



# Parkinson's Disease (Binary Classification)

The goal of this study is to train a model in order to predict whether a patient has Parkinson's Disease or not. The dataset used in this case study is found in <https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis/data>

and has 35 features and 2104 labelled samples. This dataset comprises comprehensive health information for 2,105 patients diagnosed with Parkinson's Disease, each uniquely identified with IDs ranging from 3058 to 5162. The dataset includes demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, symptoms, and a diagnosis indicator.

The dataset contains no missing values and includes several categorical features. Some of these features represent binary yes/no data, encoded as 0 for "No" and 1 for "Yes". Additionally, other categorical features contain multiple levels with corresponding numeric codes, as detailed below:

Gender:

- Male (0)
- Female (1)

Ethnicity:

- Caucasian (0)
- African American (1)
- Asian (2)
- Other (3)

Education Level:

- None (0)
- High School (1)
- Bachelor's (2)
- Higher (3)

## Step 1: Import data from file

Right click on the input spreadsheet and choose the option "Import from file". Then navigate through your files to load the one with the Parkinson's Disease data.

The screenshot shows the Isalos Analytics Platform interface. A context menu is open over a data grid, with the option 'Import from Spreadsheet' highlighted. The data grid displays a table with 12 columns labeled Col1 through Col12. The first row is labeled 'User Header'. The data rows contain various numerical values and some categorical entries like 'n' and '1'. The bottom of the screen shows a toolbar with icons for File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help, along with an 'IMPORT' button.

## Step 2: Manipulate data

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

- On the menu click on "Data Transformation" → "Data Manipulation" → "Select Column(s)".
- Select all columns except the ones that corresponds to the "PatientID" and "DoctorInCharge" columns.

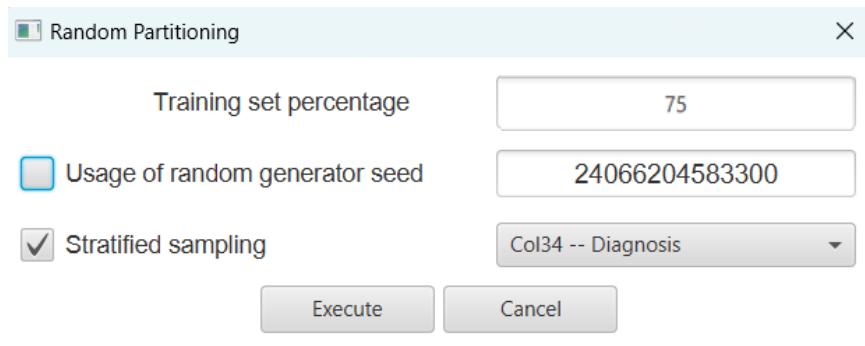
The data without the "PatientID" and "DoctorInCharge" columns will appear in the output spreadsheet.

## 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.

User Header	Col1	Col2 (I)	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholicConsumption	Col7 (I)	Col8 (D)	Col9
User Header	Col1	Col2 (I)	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholicConsumption	Col7 (I)	Col8 (D)	Col9
1	85	0	3	1	19.61987796	0	5.1082406	0	19.61987796	0	5.108240607	1.380
2	75	0	0	2	16.24733916	1	6.0276480	0	16.24733916	1	6.027648029	8.409E
3	70	1	0	0	15.36823871	0	2.2421353	0	15.45455733	0	5.997787563	1.375C
4	52	0	0	0	15.45455733	0	5.9977875	0	18.1604177	0	9.775242923	1.188
5	87	0	0	1	18.61604177	0	9.7752425	0	30.54200329	1	2.011281313	9.028E
6	68	1	2	1	39.42331141	1	13.59688	0	36.75828161	1	19.98865597	3.891
7	78	1	0	0	30.54200329	1	2.0112813	0	22.3805865	1	7.293287715	2.595E
8	70	1	0	0	36.75828161	1	19.98865	0	23.72709628	1	17.78290985	7.344E
9	80	0	2	1	22.3805865	1	7.2932877	0	38.48254474	0	6.639761993	7.8721
10	71	0	3	2	23.72708628	1	17.782905	0	30.22551208	0	3.762918066	4.311E
11	70	0	0	3	38.48254474	0	6.63976195	0	38.29830655	0	12.61599522	9.299
12	53	0	0	1	35.8967387	1	5.2129062	0	18.95878125	1	2.04712047	9.432E
13	74	0	2	2	30.22551208	0	3.762918C	0	28.04958007	1	7.260733832	4.311E
14	87	1	1	2	38.29830655	0	12.615995	0	21.85612305	0	0.255229809	4.040E
15	58	0	0	3	34.96532287	1	11.708597	0	22.54833861	0	11.57918488	8.893E
16	56	1	0	0	18.95878125	1	2.0471204	0	29.72724176	1	2.0324379	8.9341
17	54	1	0	0	28.04958007	1	7.2607338	0	33.24761783	1	9.545279952	6.9555
18	57	1	0	1	21.85612305	0	0.2552298	0	20.6172019	0	5.198675067	6.704E
19	51	1	0	0	19.0000405	0	1.532135	0	38.13363959	1	10.78134701	7.7201
20	55	0	1	2	22.54833861	0	11.579184	0	15.86360295	0	19.59717834	7.2423
21	62	1	0	2	29.72724176	1	2.0324379	0	36.90543424	0	9.89059796	7.6781
22	79	1	3	1	33.24761783	1	9.5452795	0	23.68997318	0	11.42519836	0.957C
23	74	1	0	1	20.6172019	0	5.1986750	0	25.89636742	0	17.90479619	9.869
24	60	1	n	n	38.13363959	1	10.78134701	n	36.78367747	1	14.06817123	3.067

## 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".

User Header	User Row ID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholConsumption	PhysicalActivity
1	85	0	3	1	19.61987796	0	5.108240607	1.380659917	
2	75	0	0	2	16.24733916	1	6.027648029	8.40980405	
3	52	0	0	0	15.5455733	0	5.997787563	1.375045164	
4	87	0	0	1	18.61604177	0	9.775242923	1.188607062	
5	78	1	0	0	30.5420329	1	2.011281313	9.028536304	
6	70	1	0	0	36.75828161	1	19.9886597	3.891748622	
7	80	0	2	1	22.3805865	1	7.293287715	2.595670177	
8	71	0	3	2	23.72708628	1	17.78290985	7.344899316	
9	70	0	0	3	38.48254474	0	6.639761993	7.872186697	
10	74	0	2	2	30.22551208	0	3.762910066	4.316651243	
11	87	1	1	2	38.29830655	0	12.6159952	9.299289996	
12	56	1	0	0	18.95878125	1	2.04712047	9.432830343	
13	54	1	0	0	28.04958007	1	7.260733832	4.311674743	
14	57	1	0	1	21.85612305	0	0.255229809	4.04096502	
15	55	0	1	2	22.54833861	0	11.57918488	8.893662605	
16	62	1	0	2	29.72724176	1	2.03243796	8.934190297	
17	79	1	3	1	33.24761783	1	9.545279952	6.95599901	
18	74	1	0	1	20.6172019	0	5.198675067	6.704484305	
19	60	1	0	0	38.13363959	1	10.78134701	7.720134324	
20	71	1	2	1	15.86360295	0	19.59171834	7.242315106	
21	79	1	1	2	36.90543424	0	9.89059796	7.678179303	
22	66	0	0	2	23.68997318	0	11.42519836	0.957084433	
23	78	0	1	3	25.89636742	0	17.90479619	9.86938435	
24	61	n	n	2	36.76372477	1	2.608817132	2.987486421	

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

User Header	User Row ID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholConsumption	PhysicalActivity
1	85	0	3	1	19.61987796	0	5.108240607	1.380659917	
2	75	0	0	2	16.24733916	1	6.027648029	8.40980405	
3	52	0	0	0	15.5455733	0	5.997787563	1.375045164	
4	87	0	0	1	18.61604177	0	9.775242923	1.188607062	
5	78	1	0	0	30.5420329	1	2.011281313	9.028536304	
6	70	1	0	0	36.75828161	1	19.9886597	3.891748622	
7	80	0	2	1	22.3805865	1	7.293287715	2.595670177	
8	71	0	3	2	23.72708628	1	17.78290985	7.344899316	
9	70	0	0	3	38.48254474	0	6.639761993	7.872186697	
10	74	0	2	2	30.22551208	0	3.762910066	4.316651243	
11	87	1	1	2	38.29830655	0	12.6159952	9.299289996	
12	56	1	0	0	18.95878125	1	2.04712047	9.432830343	
13	54	1	0	0	28.04958007	1	7.260733832	4.311674743	
14	57	1	0	1	21.85612305	0	0.255229809	4.04096502	
15	55	0	1	2	22.54833861	0	11.57918488	8.893662605	
16	62	1	0	2	29.72724176	1	2.03243796	8.934190297	
17	79	1	3	1	33.24761783	1	9.545279952	6.95599901	
18	74	1	0	1	20.6172019	0	5.198675067	6.704484305	
19	60	1	0	0	38.13363959	1	10.78134701	7.720134324	
20	71	1	2	1	15.86360295	0	19.59171834	7.242315106	
21	79	1	1	2	36.90543424	0	9.89059796	7.678179303	
22	66	0	0	2	23.68997318	0	11.42519836	0.957084433	
23	78	0	1	3	25.89636742	0	17.90479619	9.86938435	
24	61	n	n	2	36.76372477	1	2.608817132	2.987486421	

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 by browsing: "Analytics" → "Existing Model Utilization" → "Model (from Tab: ) NORMALISE\_TRAIN\_SET".

The screenshot shows the Isalos Analytics Platform interface. The top navigation bar includes File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help. The Analytics menu is open, with Existing Model Utilization highlighted. Below the menu is a data spreadsheet with columns labeled Col1 through Col8. The first row is labeled "User Header". The data rows contain various numerical values. At the bottom of the spreadsheet area, there are tabs for IMPORT, TRAIN\_TEST\_SPLIT, NORMALISE\_TRAIN\_SET, and NORMALISE\_TEST\_SET.

The results will appear on the output spreadsheet.

The screenshot shows the Isalos Analytics Platform interface. The top navigation bar includes File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help. The Analytics menu is open, with NORMALISE\_TRAIN\_SET highlighted. Below the menu is a data spreadsheet with columns labeled Col1 through Col9. The first row is labeled "User Header". The data rows contain various numerical values. At the bottom of the spreadsheet area, there are 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\_REGRESSION".

Import data into the input spreadsheet of the "FEATURE\_SELECTION\_REGRESSION" 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 Regression Analysis by browsing: "Data Transformation" → "Variable Selection" → "Regression Analysis". Then choose the "Diagnosis" column as the intercept column, the Significance level ( $\alpha$ ) as 0.05 and include all columns.

The results will appear on the output spreadsheet.

	Col1	Col2 (S)	Col3 (S)	Col4 (S)	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (S)
User Header	User Row ID							
1		Regression Statistics						
2		Multiple R	0.6427533978429012					
3		R Square	0.41313193043859486					
4		Adjusted R Square	0.4009845965278801					
5		Standard Error	0.37590758946897873					
6		Observations	1579					
7								
8		Regression	df	SS	MS	F	Significance F	
9		Residual	32	153.78711830 863594	4.8058474471 44873	34.010090895 26929	4.0587245879 37507E-154	
10		Total	1546	218.45987345 83048	0.1413065158 2037826			
11			1578	372.24699176 69407				
12		Coefficients		Standard Error	t Stat	P-value	Lower 95.0%	Upper 95.0%
13		Diagnosis	0.6193793540215377	0.0094559758 7989092	65.473671591 2937	0.0	0.6008236148 982906	0.6379350931 447847
14		Age	0.034944578344872954	3.6566144380 61688421	2.6412976725 26531	0.0161994307 6839E-4	0.0536897259 7307164	1667427
15		Gender	0.005905157055109047	0.0095891630 32430095	0.6158156906 017851	0.5381068188 904652	-0.012903982 619157102	0.0247142967 29375196
16		Ethnicity	-0.0051288114570836344	0.0095904324 85275994	-0.534784167 9671693	0.5928760144 825591	-0.023940441 1626286	0.0136828182 4846133
17		EducationLevel	-0.005631677102695146	0.0095434132 79615323	0.5552021963 7773302	0.024351078 015052	0.0130877244 654056724	0.024351078 48666432
18		BMI	0.006842640052181735	0.0096094414 97752863	0.7120746875 645025	0.4765259676 099952	0.012006275 824172727	0.0256915559 28536197
19		Smoking	-0.01068028611694388	0.0095538083 76531577	-1.117908764 339406	0.2637796445 924609	-0.029420077 64698911	0.0080595054 13101352
20		AlcoholConsumption	0.002461167657840956	0.0095636829 79952696	0.2573451737 1602894	0.7969465694 74437	-0.016297992 903093382	0.0212203282 18775293
21		PhysicalActivity	-0.00933461700271854	0.0095689073 55141496	-0.975488773 569992	0.3294704756 939527	-0.028103769 871170126	0.0094350464 70626419
22		DietQuality	-0.003050220046770803	0.0095468255 64645988	-0.319500971 93211996	0.7493898214 594091	-0.021776314 79393405	0.0156758747 0392445
23		SleepQuality	-0.021731024730421666	0.0095747622 37808862	-2.269615076 6658314	0.0233681600 34877315	-0.040511917 25146546	-0.002950132 2093778704
24		FamilyHistoryParkinsons	0.0101209008340667	0.0095524701 04518592	1.0590561511 1365	0.2895348303 630602	-0.008616265 675929415	0.0288580673 44062814
25		TraumaticBrainInjury	0.014589229732617963	0.0095958049 86224412	1.5203758052 150949	0.1286210332 7246324	-0.004232938 131512603	0.0334113975 9674853
26		Hypertension	0.006973590608428447	0.0095620828 29304448	0.7292961934 0431	0.4569309710 9648995	-0.011782431 257612396	0.0257296124 74469293
27		Diabetes	0.02562743849956757	0.0095871203 43858281	2.6731111721 138525	0.0075941977 85412139	-0.0068223055 58165209	0.0444325714 4096993
28		Depression	0.02723933473557805	0.0095605026 52271166	2.8491529918 7822	0.0044416491 5449658	-0.0084864123 86192469	0.045992570 8496363
29		Stroke	0.012415701060811338	0.0095488124 83364104	1.3002350902 211048	0.1937142092 2437116	-0.006314291 026668686	0.0311456931 4829136
30		SystolicBP	0.00528704609413583	0.0095272229 8344163	0.5549409416 918421	0.5790153341 23407	-0.013400598 197382648	0.0239746903 8565431
31		DiastolicBP	-0.003521415069752083	0.0095773887 41574176	-0.367680081 1546979	0.7131621231 346639	-0.022307459 476941696	0.0152646293 37437531
32		CholesterolTotal	0.007258067669105708	0.0095513189 59344359	0.7599021349 826152	0.4474289458 867108	-0.011476840 870064977	0.0259929762 80276394
33		CholesterolLDL	0.0015268790166764238	0.0095524401 46575041	0.1598417784 9762077	0.8730265930 105581	-0.017210228 855084888	0.0202639867 87645944
34		CholesterolHDL	5.411733944502038E-4	0.0095687599 01534996	0.0565562727 0607868	0.9549059762 495937	-0.018227945 54625449	0.0193102923 35154897
35		CholesterolTriglycerides	0.002268052516280527	0.0095630160 46575041	0.2371691634 9762077	0.8125569806 105581	-0.016489799 598043E-23	0.0210259048 4300933
36		UPDRS	0.18630288121195493	0.0096038801 88781513	19.3987094318 045518	3.3396373538 02613E-75	0.1674648738 4105188	0.2051408885 8285798
37		MoCA	-0.07109273757277541	0.0095563178 9932587	7.4393441404 884063	1.6703413366 346027E-13	-0.0089837451 53083589	-0.052348023 61471493
38		FunctionalAssessment	-0.11419317166623748	0.0095658700 57560244	-11.93756249 8665641	1.7403934101 474355E-31	-0.132956622 78414E-37	-0.095429721 4295439
39		Tremor	0.12423615764298186	0.0095515200 61017858	13.0069514493 259872	8.8124775789 78414E-37	0.1055008546 4295439	0.1429714606 4300933
40		Rigidity	0.10336739579051336	0.00955057635 77577548	10.196505303 891393	1.9389201436 52925E-26	0.0846625607 3165018	0.1220722308 4937655
41		Bradykinesia	0.09738441151376426	0.00955057635 81646275	1.1315824768 608197	0.0786505923 598043E-23	0.1161182306 4773905	0.1161182306 7978946
42		PosturalInstability	0.08795422774590572	0.0095343098 65763344	9.2250229942 43103	9.0117242673 54057E-20	0.0692526825 3738402	0.1066557729 5442742
43		SpeechProblems	-0.0053249866410449145	0.0095760103 43971408	-0.556075698 5185661	0.5782395755 011953	-0.024108327 321856172	0.0134583540 39766342
44		SleepDisorders	-0.01601764859514722	0.0095591144 87731708	-1.675641464 0398416	0.0940106527 4412763	-0.034767848 06031181	0.0027325508 70017371

The significant features according to the p-value are the following:

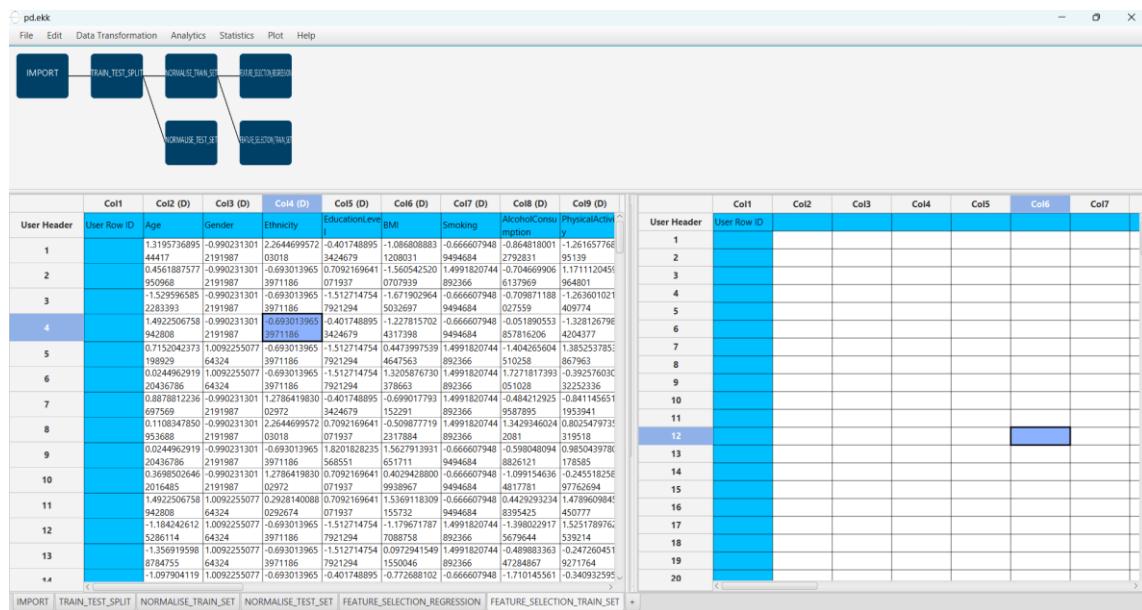
- Diagnosis (p-value = 0.0)

- Age (p-value = 2.64129767256839E-4)
- SleepQuality (p-value = 0.023368160034877315)
- Diabetes (p-value = 0.007594197785412139)
- Depression (p-value = 0.004441649154449658)
- UPDRS (p-value = 3.339637353802613E-75)
- MoCA (p-value = 1.6703413366346027E-13)
- FunctionalAssessment (p-value = 1.7403934101474355E-31)
- Tremor (p-value = 8.812477578978414E-37)
- Rigidity (p-value = 1.938920143652925E-26)
- Bradykinesia (p-value = 1.1315824768598043E-23)
- PosturalInstability (p-value = 9.011724267354057E-20)

## Step 7: Feature selection: train set

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

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



Manipulate the data by choosing the columns that correspond to the significant features (from the previous step): "Data Transformation" → "Data Manipulation" → "Select Column(s)".

The screenshot shows the Isalos Analytics Platform's Data Transformation interface. A context menu is open under 'Data Manipulation' with 'Select Column(s)' selected. A modal window titled 'Select Column(s)' is displayed, showing two lists: 'Excluded Columns' and 'Included Columns'. The 'Excluded Columns' list contains: Col2 -- Gender, Col4 -- Ethnicity, Col5 -- EducationLevel, Col6 -- BMI, Col7 -- Smoking. The 'Included Columns' list contains: Col2 -- Age, Col11 -- SleepQuality, Col15 -- Diabetes, Col16 -- Depression, Col24 -- UPDRS, Col25 -- MoCA, Col26 -- FunctionalAssessment, Col10 -- DietQuality.

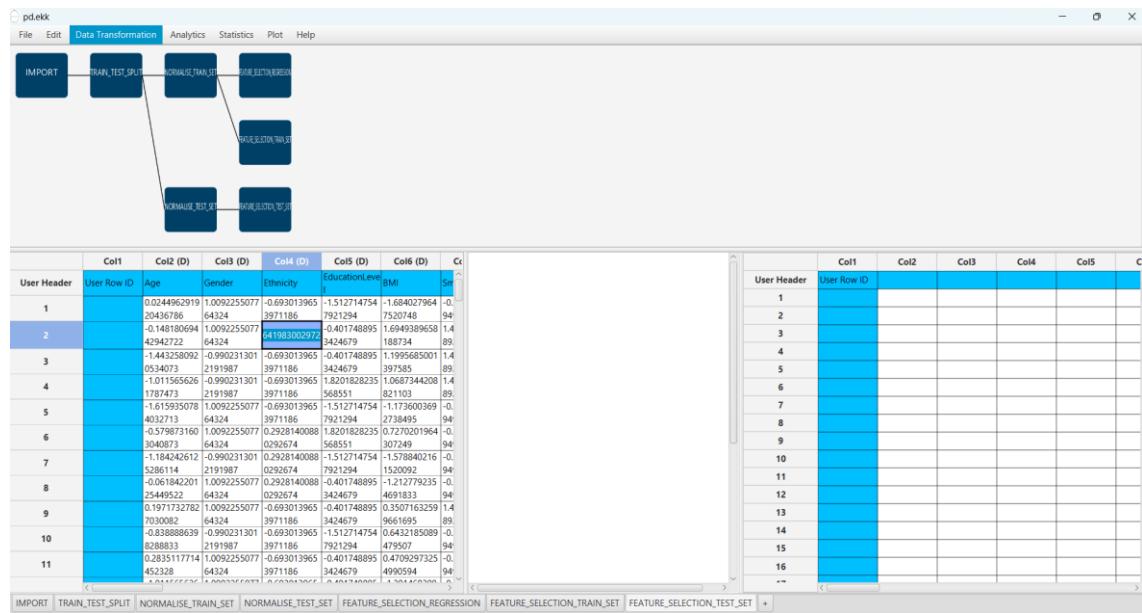
The results will appear on the output spreadsheet.

The screenshot shows the Isalos Analytics Platform's Data Transformation interface. A context menu is open under 'Data Transformation' with 'NORMALISE\_TRAIN\_SET' selected. A modal window titled 'NORMALISE TRAIN SET' is displayed, showing three tabs: 'NORMALISE TRAIN SET', 'FEATURE\_SELECTION\_REGRESSION', and 'FEATURE\_SELECTION\_TRAIN\_SET'. The 'NORMALISE TRAIN SET' tab is active.

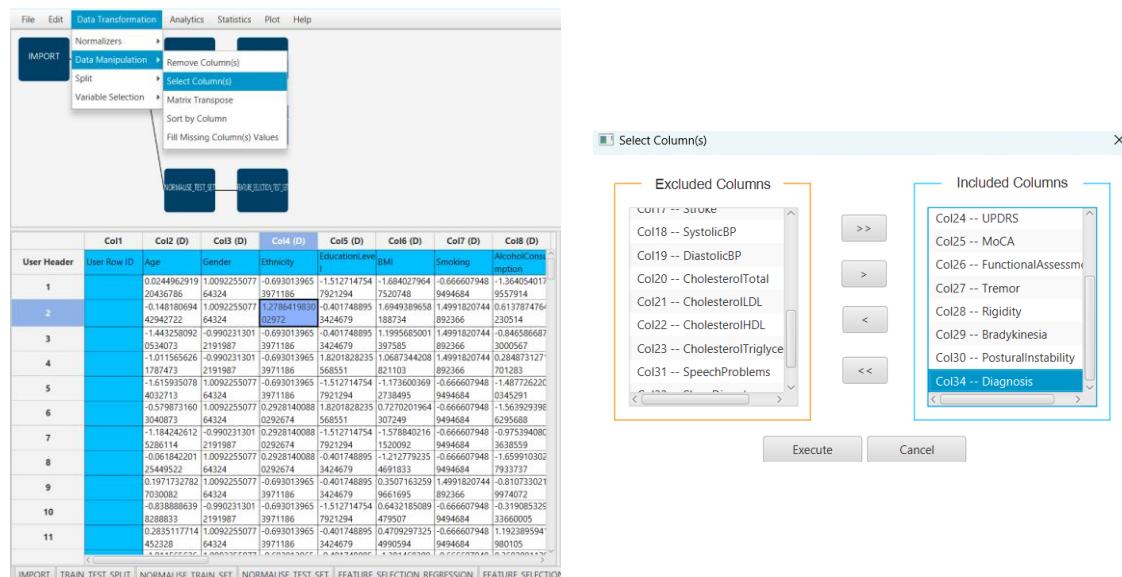
## Step 8: Feature selection: test set

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

Import data into the input spreadsheet of the "FEATURE\_SELECTION\_TEST\_SET" tab from the output of the "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 step 7): "Data Transformation" → "Data Manipulation" → "Select Column(s)".



The results will appear on the output spreadsheet.

The screenshot shows a Data Transformation workspace with the following components:

- IMPORT**: A blue rectangular node.
- TRAIN\_TEST\_SPLIT**: A blue rectangular node.
- NORMALISE\_TRAIN\_SET**: A blue rectangular node.
- FEATURE\_SELECTION\_REGRESSION**: A blue rectangular node.
- FEATURE\_SELECTION\_TRAIN\_SET**: A blue rectangular node.
- FEATURE\_SELECTION\_TEST\_SET**: A blue rectangular node.
- NORMALISE\_TEST\_SET**: A blue rectangular node.
- TRAIN\_MODEL**: A blue rectangular node.

Arrows indicate the flow from IMPORT to TRAIN\_TEST\_SPLIT, then to NORMALISE\_TRAIN\_SET, then to FEATURE\_SELECTION\_REGRESSION, then to FEATURE\_SELECTION\_TRAIN\_SET, then to FEATURE\_SELECTION\_TEST\_SET, and finally to NORMALISE\_TEST\_SET, which then connects to TRAIN\_MODEL.

The output spreadsheet has 11 rows (User Row ID 1 to 11) and columns labeled Col1 through Col8. The first few rows of data are as follows:

User Header	User Row ID	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
1	0.0244962919	1.0929255077	-0.693013965	1.512714754	-1.68427954	-0.666607948	-1.364054017		
2	20436765	64324	3971186	7921294	7520348	9494684	9557914		
3	1.489180694	1.0929255077	-0.693013965	1.401748895	1.6949389585	1.4991802744	0.613787476		
4	42942722	64324	3971186	3424679	188734	892366	230514		
5	-1.443258092	-0.990231301	-0.693013965	0.401748895	1.995685001	1.4991802744	-0.846586687		
6	0.534073	2191987	3971186	3424679	397585	892366	300567		
7	1.011565626	-0.990231301	-0.693013965	1.820182835	1.0687344208	1.4991802744	0.284873127		
8	1787473	2191987	3971186	568551	821103	892366	70126		
9	-1.615953078	1.0929255077	-0.693013965	-1.512714754	-1.173600369	-0.666607948	-1.48772622		
10	4032713	64324	3971186	7921294	728494	9494684	0.35291		
11	-0.579873160	1.0929255077	-0.9928140086	1.820182835	0.7270201964	-0.666607948	-1.563923938		

## Step 9: Train the model

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

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

The screenshot shows a Data Transformation workspace with the following components:

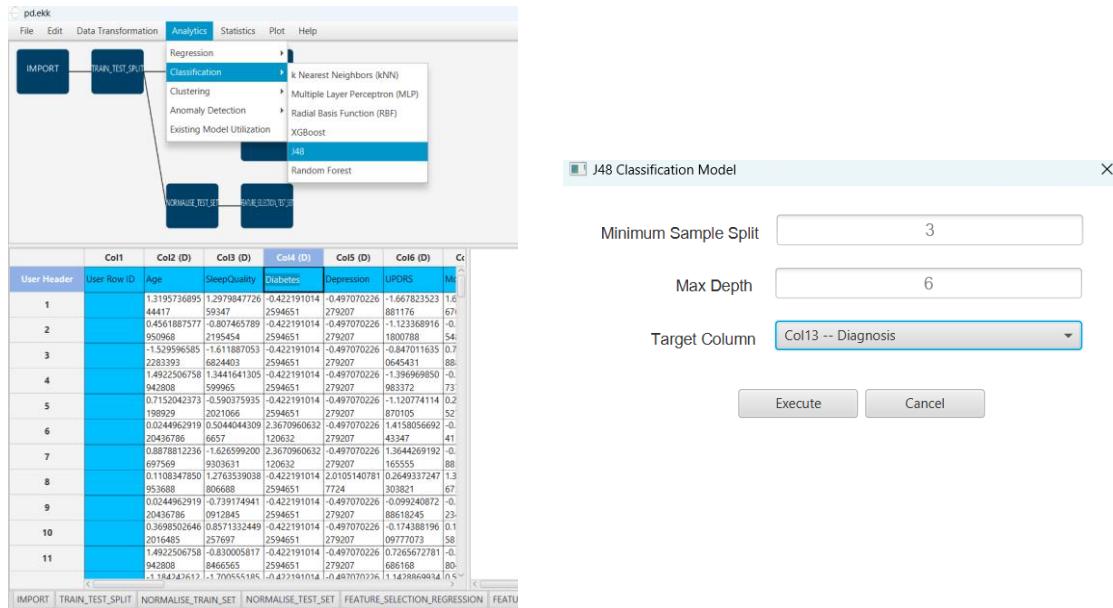
- IMPORT**: A blue rectangular node.
- TRAIN\_TEST\_SPLIT**: A blue rectangular node.
- NORMALISE\_TRAIN\_SET**: A blue rectangular node.
- FEATURE\_SELECTION\_REGRESSION**: A blue rectangular node.
- FEATURE\_SELECTION\_TRAIN\_SET**: A blue rectangular node.
- FEATURE\_SELECTION\_TEST\_SET**: A blue rectangular node.
- NORMALISE\_TEST\_SET**: A blue rectangular node.
- TRAIN\_MODEL**: A blue rectangular node.

Arrows indicate the flow from IMPORT to TRAIN\_TEST\_SPLIT, then to NORMALISE\_TRAIN\_SET, then to FEATURE\_SELECTION\_REGRESSION, then to FEATURE\_SELECTION\_TRAIN\_SET, then to FEATURE\_SELECTION\_TEST\_SET, and finally to NORMALISE\_TEST\_SET, which then connects to TRAIN\_MODEL.

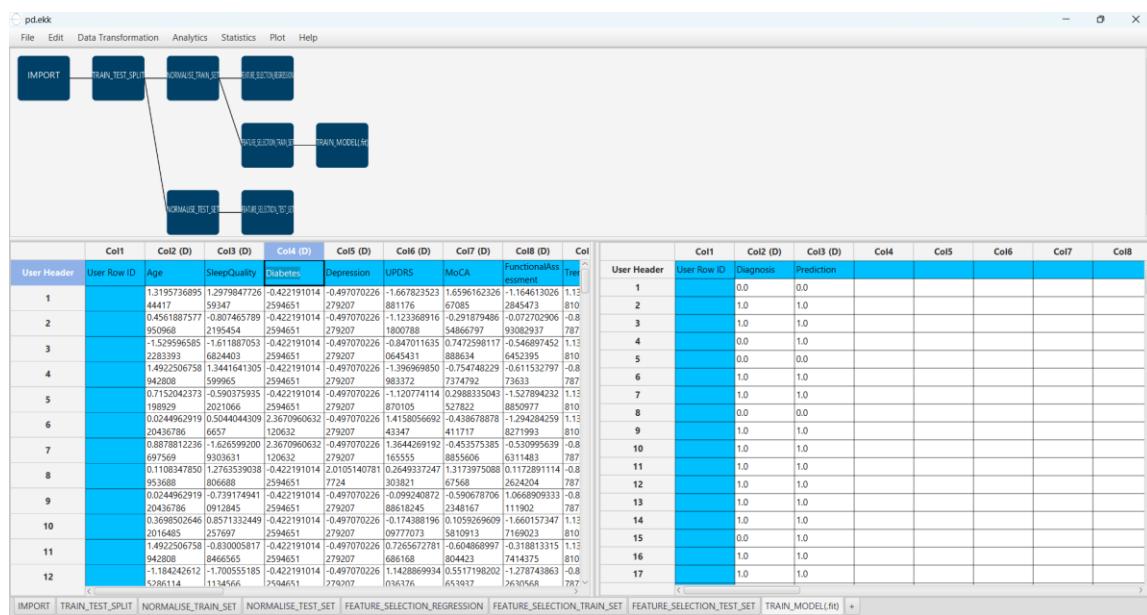
The output spreadsheet has 16 rows (User Row ID 1 to 16) and columns labeled Col1 through Col5. The first few rows of data are as follows:

User Header	User Row ID	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)
1	1.3195736895	1.2979847726	-0.422191014	-0.497070226	-1.667823523	
2	44417	59347	2594651	279207	881176	
3	1.4561887577	-0.807465789	-0.422191014	-0.497070226	-1.23368916	
4	950968	2195484	2594651	279207	180078	
5	-1.529596585	-1.611887053	-0.422191014	-0.497070226	-0.847011635	
6	2283399	6824400	2594651	279207	0645431	
7	1.4922506758	1.346141305	-0.422191014	-0.497070226	-1.369699650	
8	44406	-0.59075955	-0.422191014	-0.497070226	-0.847011635	
9	5713024273	-0.59075955	-0.422191014	-0.497070226	-1.120774114	
10	189829	2031066	2594651	279207	870105	
11	0.0244962919	0.9928140086	2.3670960623	-0.497070226	1.4158056692	

Use the J48 Method to train and fit the model by browsing: "Analytics" → "Classification" → "J48" and set the "Minimum Sample Split" as 3, the "Max Depth" as 6 and the "Target Column" as the column corresponding to "Diagnosis".



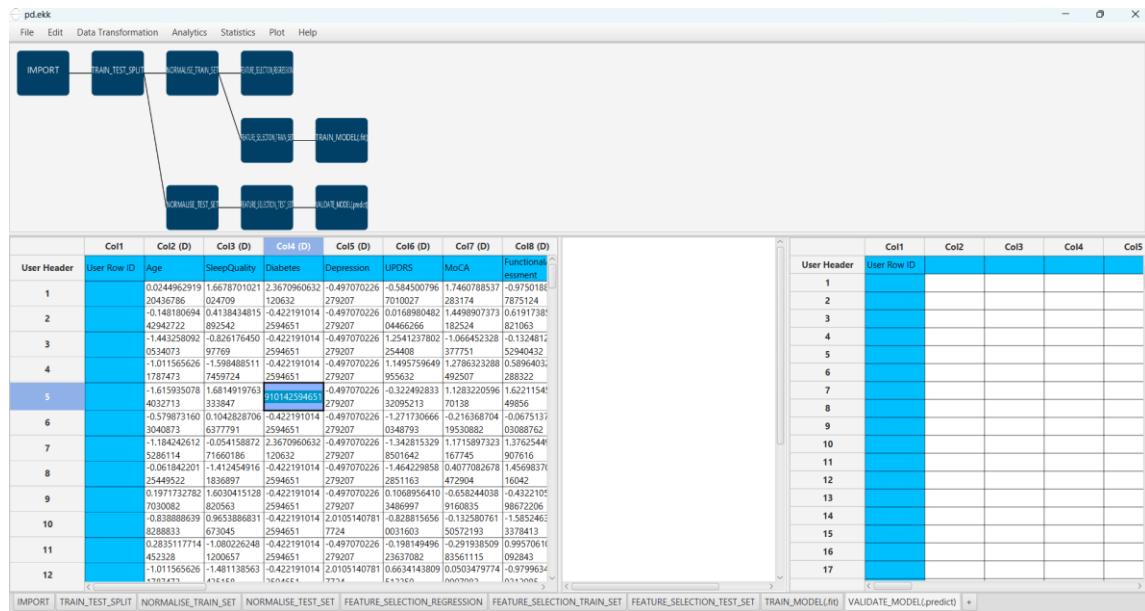
The predictions will appear on the output spreadsheet.



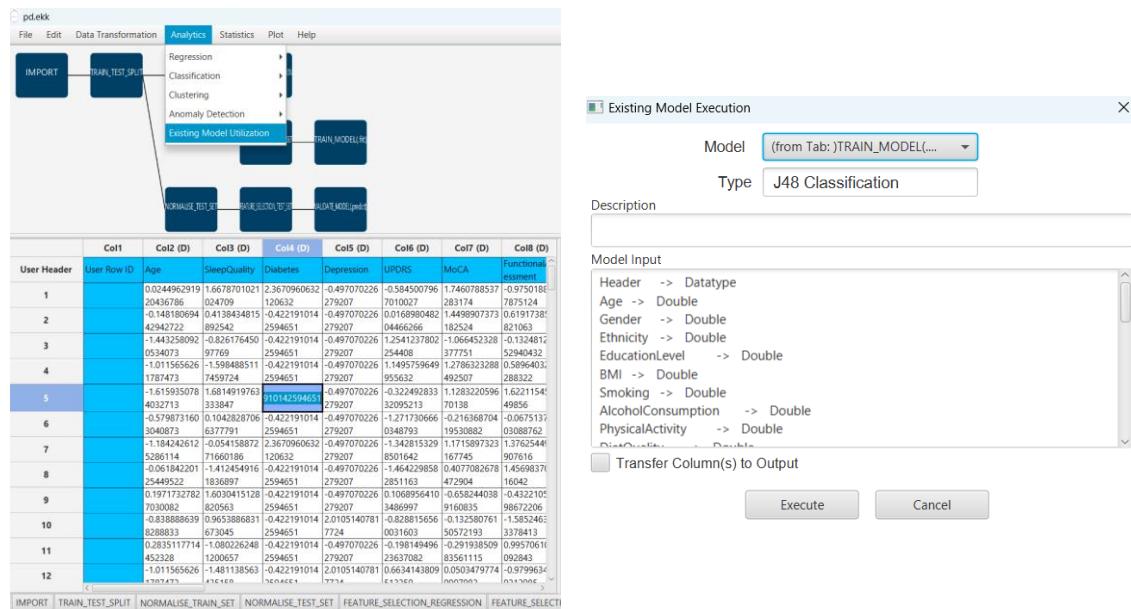
## Step 10: Validate the model

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

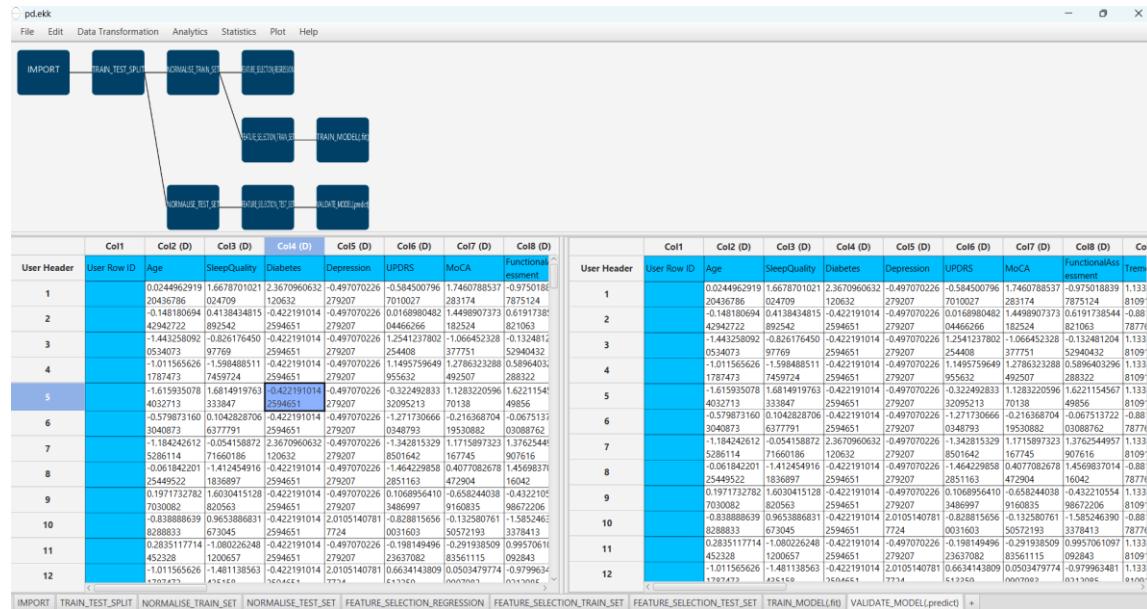
Import data into the input spreadsheet of the "VALIDATE\_MODEL(.predict)" tab from the output of the "FEATURE\_SELECTION\_TEST\_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



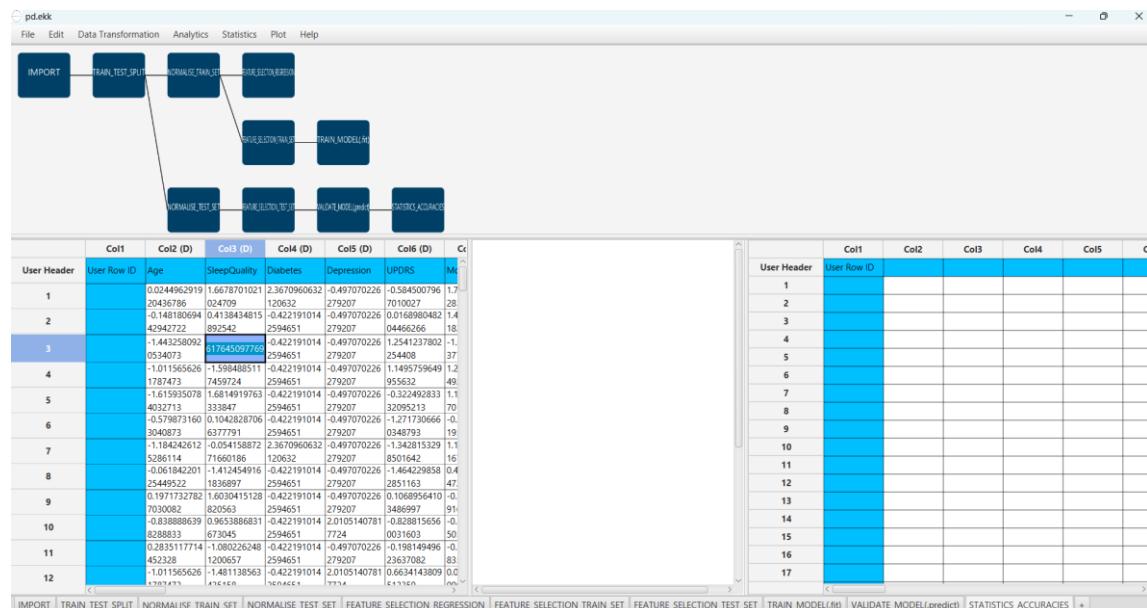
The predictions will appear on the output spreadsheet.



## Step 11: Statistics calculation

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

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



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

The screenshot shows the pd.ekk software interface. At the top, there's a menu bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help. Below the menu is a toolbar with IMPORT, TRAIN\_TEST\_SPLIT, NORMALISE\_TRAIN\_SET, and NORMALISE\_TEST\_SET buttons. A main workspace contains a data flow diagram with nodes: IMPORT, TRAIN\_TEST\_SPLIT, NORMALISE\_TRAIN\_SET, and several statistical analysis nodes like Model Metrics, Probability Distribution Functions, Descriptive Statistics, Confidence Intervals, Hypothesis Testing, Weight Cases, Random Number Generator, Design of Experiments, and Feature Selection Regression.

A modal dialog titled "Classification Statistics Metrics" is open. It has two dropdown menus: "Actual Value Column" set to "Col13 -- Diagnosis" and "Prediction Value Column" set to "Col14 -- Prediction". Below these is a text input field labeled "beta of F Score" containing the value "2". At the bottom are "Execute" and "Cancel" buttons.

User Header	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7	Col8	Col9	Col10
1	0.0244962919	1.6678701021	2.3670960632	-0.497070226	-0.584500796	1.7				
2	20436786	024709	120632	279207	7010027					
3	-0.148180694	0.4138434815	-0.422191014	-0.497070226	0.0168980482	1.4				
4	42942722	892542	2594651	279207	0466266					
5	-1.44325092	017645097769	-0.422191014	-0.497070226	1.2541237802	-1.				
6	0534073	2594651	279207	254408						
7	-1.184242612	-0.054158872	2.3670960632	-0.497070226	1.342815329	1.1				
8	5286114	7166018	120632	279207	8501642					
9	-0.061842201	-1.412454916	-0.422191014	-0.497070226	-1.464224858	0.4				
10	25449522	1836897	2594651	279207	2851163					
11	0.1971732782	1.603041512	-0.422191014	-0.497070226	0.1068956410					
12	7030082	020563	2594651	279207	3405103					
13	-0.03880639	0.4138434815	-0.422191014	-0.497070226	0.0168980482	0.0				
14	8288833	073045	2594651	7724	0031603					
15	0.2033117714	-1.08032648	-0.422191014	-0.497070226	-0.198149496	0.				
16	452328	120657	2594651	279207	23637082	83				
	-1.011565626	-1.481138563	-0.422191014	-0.497070226	0.664143809	0.0				
	1709473	2594651	7724	037566						

The results will appear on the output spreadsheet.

Accuracy: 0.93

F1-Score = 0.9199

The screenshot shows the pd.ekk software interface with a similar layout to the previous one. The data flow diagram includes nodes for IMPORT, TRAIN\_TEST\_SPLIT, NORMALISE\_TRAIN\_SET, and NORMALISE\_TEST\_SET. The main workspace displays a detailed classification statistics table.

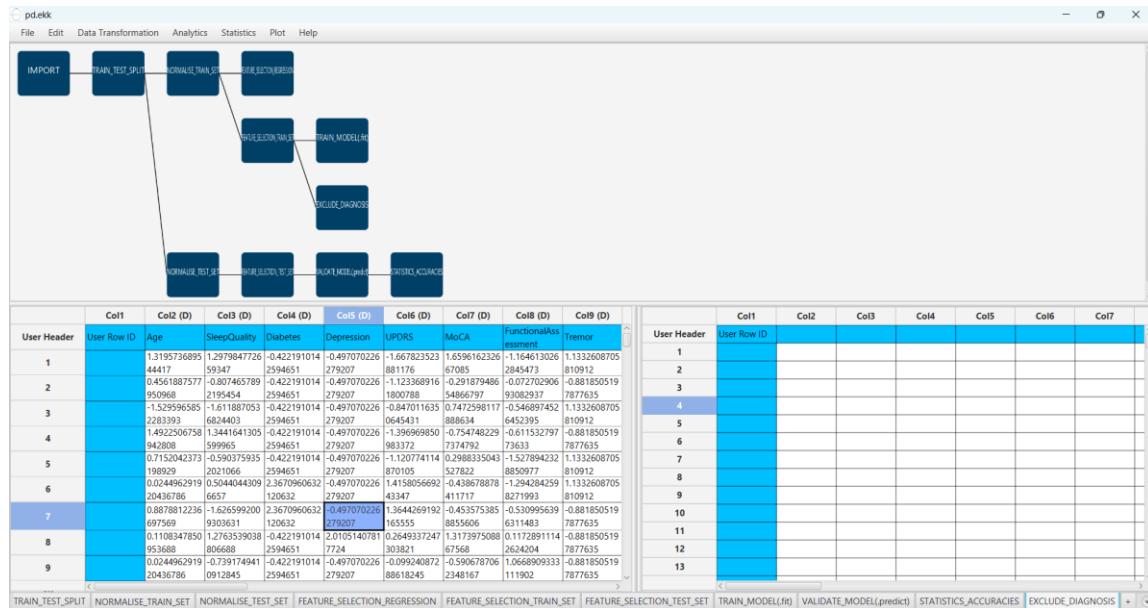
User Header	User Row ID	Col1 (S)	Col2 (D)	Col3 (S)	Col4 (S)	Col5	Col6	Col7	Col8	Col9	Col10
1						Predicted Class	Predicted Class				
2						1.0	0.0				
3	Actual Class	1.0	302	24							
4	Actual Class	0.0	15	185							
5											
6											
7	Classification Accuracy	0.9258555133	079848								
8	Precision	0.9526813880	0.8851674641	126183	148325						
9	Recall/Sensitivity	0.9263803680	981595	0.925							
10	Specificity	0.9263803680	981595								
11											
12											
13											
14											
15	F1 Score	0.9393468118	195957	0.9046454767	726161						
16											
17	F (beta=2)	0.9315237507	711289	0.9167492566	897919						
18	MCC	0.8445875112	35449								
19											

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

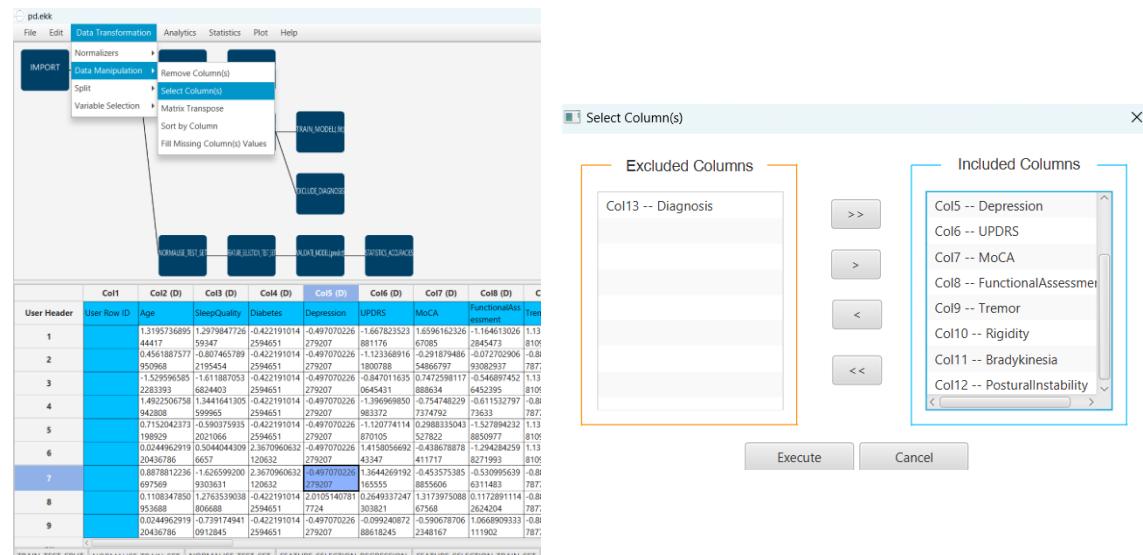
### Step 12.a: Create the domain

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

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



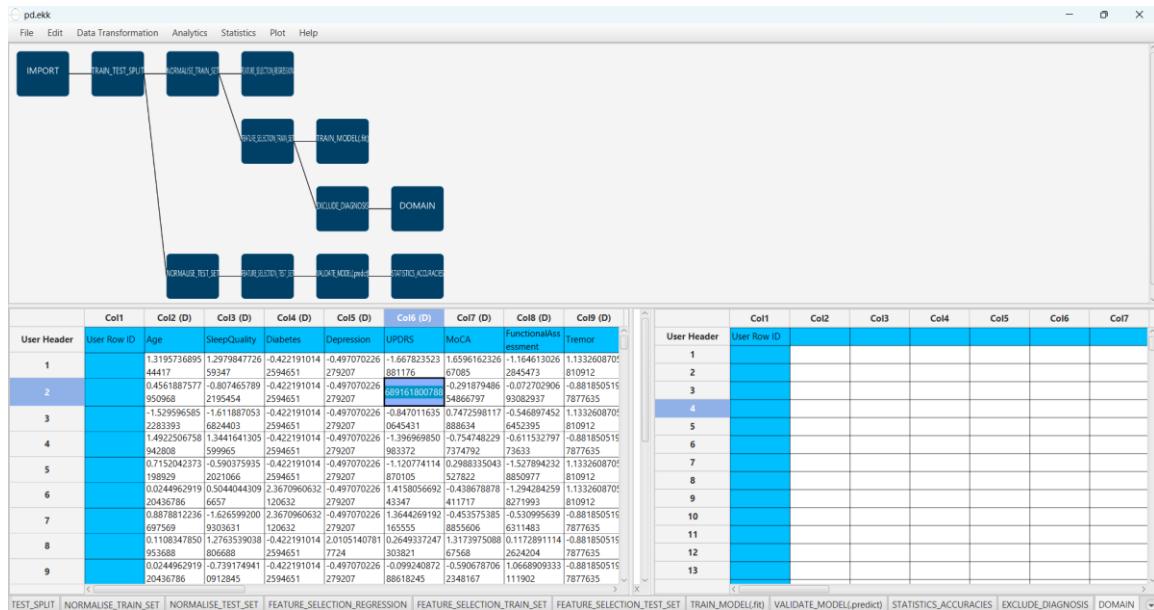
Manipulate the data to exclude the column that corresponds to the "Diagnosis" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except the "Diagnosis".



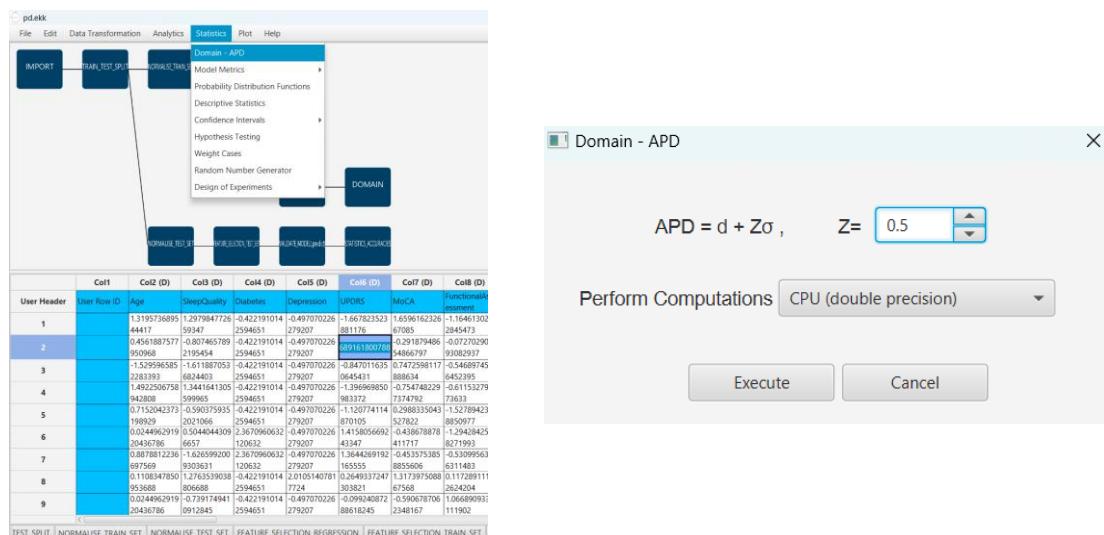
The results will appear on the output spreadsheet.

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

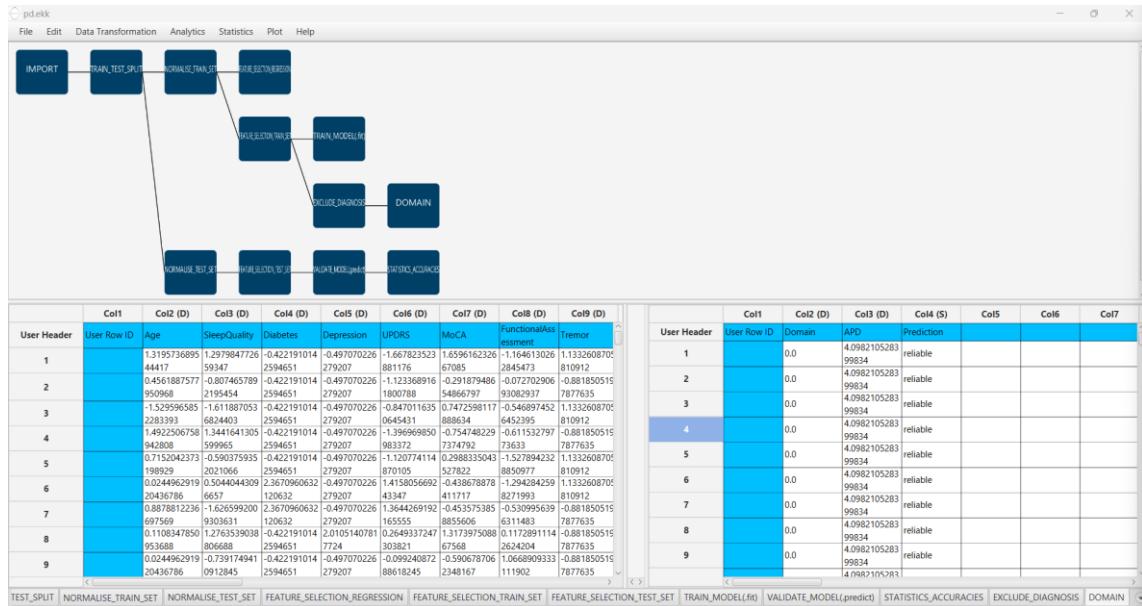
Import data into the input spreadsheet of the "DOMAIN" tab from the output of the "EXCLUDE\_DIAGNOSIS" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



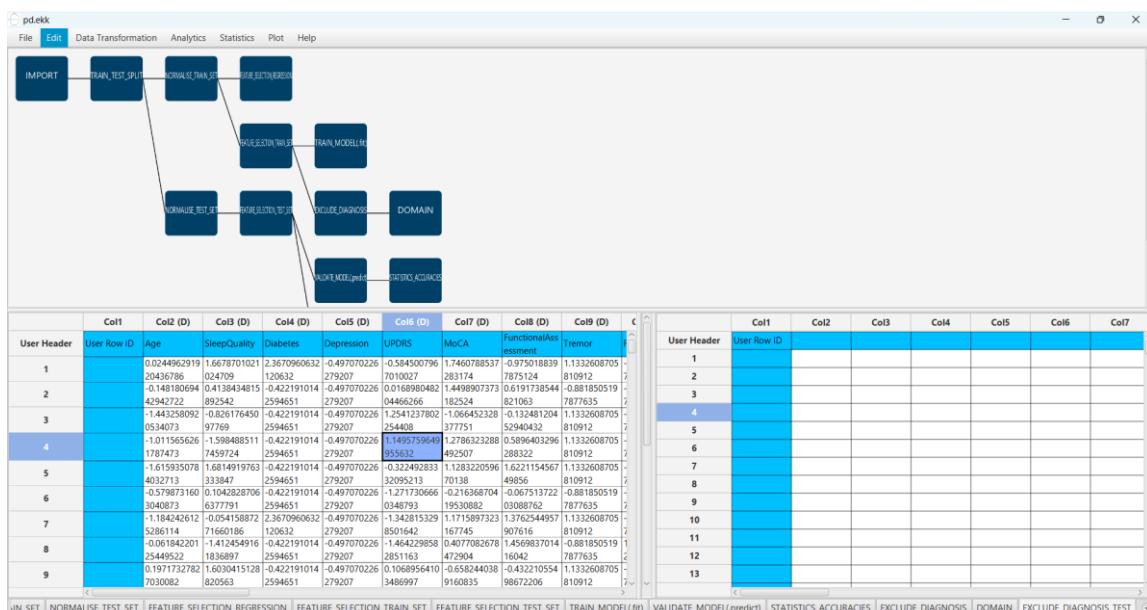
The results will appear on the output spreadsheet.



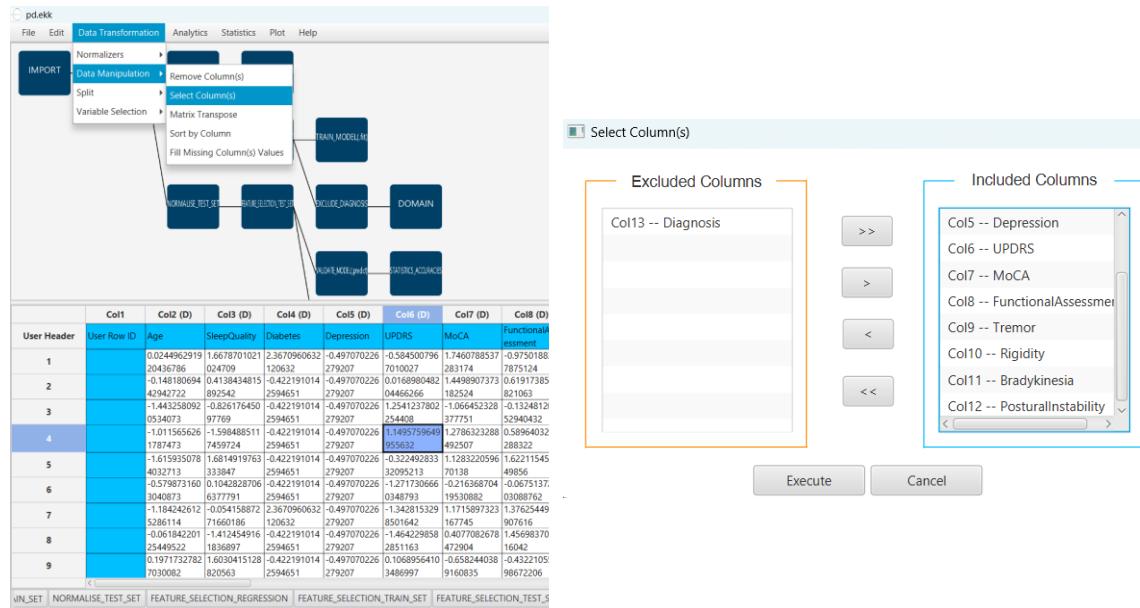
## Step 12.b: Check the test set reliability

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



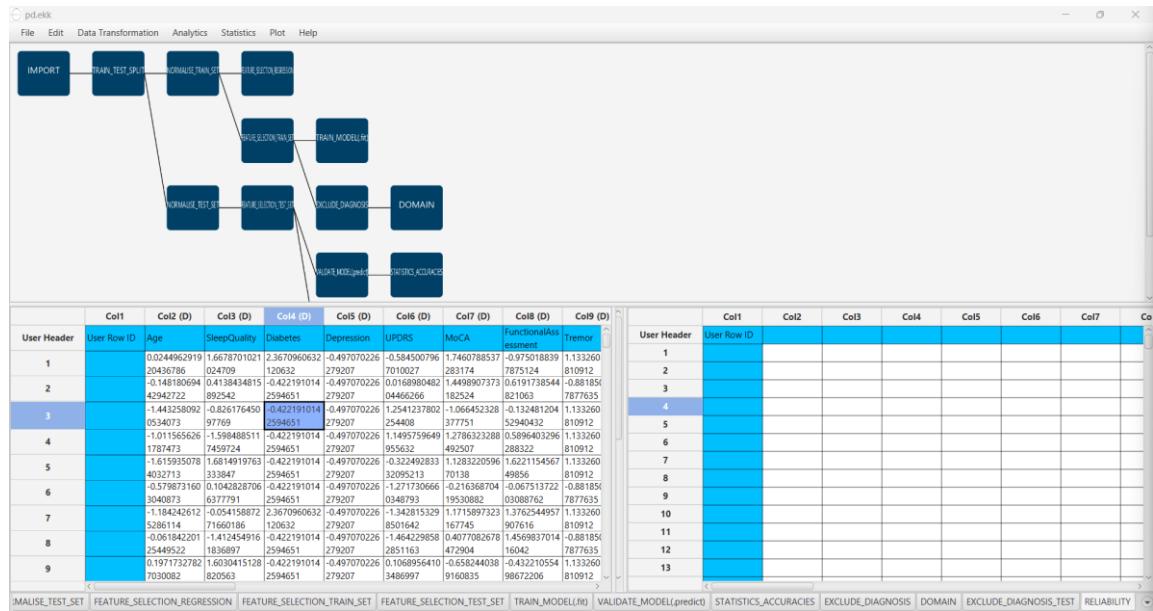
Filter the data to exclude the column that corresponds to the "Diagnosis" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except "Diagnosis".



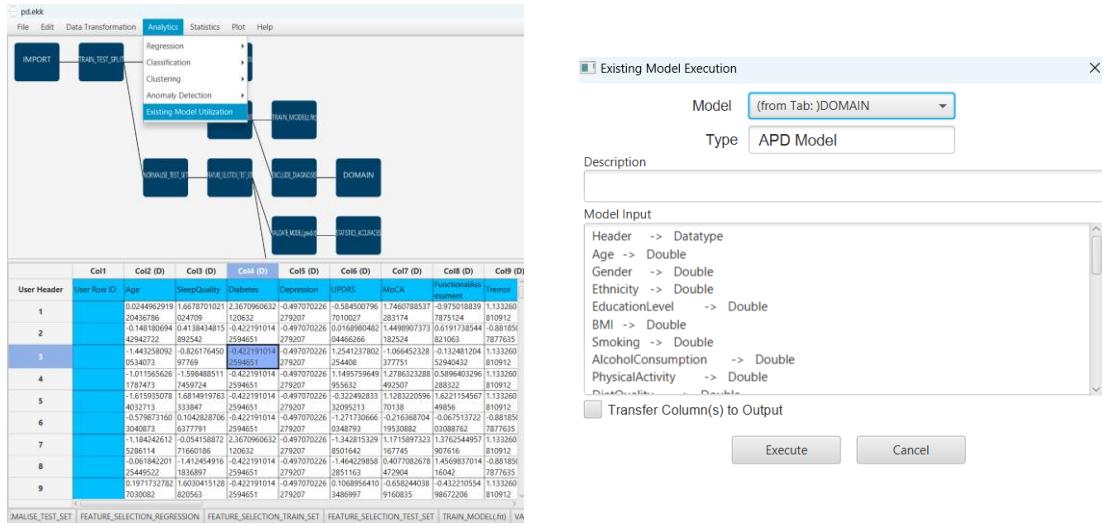
The results will appear on the output spreadsheet.

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

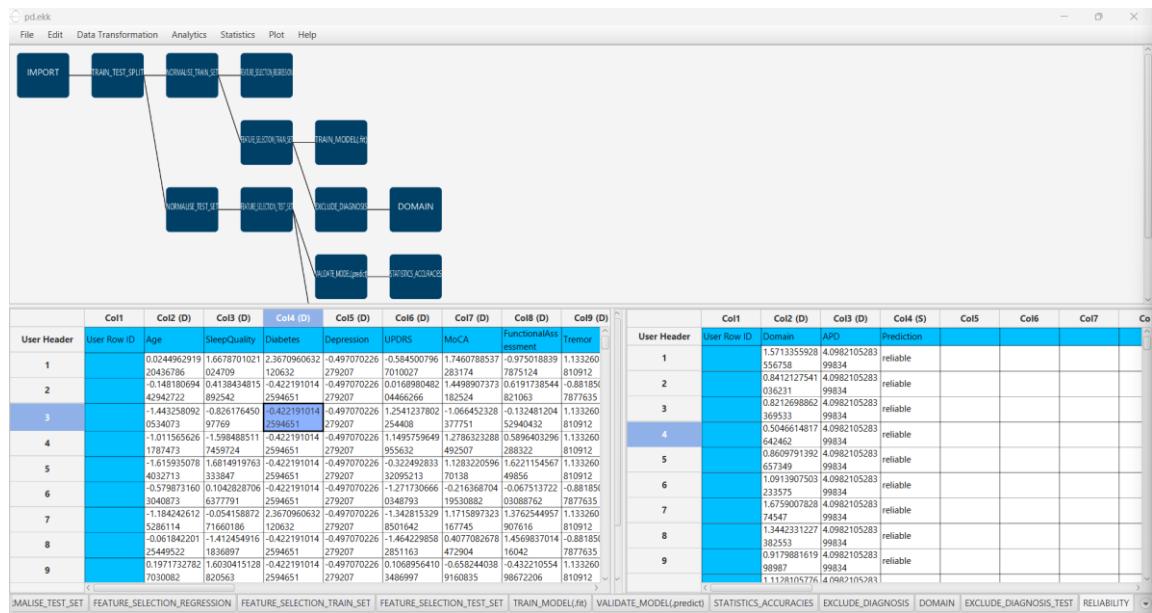
Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "EXCLUDE\_DIAGNOSIS\_TEST\_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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 are no unreliable samples in the test set.

## Final Isalos Workflow

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

