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:

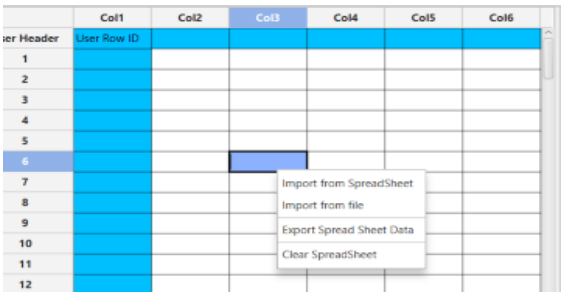
- Caucasian (0)
- African American (1)
- Asian (2)
- Other (3)

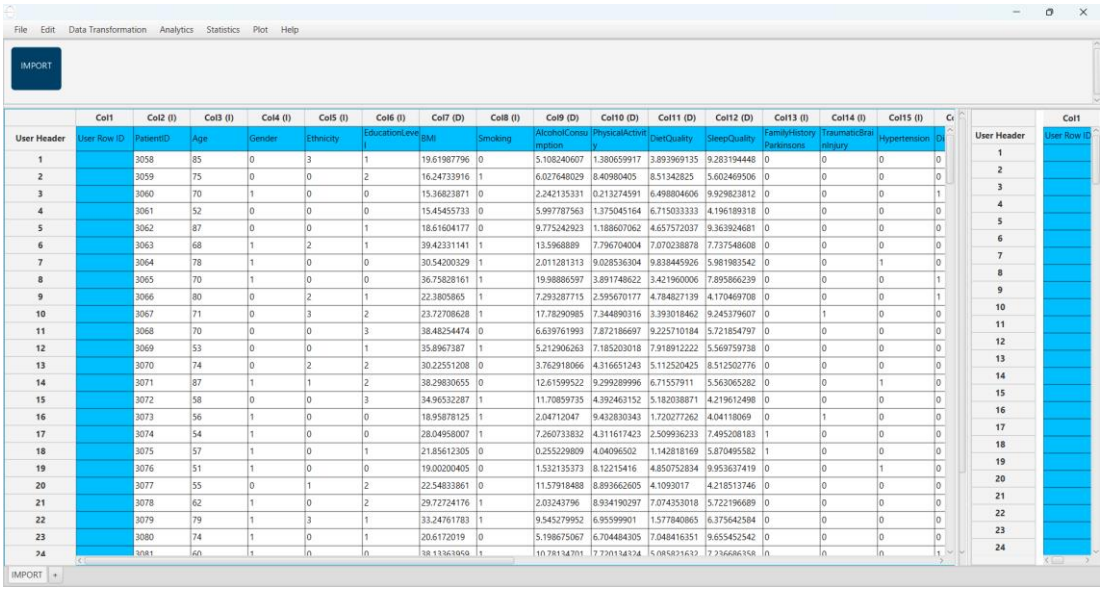
Education Level:

- 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.

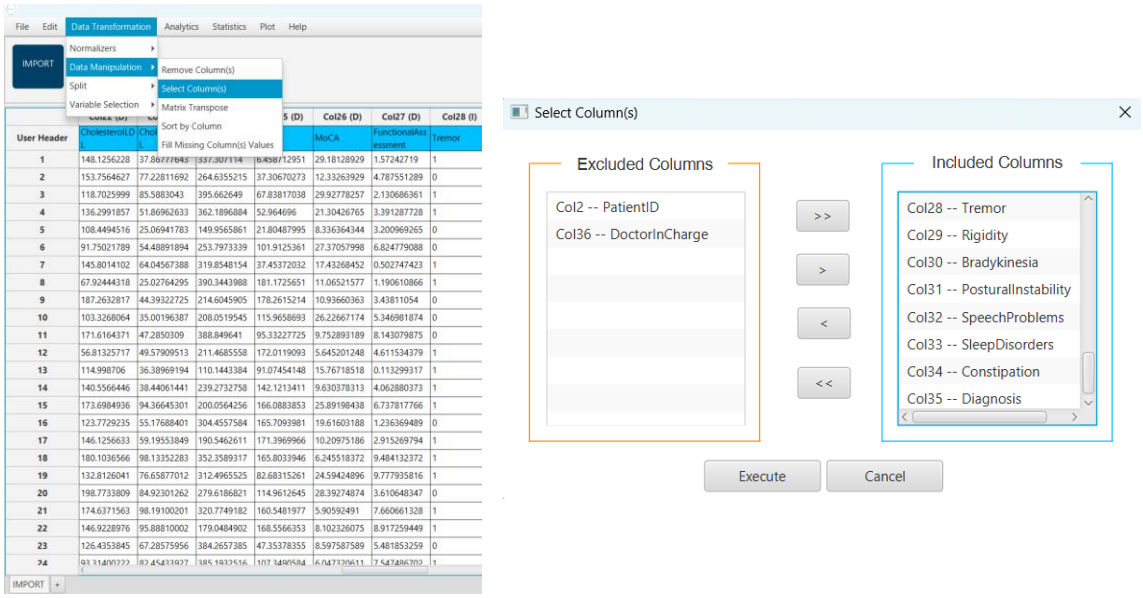




Step 2: Manipulate data

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.

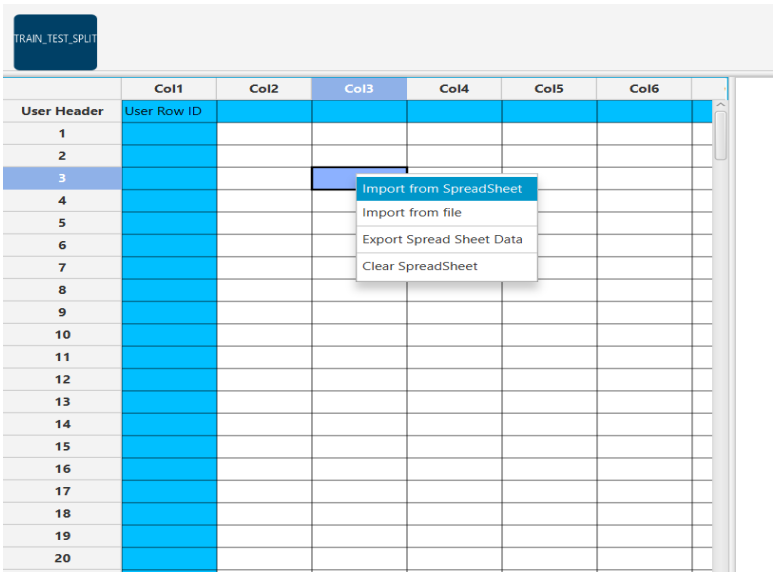


The data without the "PatientID" and "DoctorInCharge" columns will appear in the output spreadsheet.

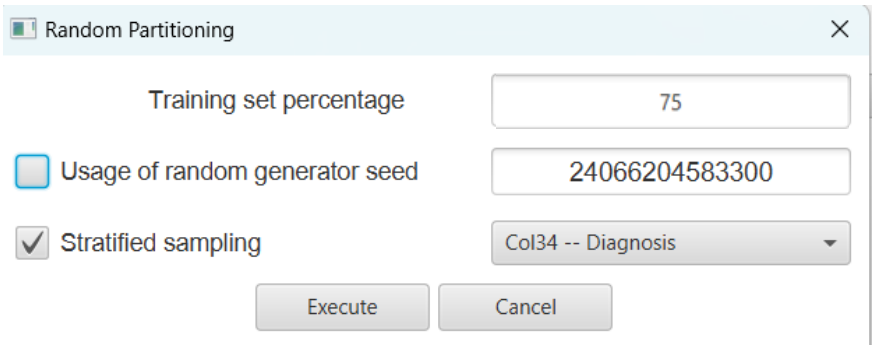
Step 3: Split data

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.

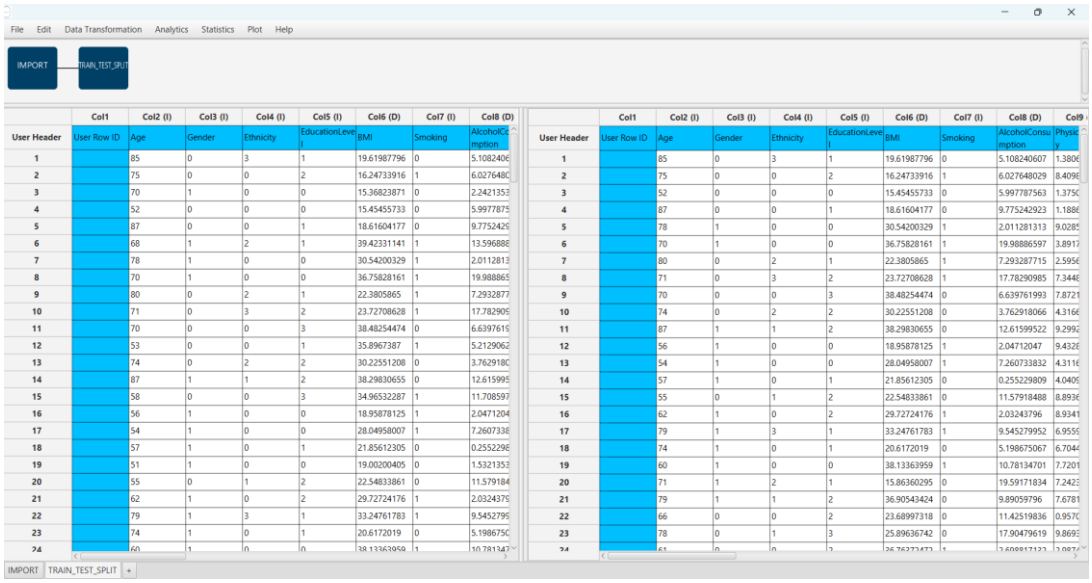
Import data into the input spreadsheet of the "TRAIN_TEST_SPLIT" tab from the output of the "IMPORT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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:



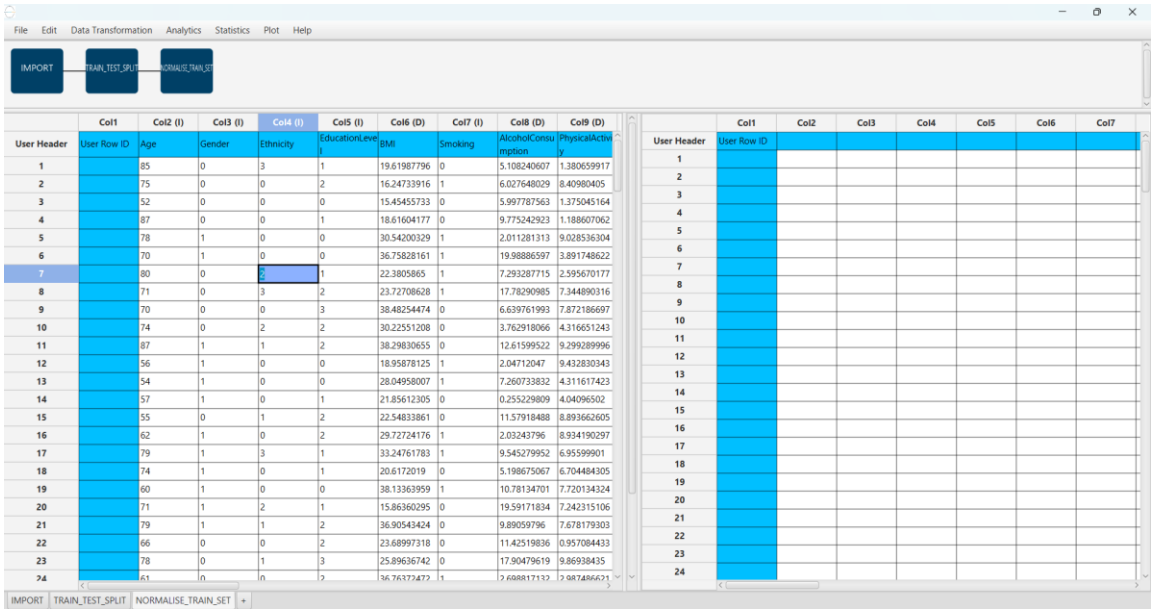
The results will appear on the output spreadsheet.



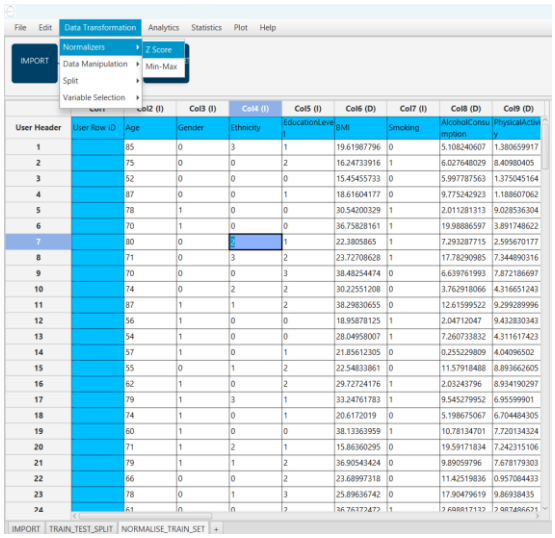
Step 4: Normalize the training set

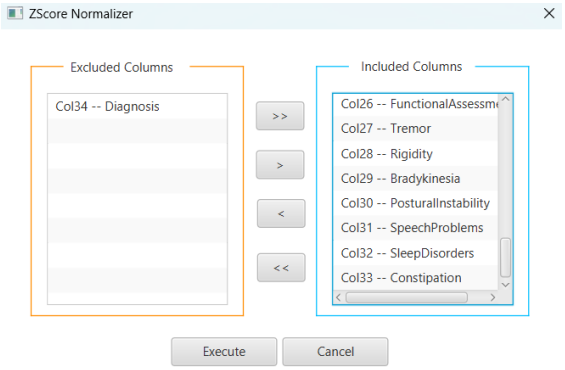
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALISE_TRAIN_SET".

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".

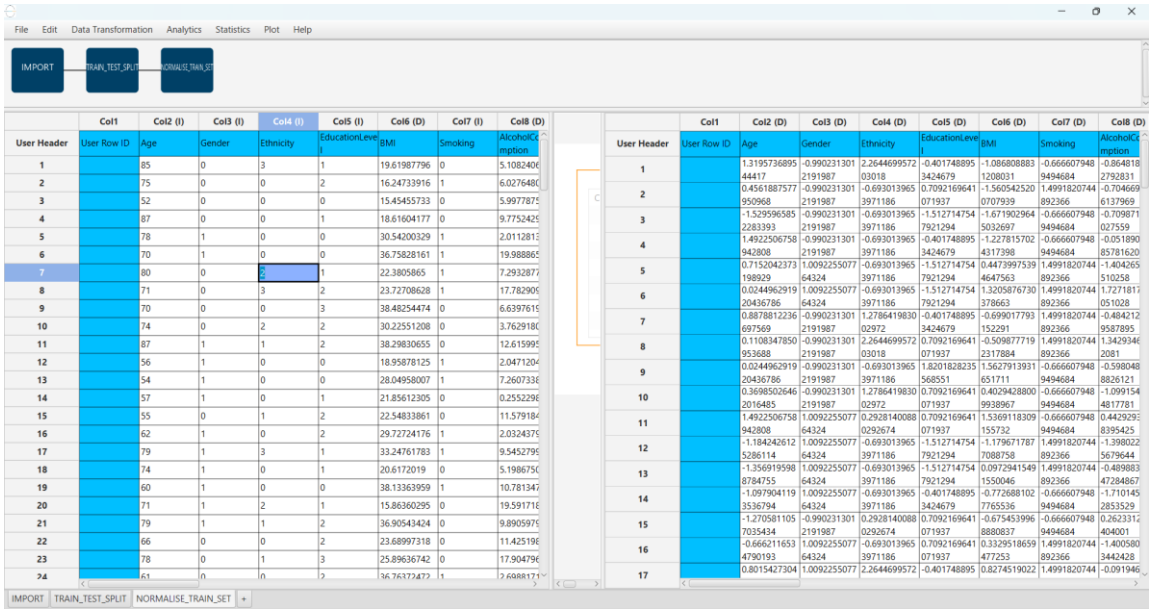


Normalize the data using Z-score by browsing: "Data Transformation" → "Normalizers" → "Z-Score". Then select all columns and click "Execute".





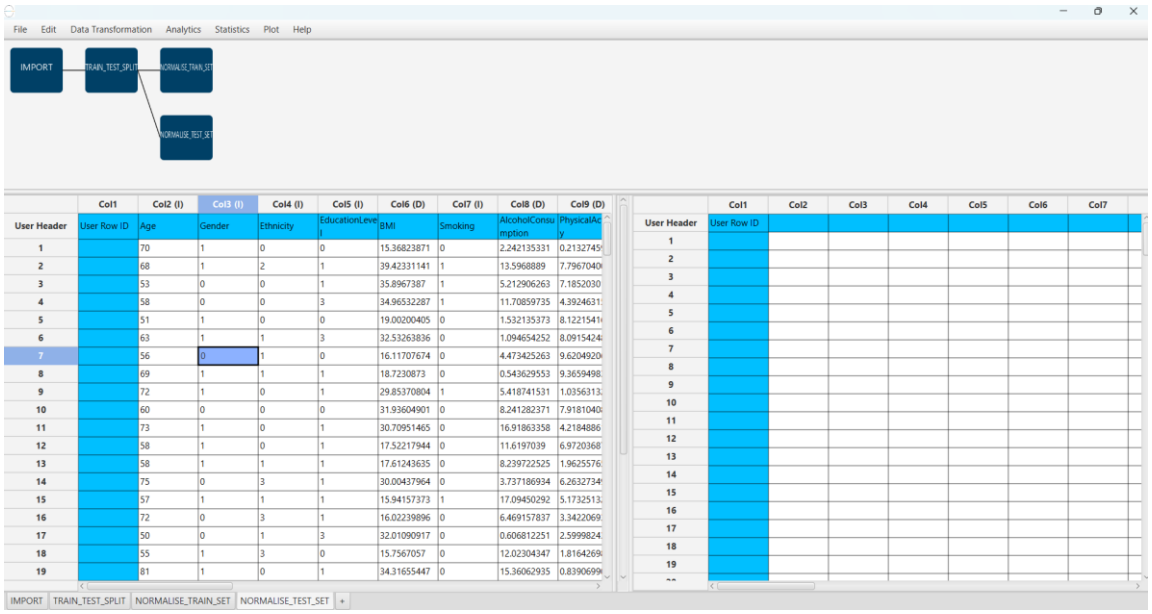
The results will appear on the output spreadsheet.



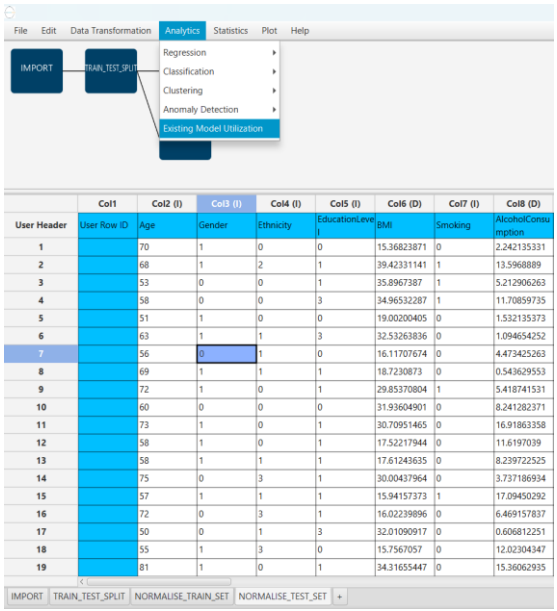
Step 5: Normalize the test set

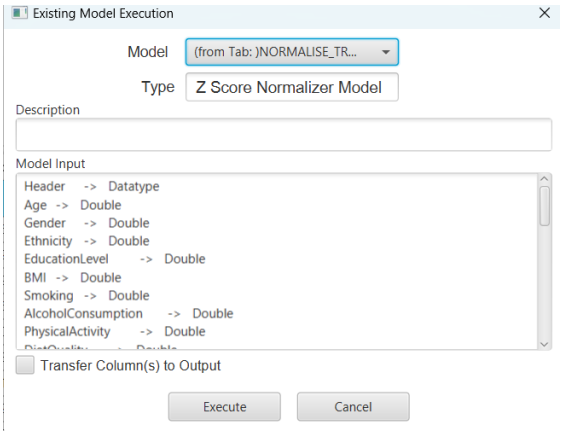
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALISE_TEST_SET".

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".

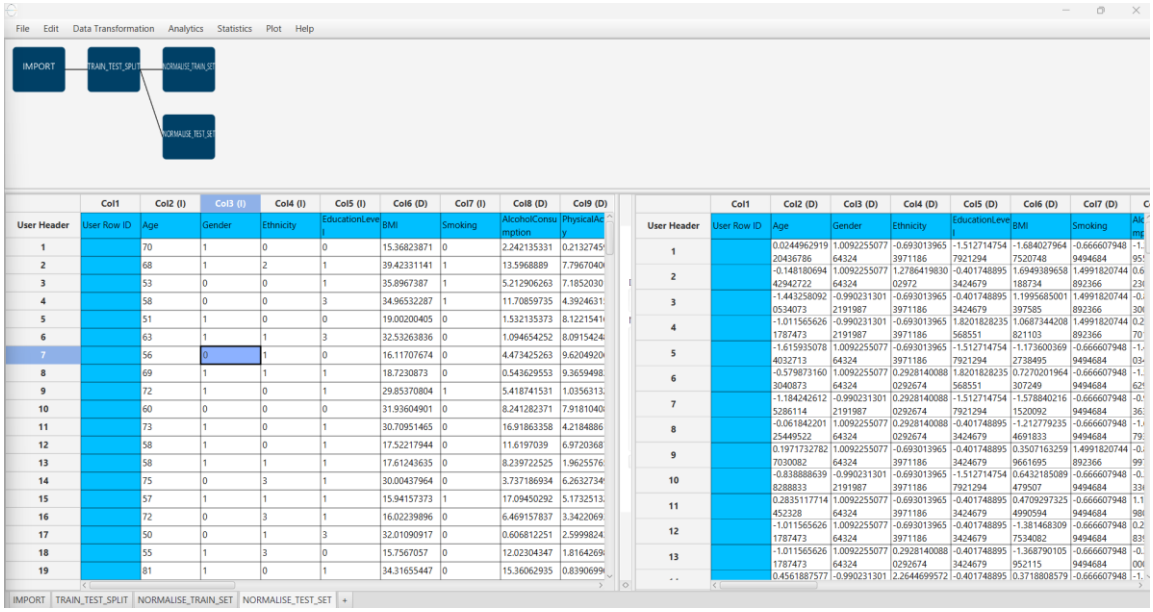


Normalize the test set using the existing normalizer of the training set by browsing: "Analytics" → "Existing Model Utilization" → "Model (from Tab:) NORMALISE_TRAIN_SET".





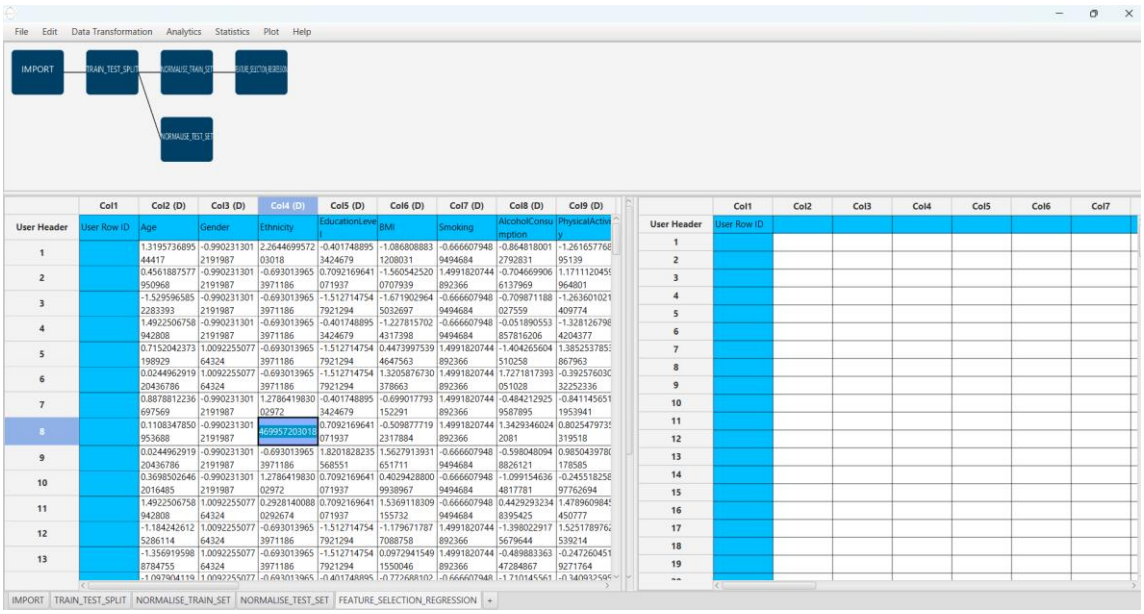
The results will appear on the output spreadsheet.



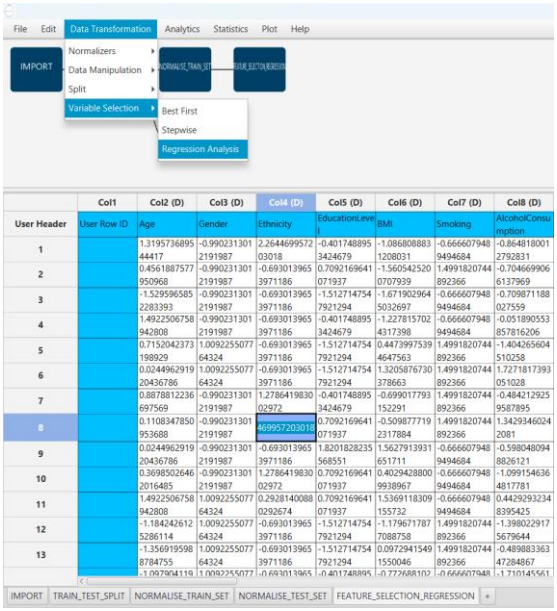
Step 6: Feature selection

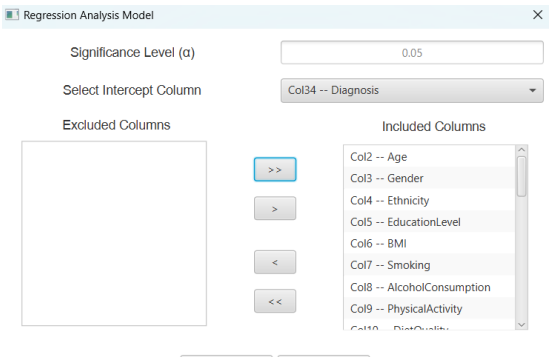
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_REGRESSION".

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 (α) 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	153.78711830863594	4.805847447144873	34.01009089526929	4.058724587937507E-154	
10		Total	1546	218.4598734583048	0.14130651582037826			
11			1578	372.2469917669407				
12			Coefficients	Standard Error	t Stat	P-value	Lower 95.0%	Upper 95.0%
13		Diagnosis	0.6193793540215377	0.009459975879892092	65.47367159129337	0.0	0.6008236148982906	0.6379350931447847
14		Age	0.034944578344872954	0.009556538961688421	3.656614438026531	2.6412976725683964E-4	0.01619943077307164	0.05368972591667427
15		Gender	0.005905157055109047	0.009589163032430095	0.6158156906017851	0.5381068188904652	-0.012903982619157102	0.024714296729375196
16		Ethnicity	-0.0051288114570836344	0.009590432485275994	-0.534784167 9671693	0.5928760144825591	-0.023940441 1626286	0.013682818 24846133
17		EducationLevel	-0.005631677102695146	0.009543413279615323	-0.590111413 7773302	0.5552021963015052	-0.024351078654056724	0.013087724448666432
18		BMI	0.006842640052181735	0.009609441497752863	0.7120746875645025	0.4765259676099952	-0.012006275824172727	0.025691555928536197
19		Smoking	-0.01068028611694388	0.009553808376531577	-1.117908764339406	0.2637796445924609	-0.029420077 64698911	0.008059505413101352
20		AlcoholConsumption	0.002461167657840956	0.009563682979952696	0.2573451737 1602894	0.7969465694 74437	-0.016297992 903093382	0.021220328218775293
21		PhysicalActivity	-0.009334361700271854	0.009568907355141496	-0.975488773 569992	0.3294704756939527	-0.028103769 871170126	0.009435046470626419
22		DietQuality	-0.003050220046770803	0.009546825564645988	-0.319500971 93211996	0.7493898214594091	-0.021776314 79393405	0.015675874700392445
23		SleepQuality	-0.021731024730421666	0.009574762237808862	-2.269615076 6658314	0.0233681600 34877315	-0.040511917 25146546	-0.002950132 2093778704
24		FamilyHistoryParkinsons	0.0101209008340667	0.009552470104518592	1.059506151113651	0.2895348303630602	-0.008616265 675929415	0.028858067344062814
25		TraumaticBrainInjury	0.014589229732617963	0.009595804986224412	1.520375805215 0949	0.1286210332 7246324	-0.004232938 131512603	0.033411397596 74853
26		Hypertension	0.006973590608428447	0.009562082829304448	0.7292961934 0431	0.4659309710 9648995	-0.011782431 257612396	0.025729612474469293
27		Diabetes	0.02562743849956757	0.009587120343858281	2.6731111721 138525	0.0075941977 85412139	0.0068223055 58165209	0.044432571440 96993
28		Depression	0.02723933473557805	0.009560502652271166	2.8491529918 7822	0.0044416491 54449658	0.0084864123 86192469	0.045992257084 96363
29		Stroke	0.012415701060811338	0.009548812483364104	1.3002350902 211048	0.1937142092 2437116	-0.006314291 026668686	0.031145693148 29136
30		SystolicBP	0.00528704609413583	0.009527222988344163	0.5549409416 914821	0.5790153341 233407	-0.013400598 197382648	0.023974690385 65431
31		DiastolicBP	-0.0035214150697520 83	0.009577388741574176	-0.367680081 1546979	0.7131621231 346639	-0.022307459 476941696	0.015264629337437531
32		CholesterolTotal	0.007258067669105708	0.009551318959344359	0.7599021349 826152	0.4474289458 867108	-0.011476840 870064977	0.025992976208276394
33		CholesterolLDL	0.0015268790166764238	0.009552440131189343	0.1598417781 9561138	0.8730265930 744	-0.017210228 700645668	0.020263986733998512
34		CholesterolHDL	5.411733944502038E-4	0.009568759901534996	0.0565562727 0607868	0.9549059762 495937	-0.018227945 54625449	0.019310292335154897
35		CholesterolTriglycerides	0.002268052516280527	0.009563016046575041	0.2371691634 9762077	0.8125569806 105581	-0.016489799 855084888	0.021025904887645944
36		UPDRS	0.18630288121195493	0.009603880188781513	19.398709433 045518	3.3396373538 02613E-75	0.1674648738 4105188	0.205140888 58285798
37		MoCA	-0.07109273757277541	0.009556317899932587	-7.439344140 884063	1.6703413366 346027E-13	-0.089837451 53083589	-0.052348023 61471493
38		FunctionalAssessment	-0.11419317166623748	0.009565870057560244	-11.93756249 866541	1.7403934101 474355E-31	-0.132956622 17907478	-0.095429721 15340018
39		Tremor	0.12423615764298186	0.009551520061017858	13.006951443 259872	8.8124775789 78414E-37	0.1055008546 4295439	0.142971460643300933
40		Rigidity	0.10336739579051336	0.009535987077577548	10.839716428 891393	1.9389201436 52925E-26	0.0846625607 3165018	0.122072230849 37655
41		Bradykinesia	0.09738441151376426	0.009550763581646275	10.196505303 608197	1.1315824768 598043E-23	0.0786505923 4773905	0.116118230679 78946
42		PosturalInstability	0.08795422774590572	0.009534309865763344	9.225022994243103	9.0117242673 54057E-20	0.0692526825 3738402	0.106655772954 42742
43		SpeechProblems	-0.0053249866410449145	0.009576010343971408	-0.556075698 5185661	0.5782395755 011953	-0.024108327 321856172	0.013458354039766342
44		SleepDisorders	-0.01601764859514722	0.009559114487731708	-1.675641464 0398416	0.0940106527 4412763	-0.034767848 06031181	0.002732550870017371

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

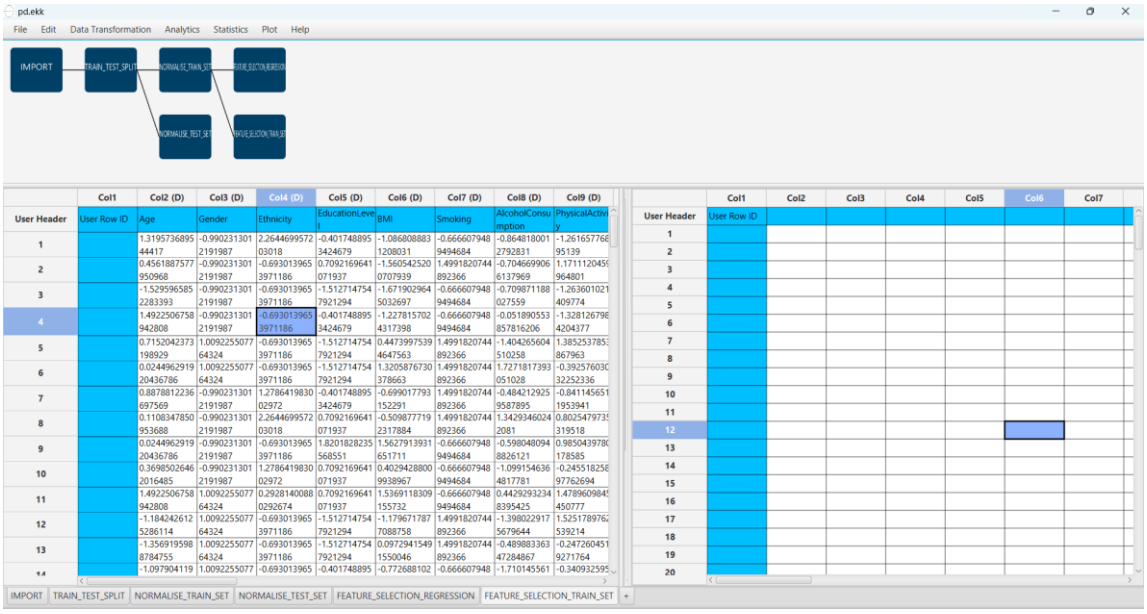
- Diagnosis (p-value = 0.0)

- 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)

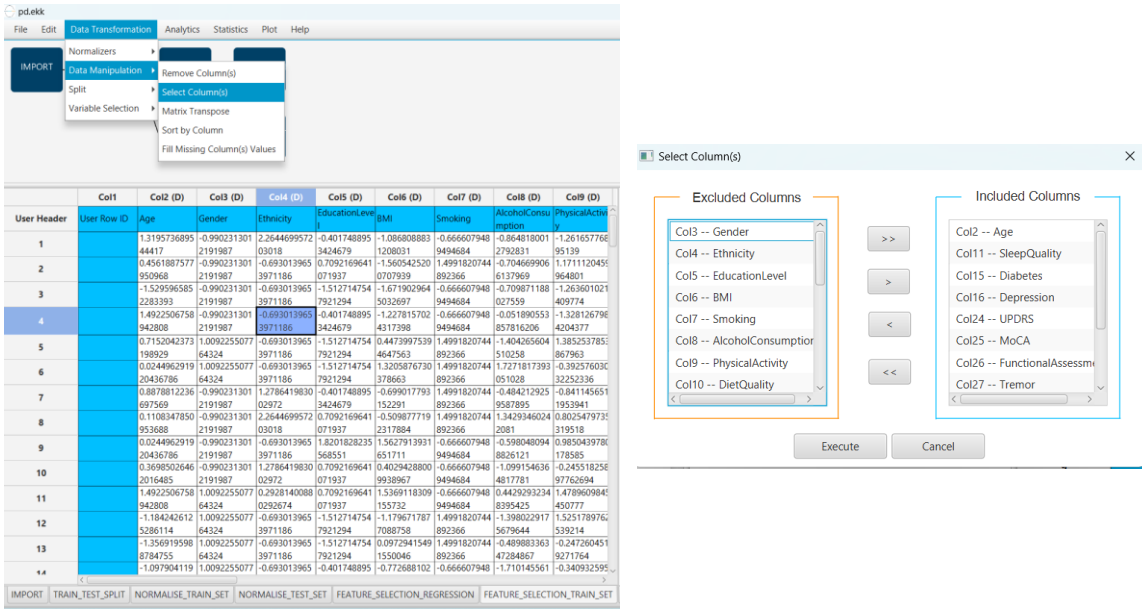
Step 7: Feature selection: train set

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

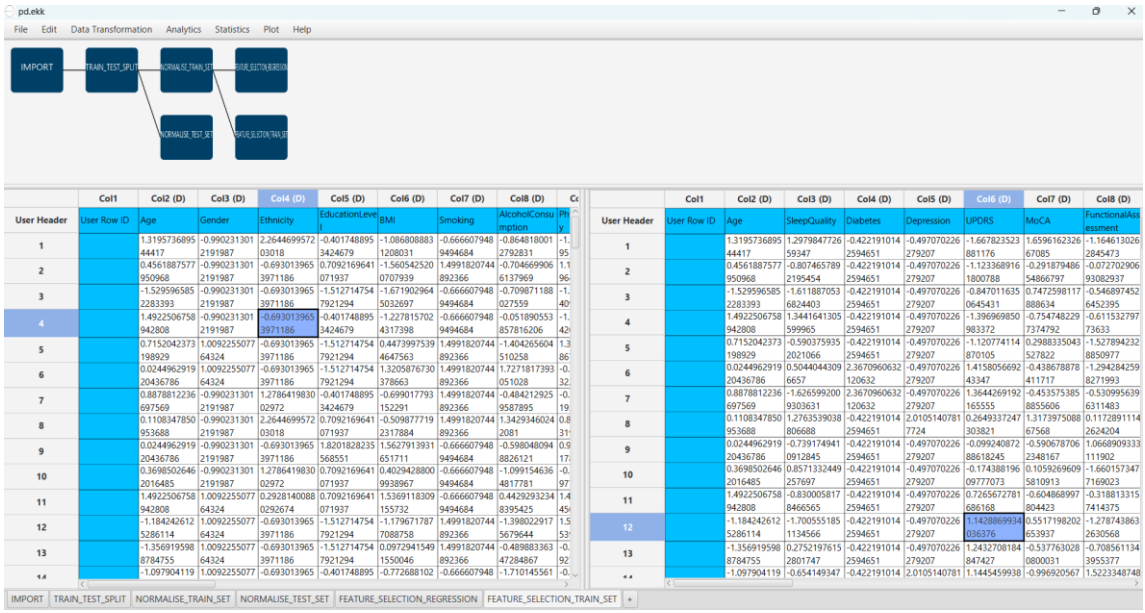
Import data into the input spreadsheet of the "FEATURE_SELECTION_TRAIN_SET" tab from the output of the "NORMALISE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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



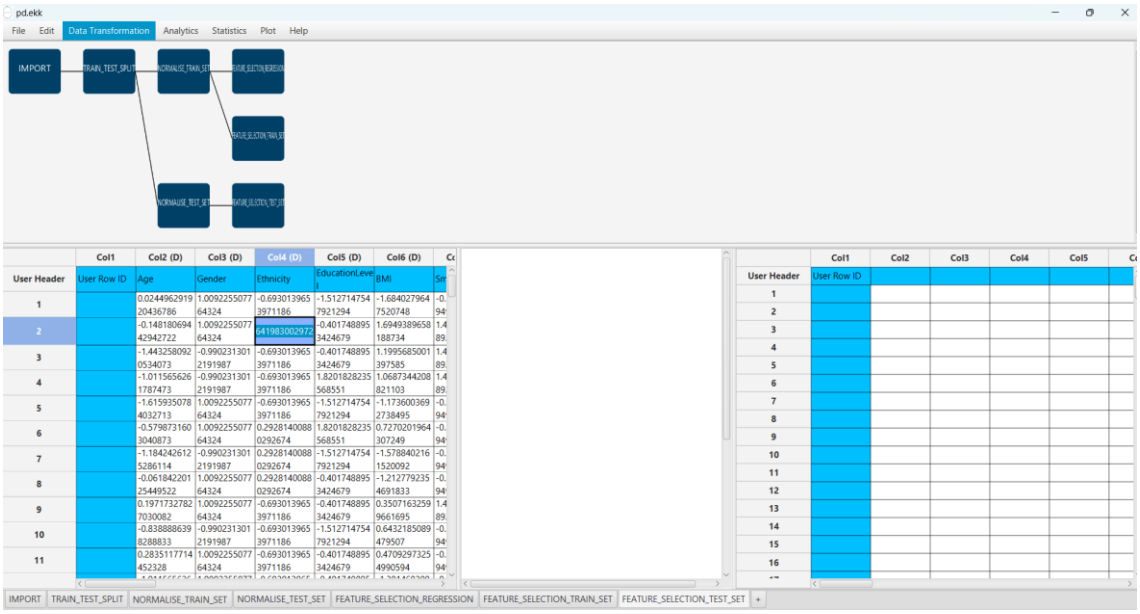
The results will appear on the output spreadsheet.



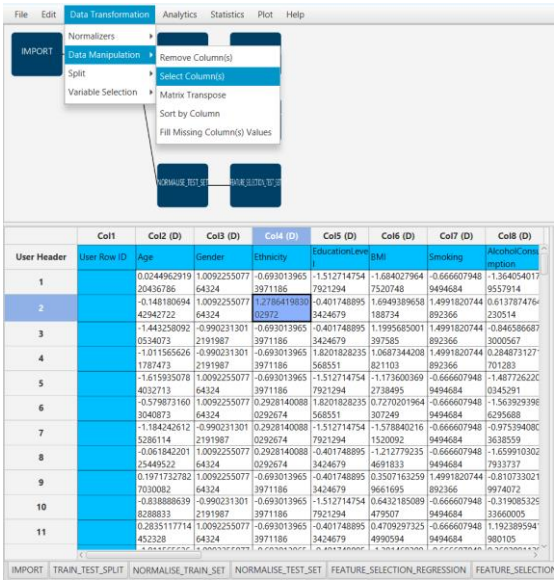
Step 8: Feature selection: test set

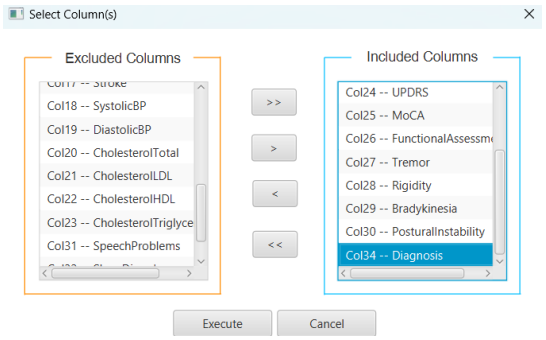
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_TEST_SET".

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".

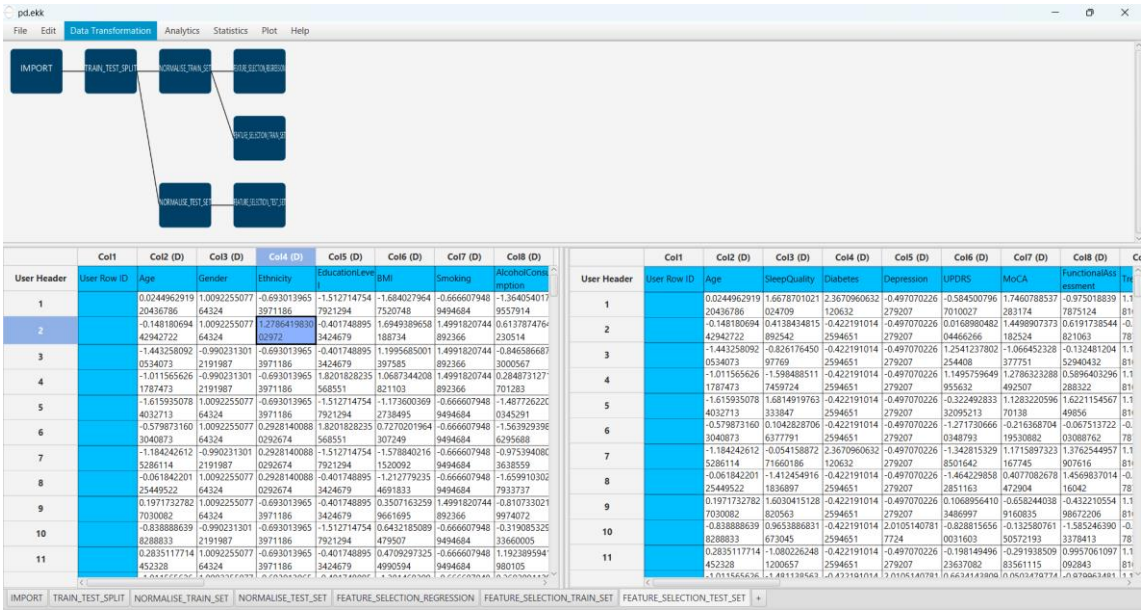


Manipulate the data by choosing the columns that correspond to the significant features (from step 7): "Data Transformation" → "Data Manipulation" → "Select Column(s)".





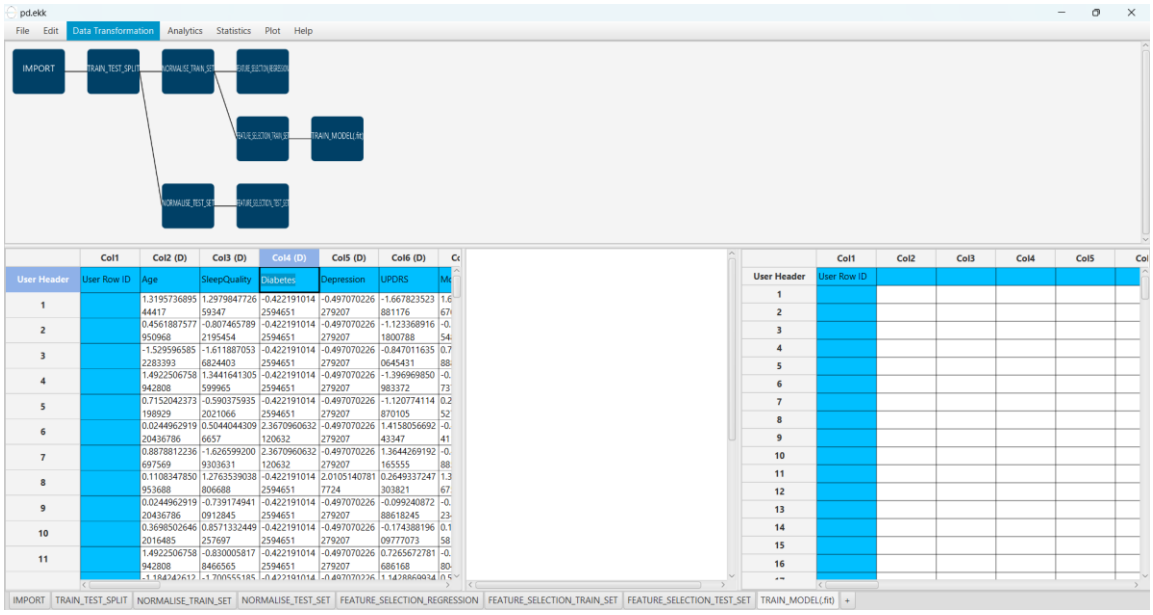
The results will appear on the output spreadsheet.



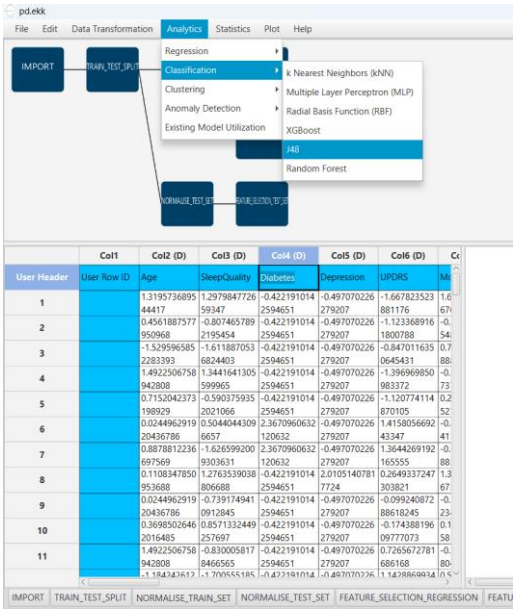
Step 9: Train the model

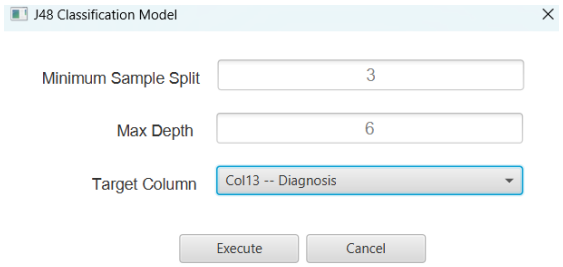
Create a new tab by pressing the "+" button on the bottom of the page with the name "TRAIN_MODEL(.fit)".

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".

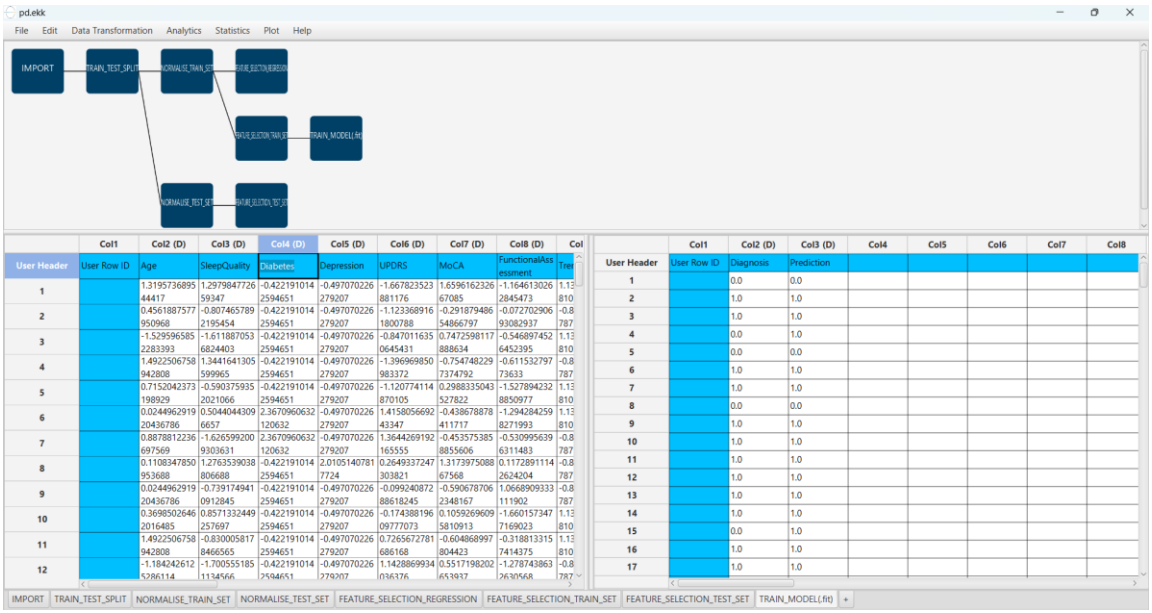


Use the J48 Method to train and fit the model by browsing: "Analytics" → "Classification" → "J48" and set the "Minimum Sample Split" as 3, the "Max Depth" as 6 and the "Target Column" as the column corresponding to "Diagnosis".





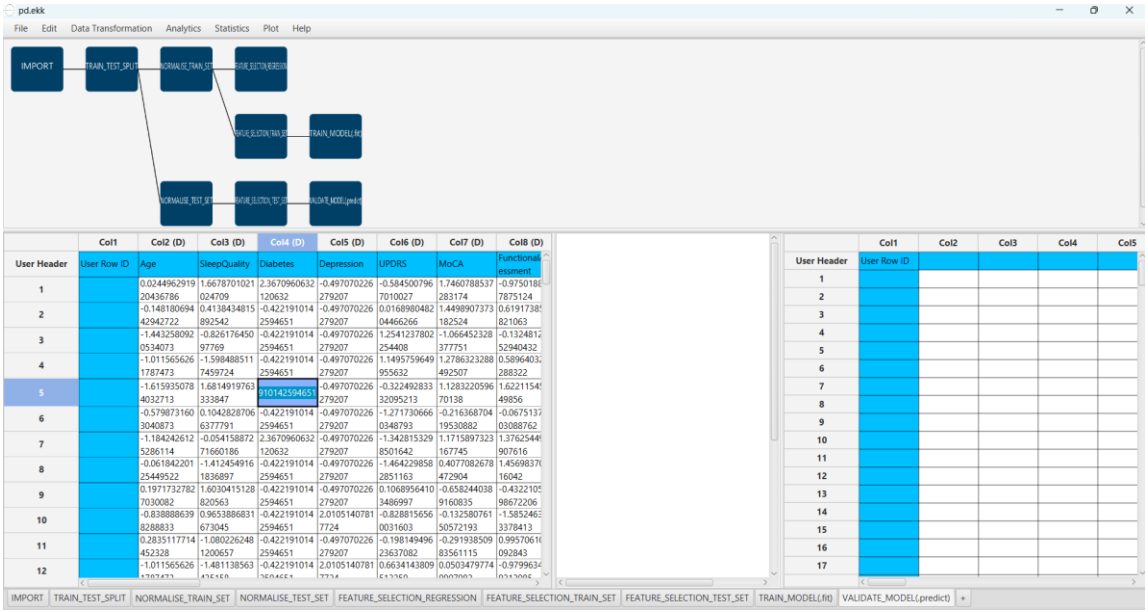
The predictions will appear on the output spreadsheet.



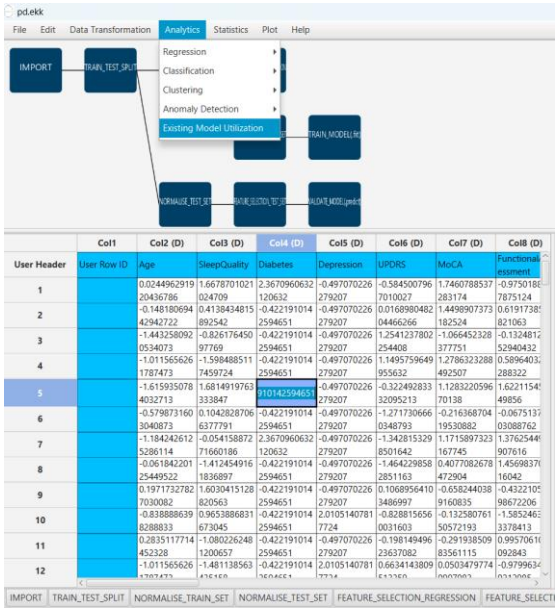
Step 10: Validate the model

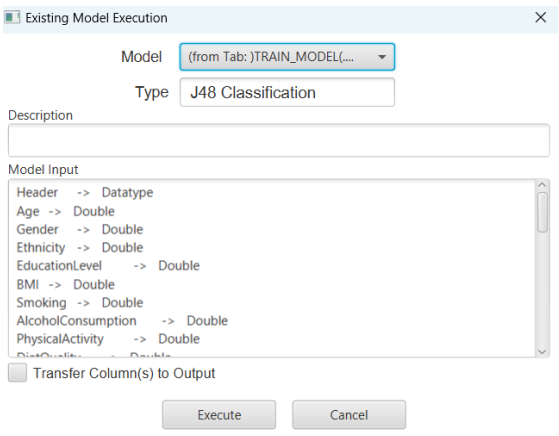
Create a new tab by pressing the "+" button on the bottom of the page with the name "VALIDATE_MODEL(.predict)".

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".

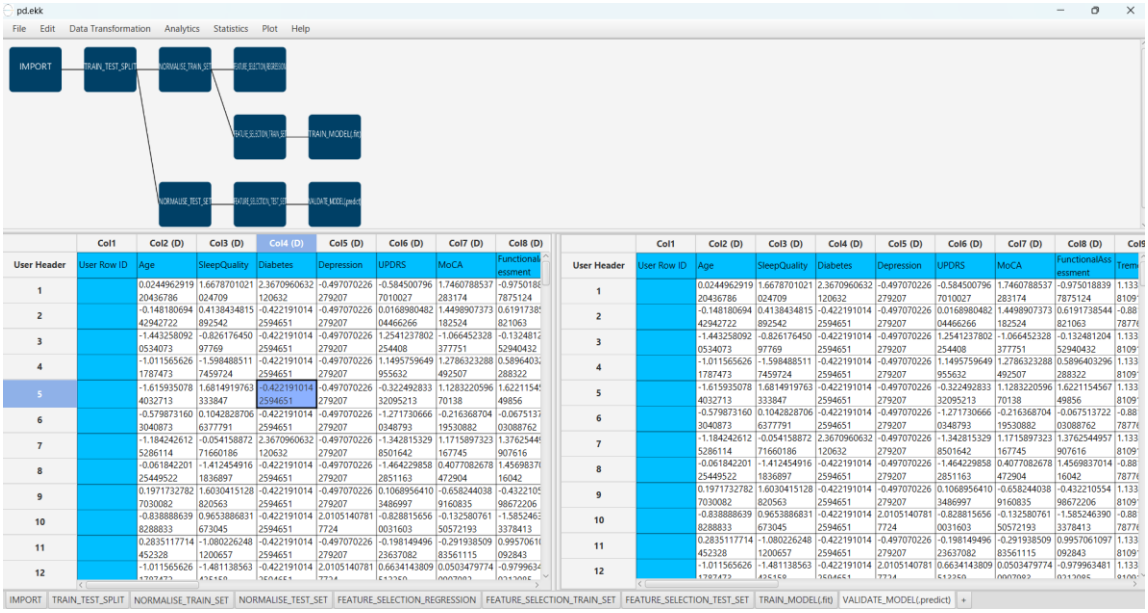


To validate the model browse: "Analytics" → "Existing Model Utilization". Then choose Model "(from Tab:) TRAIN_MODEL (.fit)".





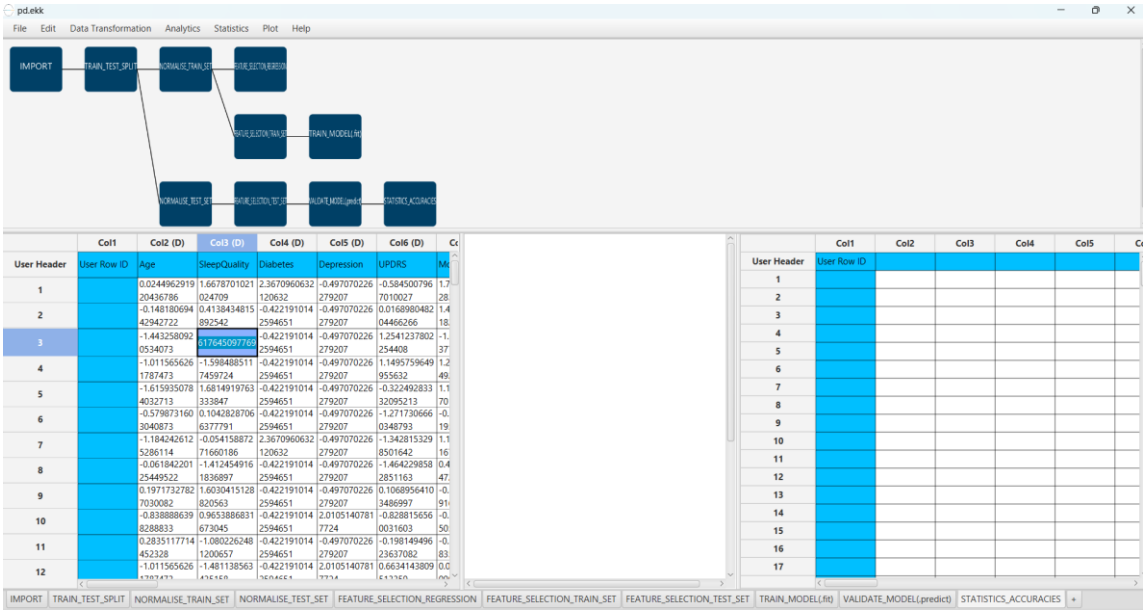
The predictions will appear on the output spreadsheet.



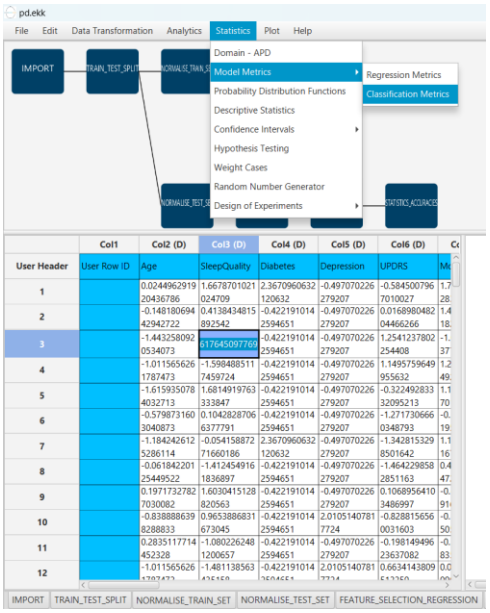
Step 11: Statistics calculation

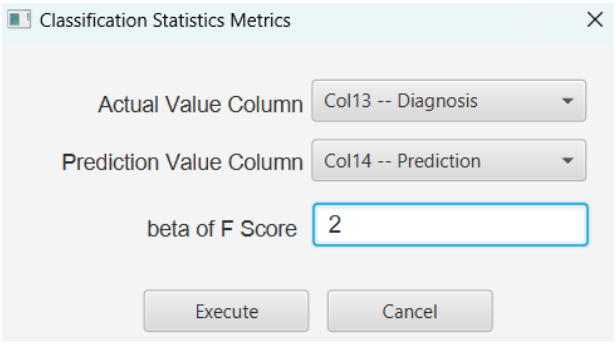
Create a new tab by pressing the "+" button on the bottom of the page with the name "STATISTICS_ACCURACIES".

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".

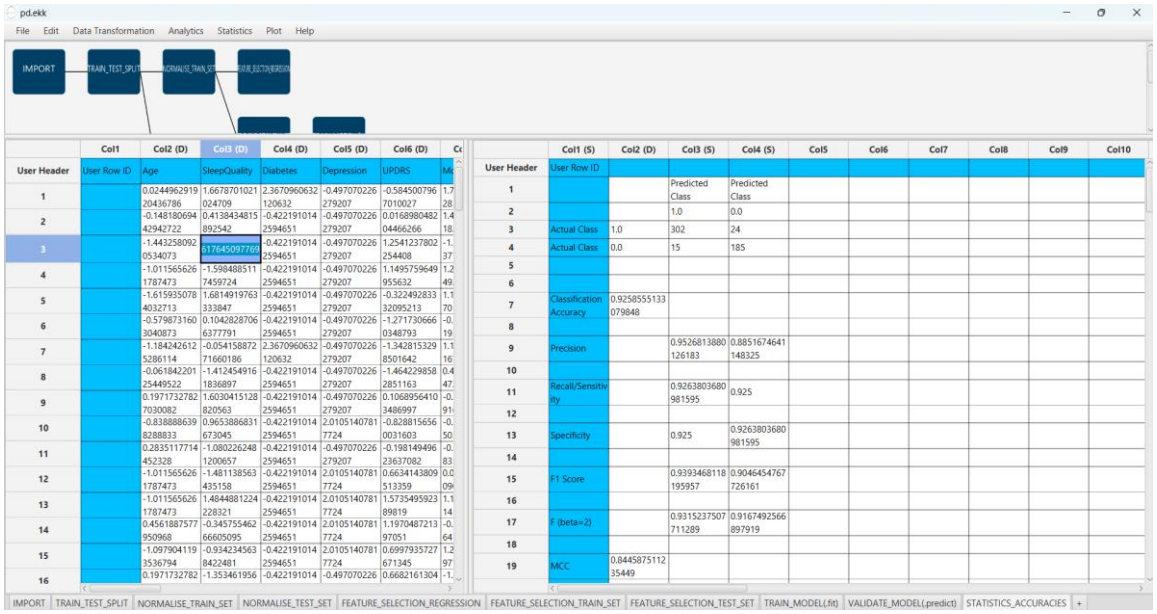




The results will appear on the output spreadsheet.

Accuracy: 0.93

F1-Score = 0.9199

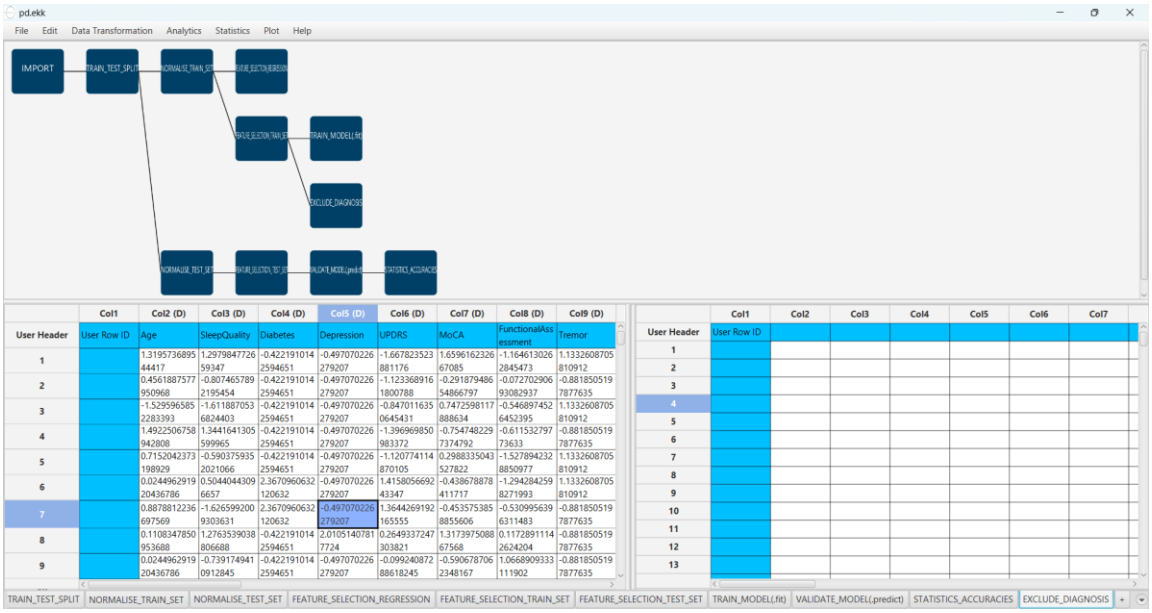


Step 12: Reliability check of each record of the test set

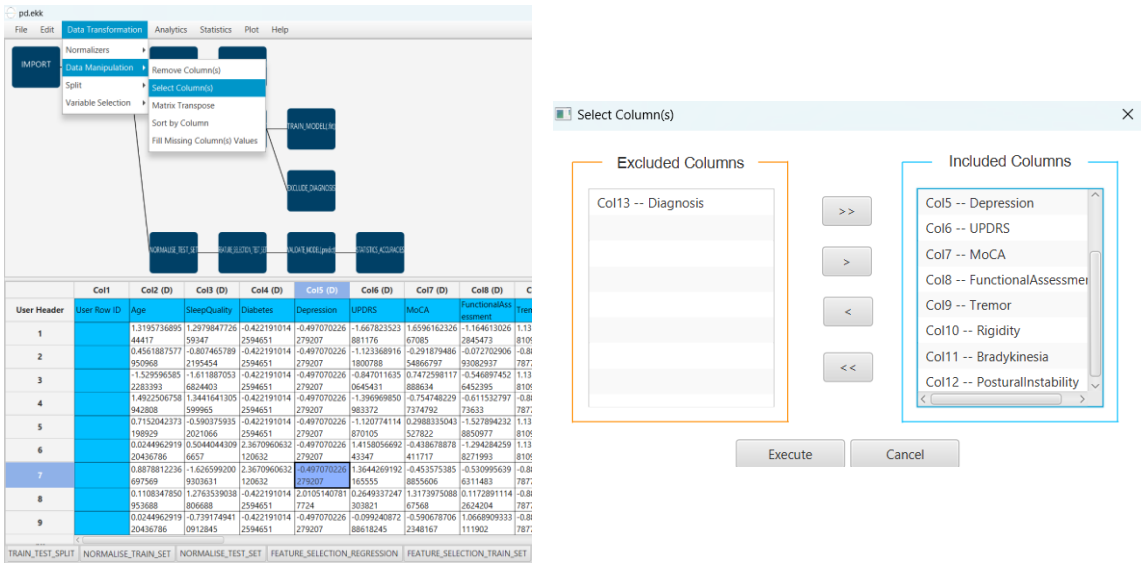
Step 12.a: Create the domain

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".



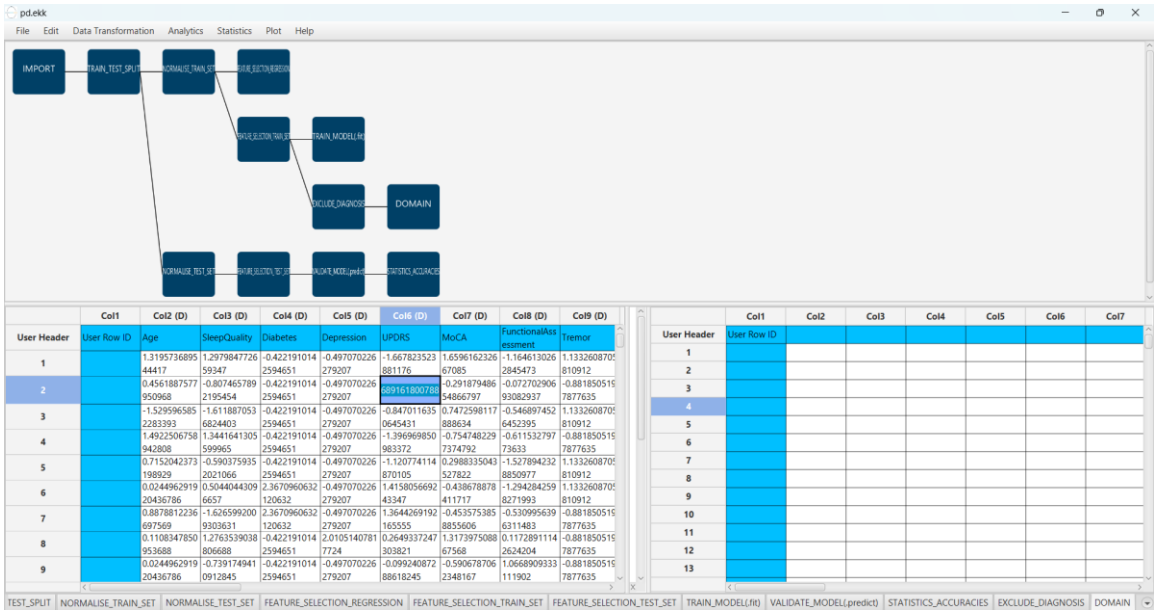
Manipulate the data to exclude the column that corresponds to the "Diagnosis" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except the "Diagnosis".



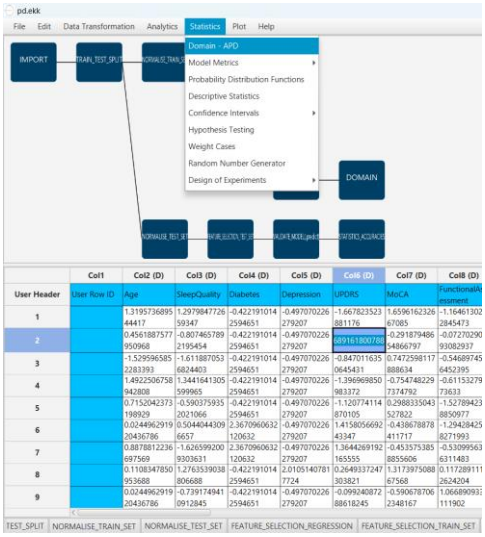
The results will appear on the output spreadsheet.

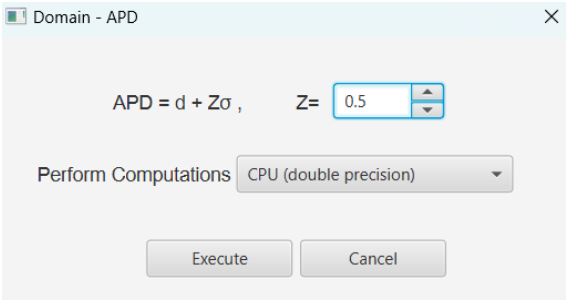
Create a new tab by pressing the "+" button on the bottom of the page with the name "DOMAIN".

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".

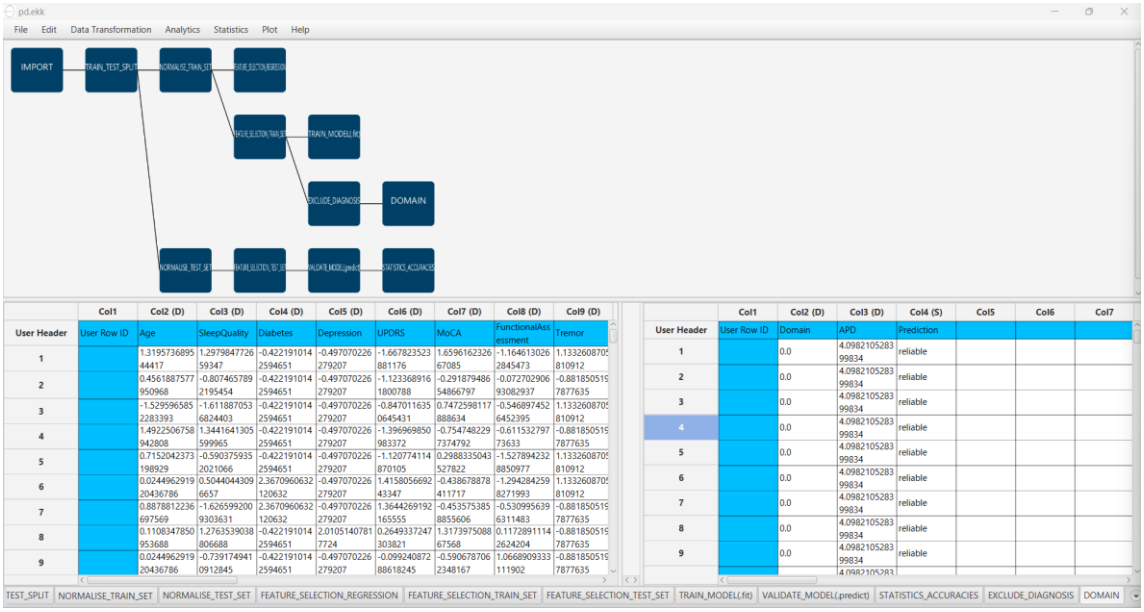


Create the domain by browsing: "Statistics" → "Domain APD".





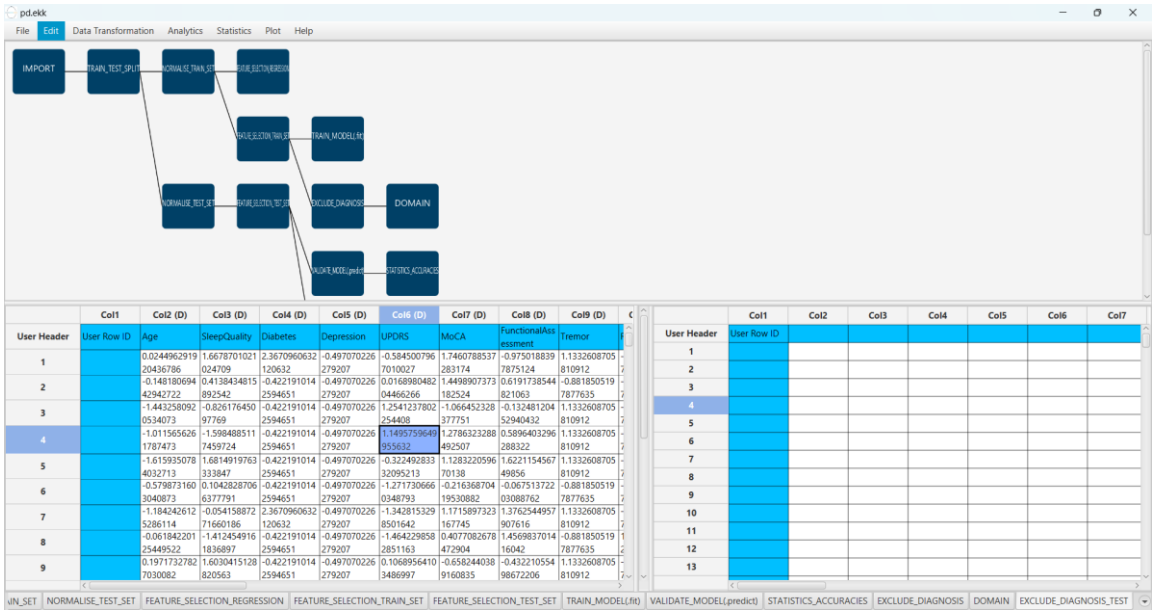
The results will appear on the output spreadsheet.



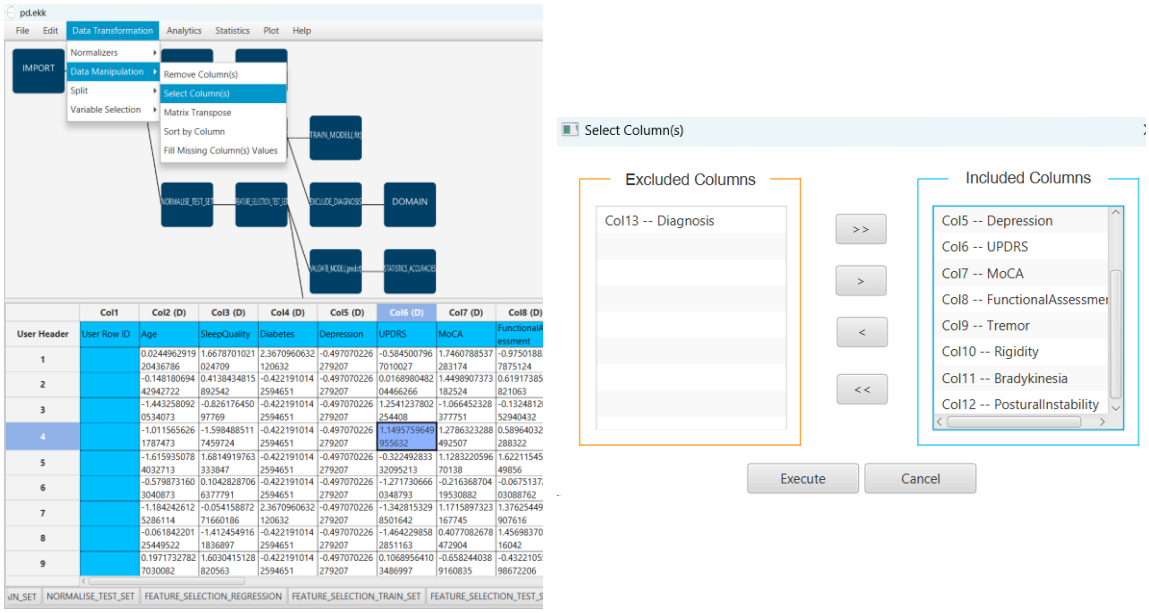
Step 12.b: Check the test set reliability

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".



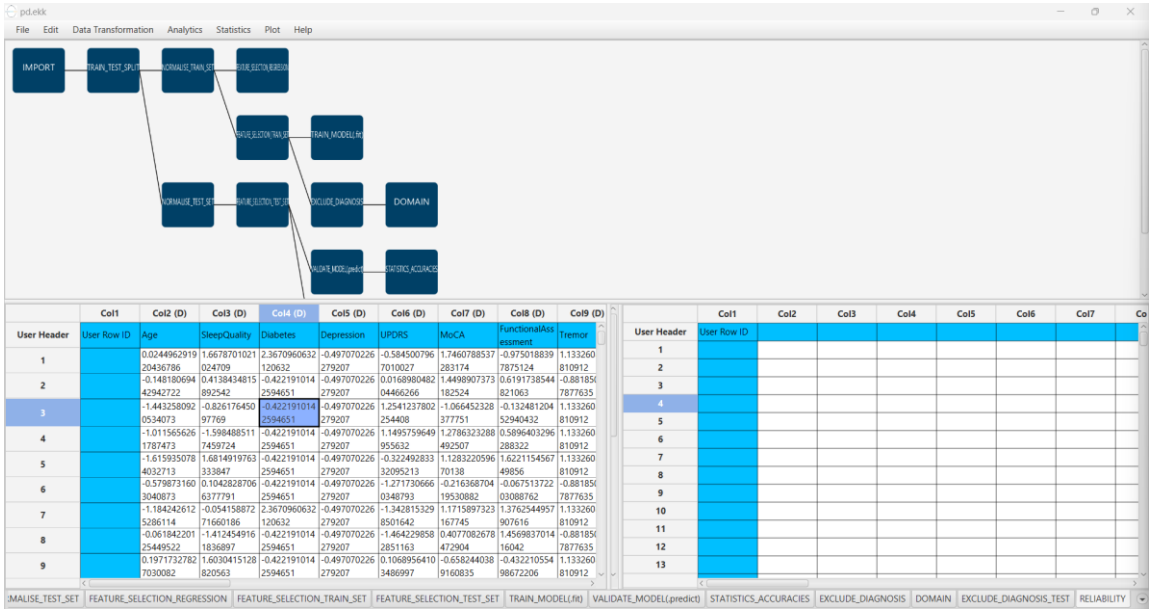
Filter the data to exclude the column that corresponds to the "Diagnosis" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except "Diagnosis".



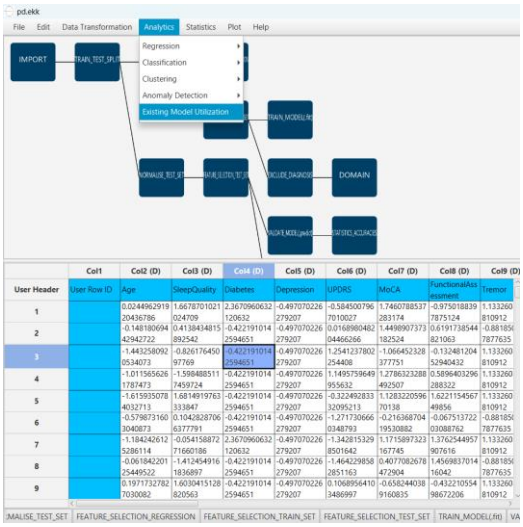
The results will appear on the output spreadsheet.

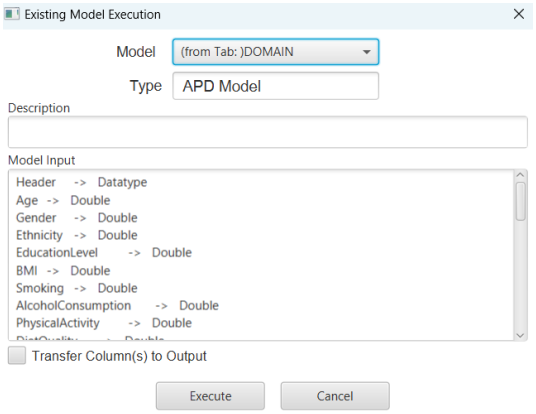
Create a new tab by pressing the "+" button on the bottom of the page with the name "RELIABILITY".

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".

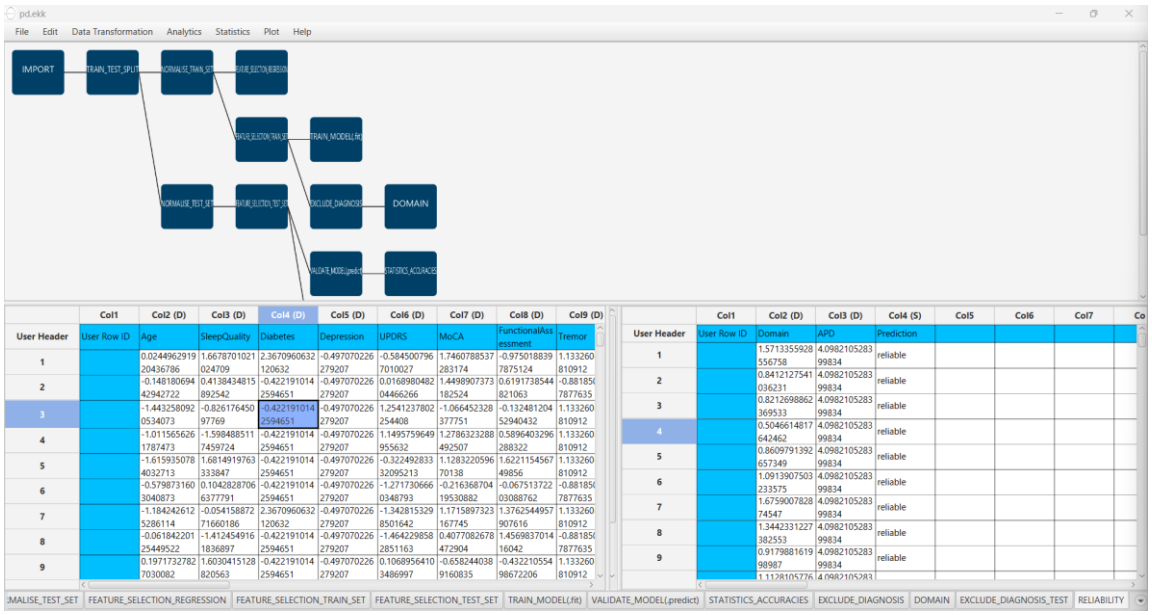


Check the Reliability of the test set predictions by browsing: "Analytics" → "Existing Model Utilization". Then select as Model "(from Tab:) DOMAIN".





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: