Parkinson's Disease Dataset Analysis

This dataset, which can be found in https://www.kaggle.com/datasets/rabieelkharoua/parkinsons-disease-dataset-analysis/data, 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. It is valuable for researchers and data scientists aiming to explore factors associated with Parkinson's Disease, develop predictive models, and conduct statistical analyses.

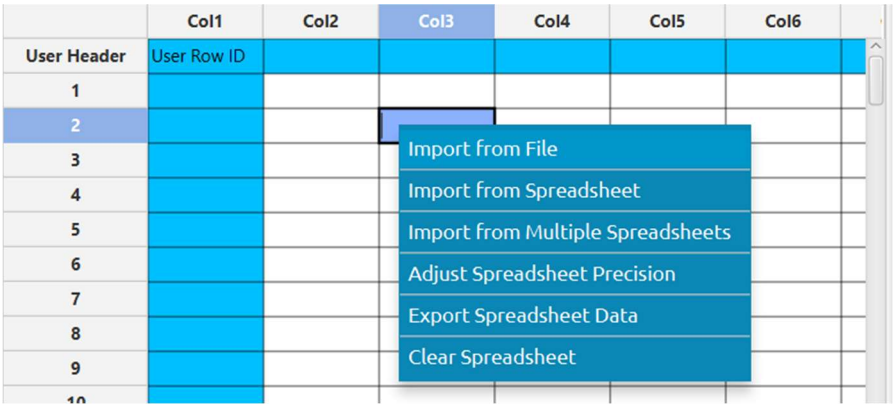
The categorical features included are encoded as:

- Gender: Male (0), Female (1)
- Ethnicity: Caucasian (0), African American (1), Asian (2), Other (3)
- EducationLevel: None (0), High School (1), Bachelor's (2), Higher (3)
- DoctorInCharge: this column contains confidential information therefore all samples have taken the value "DrXXXConfid"

Isalos version used: 2.0.6

Step 1: Import data from file

Right click on the input spreadsheet (left) and choose the option "Import from File". Then navigate through your files to load the one with the Parkinson's disease data.

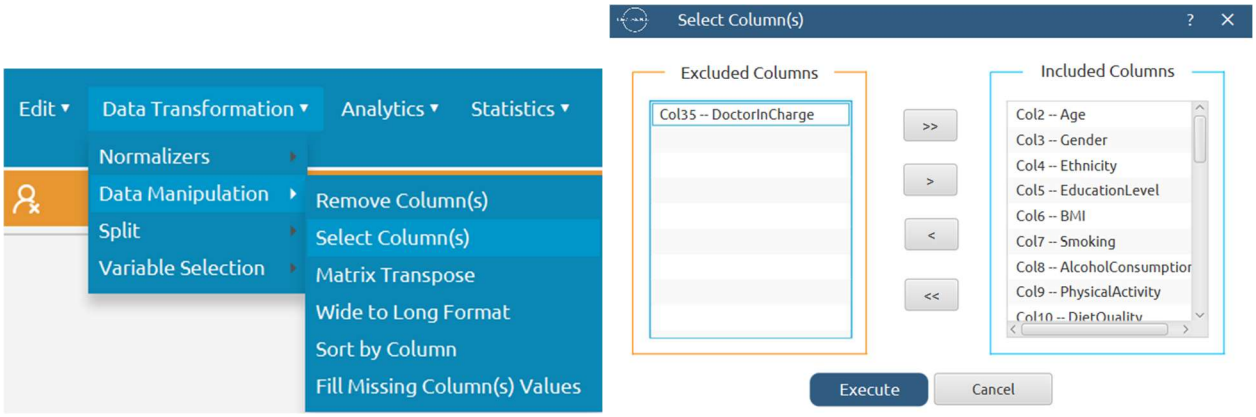


The data will appear on the left spreadsheet.

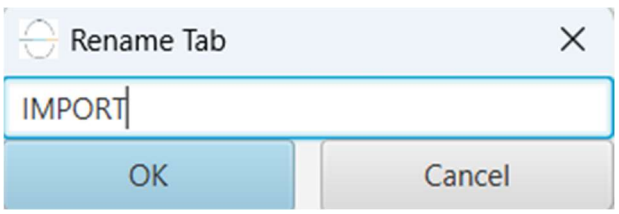
	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (I)	Col6 (D)	Col7 (I)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (I)	Col13 (I)	Col14 (I)	Col15 (I)
User Header	User Row ID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholConsumption	PhysicalActivity	DietQuality	SleepQuality	FamilyHistoryParkinsons	TraumaticBrainInjury	Hypertension	Diabetes
1	3058	85	0	3	1	19.619877964608285	0	5.108240606772179	1.3806599170830036	3.893969135156027	9.283194447541694	0	0	0	0
2	3059	75	0	0	2	16.247339156 47557	1	6.027648029307635	8.409804050283633	8.513428249596062	5.602469505671129	0	0	0	0
3	3060	70	1	0	0	15.368238711 416375	0	2.242135330530093	0.213274590901078915	6.498804606058098	9.929823812340913	0	0	0	1
4	3061	52	0	0	0	15.454557328 79956	0	5.997787562949295	1.375045164648543	6.715033333287671	4.196189318377978	0	0	0	0
5	3062	87	0	0	1	18.616041769 16242	0	9.775242922861011	1.188607062023 7166	4.657572037126733	9.363924681487411	0	0	0	0
6	3063	68	1	2	1	39.423311410 061466	1	13.59688889 6832859	7.796704003664869	7.070238878056 8075	7.737548608057438	0	0	0	0
7	3064	78	1	0	0	30.542003287 867175	1	2.011281312 5692597	9.028536304 01518	9.838445925686692	5.981983542 02343	0	0	1	0
8	3065	70	1	0	0	36.758281614 016326	1	19.98886597 2822232	3.891748622 0750854	3.421960005833149	7.895866238783689	0	0	0	1
9	3066	80	0	2	1	22.380586503 36209	1	7.293287714 899552	2.595670177 298847	4.784827138 79793	4.170469708316643	0	0	0	1
10	3067	71	0	3	2	23.727086277 31125	1	17.78290984 74483	7.344890315 666836	3.393018461974746	9.245379606923356	0	1	0	0
11	3068	70	0	0	3	38.482544735 296	0	6.639761992 8966245	7.872186696 683123	9.225710184366806	5.721854796691623	0	0	0	0
12	3069	53	0	0	1	35.896738697 9048	1	5.212906262 505421	7.185203018 161001	7.918912221561628	5.569759738189951	0	0	0	0
13	3070	74	0	2	2	30.225512078 17625	0	3.762918066 4834982	4.316651242 649941	5.112520425139752	8.512502775799907	0	0	0	0
14	3071	87	1	1	2	38.298306550 44067	0	12.61599521 504917	9.299289995 585761	6.715579110270009	5.563065282178517	0	0	1	0
15	3072	58	0	0	3	34.965322873 96592	1	11.70859735 0605928	4.392463151 845714	5.182038871201481	4.219612497925302	0	0	0	0

Step 2: Manipulate data

In our dataset there are not any empty values, so we can select all the columns to be used. However, since the column "DrXXXConfid" does not offer any significant information about the Parkinson's disease diagnosis we will exclude it. On the menu click on *Data Transformation → Data Manipulation → Select Column(s)* and select all columns except "DrXXXConfid".



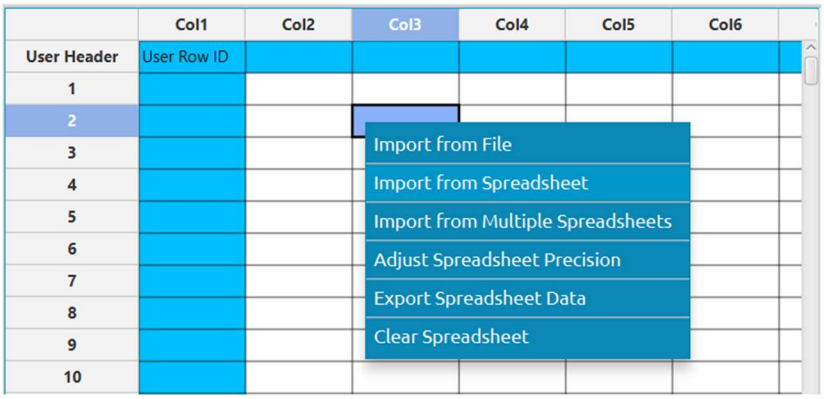
All of the data will appear in the output (right) spreadsheet. This tab can be renamed "IMPORT" by right-clicking on it and choosing the "Rename" option.



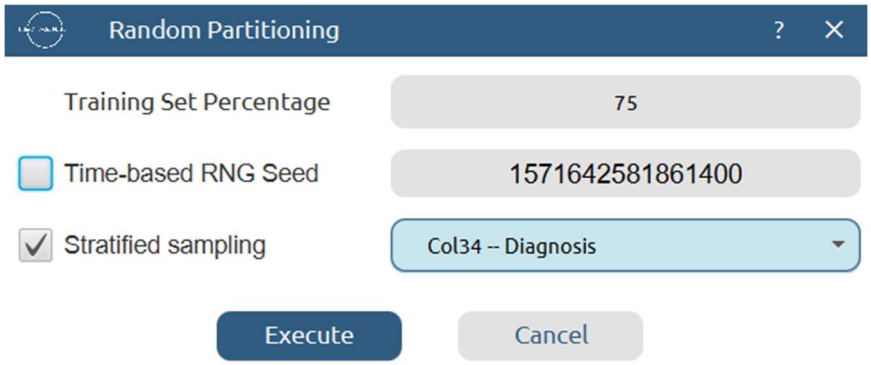
Step 3: Split data

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 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 be two separate spreadsheets, "TRAIN_TEST_SPLIT: Training Set" and "TRAIN_TEST_SPLIT: Test Set", which will be available to import into the next tabs.

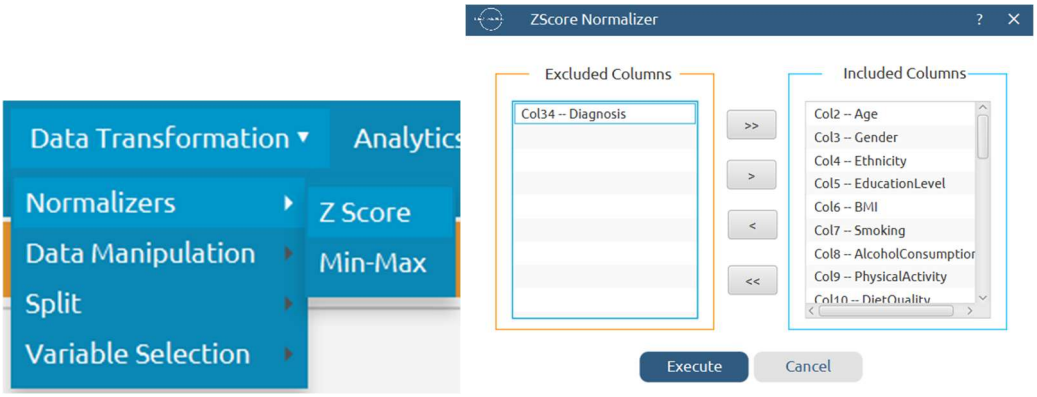
Step 4: Normalize the training set

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

Import into the input spreadsheet of the "NORMALIZE_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".

	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (I)	Col6 (D)	Col7 (I)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (I)	Col13 (I)	Col14 (I)	Col15 (I)
User Header	User Row ID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholConsumption	PhysicalActivity	DietQuality	SleepQuality	FamilyHistoryParkinsons	TraumaticBrainInjury	Hypertension	Diabetes
1	3058	85	0	3	1	19.6198780	0	5.1082406	1.3806599	3.8939691	9.2831944	0	0	0	0
2	3059	75	0	0	2	16.2473392	1	6.0276480	8.4098041	8.5134282	5.6024695	0	0	0	0
3	3061	52	0	0	0	15.4545573	0	5.9977876	1.3750452	6.7150333	4.1961893	0	0	0	0
4	3062	87	0	0	1	18.6160418	0	9.7752429	1.1886071	4.6575720	9.3639247	0	0	0	0
5	3067	71	0	3	2	23.7270863	1	17.7829098	7.3448903	3.3930185	9.2453796	0	1	0	0
6	3069	53	0	0	1	35.8967387	1	5.2129063	7.1852030	7.9189122	5.5697597	0	0	0	0
7	3070	74	0	2	2	30.2255121	0	3.7629181	4.3166512	5.1125204	8.5125028	0	0	0	0
8	3071	87	1	1	2	38.2983066	0	12.6159952	9.2992900	6.7155791	5.5630653	0	0	1	0
9	3072	58	0	0	3	34.9653229	1	11.7085974	4.3924632	5.1820389	4.2196125	0	0	0	0
10	3074	54	1	0	0	28.0495801	1	7.2607338	4.3116174	2.5099362	7.4952082	1	0	0	0
11	3076	51	1	0	0	19.0020040	0	1.5321354	8.1221542	4.8507528	9.9536374	0	0	1	0
12	3077	55	0	1	2	22.5483386	0	11.5791849	8.8936626	4.1093017	4.2185137	0	0	0	0
13	3078	62	1	0	2	29.7272418	1	2.0324380	8.9341903	7.0743530	5.7221967	0	0	0	0
14	3080	74	1	0	1	20.6172019	0	5.1986751	6.7044843	7.0484164	9.6554525	0	0	0	0
15	3081	60	1	0	0	38.1336396	1	10.7813470	7.7201343	5.0858216	7.2366864	0	0	0	1

Normalize the data using Z-score: *Data Transformation → Normalizers → Z Score* and select all columns except the "Diagnosis" target column.



The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)	Col13 (D)	Col14 (D)	Col15 (D)
User Header	User Row ID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholConsumption	PhysicalActivity	DietQuality	SleepQuality	FamilyHistoryParkinsons	TraumaticBrainInjury	Hypertension	Diabetes
1	3058	1.3104154	-0.9790044	2.3462929	-0.3688004	-1.0606881	-0.6448875	-0.8546182	-1.2534821	-0.3284443	1.3124709	-0.4127816	-0.3333450	-0.4211497	-0.4138314
2	3059	0.4536617	-0.9790044	-0.6796776	0.7489268	-1.5343834	1.5496759	-0.6920410	1.1691684	1.2692687	-0.7909199	-0.4127816	-0.3333450	-0.4211497	-0.4138314
3	3061	-1.5168719	-0.9790044	-0.6796776	-1.4865276	-1.6457348	-0.6448875	-0.6973212	-1.2554173	0.6472654	-1.5945542	-0.4127816	-0.3333450	-0.4211497	-0.4138314
4	3062	1.4817662	-0.9790044	-0.6796776	-0.3688004	-1.2016835	-0.6448875	-0.0293604	-1.3196746	-0.0643402	1.3586051	-0.4127816	-0.3333450	-0.4211497	-0.4138314
5	3067	0.1109602	-0.9790044	2.3462929	0.7489268	-0.4838037	1.5496759	1.3866211	0.8021374	-0.5017060	1.2908612	-0.4127816	2.9979952	-0.4211497	-0.4138314
6	3069	-1.4311966	-0.9790044	-0.6796776	-0.3688004	1.2255040	1.5496759	-0.8361104	0.7470999	1.0636459	-0.8096122	-0.4127816	-0.3333450	-0.4211497	-0.4138314
7	3070	0.3679863	-0.9790044	1.3376361	0.7489268	0.4289429	-0.6448875	-1.0925092	-0.2415693	0.0930109	0.8720507	-0.4127816	-0.3333450	-0.4211497	-0.4138314
8	3071	1.4817662	1.0207989	0.3289793	0.7489268	1.5628200	-0.6448875	0.4729648	1.4757368	0.6474541	-0.8134379	-0.4127816	-0.3333450	2.3729487	-0.4138314
9	3072	-1.0028197	-0.9790044	-0.6796776	1.8666540	1.0946805	1.5496759	0.3125112	-0.2154401	0.1170550	-1.5811687	-0.4127816	-0.3333450	-0.4211497	-0.4138314
10	3074	-1.3455212	1.0207989	-0.6796776	-1.4865276	0.1233190	1.5496759	-0.4739966	-0.2433043	-0.8071340	0.2907065	2.4210541	-0.3333450	-0.4211497	-0.4138314
11	3076	-1.6025473	1.0207989	-0.6796776	-1.4865276	-1.1474726	-0.6448875	-1.4869746	1.0700275	0.0024745	1.6956029	-0.4127816	-0.3333450	2.3729487	-0.4138314
12	3077	-1.2598458	-0.9790044	0.3289793	0.7489268	-0.6493666	-0.6448875	0.2896274	1.3359341	-0.2539682	-1.5817966	-0.4127816	-0.3333450	-0.4211497	-0.4138314
13	3078	-0.6601182	1.0207989	-0.6796776	0.7489268	0.3589576	1.5496759	-1.3985070	1.3499023	0.7715418	-0.7225005	-0.4127816	-0.3333450	-0.4211497	-0.4138314
14	3080	0.3679863	1.0207989	-0.6796776	-0.3688004	-0.9206074	-0.6448875	-0.8386269	0.5814163	0.7625712	1.5252019	-0.4127816	-0.3333450	-0.4211497	-0.4138314
15	3081	-0.8314689	1.0207989	-0.6796776	-1.4865276	1.5396914	1.5496759	0.1485472	0.9314682	0.0837767	0.1429713	-0.4127816	-0.3333450	-0.4211497	2.4149125

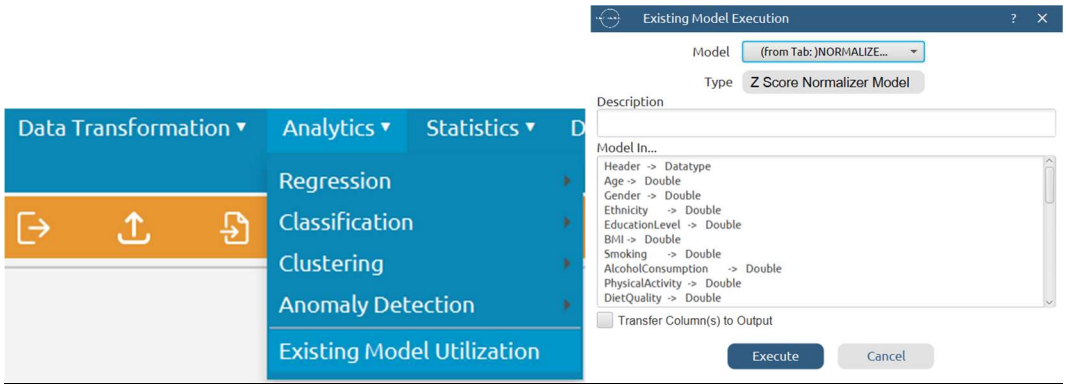
Step 5: Normalize the test set

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

Import into the input spreadsheet of the "NORMALIZE_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".

	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (I)	Col6 (D)	Col7 (I)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (I)	Col13 (I)	Col14 (I)	Col15 (I)
User Header	User Row ID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholConsumption	PhysicalActivity	DietQuality	SleepQuality	FamilyHistoryParkinsons	TraumaticBrainInjury	Hypertension	Diabetes
1	3060	70	1	0	0	15.3682387	0	2.2421353	0.2132746	6.4988046	9.9298238	0	0	0	1
2	3063	68	1	2	1	39.4233114	1	13.5968889	7.7967040	7.0702389	7.7375486	0	0	0	0
3	3064	78	1	0	0	30.5420033	1	2.0112813	9.0285363	9.8384459	5.9819835	0	0	1	0
4	3065	70	1	0	0	36.7582816	1	19.9888660	3.8917486	3.4219600	7.8958662	0	0	0	1
5	3066	80	0	2	1	22.3805865	1	7.2932877	2.5956702	4.7848271	4.1704697	0	0	0	1
6	3068	70	0	0	3	38.4825447	0	6.6397620	7.8721867	9.2257102	5.7218548	0	0	0	0
7	3073	56	1	0	0	18.9587813	1	2.0471205	9.4328303	1.7202773	4.0411807	0	1	0	0
8	3075	57	1	0	1	21.8561230	0	0.2552298	4.0409650	1.1428182	5.8704956	1	0	0	0
9	3079	79	1	3	1	33.2476178	1	9.5452800	6.9559990	1.5778409	6.3756426	0	0	0	0
10	3083	71	1	2	1	15.8636029	0	19.5917183	7.2423151	5.7027472	6.3956378	0	0	1	0
11	3084	79	1	1	2	36.9054342	0	9.8905980	7.6781793	9.7631047	8.5091359	0	1	0	0
12	3087	61	0	0	2	36.7637247	1	2.6988171	2.9874866	8.5623118	6.8286560	0	0	1	0
13	3093	74	1	0	1	23.6097394	0	13.1723855	1.5994689	1.0797522	4.9918173	0	1	0	0
14	3094	68	0	0	1	22.4658013	0	7.2068544	0.7087926	0.0000105	9.8065964	1	0	0	0
15	3095	51	1	0	2	16.8224868	1	16.0108172	8.1821264	3.4269047	9.1842093	0	0	0	0

Normalize the test set using the existing normalizer of the training set: *Analytics → Existing Model Utilization → Model (from Tab:) NORMALIZE_TRAIN_SET*



The results will appear on the output spreadsheet.

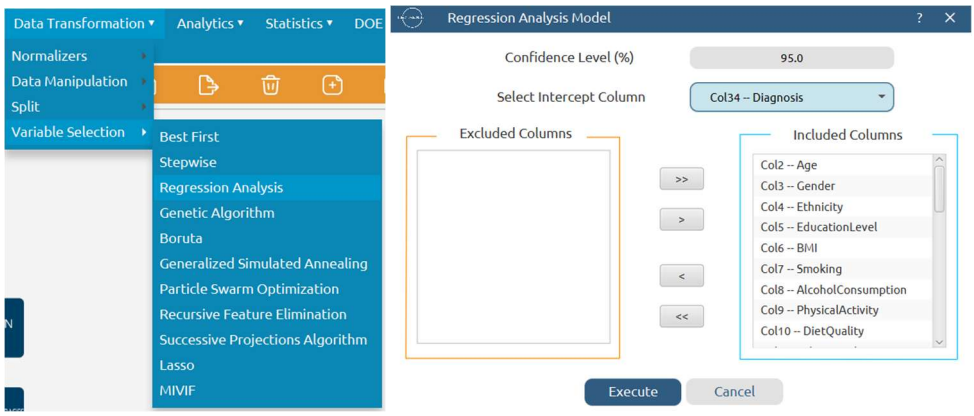
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)	Col11 (D)	Col12 (D)	Col13 (D)	Col14 (D)	Col15 (D)
User Header	User Row ID	Age	Gender	Ethnicity	EducationLevel	BMI	Smoking	AlcoholConsumption	PhysicalActivity	DietQuality	SleepQuality	FamilyHistoryParkinsons	TraumaticBrainInjury	Hypertension	Diabetes
1	3060	0.0252848	1.0207989	-0.6796776	-1.4865276	-1.6578588	-0.6448875	-1.3614266	-1.6558307	0.5724792	1.6819944	-0.4127816	-0.3333450	-0.4211497	2.4149125
2	3063	-0.1460659	1.0207989	1.3376361	-0.3688004	1.7208343	1.5496759	0.6464145	0.9578586	0.7701188	0.4291946	-0.4127816	-0.3333450	-0.4211497	-0.4138314
3	3064	0.7106878	1.0207989	-0.6796776	-1.4865276	0.4733962	1.5496759	-1.4022481	1.3824194	1.7275471	-0.5740424	-0.4127816	-0.3333450	2.3729487	-0.4138314
4	3065	0.0252848	1.0207989	-0.6796776	-1.4865276	1.3465133	1.5496759	1.7766965	-0.3880154	-0.4916962	0.5196669	-0.4127816	-0.3333450	-0.4211497	2.4149125
5	3066	0.8820386	-0.9790044	1.3376361	-0.3688004	-0.6729285	1.5496759	-0.4682402	-0.8347191	-0.0203270	-1.6092519	-0.4127816	-0.3333450	-0.4211497	2.4149125
6	3068	0.0252848	-0.9790044	-0.6796776	1.8666540	1.5886974	-0.6448875	-0.5838020	0.9838743	1.5156227	-0.7226958	-0.4127816	-0.3333450	-0.4211497	-0.4138314
7	3073	-1.1741704	1.0207989	-0.6796776	-1.4865276	-1.1535435	1.5496759	-1.3959107	1.5217625	-1.0802501	-1.6831356	-0.4127816	2.9979952	-0.4211497	-0.4138314
8	3075	-1.0884951	1.0207989	-0.6796776	-0.3688004	-0.7465928	-0.6448875	-1.7127675	-0.3365868	-1.2799734	-0.6377534	2.4210541	-0.3333450	-0.4211497	-0.4138314
9	3079	0.7963632	1.0207989	2.3462929	-0.3688004	0.8534175	1.5496759	-0.0700244	0.6681029	-1.1295140	-0.3490816	-0.4127816	-0.3333450	-0.4211497	-0.4138314
10	3083	0.1109602	1.0207989	1.3376361	-0.3688004	-1.5882816	-0.6448875	1.7064696	0.7667840	0.2971502	-0.3376551	-0.4127816	-0.3333450	2.3729487	-0.4138314
11	3084	0.7963632	1.0207989	0.3289793	0.7489268	1.3671819	-0.6448875	-0.0089624	0.9170081	1.7014891	0.8701266	-0.4127816	2.9979952	-0.4211497	-0.4138314
12	3087	-0.7457936	-0.9790044	-0.6796776	0.7489268	1.3472779	1.5496759	-1.2806723	-0.6996765	1.2861759	-0.0902021	-0.4127816	-0.3333450	2.3729487	-0.4138314
13	3093	0.3679863	1.0207989	-0.6796776	-0.3688004	-0.5002858	-0.6448875	0.5713503	-1.1780678	-1.3017858	-1.1398838	-0.4127816	2.9979952	-0.4211497	-0.4138314
14	3094	1.5674416	-0.9790044	-0.6796776	-0.3688004	-0.6609595	-0.6448875	-0.4835240	-1.4850465	-1.6752315	1.6115747	2.4210541	-0.3333450	-0.4211497	-0.4138314
15	3095	-1.6025473	1.0207989	-0.6796776	0.7489268	-1.4536001	1.5496759	1.0732651	1.0906975	-0.4899859	1.2559047	-0.4127816	-0.3333450	-0.4211497	-0.4138314

Step 6: Regression Analysis

We want to choose the features that will be the most useful for predicting the Parkinson's diagnosis. Create a new tab by pressing the "+" button on the bottom of the page with the name "REGRESSION_ANALYSIS".

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

Conduct regression analysis by choosing: *Data Transformation → Variable Selection → Regression Analysis*



The results will appear on the output spreadsheet.

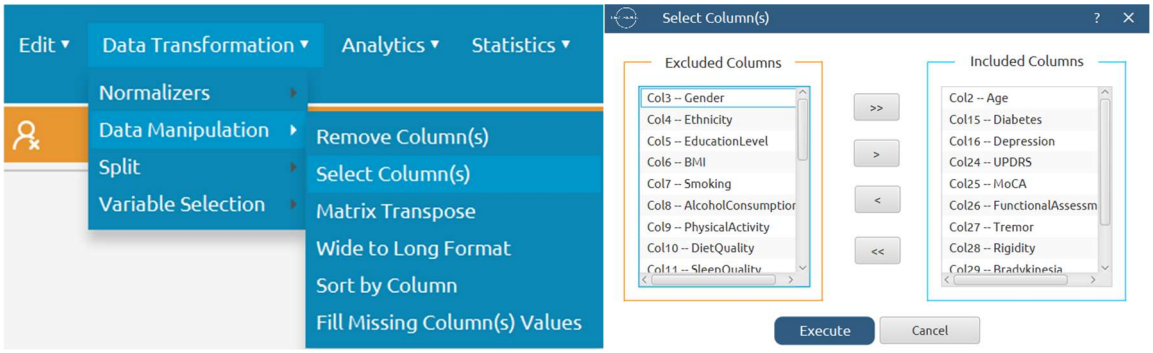
	Col1	Col2 (S)	Col3 (S)	Col4 (S)	Col5 (S)	Col6 (S)	Col7 (S)	Col8 (S)
User Header	User Row ID							
1		Regression Statistics						
2		Multiple R	0.6488624					
3		R Square	0.4210224					
4		Adjusted R Square	0.4090384					
5		Standard Error	0.3733720					
6		Observations	1579					
7								
8			Degrees of Freedom	Sum of Squares	Mean Square	F-statistic	Significance F	
9		Regression	32	156.7243279	4.8976352	35.1320087	0E-7	
10		Residual	1546	215.5226639	0.1394066			
11		Total	1578	372.2469918				
12			Coefficients	Standard Error	t-statistic	P-value	Lower 95.0%	Upper 95.0%
13		Diagnosis	0.6193794	0.0093962	65.9183095	0.0	0.6009488	0.6378099
14		Age	0.0318654	0.0094654	3.3664925	0.0007800	0.0132989	0.0504318
15		Gender	0.0067077	0.0095466	0.7026302	0.4823920	-0.0120179	0.0254334
16		Ethnicity	-0.0096292	0.0095068	-1.0128769	0.3112774	-0.0282768	0.0090184
17		EducationLevel	-0.0061676	0.0094852	-0.6502377	0.5156353	-0.0247729	0.0124376
18		BMI	0.0039260	0.0095492	0.4111293	0.6810347	-0.0148048	0.0226567
19		Smoking	-0.0035615	0.0095095	-0.3745179	0.7080704	-0.0222144	0.0150914
20		AlcoholConsumption	0.0138605	0.0094733	1.4631186	0.1436382	-0.0047213	0.0324424
21		PhysicalActivity	0.0033893	0.0095047	0.3565983	0.7214412	-0.0152540	0.0220327
22		DietQuality	-0.0103349	0.0094643	-1.0919957	0.2750051	-0.0288991	0.0082292
23		SleepQuality	-0.0157571	0.0094695	-1.6639748	0.0963201	-0.0343316	0.0028174
24		FamilyHistoryParkinsons	0.0147373	0.0094975	1.5517018	0.1209383	-0.0038921	0.0333667
25		TraumaticBrainInjury	0.0130664	0.0095600	1.3667840	0.1718917	-0.0056855	0.0318184
26		Hypertension	0.0068568	0.0095057	0.7213385	0.4708104	-0.0117886	0.0255023

Step 7: Feature Selection: Train set

We need to select the features of the train set that the regression analysis indicated. Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_TRAIN".

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

Select the columns that correspond to the important features: *Data Transformation → Data Manipulation → Select Column(s)*



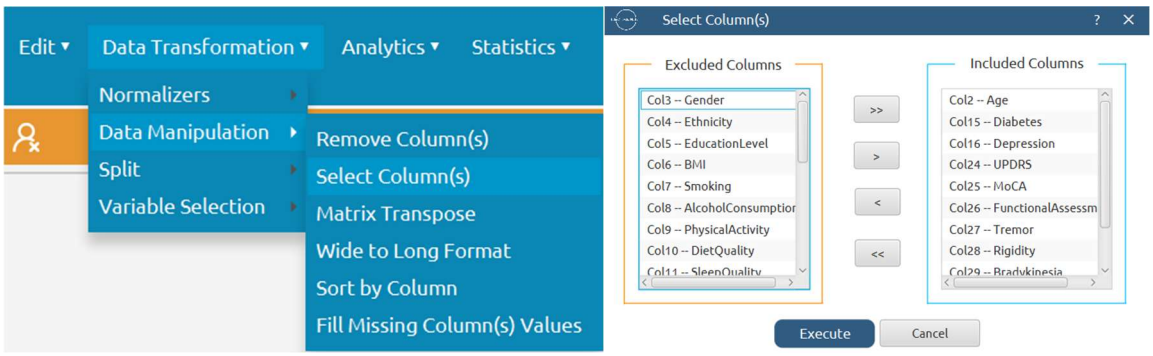
The results will appear on the output spreadsheet.

Step 8: Feature Selection: Test set

We need to select the features of the test set that the regression analysis indicated. Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_TEST".

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

Select the columns that correspond to the important features: *Data Transformation → Data Manipulation → Select Column(s)*



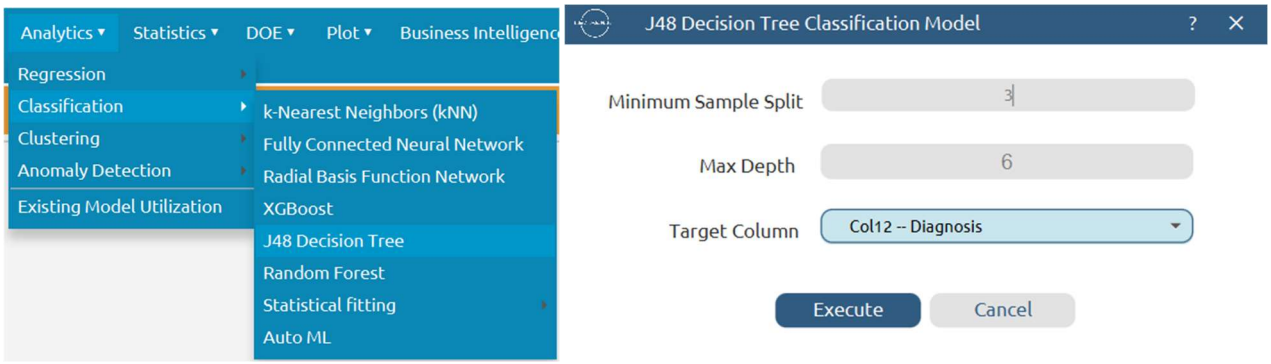
The results will appear on the output spreadsheet.

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" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

Use the J48 decision tree method to train and fit the model: *Analytics → Classification → J48 Decision Tree*



The predictions will appear on the output spreadsheet.

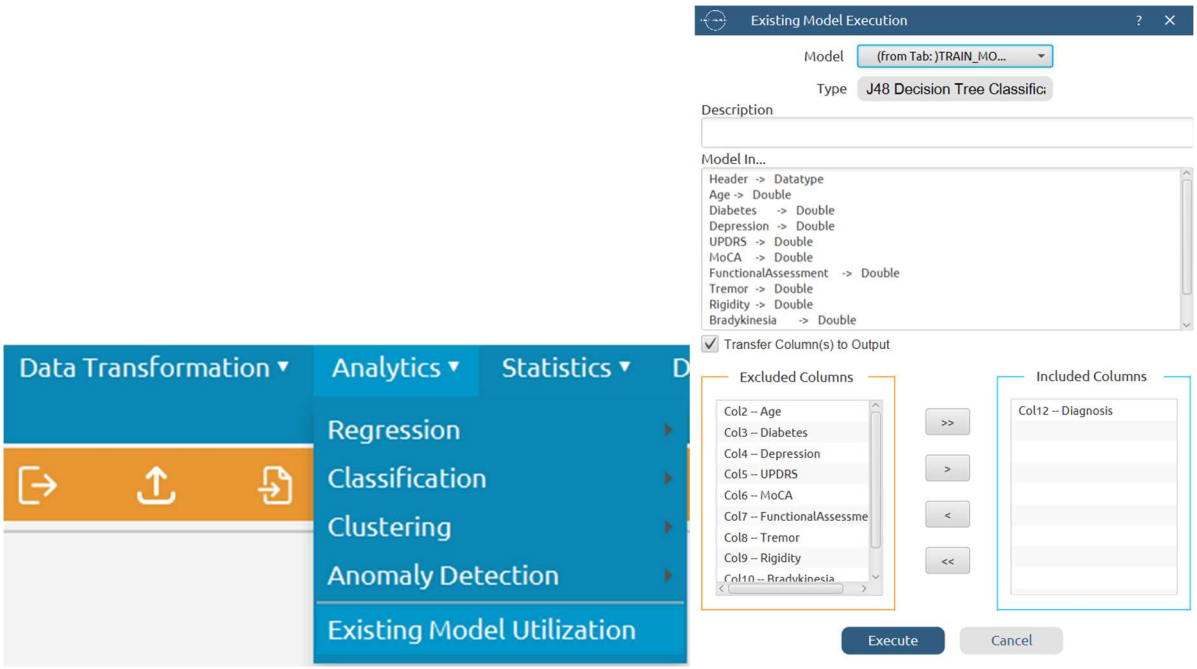
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Diagnosis	Prediction
1	3058	0.0	0.0
2	3059	1.0	1.0
3	3061	1.0	1.0
4	3062	0.0	1.0
5	3067	0.0	0.0
6	3069	1.0	1.0
7	3070	1.0	1.0
8	3071	1.0	1.0
9	3072	1.0	1.0
10	3074	1.0	1.0
11	3076	1.0	1.0
12	3077	0.0	1.0
13	3078	1.0	1.0
14	3080	0.0	0.0
15	3081	1.0	1.0

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" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

To validate the model: *Analytics → Existing Model Utilization→ Model (from Tab:) TRAIN_MODEL(.fit)*. Choose the column "Diagnosis" to be transferred to the output spreadsheet.



The predictions will appear on the output spreadsheet.

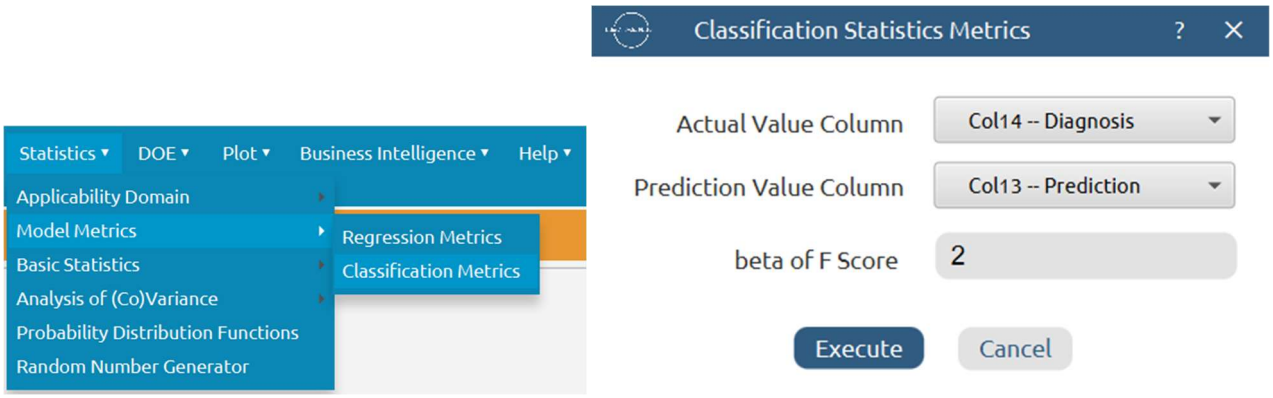
	Col13 (D)	Col14 (D)
User Header	Prediction	Diagnosis
1	1.0	1.0
2	0.0	0.0
3	0.0	0.0
4	1.0	1.0
5	1.0	1.0
6	1.0	1.0
7	1.0	1.0
8	1.0	1.0
9	1.0	1.0
10	1.0	1.0
11	1.0	1.0
12	1.0	1.0
13	1.0	1.0
14	1.0	1.0
15	1.0	0.0

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: *Statistics → Model Metrics → Classification Metrics*



The results will appear on the output spreadsheet.

	Col1 (S)	Col2 (D)	Col3 (S)	Col4 (S)
User Header	User Row ID			
1			Predicted Class	Predicted Class
2			1.0	0.0
3	Actual Class	1.0	302	24
4	Actual Class	0.0	29	171
5				
6				
7	Classification Accuracy	0.8992395		
8				
9	Precision		0.9123867	0.8769231
10				
11	Recall/Sensitivity		0.9263804	0.855
12				
13	Specificity		0.855	0.9263804
14				
15	F1 Score		0.9193303	0.8658228
16				
17	F (beta=2)		0.9235474	0.8592965
18				
19	MCC	0.7853351		

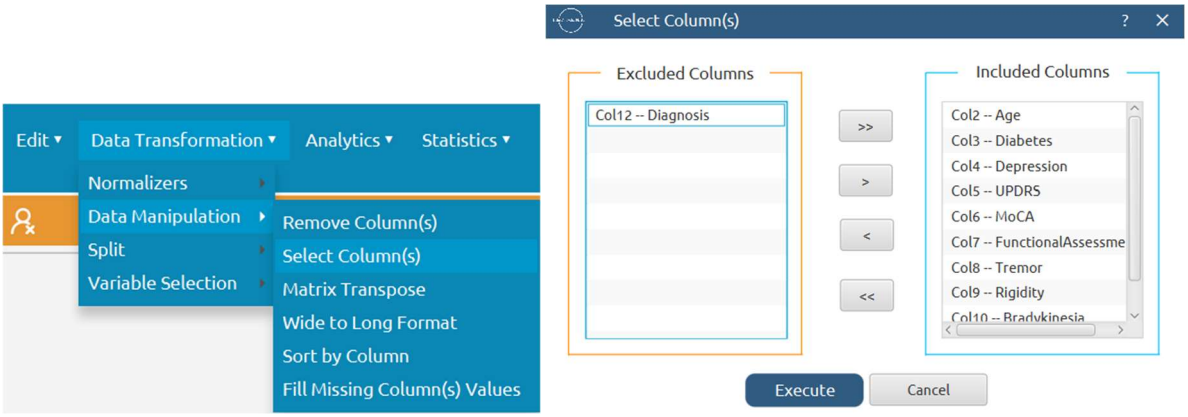
Step 12: Reliability check for each record of the test set

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" tab by right-clicking on the input spreadsheet and then choosing "Import from Spreadsheet".

Manipulate the data to exclude the target column "Diagnosis": *Data Transformation → Data Manipulation → Select Column(s)*

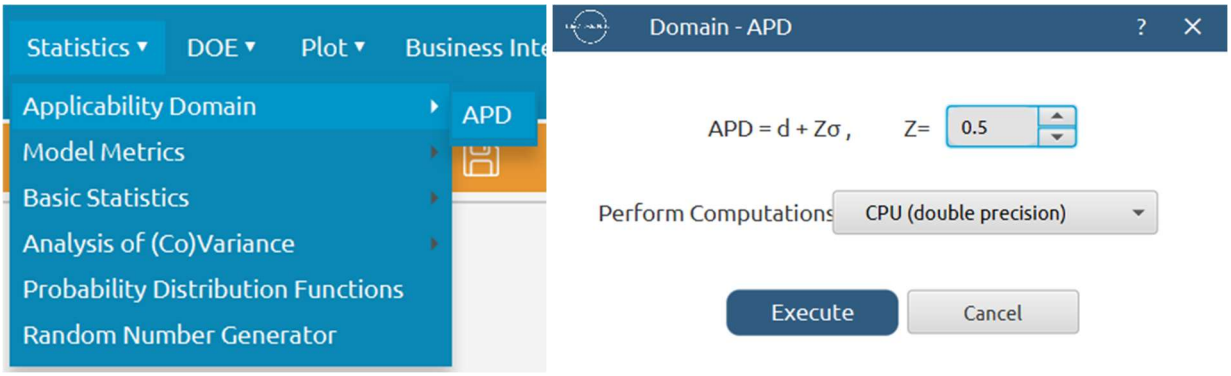


The results will appear on the output spreadsheet.

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

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

Create the domain: *Statistics → Applicability Domain → APD*



The results will appear on the output spreadsheet.

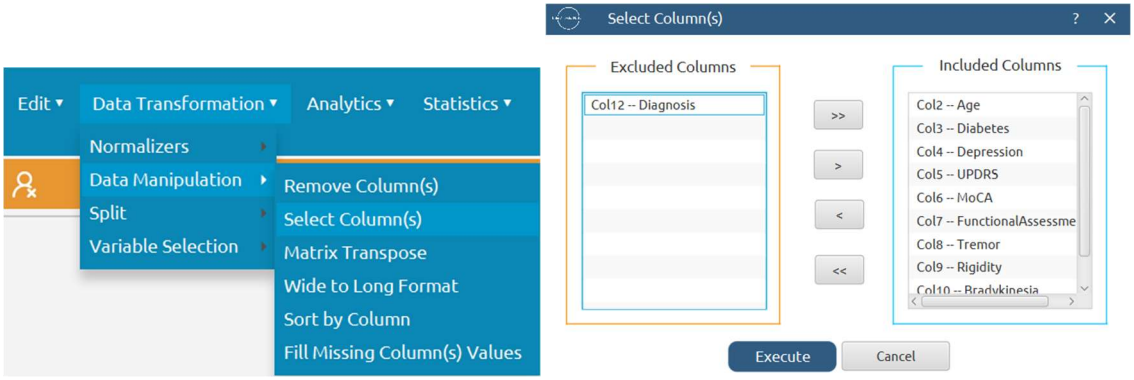
	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1	3058	0.0	3.8564136	reliable
2	3059	0.0	3.8564136	reliable
3	3061	0.0	3.8564136	reliable
4	3062	0.0	3.8564136	reliable
5	3067	0.0	3.8564136	reliable
6	3069	0.0	3.8564136	reliable
7	3070	0.0	3.8564136	reliable
8	3071	0.0	3.8564136	reliable
9	3072	0.0	3.8564136	reliable
10	3074	0.0	3.8564136	reliable
11	3076	0.0	3.8564136	reliable
12	3077	0.0	3.8564136	reliable
13	3078	0.0	3.8564136	reliable
14	3080	0.0	3.8564136	reliable
15	3081	0.0	3.8564136	reliable

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".

Manipulate the data to exclude the target column "Diagnosis": *Data Transformation → Data Manipulation → Select Column(s)*

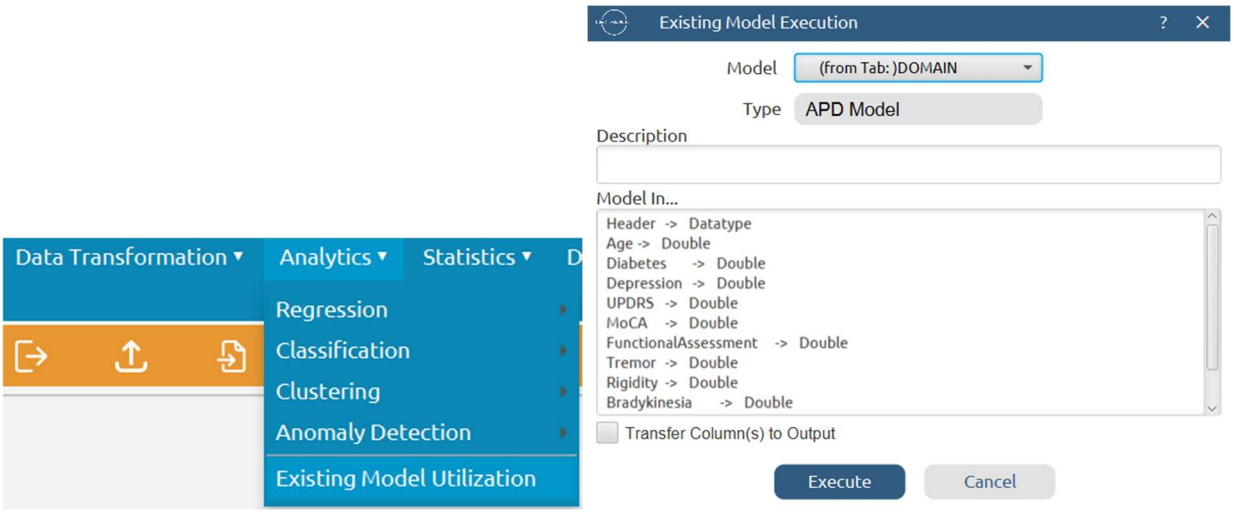


The results will appear on the output spreadsheet.

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

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

Check the Reliability: *Analytics → Existing Model Utilization → Model (from Tab:) DOMAIN*



The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1	3060	1.1839593	3.8564136	reliable
2	3063	0.6034224	3.8564136	reliable
3	3064	0.6159510	3.8564136	reliable
4	3065	0.9854746	3.8564136	reliable
5	3066	0.7865596	3.8564136	reliable
6	3068	1.2343587	3.8564136	reliable
7	3073	0.8994424	3.8564136	reliable
8	3075	0.8175593	3.8564136	reliable
9	3079	0.4345935	3.8564136	reliable
10	3083	0.3292491	3.8564136	reliable
11	3084	0.4777645	3.8564136	reliable
12	3087	0.8664361	3.8564136	reliable
13	3093	0.6531547	3.8564136	reliable
14	3094	0.4522236	3.8564136	reliable
15	3095	0.6650592	3.8564136	reliable

Final Isalos Workflow

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