

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](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:

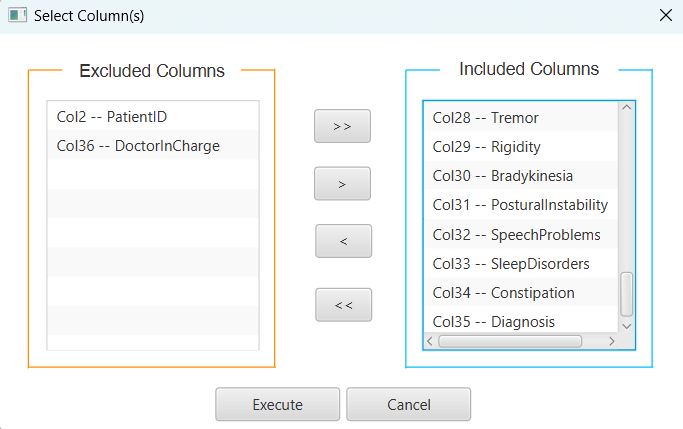
* None (0)
* High School (1)
* Bachelor’s (2)
* Higher (3)

# 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.

# A screenshot of a spreadsheet Description automatically generatedA screenshot of a computer Description automatically generatedStep 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:

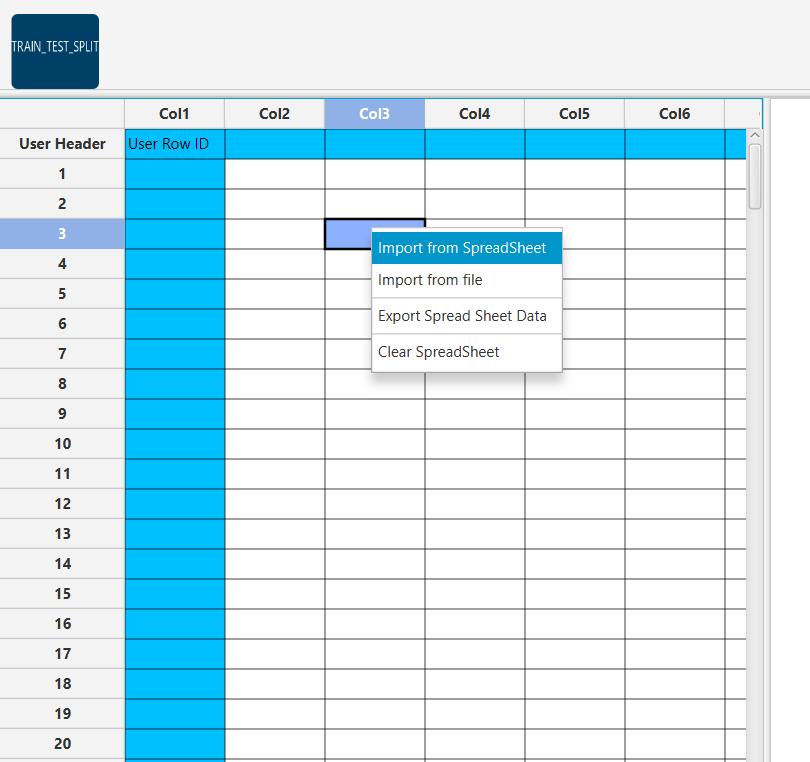
* On the menu click on “Data Transformation” → “Data Manipulation” → “Select Column(s)”
* A screenshot of a computer

  Description automatically generatedSelect 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

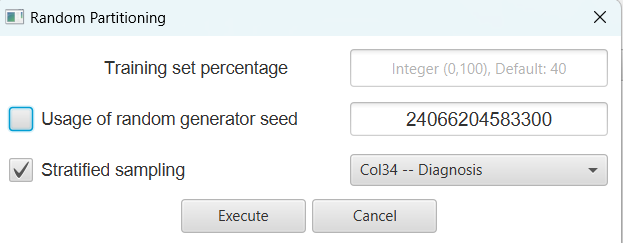
Create a new tab by pressing the “+” button on the bottom of the page with the name “TRAIN\_TEST\_SPLIT” which we will use for splitting to create the train and test set.

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.

A screenshot of a computer

Description automatically generated

# 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”.

A screenshot of a computer

Description automatically generatedImport data into the input spreadsheet of the “NORMALISE\_TRAIN\_SET” tab the train set from the output of the “TRAIN\_TEST\_SPLIT” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”. From the available Select input tab options choose “TRAIN\_TEST\_SPLIT: Training Set”

Normalize the data using Z-score:

“Data Transformation” → “Normalizers” → “Z-Score”

Screens screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThen select all columns and click “Execute”.

A screenshot of a computer

Description automatically generatedThe 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”.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedNormalize the test set using the existing normalizer of the training set:

A screenshot of a computer

Description automatically generated  
“Analytics” → “Existing Model Utilization” → “Model (from Tab: ) NORMALISE\_TRAIN\_SET”.

The results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

# Step 6: Feature selection

Create a new tab by pressing the “+” button on the bottom of the page with the name “FEATURE\_SELECTION\_REGRESSION”.

A screenshot of a computer

Description automatically generatedImport 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:

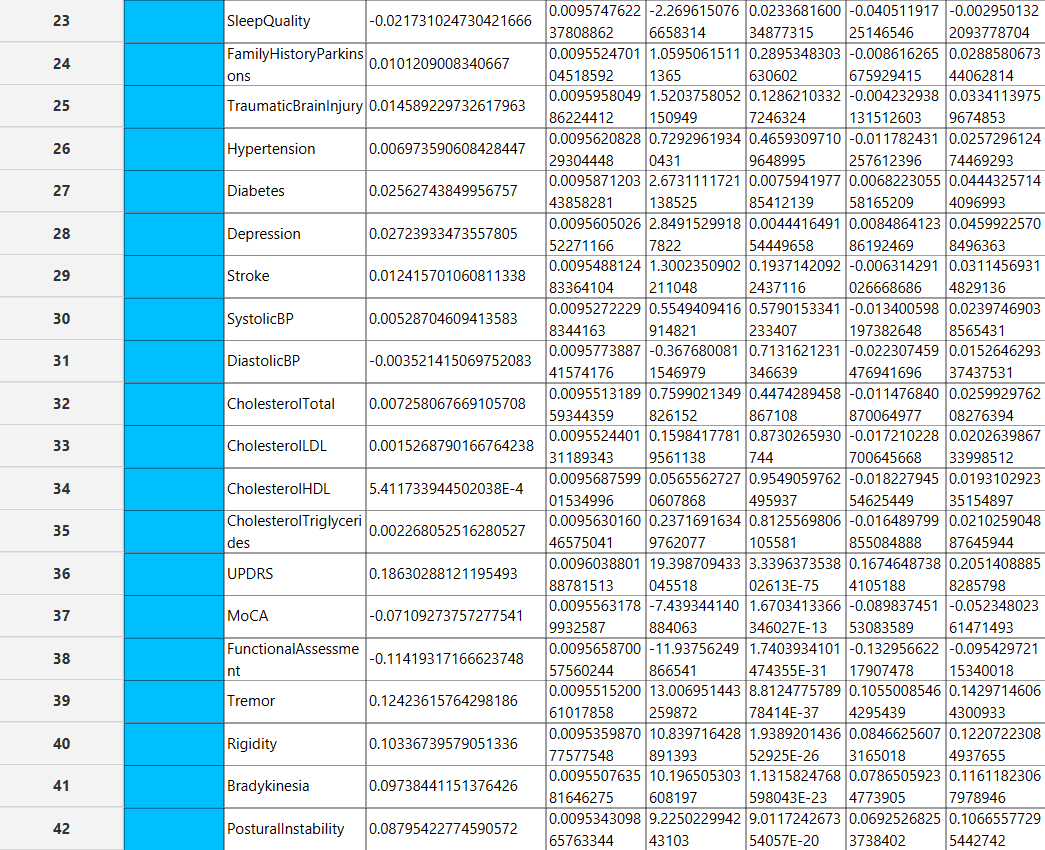
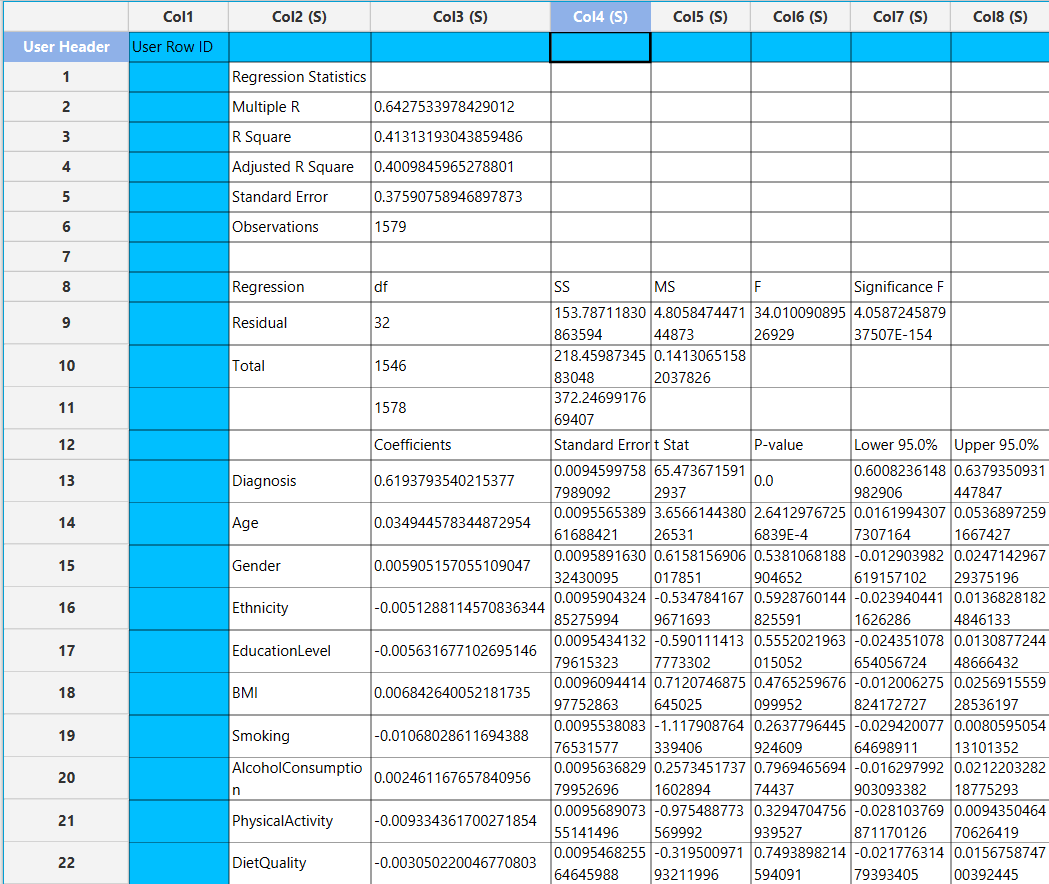
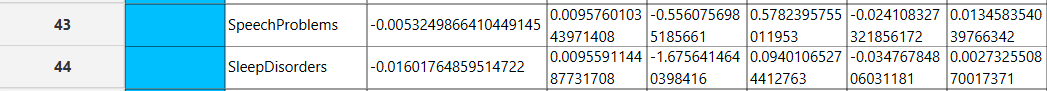
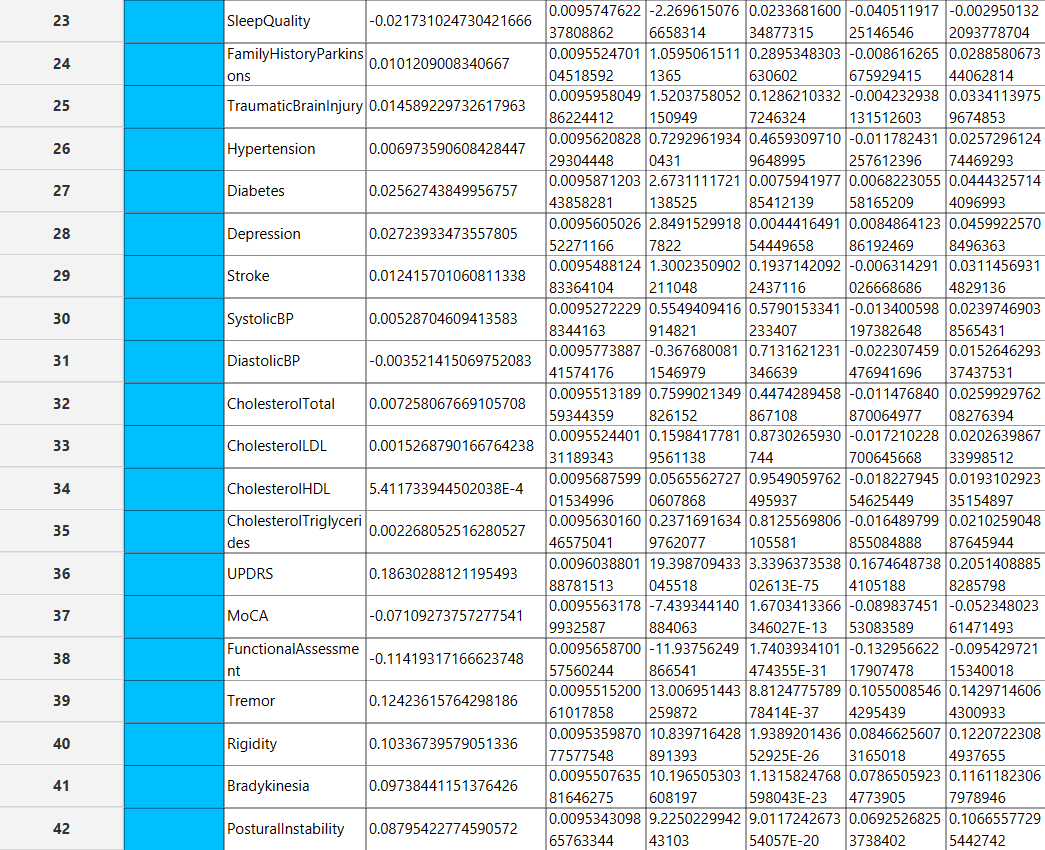
“Data Transformation” → “Variable Selection” → “Regression Analysis”.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThen 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.



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”.

A screenshot of a computer

Description automatically generated

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

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated“Data Transformation” → “Data Manipulation” → “Select Column(s)”.

The results will appear on the output spreadsheet.

A screenshot of a computer

Description automatically generated

# 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”.

A screenshot of a computer

Description automatically generatedImport data into the input spreadsheet of the “FEATURE\_SELECTION\_TEST\_SET” tab from the output of the “NORMALISE\_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from SpreadSheet”.

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

“Data Transformation” → “Data Manipulation” → “Select Column(s)”.

Screens screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe 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”.

A screenshot of a computer

Description automatically generated

Use the J48 Method to train and fit the model:

“Analytics” → “Classification” → “J48”

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedand 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.

A screenshot of a computer

Description automatically generated

# 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)”.

A screenshot of a computer

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

To validate the model:

“Analytics” → “Existing Model Utilization”. Then choose Model “(from Tab: ) TRAIN\_MODEL (.fit)”.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe 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”.

A screenshot of a computer

Description automatically generated

Calculate the statistical metrics for the classification:

A screenshot of a statistics tool

Description automatically generatedA screenshot of a computer

Description automatically generated“Statistics” → “Model Metrics” → “Classification Metrics”.

The results will appear on the output spreadsheet.

Accuracy: 0.93

F1-Score = 0.9199

A screenshot of a computer

Description automatically generated

# 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”.

A screenshot of a computer

Description automatically generatedImport 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”

“Data Transformation” → “Data Manipulation” → “Select Columns”

Screens screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThen 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”.

A screenshot of a computer

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

Create the domain: “Statistics” → “Domain APD”

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedThe 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”.

A screenshot of a computer

Description automatically generated

Filter the data to exclude the column that corresponds to the “Diagnosis”

“Data Transformation” → “Data Manipulation” → “Select Columns”.

A screenshot of a computer

Description automatically generatedScreens screenshot of a computer

Description automatically generatedThen 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”.

A screenshot of a computer

Description automatically generated

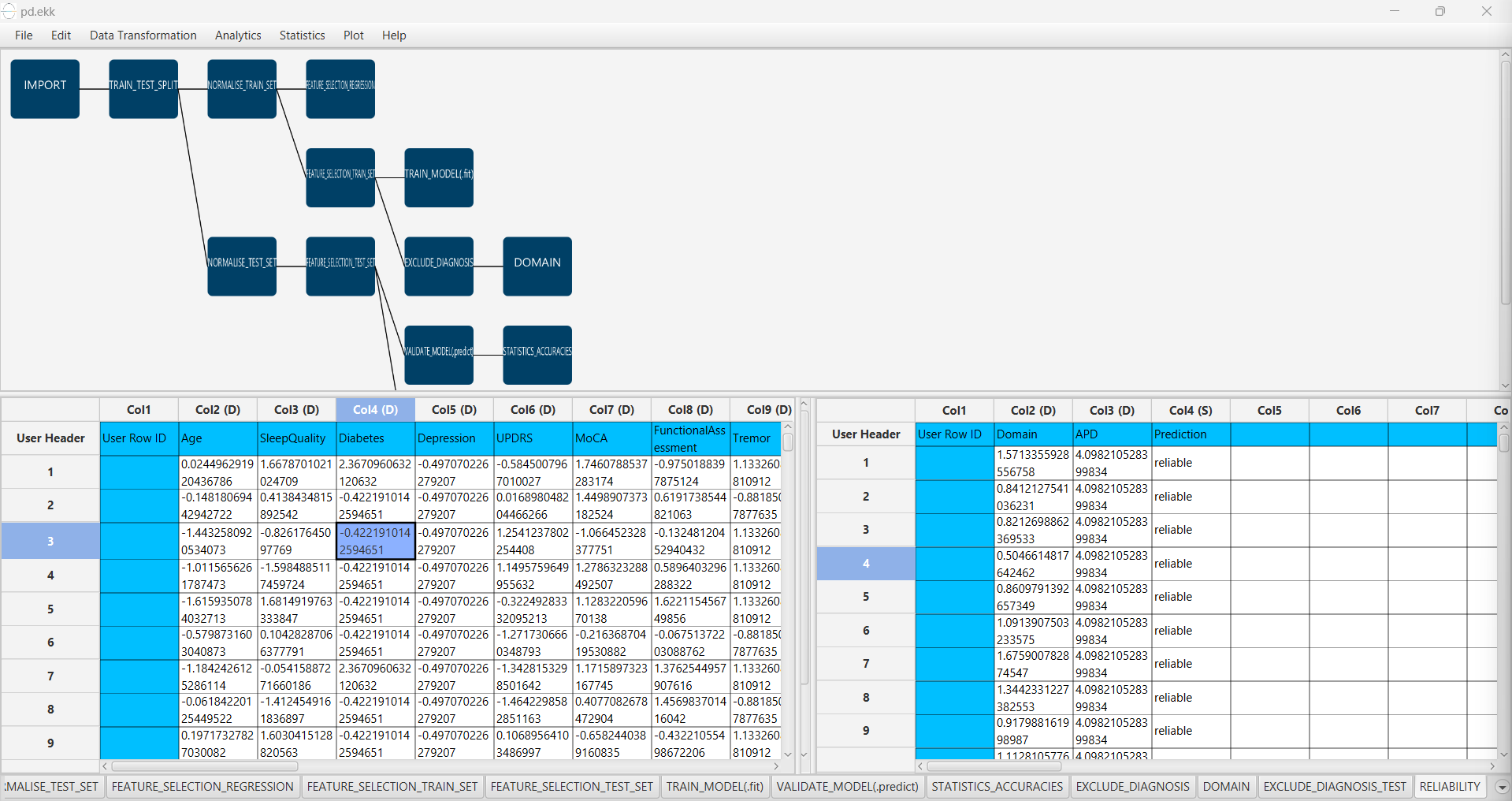
Check the Reliability: “Analytics” → “Existing Model Utilization”.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedThen 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:

