	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419
2	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812
3	84300903	M	19.69	21.25	130	1203	0.1095	0.1599	0.1974	0.1279	0.2069
4	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597
5	84358402	M	20.29	14.34	135.1	1297	0.1009	0.1328	0.194	0.1043	0.1809
6	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087
7	844359	M	10.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794
8	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196
9	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235
10	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203
11	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528
12	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842
13	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397
14	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847
15	84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069
16	84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303
17	849406	M	14.68	20.13	94.74	684.5	0.09967	0.072	0.07395	0.05259	0.1586
18	84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164
19	849014	M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582
20	8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885
21	8510453	R	13.08	15.71	85.62	520	0.1075	0.177	0.04568	0.03111	0.1947

## Step 2: Manipulate data

In our Dataset there are not empty values, and the only categorical feature is the label ("Diagnosis") which has two categories and the number of samples in each category are:

- Benign (B): 357
- Malignant (M): 212

In order to use the data for training we have to exclude any columns that do not contain features, like the "id" column. We follow these steps to execute this:

- On the menu click on "Data Transformation" → "Data Manipulation" → "Select Column(s)"
- Select all columns except the one that corresponds to the id.

User Header	User Row ID	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)
1		842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	
2		842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	
3		84309903	M	19.69	21.25	130	1203	0.1096	0.1599	
4		84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2639	
5		84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	
6		843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	
7		844359	M	18.25	19.98	119.6	1040	0.09463	0.109	
8		84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	
9		844981	M	13	21.82	87.5	519.8	0.1273	0.1932	
10		84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	
11		845636	M	16.02	23.24	102.7	797.8	0.08205	0.06669	
12		84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	
13		846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	
14		846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	
15		84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	
16		84799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	
17		848406	M	14.68	20.13	94.74	684.5	0.09867	0.072	
18		84862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	
19		849014	M	19.81	22.15	130	1260	0.09831	0.1027	
20		8510426	B	13.54	14.36	87.46	566.3	0.09779	0.08129	
21		8510653	B	13.08	15.71	85.63	520	0.1075	0.127	
22		8510824	B	9.504	12.44	60.34	273.9	0.1024	0.06492	
23		8511133	M	15.34	14.26	102.5	704.4	0.1073	0.2135	

The data without the "id" column 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".

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

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

**Random Partitioning**

Training set percentage	75
<input type="checkbox"/> Usage of random generator seed	6615372866300
<input checked="" type="checkbox"/> Stratified sampling	Col2 -- diagnosis
Execute	Cancel

The results will appear on the output spreadsheet.

The screenshot shows a workflow diagram at the top with nodes: IMPORT, TRAIN\_TEST\_SPLIT, and NORMALIZE\_TRAIN\_SET. Below it is a large spreadsheet table with columns: Col1, Col2 (S), Col3 (D), Col4 (D), Col5 (D), and Col6. A context menu is open over the first few rows, showing options: Import from SpreadSheet, Import from file, Export Spread Sheet Data, and Clear SpreadSheet. In the center, a 'Random Partitioning' dialog box is displayed with the following settings: Training set percentage (75), Usage of random generator seed (6615372866300), and Stratified sampling (Col2 -- diagnosis). At the bottom right of the dialog is a 'Reconfigure' button.

User Header	Col1	Col2 (S)	Col3 (D)	Col4 (D)	Col5 (D)	Col6
1	M	17.99	10.38	122.8	1001	
2	M	20.57	17.77	132.9	1326	
3	M	19.69	21.25	130	1203	
4	M	11.42	20.38	77.58	386.1	
5	M	20.29	14.34	135.1	1040	
6	M	12.45	15.7	82.57	577.9	
7	M	18.25	19.98	119.6	519.8	
8	M	13.71	20.83	90.2	797.8	
9	M	13	21.82	87.5	1123	
10	M	12.46	24.04	83.97	782.7	
11	M	16.02	23.24	102.7	578.3	
12	M	15.78	17.89	103.6	684.5	
13	M	19.17	24.8	132.4	798.8	
14	M	15.85	23.95	103.7	1260	
15	M	13.73	22.61	93.6	566.3	
16	M	14.54	27.54	96.73	520	
17	M	14.68	20.13	94.74	273.9	
18	M	16.13	20.68	108.1	704.4	

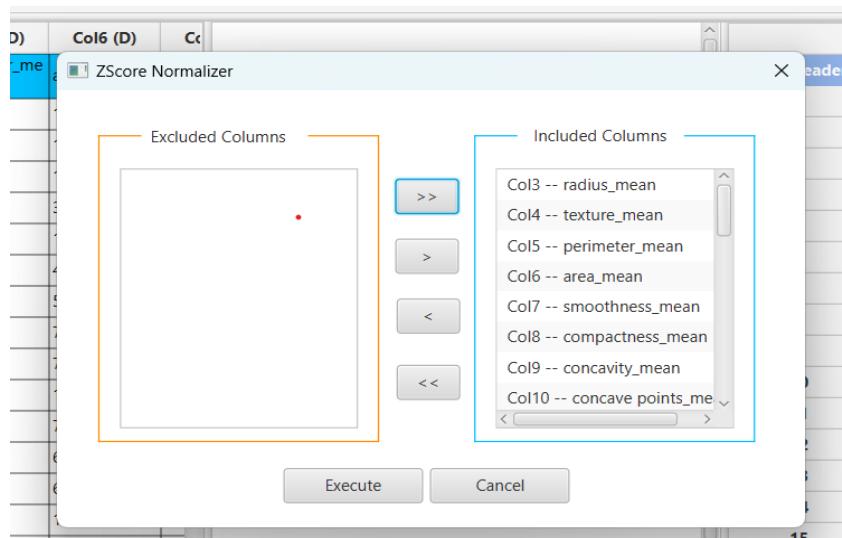
## Step 4: Normalize the training set

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

Import data into the input spreadsheet of the "NORMALISE\_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 two tabs open in the platform. The left tab is 'TRAIN\_TEST\_SPLIT' containing a 21x10 grid of data. The right tab is 'NORMALISE\_TRAIN\_SET' containing a 21x2 grid of data. The columns in the TRAIN\_TEST\_SPLIT tab are labeled Col1 through Col10. The columns in the NORMALISE\_TRAIN\_SET tab are labeled Col1 and Col2.

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



The results will appear on the output spreadsheet.

User Header	User Row ID	diagnosis	radius_mean	texture_mean
1	M	17.99	10.38	
2	M	20.57	17.77	
3	M	19.69	21.25	
4	M	11.42	20.38	
5	M	20.29	14.34	
6	M	12.45	15.7	
7	M	13	21.82	
8	M	16.02	23.24	
9	M	15.78	17.89	
10	M	19.17	24.8	
11	M	15.85	23.95	
12	M	14.54	27.54	
13	M	14.68	20.13	
14	M	19.81	22.15	
15	B	13.54	14.36	
16	B	13.08	15.71	
17	M	15.34	14.26	
18	M	17.14	16.4	
19	M	14.58	21.53	
20	M	17.57	15.05	
21	M	18.63	25.11	

## Step 5: Normalize the test set

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

Import data into the input spreadsheet of the "NORMALISE\_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".

User Header	User Row ID	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
1	M	20.29	14.34	135.1	1297	0.1003	0.1328	
2	M	12.45	15.7	82.57	477.1	0.1278	0.17	
3	M	12.46	24.04	83.97	475.9	0.1186	0.2396	
4	M	15.78	17.89	103.6	781	0.0971	0.1292	
5	M	14.54	27.54	96.73	658.8	0.1139	0.1595	
6	M	21.16	23.04	137.2	1404	0.09428	0.1022	
7	M	18.61	20.25	122.1	1094	0.0944	0.1066	
8	M	19.27	26.47	127.9	1162	0.09401	0.1719	
9	M	16.13	17.88	107	807.2	0.104	0.1559	
10	M	14.25	21.72	93.63	633	0.09823	0.1098	
11	B	13.03	18.42	82.61	523.8	0.08983	0.03766	
12	M	14.99	25.2	95.54	698.8	0.09387	0.05131	
13	M	13.44	21.58	86.18	563	0.08162	0.06031	
14	M	13.17	21.81	85.42	531.5	0.09714	0.1047	
15	B	13.49	22.3	86.91	561	0.08752	0.07698	
16	B	11.76	21.6	74.72	427.9	0.08637	0.04966	
17	B	13.64	16.34	87.21	571.8	0.07685	0.06059	
18	B	11.94	18.24	75.71	437.6	0.08261	0.04751	

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

The screenshot shows the Isalos Analytics Platform interface. At the top, there's a navigation bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help. The Analytics tab is selected. Below the navigation bar is a data grid displaying a dataset with columns: User Header, User Row ID, diagnosis, radius\_mean, texture\_mean, perimeter\_mean, area\_mean, smoothness\_mean, compactness\_mean, and Col8 (D). A data flow diagram is visible on the left, starting with an IMPORT node, followed by TRAIN\_TEST\_SPLIT, and then Existing Model Utilization. A context menu is open over the Existing Model Utilization node, showing options like Regression, Classification, Clustering, Anomaly Detection, and Existing Model Utilization. To the right, a modal window titled "Existing Model Execution" is open, showing a Model dropdown set to "(from Tab:)NORMALISE\_TR...", a Type dropdown set to "Z Score Normalizer Model", and a Description section with detailed input headers. Buttons for Execute and Cancel are at the bottom.

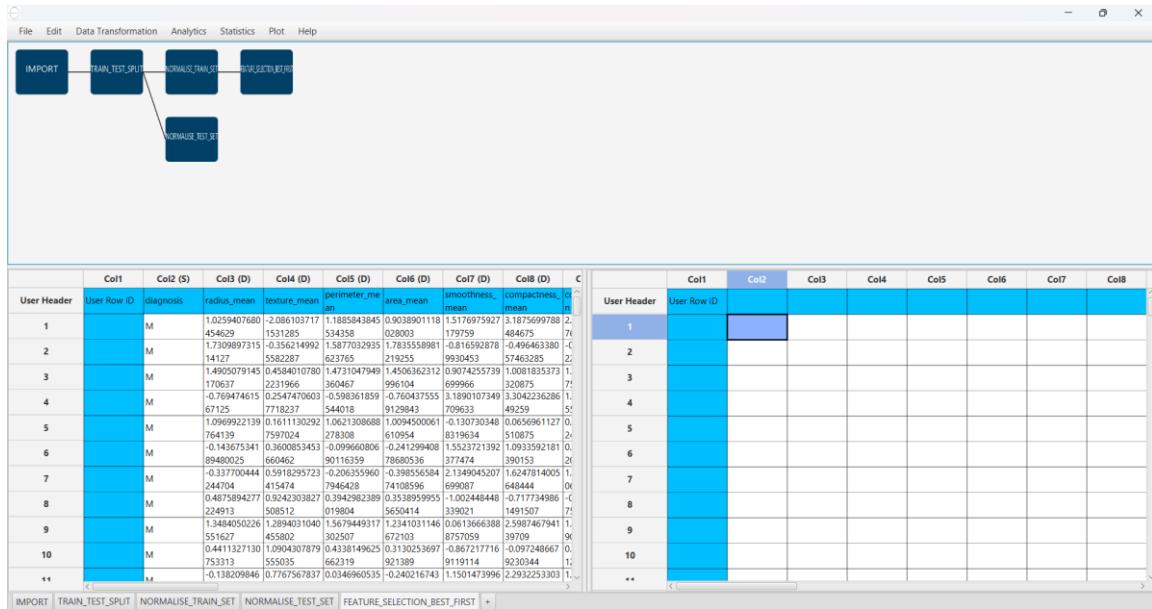
The results will appear on the output spreadsheet.

The screenshot shows the Isalos Analytics Platform interface again. The data flow diagram now includes a NORMALISE\_TRAIN\_SET node between TRAIN\_TEST\_SPLIT and Existing Model Utilization. The data grid on the left remains the same. On the right, there are two side-by-side spreadsheets. The left spreadsheet is identical to the one shown in the first screenshot. The right spreadsheet has 13 rows and 8 columns, with columns labeled Col1 through Col8. The data values in the right spreadsheet are scaled versions of the original data, reflecting the normalization process. The bottom of the screen shows tabs for IMPORT, TRAIN\_TEST\_SPLIT, NORMALISE\_TRAIN\_SET, and NORMALISE\_TEST\_SET.

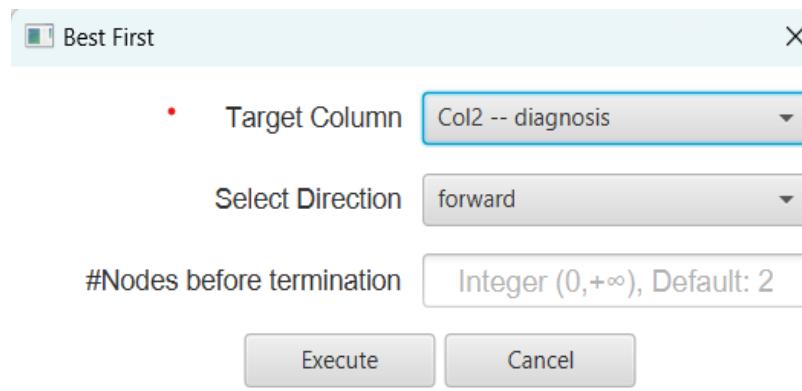
## Step 6: Feature selection

Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE\_SELECTION\_BEST\_FIRST".

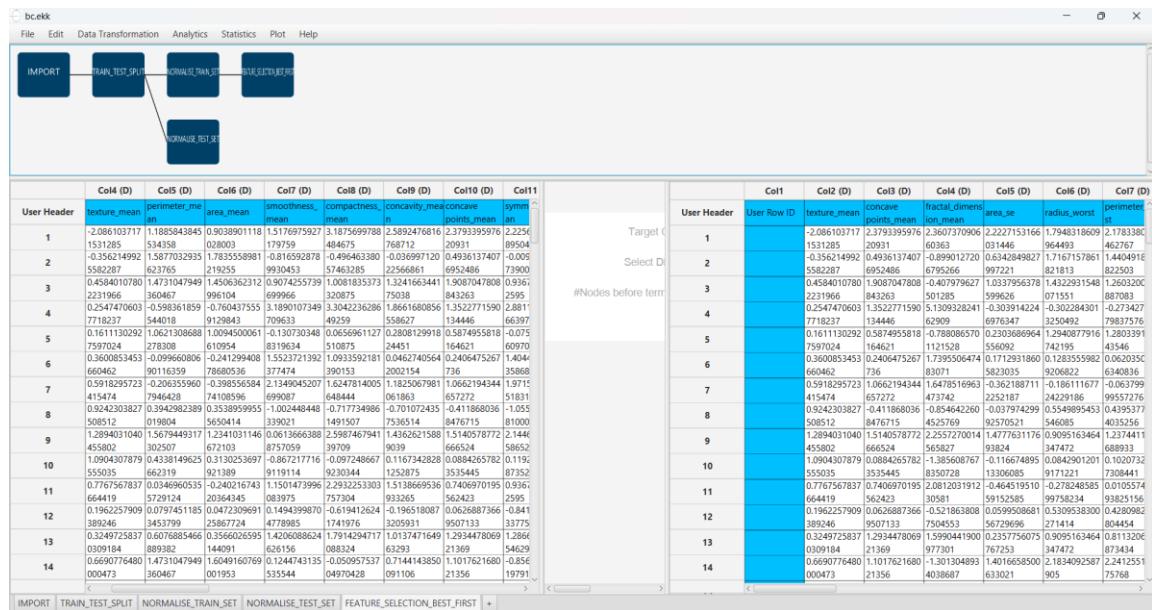
Import data into the input spreadsheet of the "FEATURE\_SELECTION\_BEST\_FIRST" tab from the output of the "NORMALISE\_TRAIN\_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



Choose the most important features for the classification using the Best First Function by browsing: "Data Transformation" → "Variable Selection" → "Best First". Then choose the "diagnosis" column as the target variable and the direction as forward.



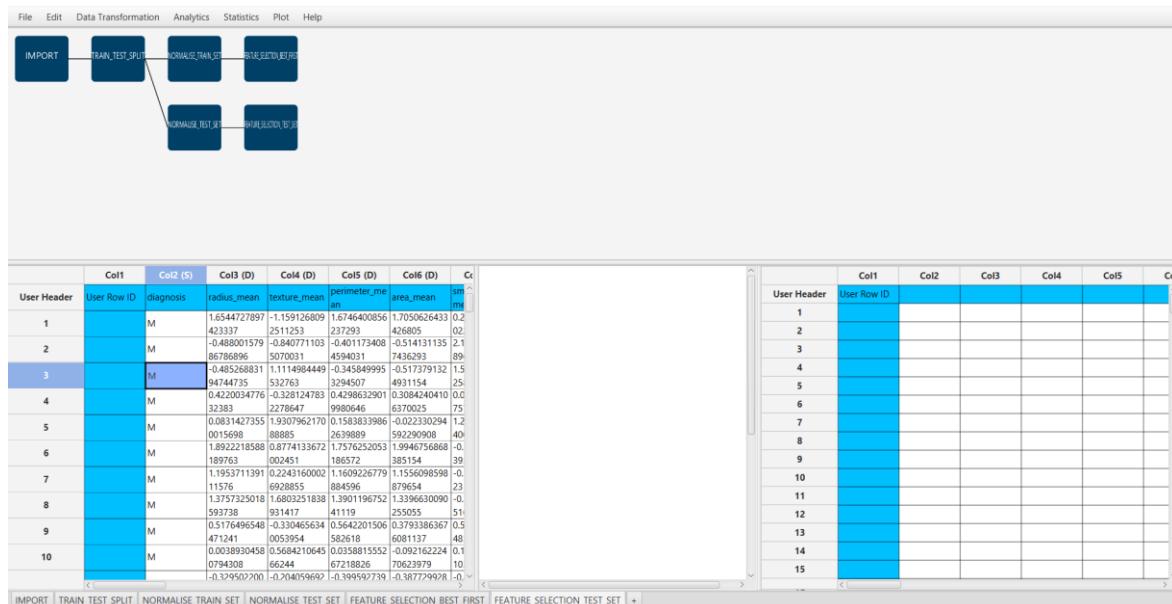
The results will appear on the output spreadsheet.



## Step 7: 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 "NORMALISE\_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 previous step): "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 options like Import, Data Manipulation (with Remove Column(s) and Select Column(s) highlighted), Split, Variable Selection, Matrix Transpose, Sort by Column, and Fill Missing Column(s) Values.

A modal dialog titled "Select Column(s)" is displayed, containing two lists: Excluded Columns and Included Columns. The Excluded Columns list contains columns like Col19 -- concavity\_se, Col20 -- concave points\_se, etc. The Included Columns list contains columns like Col12 -- fractal\_dimension\_r^, Col16 -- area\_se, etc. Buttons for >>, >, <, <<, Execute, and Cancel are at the bottom.

The main workspace shows a spreadsheet with columns labeled Col1 through Col31. The rows represent data points with User Row ID, diagnosis, and various numerical features. Below the spreadsheet are tabs for IMPORT, TRAIN\_TEST\_SPLIT, NORMALISE\_TRAIN\_SET, NORMALISE\_TEST\_SET, and FEATURE\_SELECTION\_BEST\_FIRST.

The results will appear on the output spreadsheet.

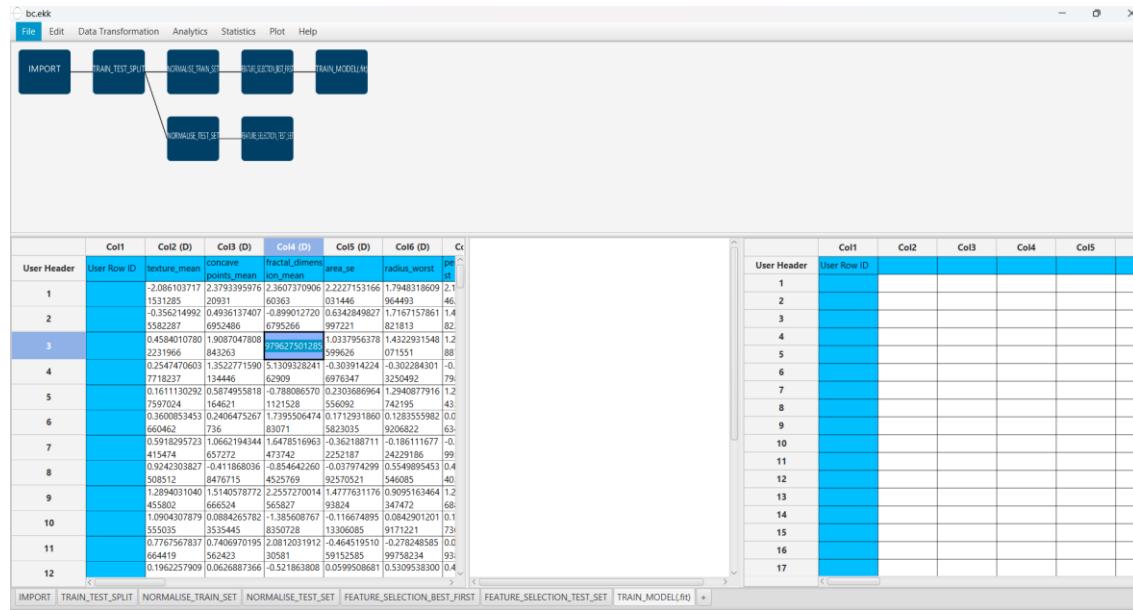
The screenshot shows a data flow diagram where the "FEATURE\_SELECTION\_BEST\_FIRST" tab from the previous step is connected to a "NORMALISE\_TRAIN\_SET" node, which then connects to a "TRAIN\_TEST\_SPLIT" node. The "TRAIN\_TEST\_SPLIT" node has two outputs: "NORMALISE\_TRAIN\_SET" and "NORMALISE\_TEST\_SET".

The main workspace shows a large output spreadsheet with many rows of data. The columns are labeled Col1 through Col31. The data includes rows for users 1 through 12, with columns for User Row ID, diagnosis, and various numerical features like radius\_mean, texture\_mean, perimeter\_mean, etc.

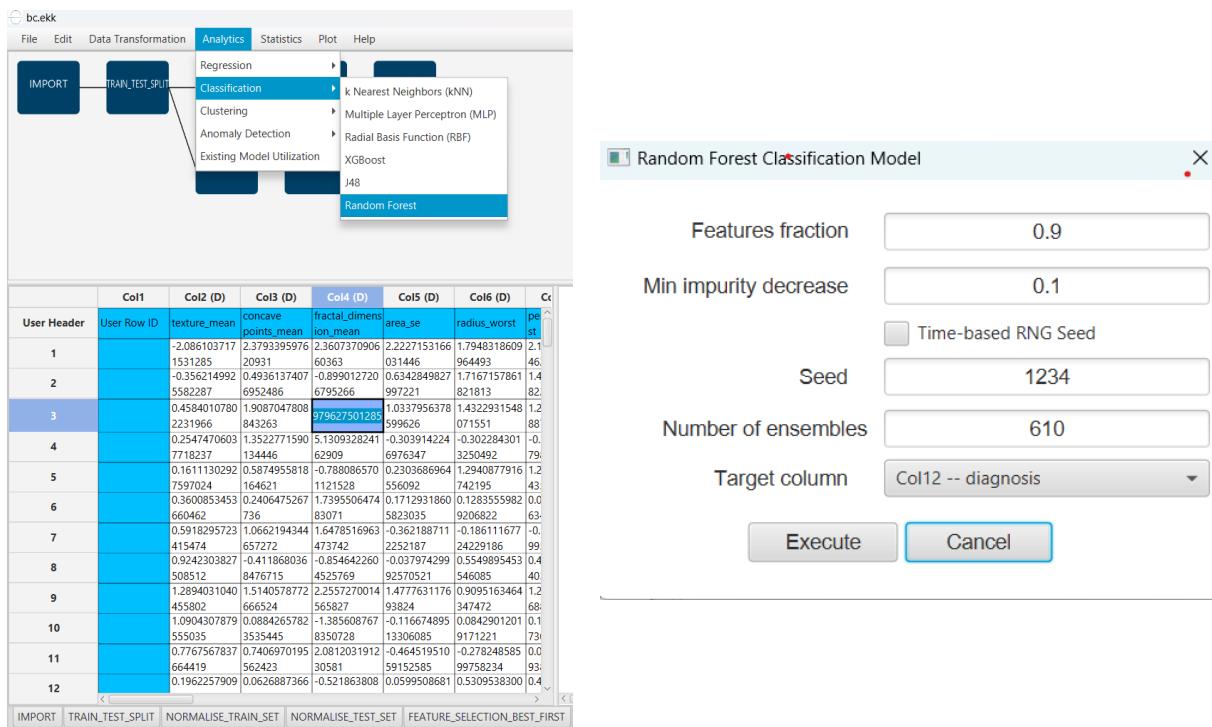
## Step 8: Train the model

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

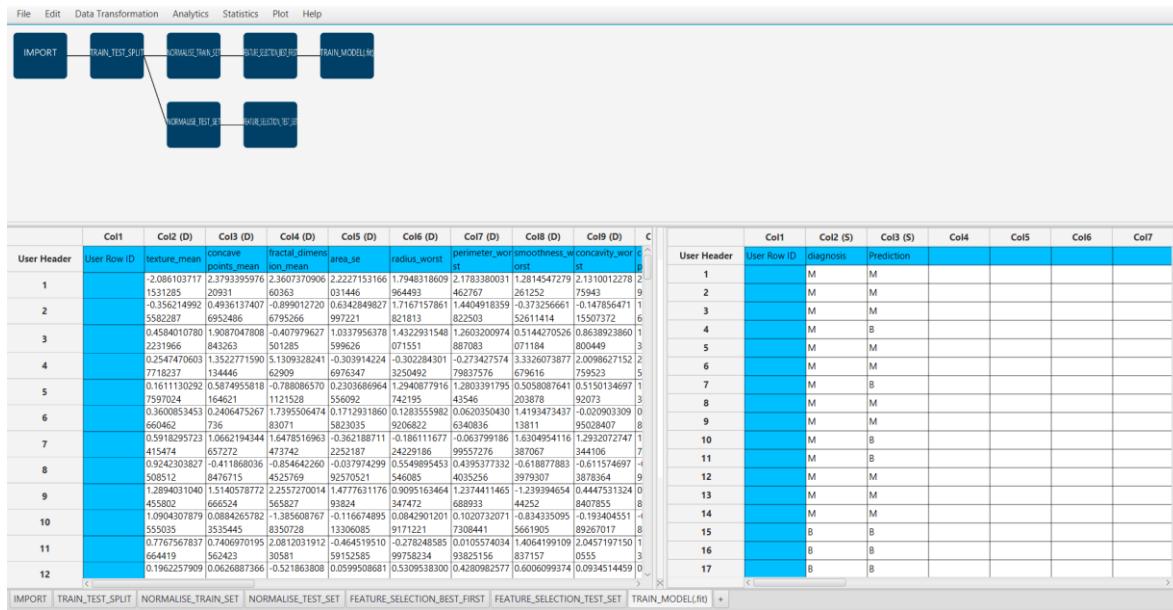
Import data into the input spreadsheet of the "TRAIN\_MODEL(.fit)" tab from the output of the "FEATURE\_SELECTION\_BEST\_FIRST" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



Use the Random Forest Method to train and fit the model by browsing: "Analytics" → "Classification" → "Random Forest" and adjust the model parameters based on training set performance.



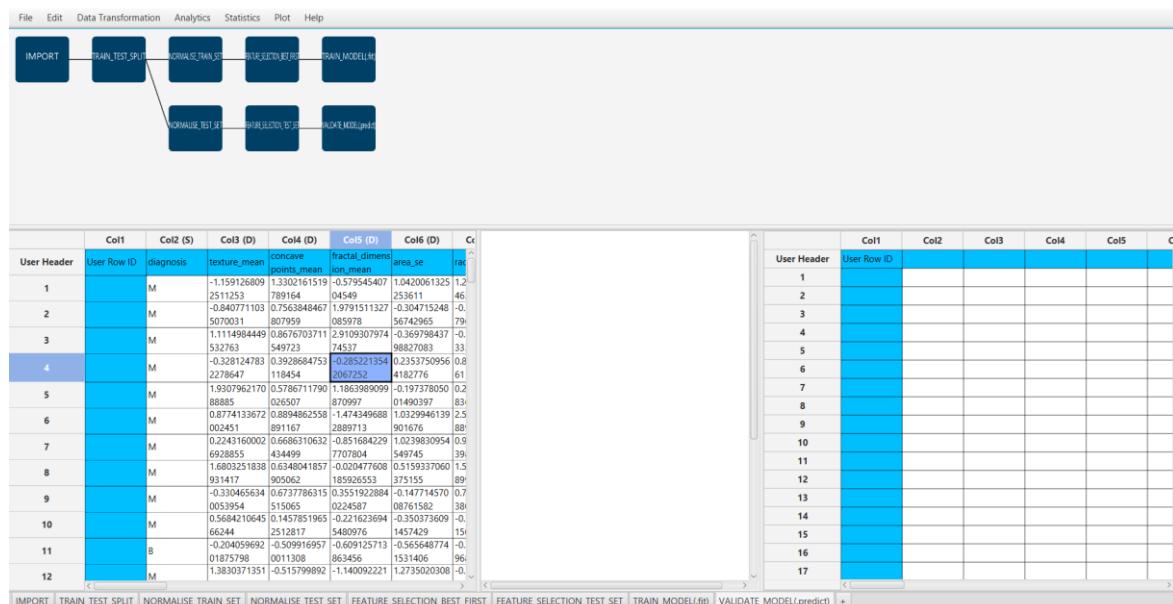
The predictions will appear on the output spreadsheet.



## Step 9: 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".



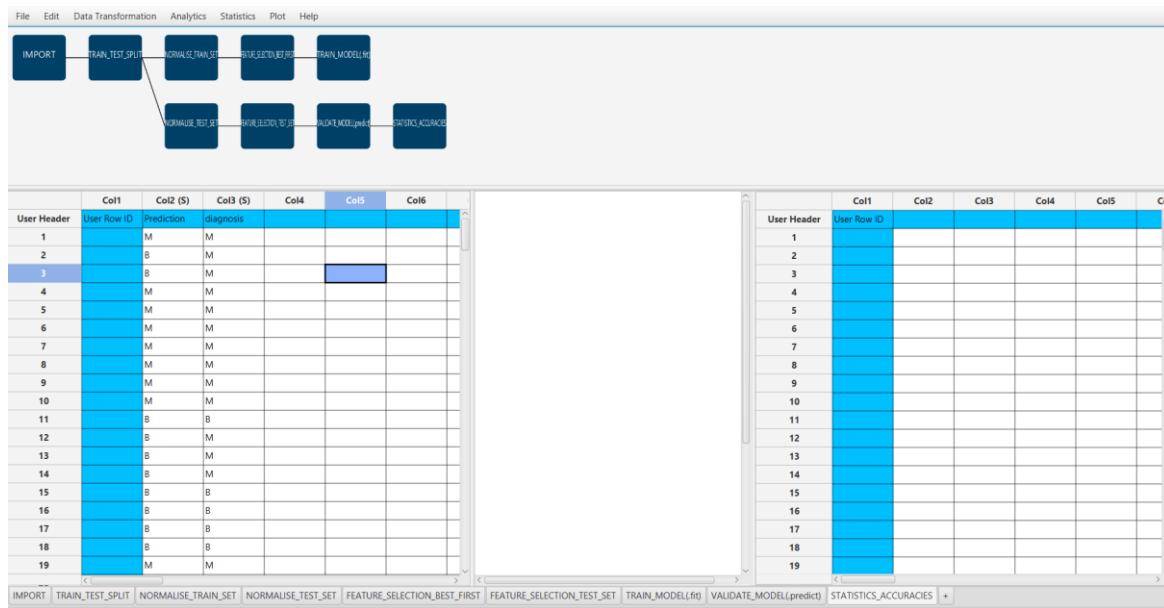
To validate the model browse: "Analytics" → "Existing Model Utilization". Then choose Model "(from Tab:) TRAIN\_MODEL (.fit)".

The predictions will appear on the output spreadsheet.

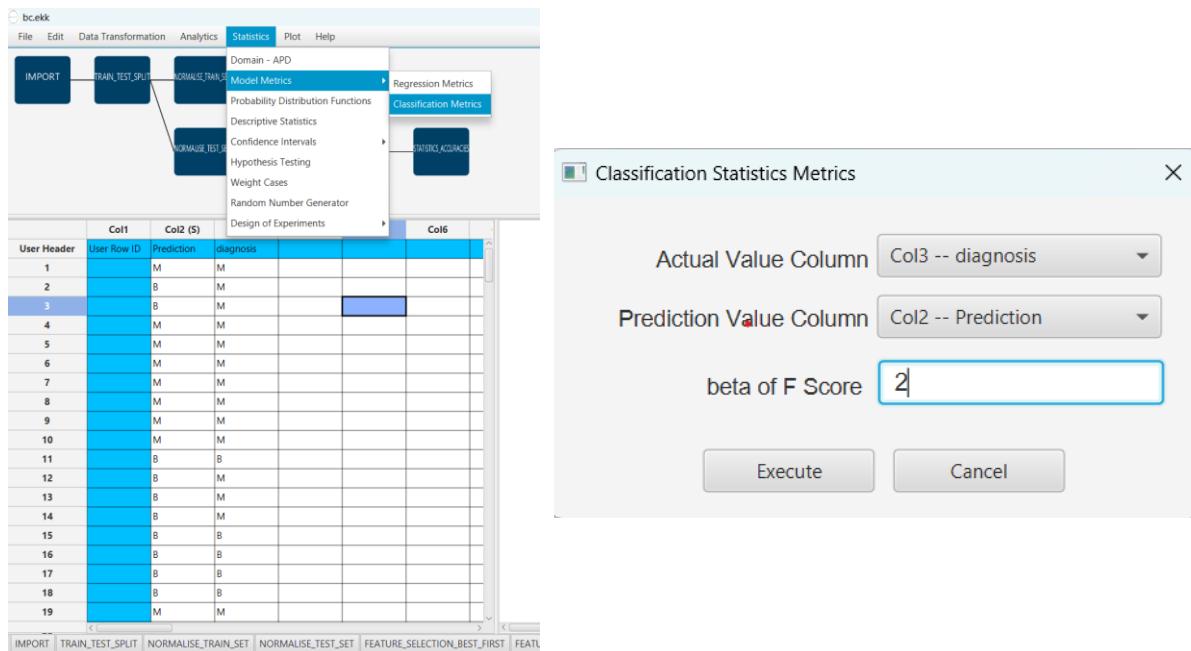
## Step 10: Statistics calculation

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".



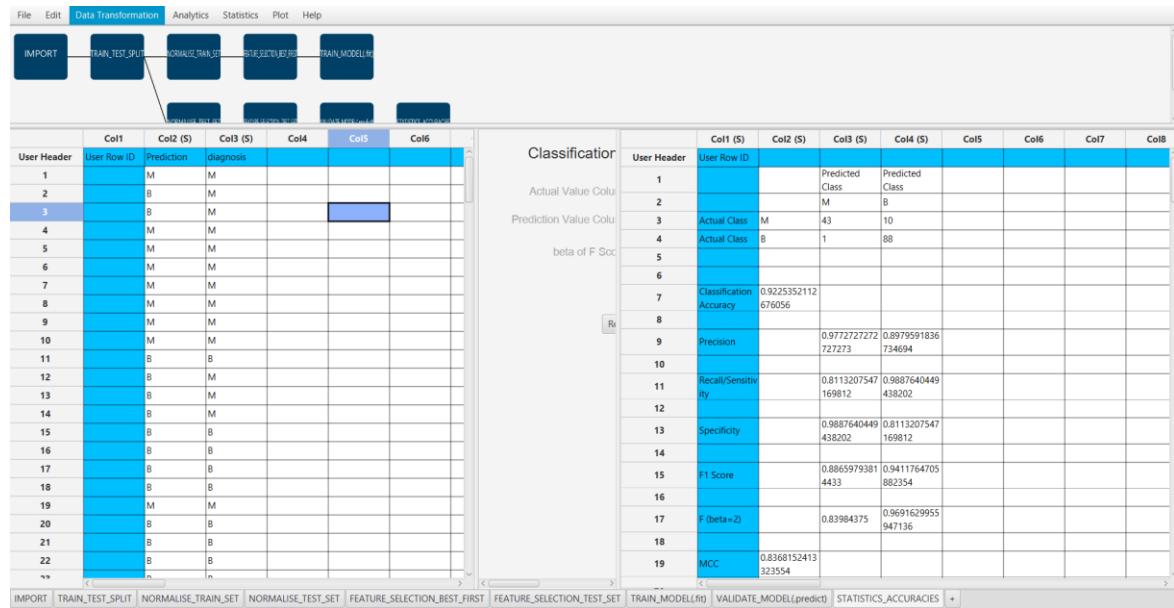
Calculate the statistical metrics for the classification by browsing: "Statistics" → "Model Metrics" → "Classification Metrics".



The results will appear on the output spreadsheet.

Accuracy: 0.923

F1-Score = 0.914

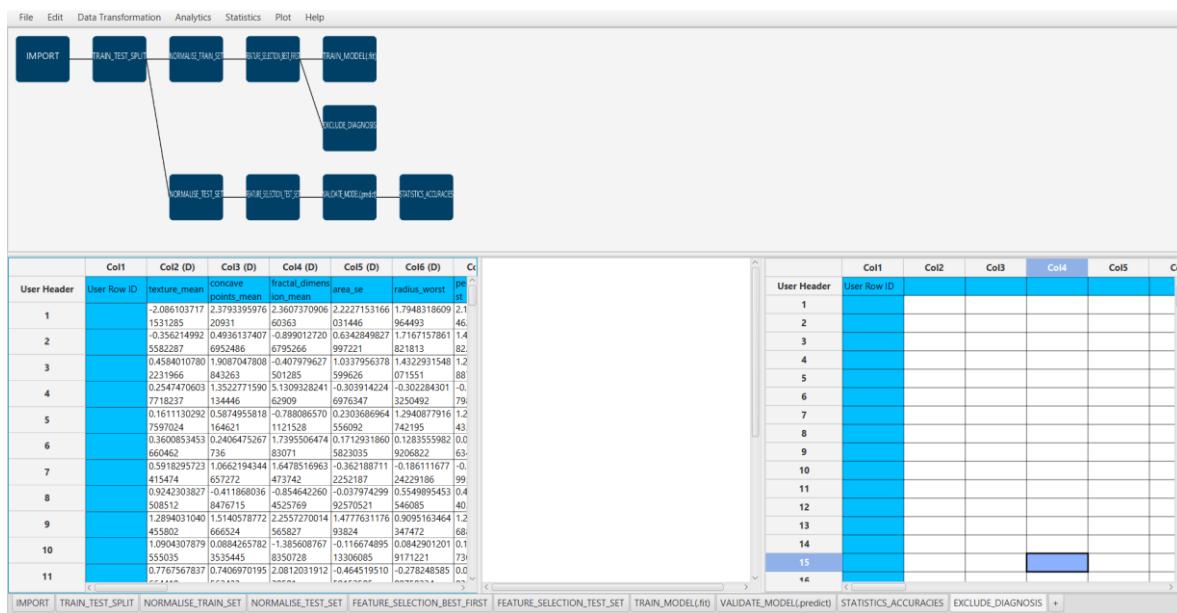


## Step 11: Reliability check of each record of the test set

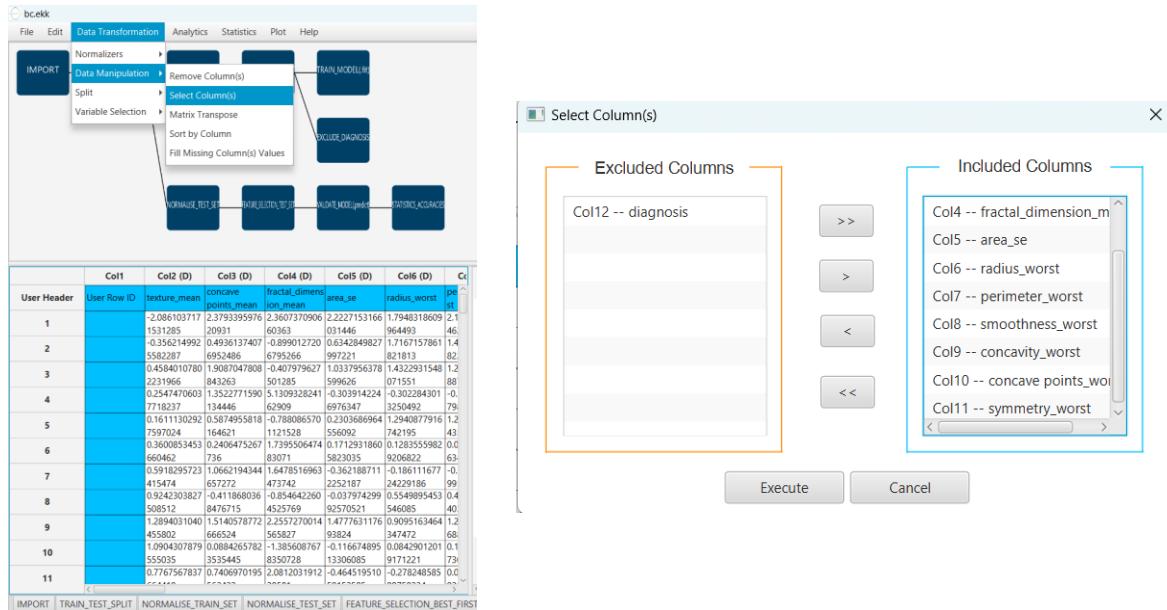
### Step 11.a: Create the domain

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

Import data into the input spreadsheet of the "EXCLUDE\_DIAGNOSIS" tab from the output of the "FEATURE\_SELECTION\_BEST\_FIRST" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



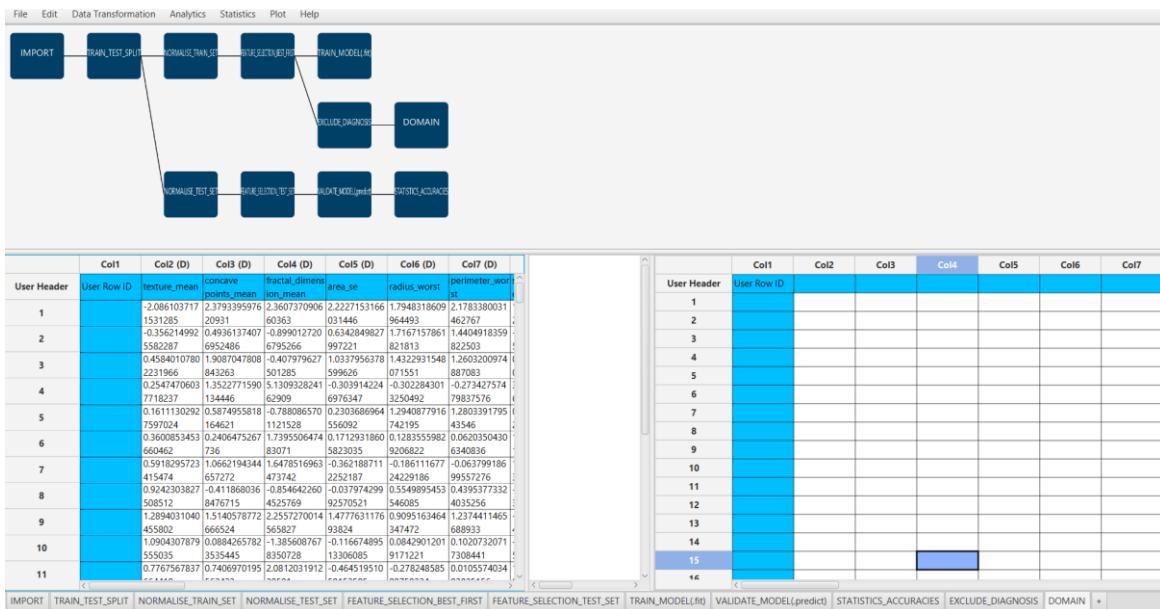
Manipulate the data to exclude the column that corresponds to the diagnosis by browsing: "Data Transformation" → "Data Manipulation" → "Select Column(s)". Then select all the columns except the "diagnosis".



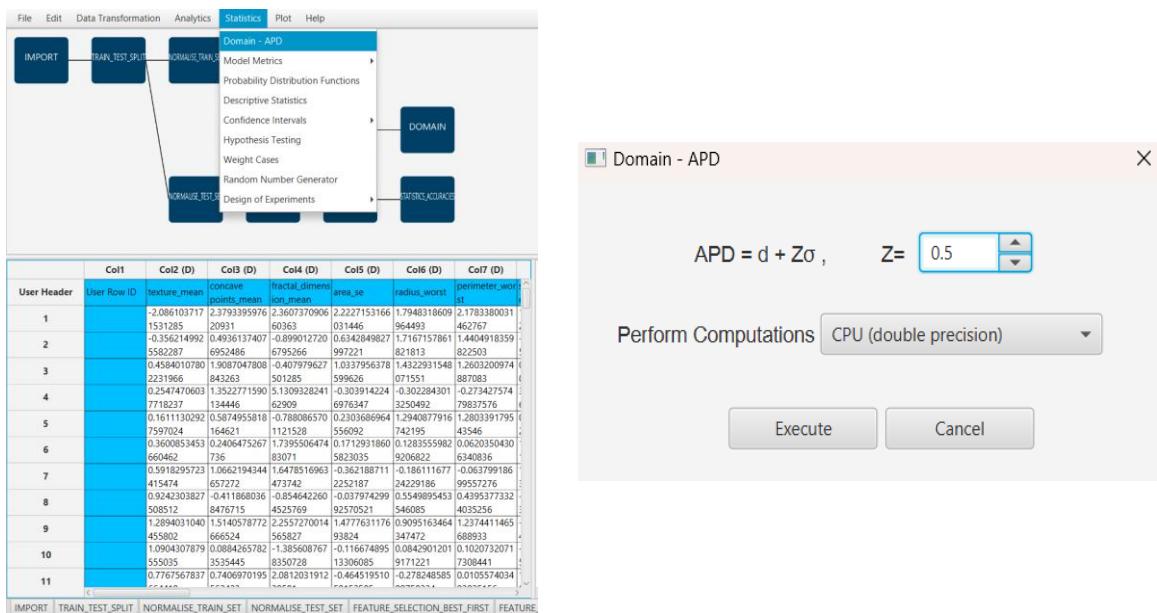
The results will appear on the output spreadsheet.

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

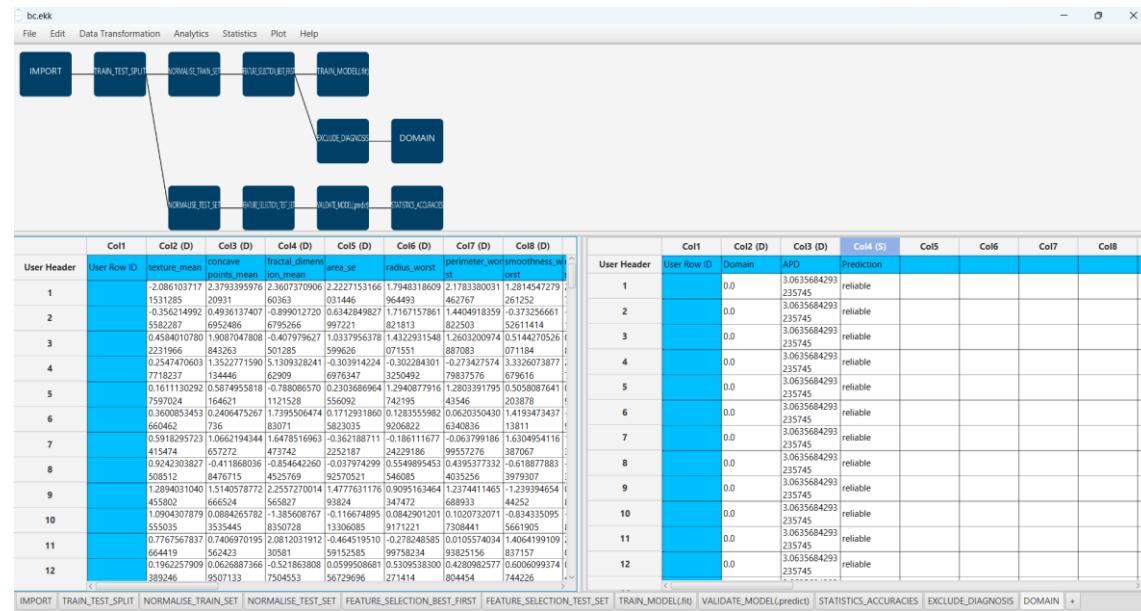
Import data into the input spreadsheet of the "DOMAIN" tab from the output of the "EXCLUDE\_DIAGNOSIS" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



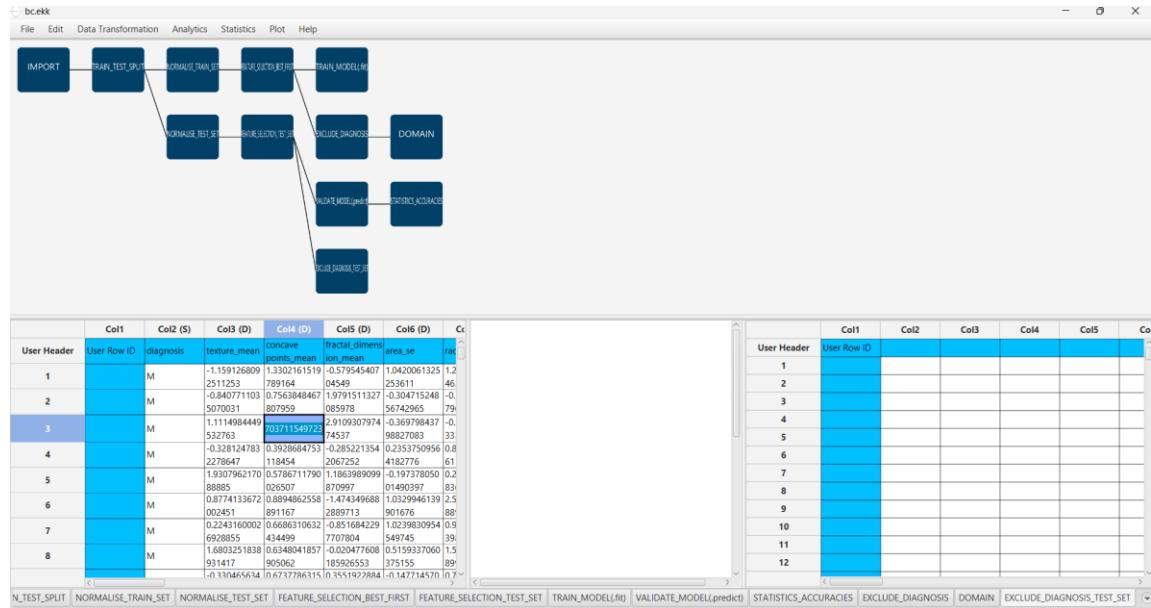
The results will appear on the output spreadsheet.



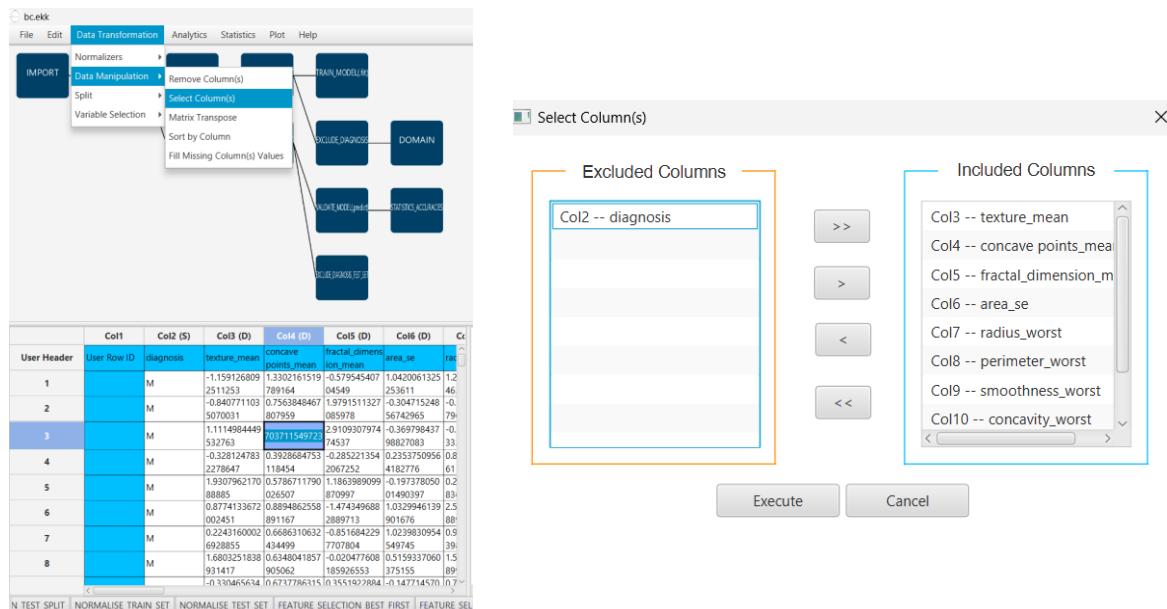
## Step 11.b: Check the test set reliability

Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE\_DIAGNOSIS\_TEST\_SET".

Import data into the input spreadsheet of the "EXCLUDE\_DIAGNOSIS\_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 diagnosis by browsing: "Data Transformation" → "Data Manipulation" → "Select Column(s)". Then select all the columns except diagnosis.



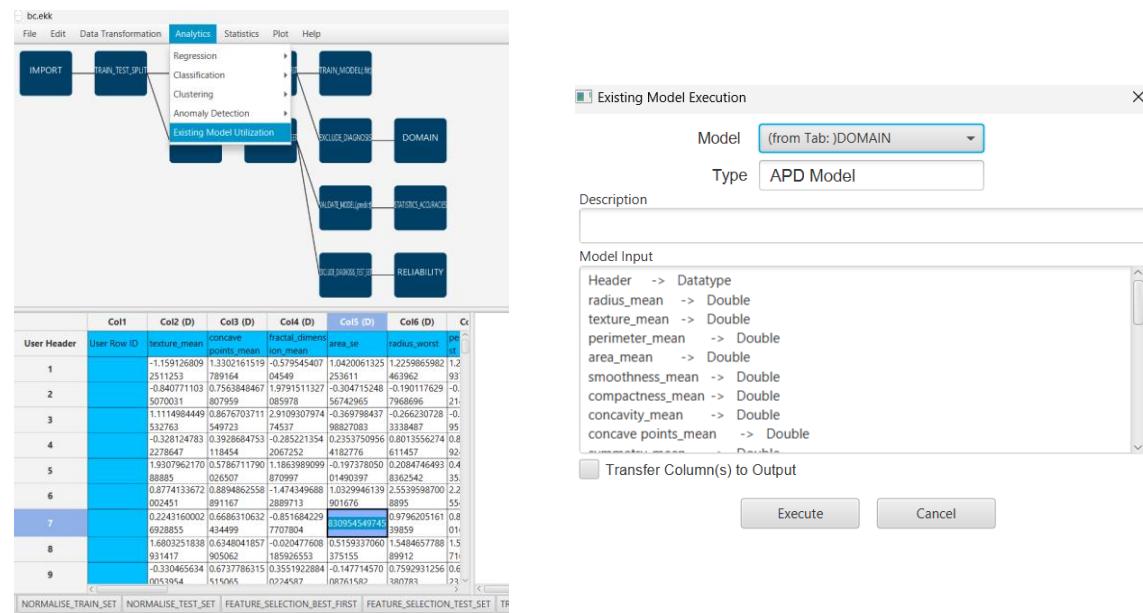
The results will appear on the output spreadsheet.

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

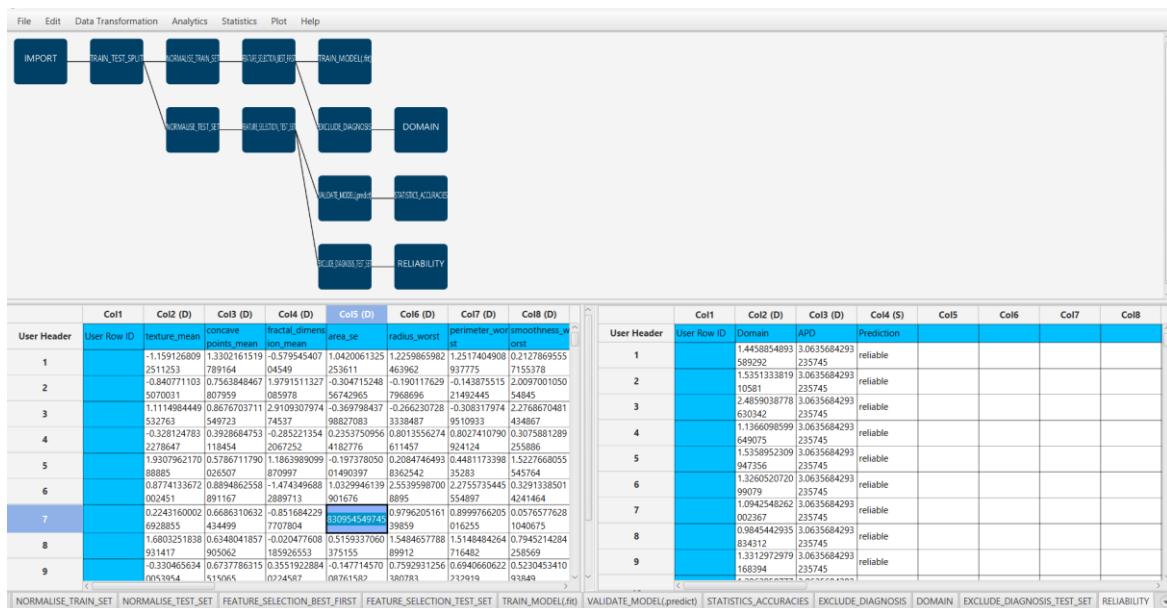
Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "EXCLUDE\_DIAGNOSIS\_TEST\_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



The results will appear on the output spreadsheet. There is one unreliable sample in the test set.



## Final Isalos Workflow

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

