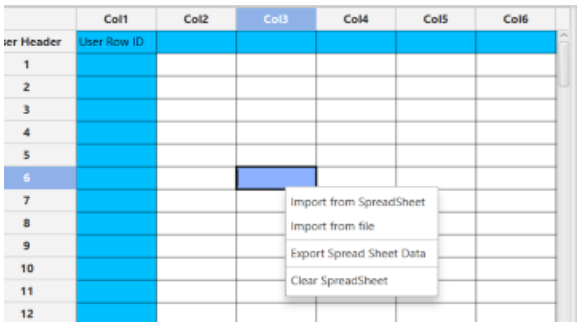
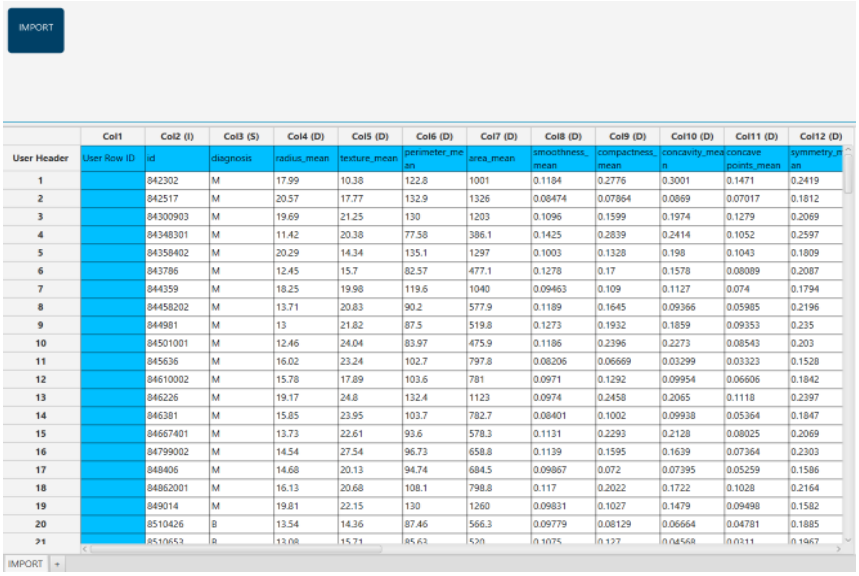
Breast Cancer (Binary Classification)

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.





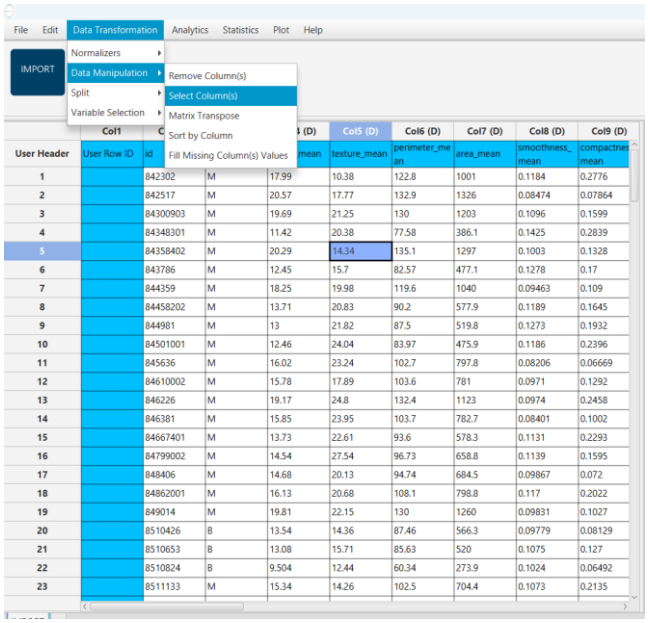
Step 2: Manipulate data

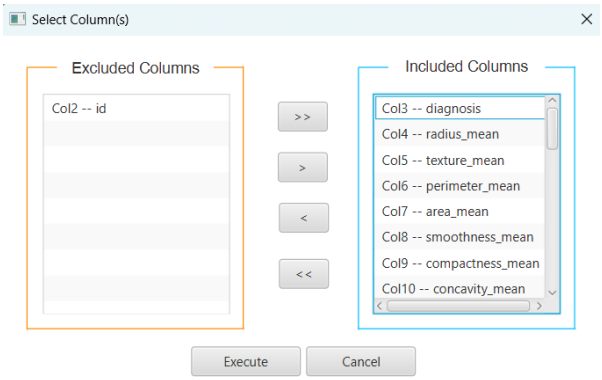
In our Dataset there are not empty values, and the only categorical feature is the label ("Diagnosis") which has two categories and the number of samples in each category are:

- 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.



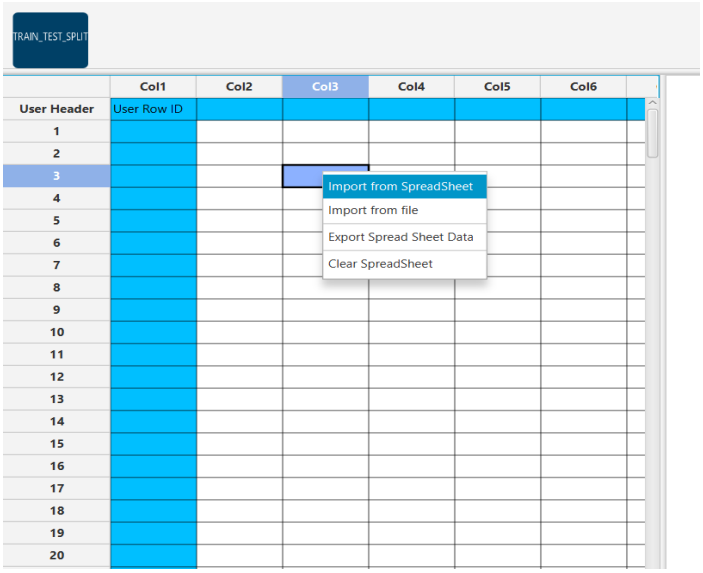


The data without the "id" column will appear in the output spreadsheet.

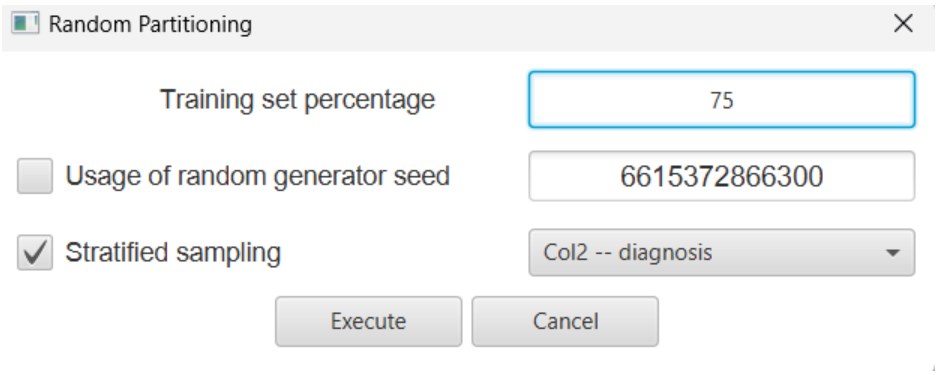
Step 3: Split data

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.

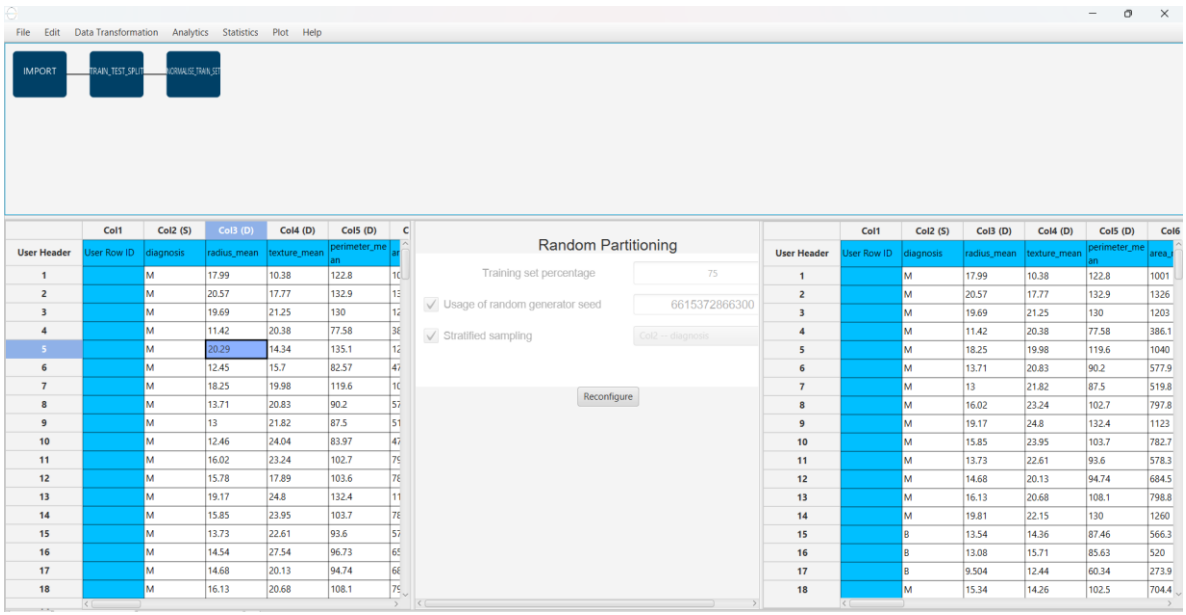
Import data into the input spreadsheet of the "TRAIN_TEST_SPLIT" tab from the output of the "IMPORT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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:



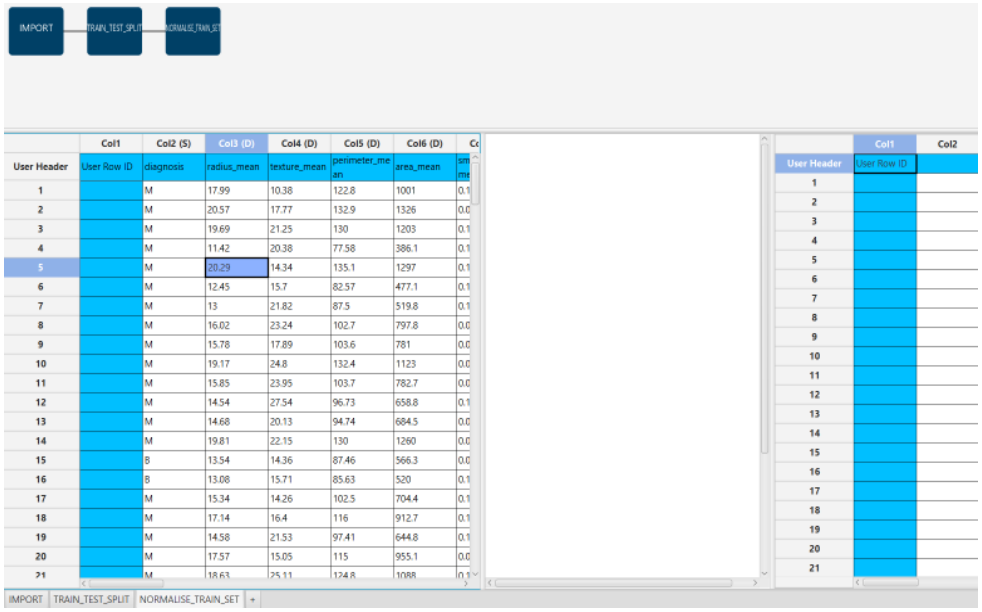
The results will appear on the output spreadsheet.



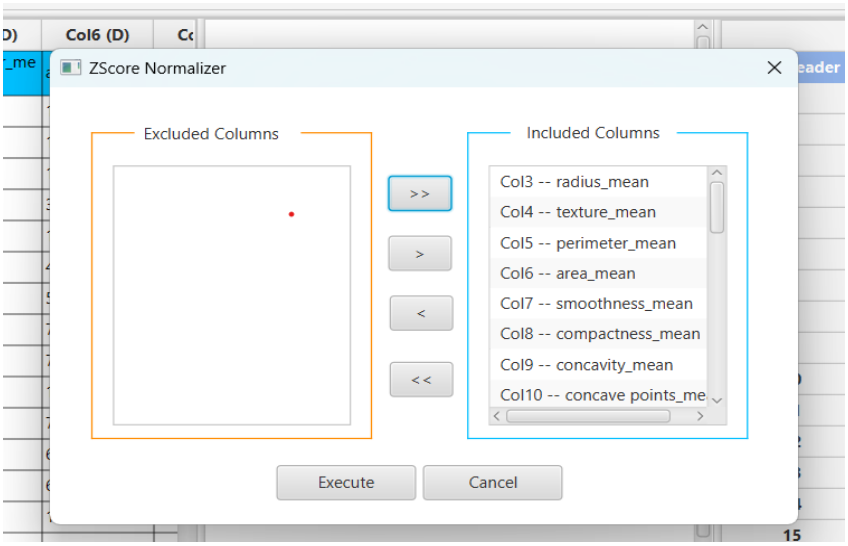
Step 4: Normalize the training set

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

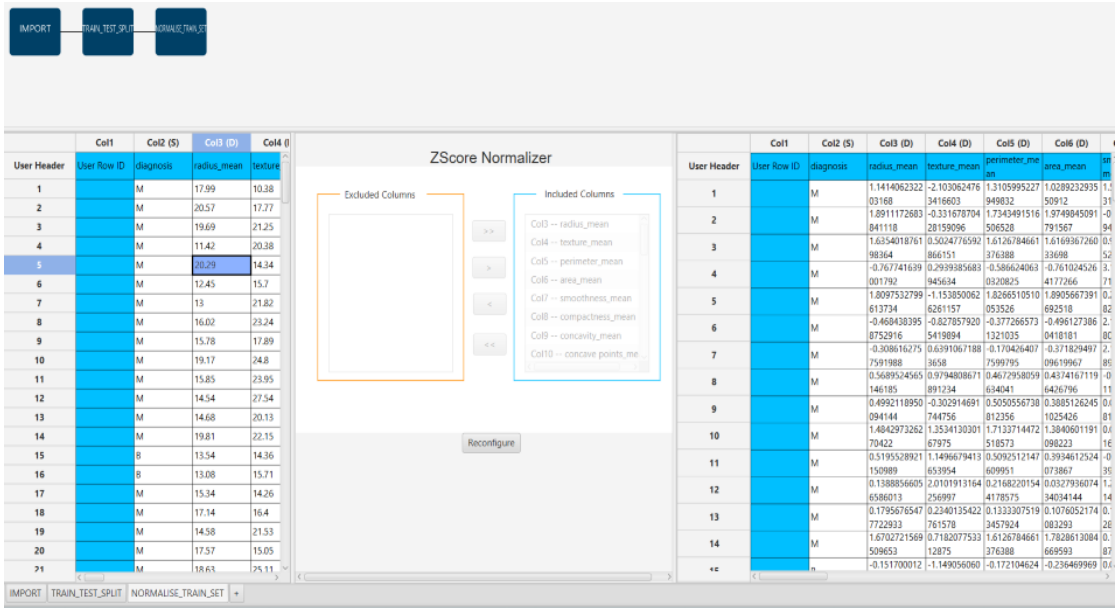
Import data into the input spreadsheet of the "NORMALISE_TRAIN_SET" tab the train set from the output of the "TRAIN_TEST_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet". From the available Select input tab options choose "TRAIN_TEST_SPLIT: Training Set"



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



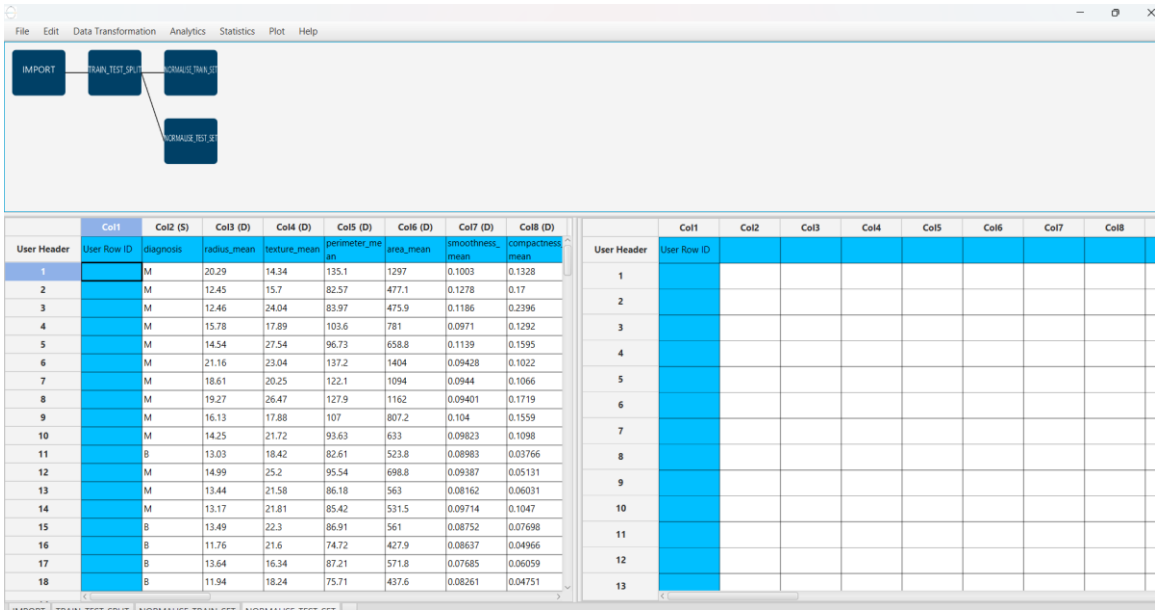
The results will appear on the output spreadsheet.



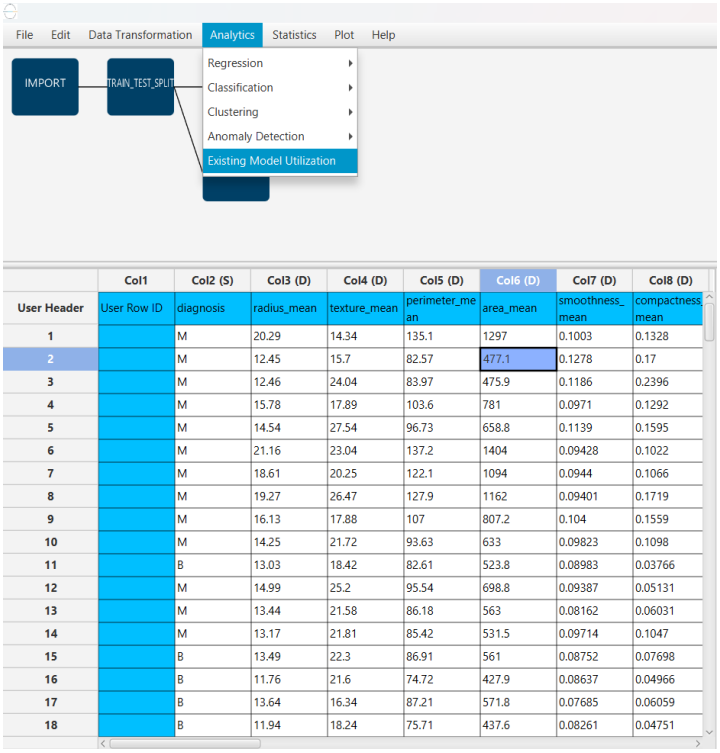
Step 5: Normalize the test set

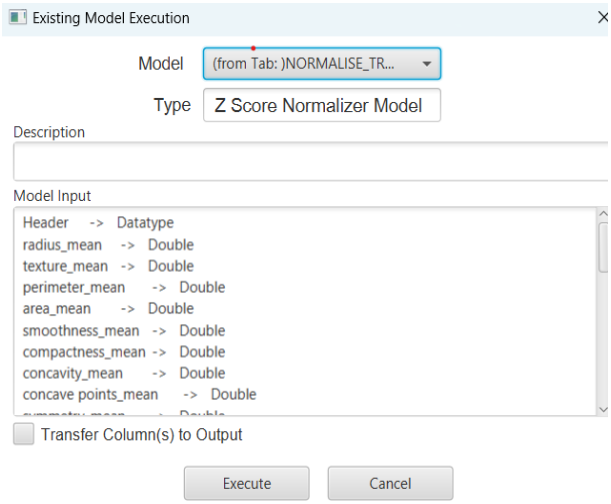
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALISE_TEST_SET".

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".



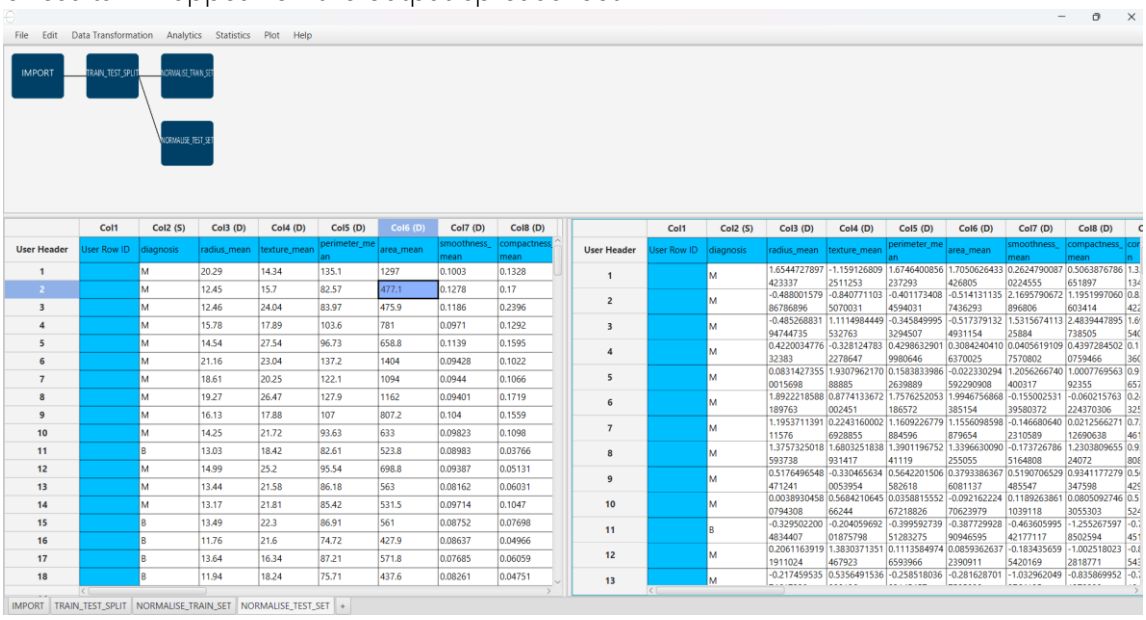
Normalize the test set using the existing normalizer of the training set: "Analytics" → "Existing Model Utilization" → "Model (from Tab:) NORMALISE_TRAIN_SET".





	Col1	Col2 (S)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
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

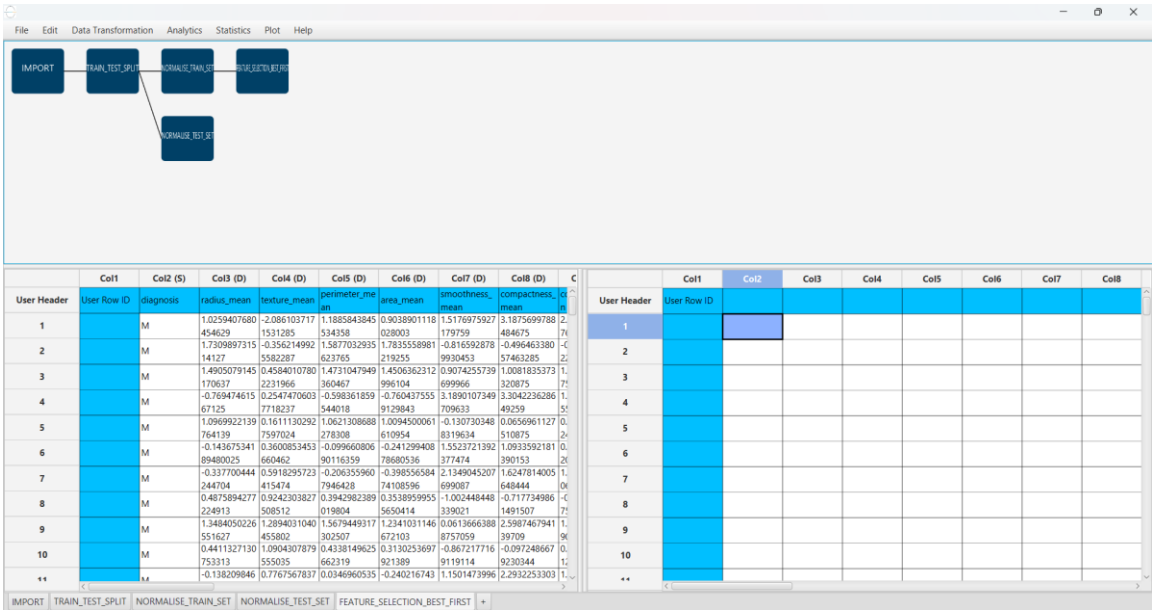
The results will appear on the output spreadsheet.



Step 6: Feature selection

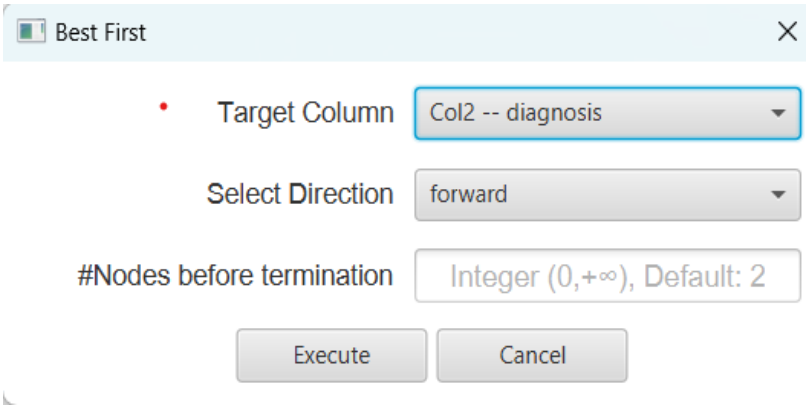
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_BEST_FIRST".

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".

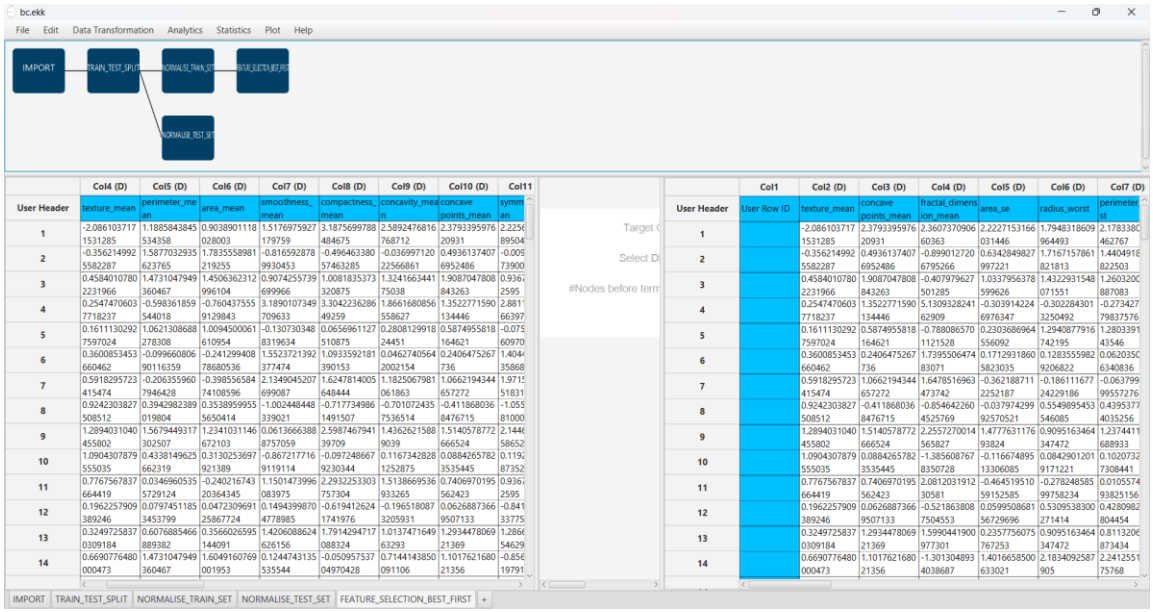


Choose the most important features for the classification using the Best First Function: "Data Transformation" → "Variable Selection" → "Best First".

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



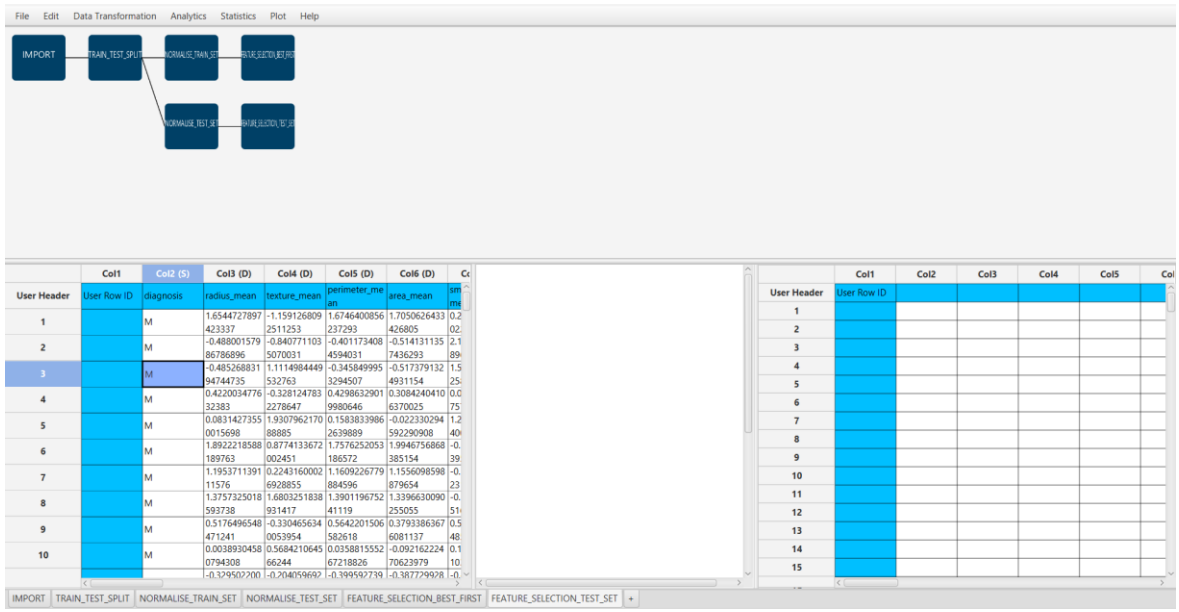
The results will appear on the output spreadsheet.



Step 7: Feature selection: test set

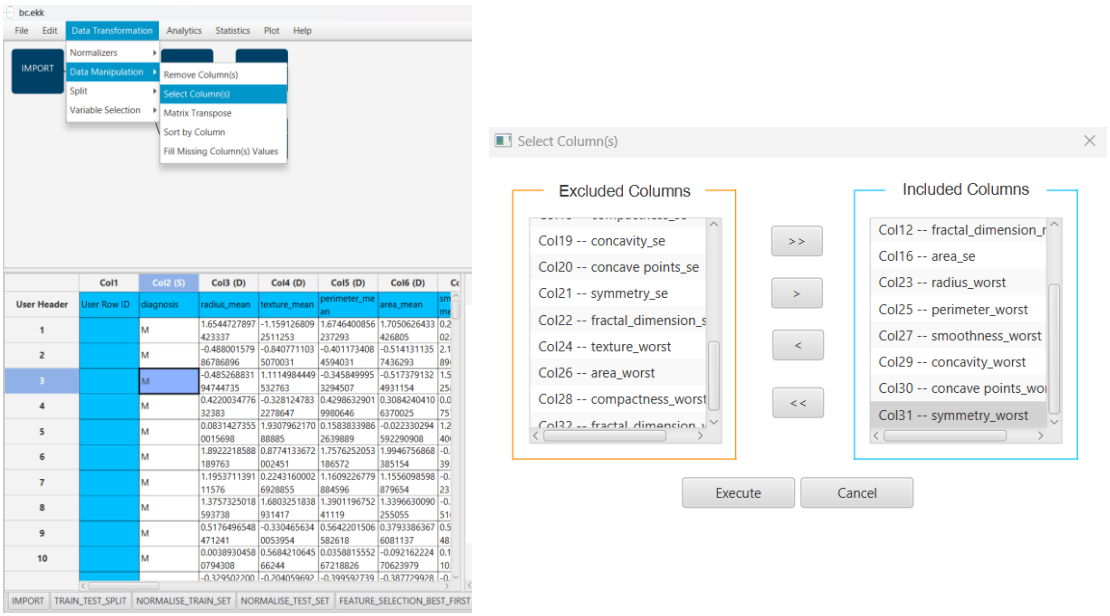
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_TEST_SET".

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".

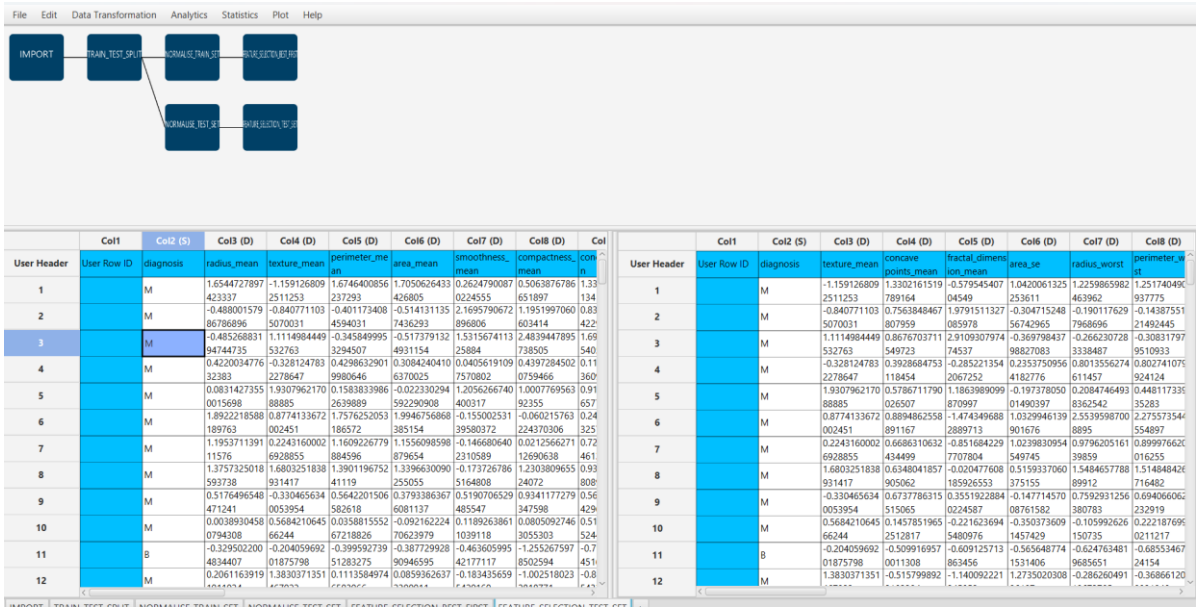


Manipulate the data by choosing the columns that correspond to the significant features (from the previous step):

"Data Transformation" → "Data Manipulation" → "Select Column(s)".



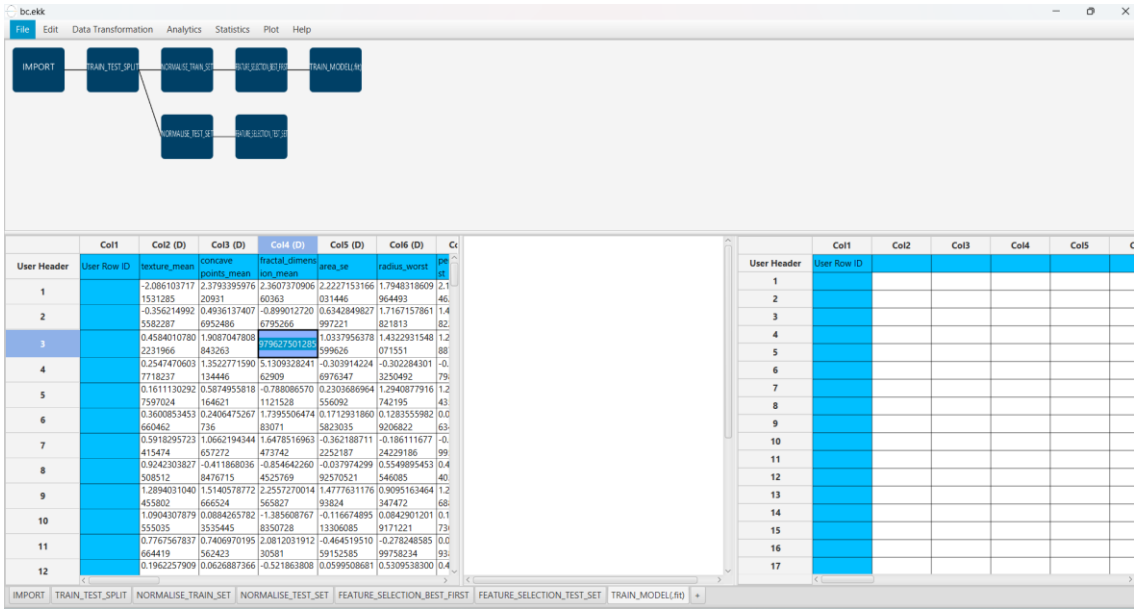
The results will appear on the output spreadsheet.



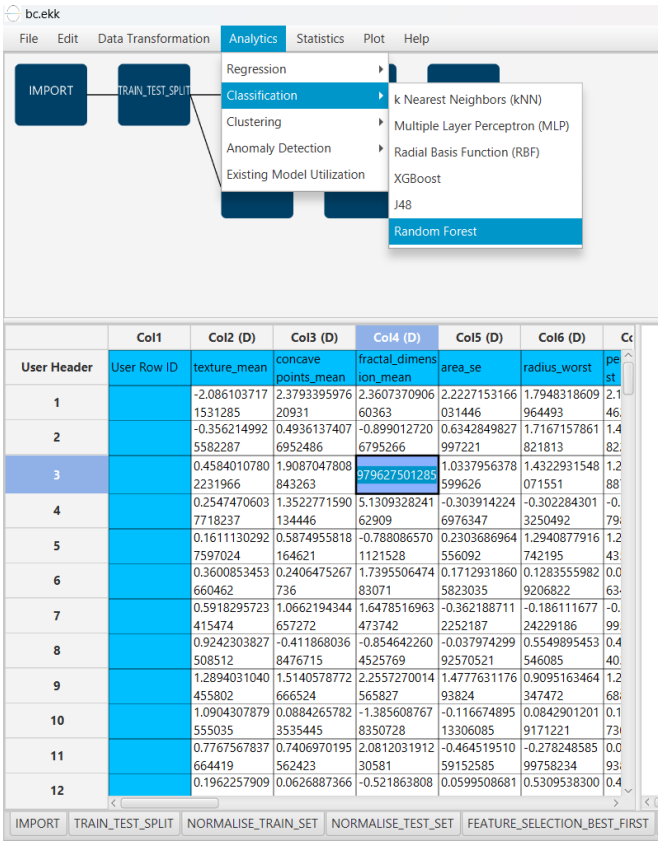
Step 8: Train the model

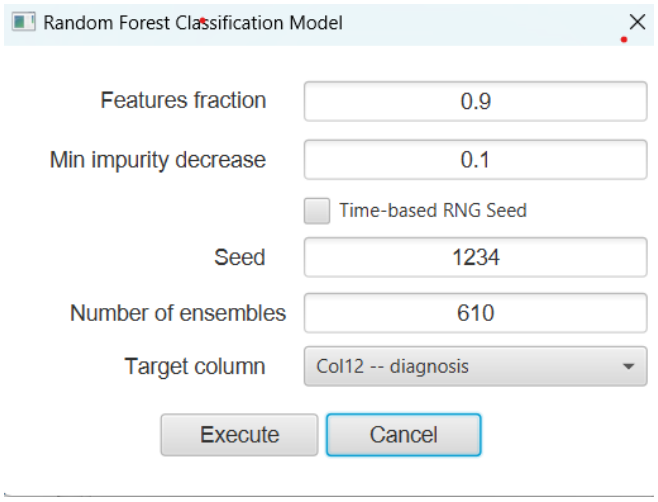
Create a new tab by pressing the "+" button on the bottom of the page with the name "TRAIN_MODEL(.fit)".

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".

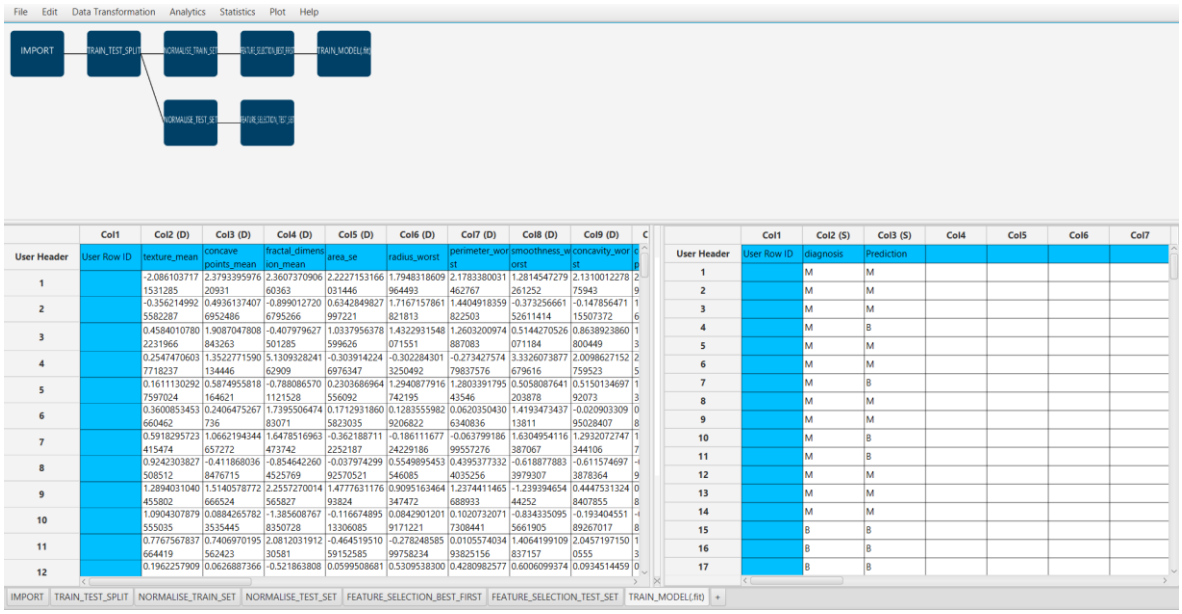


Use the Random Forest Method to train and fit the model: "Analytics" → "Classification" → "Random Forest" and adjust the model parameters based on training set performance.





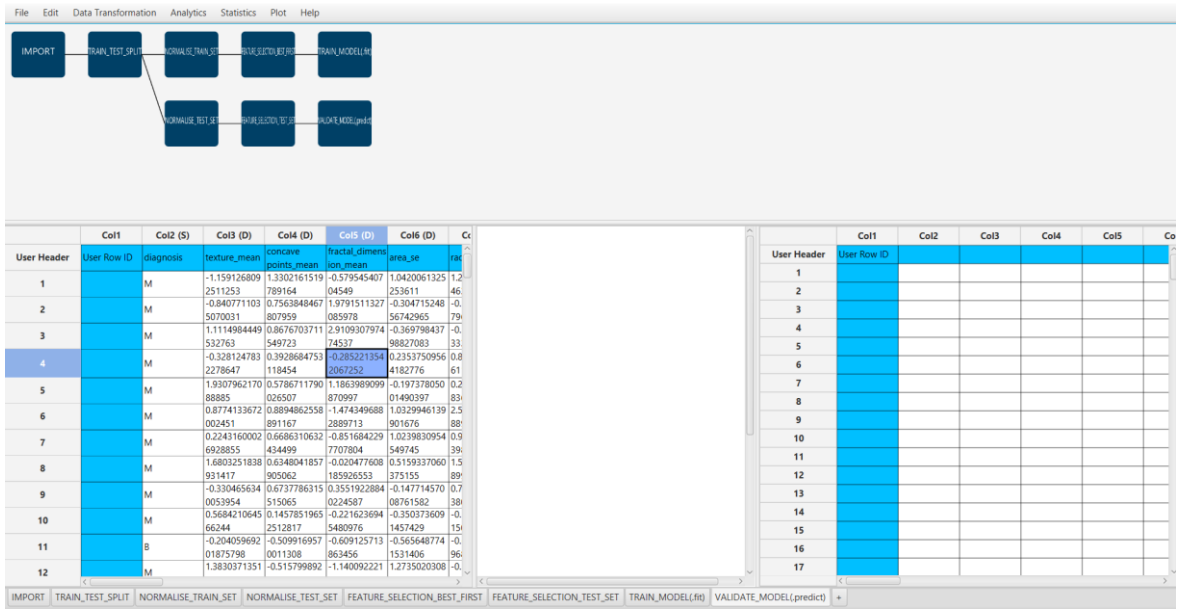
The predictions will appear on the output spreadsheet.



Step 9: Validate the model

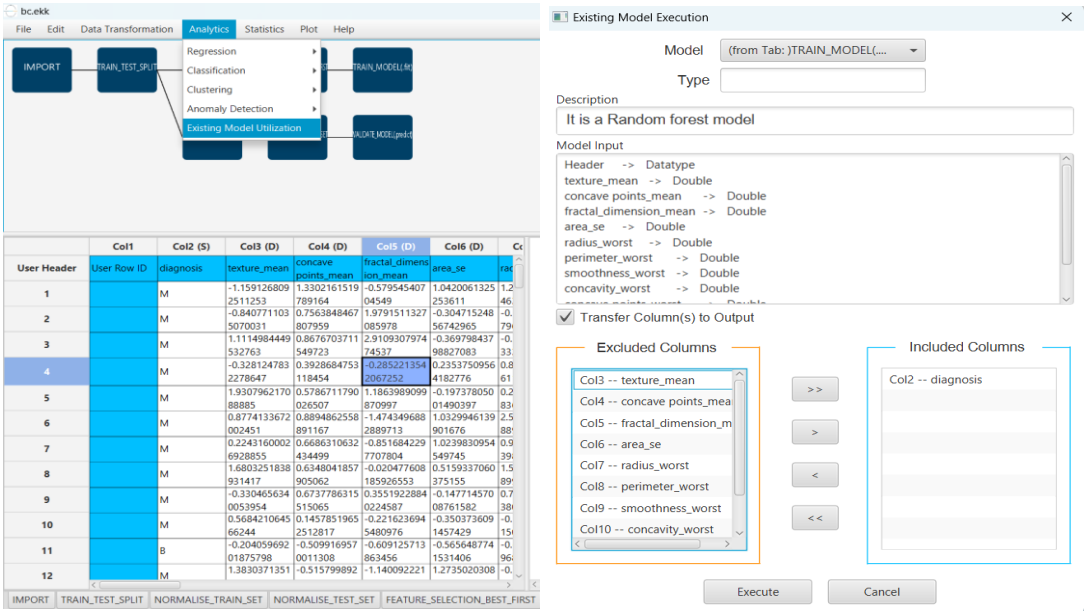
Create a new tab by pressing the "+" button on the bottom of the page with the name "VALIDATE_MODEL(.predict)".

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".

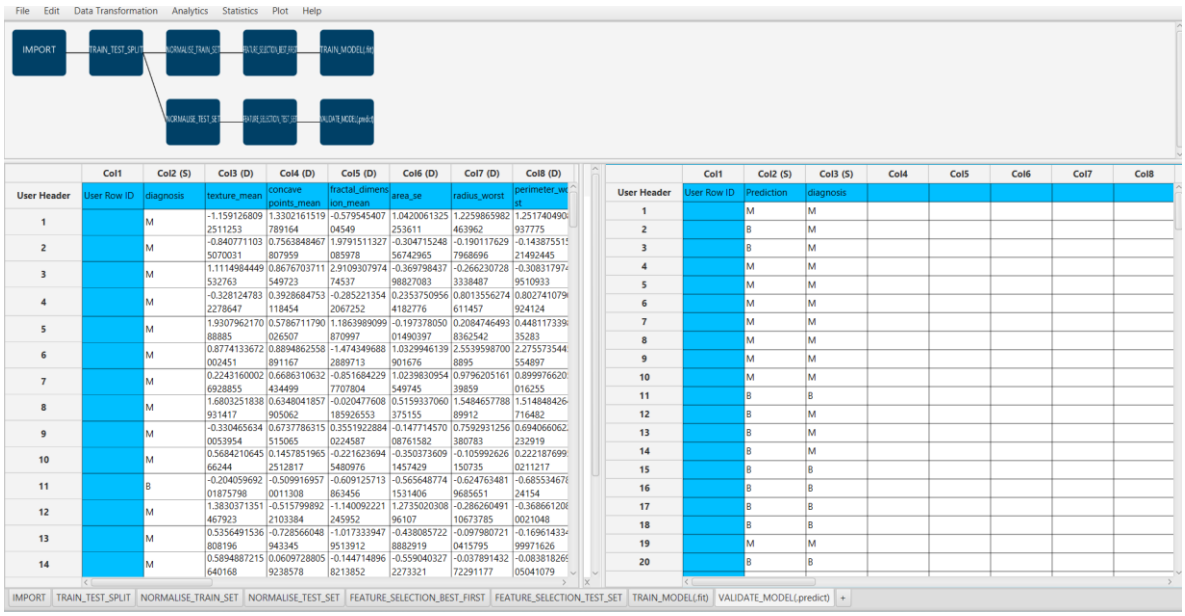


To validate the model:

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



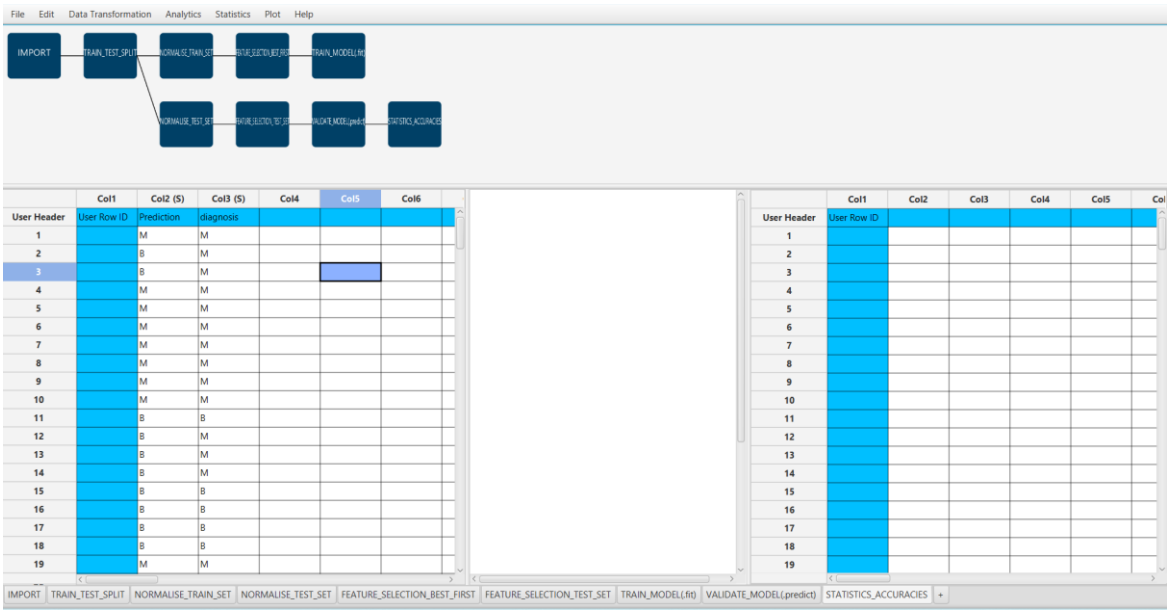
The predictions will appear on the output spreadsheet.



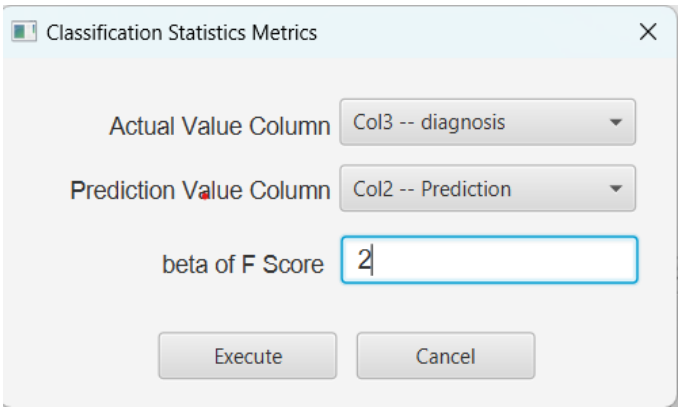
Step 10: Statistics calculation

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

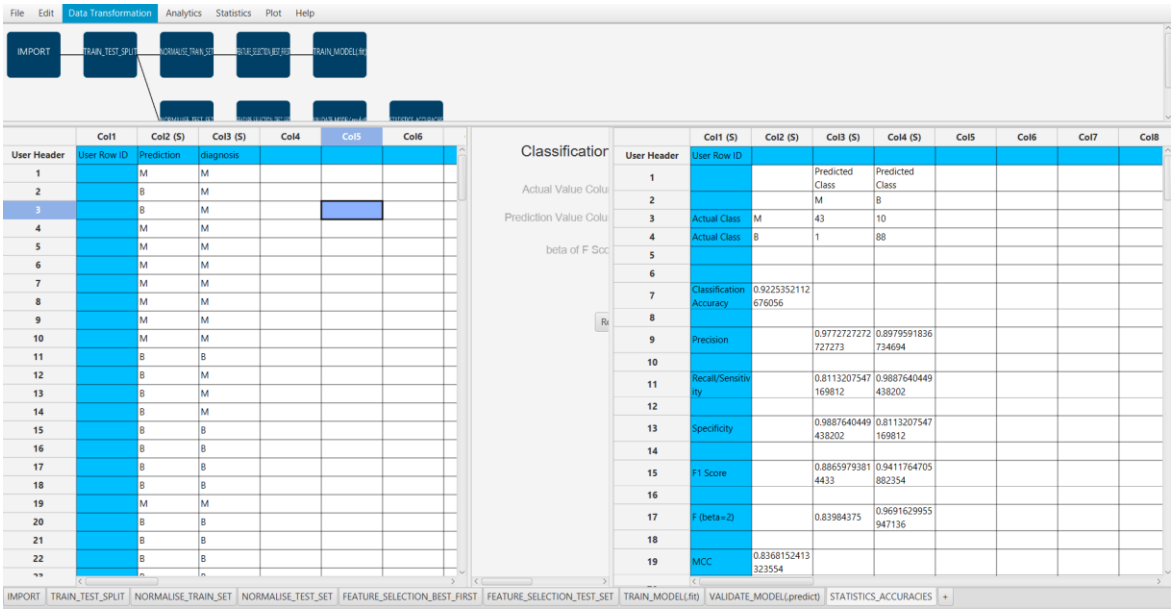
Import data into the input spreadsheet of the "STATISTICS_ACCURACIES" tab from the output of the "VALIDATE_MODEL(.predict)" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



The results will appear on the output spreadsheet.



Accuracy: 0.923

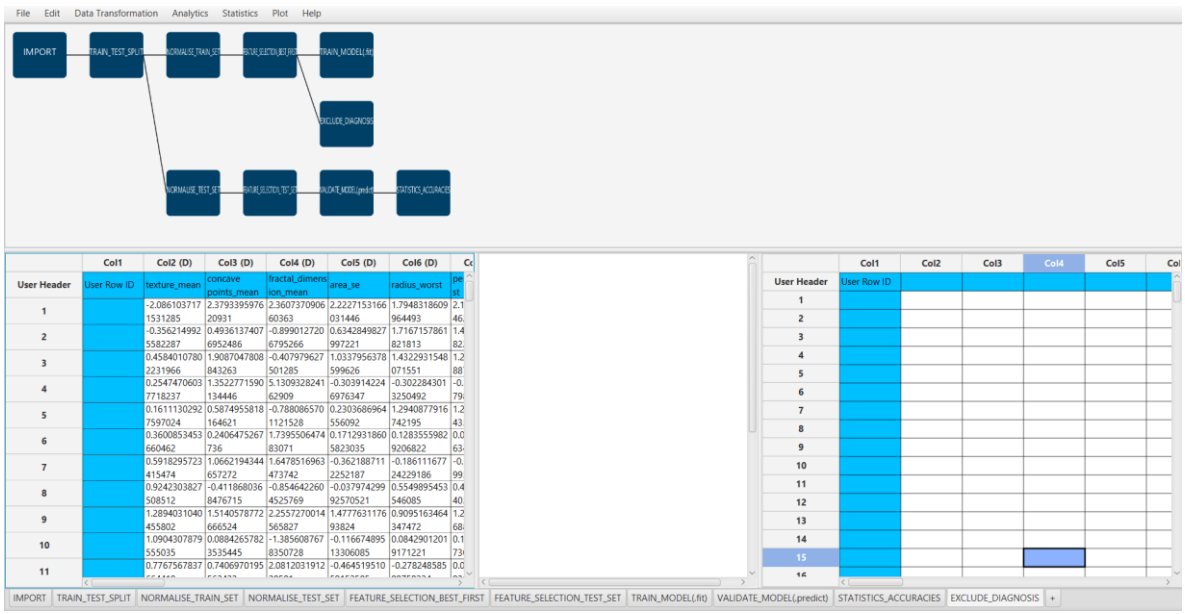
F1-Score = 0.914

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

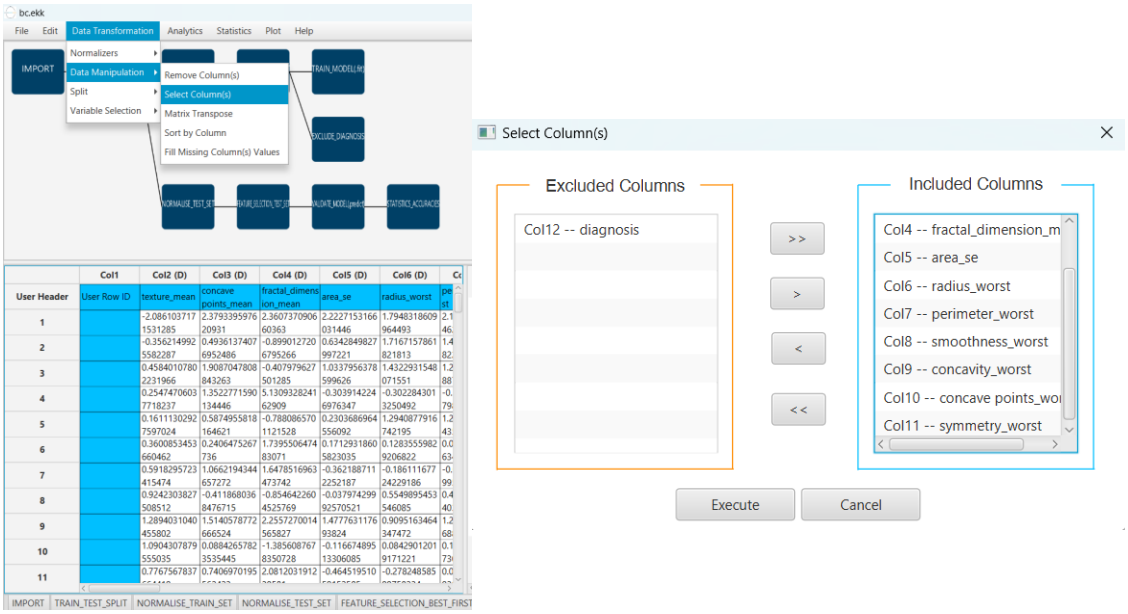
Step 11.a: Create the domain

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".



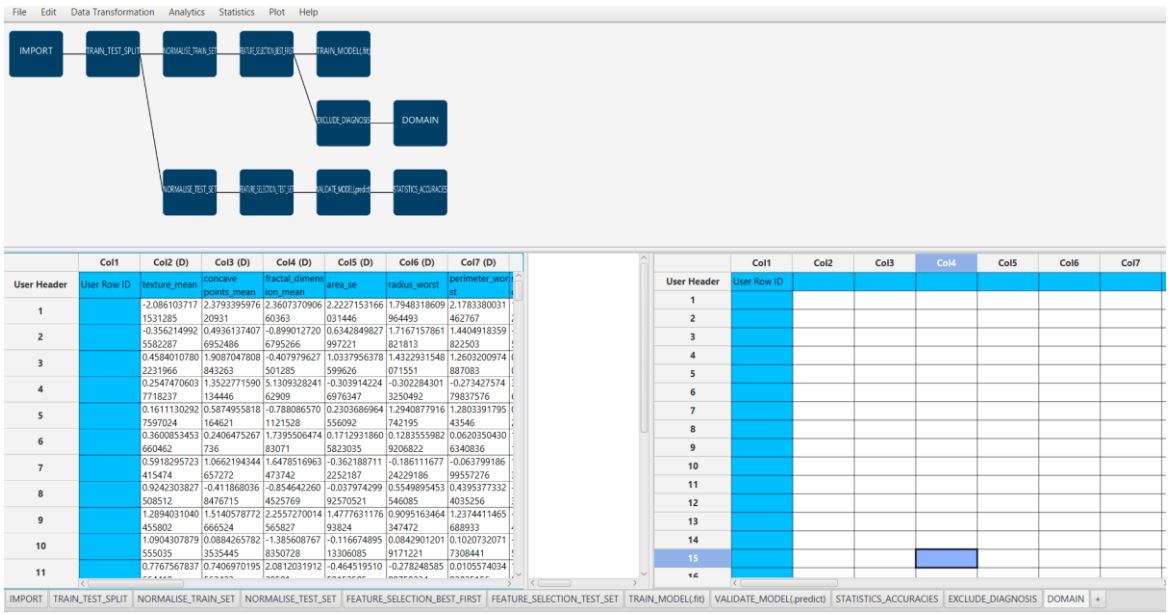
Manipulate the data to exclude the column that corresponds to the diagnosis "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except the "diagnosis".



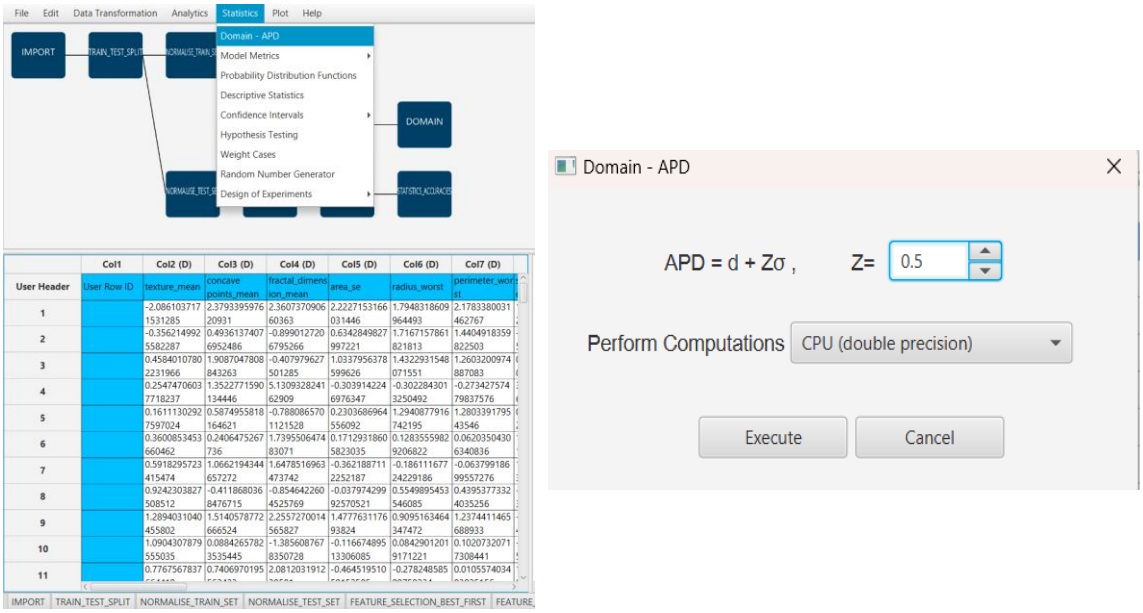
The results will appear on the output spreadsheet.

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

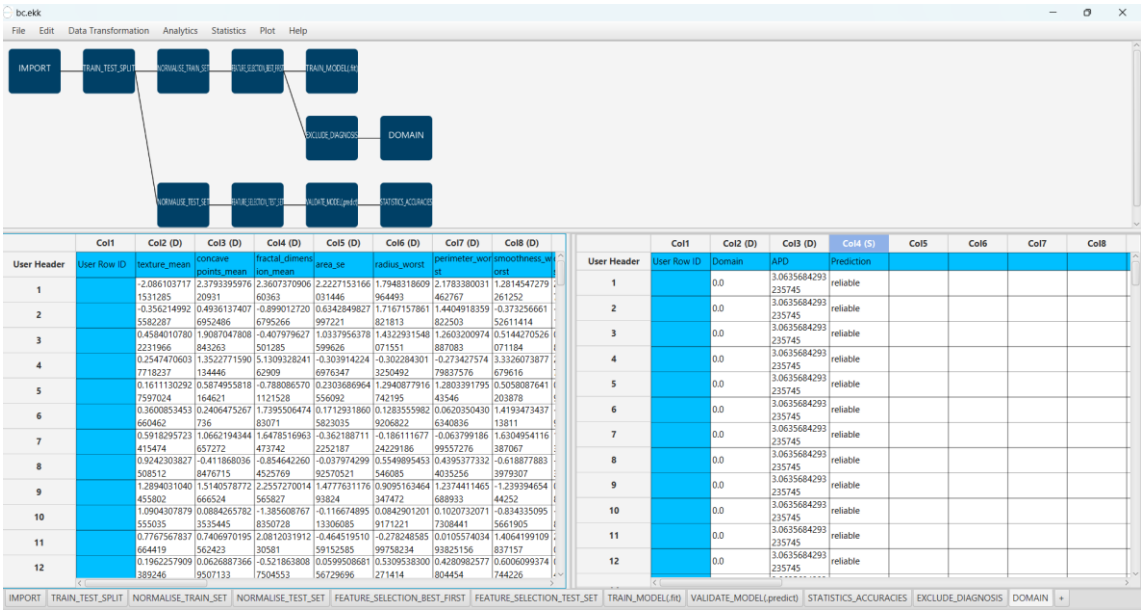
Import data into the input spreadsheet of the "DOMAIN" tab from the output of the "EXCLUDE_DIAGNOSIS" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



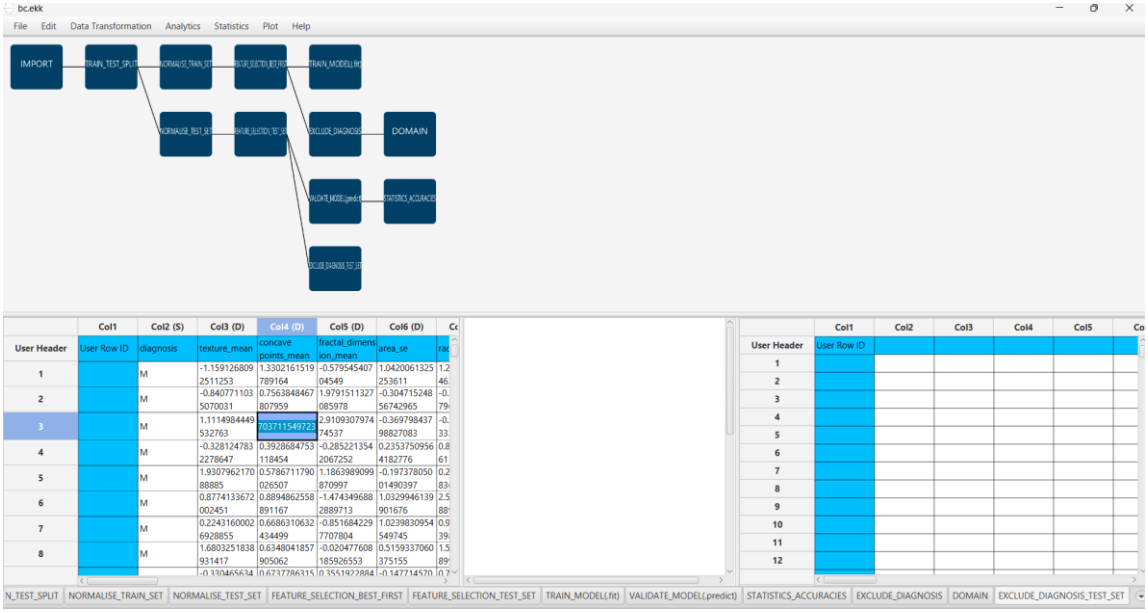
The results will appear on the output spreadsheet.



Step 11.b: Check the test set reliability

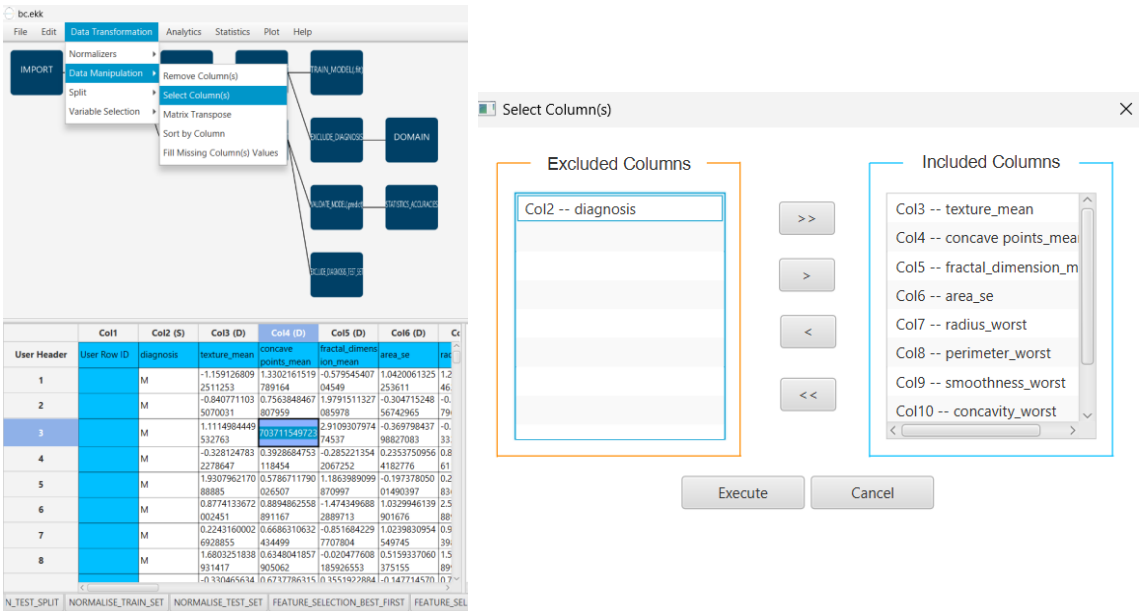
Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_DIAGNOSIS_TEST_SET".

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

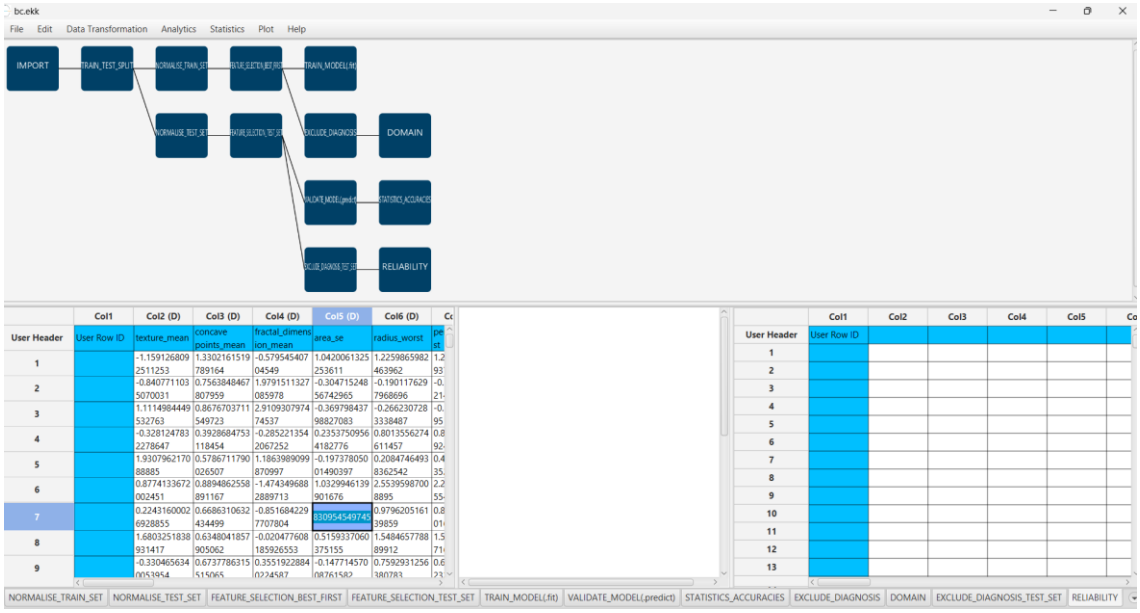
"Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except diagnosis.



The results will appear on the output spreadsheet.

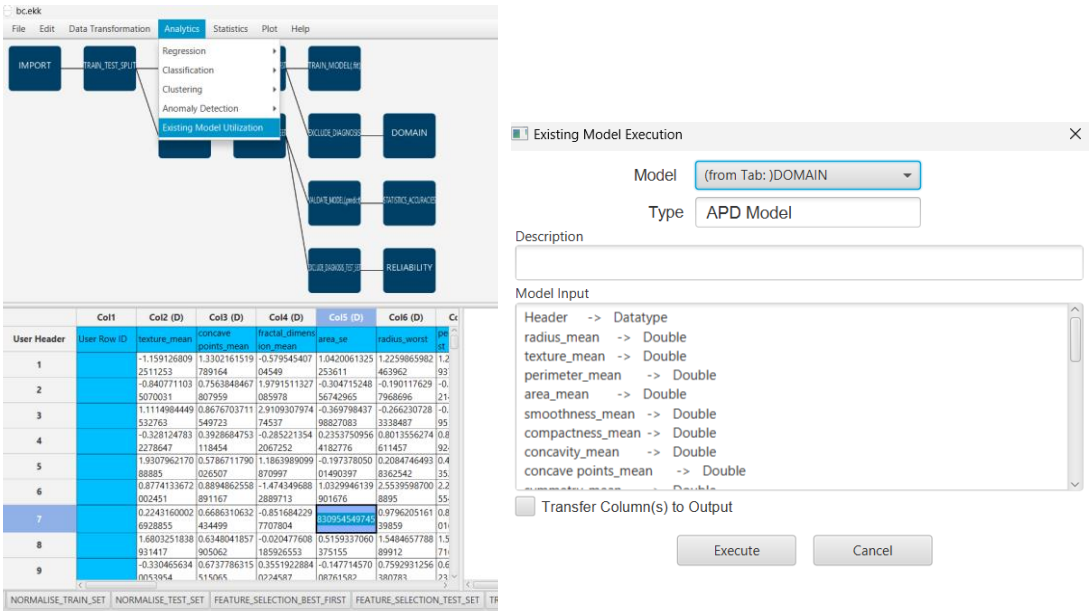
Create a new tab by pressing the "+" button on the bottom of the page with the name "RELIABILITY".

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".

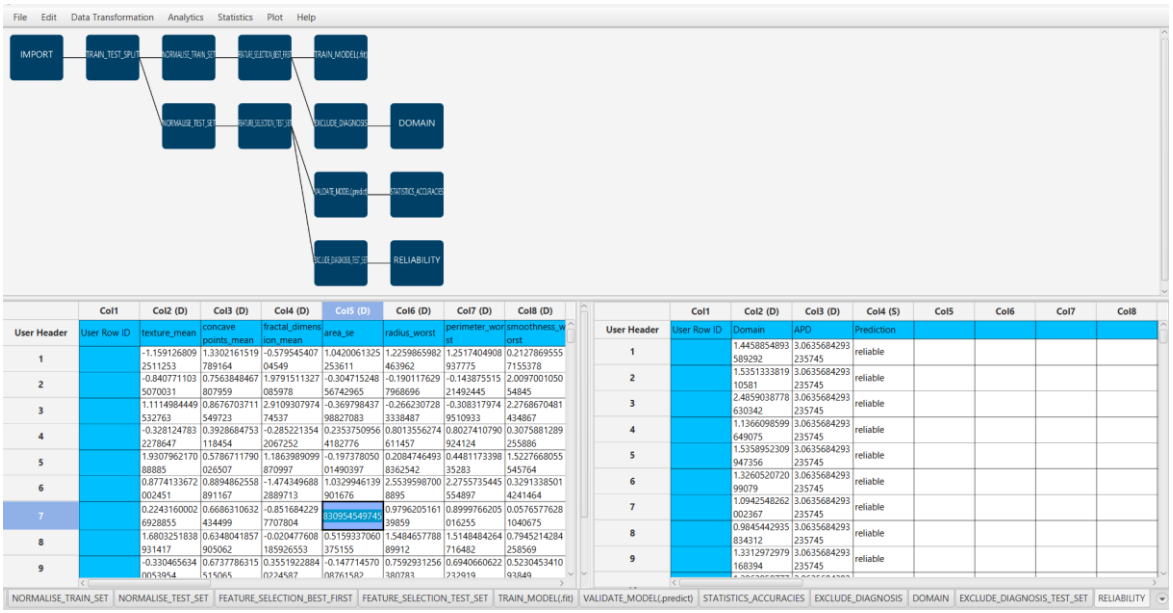


Check the Reliability:

"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: