

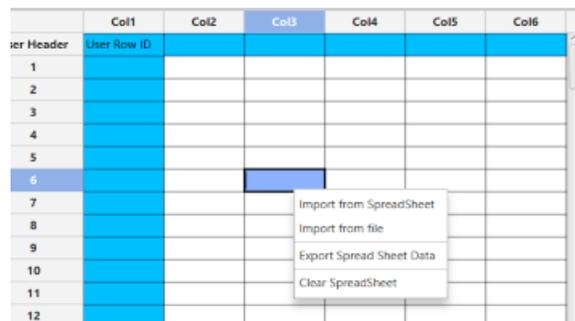


Breast Cancer (Binary Classification)

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.



IMPORT

User Header	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	convexity_mean	symmetry_mean
1	042302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	
2	042517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	
3	04300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	
4	04348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	
5	04358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	
6	043786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	
7	044359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	
8	04458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	
9	044981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	
10	04501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	
11	045636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	
12	04610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	
13	046226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	
14	046381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	
15	04667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	
16	04799002	M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	
17	048406	M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	
18	04862001	M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	
19	049014	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	8510653	B	13.08	15.71	85.63	530	0.1075	0.127	0.04568	0.0311	0.1967	

IMPORT

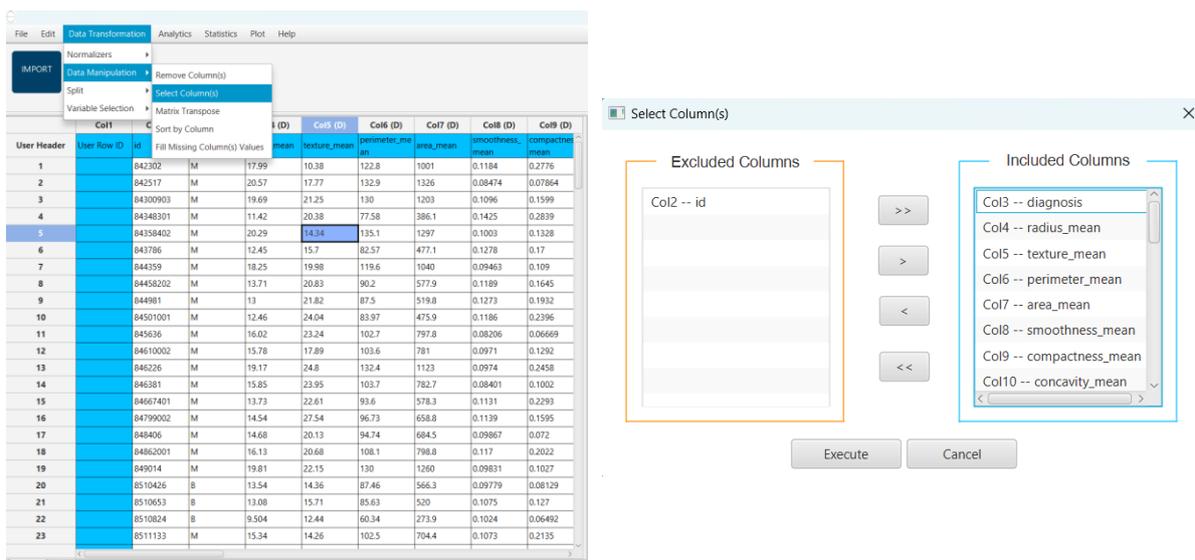
Step 2: Manipulate data

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".

	Col1	Col2	Col3	Col4	Col5	Col6
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						

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:

Usage of random generator seed:

Stratified sampling:

The results will appear on the output spreadsheet.

	Col1	Col2 (S)	Col3 (D)	Col4 (D)	Col5 (D)	Col6
User Header	User Row ID	diagnosis	radius_mean	texture_mean	perimeter_mean	area
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	519.8	
10	M	12.46	24.04	83.97	47	
11	M	16.02	23.24	102.7	75	
12	M	15.78	17.89	103.6	76	
13	M	19.17	24.8	132.4	11	
14	M	15.85	23.95	103.7	76	
15	M	13.73	22.61	93.6	57	
16	M	14.54	27.54	96.73	66	
17	M	14.68	20.13	94.74	66	
18	M	16.13	20.68	108.1	75	

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".

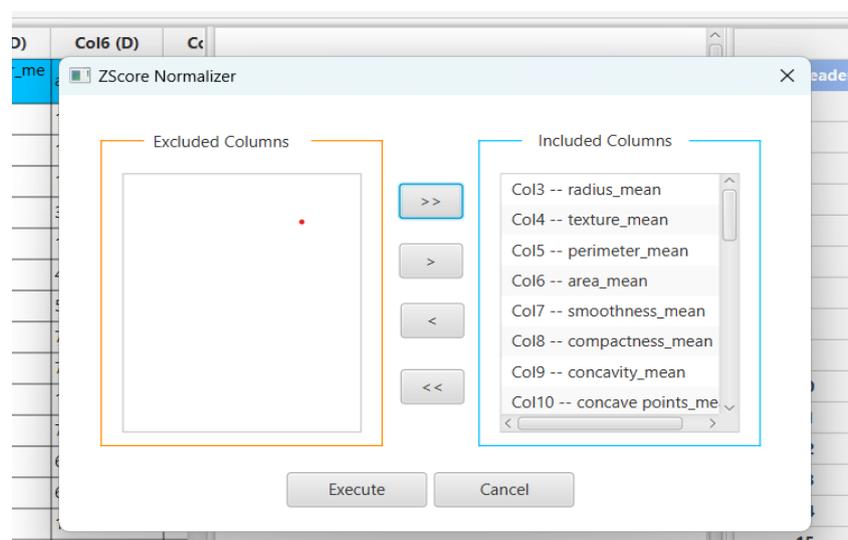
Import data into the input spreadsheet of the "NORMALISE_TRAIN_SET" tab the train set from the output of the "TRAIN_TEST_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet". From the available Select input tab options choose "TRAIN_TEST_SPLIT: Training Set"

The screenshot shows the Isalos Analytics Platform interface. At the top, there are three tabs: "IMPORT", "TRAIN_TEST_SPLIT", and "NORMALISE_TRAIN_SET". Below the tabs is a spreadsheet with 21 rows and 10 columns. The columns are labeled: "User Header", "Col1", "Col2 (S)", "Col3 (D)", "Col4 (D)", "Col5 (D)", "Col6 (D)", "Col7 (D)", "Col8 (D)", and "Col9 (D)". The data in the spreadsheet is as follows:

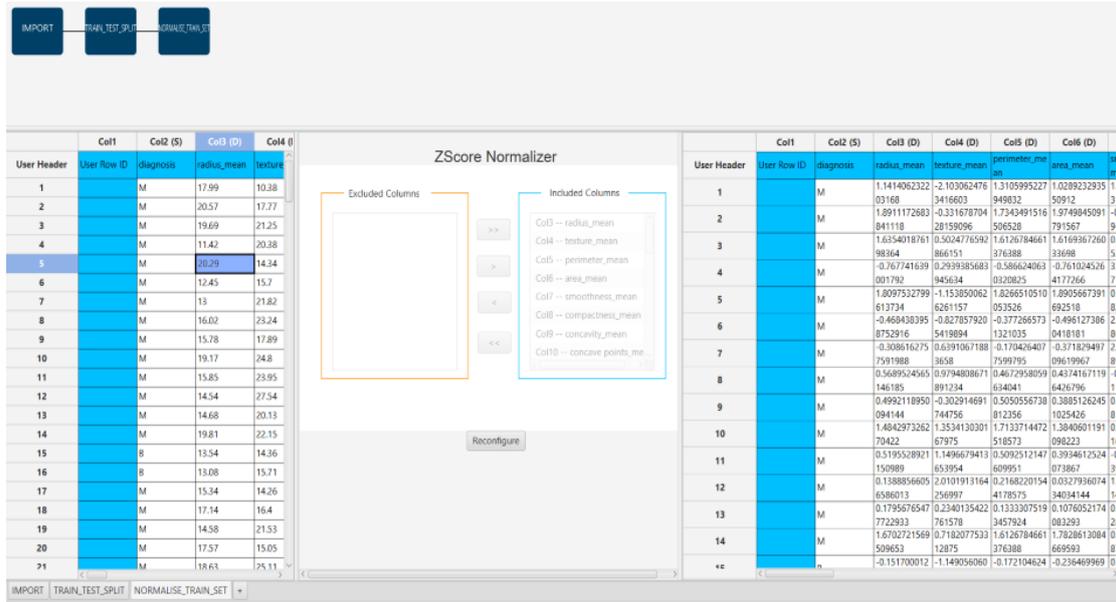
User Header	Col1	Col2 (S)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)
1	M	17.99	10.38	122.8	1001	0.1			
2	M	20.57	17.77	132.9	1326	0.0			
3	M	19.69	21.25	130	1203	0.1			
4	M	11.42	20.38	77.58	386.1	0.1			
5	M	20.29	14.34	135.1	1297	0.1			
6	M	12.45	15.7	82.57	477.1	0.1			
7	M	13	21.82	87.5	519.8	0.1			
8	M	16.02	23.24	102.7	797.8	0.0			
9	M	15.78	17.89	103.6	781	0.0			
10	M	19.17	24.8	132.4	1123	0.0			
11	M	15.85	23.95	103.7	782.7	0.0			
12	M	14.54	27.54	96.73	658.8	0.1			
13	M	14.68	20.13	94.74	684.5	0.0			
14	M	19.81	22.15	130	1260	0.0			
15	B	13.54	14.36	87.46	566.3	0.0			
16	B	13.08	15.71	85.63	520	0.1			
17	M	15.34	14.26	102.5	704.4	0.1			
18	M	17.14	16.4	116	912.7	0.1			
19	M	14.58	21.53	97.41	644.8	0.1			
20	M	17.57	15.05	115	955.1	0.0			
21	M	18.63	25.11	124.8	1088	0.1			

Overlaid on the spreadsheet is a "ZScore Normalizer" dialog box. The dialog has two sections: "Excluded Columns" (empty) and "Included Columns" (containing a list of columns: Col3 -- radius_mean, Col4 -- texture_mean, Col5 -- perimeter_mean, Col6 -- area_mean, Col7 -- smoothness_mean, Col8 -- compactness_mean, Col9 -- concavity_mean, and Col10 -- concave points_me). There are navigation buttons (>>, >, <, <<) and "Execute" and "Cancel" buttons at the bottom.

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



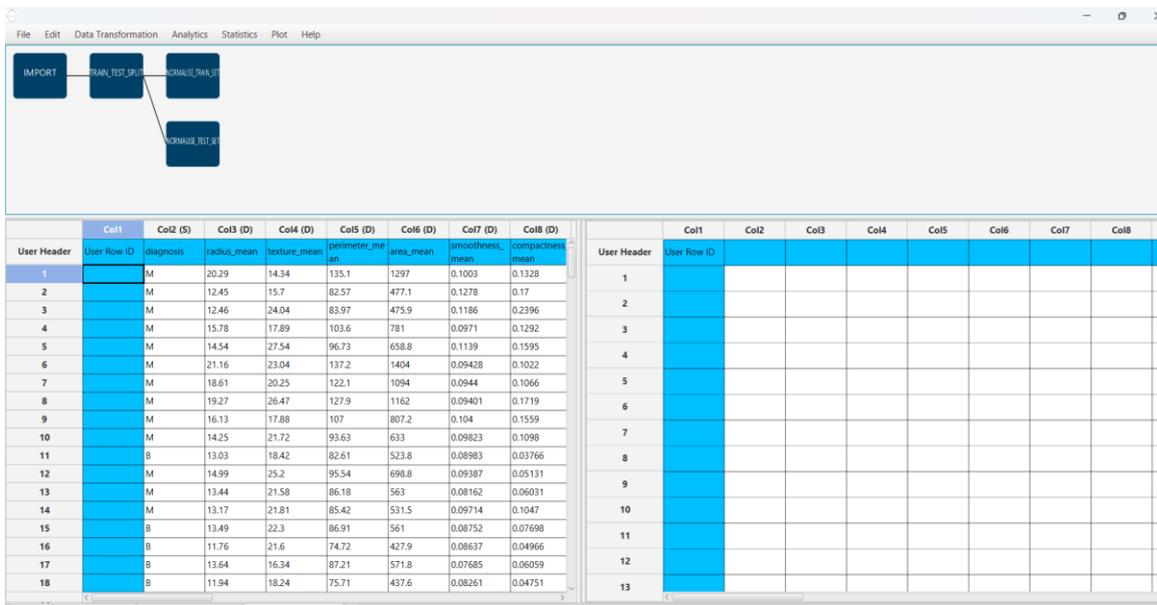
The results will appear on the output spreadsheet.



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".

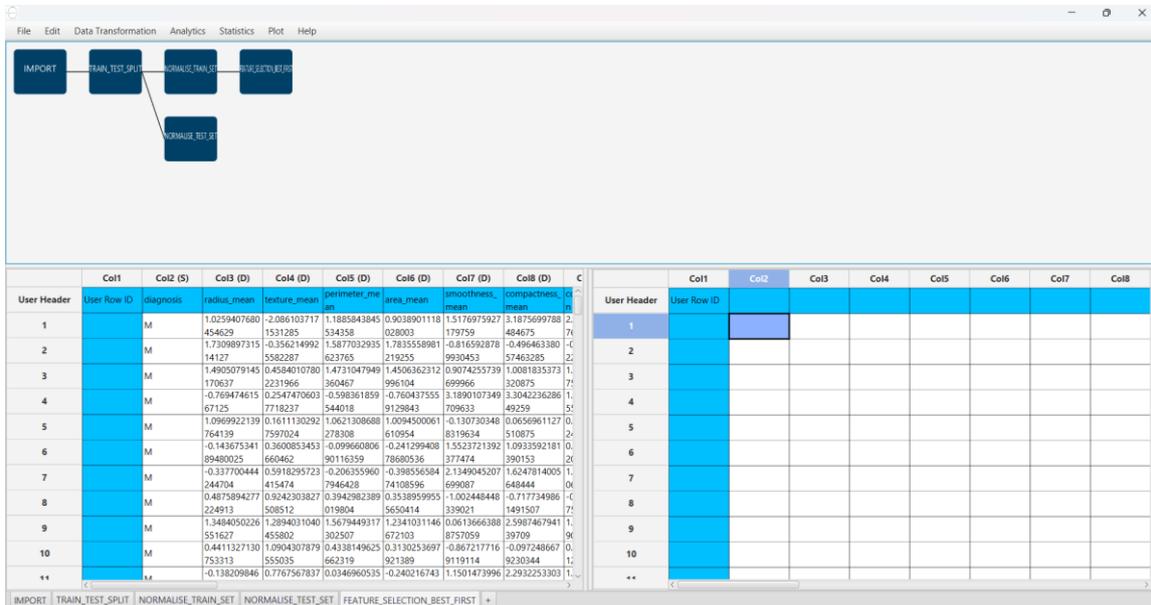


Normalize the test set using the existing normalizer of the training set: "Analytics" → "Existing Model Utilization" → "Model (from Tab:) NORMALISE_TRAIN_SET".

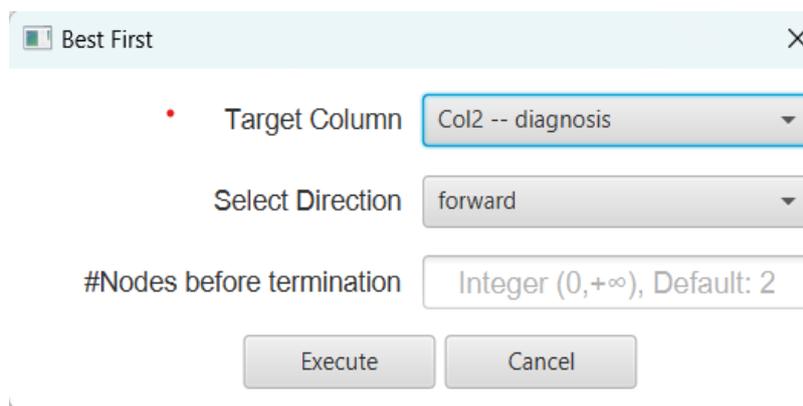
Step 6: Feature selection

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.

Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11
texture_mean	perimeter_me	area_mean	smoothness	compactness	convexity_me	concave	symm
1	2.086103717	1.188584345	0.9038901118	1.5176975927	3.1875699788	2.5892476816	2.3793395976
2	1531285	534358	028003	179759	484675	768712	20931
3	0.4584010780	1.4731047949	1.4506362312	0.9074255739	1.0081835373	1.3241663441	1.5087047808
4	0.2547470603	-0.598361859	-0.760437555	1.1890107349	3.3042236286	1.8661680856	1.3522771590
5	0.1611130292	1.0621308688	1.0094500061	-0.130730348	0.0656961127	0.2808129918	0.5874955818
6	0.3600853453	-0.099668086	-0.241299408	1.5523721392	1.0933592181	0.0462740564	0.2406475267
7	0.5918295723	-0.206355960	-0.398556584	2.1349045207	1.6247814005	1.1825067981	1.0662194344
8	0.3242303827	0.3942803289	0.3538959955	-1.002448448	0.717724986	-0.701072435	-0.411866036
9	1.2894031040	1.5679449317	1.2341031146	0.0613666388	2.5987467941	1.4362621588	1.5140578772
10	0.455802	0.302507	0.672103	0.8757059	39709	0.9039	0.666524
11	0.1962257909	0.0797451185	0.0472309691	0.1494399870	-0.619412624	-0.196518087	0.0626887366
12	0.324975837	0.6076885466	0.3564026595	1.4206988624	1.7914294717	1.01032741649	1.2934487069
13	0.6690776480	1.4731047949	1.6049160769	0.1244743135	-0.050957537	0.1144143850	1.1017621680
14	0.00473	0.360467	0.01953	0.535544	0.4970428	0.91106	0.21356

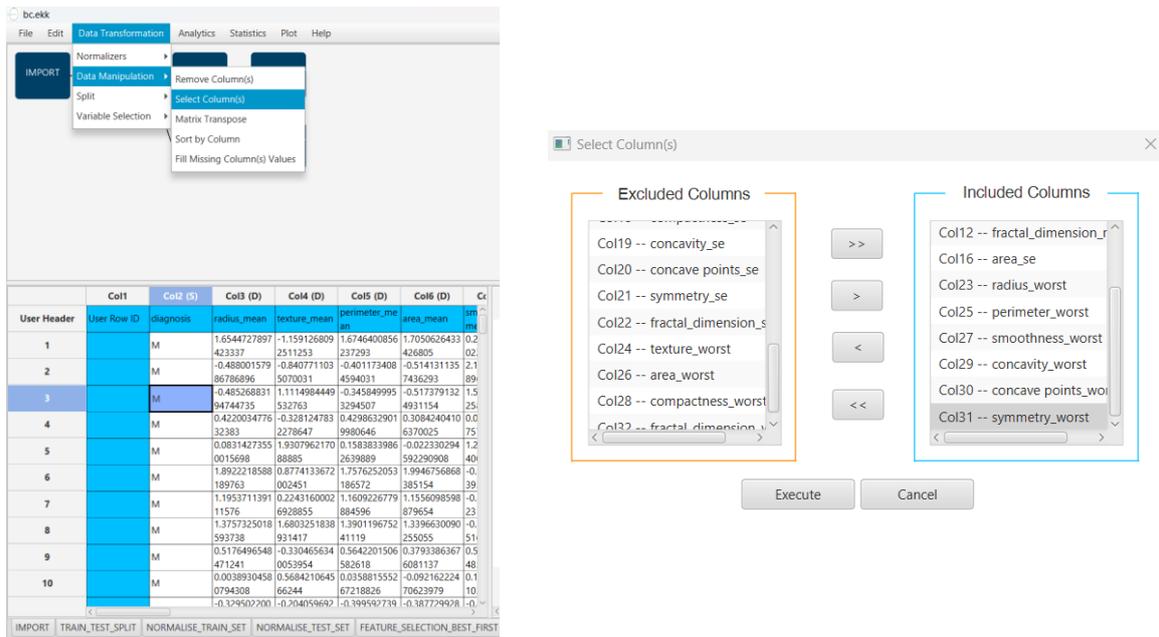
Step 7: Feature selection: test set

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

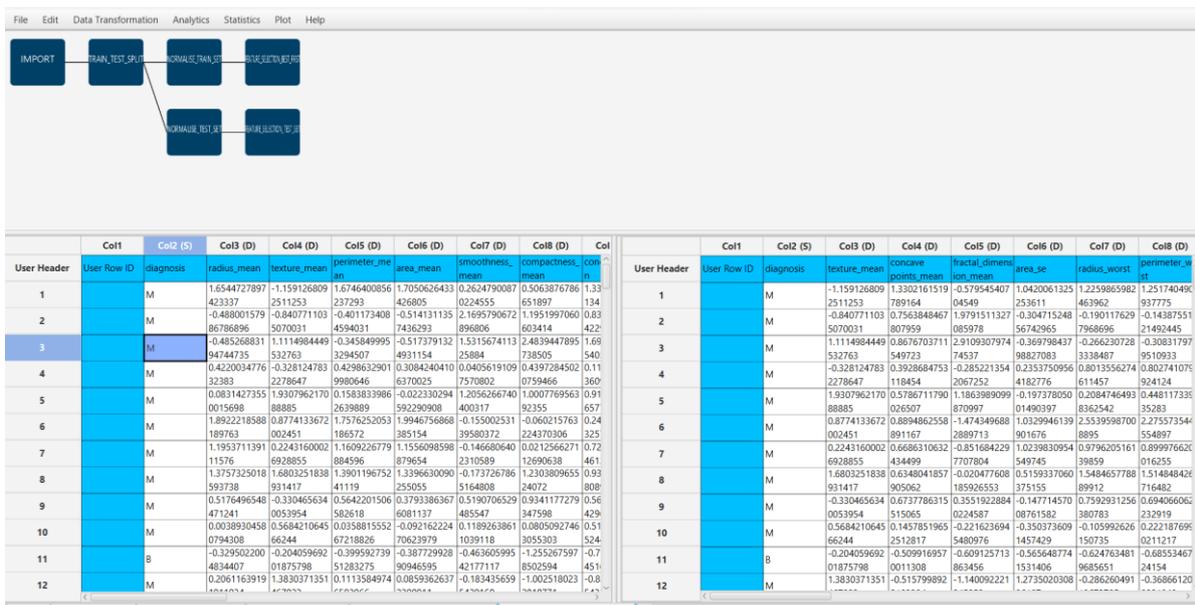
Import data into the input spreadsheet of the "FEATURE_SELECTION_TEST_SET" tab from the output of the "NORMALISE_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".

Col1	Col2 (S)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)
diagnosis	radius_mean	texture_mean	perimeter_me	area_mean	sm	ms
1	M	1.6544727897	-1.159126809	1.6746400856	1.7050626433	0.2
2	M	423337	2511253	237293	426805	0.02
3	M	0.488001579	-0.840771103	-0.401173408	-0.514131135	2.1
4	M	86786896	5070031	4594031	7436293	89
5	M	-0.485268831	1.1114984449	-0.345849995	-0.517379132	1.5
6	M	94744735	532763	3294507	4931154	25
7	M	0.4220034776	-0.328124783	0.4298632901	0.3084240410	0.1
8	M	32383	2278647	9980646	6370025	75
9	M	0.0831427355	1.9307962170	0.1583833986	-0.022330294	1.2
10	M	0015698	88885	2639889	592290908	40
11	M	1.8922218588	0.8774133672	1.7576252053	1.9946756868	-0.
12	M	189763	002451	186572	385154	39
13	M	1.1953711391	0.2243160002	1.1609226779	1.1556098598	-0.
14	M	11576	6928855	884956	879654	23
15	M	1.3757325018	1.6803251838	1.3901196752	1.3396630090	-0.
16	M	593738	931417	41119	255055	51
17	M	0.5176496548	-0.330465634	0.5642201506	0.3793386367	0.5
18	M	471241	0053954	582618	6081137	48
19	M	0.003893458	0.5684210645	0.0358815552	-0.092162234	0.1
20	M	0794308	66244	67218826	70623079	10
21	M	-0.329502700	-0.704059692	-0.399547739	-0.387724928	-0.

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".

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)
1	-2.086103717	2.3793395976	2.3607370906	2.2227153166	1.7948318609	2.1	2.1
2	1531285	20931	60363	031446	964493	46	46
3	0.4584010780	1.9087047808	1.3522771590	5.1309328241	-0.303914224	-0.302284301	-0.
4	0.2547470603	1.34446	62909	6976347	3250492	79	79
5	0.1611130292	0.5874955818	-0.78806570	0.2303686964	1.2940877916	1.2	1.2
6	0.3600853453	0.2406475267	1.7395506474	0.1712931860	0.1283555982	0.0	0.0
7	0.5918295723	1.0662194344	1.6478516963	-0.362188711	-0.186111677	-0.	-0.
8	0.9242303827	-0.411868036	-0.854642260	-0.037974299	0.5549895453	0.4	0.4
9	1.2894031040	1.5140578772	2.2557270014	1.4777631176	0.9095163464	1.2	1.2
10	0.7767567837	0.7406970195	2.0812031912	-0.464519510	-0.278248585	0.0	0.0
11	664419	562423	30581	59152585	99758234	93	93
12	0.1962257909	0.0626887366	-0.521863808	0.0599508681	0.5309538300	0.4	0.4

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.

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)
1	-2.086103717	2.3793395976	2.3607370906	2.2227153166	1.7948318609	2.1	2.1
2	1531285	20931	60363	031446	964493	46	46
3	0.4584010780	1.9087047808	1.3522771590	5.1309328241	-0.303914224	-0.302284301	-0.
4	0.2547470603	1.34446	62909	6976347	3250492	79	79
5	0.1611130292	0.5874955818	-0.78806570	0.2303686964	1.2940877916	1.2	1.2
6	0.3600853453	0.2406475267	1.7395506474	0.1712931860	0.1283555982	0.0	0.0
7	0.5918295723	1.0662194344	1.6478516963	-0.362188711	-0.186111677	-0.	-0.
8	0.9242303827	-0.411868036	-0.854642260	-0.037974299	0.5549895453	0.4	0.4
9	1.2894031040	1.5140578772	2.2557270014	1.4777631176	0.9095163464	1.2	1.2
10	0.7767567837	0.7406970195	2.0812031912	-0.464519510	-0.278248585	0.0	0.0
11	664419	562423	30581	59152585	99758234	93	93
12	0.1962257909	0.0626887366	-0.521863808	0.0599508681	0.5309538300	0.4	0.4

Random Forest Classification Model

Features fraction: 0.9

Min impurity decrease: 0.1

Time-based RNG Seed

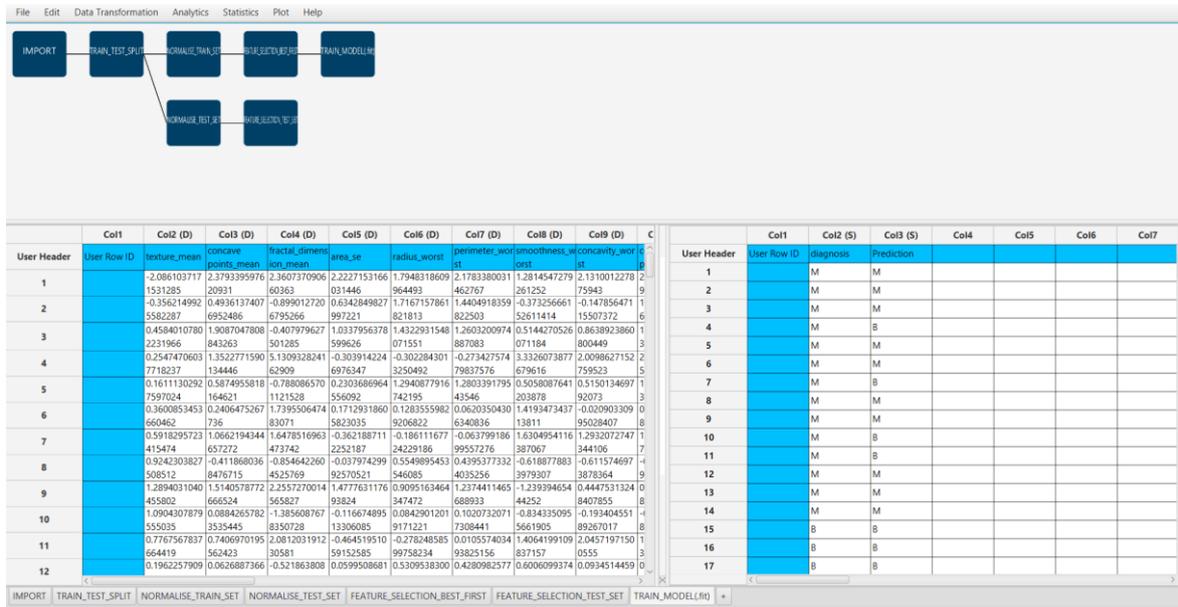
Seed: 1234

Number of ensembles: 610

Target column: Col12 -- diagnosis

Execute Cancel

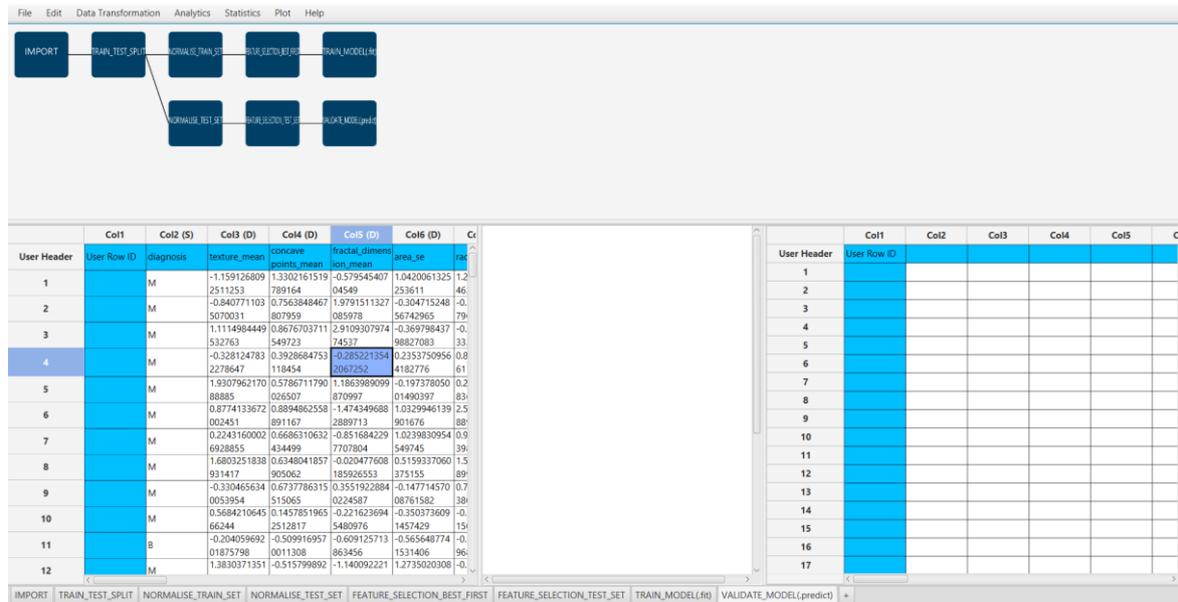
The predictions will appear on the output spreadsheet.



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 screenshot shows the 'bc.ekk' application window. At the top, there is a menu bar with 'File', 'Edit', 'Data Transformation', 'Analytics', 'Statistics', 'Plot', and 'Help'. Below the menu is a workflow diagram with nodes: 'IMPORT', 'TRAIN_TEST_SPLIT', 'NORMALISE_TRAIN_SET', 'NORMALISE_TEST_SET', 'FEATURE_SELECTION_BEST_FIRST', 'TRAIN_MODEL', and 'VALIDATE_MODEL(predict)'. The 'Analytics' menu is open, showing options like 'Regression', 'Classification', 'Clustering', 'Anomaly Detection', and 'Existing Model Utilization'. Below the workflow is a data table with columns: User Header, User Row ID, diagnosis, texture_mean, concave_points_mean, fractal_dimension_mean, area_se, radius_worst, perimeter_worst, smoothness_worst, concavity_worst, and prediction.

User Header	Col1	Col2 (S)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)
1	1	M	-1.159126809	1.3302161519	-0.579545407	1.0420061325	1.2259865982	1.2517404900	0.463962	0.937775	0.143875511
2	2	M	-0.840771103	0.7563848467	1.0791511327	-0.304715248	-0.190117629	-0.143875511	0.7968896	2.1482445	0.802741079
3	3	M	1.1114984449	0.8676703711	2.9109307974	-0.369798437	-0.266230728	-0.308317974	33.38487	95.10933	0.802741079
4	4	M	-0.328124783	0.3928684753	-0.285221354	0.2353750956	0.8013556274	0.802741079	61.1457	95.10933	0.802741079
5	5	M	1.9307962170	0.5786711790	1.1863989099	-0.197378050	0.204746493	0.448117339	88.885	0.26507	0.870997
6	6	M	0.8774133672	0.8894862558	-1.474349688	1.0329946139	2.5539598700	2.275573544	0.02451	89.1167	2889713
7	7	M	0.2243160002	0.6686310632	-0.851684229	1.0239830954	0.9796205161	0.899976620	69.28855	434499	7707804
8	8	M	1.6803251838	0.6348041857	-0.020477608	0.515937060	1.5484657788	1.514848426	93.1417	905062	185926553
9	9	M	-0.330465634	0.6737786315	0.3551922884	-0.147714570	0.759291256	0.694066062	0.003954	515065	0.224587
10	10	M	0.5684210645	0.1457851965	-0.221623694	-0.350373609	-0.105992626	0.222187699	66.244	25.12817	5480976
11	11	B	-0.204059692	-0.509916957	-0.609125713	-0.565648774	-0.624763481	-0.68553467	0.1875798	0.011308	863456
12	12	M	1.3830371351	-0.515799892	-1.140092221	1.2735020308	-0.286260491	-0.36866120	46.7923	2103384	245952
13	13	M	0.5356491536	-0.728566048	-1.017333947	-0.438085722	-0.097980721	-0.16961433	808196	943345	9513912
14	14	M	0.589488215	0.060972805	-0.144714896	-0.559040327	-0.037891432	-0.08381826	640168	9238378	8213852

The 'Existing Model Execution' dialog box shows a dropdown menu for 'Model' set to '(from Tab:) TRAIN_MODEL(...)' and a 'Type' field. The 'Description' field contains 'It is a Random forest model'. The 'Model Input' section lists various features: texture_mean, concave_points_mean, fractal_dimension_mean, area_se, radius_worst, perimeter_worst, smoothness_worst, and concavity_worst. A checkbox 'Transfer Column(s) to Output' is checked. Below, there are two columns: 'Excluded Columns' (containing texture_mean, concave_points_mean, fractal_dimension_mean, area_se, radius_worst, perimeter_worst, smoothness_worst, and concavity_worst) and 'Included Columns' (containing diagnosis). Navigation buttons '>>', '>', '<', and '<<' are between the columns. 'Execute' and 'Cancel' buttons are at the bottom.

The predictions will appear on the output spreadsheet.

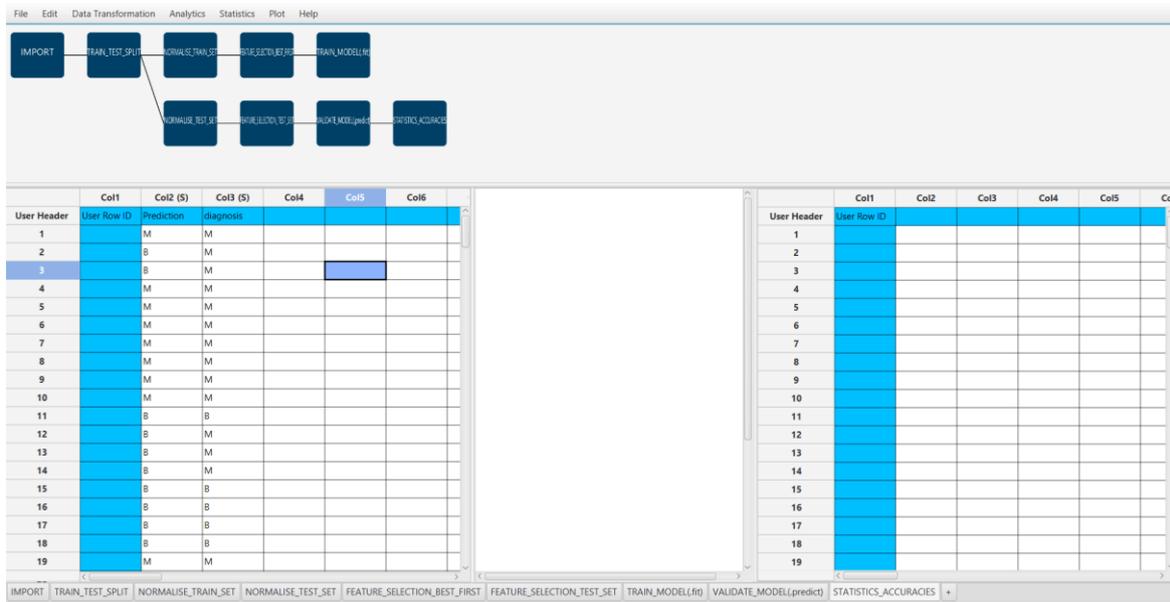
The screenshot shows the 'bc.ekk' application window with a workflow diagram and a data table. The workflow diagram includes nodes: 'IMPORT', 'TRAIN_TEST_SPLIT', 'NORMALISE_TRAIN_SET', 'NORMALISE_TEST_SET', 'FEATURE_SELECTION_BEST_FIRST', 'TRAIN_MODEL', 'VALIDATE_MODEL(predict)', and 'TRAIN_MODEL(...)'.

User Header	Col1	Col2 (S)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)
1	1	M	M	M	M	M	M	M	M	M	M	M
2	2	B	M	M	M	M	M	M	M	M	M	M
3	3	B	M	M	M	M	M	M	M	M	M	M
4	4	M	M	M	M	M	M	M	M	M	M	M
5	5	M	M	M	M	M	M	M	M	M	M	M
6	6	M	M	M	M	M	M	M	M	M	M	M
7	7	M	M	M	M	M	M	M	M	M	M	M
8	8	M	M	M	M	M	M	M	M	M	M	M
9	9	M	M	M	M	M	M	M	M	M	M	M
10	10	M	M	M	M	M	M	M	M	M	M	M
11	11	B	B	B	B	B	B	B	B	B	B	B
12	12	B	M	M	M	M	M	M	M	M	M	M
13	13	B	M	M	M	M	M	M	M	M	M	M
14	14	B	M	M	M	M	M	M	M	M	M	M
15	15	B	B	B	B	B	B	B	B	B	B	B
16	16	B	B	B	B	B	B	B	B	B	B	B
17	17	B	B	B	B	B	B	B	B	B	B	B
18	18	B	B	B	B	B	B	B	B	B	B	B
19	19	M	M	M	M	M	M	M	M	M	M	M
20	20	B	B	B	B	B	B	B	B	B	B	B

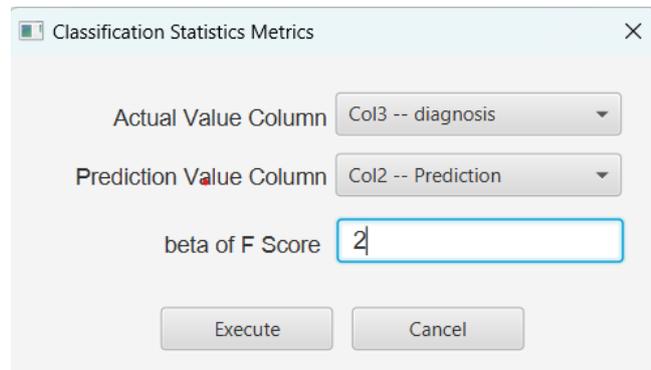
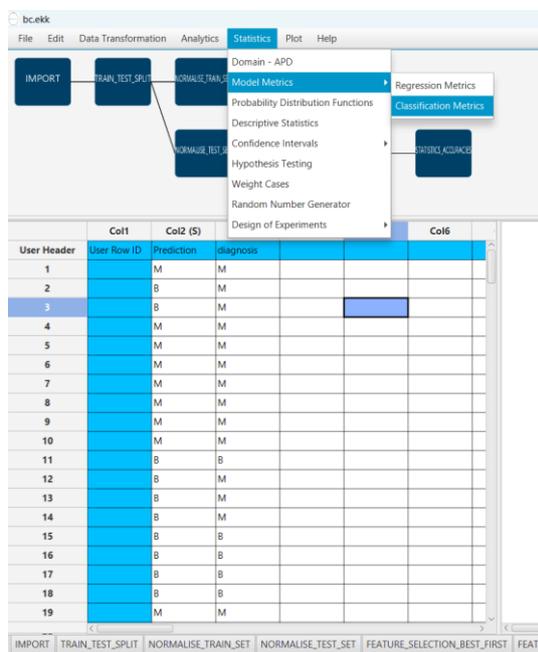
Step 10: Statistics calculation

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".



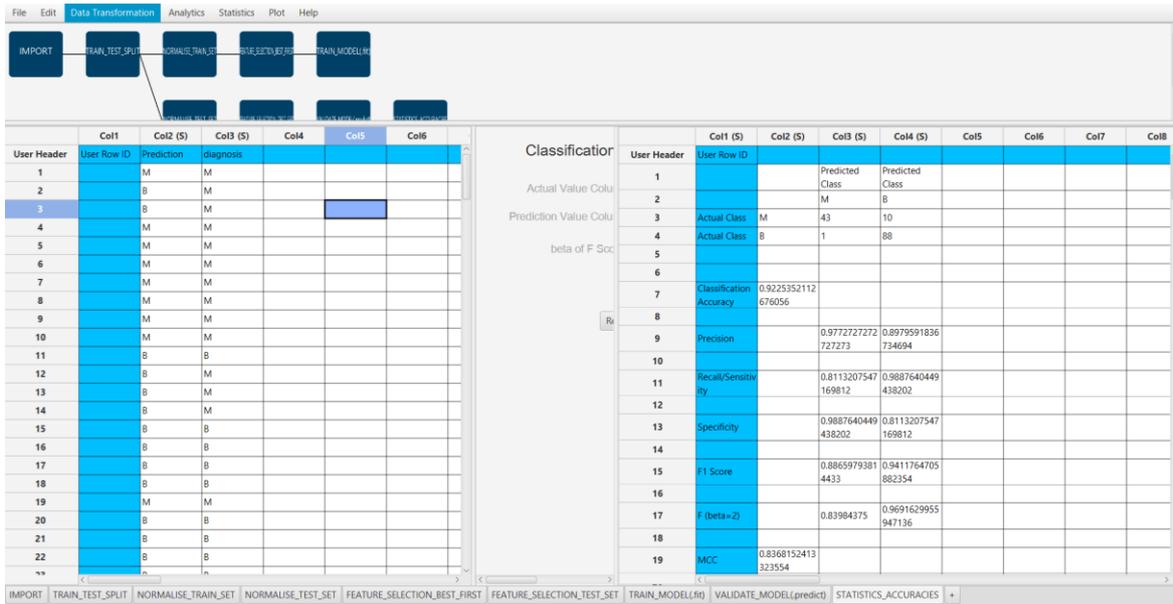
Calculate the statistical metrics for the classification by browsing: "Statistics" → "Model Metrics" → "Classification Metrics".



The results will appear on the output spreadsheet.

Accuracy: 0.923

F1-Score = 0.914

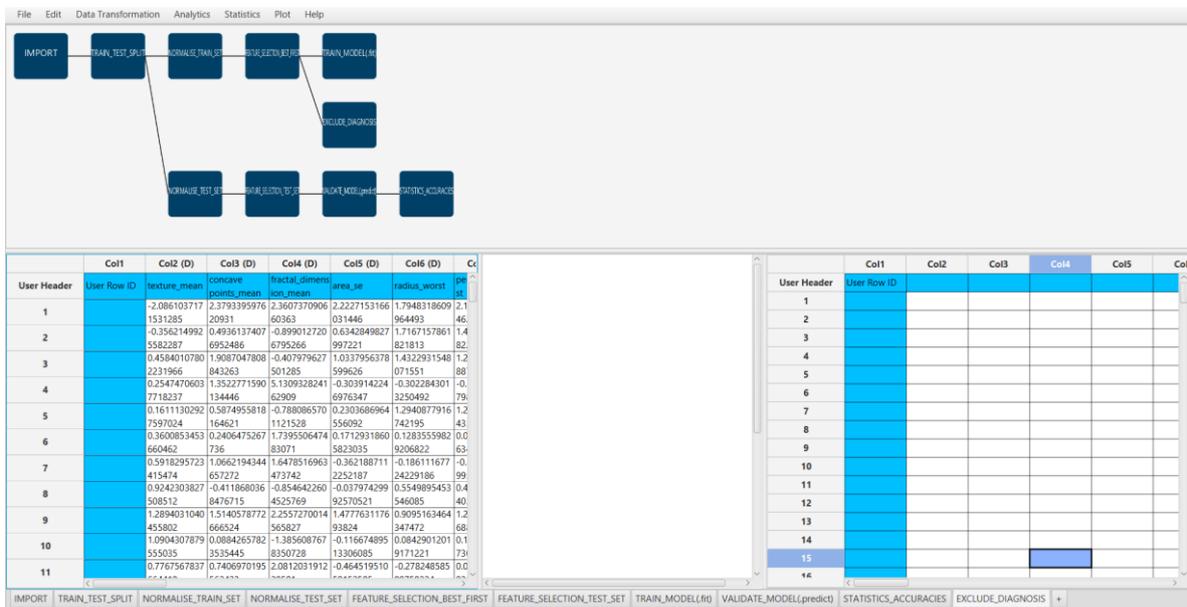


Step 11: Reliability check of each record of the test set

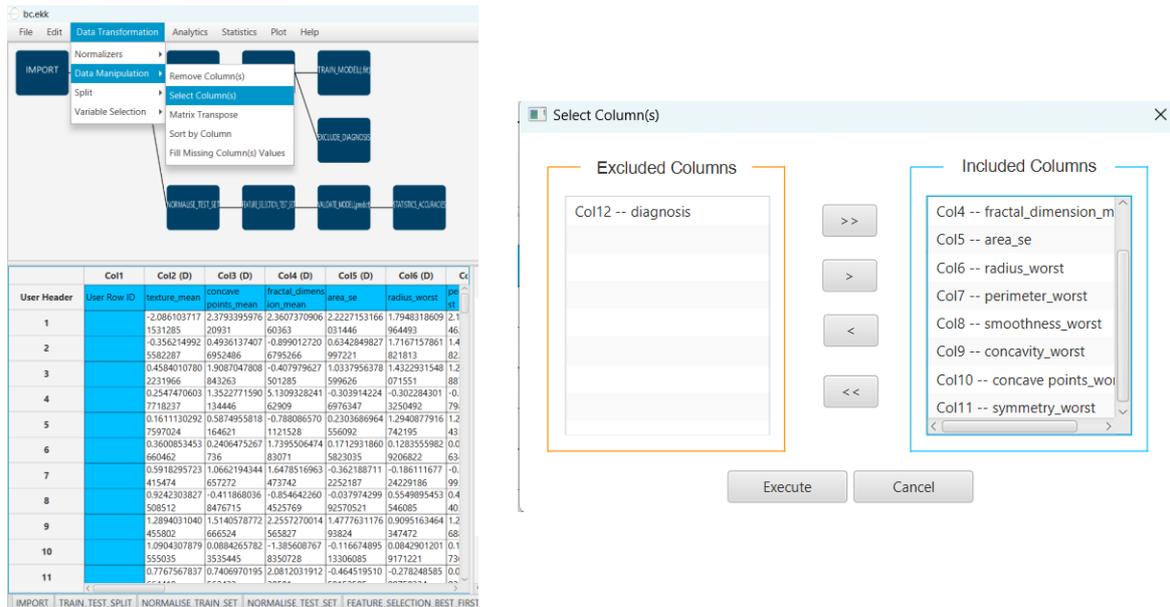
Step 11.a: Create the domain

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".



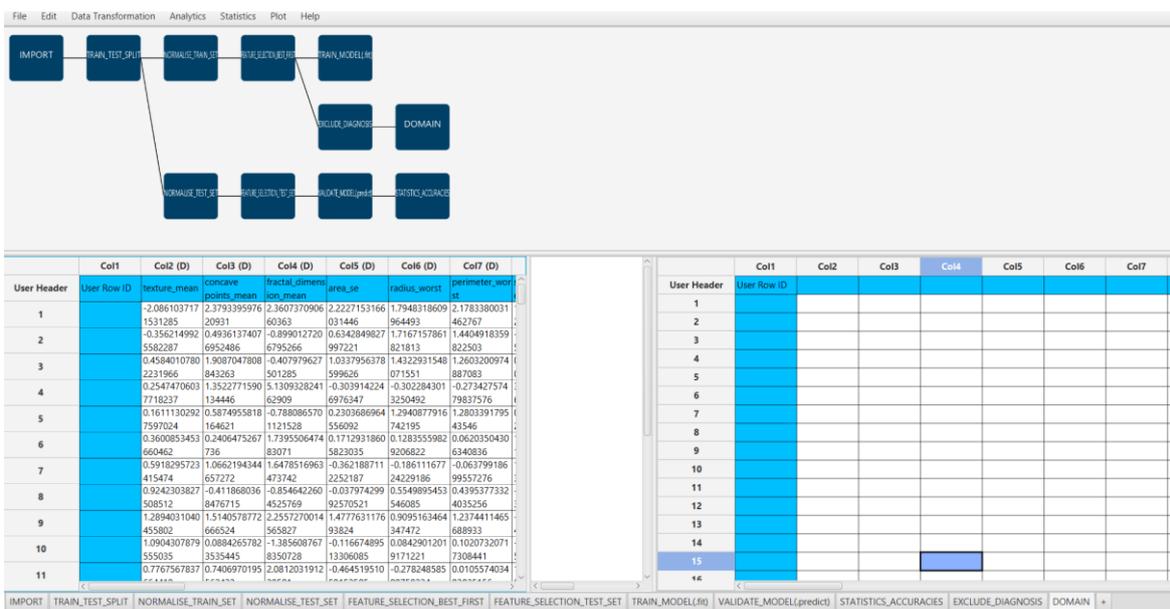
Manipulate the data to exclude the column that corresponds to the diagnosis by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except the "diagnosis".



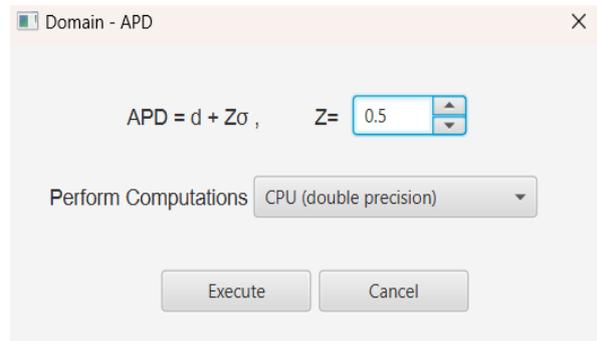
The results will appear on the output spreadsheet.

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

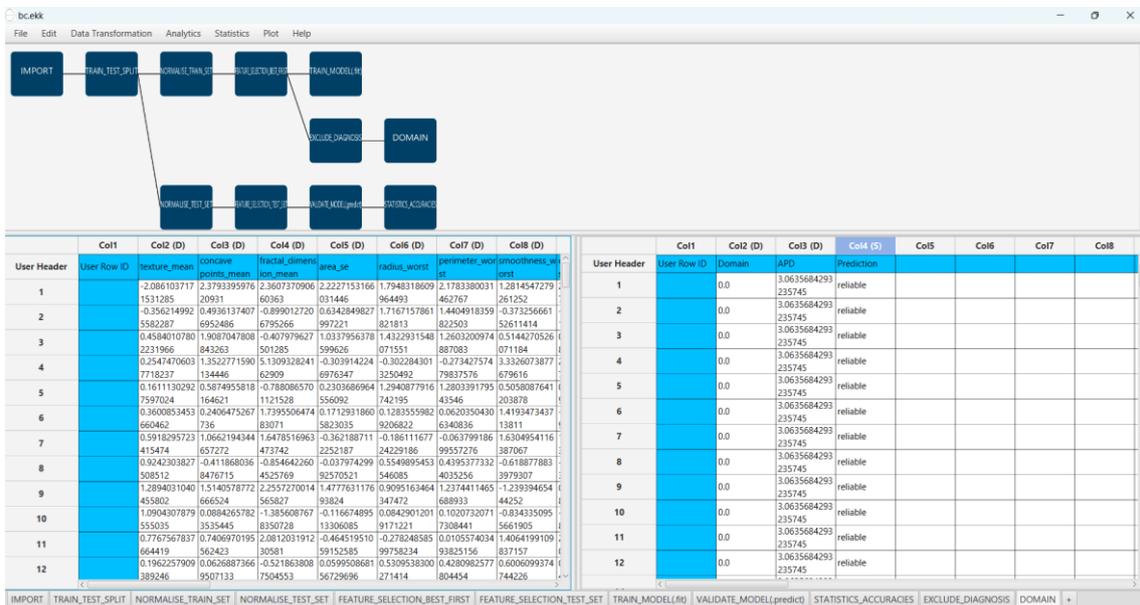
Import data into the input spreadsheet of the "DOMAIN" tab from the output of the "EXCLUDE_DIAGNOSIS" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".



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



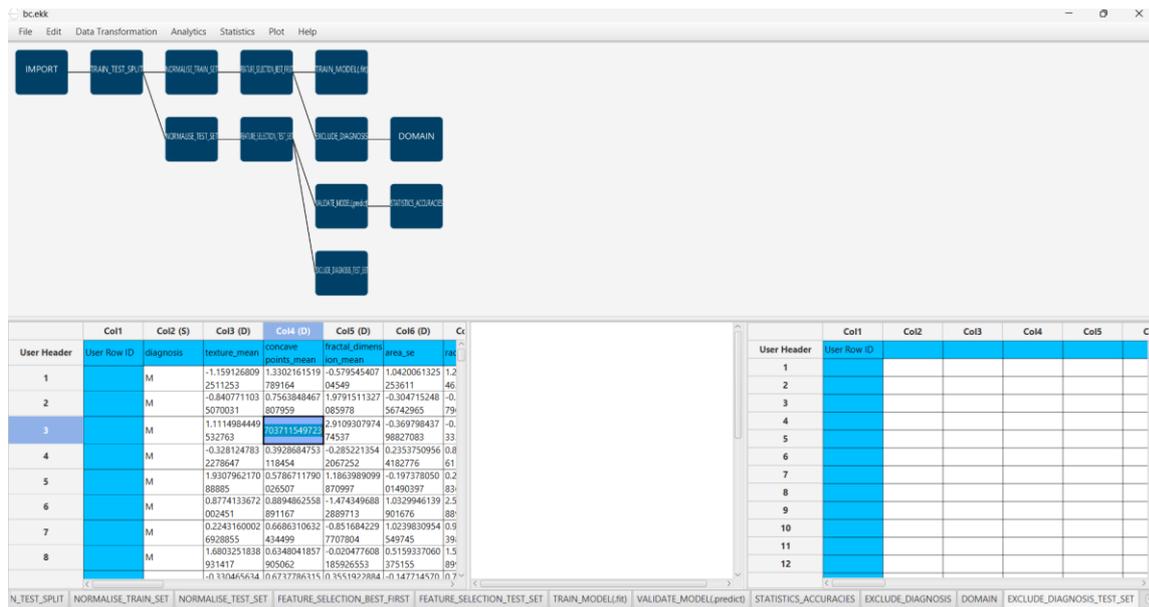
The results will appear on the output spreadsheet.



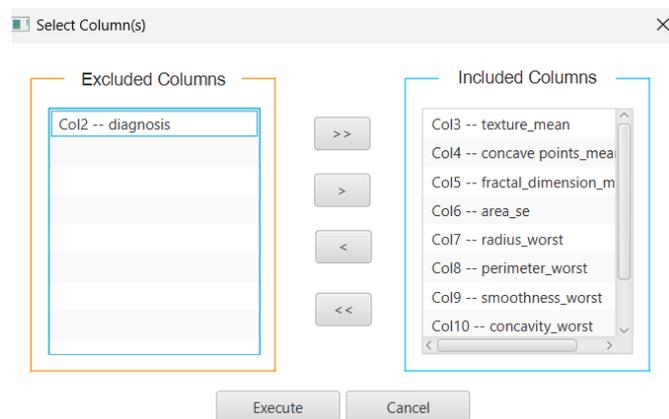
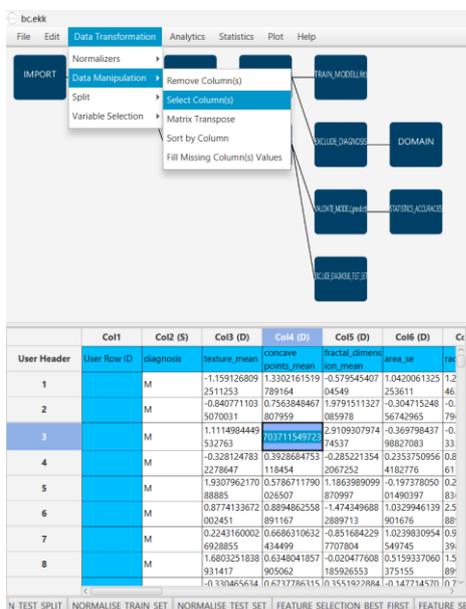
Step 11.b: Check the test set reliability

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 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”.

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)
1	-1.159126809	1.3302161519	-0.579545407	1.0420061325	1.2259865982	1.2	1.2
2	2511253	789164	04549	253611	463962	93	93
3	-0.84071103	0.7563848467	1.9791511327	-0.304715248	-0.190117629	-0.	-0.
4	5070031	807959	085978	56742965	7966696	21.	21.
5	1.1114984449	0.8676703711	2.9109307974	-0.369798437	-0.266230728	-0.	-0.
6	532763	549723	74537	98827083	3338487	95	95
7	-0.328124783	0.3928684753	-0.285221354	0.2353750956	0.8013556274	0.8	0.8
8	2278647	118454	2067252	4182776	611457	92.	92.
9	1.9307962170	0.5786711790	1.1863989099	-0.197378050	0.2084746493	0.4	0.4
10	88885	026507	870997	01490397	8362542	35.	35.
11	0.8774133672	0.8894862558	-1.474349688	1.0329946139	2.5539598700	2.2	2.2
12	002451	891167	2889713	901676	8895	55.	55.
13	0.2243160002	0.6686310632	-0.851684229	0.3095454974	0.9796205161	0.8	0.8

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

Existing Model Execution

Model: (from Tab:) DOMAIN

Type: APD Model

Description:

Model Input

- Header -> Datatype
- radius_mean -> Double
- texture_mean -> Double
- perimeter_mean -> Double
- area_mean -> Double
- smoothness_mean -> Double
- compactness_mean -> Double
- concavity_mean -> Double
- concave points_mean -> Double

Transfer Column(s) to Output

Execute Cancel

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

The screenshot shows the Isalos Analytics Platform interface. At the top, there is a menu bar with options: File, Edit, Data Transformation, Analytics, Statistics, Plot, Help. Below the menu is a workflow diagram with nodes: IMPORT, TRAIN_TEST_SPLIT, NORMALISE_TRAIN_SET, FEATURE_SELECTION_BEST_FIRST, TRAIN_MODEL(.fit), NORMALISE_TEST_SET, FEATURE_SELECTION_TEST_SET, EXCLUDE_DIAGNOSIS, DOMAIN, VALIDATE_MODEL(.predict), STATISTICS_ACCURACIES, EXCLUDE_DIAGNOSIS_TEST_SET, and RELIABILITY. Below the workflow is a data table with columns labeled Col1 through Col8. The table contains numerical data for 9 rows. At the bottom of the interface, there is a row of buttons corresponding to the workflow steps: NORMALISE_TRAIN_SET, NORMALISE_TEST_SET, FEATURE_SELECTION_BEST_FIRST, FEATURE_SELECTION_TEST_SET, TRAIN_MODEL(.fit), VALIDATE_MODEL(.predict), STATISTICS_ACCURACIES, EXCLUDE_DIAGNOSIS, DOMAIN, EXCLUDE_DIAGNOSIS_TEST_SET, and RELIABILITY.

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
1	-1.159126809	1.3302161519	0.0646	1.0420061325	1.2259865982	1.2517404908	0.2127869555	
2	2511253	789164	04549	253611	463962	937775	7155378	
3	-0.840771103	0.7563848467	1.9791511327	-0.304715248	-0.190117629	-0.143875515	2.0097001050	
4	5070031	807959	085978	36742965	7968996	21482445	54845	
5	1.1114984449	0.0676703711	2.9109307974	-0.369798437	-0.266230728	-0.308317974	2.2768670481	
6	532763	549723	74537	98827083	3338487	9510933	434867	
7	-0.328124783	0.3928684753	-0.285221354	0.2353750956	0.8013556274	0.8027410790	0.3075881289	
8	2278647	118454	2067252	4182776	611457	924124	255886	
9	1.9307962170	0.5786711790	1.1863898099	-0.197378050	0.2084746493	0.4481173398	1.5227668055	
0	88085	028507	870997	01480397	8362542	35283	545764	
1	0.8774133672	0.8894862550	-1.474349688	1.0329946139	2.5539598700	2.2755735445	0.3291338501	
2	002451	891167	2889713	901676	8895	554897	4241464	
3	0.2243160002	0.6686310632	-0.851684229	3.00954549745	0.9796205161	0.8999766205	0.0576577628	
4	6928855	434499	7707804	39859	016255	1040675		
5	1.6803251838	0.6348041857	-0.020477608	0.5159337060	1.5484657788	1.5148484264	0.7945214284	
6	931417	905062	185926553	375155	89912	716482	258569	
7	-0.330465634	0.6737786315	0.3551922884	-0.147714570	0.7592931256	0.6940660622	0.5230453410	
8	0053954	515065	0274587	08761582	380783	732919	93849	

Final Isalos Workflow

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

