Students' Performance (Multilabel Classification)

The goal of this study is to train a model in order to predict the grade class of high school students. The dataset used in this case study is found in https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset and has 15 features and 2392 labelled samples. The dataset includes demographic details, study habits, parental involvement, extracurricular activities and academic performance.

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:

"GradeClass":

- 'A' - GPA >= 3.5 (0)
- 'B' - 3.0 <= GPA < 3.5 (1)
- 'C' - 2.5 <= GPA < 3.0 (2)
- 'D' - 2.0 <= GPA < 2.5 (3)
- 'F' - GPA < 2.0 (4)
-

"Gender":

- Male (0)
- Female (1)

"Ethnicity":

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

"ParentalEducation":

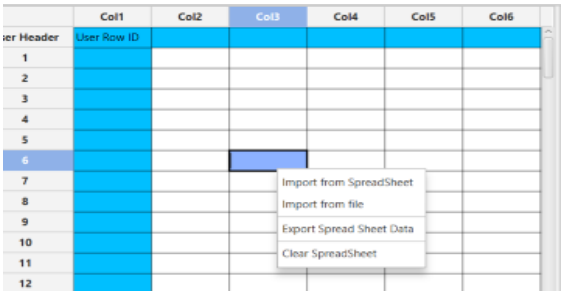
- None (0)
- High School (1)
- Some College (2)
- Bachelor's (3)
- Higher (4)

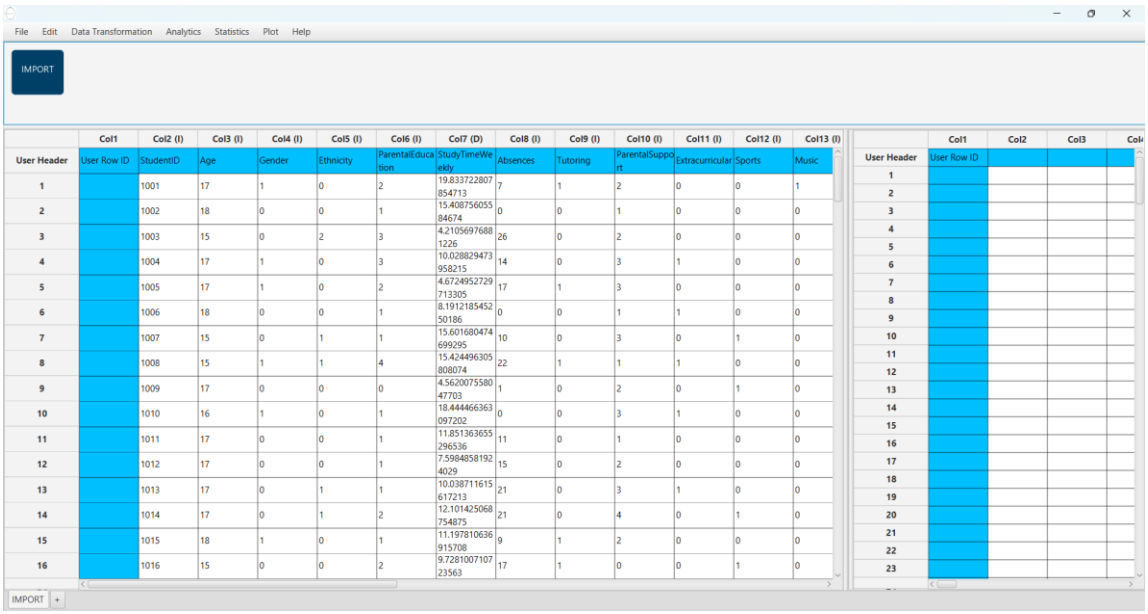
"ParentalSupport":

- None (0)
- Low (1)
- Moderate (2)
- High (3)
- Very High (4)

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 Students' Performance data.

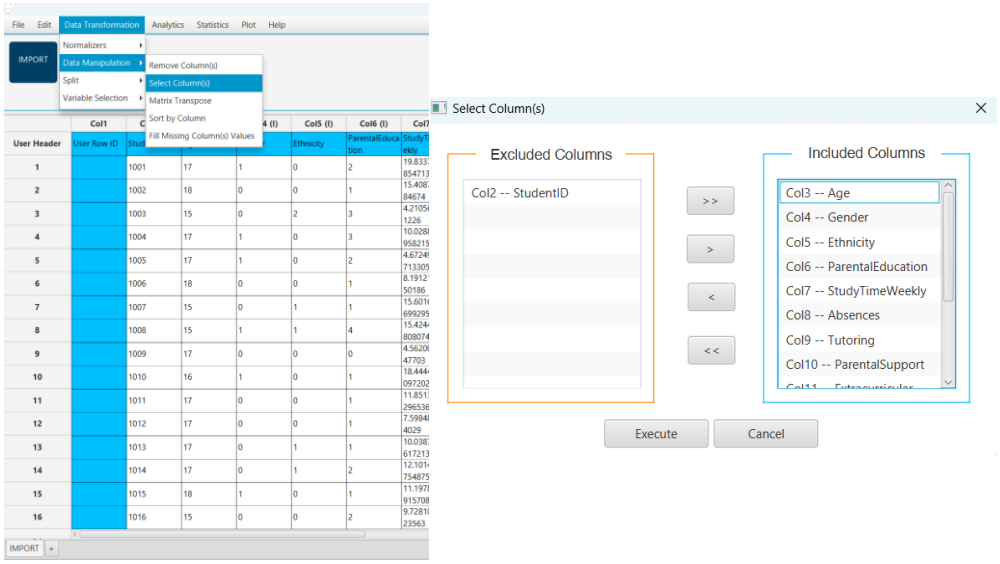




Step 2: Manipulate data

In order to use the data for training we have to exclude any columns that do not contain features, like "StudentID". We follow these steps to execute this:

- On the menu click on "Data Transformation" → "Data Manipulation" → "Select Column(s)"
- Select all columns except the one that corresponds to the "StudentID".

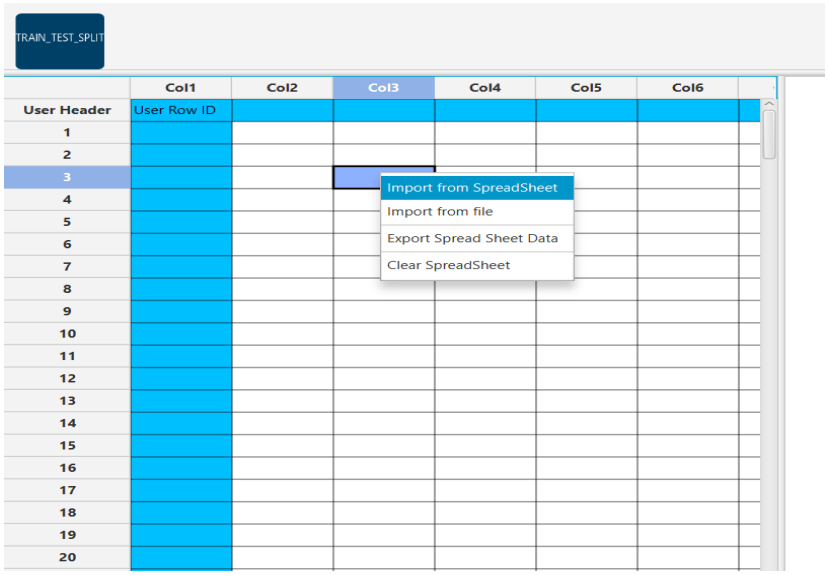


The data without the "StudentID" column will appear in the output spreadsheet.

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.

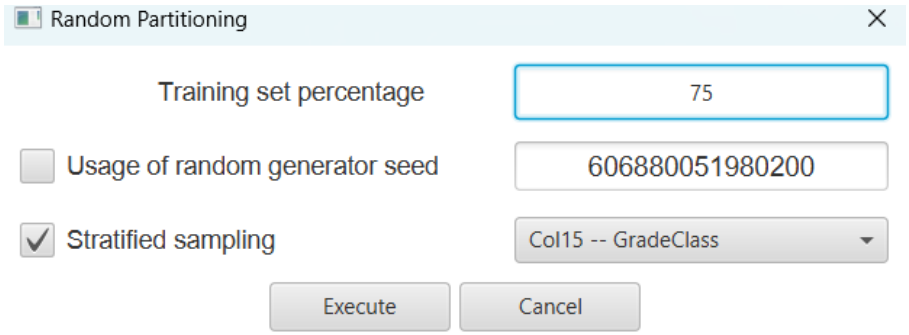
Import data into the input spreadsheet of the "TRAIN_TEST_SPLIT" tab from the output of the "IMPORT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



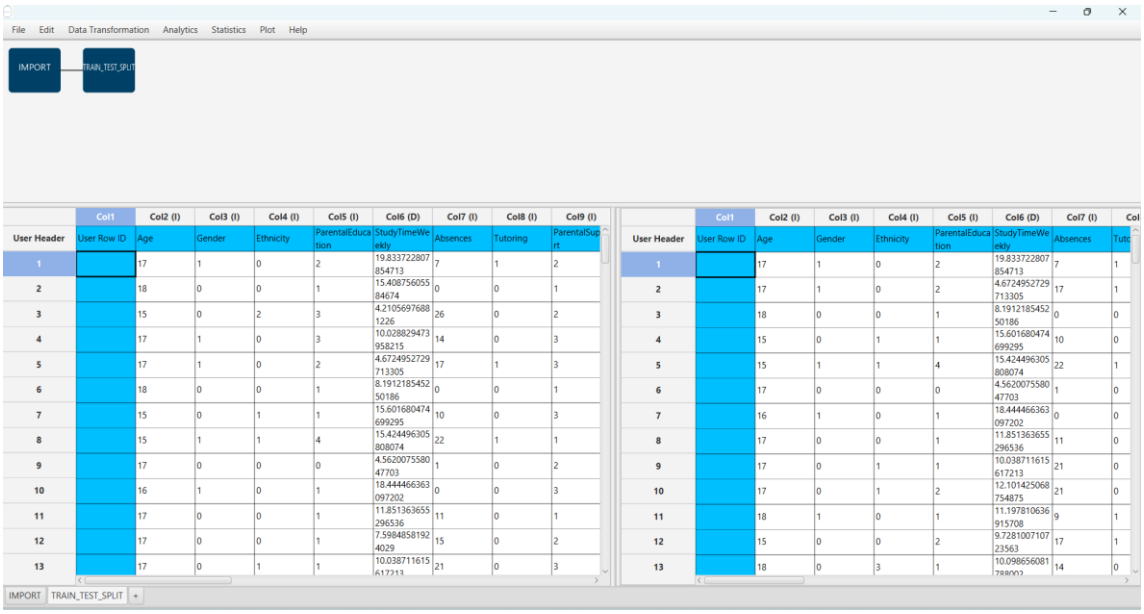
Split the dataset by choosing

"Data Transformation" → "Split" → "Random Partitioning"

Then choose the "Training set percentage" and the column for the sampling as shown below:



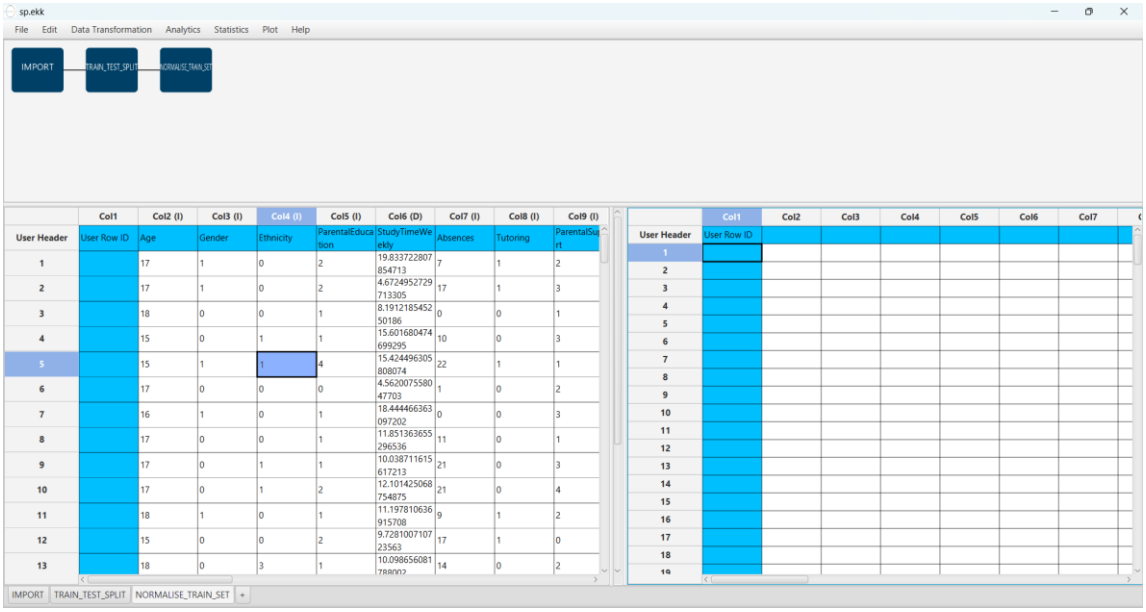
The results will appear on the output spreadsheet.



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".

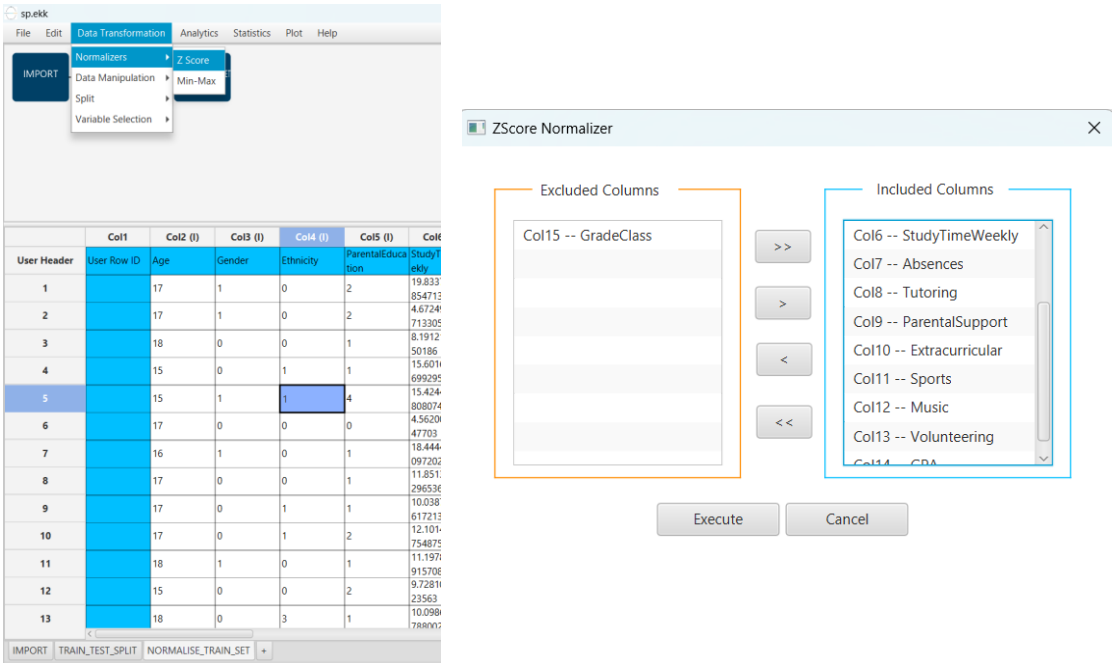
Import data into the input spreadsheet of the "NORMALISE_TRAIN_SET" tab the train set from the output of the "TRAIN_TEST_SPLIT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet". From the available Select input tab options choose "TRAIN_TEST_SPLIT: Training Set"



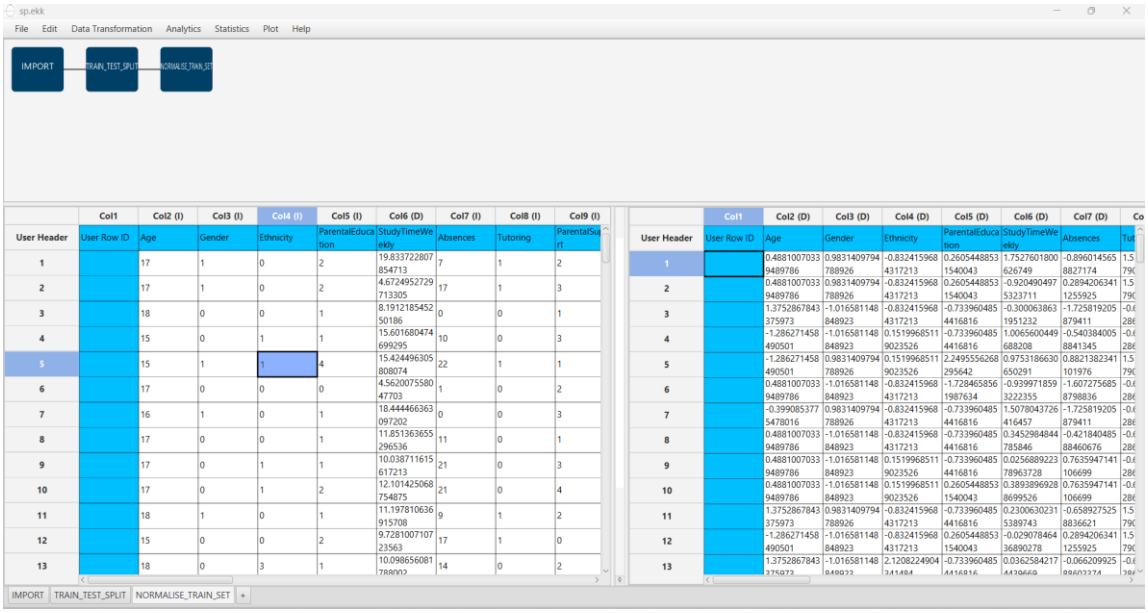
Normalize the data using Z-score:

"Data Transformation" → "Normalizers" → "Z-Score"

Then select all columns and click "Execute".



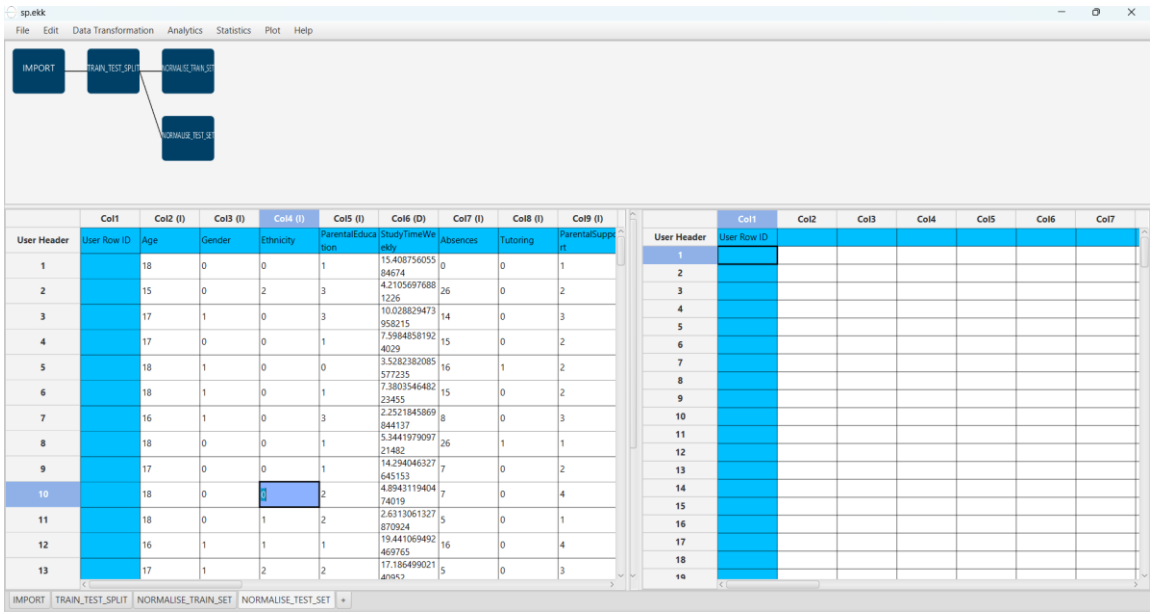
The results will appear on the output spreadsheet.



Step 5: Normalize the test set

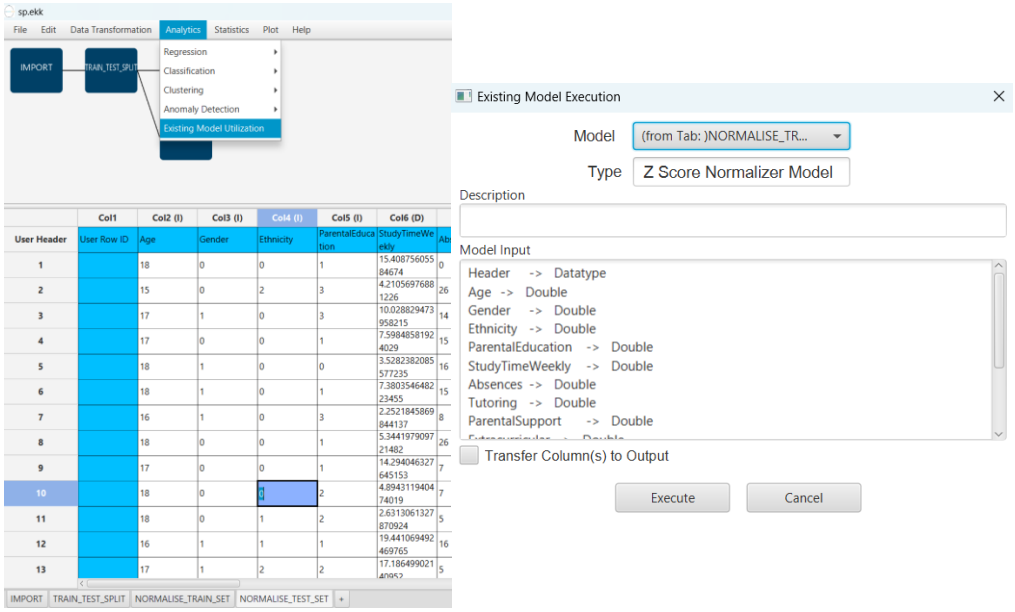
Create a new tab by pressing the "+" button on the bottom of the page with the name "NORMALISE_TEST_SET".

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".

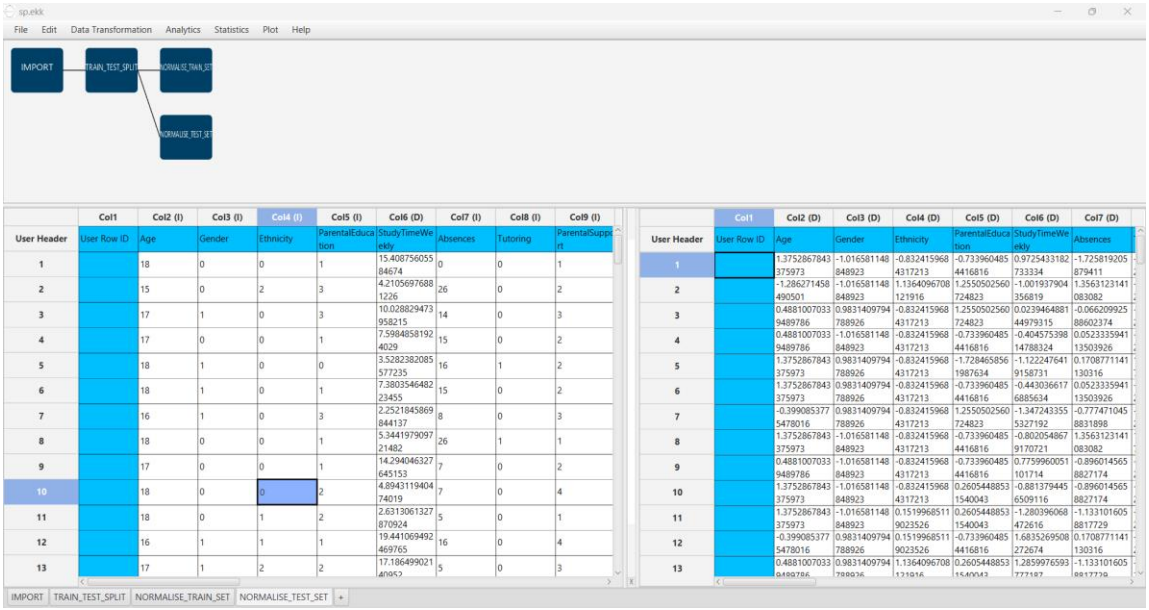


Normalize the test set using the existing normalizer of the training set:

"Analytics" → "Existing Model Utilization" → "Model (from Tab:) NORMALISE TRAIN_SET".



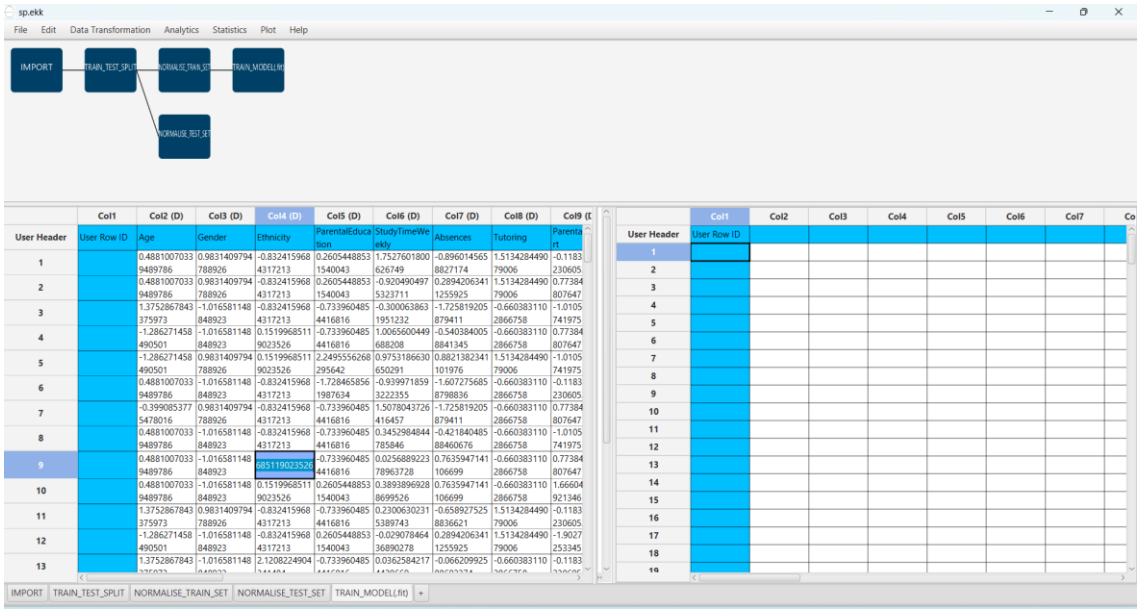
The results will appear on the output spreadsheet.



Step 6: Train the model

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

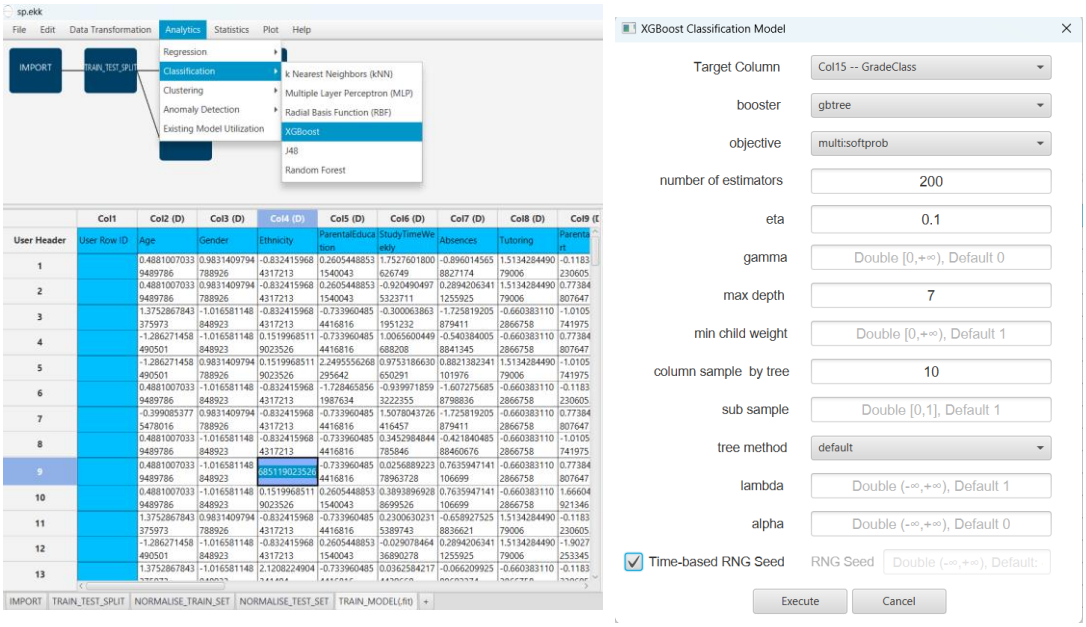
Import data into the input spreadsheet of the "TRAIN_MODEL(.fit)" tab from the output of the "NORMALISE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



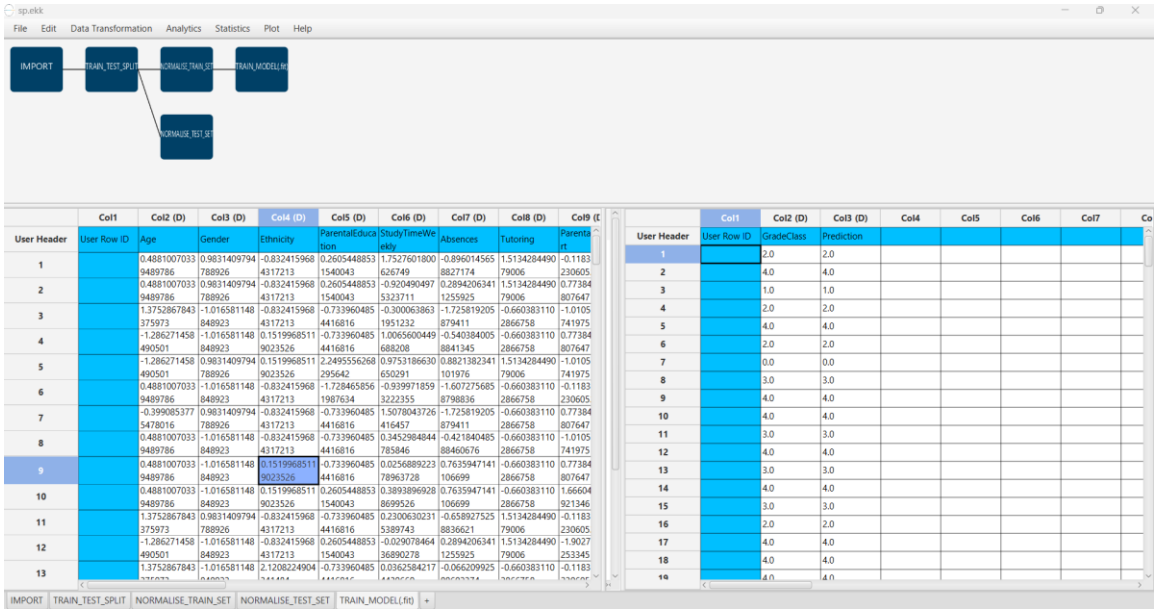
Use the XGBoost Method to train and fit the model:

"Analytics" → "Classification" → "XGBoost"

and adjust model parameters based on training set performance.



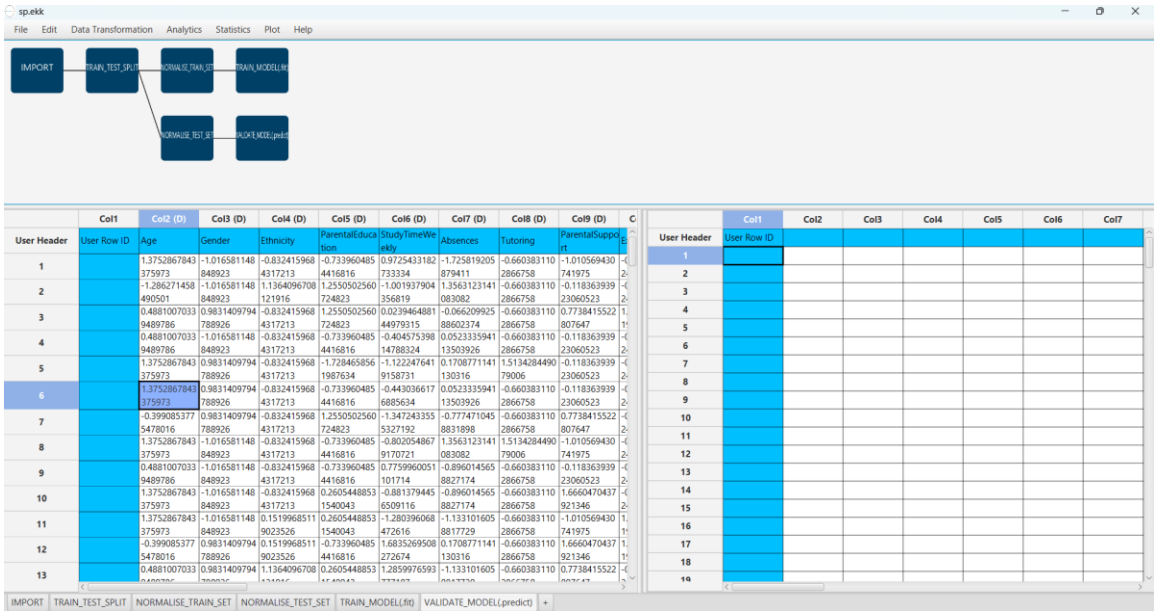
The predictions will appear on the output spreadsheet.



Step 7: Validate the model

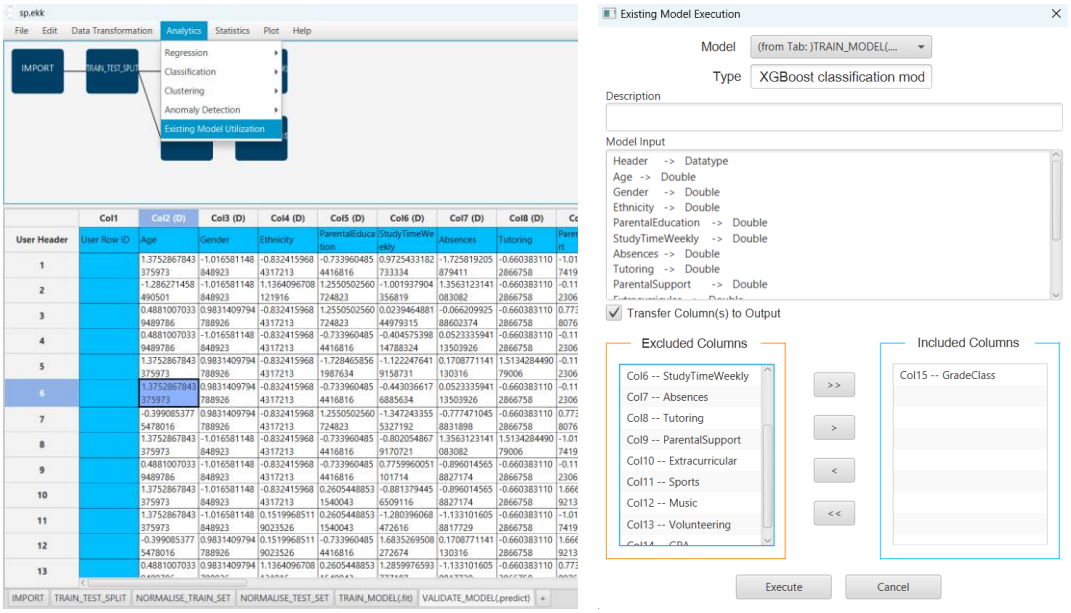
Create a new tab by pressing the "+" button on the bottom of the page with the name "VALIDATE_MODEL(.predict)".

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".

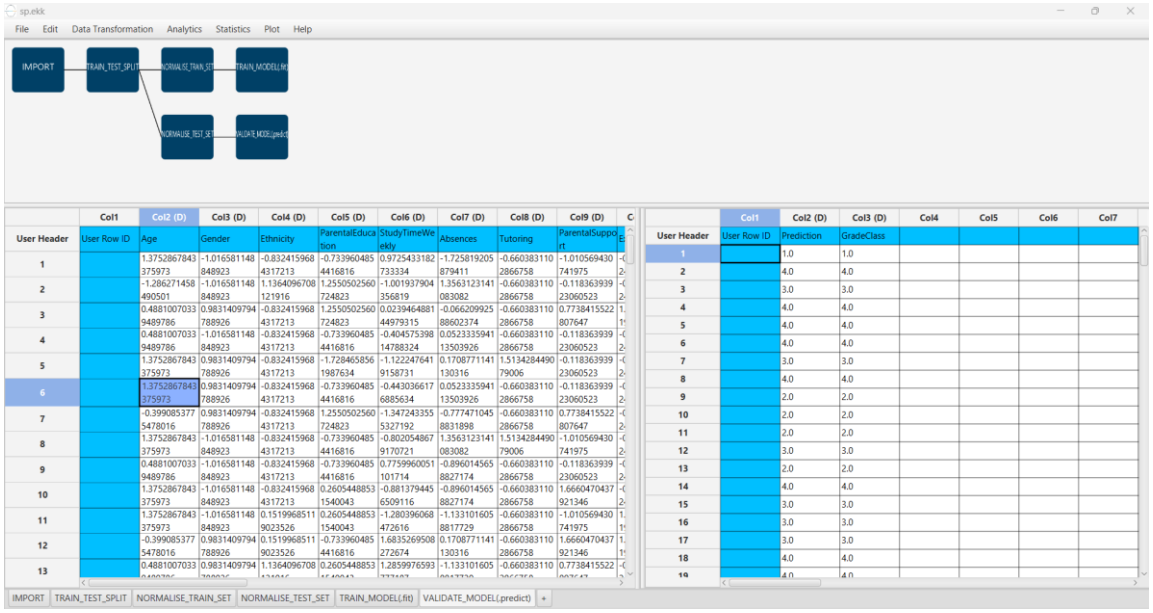


To validate the model:

"Analytics" → "Existing Model Utilization". Then choose Model "(from Tab:) TRAIN_MODEL (.fit)".



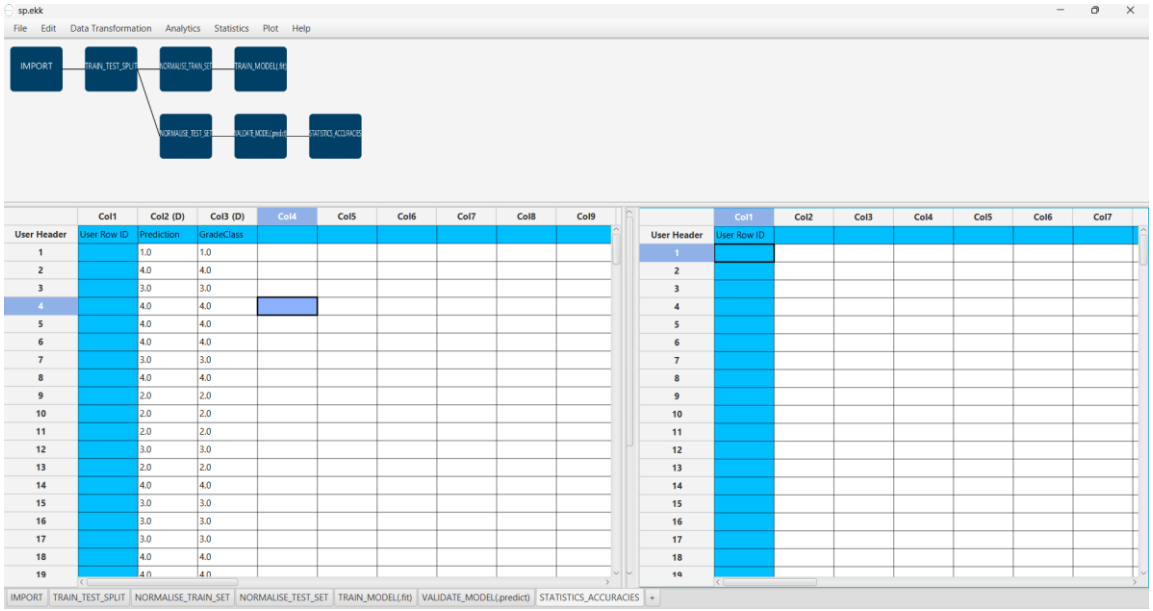
The predictions will appear on the output spreadsheet.



Step 8: Statistics calculation

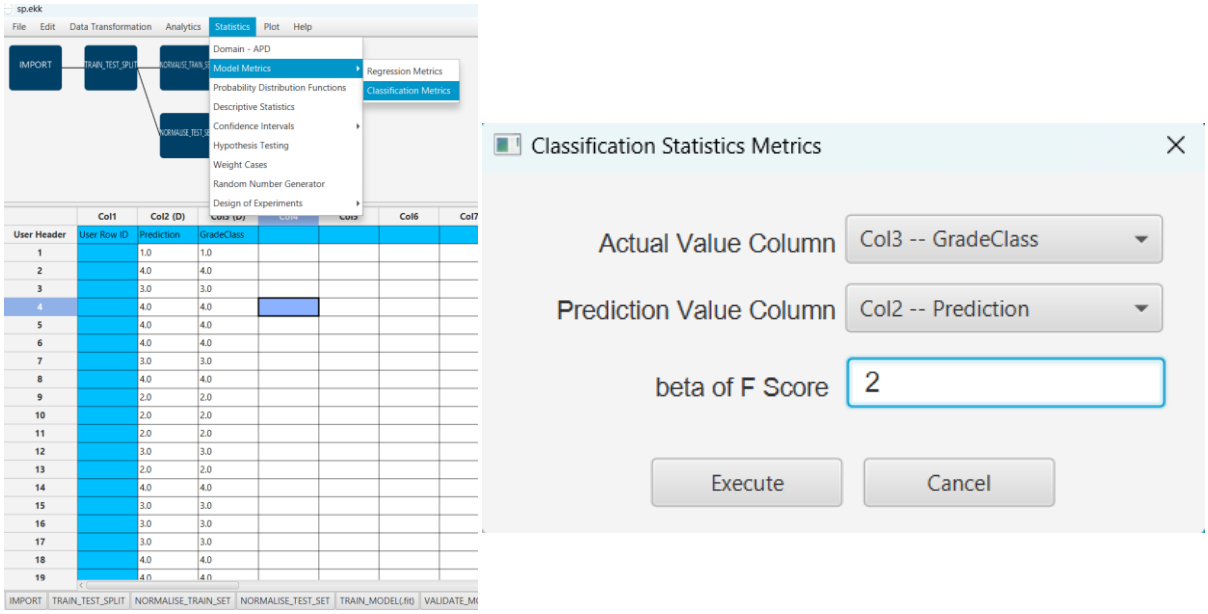
Create a new tab by pressing the "+" button on the bottom of the page with the name "STATISTICS_ACCURACIES".

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:

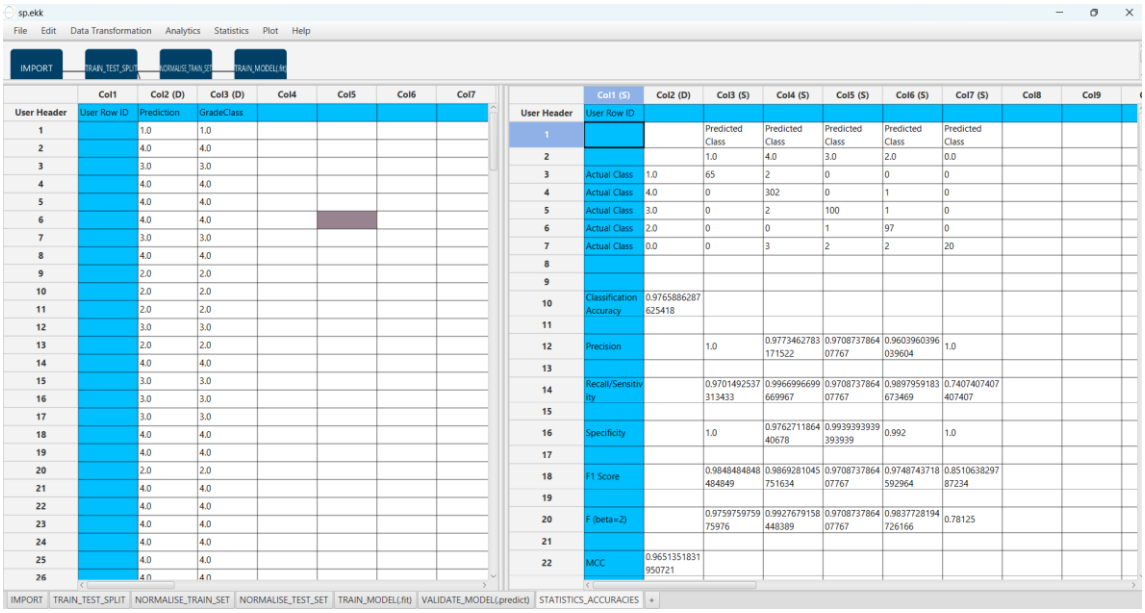
"Statistics" → "Model Metrics" → "Classification Metrics".



The results will appear on the output spreadsheet.

Accuracy: 0.977

F1-Score = 0.954

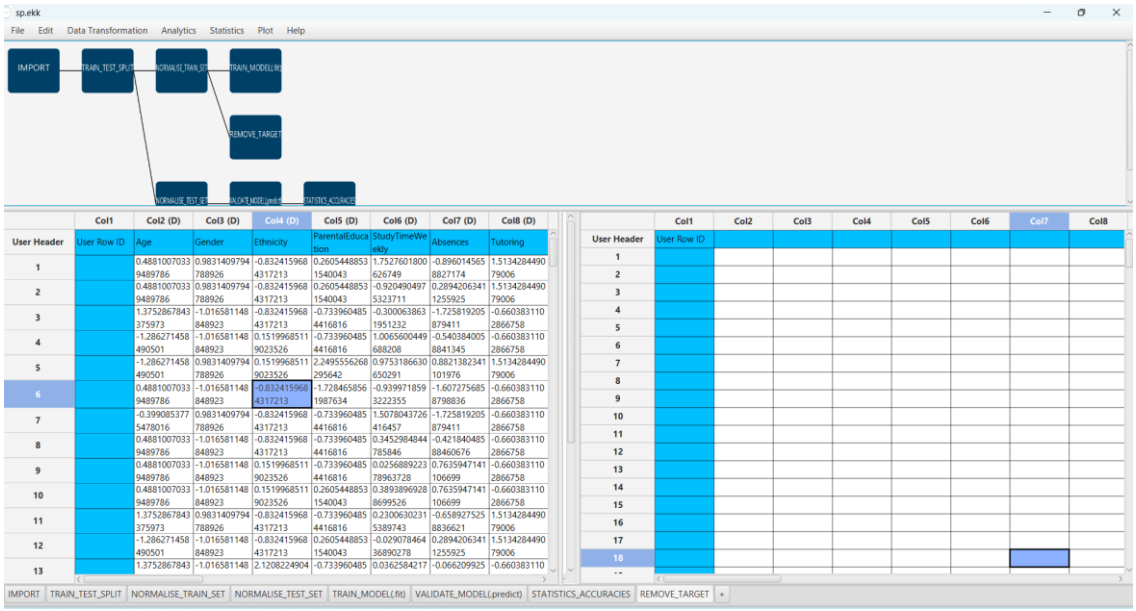


Step 9: Reliability check of each record of the test set

Step 9.a: Create the domain

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

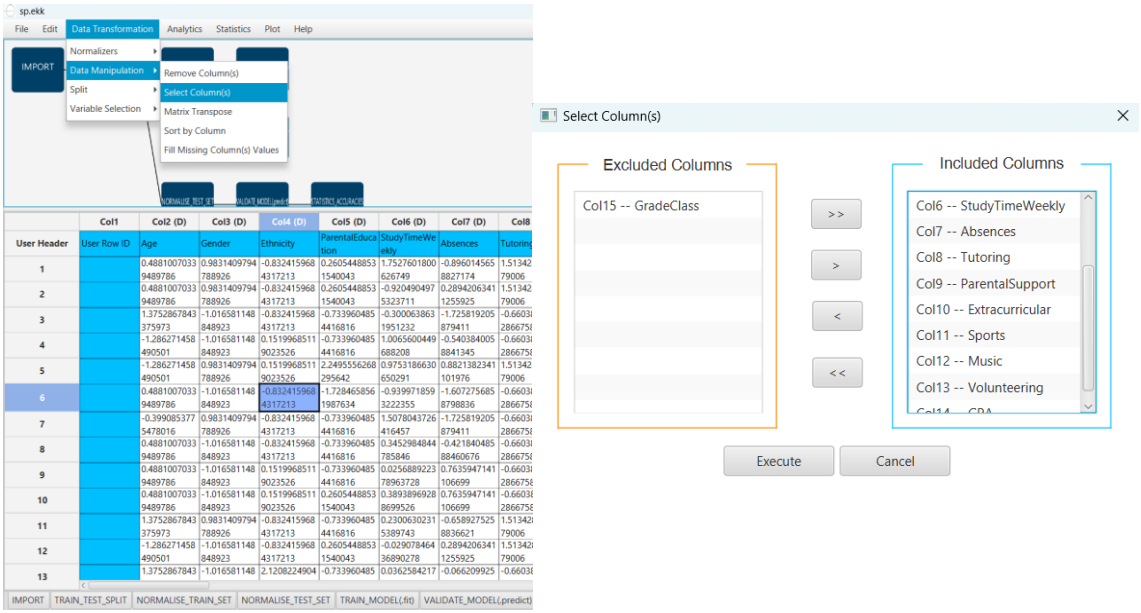
Import data into the input spreadsheet of the "REMOVE_TARGET" tab from the output of the "NORMALISE_TRAIN_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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

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

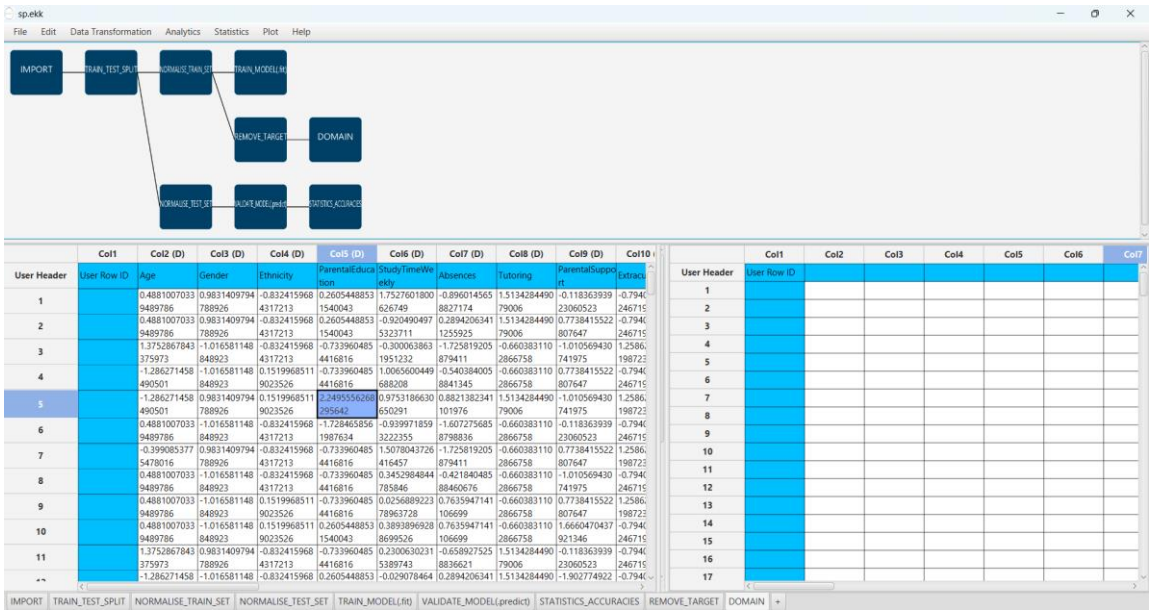
Then select all the columns except the "GradeClass".



The results will appear on the output spreadsheet.

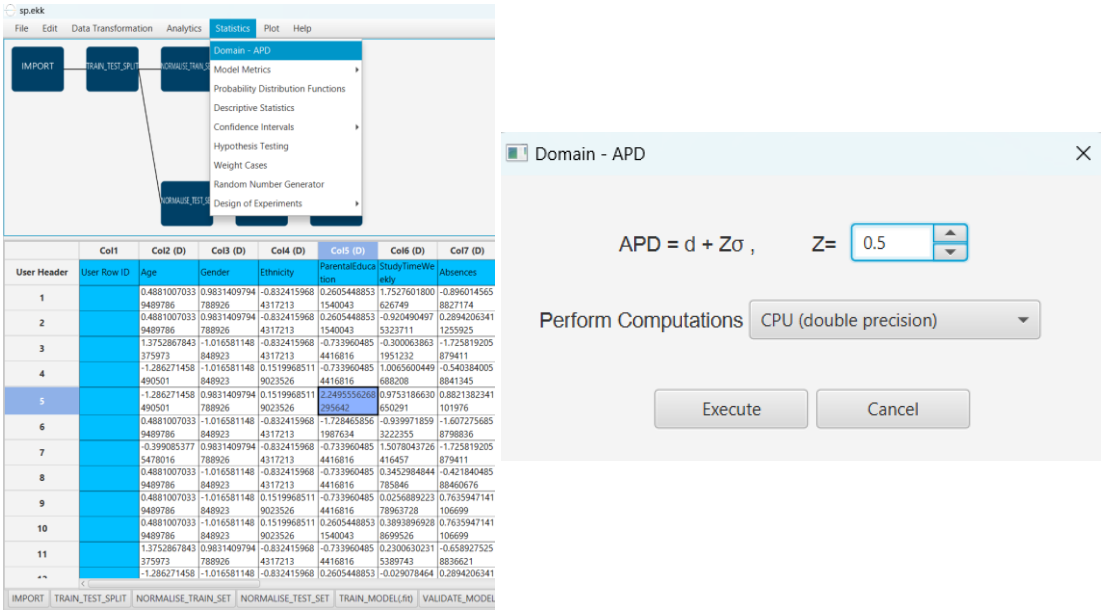
Create a new tab by pressing the "+" button on the bottom of the page with the name "DOMAIN".

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

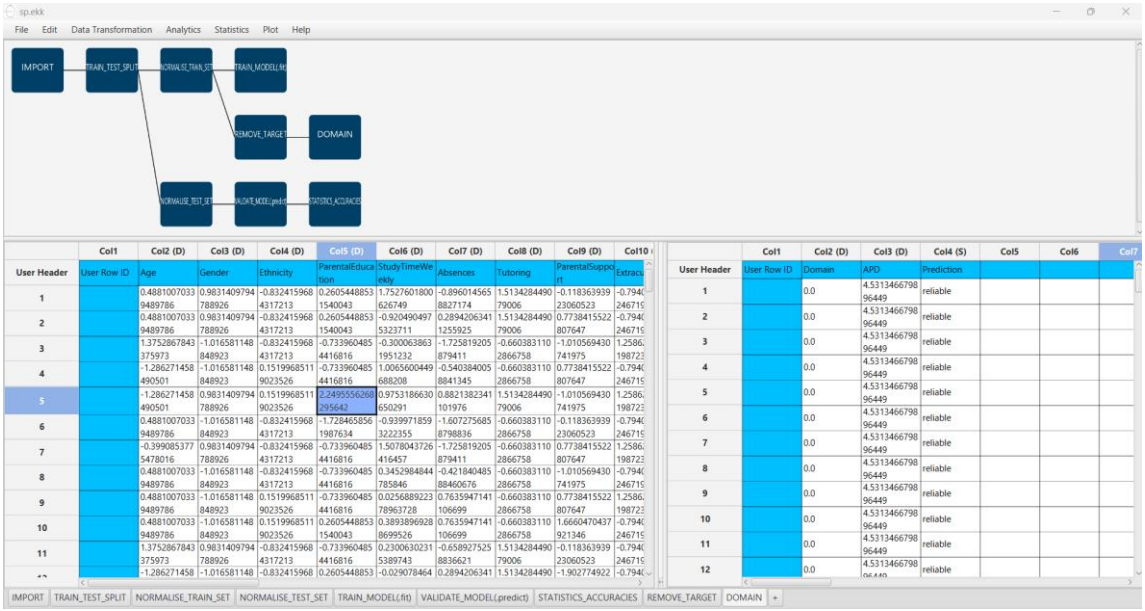


Create the domain:

"Statistics" → "Domain APD"



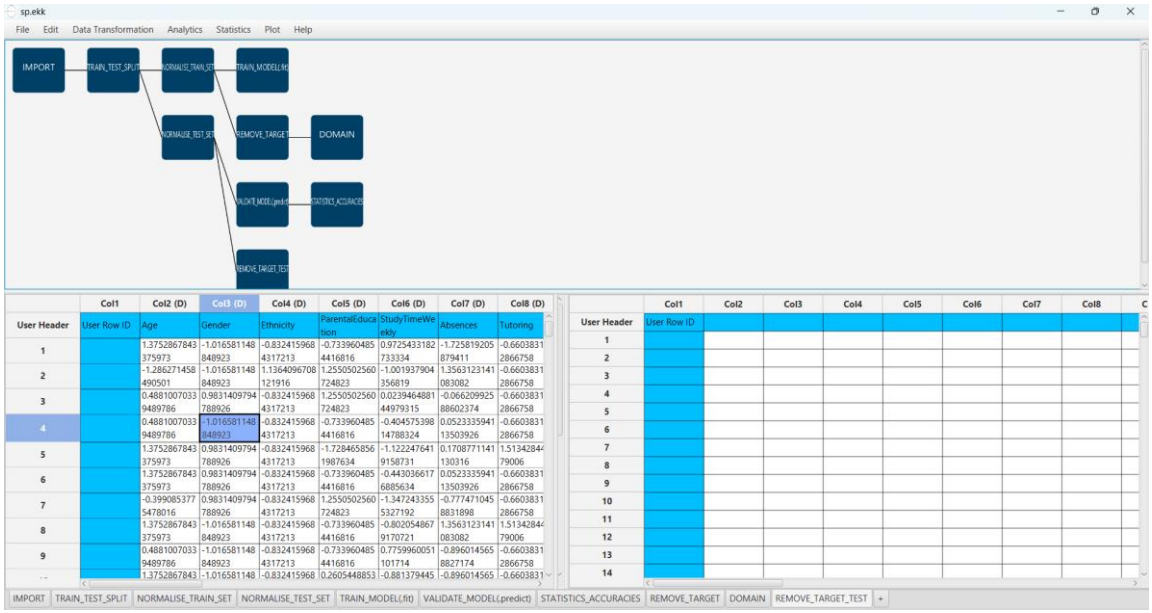
The results will appear on the output spreadsheet.



Step 9.b: Check the test set reliability

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

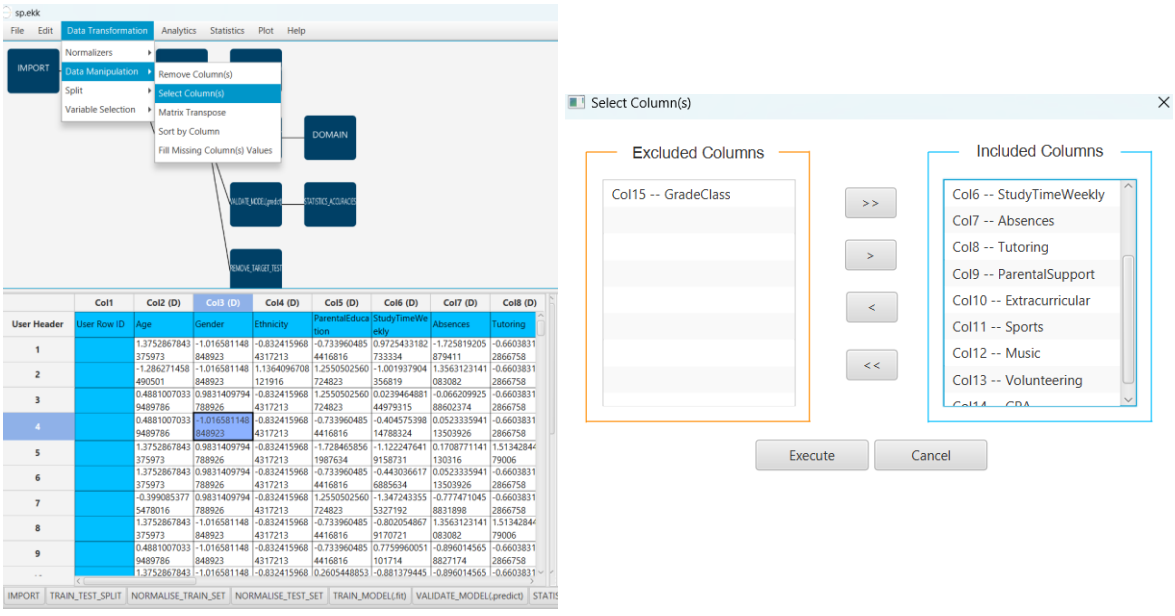
Import data into the input spreadsheet of the "REMOVE_TARGET_TEST" tab from the output of the "NORMALISE_TEST_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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

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

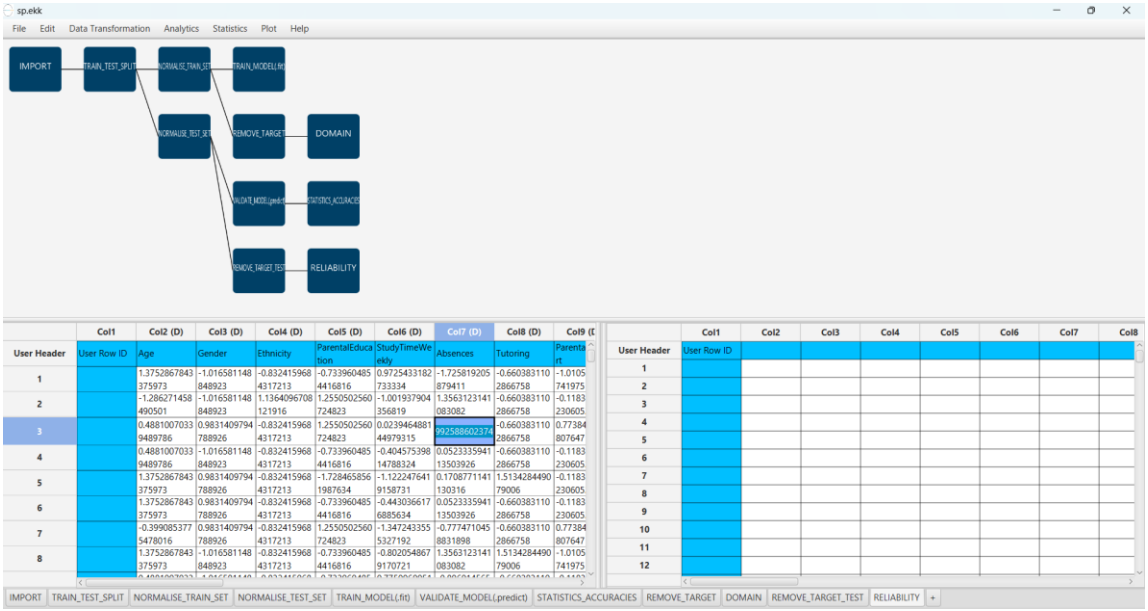
Then select all the columns except "GradeClass".



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".

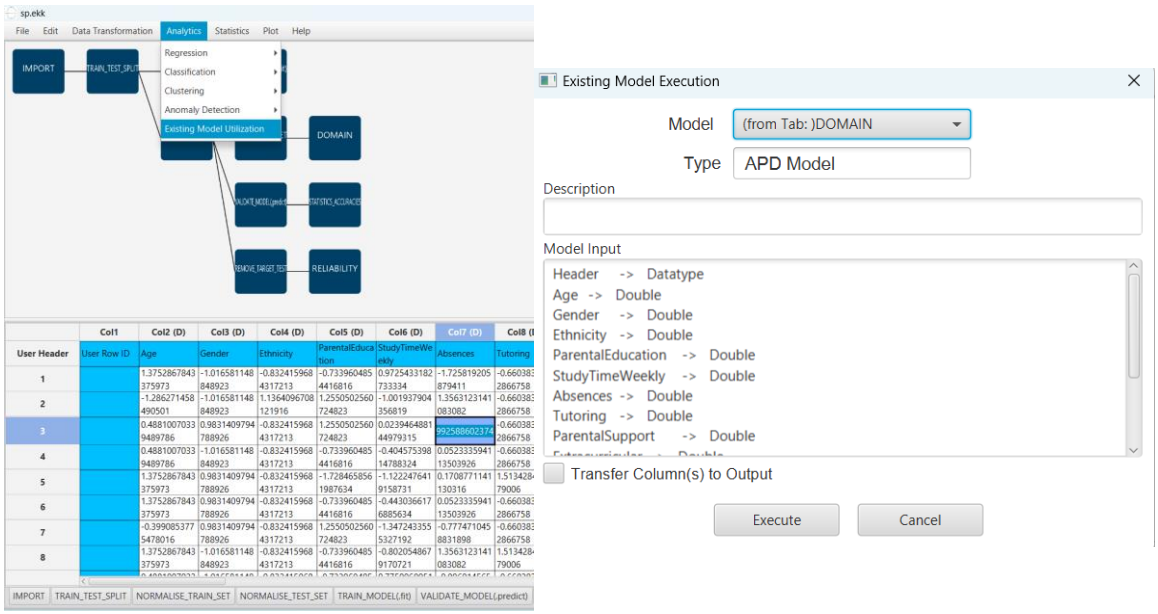
Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "REMOVE_TARGET_TEST" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



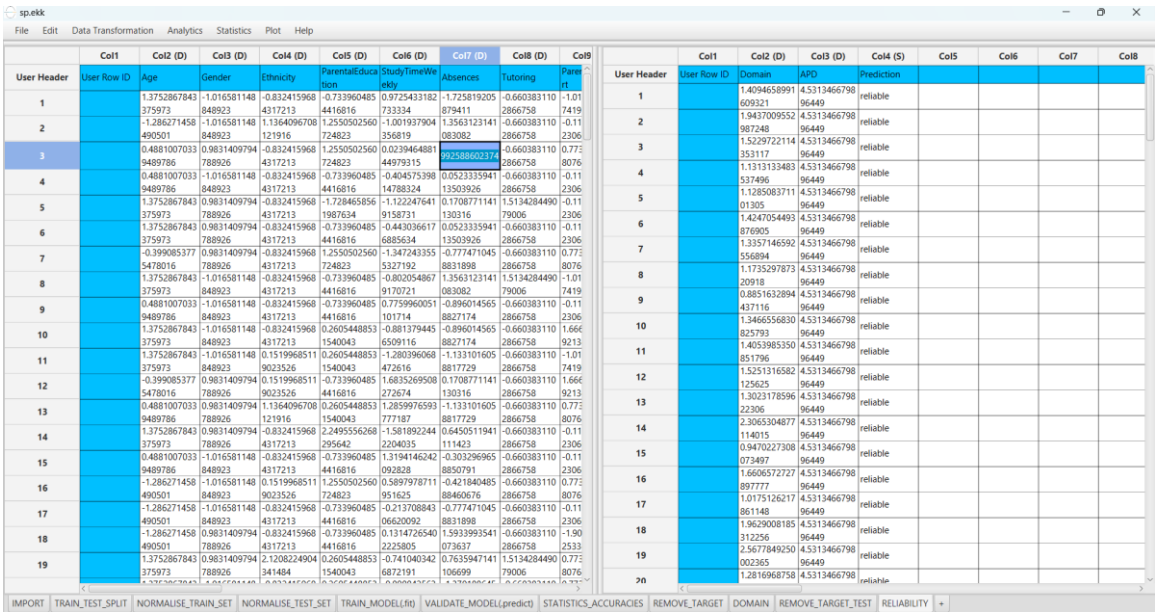
Check the Reliability:

"Analytics" → "Existing Model Utilization".

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



The results will appear on the output spreadsheet.



There are no unreliable samples in the test set.

Final Isalos Workflow

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