



# 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  $\geq 3.5$  (0)
- 'B' -  $3.0 \leq \text{GPA} < 3.5$  (1)
- 'C' -  $2.5 \leq \text{GPA} < 3.0$  (2)
- 'D' -  $2.0 \leq \text{GPA} < 2.5$  (3)
- 'F' -  $\text{GPA} < 2.0$  (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.

User Header	Col1	Col2	Col3	Col4	Col5	Col6
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						

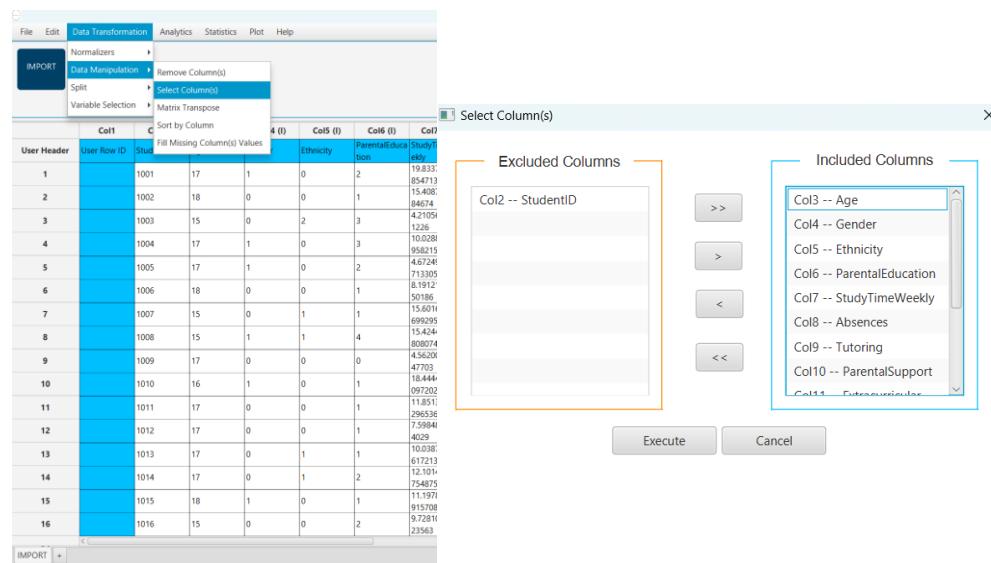
  

User Header	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (I)	Col6 (I)	Col7 (D)	Col8 (I)	Col9 (I)	Col10 (I)	Col11 (I)	Col12 (I)	Col13 (I)
1	1001	17	1	0	2	19.831722807	7	1	2	0	0	0	1
2	1002	18	0	0	1	15.408756055	0	0	1	0	0	0	0
3	1003	15	0	2	3	4.2105697688	26	0	2	0	0	0	0
4	1004	17	1	0	3	10.028829473	14	0	3	1	0	0	0
5	1005	17	1	0	2	4.6724957279	17	1	3	0	0	0	0
6	1006	18	0	0	1	8.9171815452	0	0	1	1	0	0	0
7	1007	15	0	1	1	15.601680474	10	0	3	0	1	0	0
8	1008	15	1	1	4	15.424496305	22	1	1	1	0	0	0
9	1009	17	0	0	0	4.5620075580	1	0	2	0	1	0	0
10	1010	16	1	0	1	18.44466363	0	0	3	1	0	0	0
11	1011	17	0	0	1	11.851363655	11	0	1	0	0	0	0
12	1012	17	0	0	1	26.93348	15	0	2	0	0	0	0
13	1013	17	0	1	1	7.5984858192	15	0	3	1	0	0	0
14	1014	17	0	1	2	10.038711615	21	0	4	0	1	0	0
15	1015	18	1	0	1	12.101425068	21	0	1	2	0	0	0
16	1016	15	0	0	2	9.197810636	9	1	0	0	1	0	0

## 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".

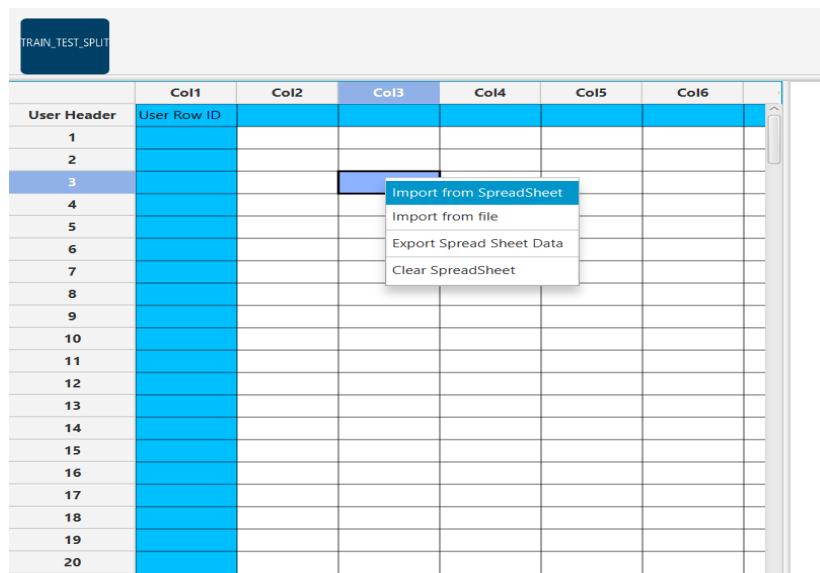


The data without the "StudentID" column will appear in the output spreadsheet.

## 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.

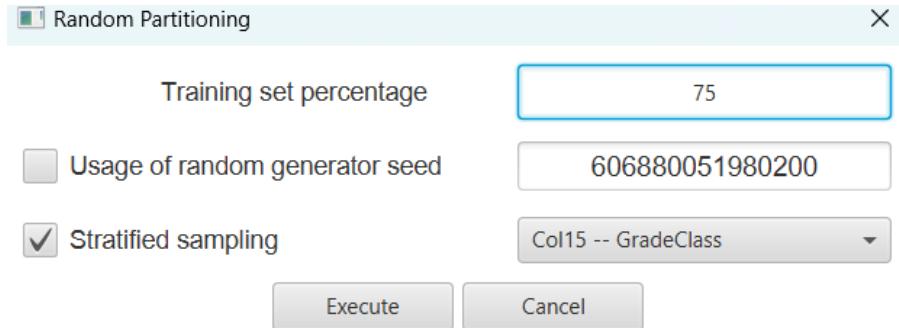
Import data into the input spreadsheet of the "TRAIN\_TEST\_SPLIT" tab from the output of the "IMPORT" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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:



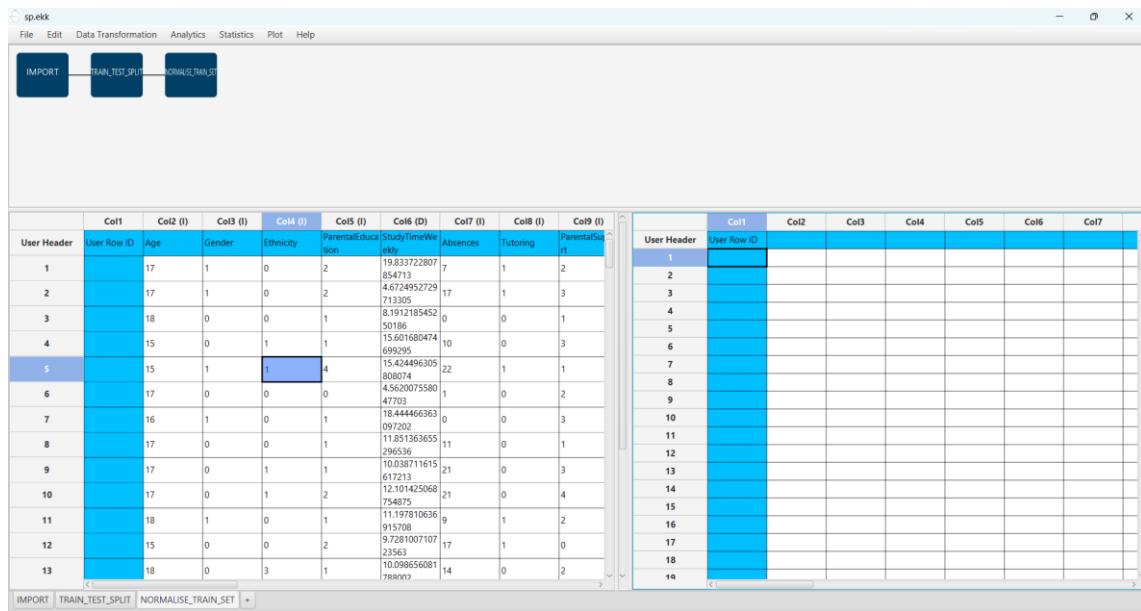
The results will appear on the output spreadsheet.

User Header	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (I)	Col6 (D)	Col7 (I)	Col8 (I)	Col9 (I)	Col10 (I)	Col11 (I)	Col12 (I)	Col13 (I)	Col14 (I)	Col15 (I)	Col16 (D)	Col17 (I)	Col18 (I)	Col19 (I)	Col20 (I)
User Header	User Row ID	Age	Gender	Ethnicity	ParentalEducation	StudyTimeWork	Absences	Tutoring	ParentalSupport	Col10	Col11	Col12	Col13	Col14	Col15	Col16	Col17	Col18	Col19	Col20
1	17	1	0	2	19.633722807 914713	7	1	2												
2	18	0	0	1	15.408756055 84674	0	0	1												
3	15	0	2	3	4.2105697688 1226	26	0	2												
4	17	1	0	3	10.028829473 958215	14	0	3												
5	17	1	0	2	4.6724952729 11300	17	1	3												
6	18	0	0	1	9.0121785452 50186	0	0	1												
7	15	0	1	1	15.601680474 699295	10	0	3												
8	15	1	1	4	15.424496305 808074	22	1	1												
9	17	0	0	0	4.5620075580 47703	1	0	2												
10	16	1	0	1	18.444466363 99709	0	0	3												
11	17	0	0	1	11.651363655 296436	11	0	1												
12	17	0	0	1	7.5984858192 4029	15	0	2												
13	17	0	1	1	10.038711615 617713	21	0	3												

## 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"



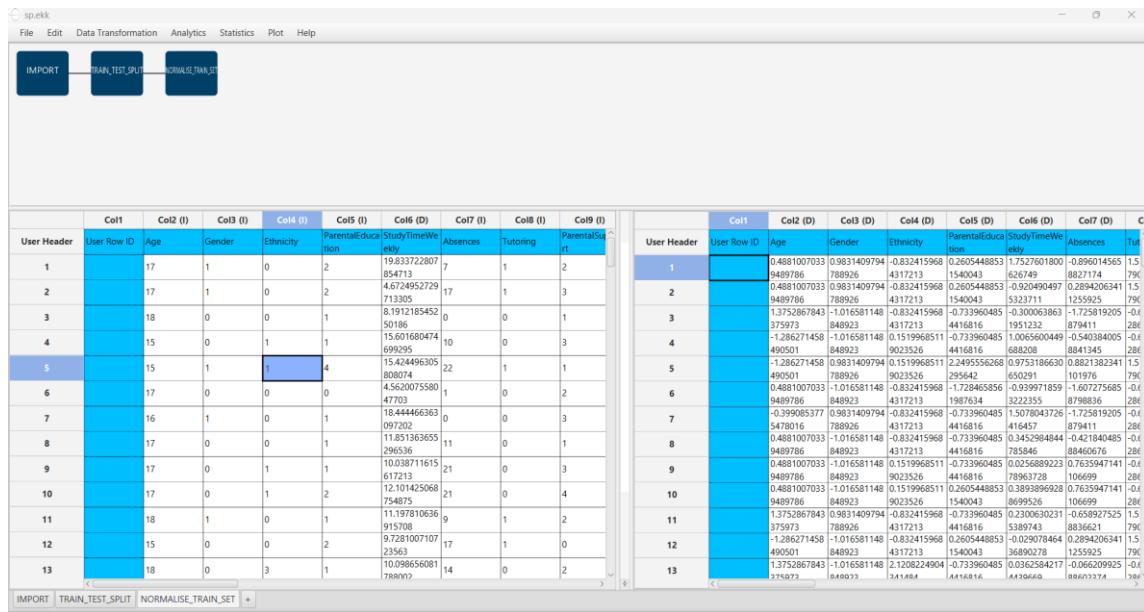
Normalize the data using Z-score:

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

Then select all columns and click "Execute".

The screenshot shows the Isalos Analytics Platform interface with the Data Transformation menu selected. Under the Data Transformation menu, the Normalizers option is highlighted, and the Z-Score option is selected. A dialog box titled "ZScore Normalizer" is open, showing two lists of columns: Excluded Columns and Included Columns. The Excluded Columns list contains "Col6 -- GradeClass". The Included Columns list contains "Col6 -- StudyTimeWeekly", "Col7 -- Absences", "Col8 -- Tutoring", "Col9 -- ParentalSupport", "Col10 -- Extracurricular", "Col11 -- Sports", "Col12 -- Music", "Col13 -- Volunteering", and "Col14 -- GPA". At the bottom of the dialog box are "Execute" and "Cancel" buttons. In the background, the main spreadsheet view is visible, showing the same 14 rows and 10 columns as the previous screenshot.

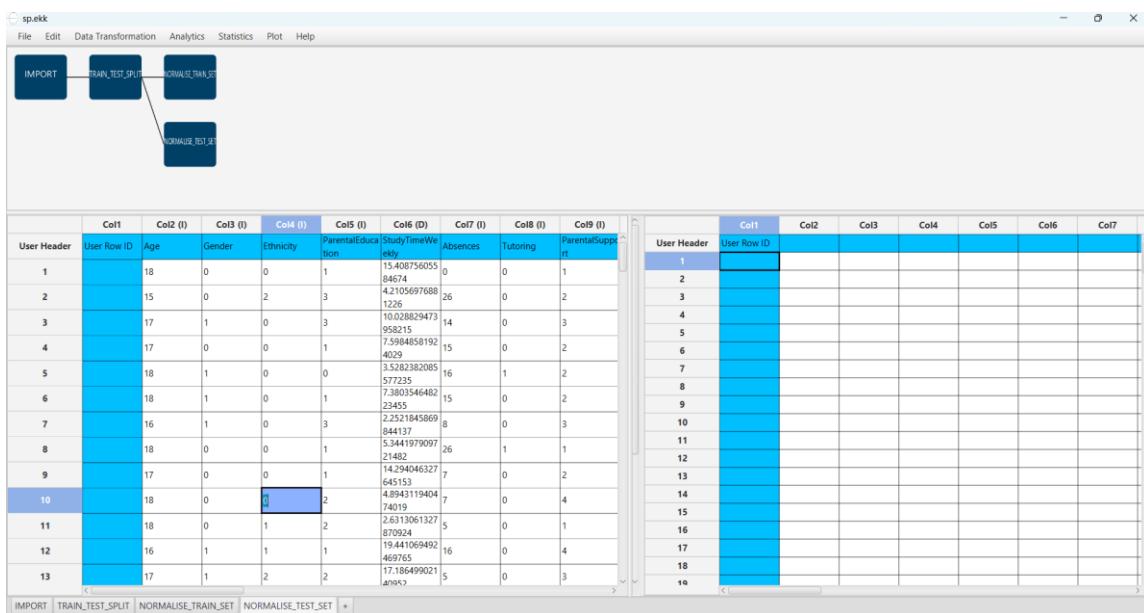
The results will appear on the output spreadsheet.



## 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 screenshot shows the Isalos Analytics Platform interface. In the top navigation bar, the 'Analytics' tab is selected. A context menu is open over a data transformation node, with 'Existing Model Utilization' highlighted. Below the menu, a data spreadsheet is visible with columns labeled Col1 through Col6. To the right, an 'Existing Model Execution' dialog box is open. Inside the dialog, the 'Model' dropdown is set to '(from Tab:) NORMALISE\_TRAIN\_SET' and the 'Type' is 'Z Score Normalizer Model'. The 'Model Input' section contains a detailed list of column mappings:

- User Header → Datatype
- Age → Double
- Gender → Double
- Ethnicity → Double
- ParentalEducation → Double
- StudyTimeWeekly → Double
- Absences → Double
- Tutoring → Double
- ParentalSupport → Double

At the bottom of the dialog are 'Execute' and 'Cancel' buttons.

The results will appear on the output spreadsheet.

The screenshot shows the Isalos Analytics Platform interface with a new tab named 'NORMALISE\_TRAIN\_SET' added to the bottom. The main area displays a data spreadsheet with columns Col1 through Col7. The data in the spreadsheet is identical to the input spreadsheet shown in the previous screenshot, but the values have been normalized according to the Z Score Normalizer Model. The columns are labeled User Header, User Row ID, Age, Gender, Ethnicity, ParentalEducation, StudyTimeWeekly, Absences, Tutoring, and ParentalSupport.

## 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".

The screenshot shows the Isalos Analytics Platform interface. At the top, there's a navigation bar with 'File', 'Edit', 'Data Transformation', 'Analytics', 'Statistics', 'Plot', and 'Help'. Below the navigation bar is a toolbar with icons for 'IMPORT', 'TRAIN\_TEST\_SPLIT', 'NORMALISE\_TRAIN\_SET', and 'TRAIN\_MODEL(R0)'. A data flow diagram is displayed, starting with 'IMPORT' followed by 'TRAIN\_TEST\_SPLIT', then branching into 'NORMALISE\_TRAIN\_SET' and 'TRAIN\_MODEL(R0)'. Below the diagram are two spreadsheets. The left spreadsheet has columns labeled 'User Header', 'Col1' through 'Col9 (I)', and rows numbered 1 to 13. The right spreadsheet has columns labeled 'User Header', 'Col1' through 'Col7', and rows numbered 1 to 19. Both spreadsheets contain numerical data corresponding to the rows and columns defined.

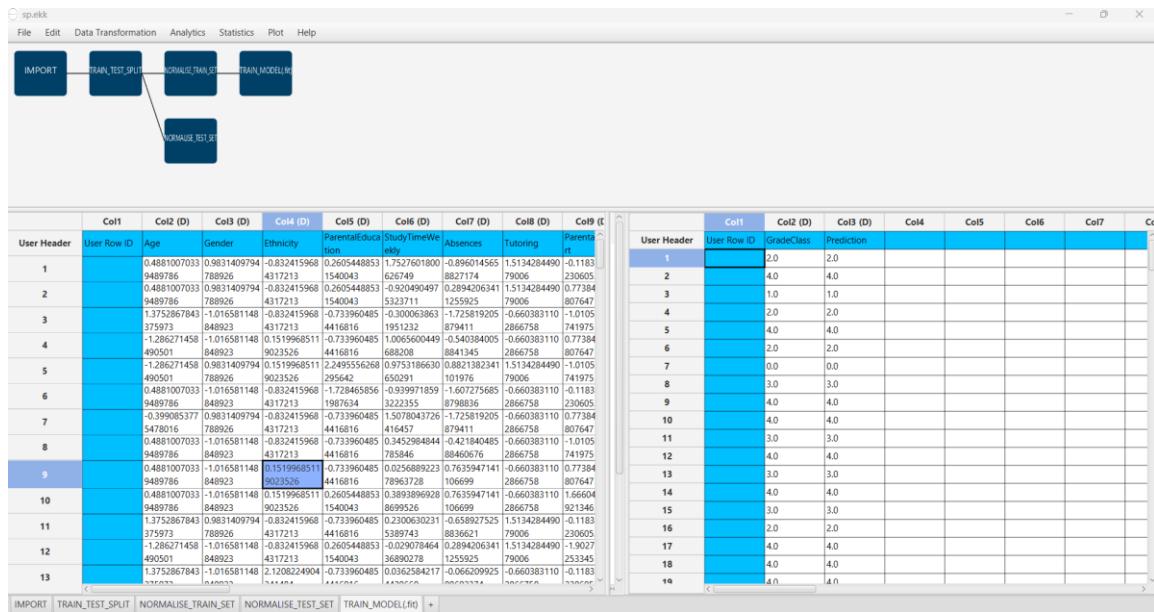
Use the XGBoost Method to train and fit the model:

"Analytics" → "Classification" → "XGBoost"

and adjust model parameters based on training set performance.

The screenshot shows the Isalos Analytics Platform interface with the 'Analytics' tab selected. Under the 'Classification' section, 'XGBoost' is highlighted. To the right, a detailed configuration window for the 'XGBoost Classification Model' is open. It includes fields for 'Target Column' (set to 'Col15 -- GradeClass'), 'booster' (set to 'gbtree'), 'objective' (set to 'multisoftprob'), 'number of estimators' (set to 200), 'eta' (set to 0.1), 'gamma' (set to 'Double [0, +∞), Default 0'), 'max depth' (set to 7), 'min child weight' (set to 'Double [0, +∞), Default 1'), 'column sample by tree' (set to 10), 'sub sample' (set to 'Double [0, 1], Default 1'), 'tree method' (set to 'default'), 'lambda' (set to 'Double (-∞, +∞), Default 1'), and 'alpha' (set to 'Double (-∞, +∞), Default 0'). A checkbox for 'Time-based RNG Seed' is checked, and the 'RNG Seed' field is set to 'Double (-∞, +∞), Default:'. At the bottom are 'Execute' and 'Cancel' buttons. Below the configuration window, the same data pipeline and spreadsheets are visible as in the first screenshot.

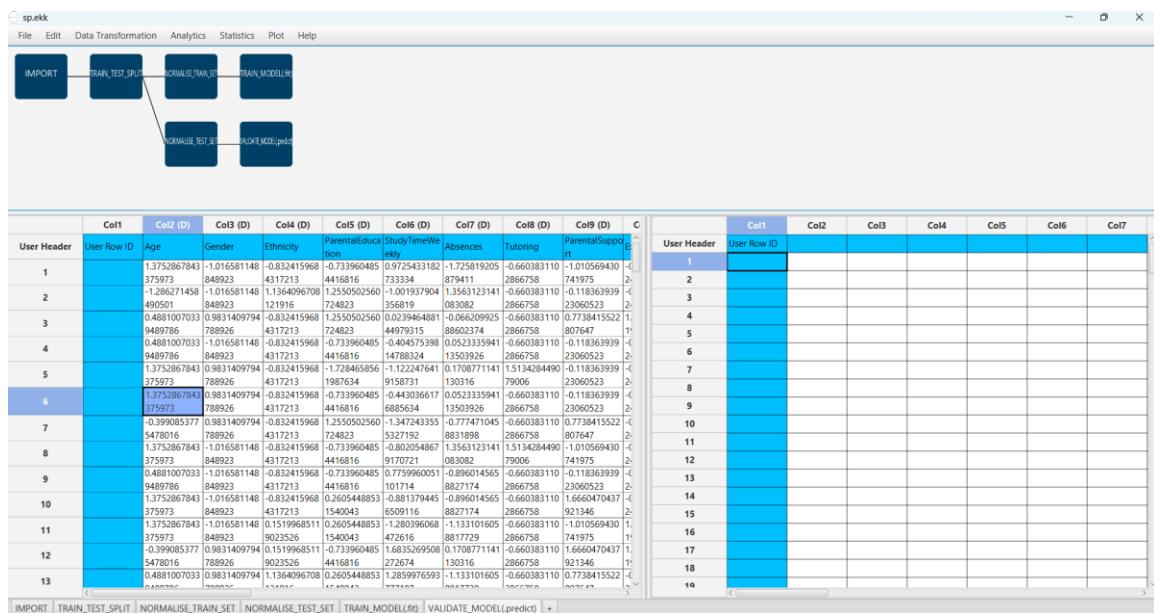
The predictions will appear on the output spreadsheet.



## 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)".

User Row ID	Age	Gender	Ethnicity	ParentalEducation	StudyTimeWeekly	Absences	Tutoring	ParentalSupport	Col6	Col7	Col8	Col9	Col10
1	1.3752867843	-1.016581148	-0.832415968	-0.733960485	0.9725433182	1.72519205	-0.660383110	-1.01					
2	-1.286271458	-1.016581148	1.1364096708	1.2550502560	-1.001937904	1.3563123141	-0.660383110	-0.11					
3	49050	121916	724823	356819	0.93082	2866758	2306						
4	0.488100703	0.9831409794	-0.832415968	1.2550502560	0.0239464881	-0.066029998	-0.660383110	0.773					
5	0.489786	4317213	724823	44979313	0.8602373	2866758	8076						
6	0.489786	4317213	724823	44979313	0.8602373	2866758	2306						
7	375973	788926	4317213	1987634	0.9518711	1.30316	79006	2306					
8	-0.390085377	0.9831409794	-0.832415968	-0.733960485	-1.442036617	0.0523335941	-0.660383110	-0.11					
9	1.3752867843	-1.016581148	-0.832415968	-1.728465856	-1.12247641	0.1738971141	1.5134284490	1.11					
10	375973	788926	4317213	1987634	0.9518711	1.30316	79006	2306					
11	1.3752867843	-1.016581148	-0.832415968	-0.733960485	-0.832415968	0.2605448851	-0.091164464	-0.443026617	-0.0523335941	-0.660383110	1.666	2913	
12	375973	788926	4317213	1987634	0.9518711	1.30316	79006	2306					
13	49050	121916	724823	44979313	0.8602373	2866758	2306						

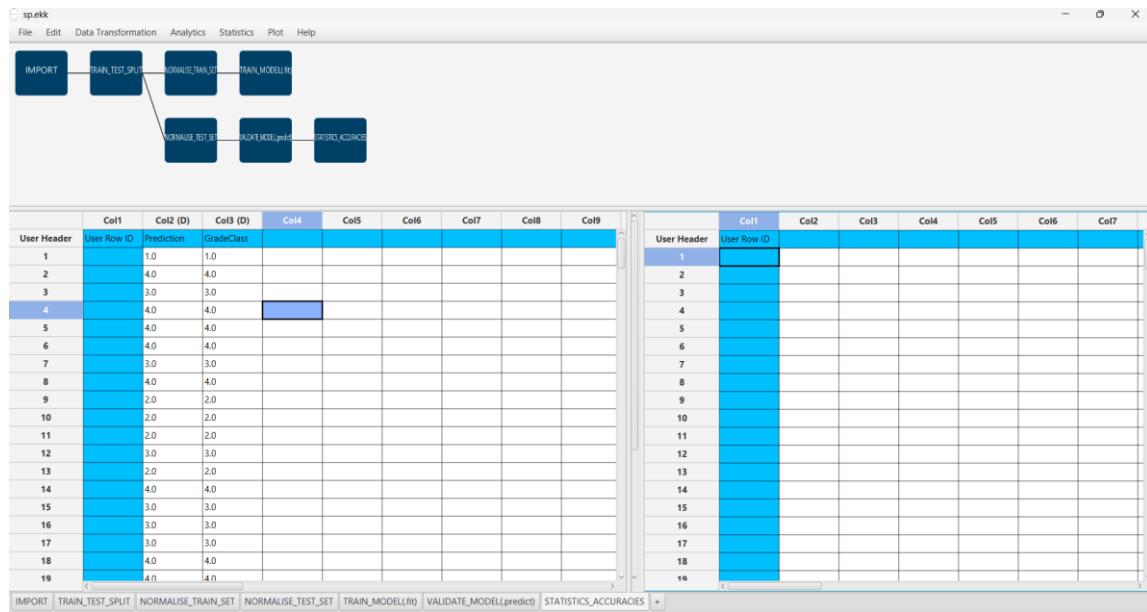
The predictions will appear on the output spreadsheet.

User Row ID	Age	Gender	Ethnicity	ParentalEducation	StudyTimeWeekly	Absences	Tutoring	ParentalSupport	Col6	Col7	Col8	Col9	Col10
1	1.0	1.0											
2	4.0	4.0											
3	3.0	3.0											
4	4.0	4.0											
5	4.0	4.0											
6	4.0	4.0											
7	3.0	3.0											
8	4.0	4.0											
9	2.0	2.0											
10	2.0	2.0											
11	2.0	2.0											
12	3.0	3.0											
13	4.0	4.0											

## 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 screenshot shows the Isalos Analytics Platform interface. The menu bar is visible at the top. A context menu is open over the data flow diagram, with "Classification Metrics" selected. A dialog box titled "Classification Statistics Metrics" is displayed on the right side of the screen. The dialog contains the following fields:

- Actual Value Column: Col3 -- GradeClass
- Prediction Value Column: Col2 -- Prediction
- beta of F Score: 2

At the bottom of the dialog are "Execute" and "Cancel" buttons.

The results will appear on the output spreadsheet.

Accuracy: 0.977

F1-Score = 0.954

The screenshot shows the Isalos Analytics Platform interface. At the top, there's a menu bar with File, Edit, Data Transformation, Analytics, Statistics, Plot, and Help. Below the menu is a toolbar with buttons for IMPORT, TRAIN\_TEST\_SPLIT, NORMALISE\_TRAIN\_SET, and TRAIN\_MODEL. The main area has two tabs: 'User Header' and 'User Row ID'. The 'User Header' tab displays a table with columns Col1 through Col7. The 'User Row ID' tab displays a table with columns Col1 through Col9, including headers for Predicted Class and Classification Accuracy.

## 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".

The screenshot shows the Isalos Analytics Platform interface with a flowchart at the top. The flowchart consists of several nodes connected by arrows: IMPORT, TRAIN\_TEST\_SPLIT, NORMALISE\_TRAIN\_SET, TRAIN\_MODEL, REMOVE\_TARGET, and NORMALISE\_TEST\_SET. The REMOVE\_TARGET node is highlighted. Below the flowchart is a table with columns Col1 through Col8. The table contains data for multiple users, with User Row ID as the primary key.

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:

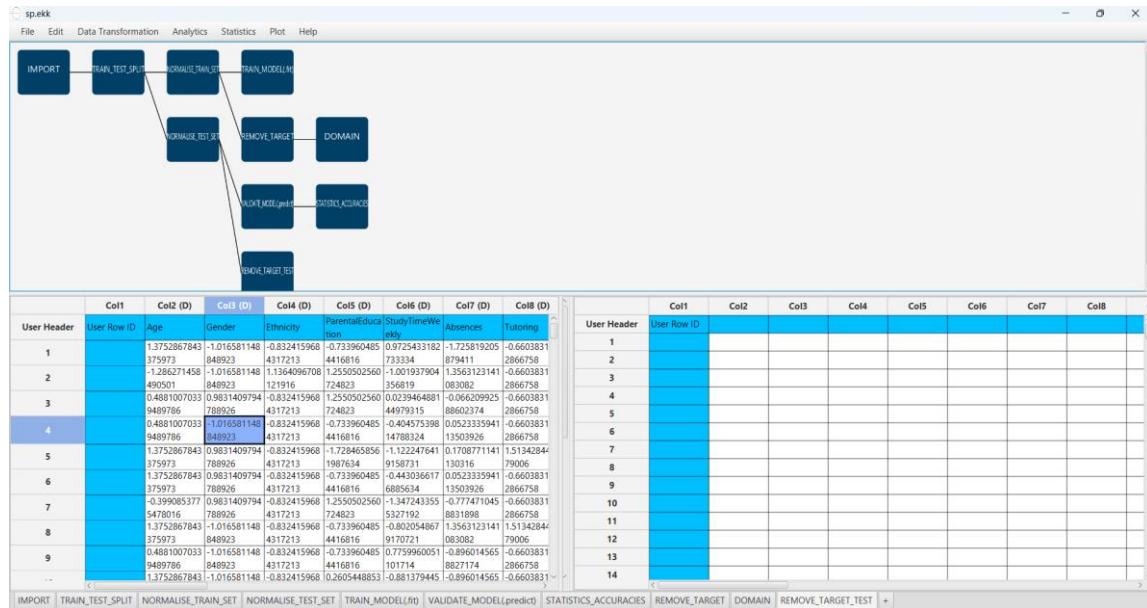
"Statistics" → "Domain APD"

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".

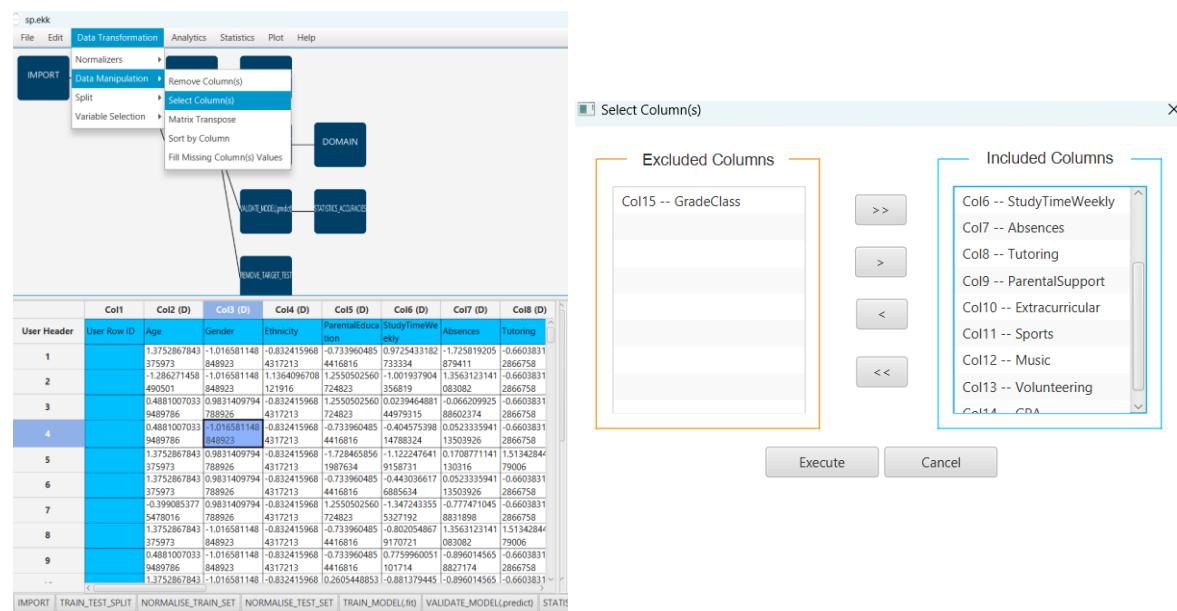
Import data into the input spreadsheet of the "REMOVE\_TARGET\_TEST" tab from the output of the "NORMALISE\_TEST\_SET" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



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

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

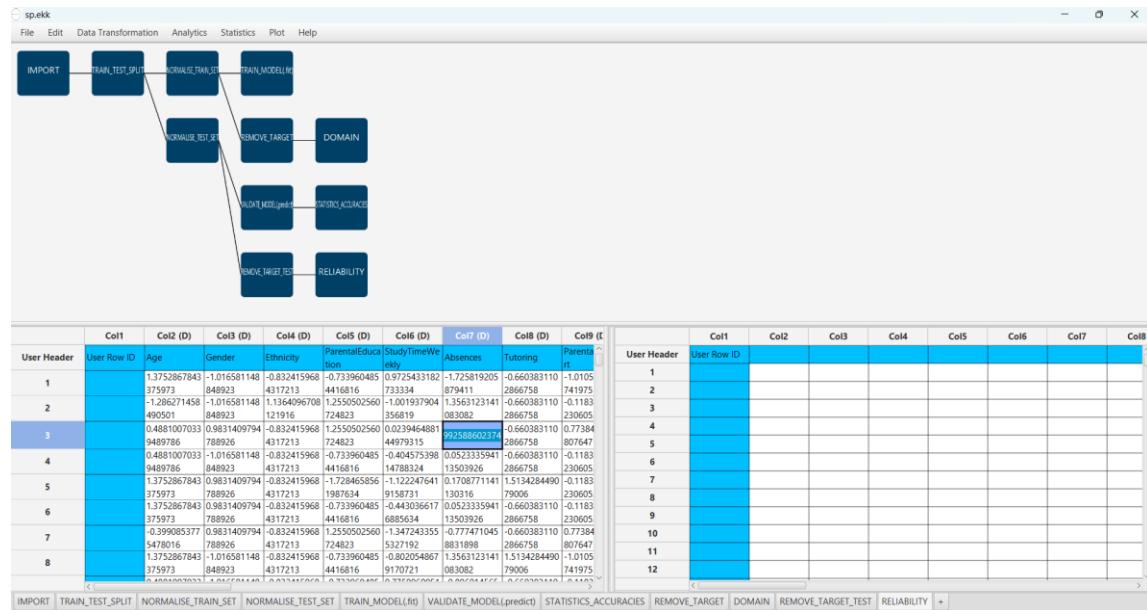
Then select all the columns except "GradeClass".



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".

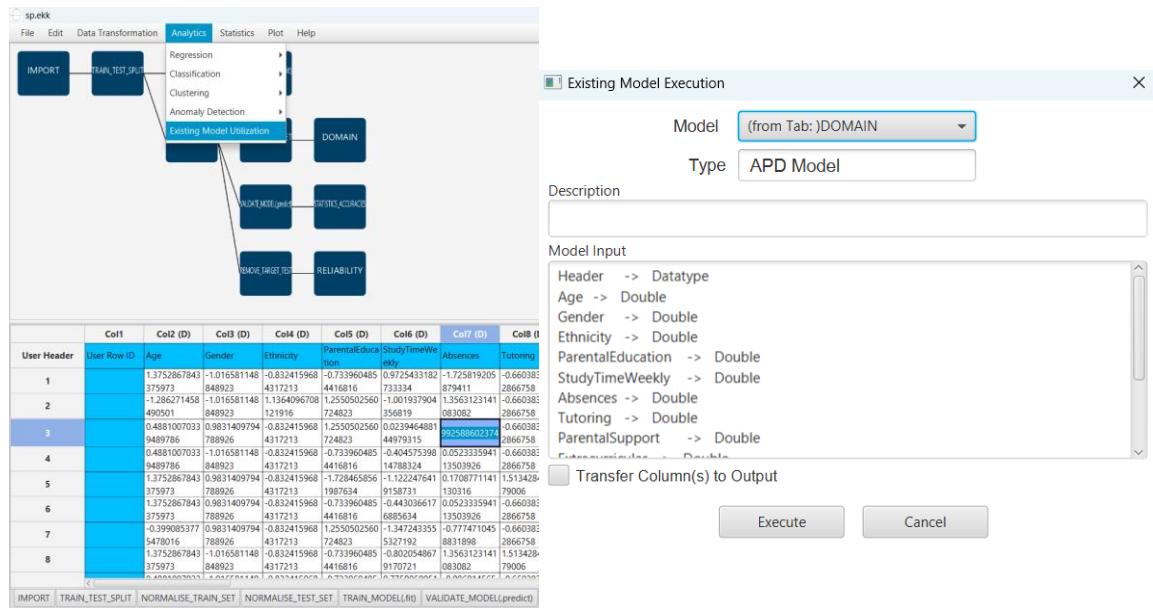
Import data into the input spreadsheet of the "RELIABILITY" tab from the output of the "REMOVE\_TARGET\_TEST" tab by right-clicking on the input spreadsheet and then choosing "Import from SpreadSheet".



Check the Reliability:

"Analytics" → "Existing Model Utilization".

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



The results will appear on the output spreadsheet.

sp.ekk									
	Col1	Col2 (D)	Col3 (D)	Co4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9
User Header	User Row ID	Age	Gender	Ethnicity	ParentEducationLevel	StudyTimeWeek	Abseances	TutoringRate	ParentIncome
1	1.3752867843	1.016581148	-0.832415968	-0.733960485	0.972533182	-1.725819205	-0.660383110	1.01	375973
2	375973	848923	4317213	4416816	733334	879411	2866758	7419	-1.286271458
3	490501	848923	121916	724823	356819	083082	2866758	2306	0.4881007033
4	548786	1.016581148	-0.832415968	1.25505256	0.023964881	2515860374	-0.660383110	0.773	0.4881007033
5	348976	848923	4317213	4416816	14788324	13503926	2866758	8076	1.3752867843
6	375973	788926	4317213	1987624	9158731	130216	79006	2306	1.3752867843
7	375973	788926	4317213	4416816	6885634	13503926	2866758	2306	1.3752867843
8	5478016	788926	4317213	724823	5327193	883180	2866758	8076	1.3752867843
9	375973	848923	4317213	4416816	9170271	083082	79006	7419	0.4881007033
10	375973	788926	4317213	150403	6509116	8821714	2866758	0213	1.3752867843
11	375973	848923	1.016581148	-0.832415968	0.260548853	0.8817945	-0.660383110	1.666	0.390905377
12	5478016	788926	4317213	150403	472616	8817229	2866758	7419	0.9831409794
13	375973	788926	4317213	121916	777177	1.133101605	-0.660383110	0.773	0.4881007033
14	375973	788926	4317213	295642	220403	111423	2866758	2306	1.3752867843
15	375973	788926	4317213	150403	1.280396069	-1.33101605	-0.660383110	1.01	1.3752867843
16	375973	848923	4317213	150403	472616	8817229	2866758	7419	-1.286271458
17	490501	848923	9023526	724823	051625	88460676	2866758	8076	0.4881007033
18	375973	788926	4317213	4416816	06620092	883180	2866758	2306	-1.286271458
19	490501	788926	4317213	4416816	222580	073637	2866758	2533	0.4881007033
20	375973	788926	341484	150403	6872191	106699	79006	8076	1.3752867843

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:

