



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 toolbar with various icons. Below the toolbar, a large table represents the dataset. The columns are labeled "Col1" through "Col6". The first row is labeled "User Header" and contains "User Row ID". Rows 1 through 6 are empty. Row 7 contains the first data point: "User Row ID", "id", "diagnosis", "radius_mean", "texture_mean", "perimeter_mean", "area_mean", "smoothness_mean", "compactness_mean", "concavity_mean", "concave_points_mean", and "symmetry_mean". A context menu is open over this row, with options: "Import from SpreadSheet", "Import from file", "Export Spread Sheet Data", and "Clear SpreadSheet". The bottom portion of the interface shows a detailed view of the first 21 rows of the dataset, with columns labeled "Col1" through "Col12(D)". The "User Header" row is also present at the top of this section. The "diagnosis" column shows values "M" (malignant) and "B" (benign). Other columns show numerical values for various breast cancer features like radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension.

User Header	User Row ID	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave_points_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	0.198	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	18.25	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	0.1185	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	846301	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	0.1995	0.1639	0.07364	0.2303	
17	848406	M	14.68	20.13	94.74	684.5	0.09867	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	8510552	B	13.08	15.71	85.63	452.1	0.1075	0.127	0.14568	0.03111	0.1967	

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)	Col10 (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	84300903	M	19.69	21.25	130	1203	0.1096	0.1599			
4	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839			
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	0.08206	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						

Import from SpreadSheet
Import from file
Export Spread Sheet Data
Clear SpreadSheet

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.

Random Partitioning

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

Random Partitioning Results

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	

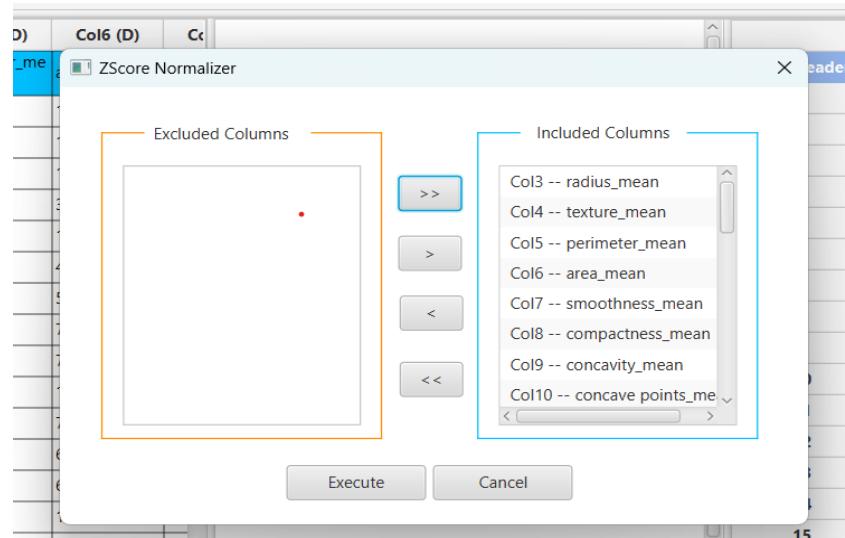
Step 4: Normalize the training set

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 the Isalos Analytics Platform interface. At the top, there are three tabs: "IMPORT", "TRAIN_TEST_SPLIT", and "NORMALISE_TRAIN_SET". The "TRAIN_TEST_SPLIT" tab is currently selected, showing a table with 21 rows of data. The columns are labeled "User Header", "User Row ID", "diagnosis", "radius_mean", "texture_mean", "perimeter_mean", "area_mean", "smoothness_mean", "compactness_mean", "concavity_mean", and "concave points_mean". The data consists of mostly 'M' values (Malignant) with some 'B' values (Benign). The "NORMALISE_TRAIN_SET" tab is also visible at the bottom.

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



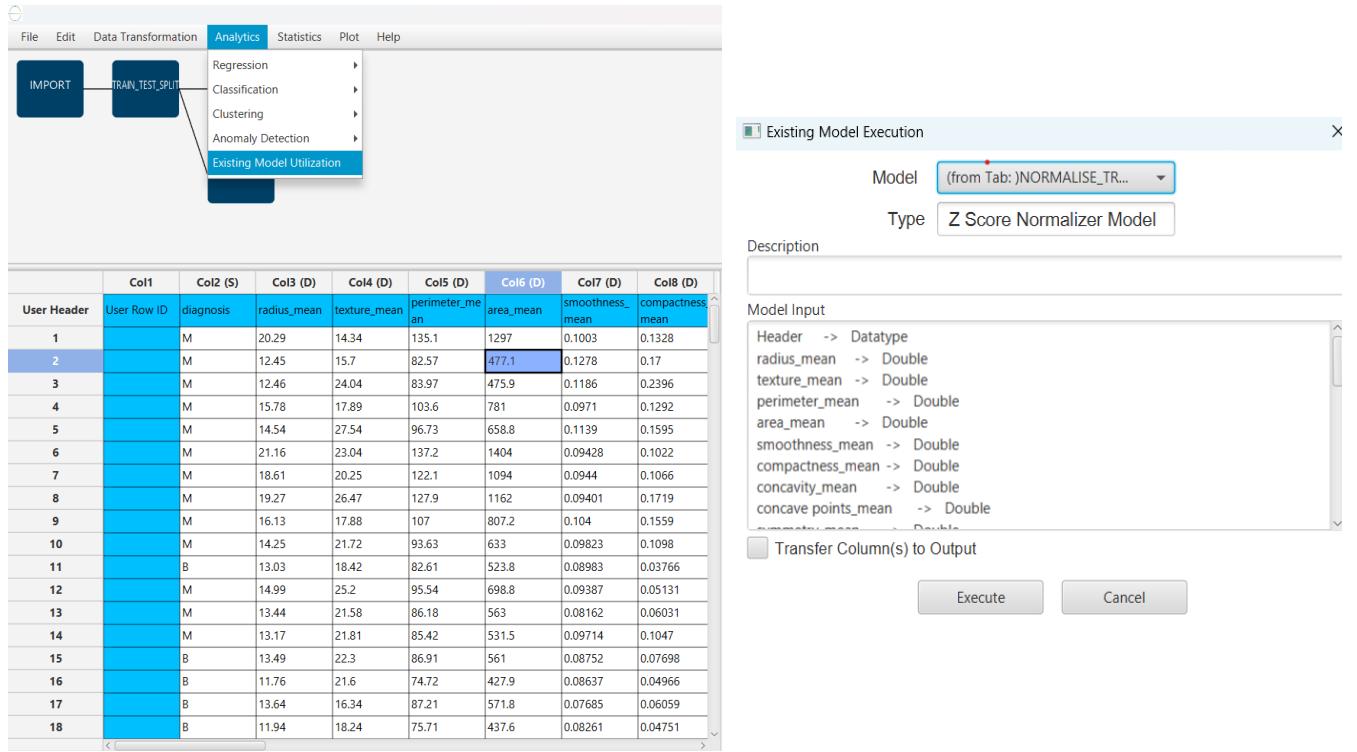
The results will appear on the output spreadsheet.

Step 5: Normalize the test set

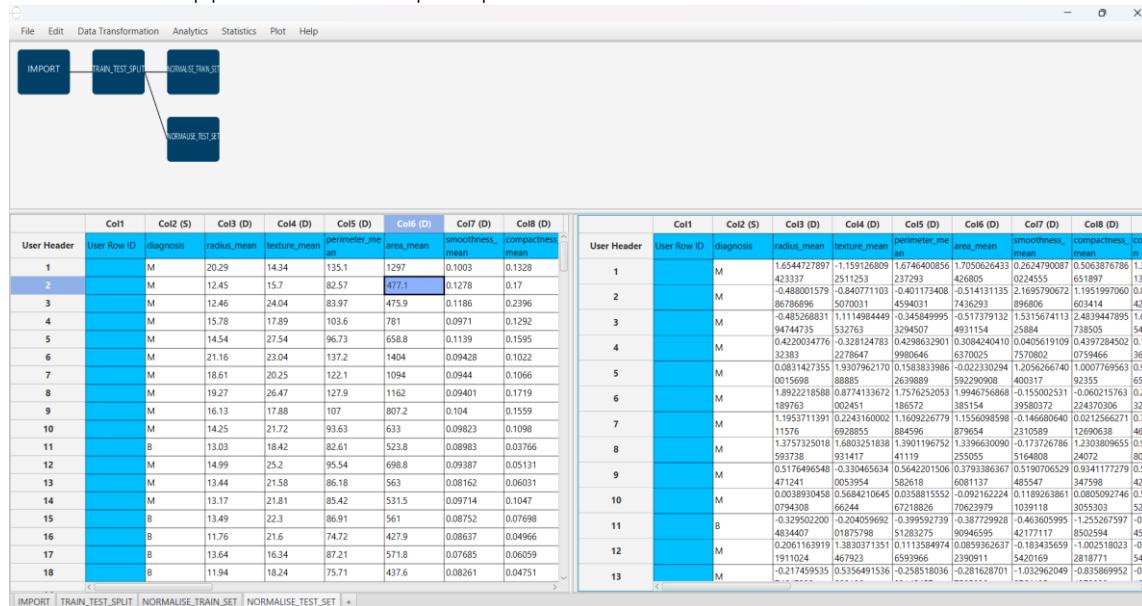
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALISE_TEST_SET".

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".

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



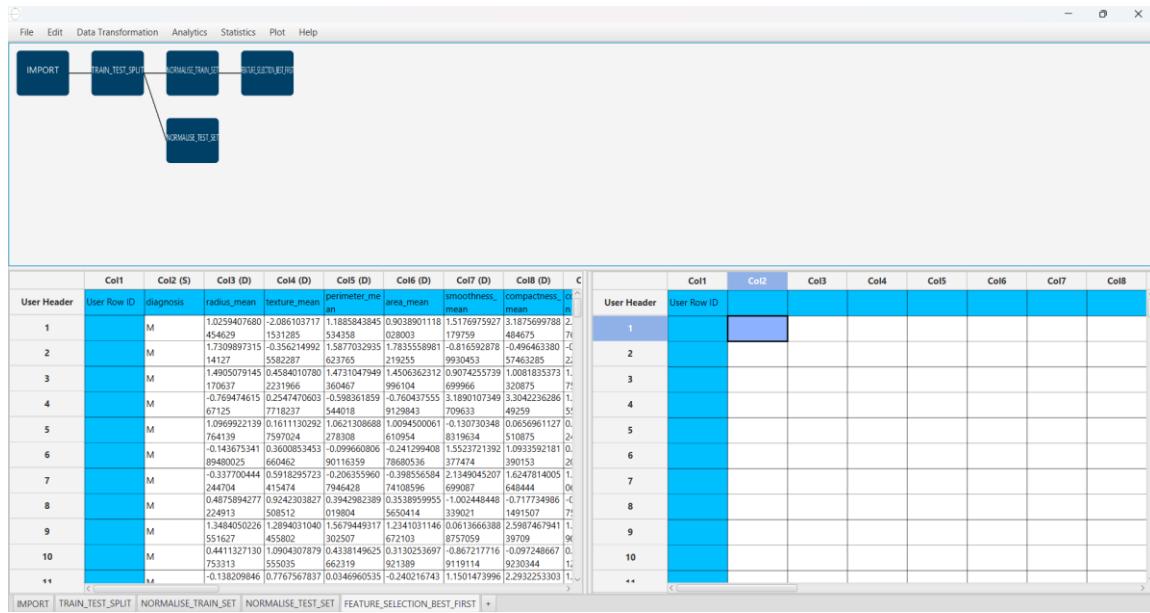
The results will appear on the output spreadsheet.



Step 6: Feature selection

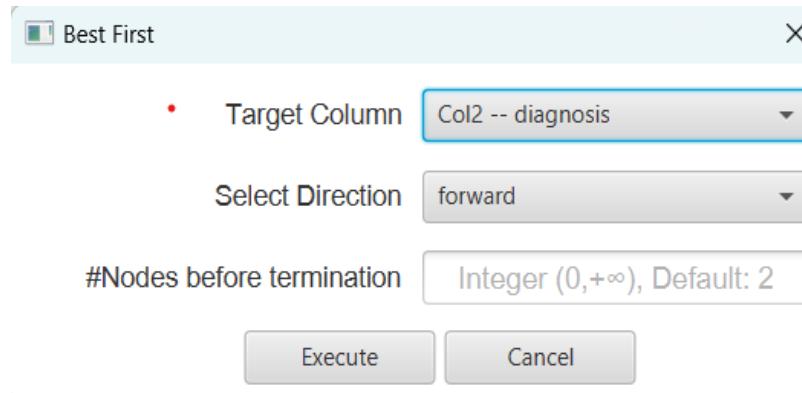
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_BEST_FIRST".

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".

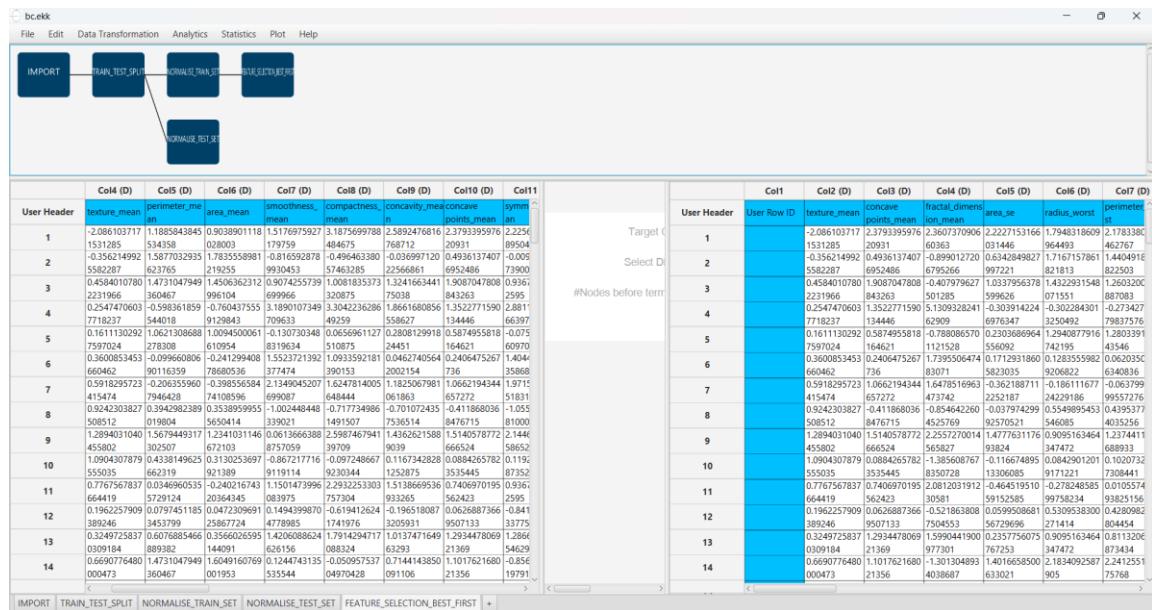


Choose the most important features for the classification using the Best First Function: "Data Transformation" → "Variable Selection" → "Best First".

Then choose the "diagnosis" column as the target variable and the direction as forward.



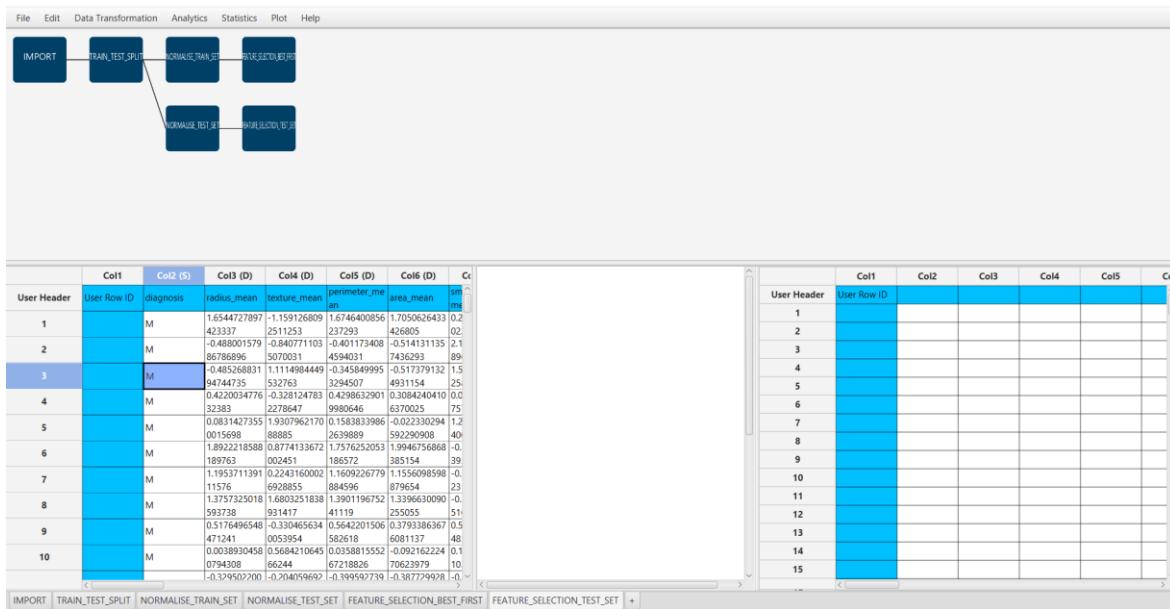
The results will appear on the output spreadsheet.



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 Orange data mining interface. The 'Data Transformation' menu is open, with 'Select Column(s)' highlighted. The 'Select Column(s)' dialog is displayed, showing the 'Excluded Columns' and 'Included Columns' lists.

Excluded Columns

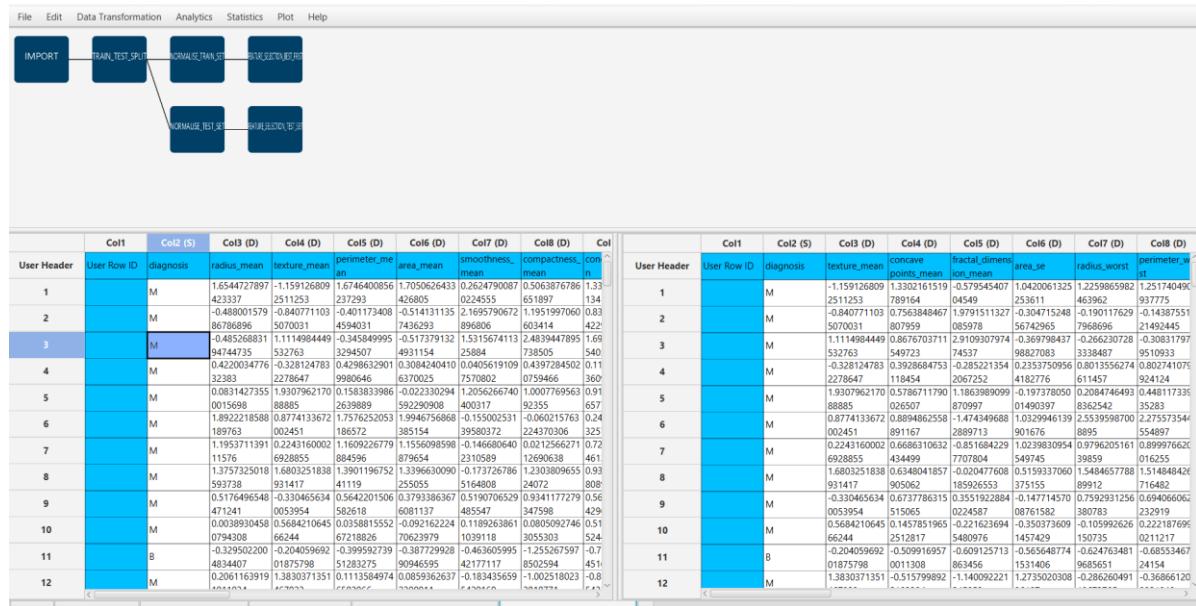
- Col19 -- concavity_se
- Col20 -- concave points_se
- Col21 -- symmetry_se
- Col22 -- fractal_dimension_s
- Col24 -- texture_worst
- Col26 -- area_worst
- Col28 -- compactness_worst
- Col32 -- fractal_dimension_w

Included Columns

- Col12 -- fractal_dimension_r
- Col16 -- area_se
- Col23 -- radius_worst
- Col25 -- perimeter_worst
- Col27 -- smoothness_worst
- Col29 -- concavity_worst
- Col30 -- concave points_w
- Col31 -- symmetry_worst

Buttons: >>, >, <, <<, Execute, Cancel.

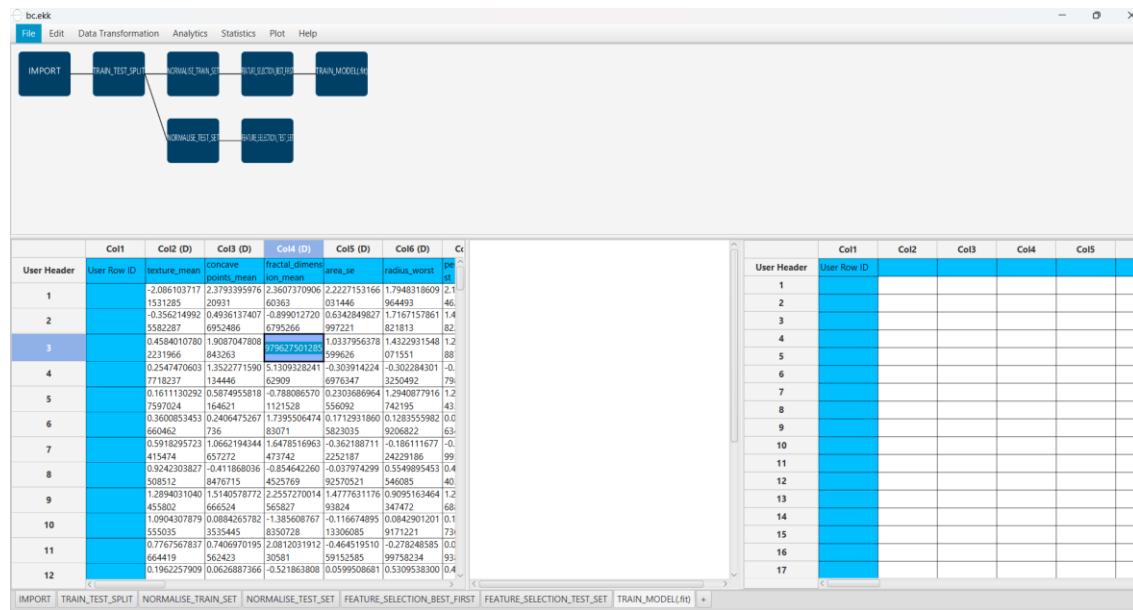
The results will appear on the output spreadsheet.



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".

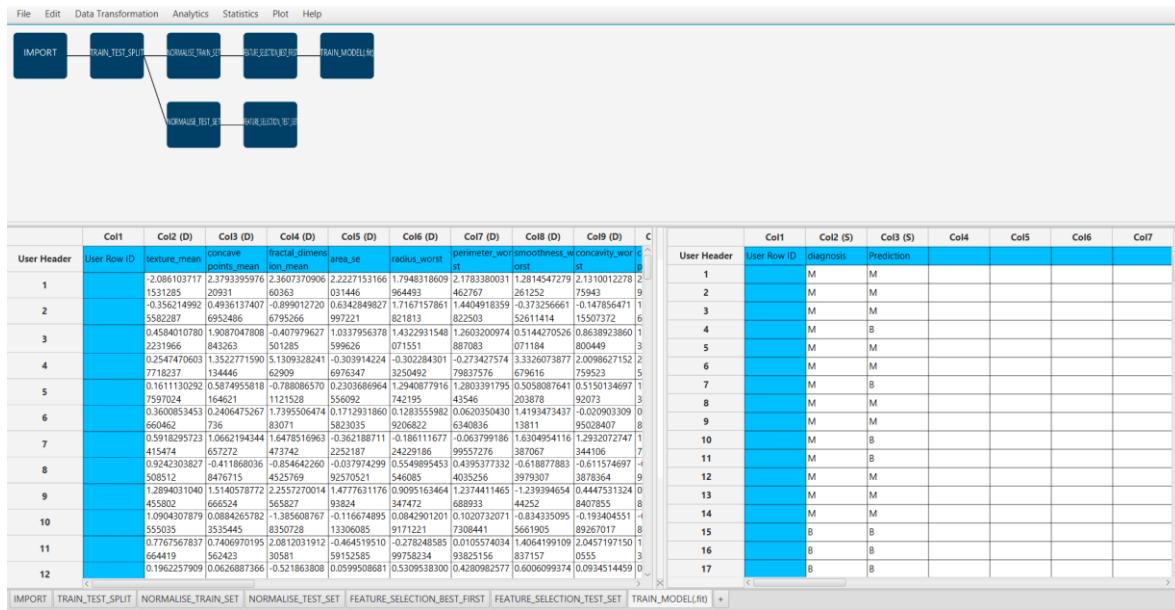


Use the Random Forest Method to train and fit the model: "Analytics" → "Classification" → "Random Forest" and adjust the model parameters based on training set performance.

The screenshot shows the classification process and model configuration:

- Classification Menu:** Under the "Analytics" tab, the "Classification" option is selected, revealing sub-options: k Nearest Neighbors (kNN), Multiple Layer Perceptron (MLP), Radial Basis Function (RBF), XGBoost, J48, and Random Forest.
- Model Configuration Dialog:** A modal dialog titled "Random Forest Classification Model" is open, containing the following settings:
 - Features fraction: 0.9
 - Min impurity decrease: 0.1
 - Time-based RNG Seed: 1234
 - Seed: 1234
 - Number of ensembles: 610
 - Target column: Col12 -- diagnosis
- Data Preview Table:** Below the dialog is a table showing the same 12 rows of data as the previous screenshot, with columns Col1 through Col12.
- Toolbar:** At the bottom of the interface are buttons: IMPORT, TRAIN_TEST_SPLIT, NORMALISE_TRAIN_SET, NORMALISE_TEST_SET, and FEATURE_SELECTION_BEST_FIRST.

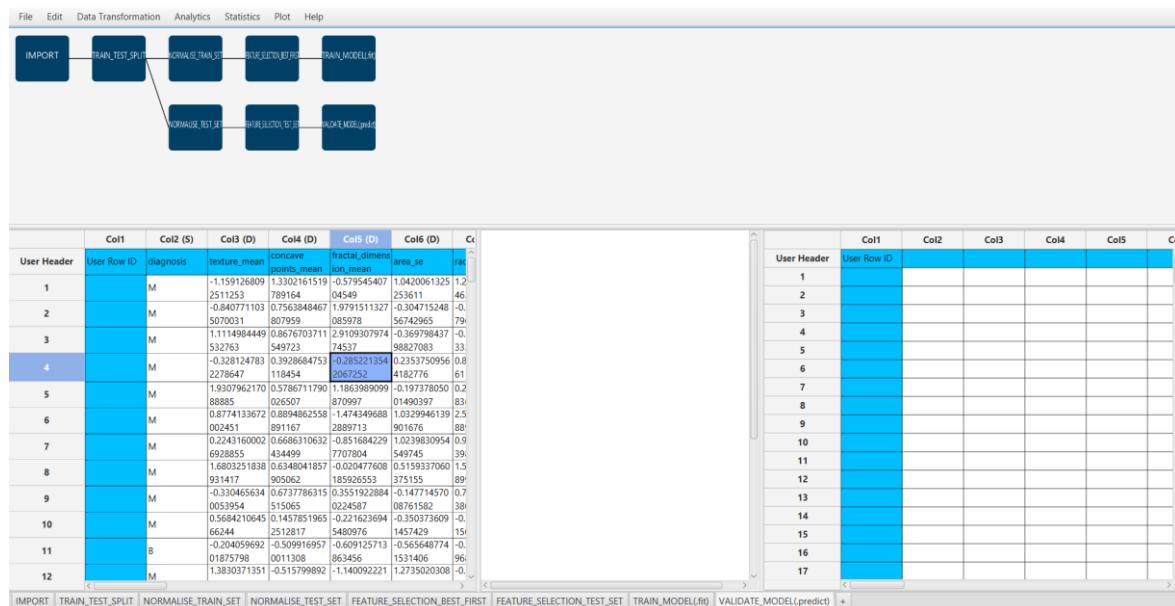
The predictions will appear on the output spreadsheet.



Step 9: Validate the model

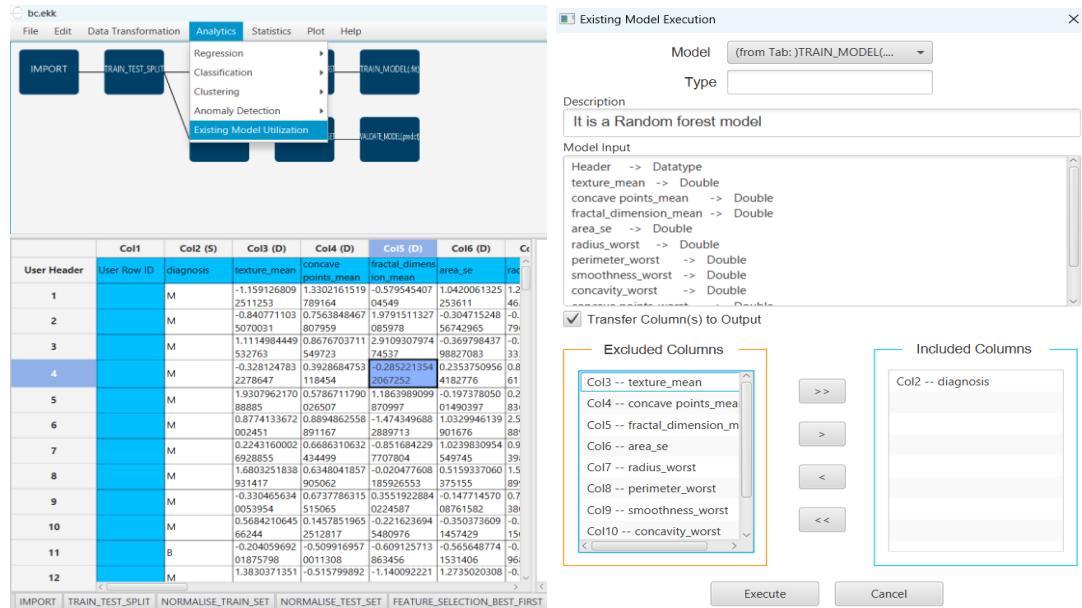
Create a new tab by pressing the "+" button on the bottom of the page with the name "VALIDATE_MODEL(.predict)".

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".

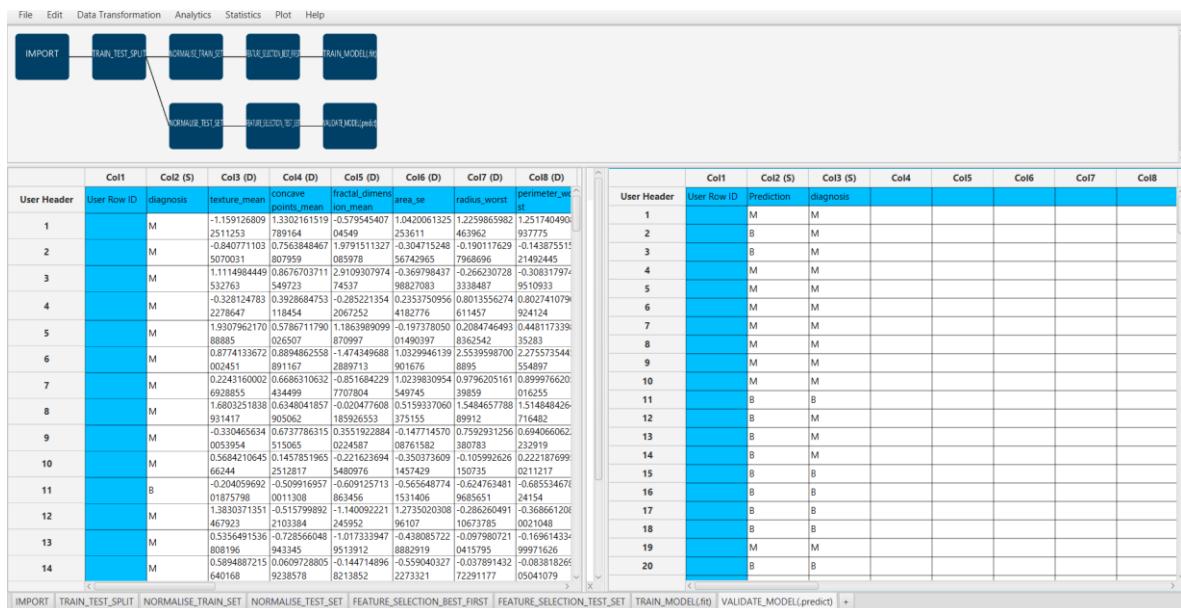


To validate the model:

"Analytics" → "Existing Model Utilization". Then choose Model "(from Tab:) TRAIN_MODEL (.fit)".



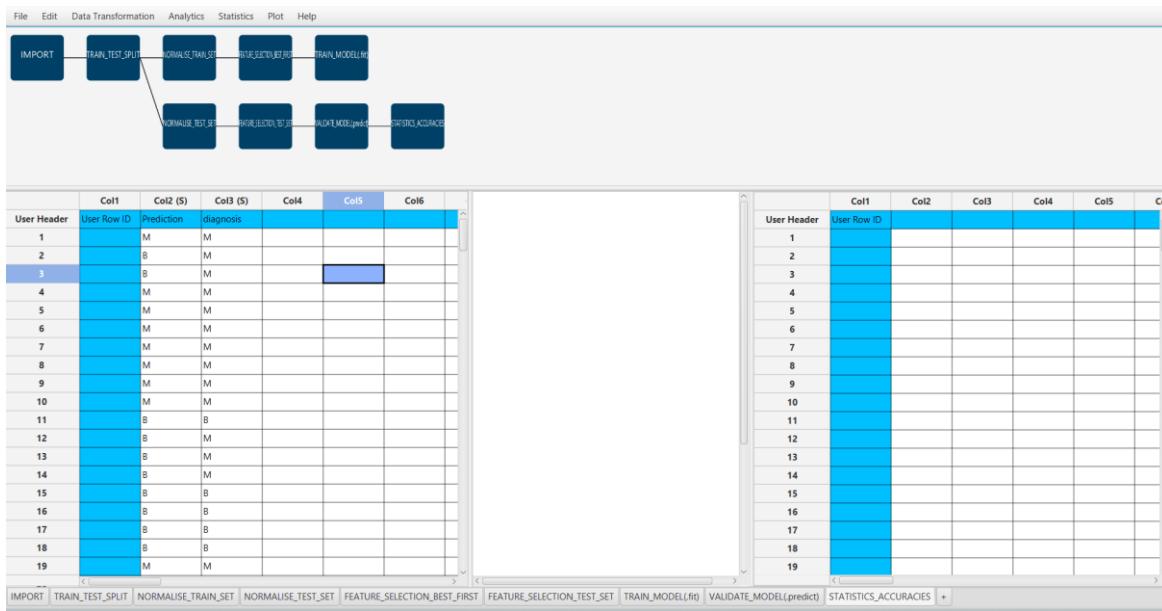
The predictions will appear on the output spreadsheet.



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: "Statistics" → "Model Metrics" → "Classification Metrics".

The screenshot shows a data flow diagram on the left and a dialog box on the right.

Data Flow Diagram:

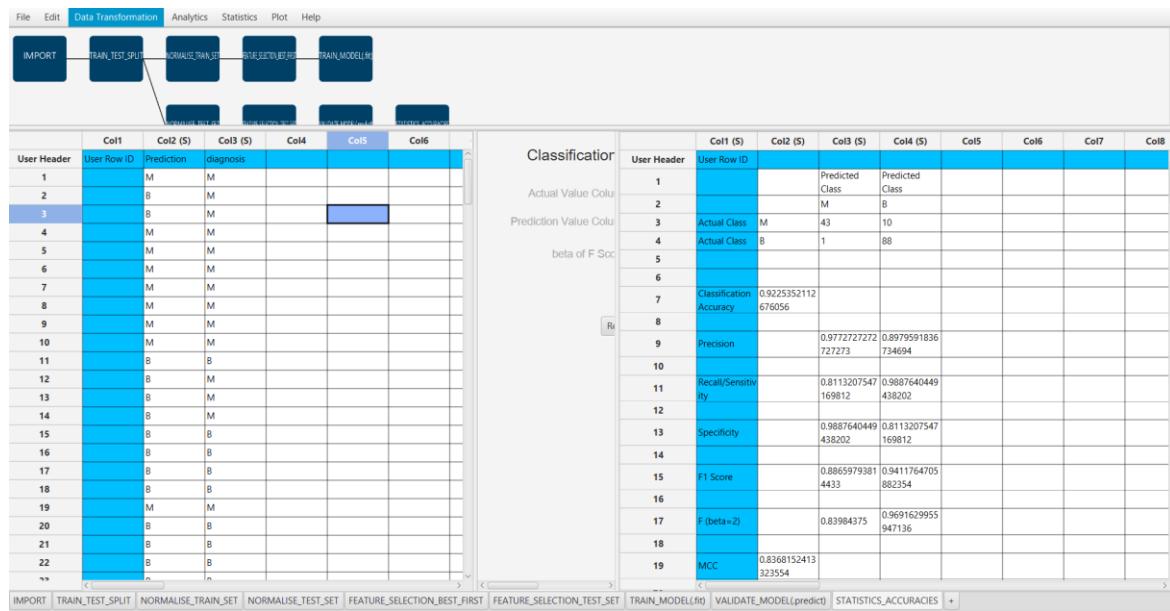
Classification Statistics Metrics Dialog:

The dialog box is titled "Classification Statistics Metrics". It contains the following fields:

- Actual Value Column: Col3 -- diagnosis
- Prediction Value Column: Col2 -- Prediction
- beta of F Score: 2

Buttons at the bottom: Execute and Cancel.

The results will appear on the output spreadsheet.



Accuracy: 0.923

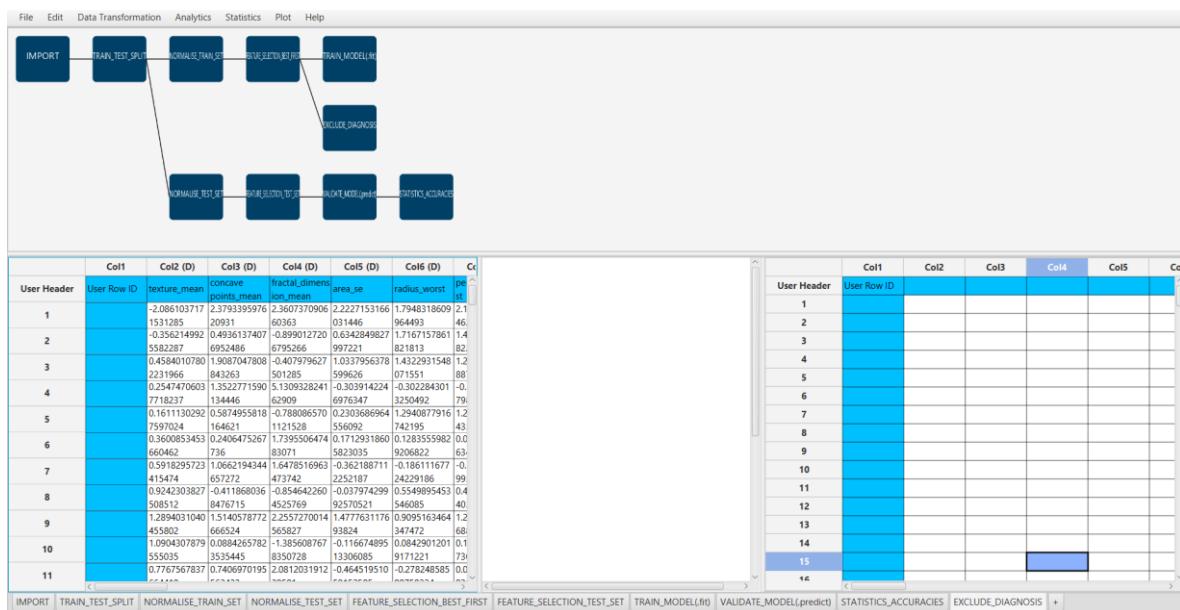
F1-Score = 0.914

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

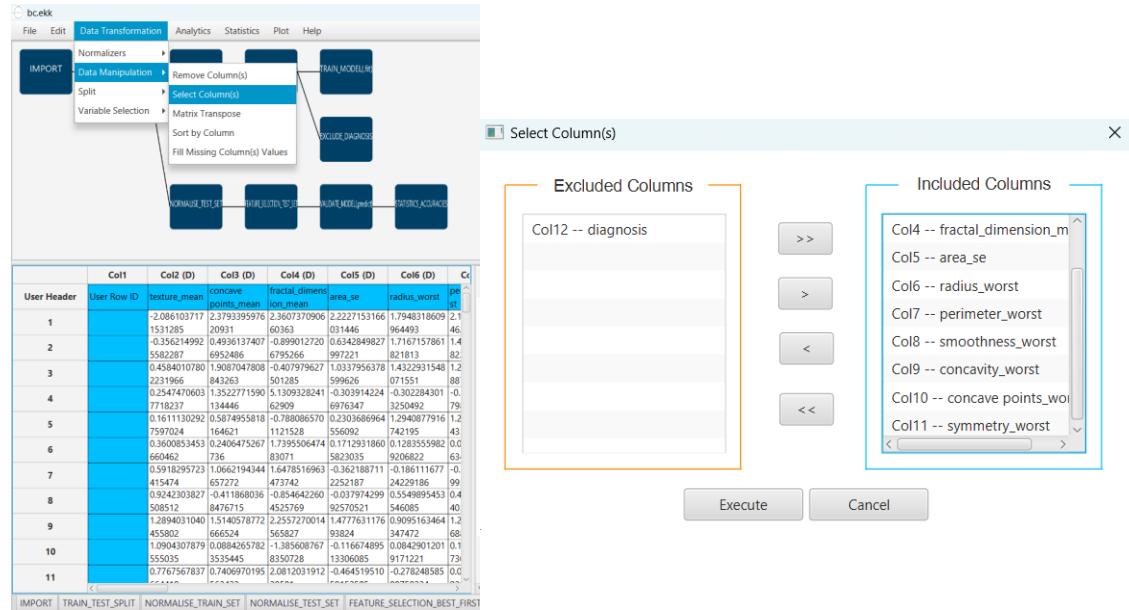
Step 11.a: Create the domain

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".



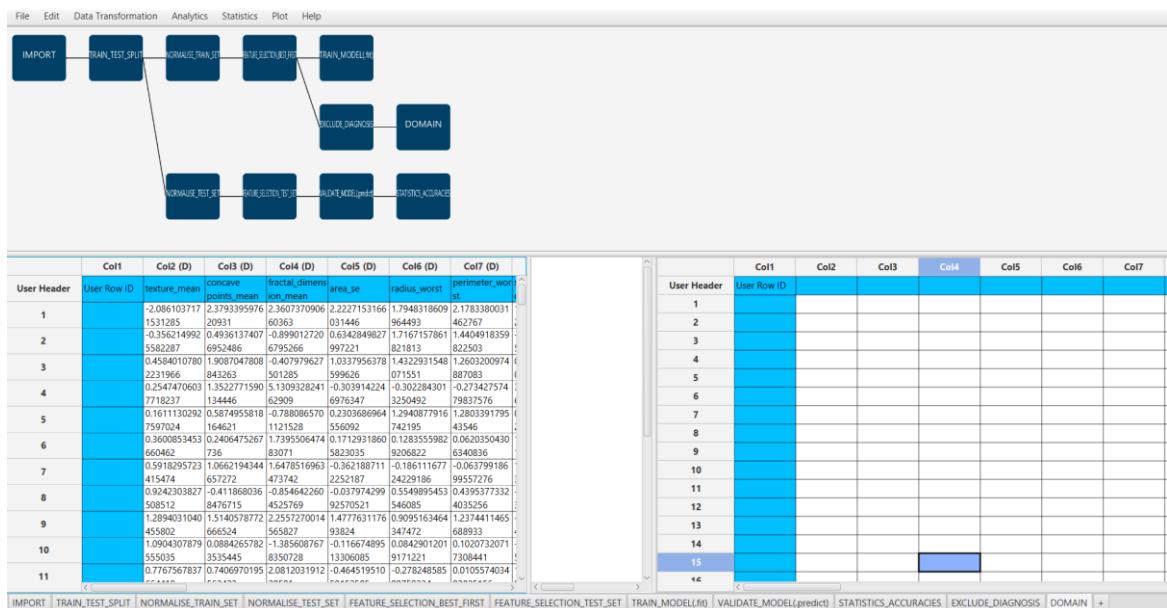
Manipulate the data to exclude the column that corresponds to the diagnosis "Data Transformation" → "Data Manipulation" → "Select Column(s)". Then select all the columns except the "diagnosis".



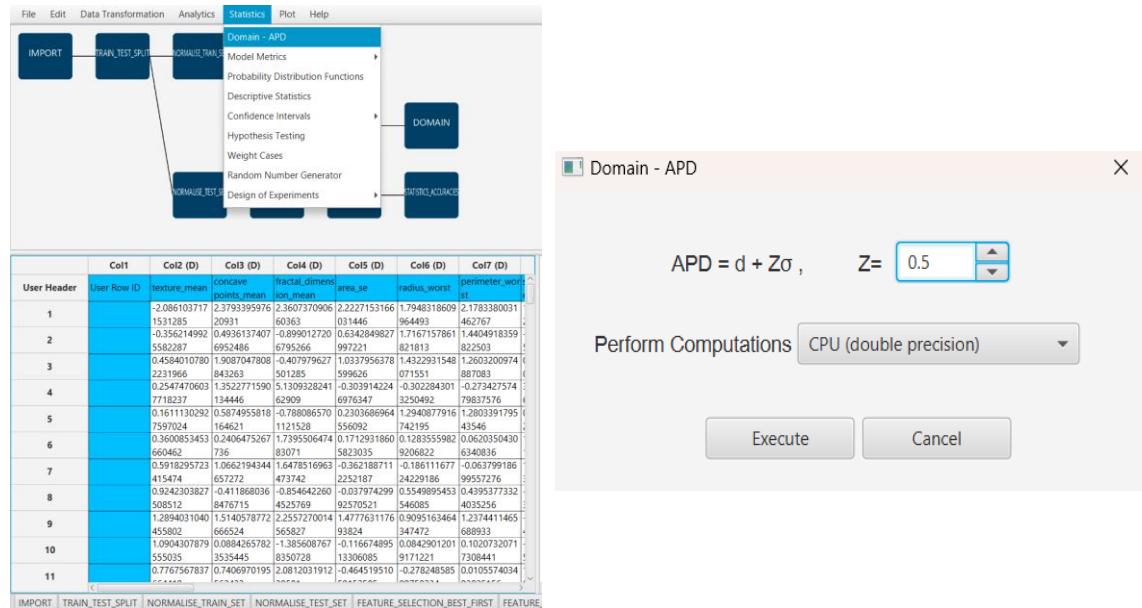
The results will appear on the output spreadsheet.

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

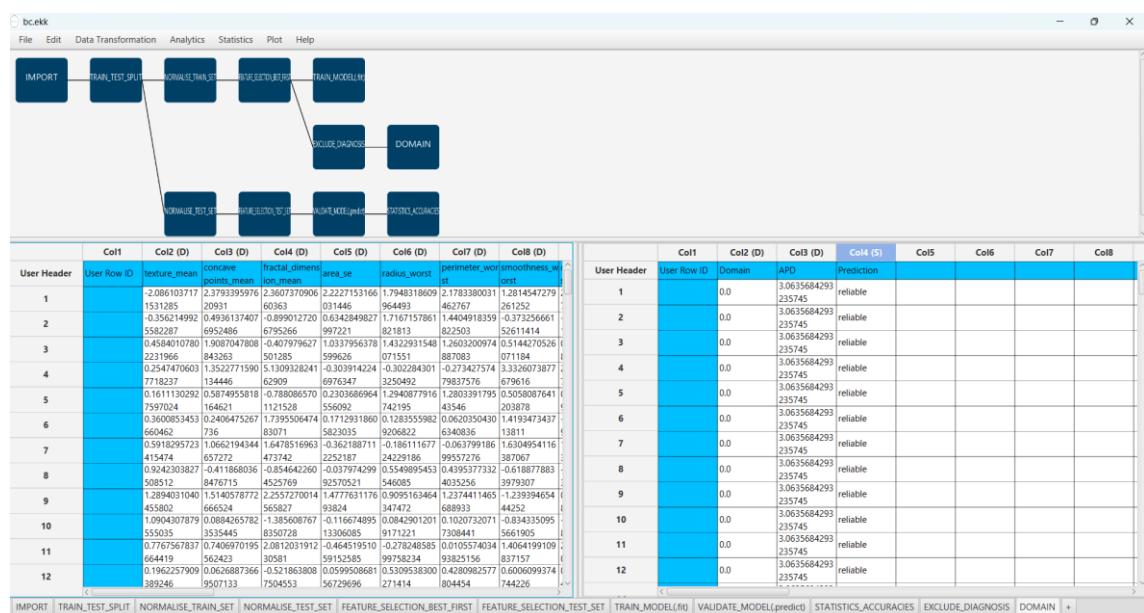
Import data into the input spreadsheet of the "DOMAIN" tab from the output of the "EXCLUDE_DIAGNOSIS" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



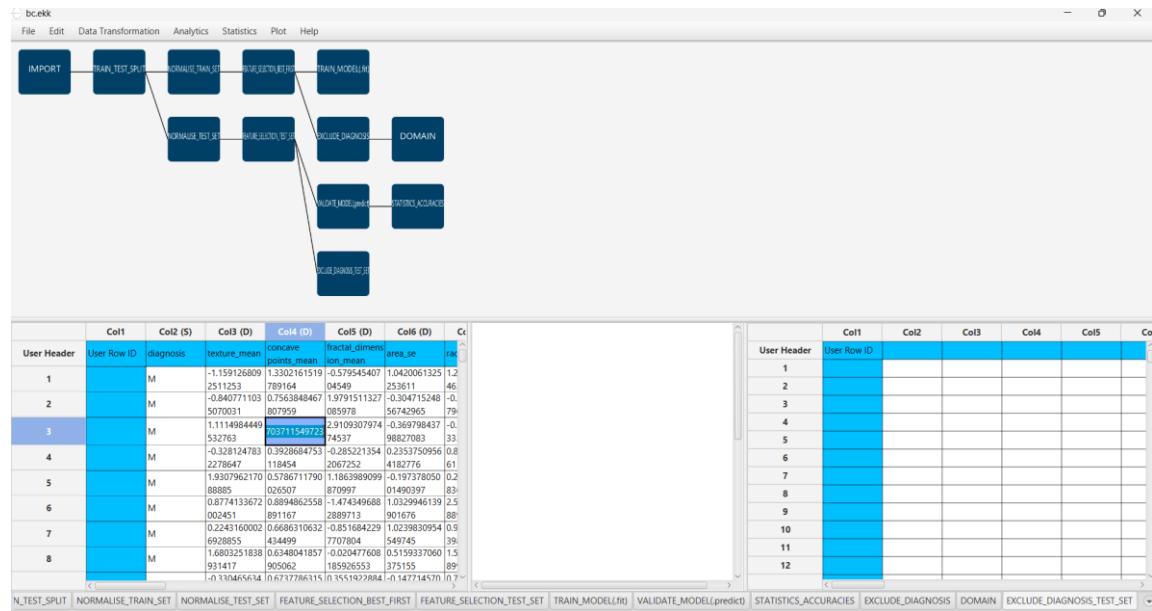
The results will appear on the output spreadsheet.



Step 11.b: Check the test set reliability

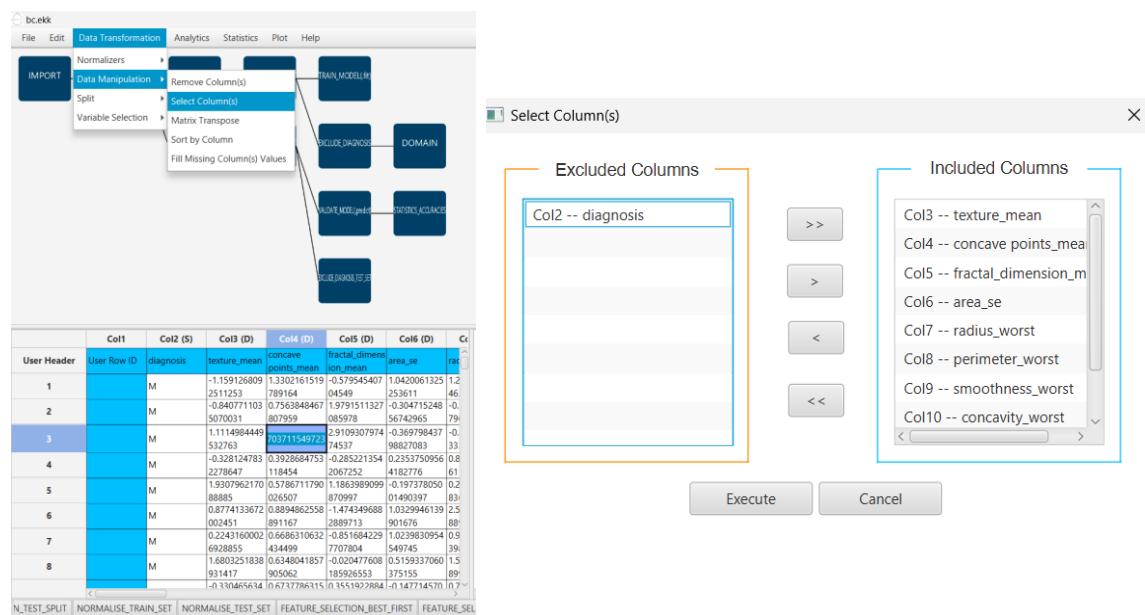
Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_DIAGNOSIS_TEST_SET".

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

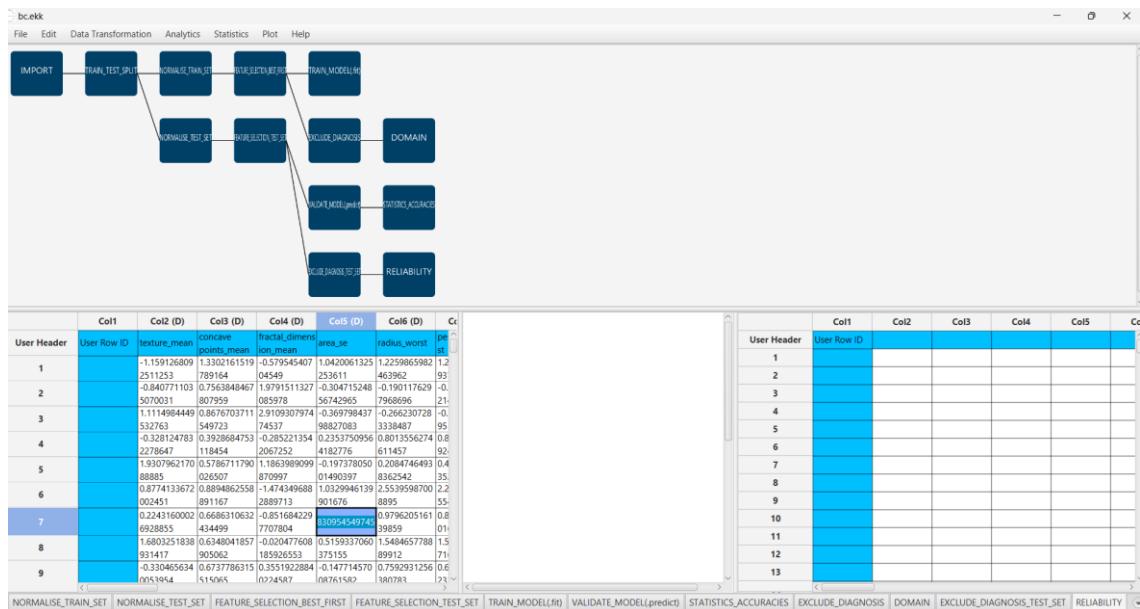
"Data Transformation" → "Data Manipulation" → "Select Column(s)". Then select all the columns except diagnosis.



The results will appear on the output spreadsheet.

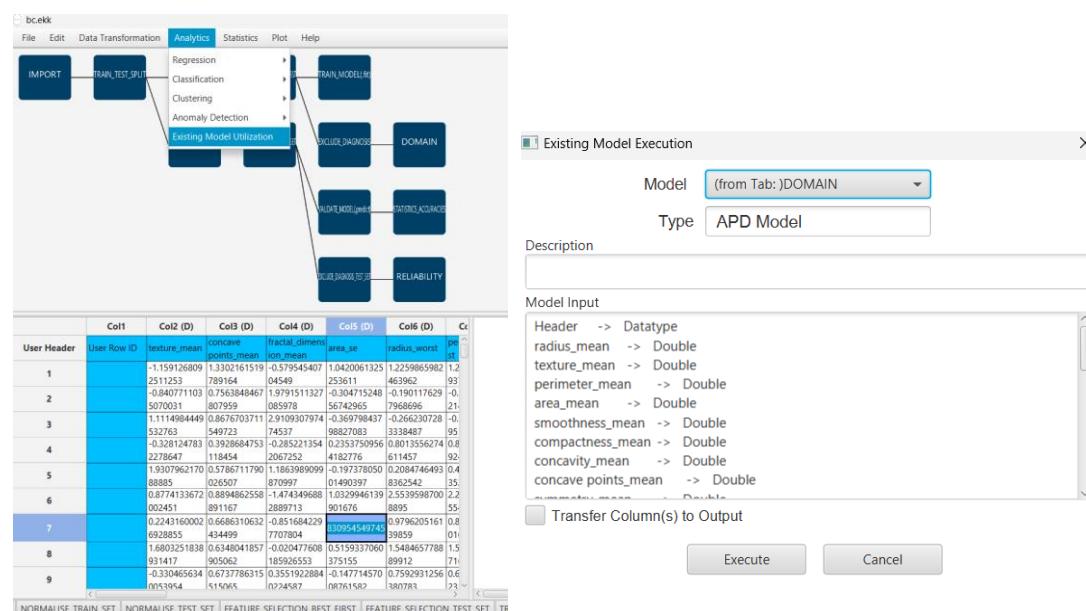
Create a new tab by pressing the "+" button on the bottom of the page with the name "RELIABILITY".

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".

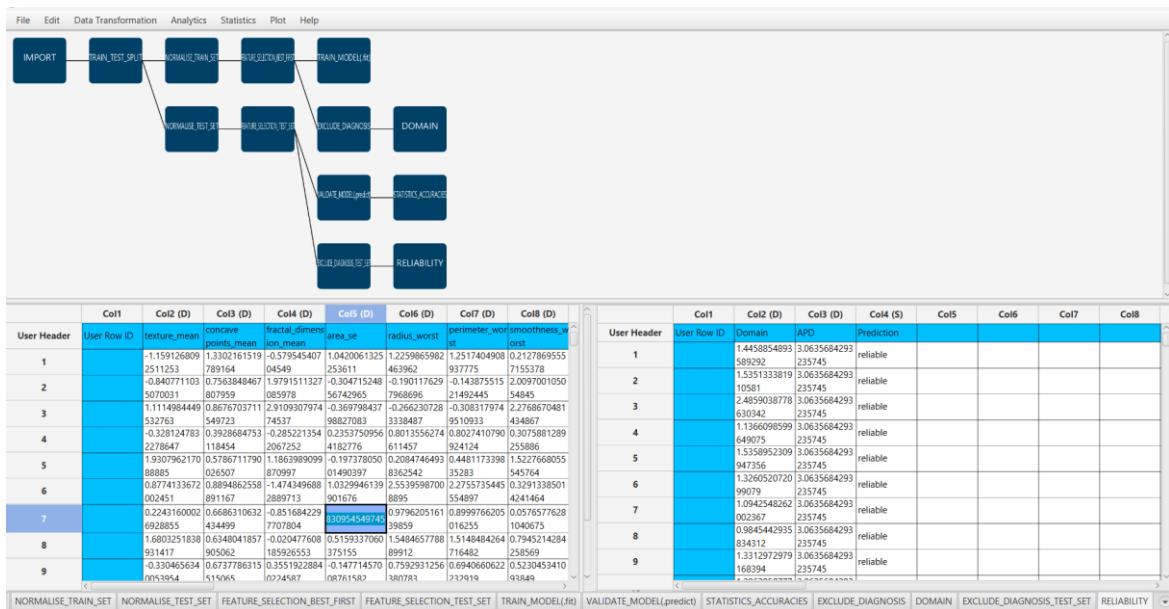


Check the Reliability:

"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:

