Salary Prediction (Regression)

The goal of this study is to train a model in order to predict a person's salary . The dataset used in this case study is found in https://www.kaggle.com/datasets/mrsimple07/salary-prediction-data/data and has 7 features and 1000 samples. It contains simulated data reflecting various factors influencing salary levels such as education, experience, location, job title, age, and gender.

The dataset contains no missing values and includes several categorical features. Categorical features contain multiple levels and the data was transformed to corresponding numeric codes, as detailed below:

Gender:

- Female (0)
- Male (1)

Education:

- High School (0)
- Bachelor (1)
- Master (2)
- PhD (3)

Location:

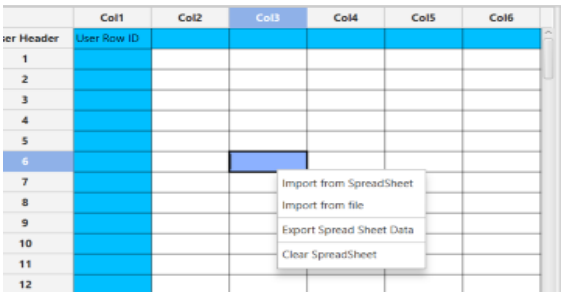
- Rural (0)
- Suburban (1)
- Urban (2)

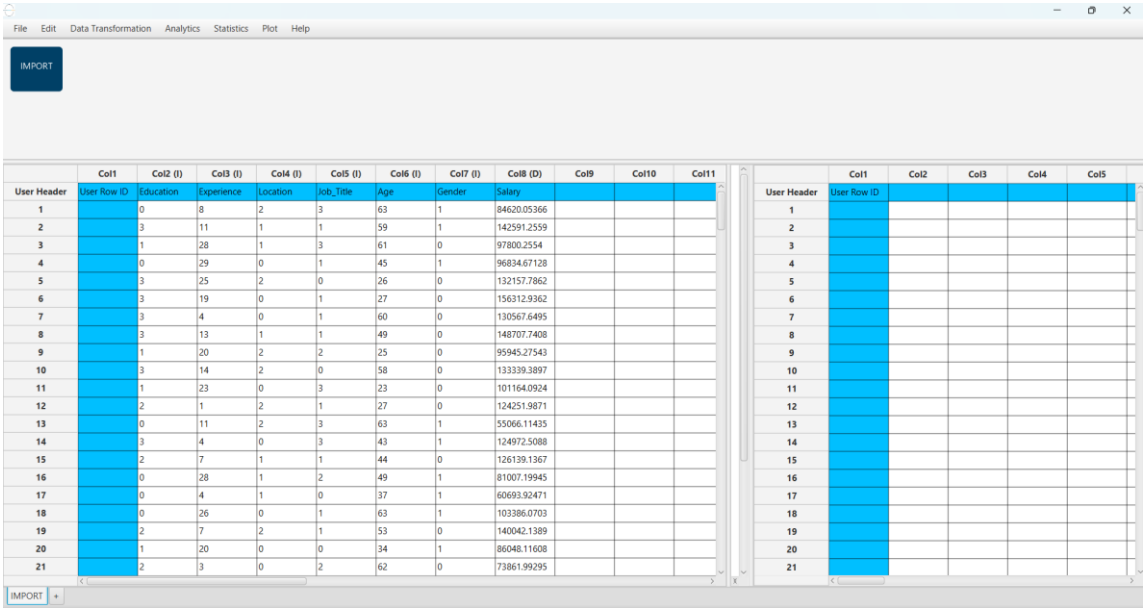
Job_Title:

- Analyst (0)
- Director (1)
- Engineer (2)
- Manager (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 salary data.

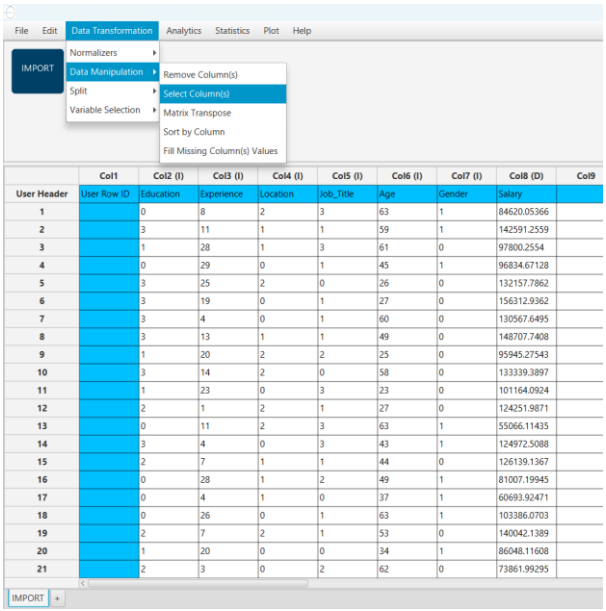


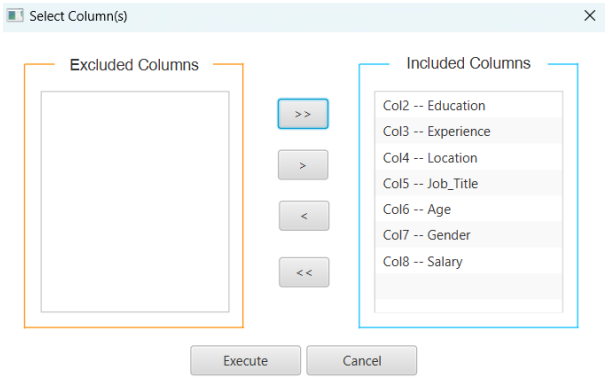


Step 2: Manipulate data

In order to use the data for training we have to exclude any columns that do not contain features. In our dataset there are no such columns. Therefore, we will include all columns in the training. We follow these steps to execute this:

- On the menu click on "Data Transformation" → "Data Manipulation" → "Select Column(s)"
- Select all columns.



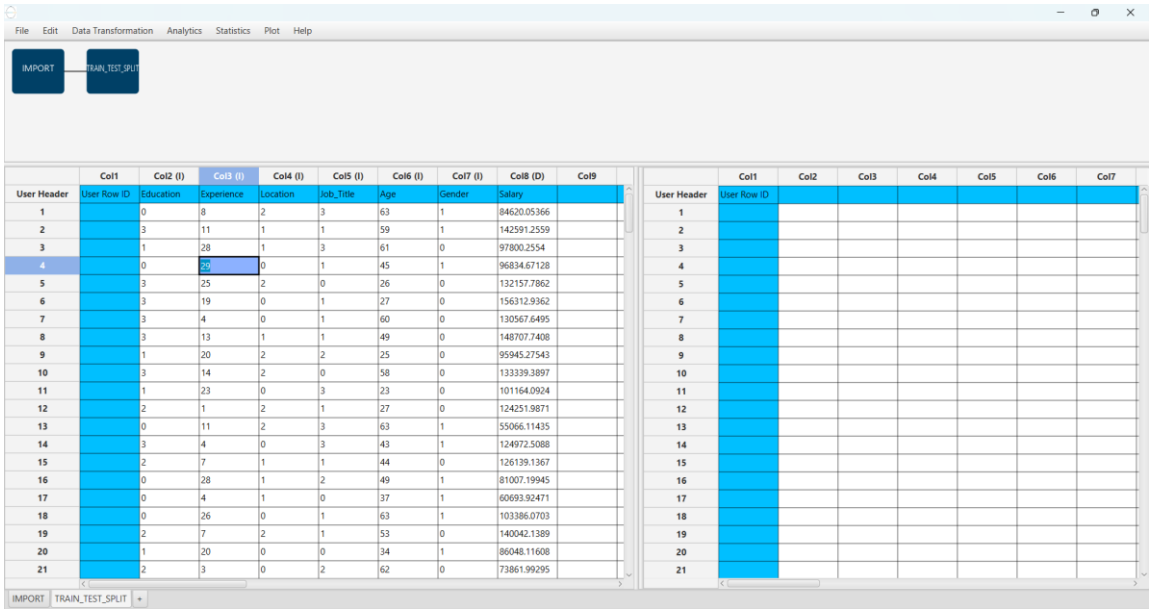


The data will appear in the output spreadsheet.

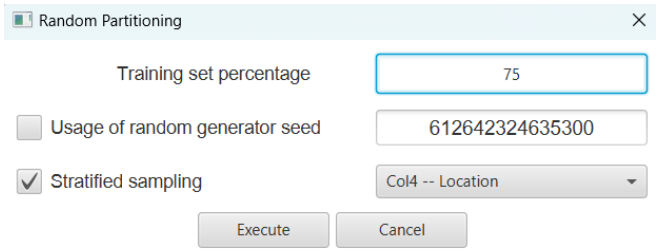
Step 3: Split data

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.

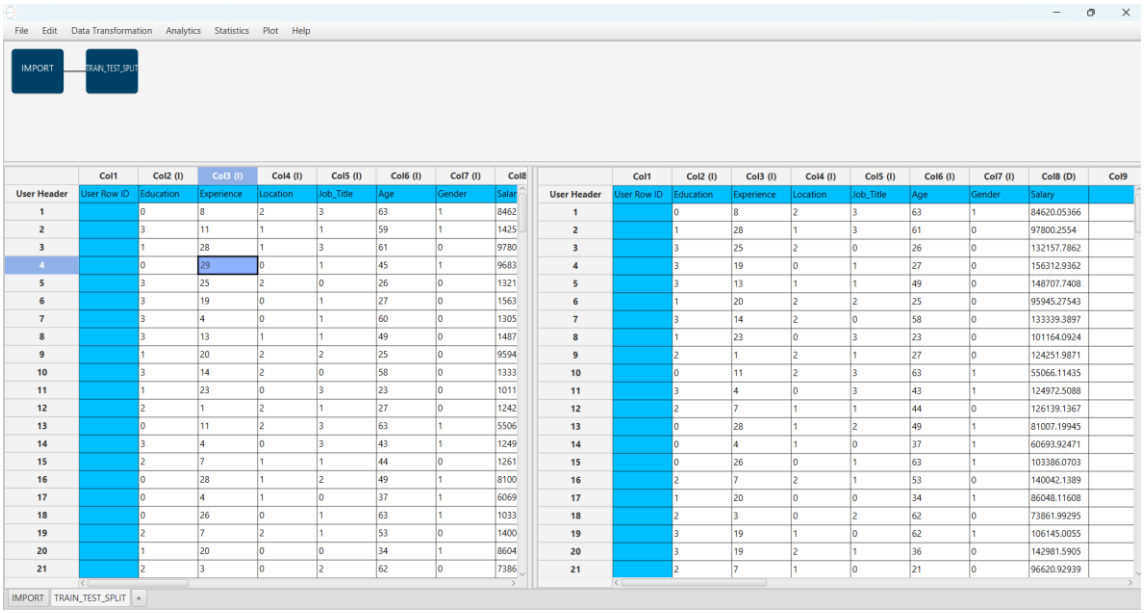
Import data into the input spreadsheet of the "TRAIN_TEST_SPLIT" tab from the output of the "IMPORT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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:



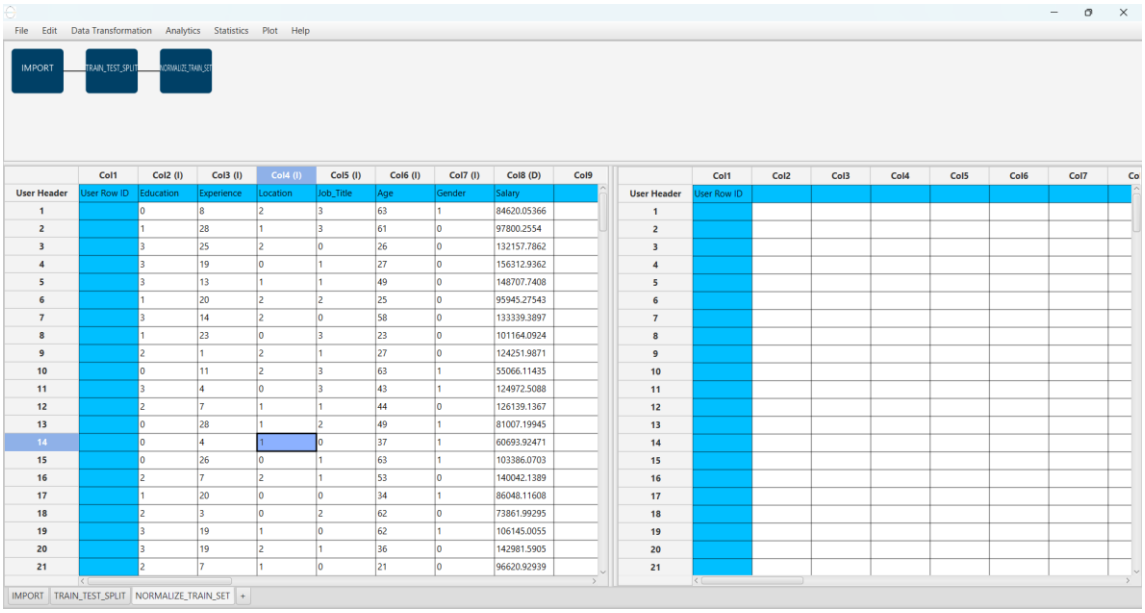
The results will appear on the output spreadsheet.



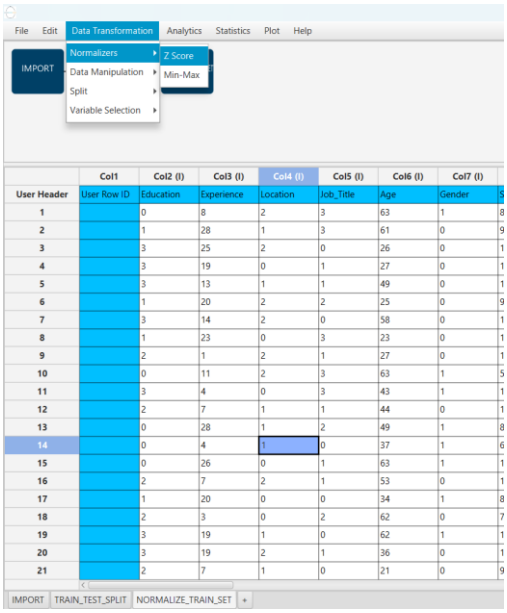
Step 4: Normalize the training set

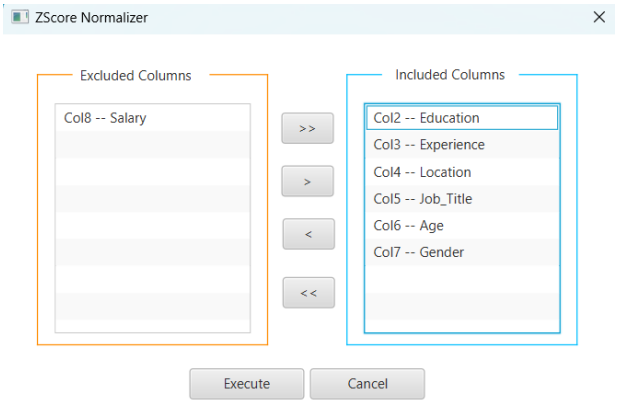
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALIZE_TRAIN_SET".

Import data 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"

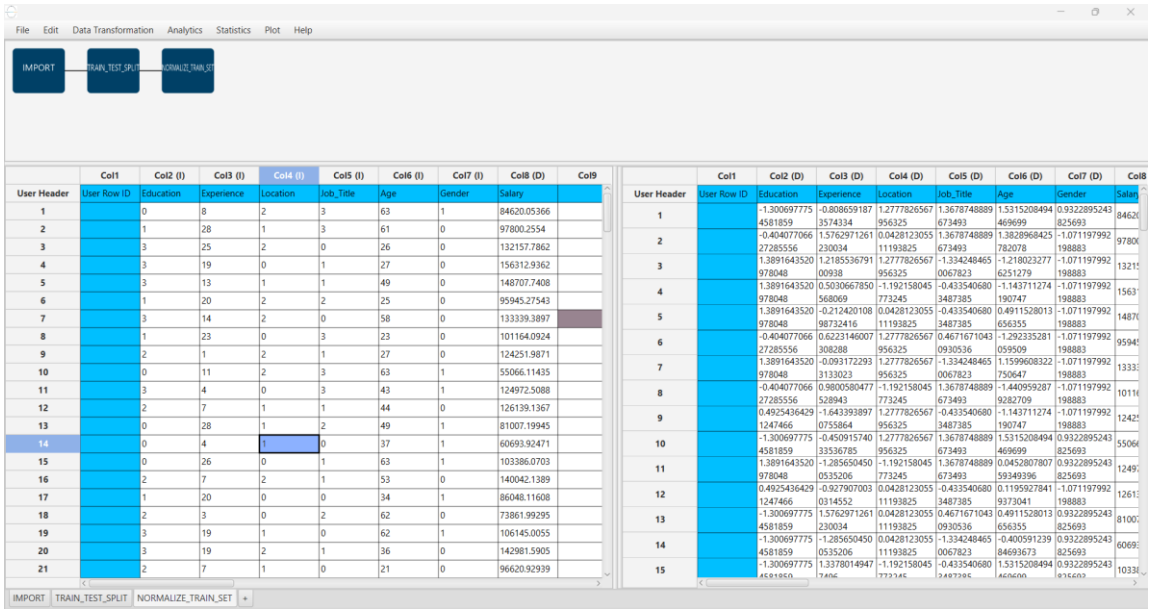


Normalize the data using Z-score by browsing: "Data Transformation" → "Normalizers" → "Z-Score". Then select all columns except "Salary" and click "Execute".





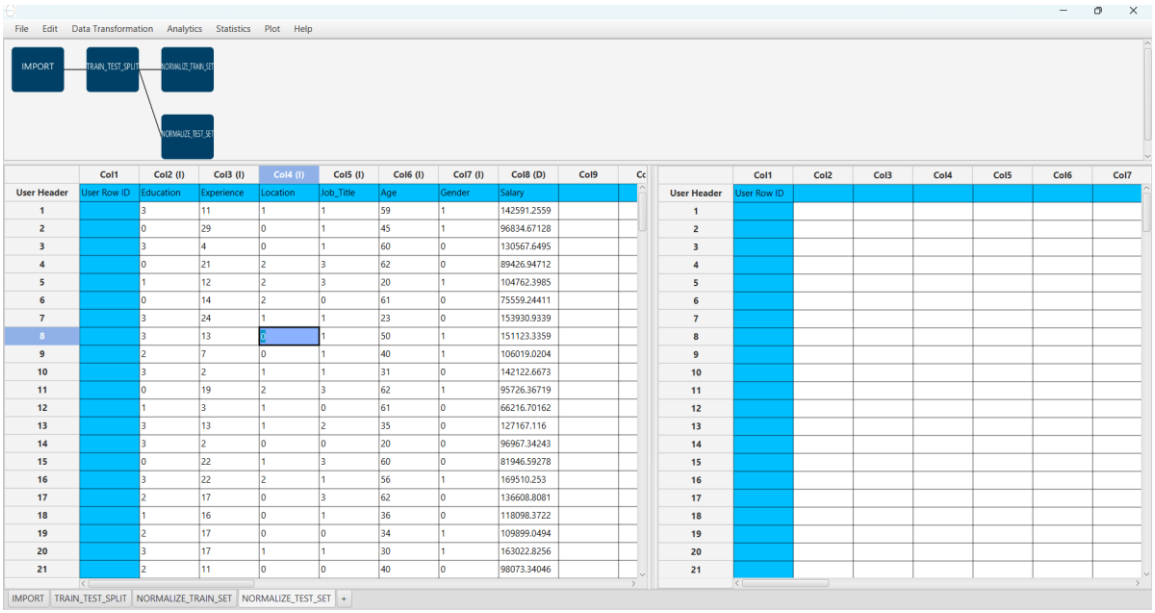
The results will appear on the output spreadsheet.



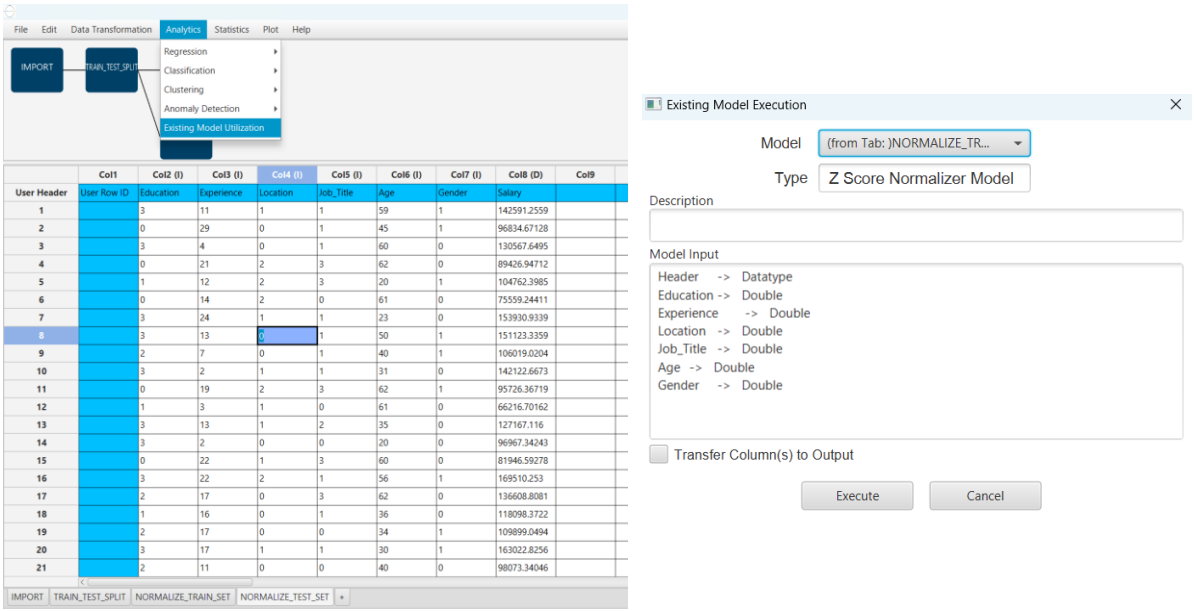
Step 5: Normalize the test set

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

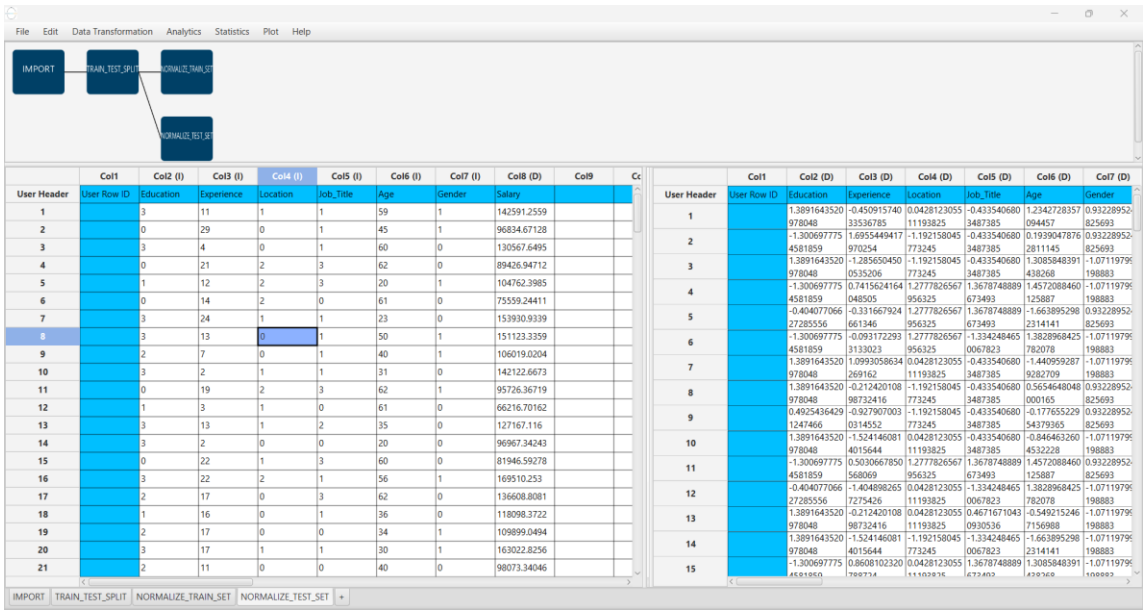
Import data into the input spreadsheet of the "NORMALIZE_TEST_SET" tab the test set from the output of the "TRAIN_TEST_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet". From the available Select input tab options choose "TRAIN_TEST_SPLIT: Test Set".



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



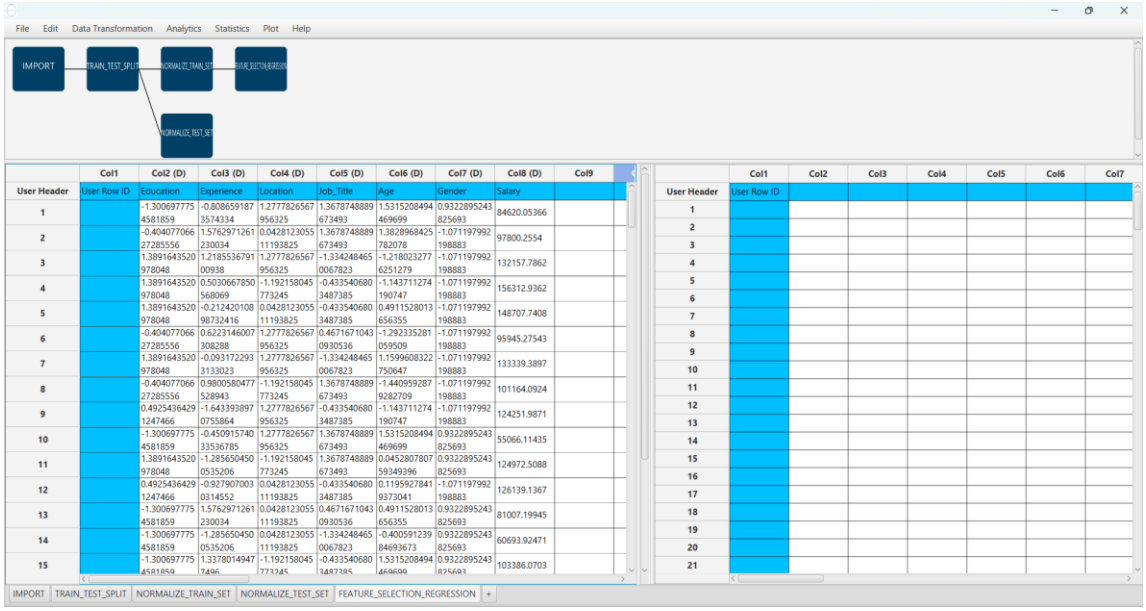
The results will appear on the output spreadsheet.



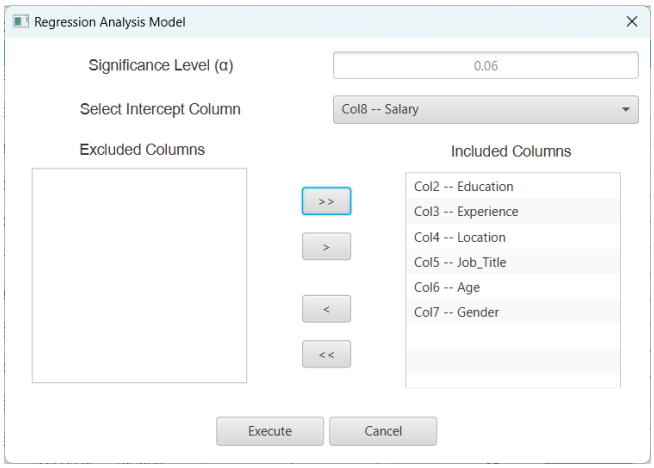
Step 6: Feature selection

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

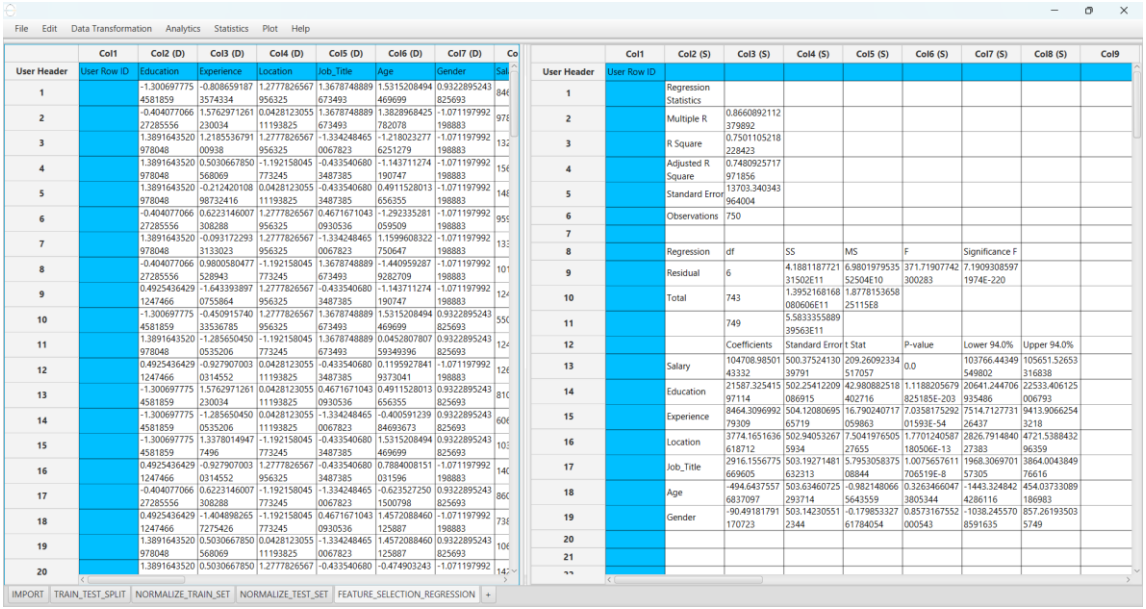
Import data into the input spreadsheet of the "FEATURE_SELECTION_REGRESSION" tab from the output of the "NORMALIZE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



Choose the most important features using the Regression Analysis by browsing: "Data Transformation" → "Variable Selection" → "Regression Analysis". Then choose the "Salary" column as the intercept column, the Significance level (α) as 0.05 and include all columns.



The results will appear on the output spreadsheet.



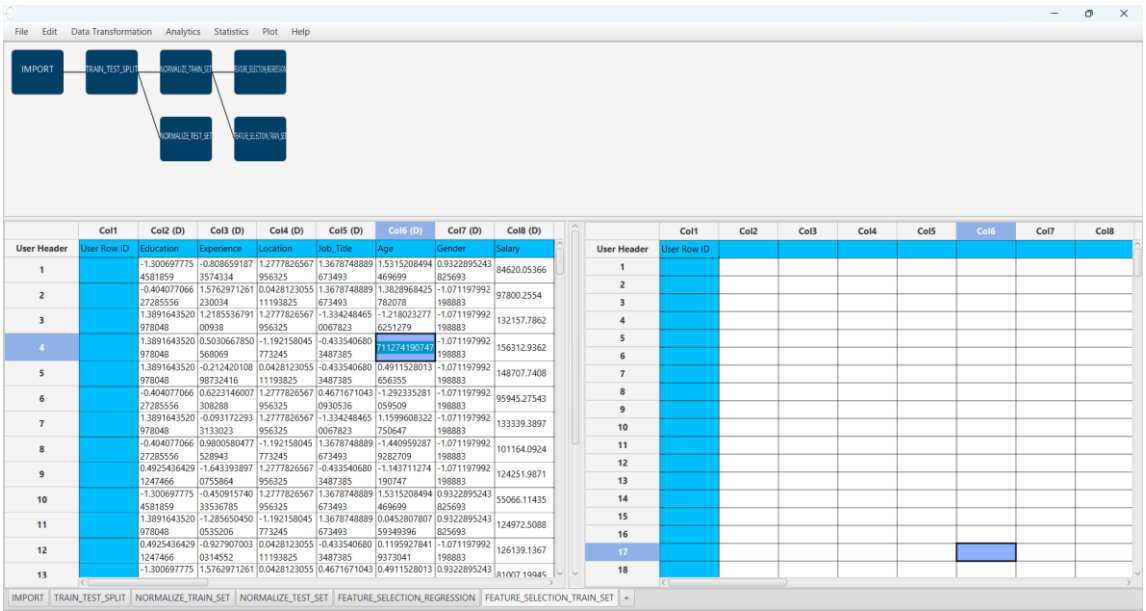
The significant features according to the p-value are the following:

- Salary (p-value = 0.0)
- Education (p-value = 1.1188205679825185E-203)
- Experience (p-value = 7.035817529201593E-54)
- Location (p-value = 1.7701240587180506E-13)
- Job_Title (p-value = 1.0075657611706519E-8)

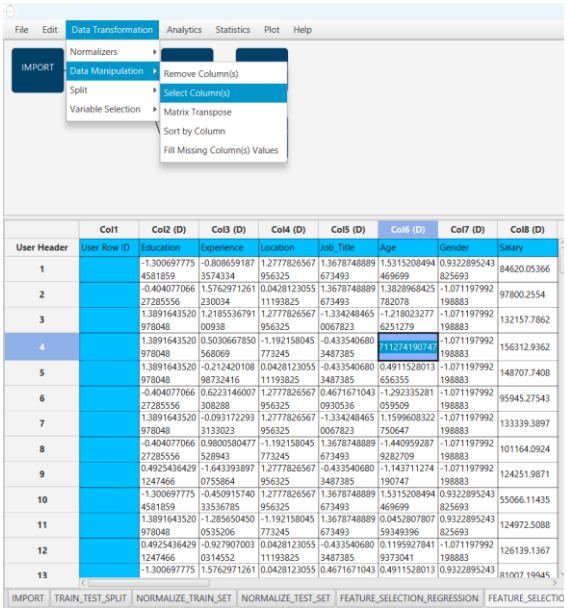
Step 7: Feature selection: train set

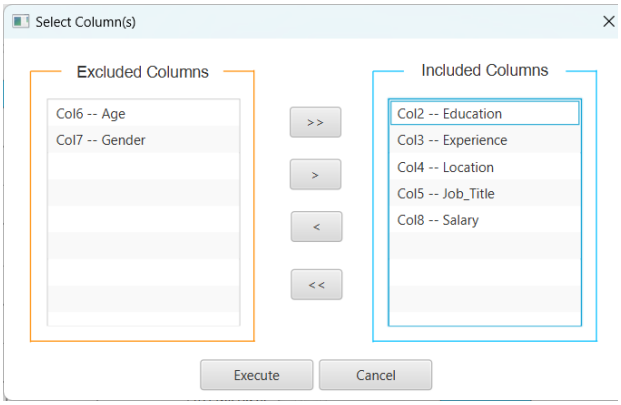
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_TRAIN_SET".

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

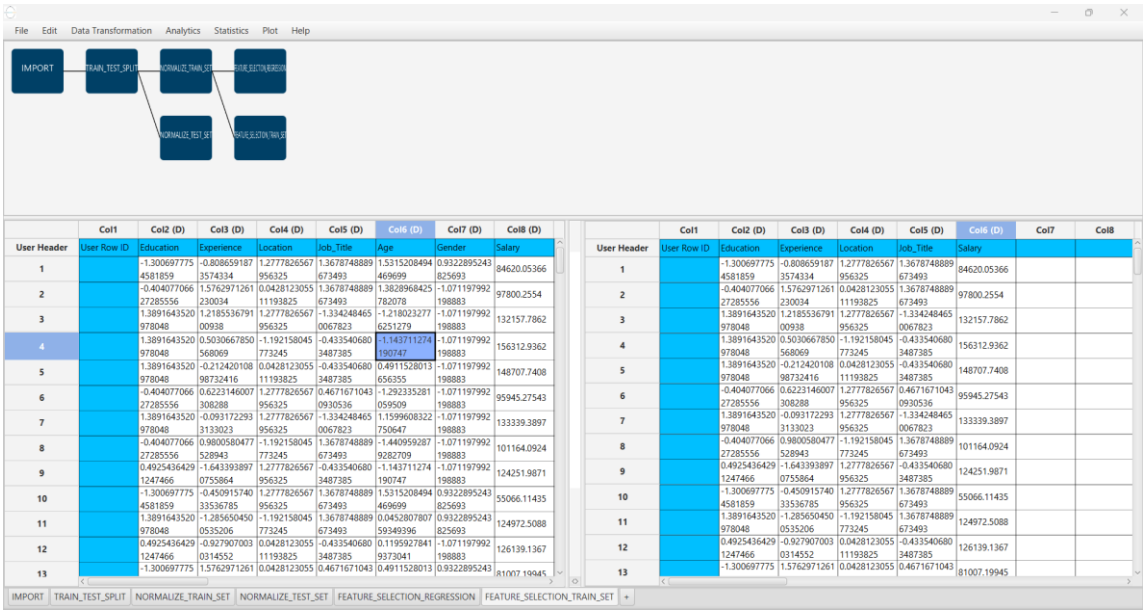


Manipulate the data by choosing the columns that correspond to the significant features (from the previous step): "Data Transformation" → "Data Manipulation" → "Select Column(s)".





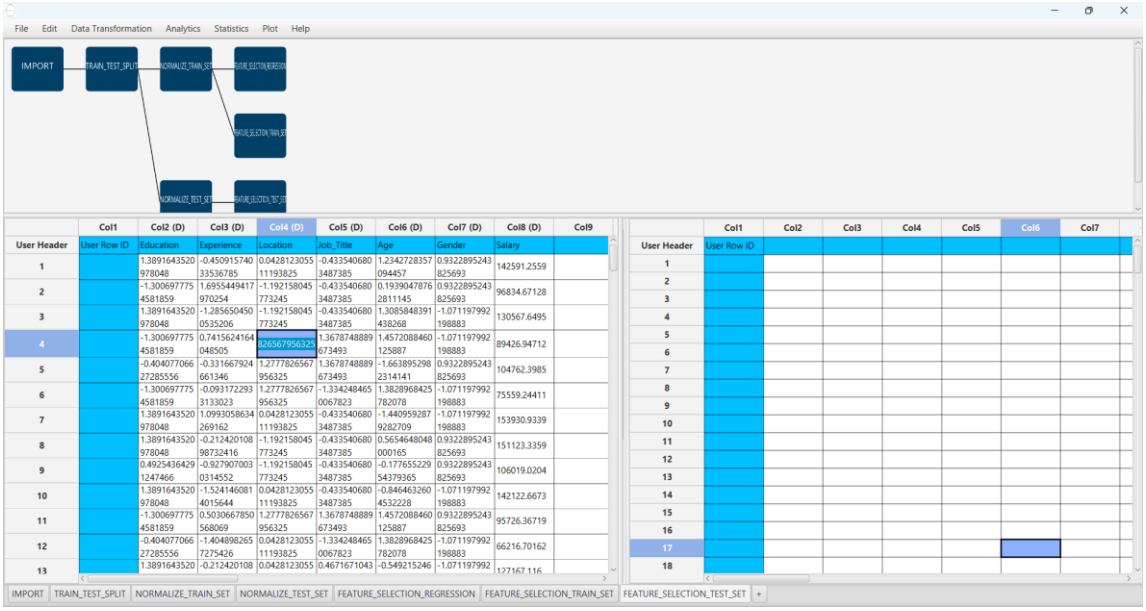
The results will appear on the output spreadsheet.



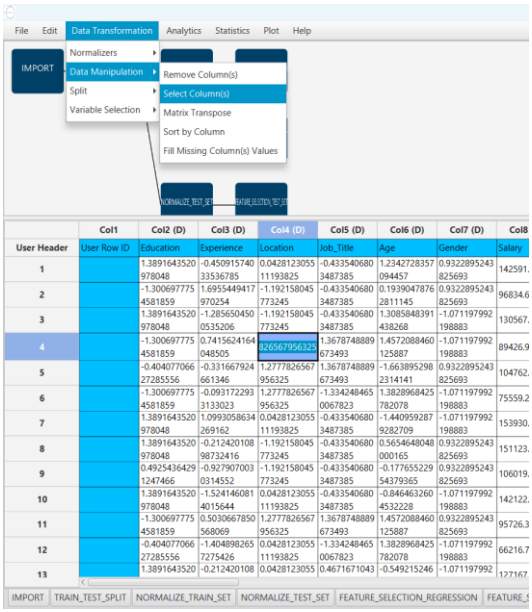
Step 8: Feature selection: test set

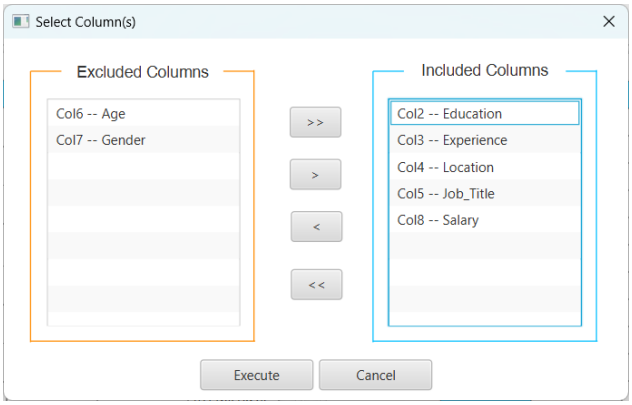
Create a new tab by pressing the "+" button on the bottom of the page with the name "FEATURE_SELECTION_TEST_SET".

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

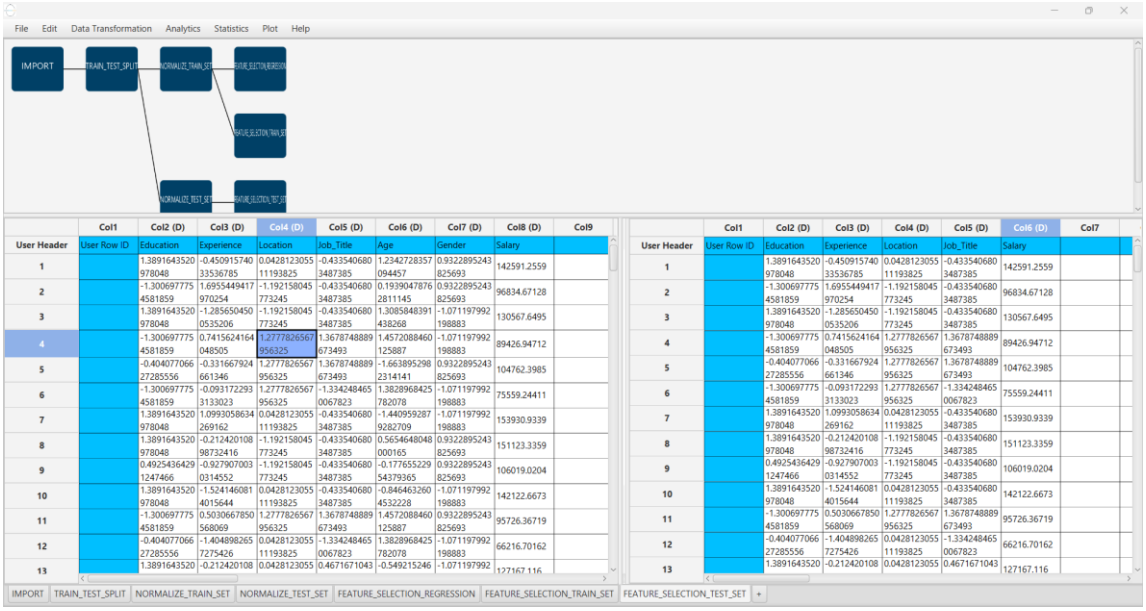


Manipulate the data by choosing the columns that correspond to the significant features (from step 7): "Data Transformation" → "Data Manipulation" → "Select Column(s)".





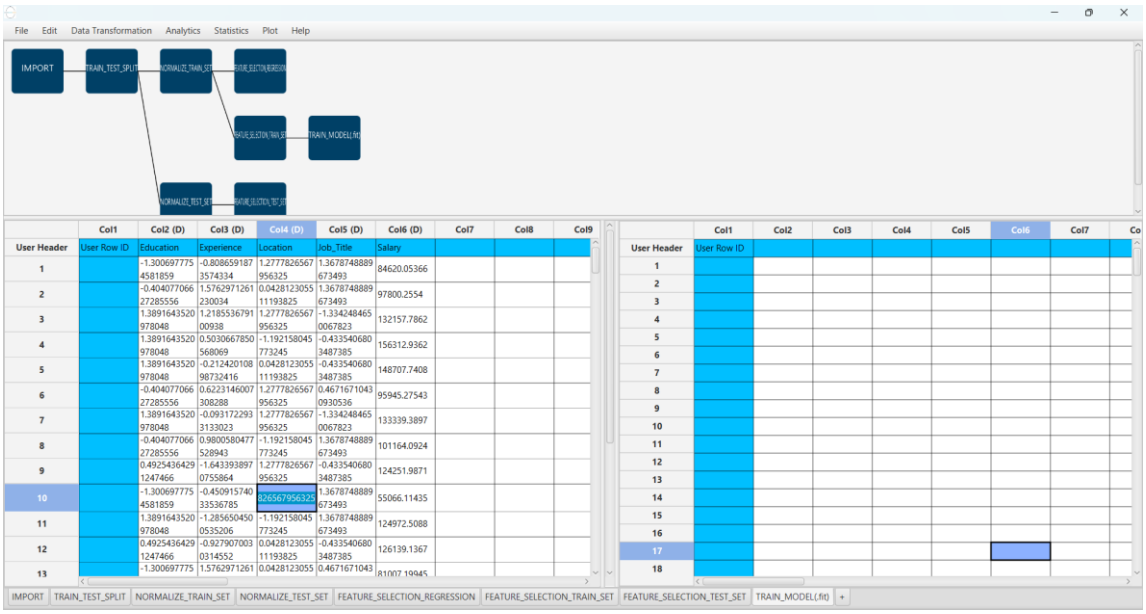
The results will appear on the output spreadsheet.



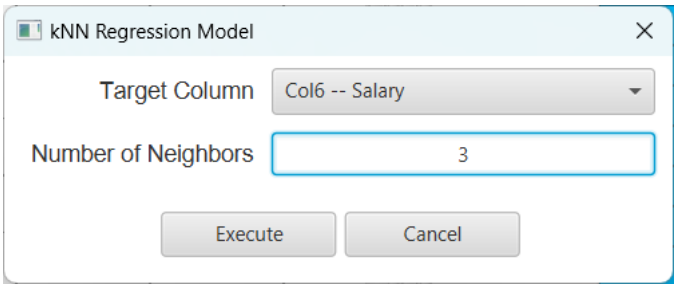
Step 9: Train the model

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

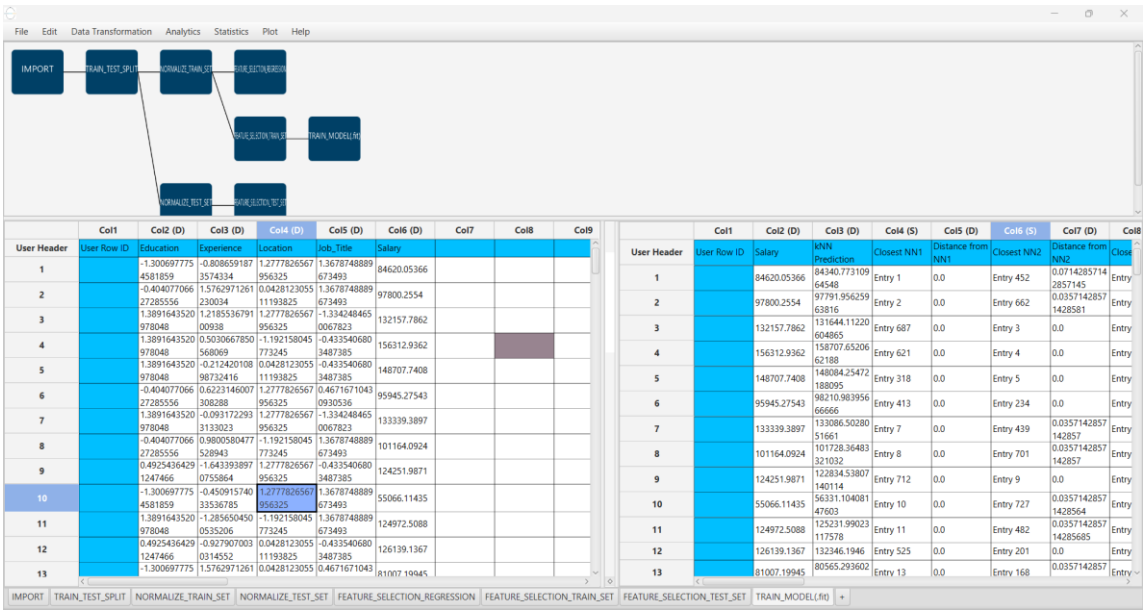
Import data into the input spreadsheet of the "TRAIN_MODEL(.fit)" tab from the output of the "FEATURE_SELECTION_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



Use the k Nearest Neighbors (kNN) method to train and fit the model by browsing: "Analytics" → "Regression" → "k Nearest Neighbors (kNN)" and set the "Target Column" as the column corresponding to "Salary" and the "Number of Neighbors" to 3.



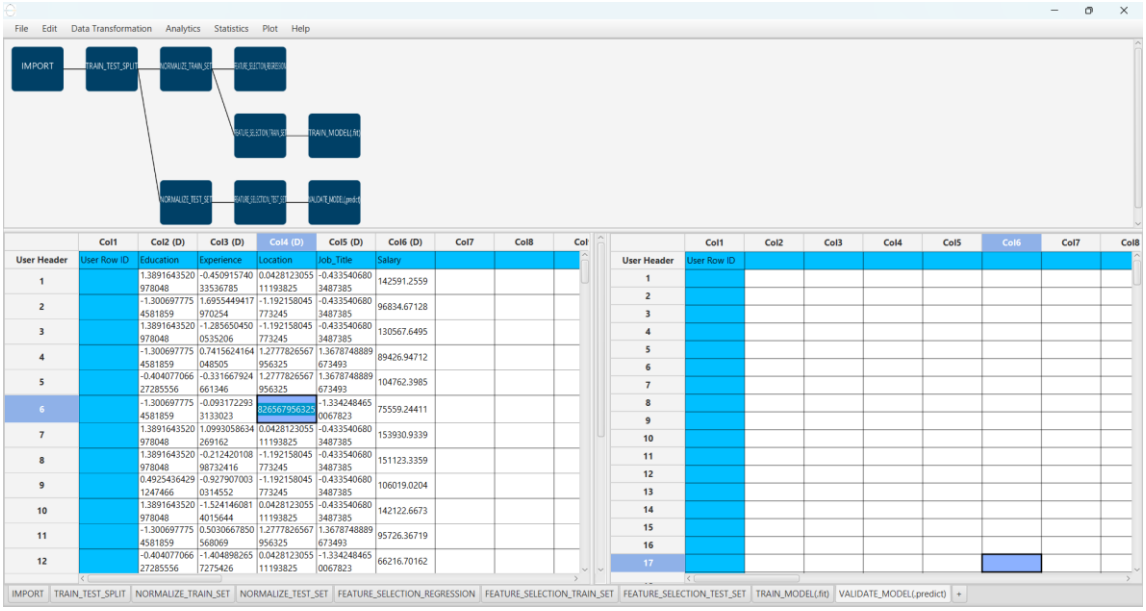
The predictions will appear on the output spreadsheet.



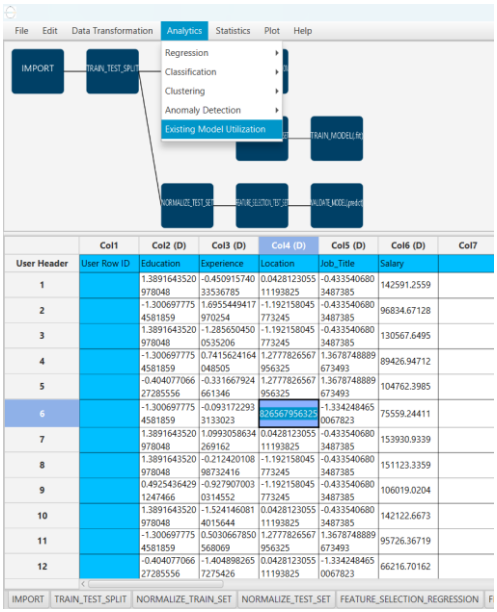
Step 10: Validate the model

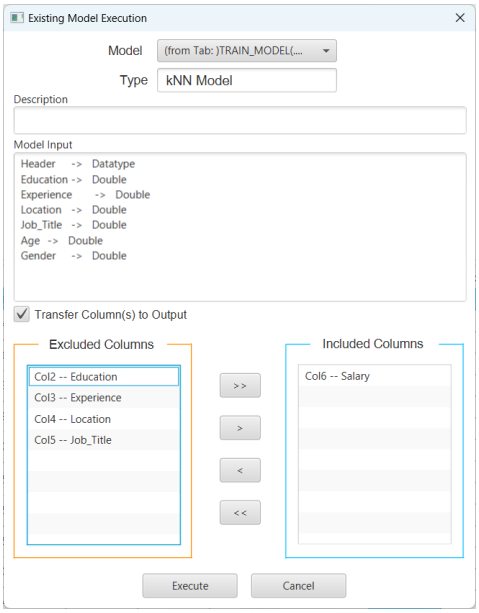
Create a new tab by pressing the "+" button on the bottom of the page with the name "VALIDATE_MODEL(.predict)".

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".

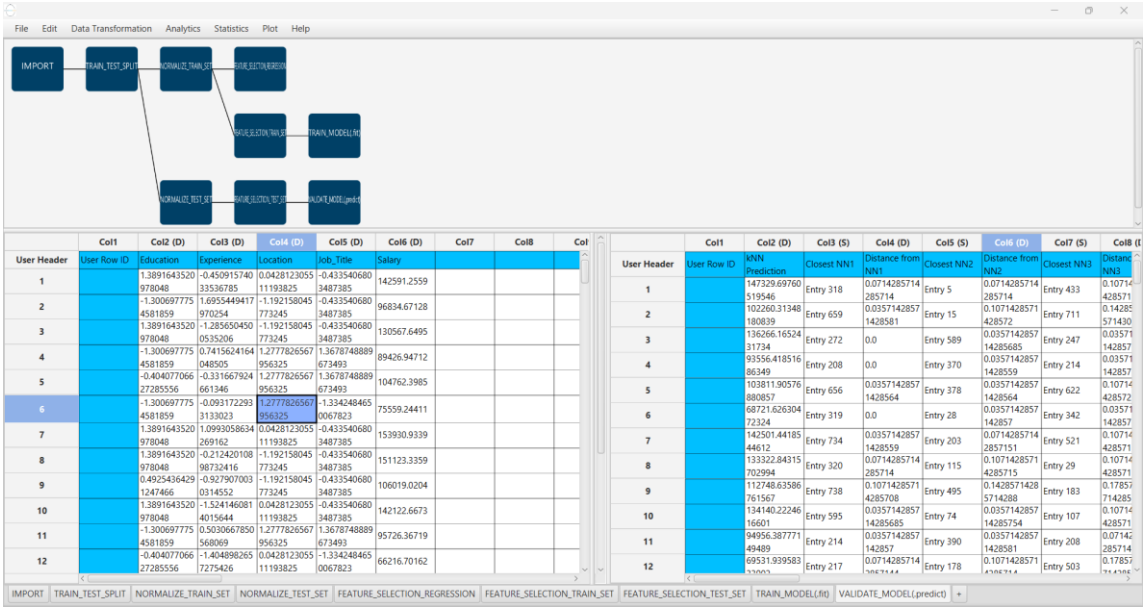


To validate the model browse: "Analytics" → "Existing Model Utilization". Then choose Model "(from Tab:) TRAIN_MODEL (.fit)" and transfer the "Salary" column in the output.





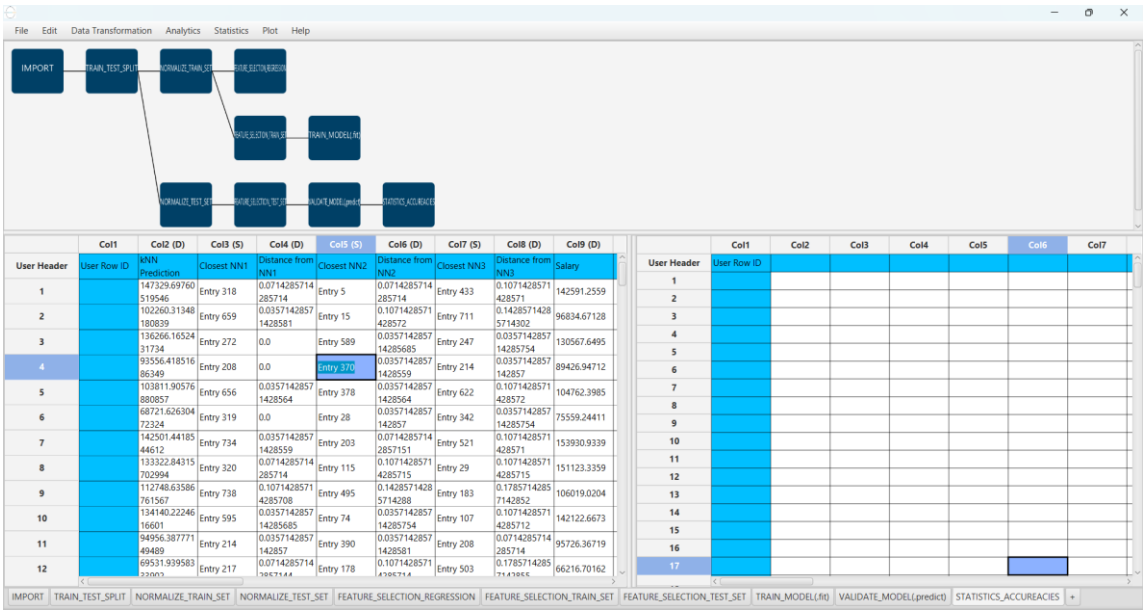
The predictions will appear on the output spreadsheet.



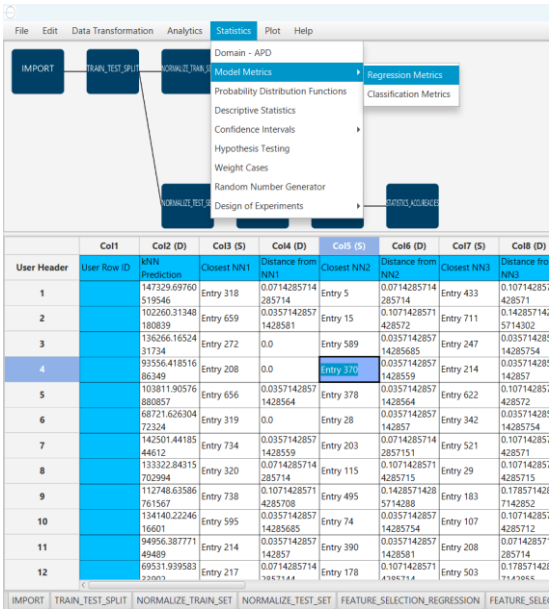
Step 11: Statistics calculation

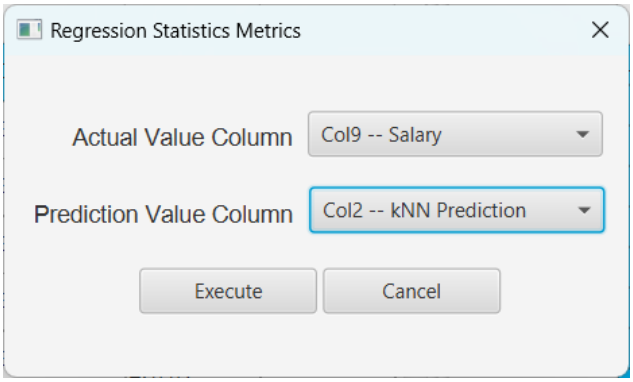
Create a new tab by pressing the "+" button on the bottom of the page with the name "STATISTICS_ACCURACIES".

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".

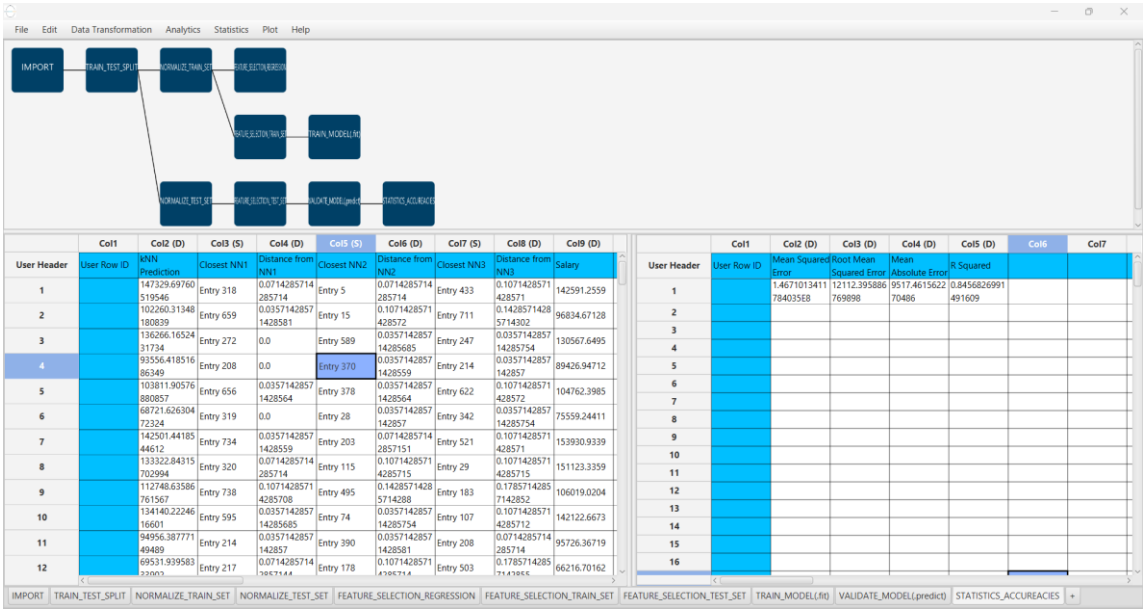


Calculate the statistical metrics for the regression by browsing: "Statistics" → "Model Metrics" → "Regression Metrics".





The results will appear on the output spreadsheet.

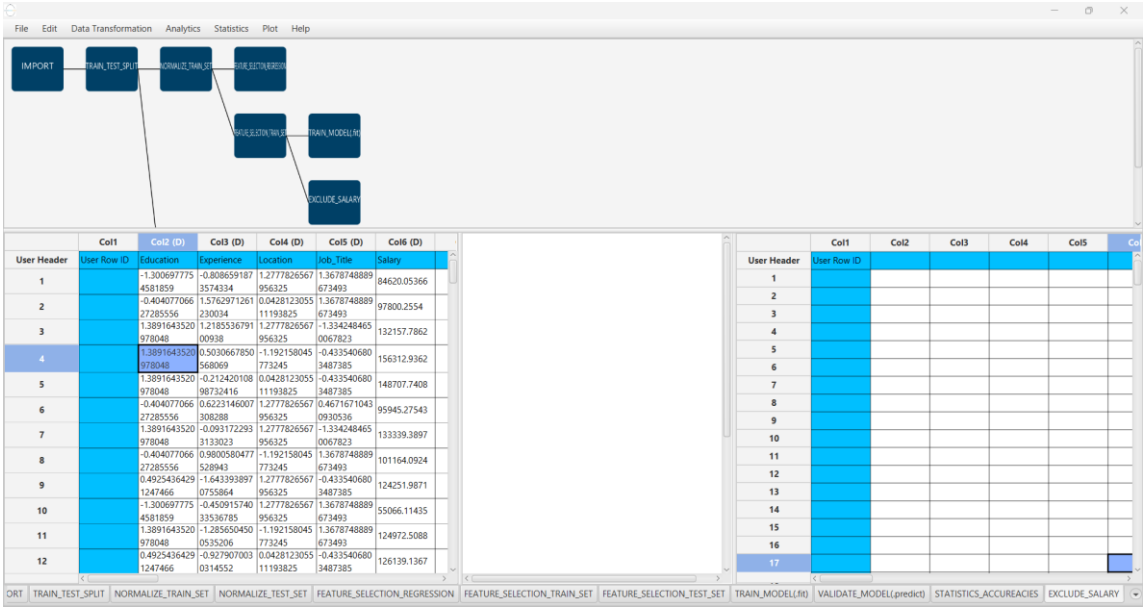


Step 12: Reliability check of each record of the test set

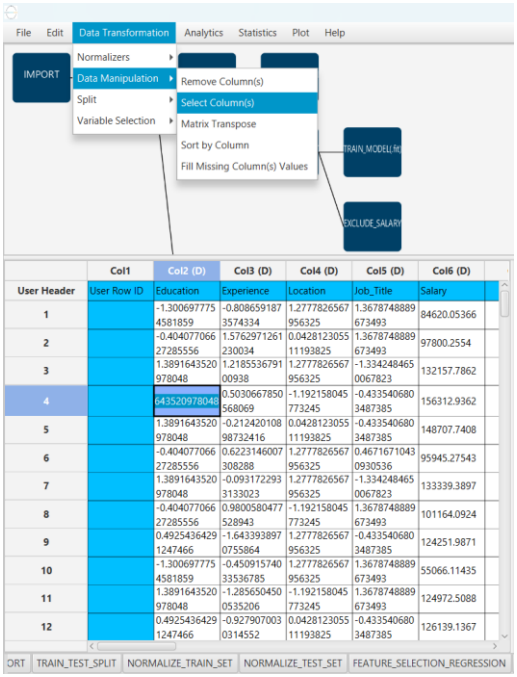
Step 12.a: Create the domain

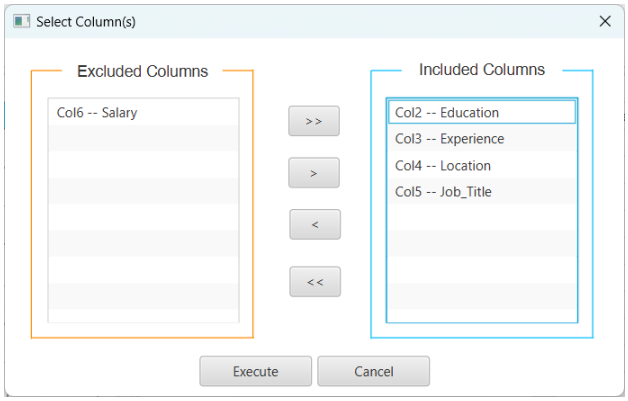
Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_SALARY".

Import data into the input spreadsheet of the "EXCLUDE_SALARY" 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 "Salary" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except the "Salary".

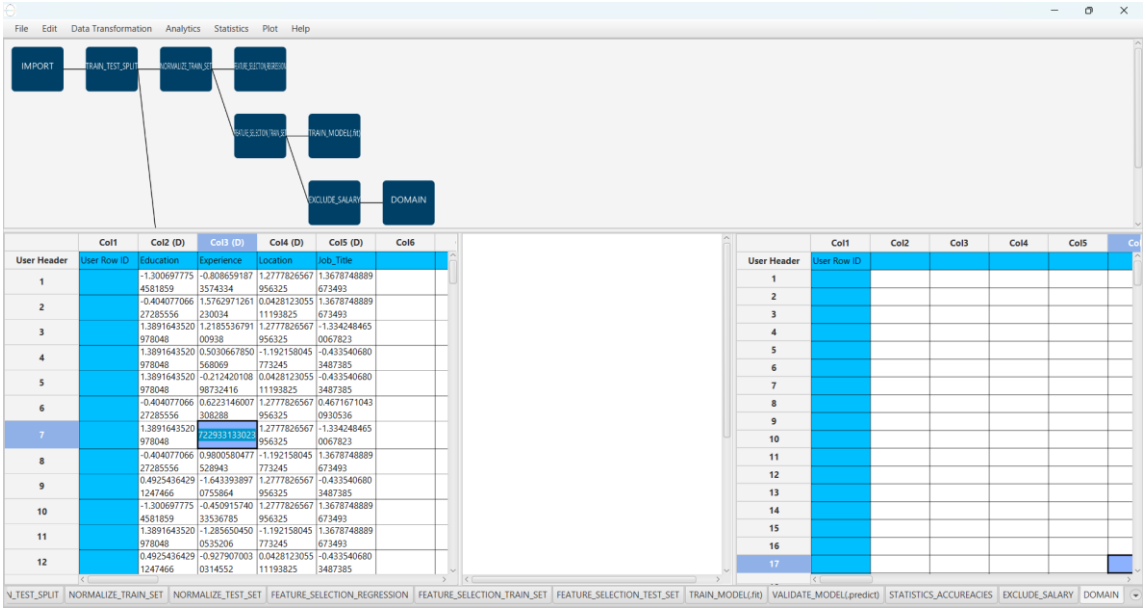




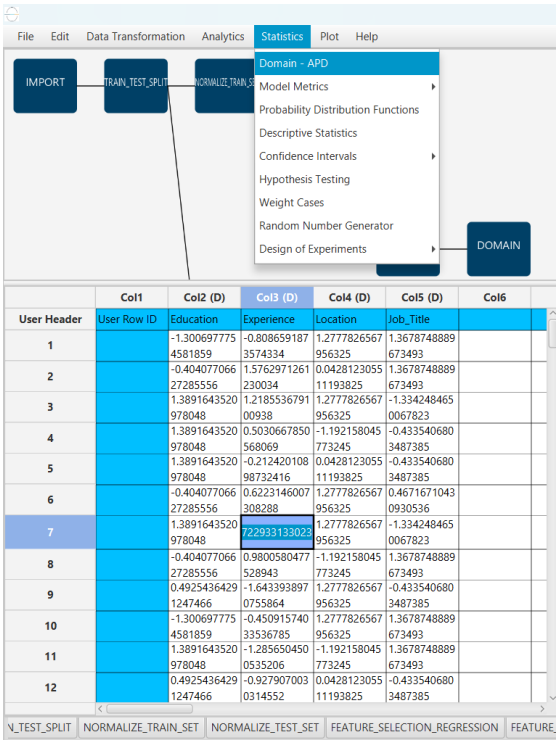
The results will appear on the output spreadsheet.

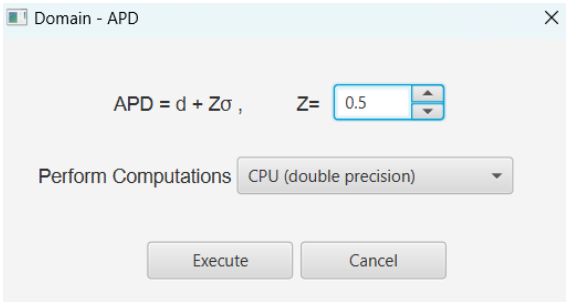
Create a new tab by pressing the "+" button on the bottom of the page with the name "DOMAIN".

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

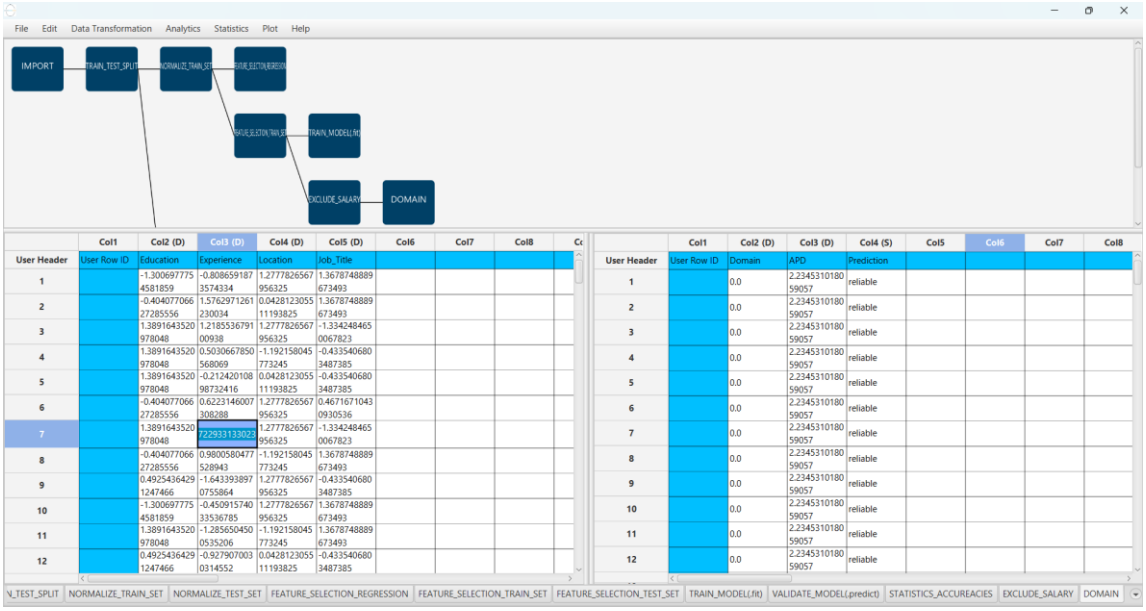


Create the domain by browsing: "Statistics" → "Domain APD".





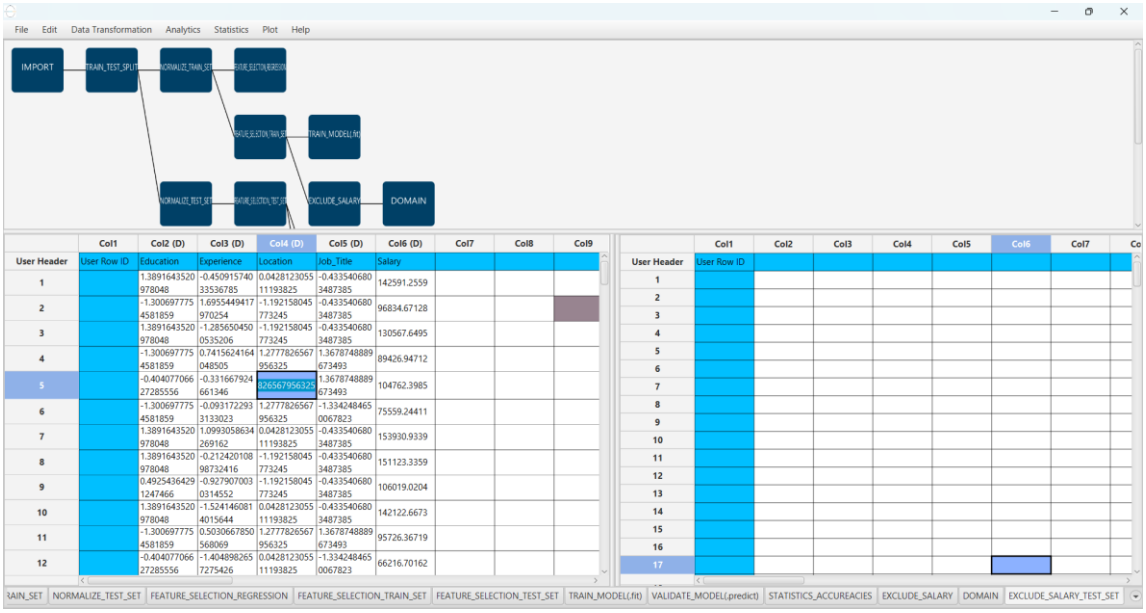
The results will appear on the output spreadsheet.



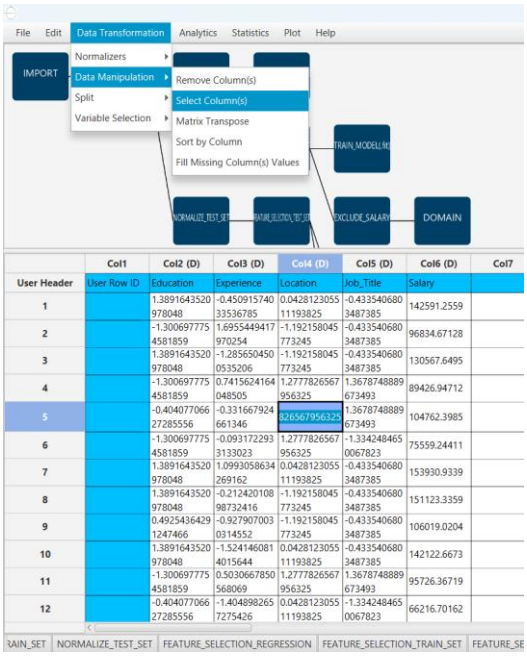
Step 12.b: Check the test set reliability

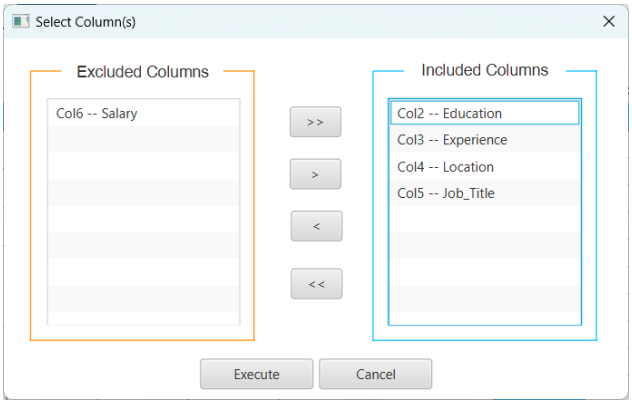
Create a new tab by pressing the "+" button on the bottom of the page with the name "EXCLUDE_SALARY_TEST_SET".

Import data into the input spreadsheet of the "EXCLUDE_SALARY_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 "Salary" by browsing: "Data Transformation" → "Data Manipulation" → "Select Columns". Then select all the columns except "Salary".

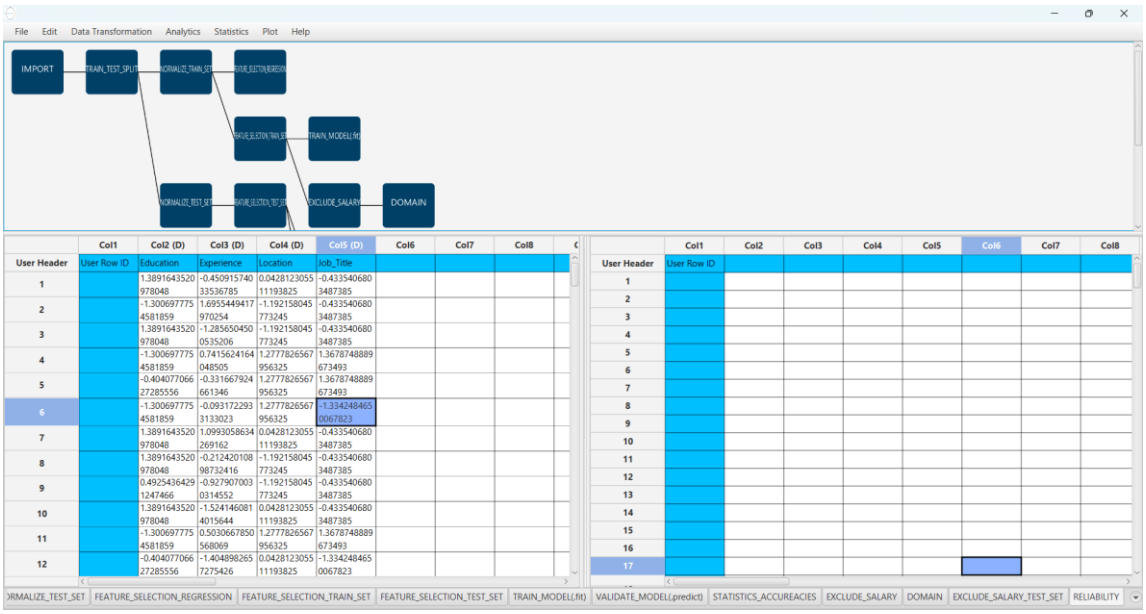




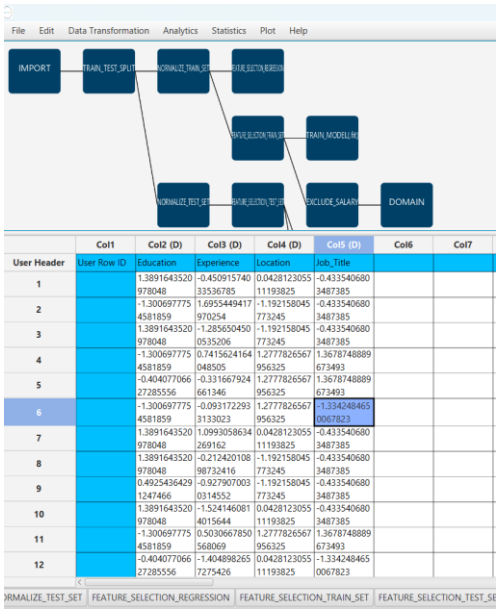
The results will appear on the output spreadsheet.

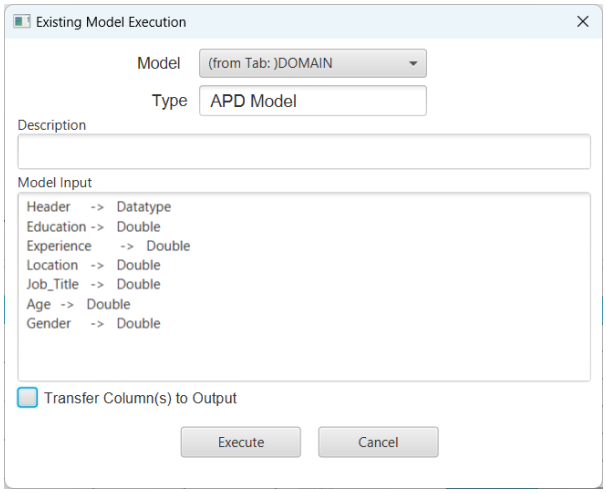
Create a new tab by pressing the "+" button on the bottom of the page with the name "RELIABILITY".

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

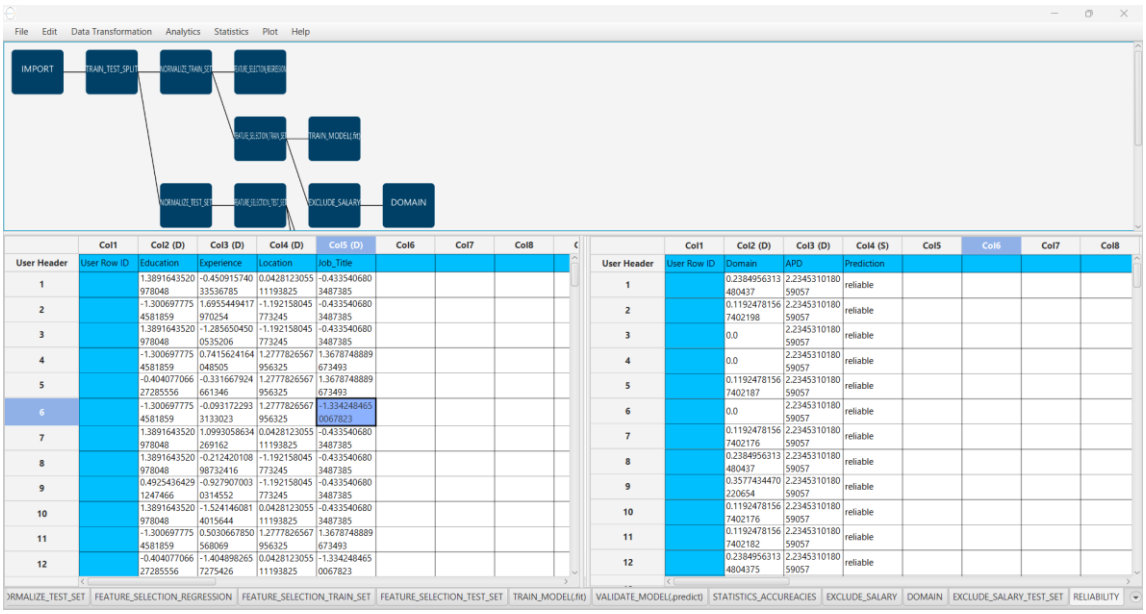


Check the Reliability by browsing: "Analytics" → "Existing Model Utilization". Then select as Model "(from Tab:) DOMAIN".





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: