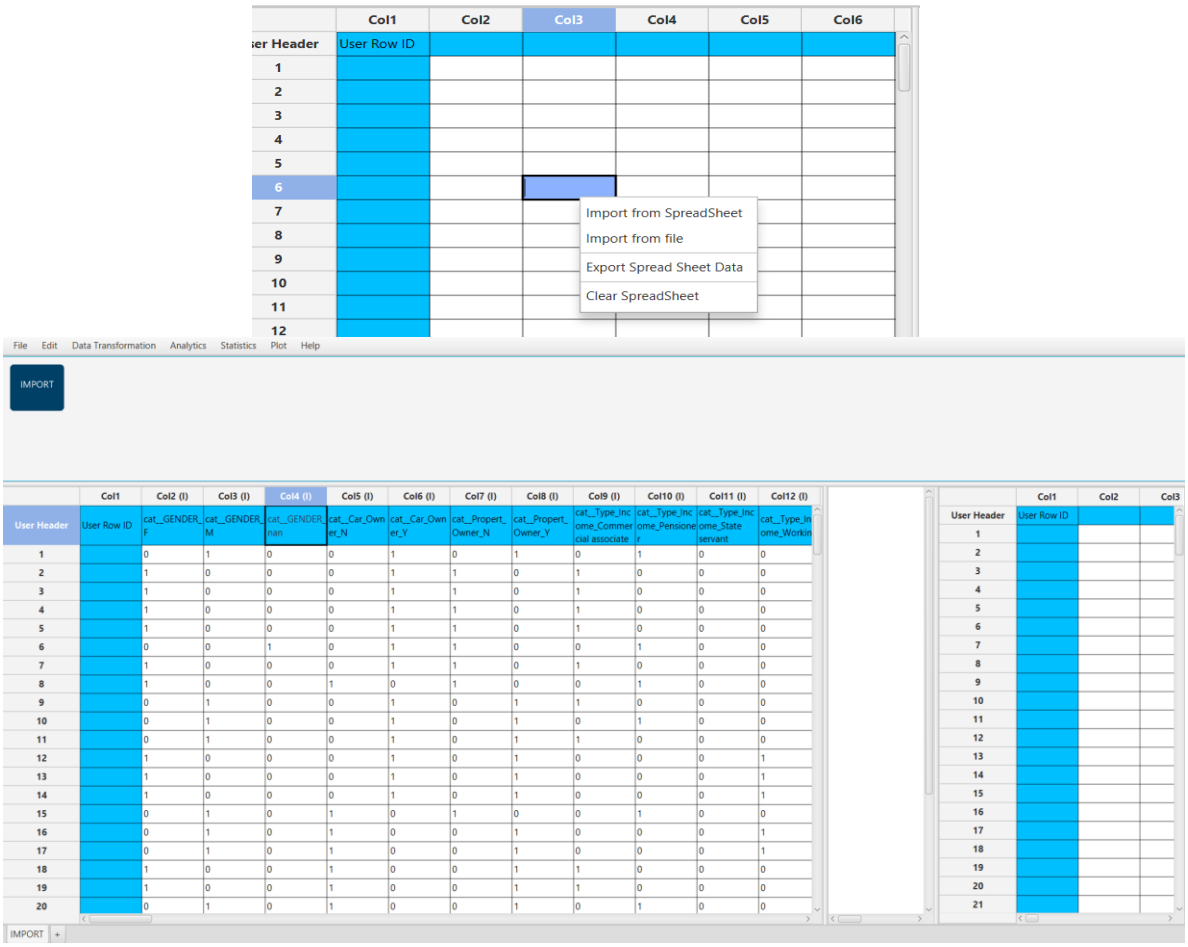
Credit Card (Binary Classification)

The goal of this study is to train a model in order to predict whether the application is Approved (0) or Rejected (1). The dataset used in this case study is found in https://www.kaggle.com/datasets/rohitudageri/credit-card-details?select=Credit_card_label.csv and has 20 features and 1458 labelled samples.

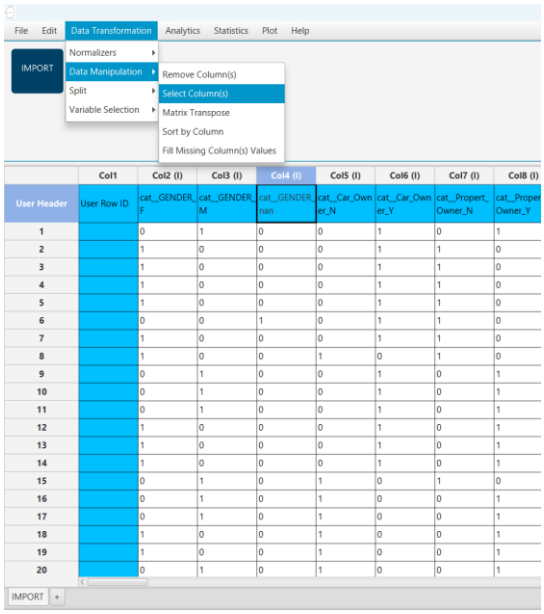
Step 1: Import Data from the file

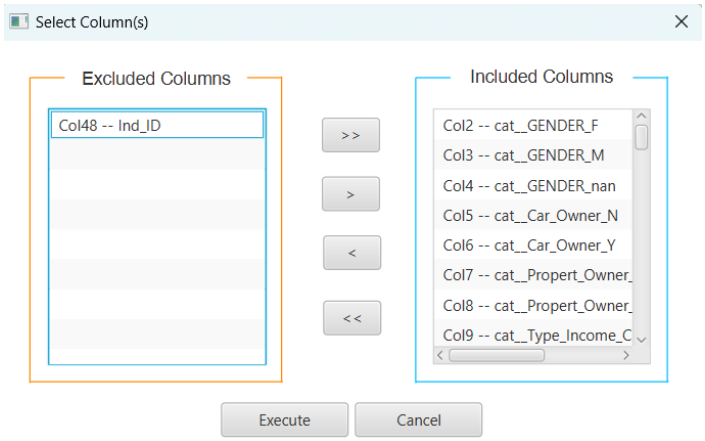
Right click on the input spreadsheet and choose the option "Import from file". Then navigate through your files to find the one with the credit card data.



Step 2: Manipulate Data

In order to use the data for training we have to exclude any columns that do not represent factor, like Ind_ID. We follow these steps to execute this:
Data Transformation → Data Manipulation → Select Column(s)
Select all columns except the one that corresponds to the Ind_ID.





The data without the Ind_ID column will appear in the output spreadsheet.

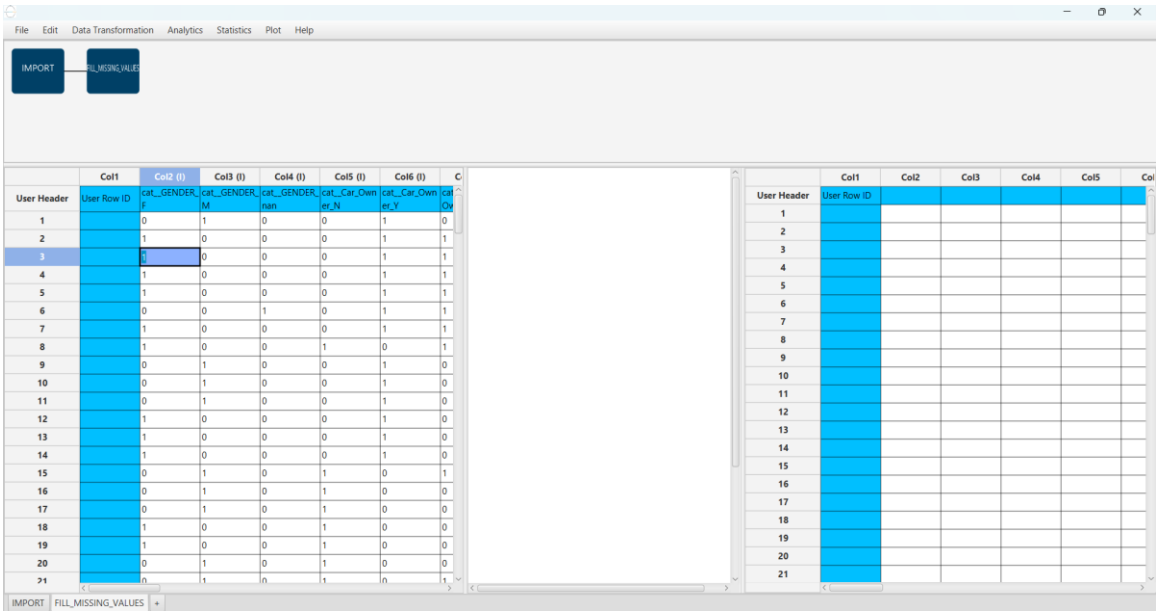
Step 3: Fill missing values

There are empty values in the Dataset. Specifically, we show below how many missing values there are for each feature:

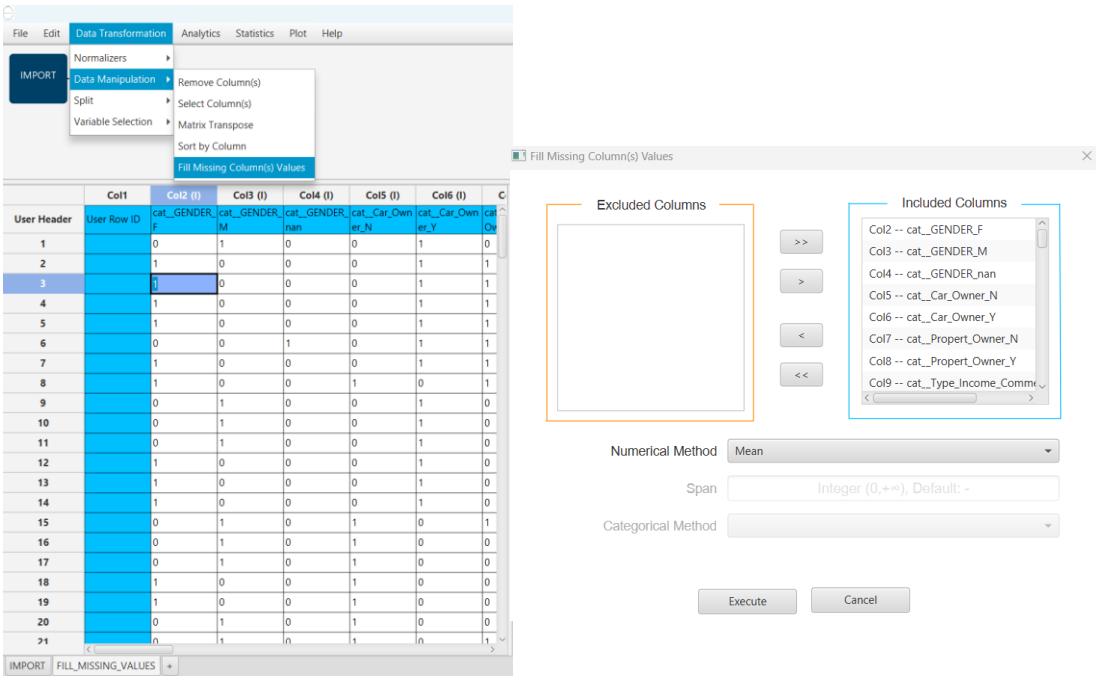
```
Empty data:
Ind_ID              0
GENDER              7
Car_Owner           0
Propert_Owner       0
CHILDREN            0
Annual_income      23
Type_Income         0
EDUCATION           0
Marital_status      0
Housing_type        0
Birthday_count     22
Employed_days       0
Mobile_phone        0
Work_Phone          0
Phone               0
EMAIL_ID            0
Type_Occupation   488
Family_Members      0
dtype: int64
```

Create a new action by pressing the + button on the bottom of the page with the name FILL_MISSING VALUES which will be used to fill the missing values.

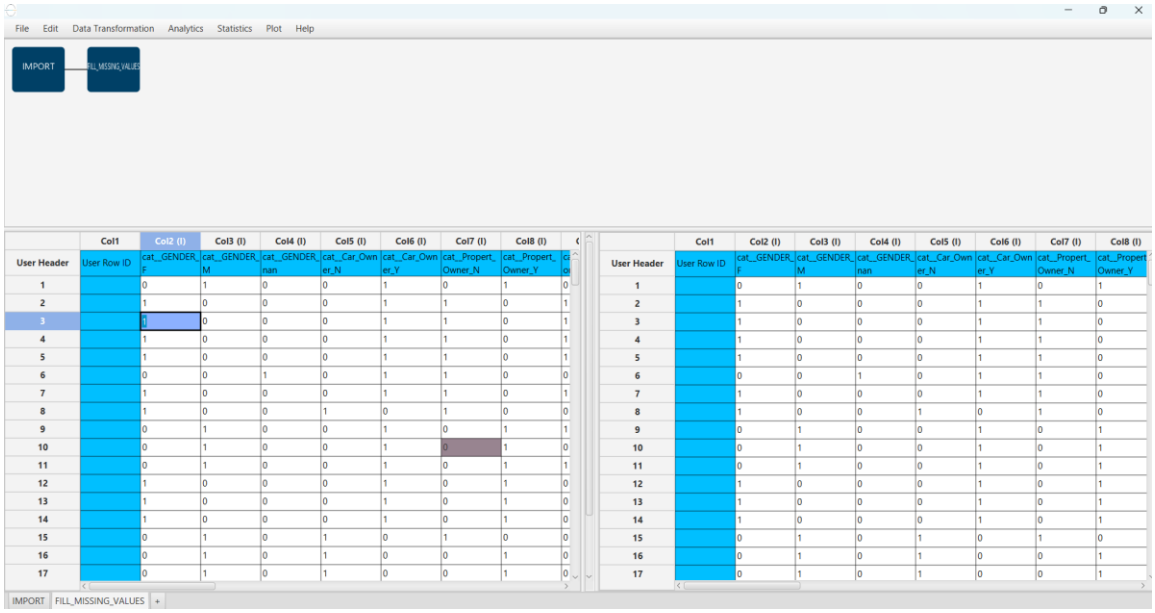
Import Data into the input spreadsheet of the FILL_MISSING_VALUES action from the output of the IMPORT action by right-clicking on the input spreadsheet and then choosing Import from SpreadSheet.



Handle missing columns values:
Data Transformation → Data Manipulation → Fill missing column(s) Values:
Then choose the Mean as the Numerical Method



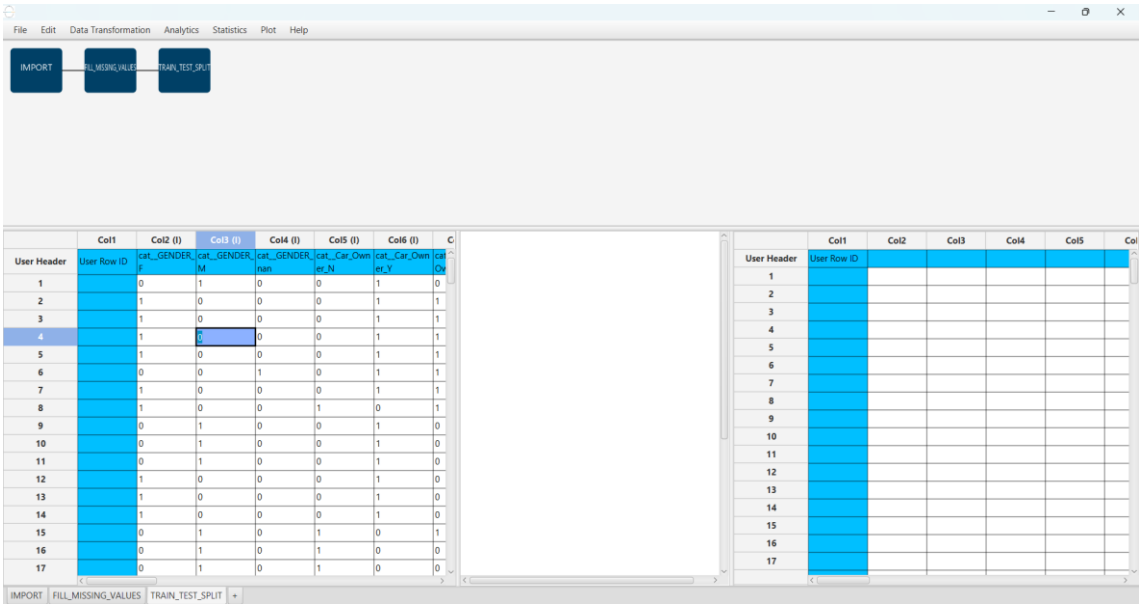
The results will appear on the output spreadsheet.



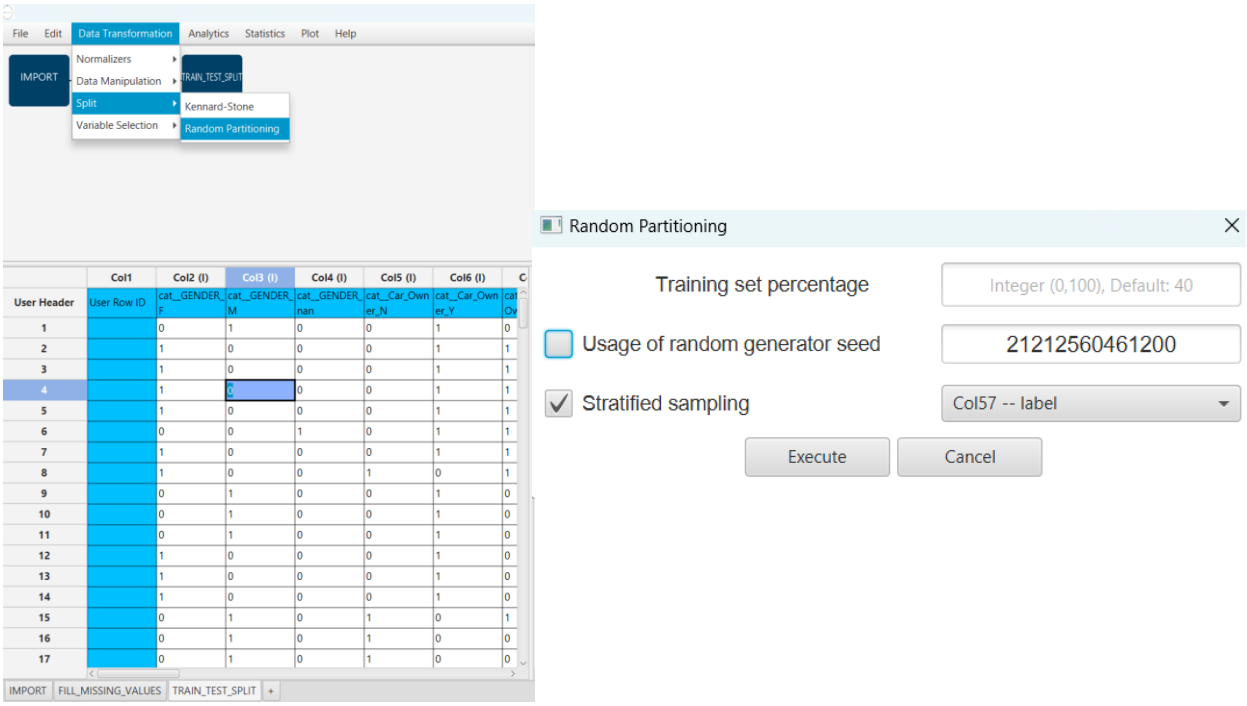
Step 4: Split Data

Create a new action 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.

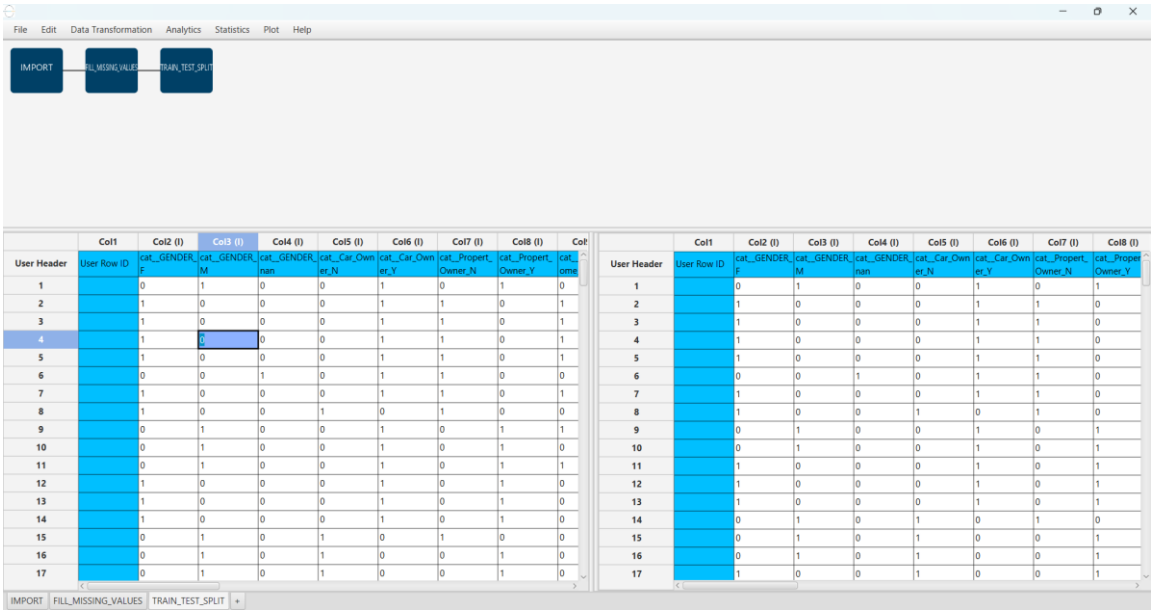
Import Data into the input spreadsheet of the TRAIN_TEST_SPLIT action from the output of the FILL_MISSING_VALUES action by right-clicking on the input spreadsheet and then choosing Import from SpreadSheet.



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



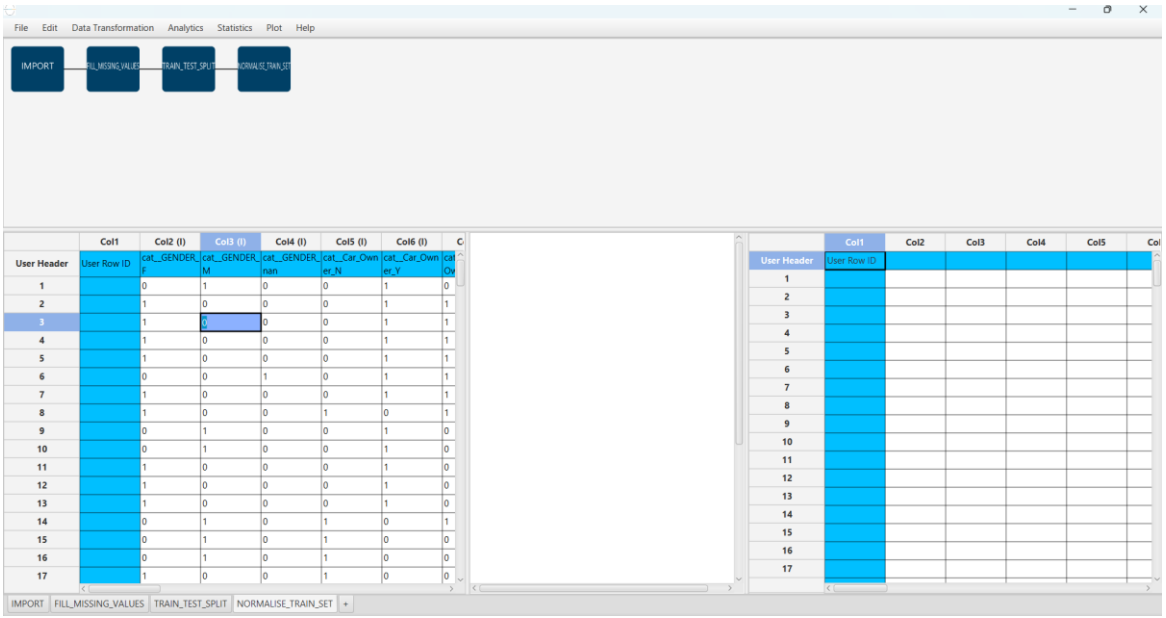
The results will appear on the output spreadsheet.



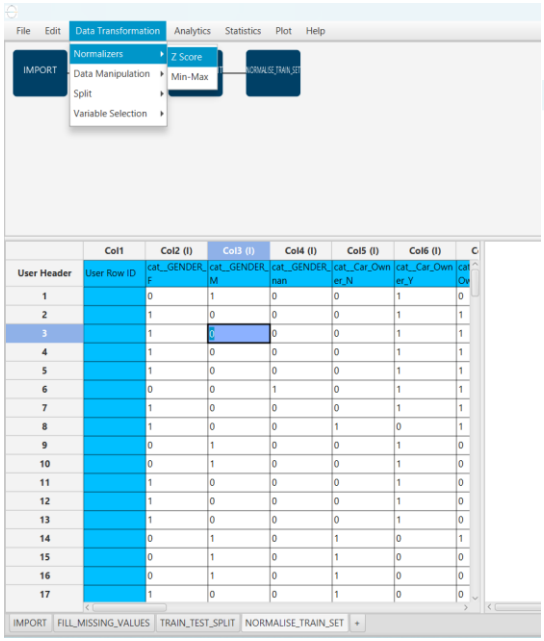
Step 5: Normalize the Training Set

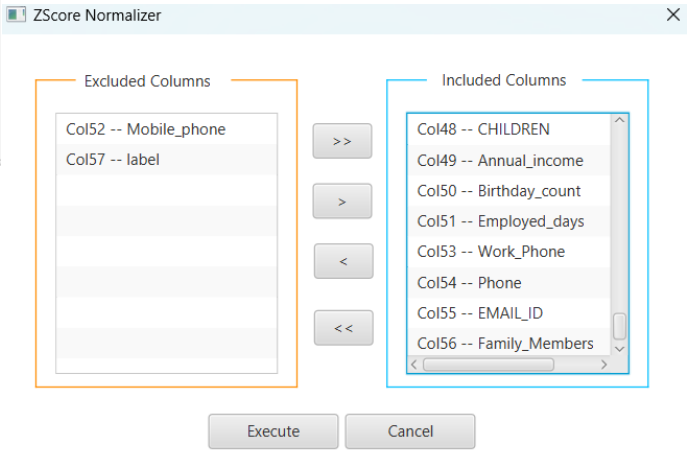
Create a new action by pressing the + button on the bottom of the page with the name NORMALISE_TRAIN_SET.

Import Data into the input spreadsheet of the NORMALISE_TRAIN_SET action the train set from the output of the TRAIN_TEST_SPLIT action 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

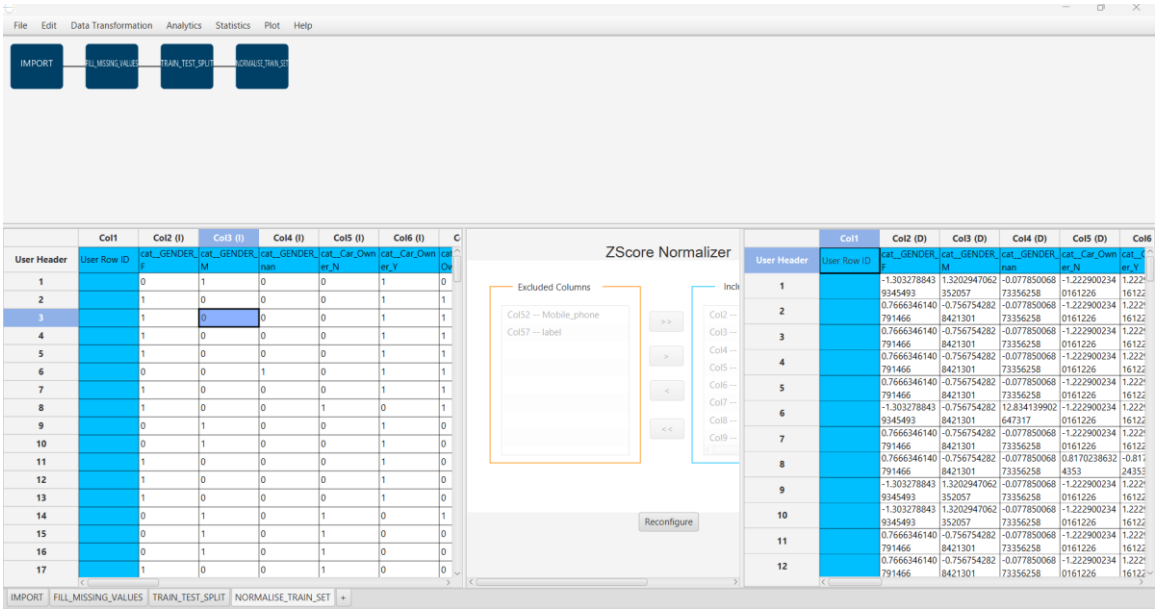


Normalize the Data using Z-score: Data Transformation → Normalize → Z-Score
Then select all columns excluding Mobile_phone and Label and click Execute.





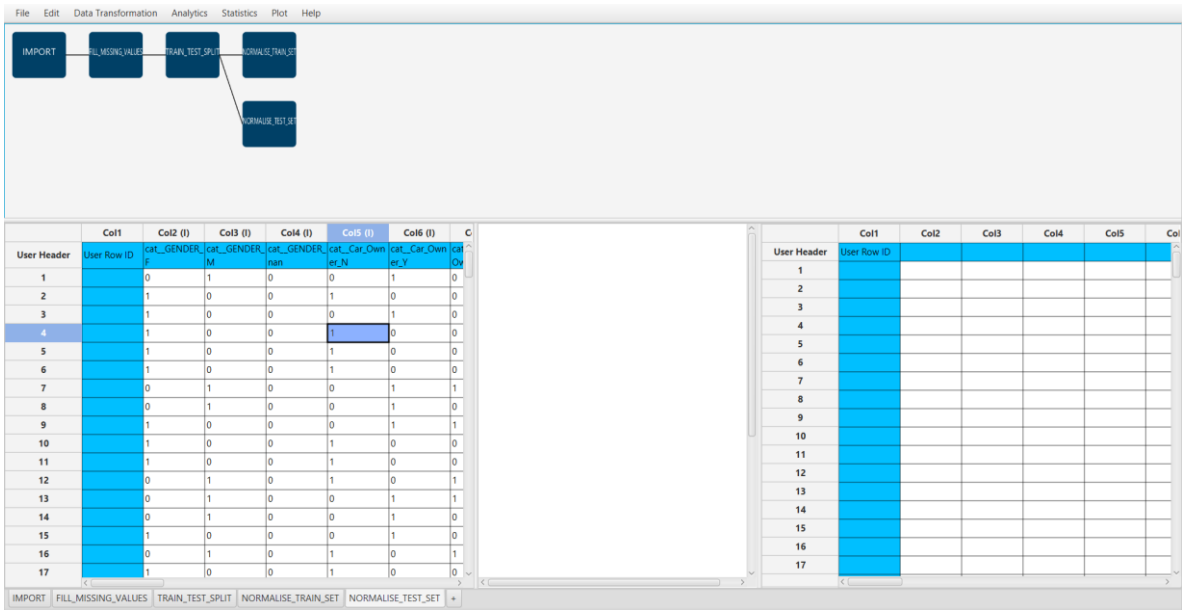
The results will appear on the output spreadsheet.



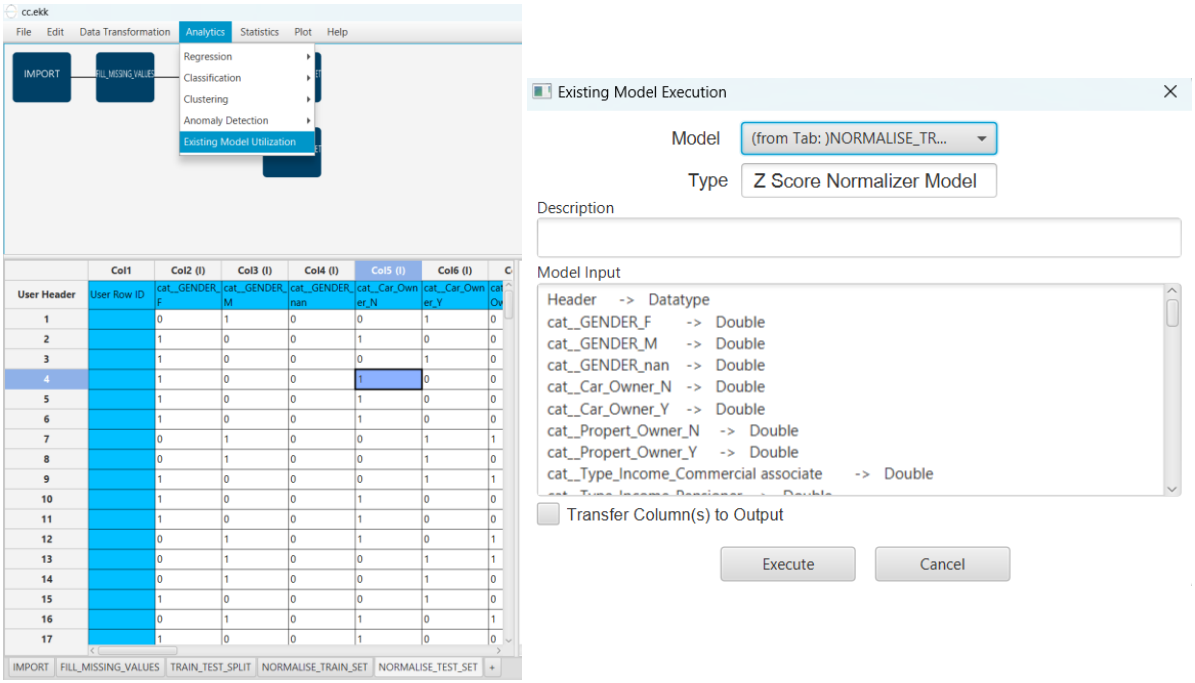
Step 6: Normalize the Test Set

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

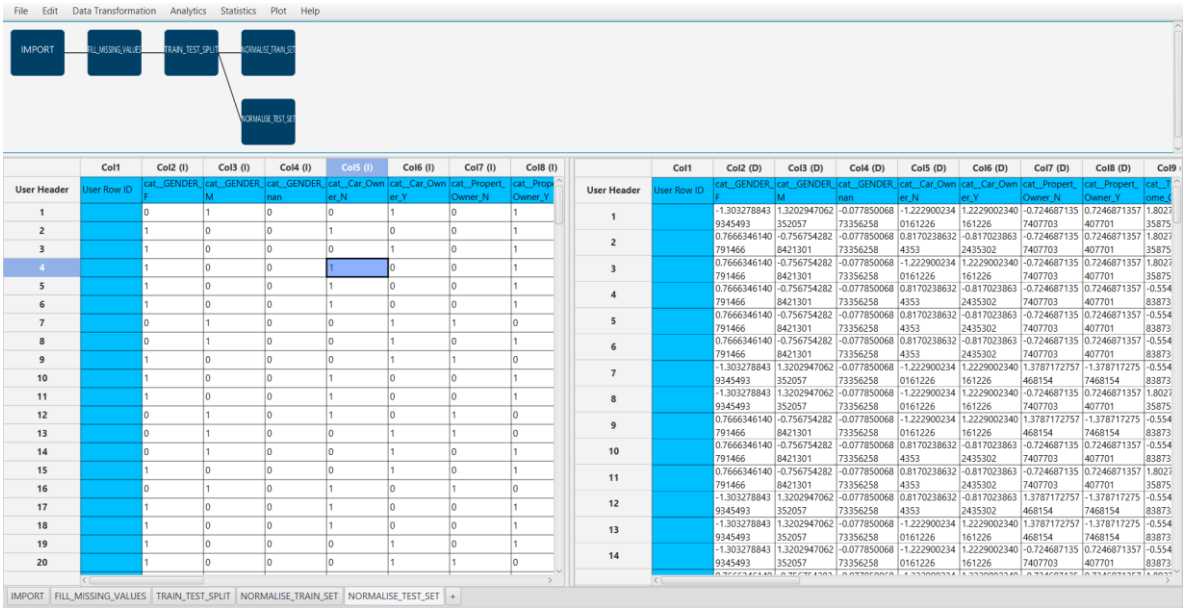
Import Data into the input spreadsheet of the NORMALISE_TEST_SET action the test set from the output of the TRAIN_TEST_SPLIT action by right-clicking on the input spreadsheet and then choosing Import from SpreadSheet. From the available Select input tab options choose TRAIN_TEST_SPLIT: Test Set



Normalize the test set using the existing normalizer of the training set:
Analytics → Existing Model Utilization → Model: NORMALIZE_TRAIN_SET



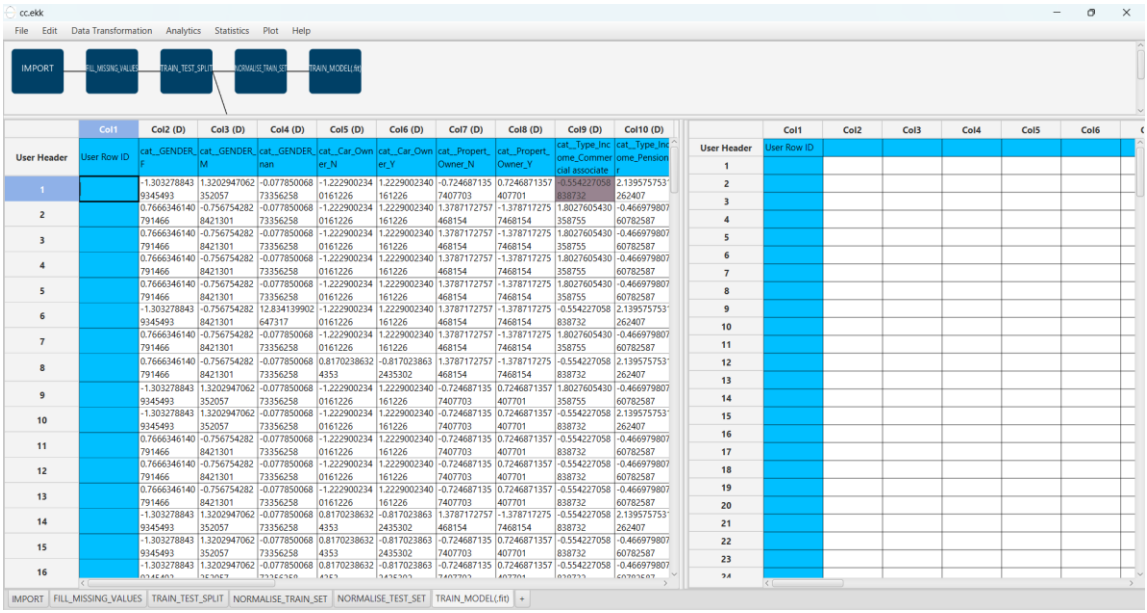
The results will appear on the output spreadsheet.



Step 7: Train the model

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

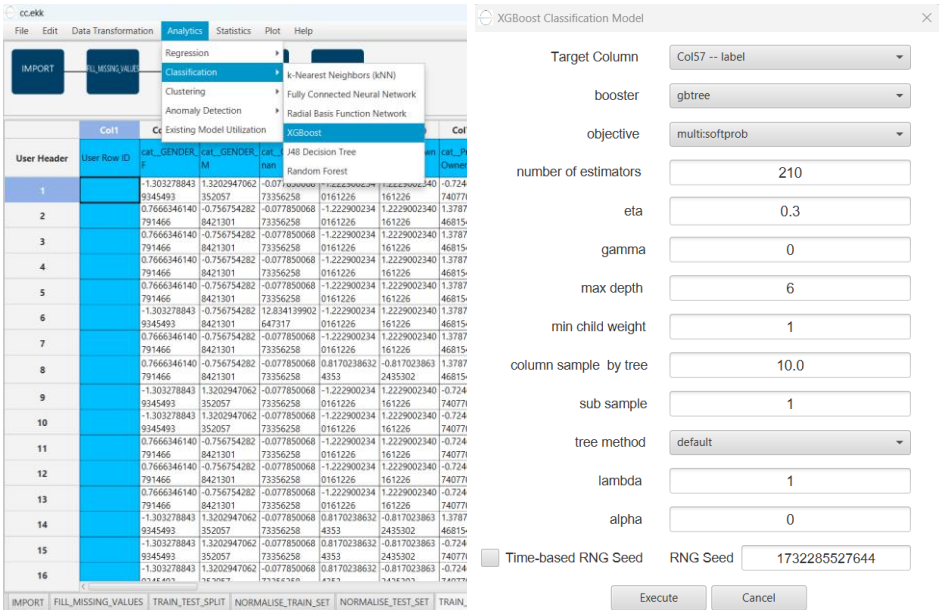
Import data into the input spreadsheet of the "TRAIN_MODEL(.fit)" tab from the output of the "NORMALISE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



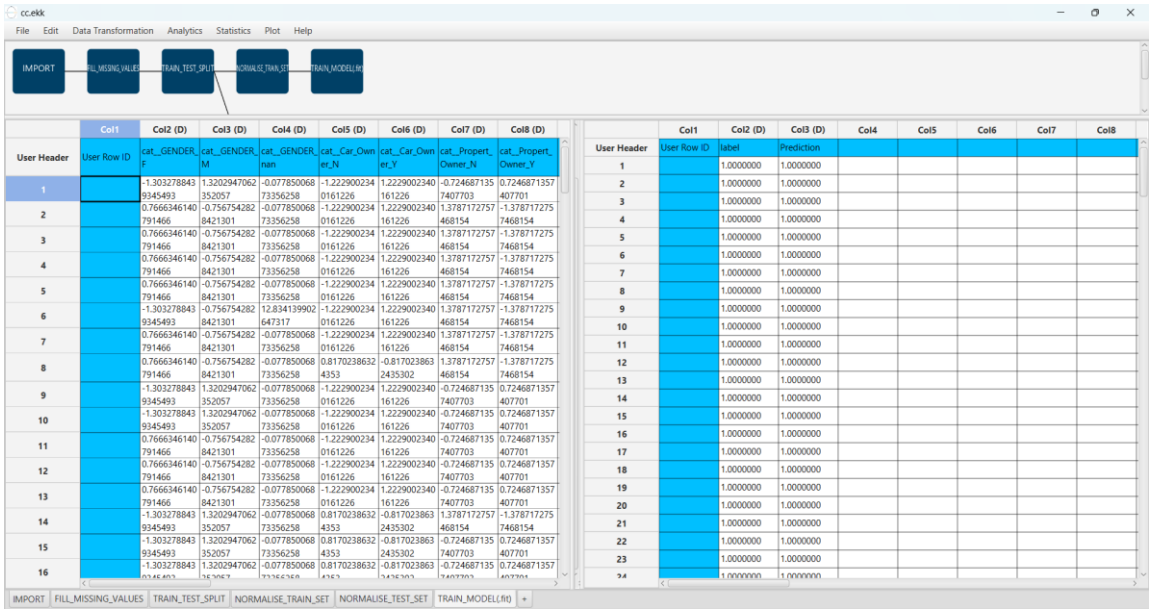
Use the XGBoost Method to train and fit the model:

"Analytics" → "Classification" → "XGBoost"

and set the "number of estimators" as 210, the "column sample by tree" as 10, the "Target Column" as the column corresponding to "Label" and use the following "RNG Seed": 1732285527644.



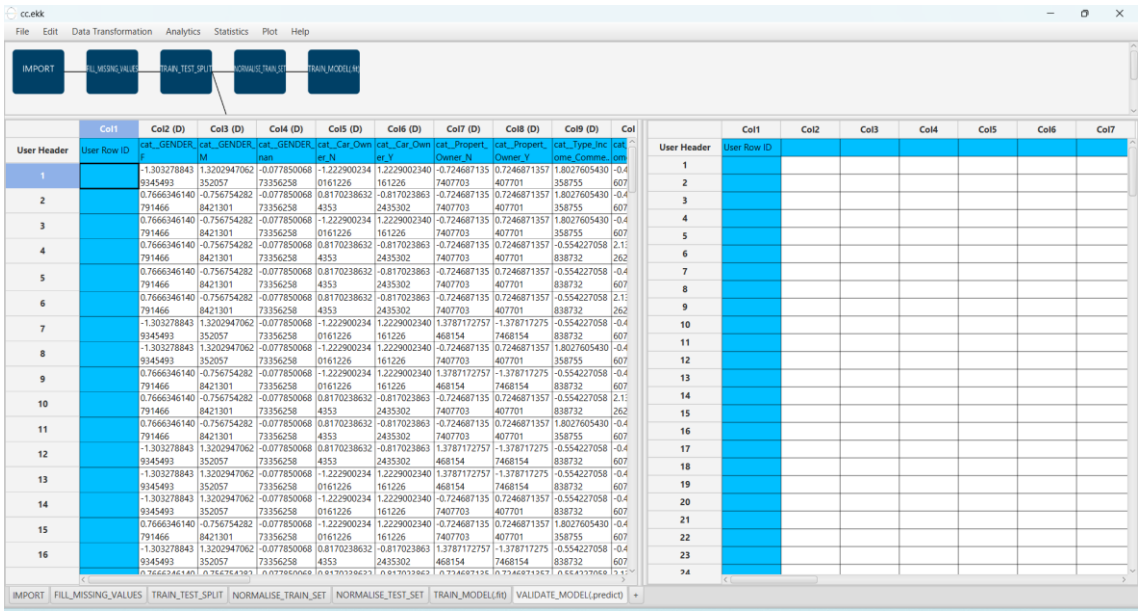
The predictions will appear on the output spreadsheet.



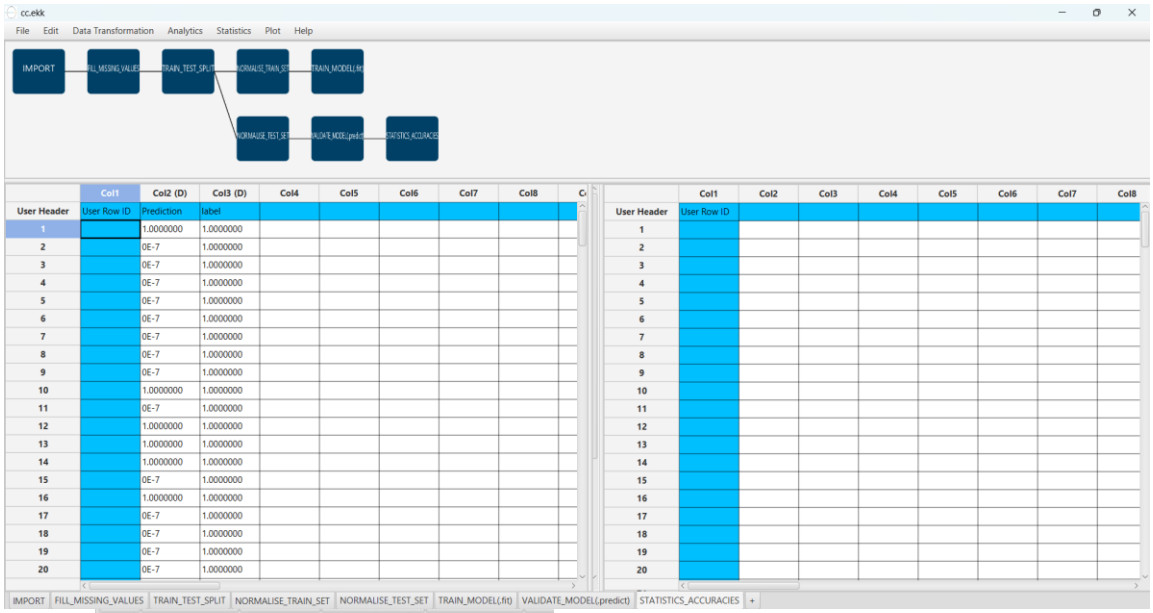
Step 8: Validate the model

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

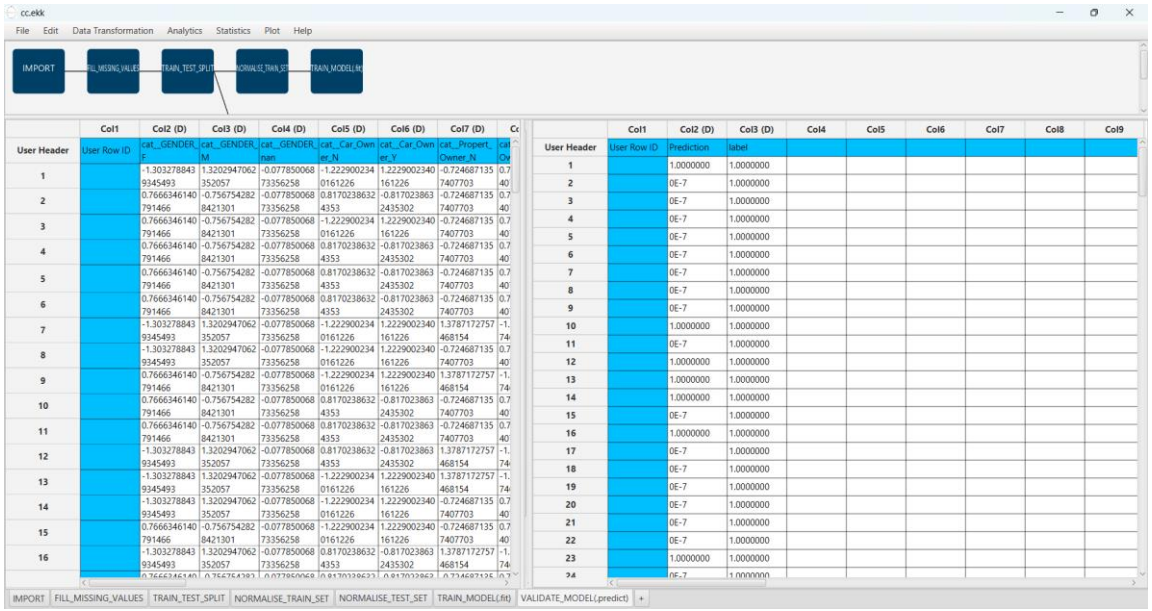
Import data into the input spreadsheet of the "VALIDATE_MODEL(.predict)" tab from the output of the "NORMALISE _TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



To validate the model: "Analytics" → "Existing Model Utilization". Then choose Model "(from Tab:) TRAIN_MODEL (.fit)". and transfer the "label" column to the output.



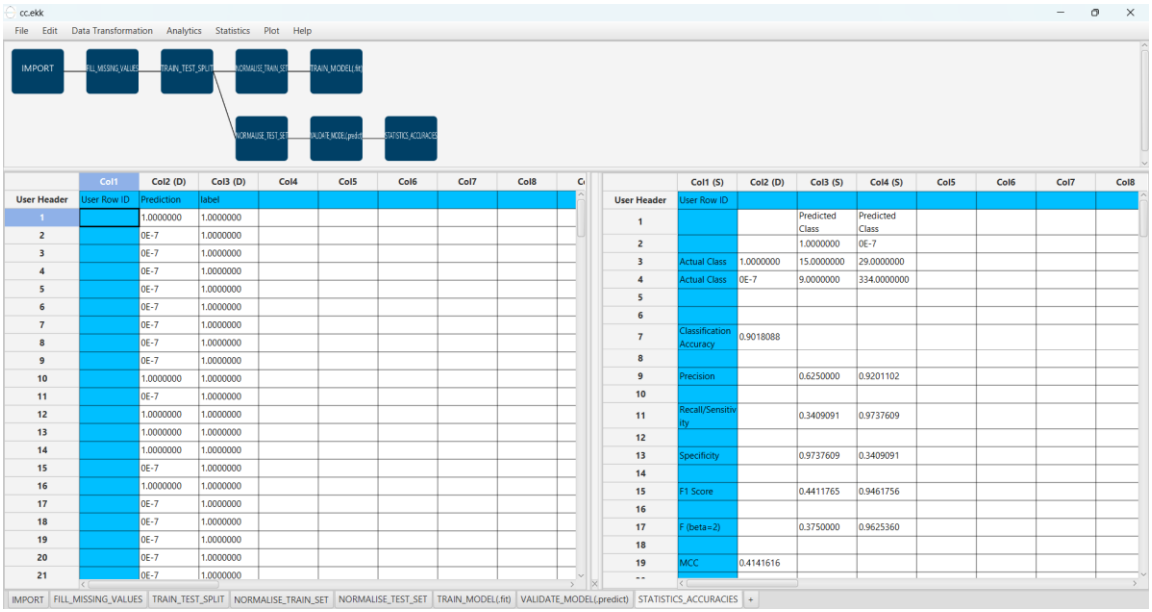
The predictions will appear on the output spreadsheet.



Step 9: Statistics calculation

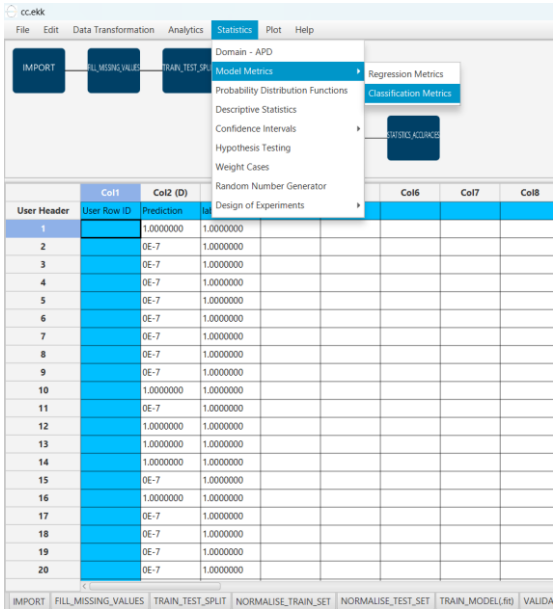
Create a new tab by pressing the "+" button on the bottom of the page with the name "STATISTICS_ACCURACIES".

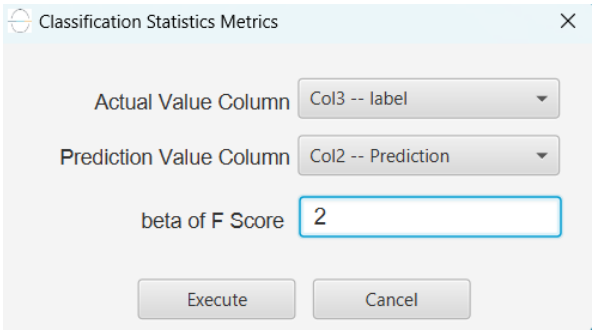
Import data into the input spreadsheet of the "STATISTICS_ACCURACIES" tab from the output of the "VALIDATE_MODEL(.predict)" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



Calculate the statistical metrics for the classification:

"Statistics" → "Model Metrics" → "Classification Metrics".





The results will appear on the output spreadsheet.

Accuracy: 0.902

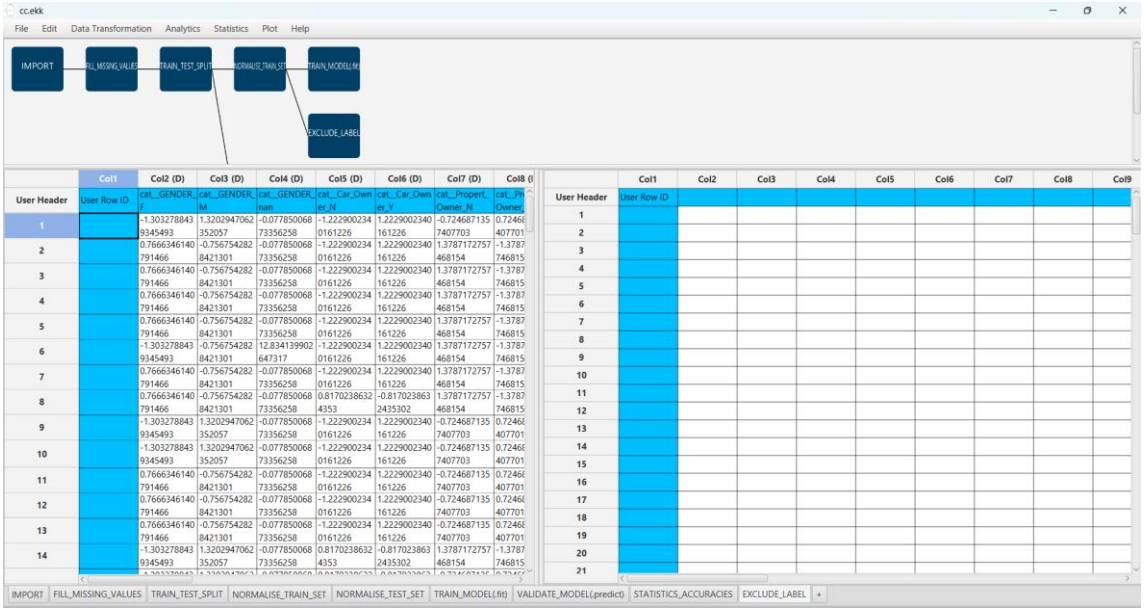
F1-Score = 0.694

Step 10: Reliability check of each record of the test set

Step 10.a: Create the domain

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

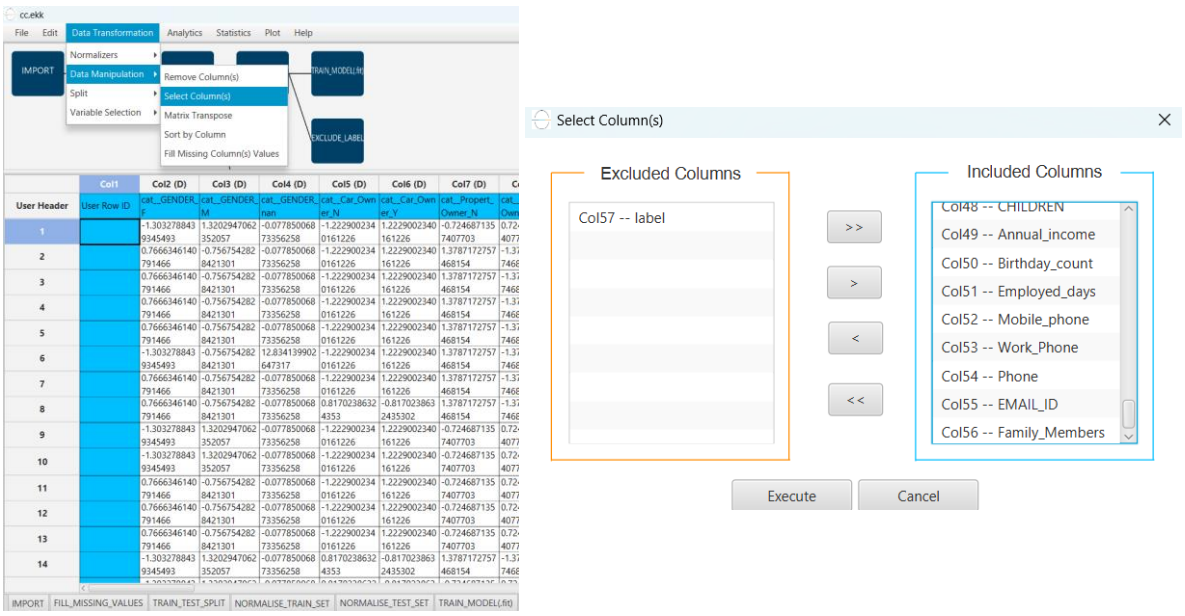
Import data into the input spreadsheet of the "EXCLUDE_LABEL" tab from the output of the "NORMALISE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



Manipulate the data to exclude the column that corresponds to the "label"

"Data Transformation" → "Data Manipulation" → "Select Columns"

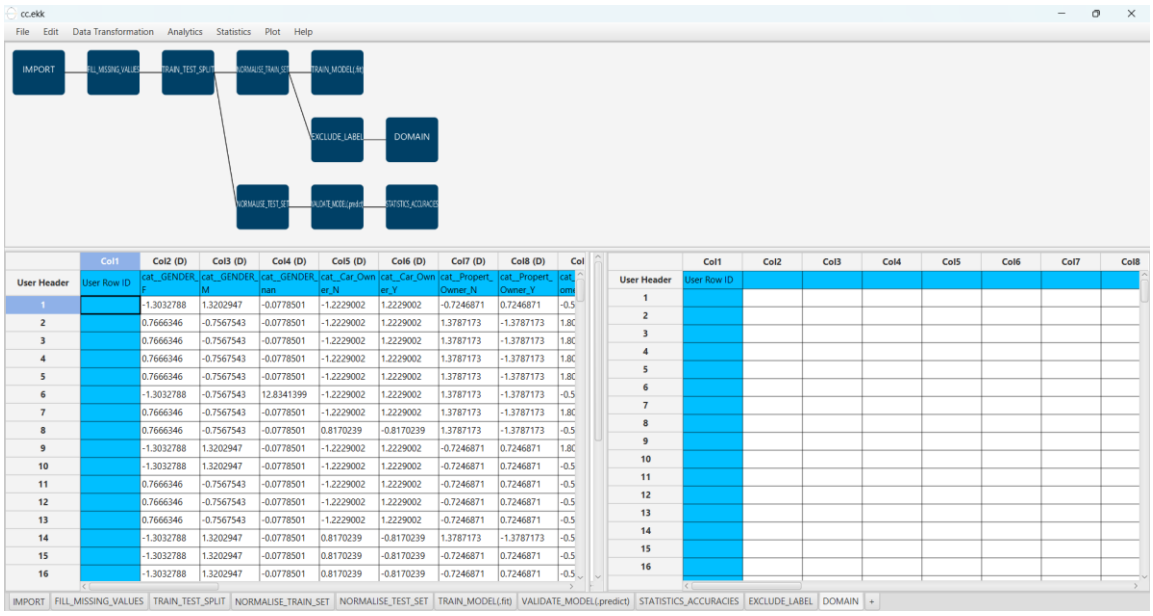
Then select all the columns except the "label".



The results will appear on the output spreadsheet.

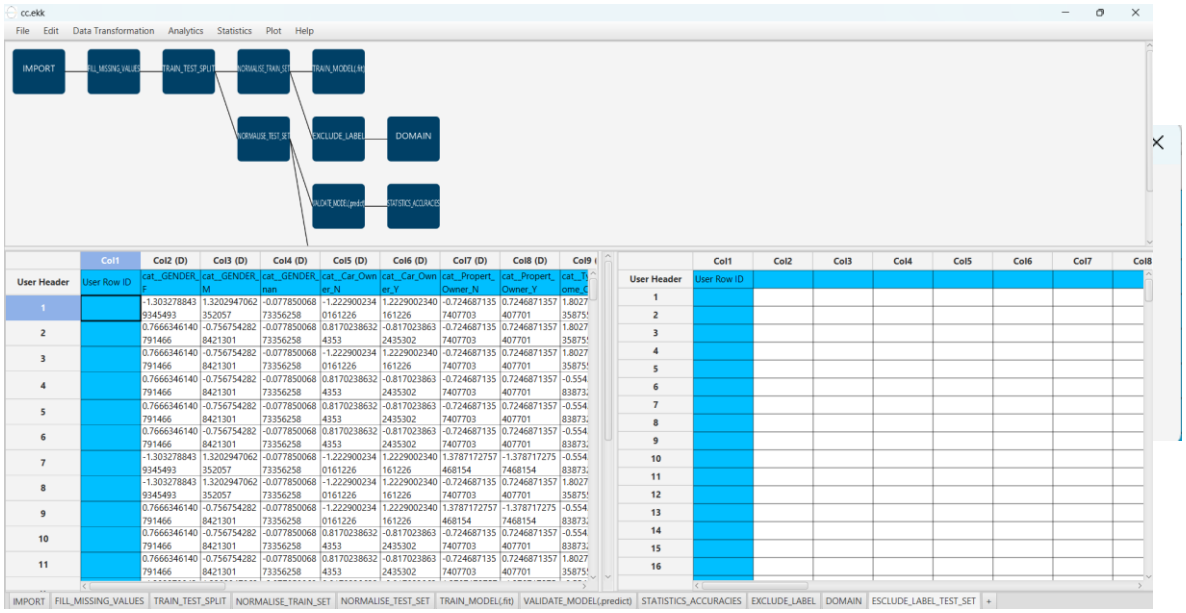
Create a new tab by pressing the "+" button on the bottom of the page with the name "DOMAIN".

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

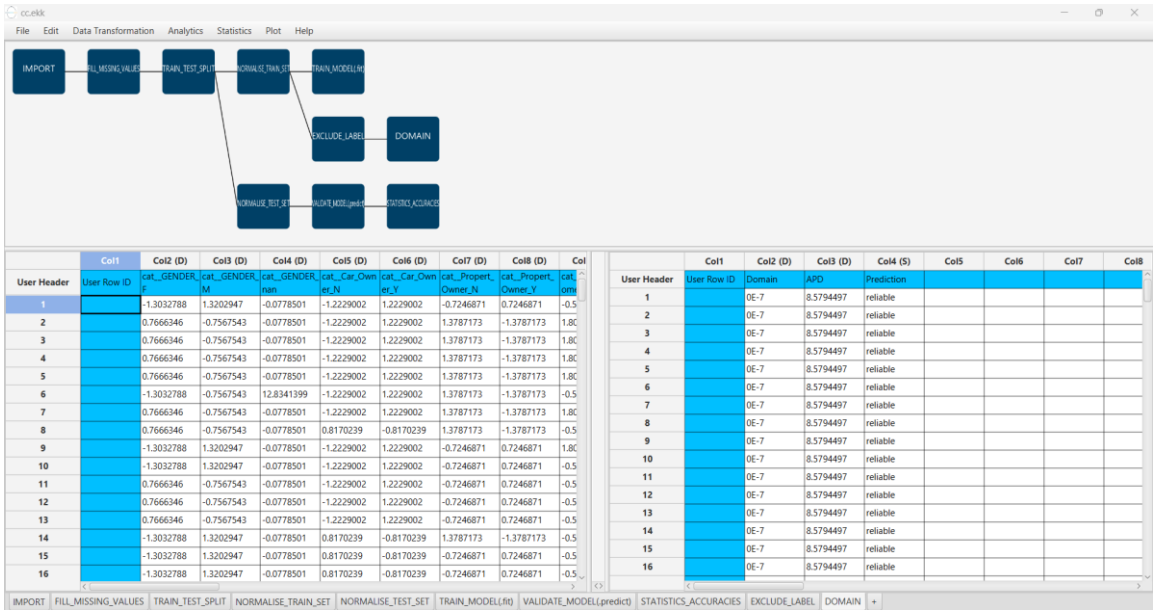


Create the domain:

"Statistics" → "Domain APD"



The results will appear on the output spreadsheet.



Step 10.b: Check the test set reliability

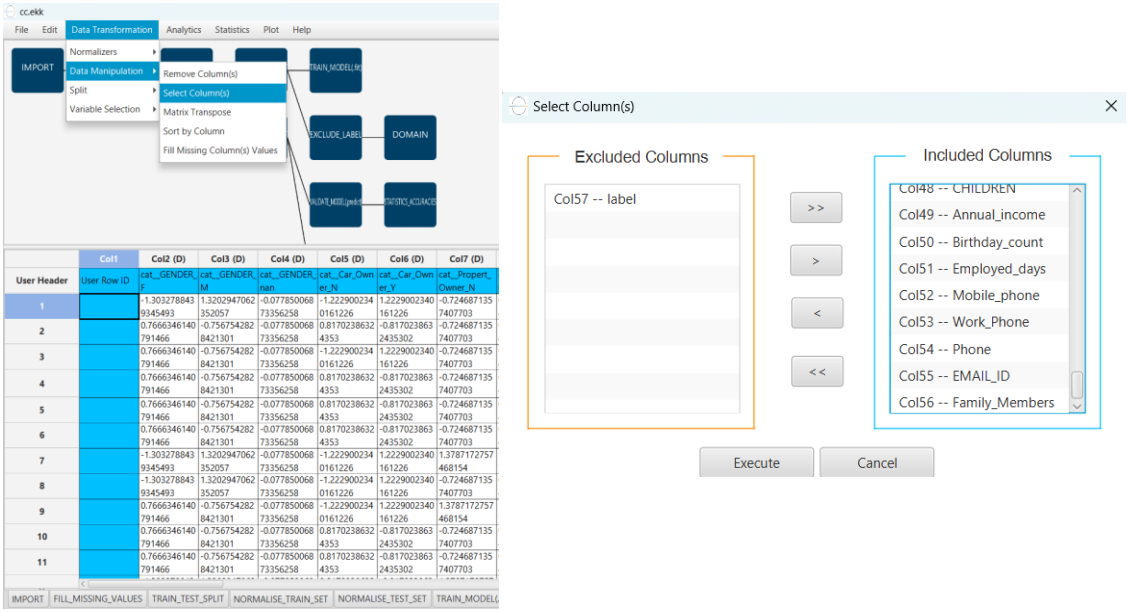
Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_LABEL_TEST_SET".

Import data into the input spreadsheet of the "EXCLUDE_LABEL_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".

Filter the data to exclude the column that corresponds to the "label"

"Data Transformation" → "Data Manipulation" → "Select Columns".

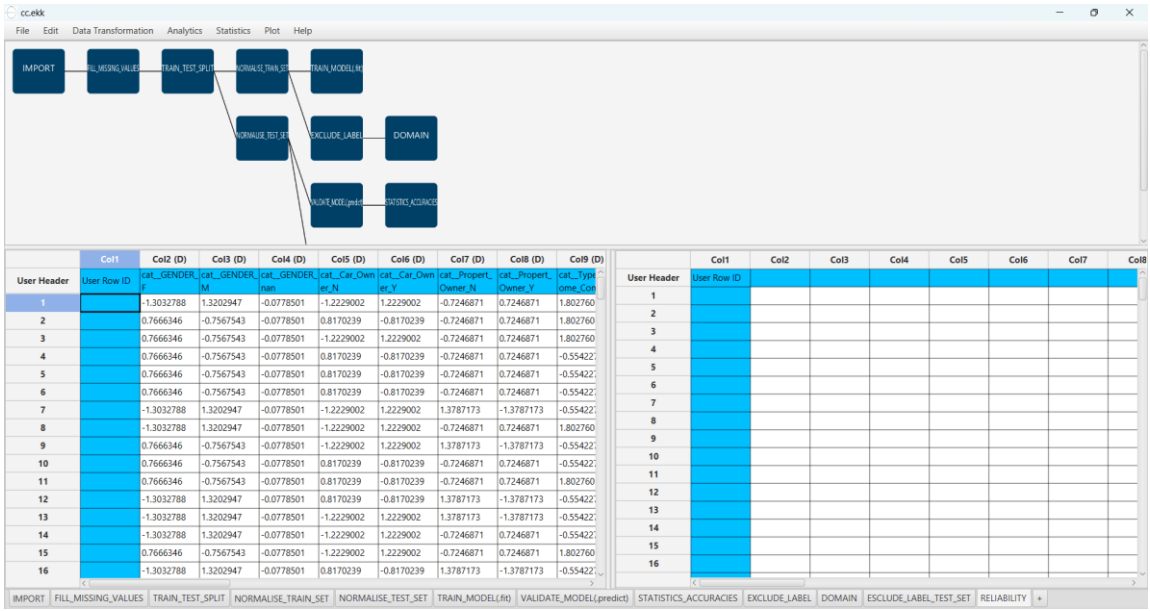
Then select all the columns except "label".



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".

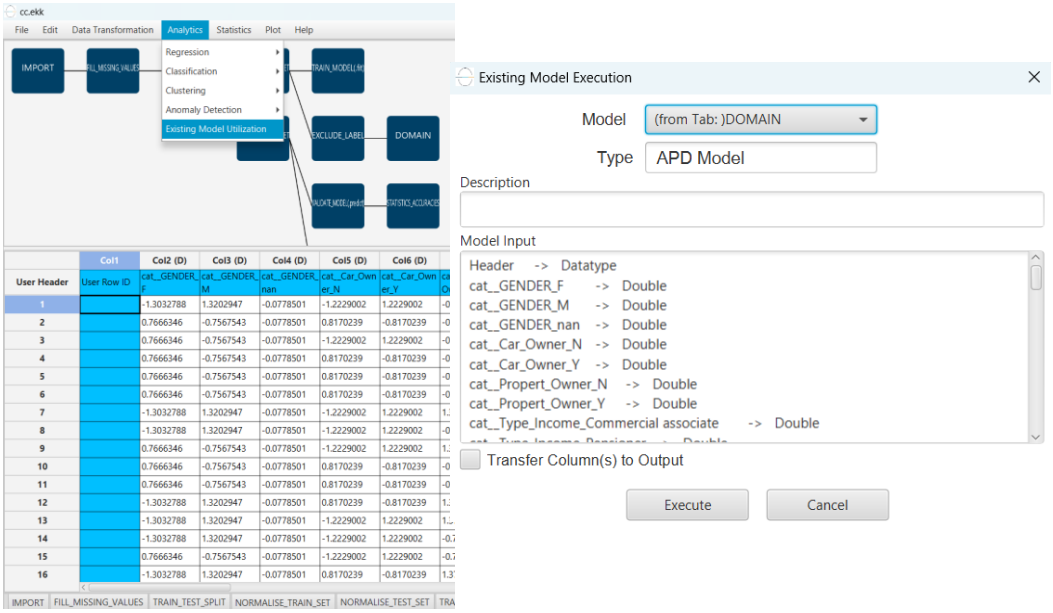
Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "EXCLUDE_LABEL_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



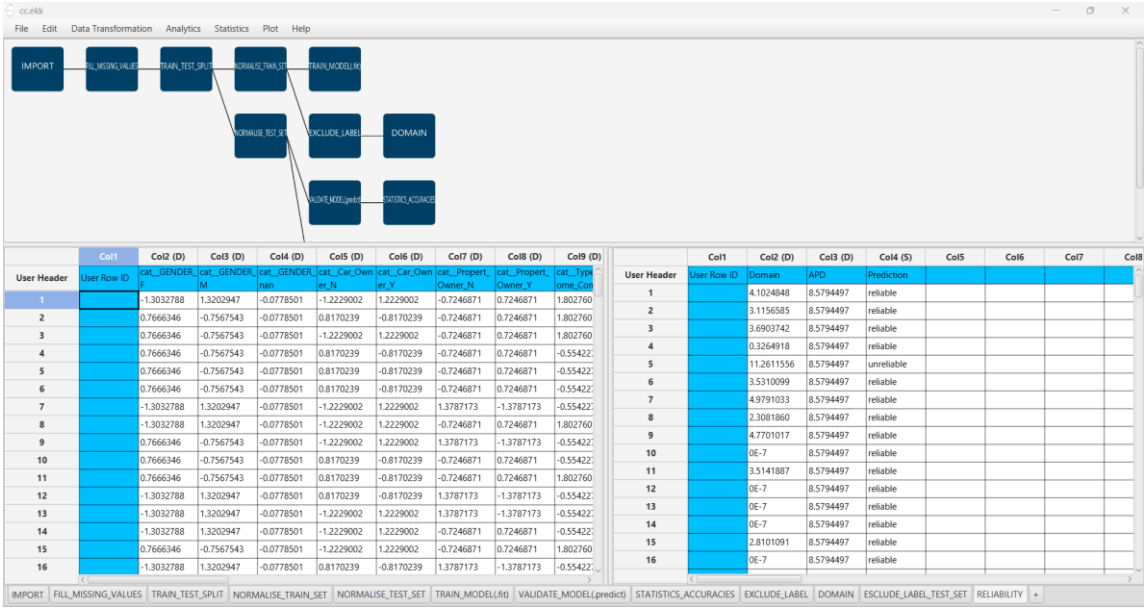
Check the Reliability:

"Analytics" → "Existing Model Utilization".

Then select as Model "(from Tab:) DOMAIN".



The results will appear on the output spreadsheet.



There are four unreliable samples in the test set.

Final Isalos Workflow

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