



# Credit card Details Binary Classification Problem

This dataset, which can be found in [https://www.kaggle.com/datasets/rohitudageri/credit-card-details?select=Credit\\_card\\_label.csv](https://www.kaggle.com/datasets/rohitudageri/credit-card-details?select=Credit_card_label.csv), contains information about credit card applicants. It consists of 53 features and 1548 samples and it is used to predict whether an application for a credit card will be approved or denied. The dataset contains some categorical features which are encoded outside of Isalos using Python code.

*Isalos version used: 2.0.6*

## Step 1: Import data from file

Right click on the input spreadsheet (left) and choose the option “Import from File”. Then navigate through your files to load the one with the credit card data.

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

The data will appear on the left spreadsheet.

	Col1	Col2 (S)	Col3 (S)	Col4 (S)	Col5 (I)	Col6 (D)	Col7 (S)	Col8 (S)	Col9 (S)	Col10 (S)	Col11 (D)	Col12 (I)	Col13 (I)	Col14 (I)	Col15 (I)	Col16 (I)	Col17 (S)	Col18 (I)	Col19 (I)
User Header	User Row ID	GENDER	Car_Owner	Propert_Owner	CHILDREN	Annual_Income	Type_Income	EDUCATION	Marital_status	Housing_type	Birthday_count	Employed_days	Mobile_phone	Work_Phone	Phone	EMAIL_ID	Type_Occupation	Family_Members	label
1	5008627	M	Y	Y	0	180000.0	Pensioner	Higher education	Married	House / apartment	-18772.0	365243	1	0	0	0		2	1
2	5009744	F	Y	N	0	315000.0	Commercial associate	Higher education	Married	House / apartment	-13557.0	-586	1	1	1	0		2	1
3	5009746	F	Y	N	0	315000.0	Commercial associate	Higher education	Married	House / apartment		-586	1	1	1	0		2	1
4	5009749	F	Y	N	0		Commercial associate	Higher education	Married	House / apartment	-13557.0	-586	1	1	1	0		2	1
5	5009752	F	Y	N	0	315000.0	Commercial associate	Higher education	Married	House / apartment	-13557.0	-586	1	1	1	0		2	1
6	5009753		Y	N	0	315000.0	Pensioner	Higher education	Married	House / apartment	-13557.0	-586	1	1	1	0		2	1
7	5009754	F	Y	N	0	315000.0	Commercial associate	Higher education	Married	House / apartment	-13557.0	-586	1	1	1	0		2	1
8	5009894	F	N	N	0	180000.0	Pensioner	Secondary / secondary special	Married	House / apartment	-22134.0	365243	1	0	0	0		2	1
9	5010864	M	Y	Y	1	450000.0	Commercial associate	Secondary / secondary special	Married	House / apartment	-18173.0	-678	1	0	1	1	Core staff	3	1
10	5010868	M	Y	Y	1	450000.0	Pensioner	Secondary / secondary special	Married	House / apartment	-18173.0	-678	1	0	1	1	Core staff	3	1
11	5010869	M	Y	Y	1	450000.0	Commercial associate	Secondary / secondary special	Single / not married	House / apartment	-18173.0	-678	1	0	1	1	Core staff	1	1
12	5016498	F	Y	Y	0	90000.0	Working	Secondary / secondary special	Married	House / apartment	-18950.0	-1002	1	1	1	0	Cooking staff	2	1
13	5016501	F	Y	Y	0		Working	Secondary / secondary special	Married	House / apartment	-18950.0	-1002	1	1	1	0	Cooking staff	2	1
14	5016503	F	Y	Y	0	90000.0	Working	Secondary / secondary special	Married	House / apartment	-18950.0	-1002	1	1	1	0	Cooking staff	2	1
15	5021303	M	N	N	1	472500.0	Pensioner	Higher education	Married	With parents	-8907.0	-913	1	0	0	1		3	1

## Step 2: Manipulate data

We can select all the columns to be used. On the menu click on **Data Transformation** → **Data Manipulation** → **Select Column(s)** and select all columns.

The screenshot shows the 'Data Manipulation' menu with the following options: Normalizers, Data Manipulation (selected), Split, and Variable Selection. The 'Data Manipulation' submenu is open, showing: Remove Column(s), Select Column(s) (selected), Matrix Transpose, Wide to Long Format, Sort by Column, and Fill Missing Column(s) Values. The 'Select Column(s)' dialog box is displayed, showing 'Excluded Columns' (empty) and 'Included Columns' (listing all columns from Col2 to Col10). The 'Execute' button is highlighted.

All of the data will appear in the output (right) spreadsheet. This tab can be renamed “IMPORT” by right-clicking on it and choosing the “Rename” option.

The 'Rename Tab' dialog box is shown with the text 'IMPORT' entered in the input field. The 'OK' button is highlighted.

## Step 3: Fill missing data

This dataset includes some missing values therefore we need to fill them appropriately. Create a new tab by pressing the “+” button on the bottom of the page with the name “FILL\_MISSING” which we will use for filling the missing values.

Import data into the input spreadsheet of the “FILL\_MISSING” tab from the output of the “IMPORT” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

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

To fill the missing data, choose: *Data Transformation* → *Data Manipulation* → *Fill Missing Column(s) Values*. Choose to fill the numerical features with their mean values and the categorical features with their most common values.

The screenshot shows the 'Fill Missing Column(s) Values' dialog box. On the left, a menu is open with 'Data Manipulation' selected, and 'Fill Missing Column(s) Values' is highlighted. The dialog box has two main sections: 'Excluded Columns' (empty) and 'Included Columns' (containing a list of columns). Below these are settings for 'Numerical Method' (Mean), 'Span' (Integer (0,+∞), Default: -), and 'Categorical Method' (Most Common Value). At the bottom are 'Execute' and 'Cancel' buttons.

The results will appear on the output spreadsheet.

Export the results in a csv file by right-clicking on the output spreadsheet and selecting “Export Spreadsheet Data”.

The screenshot shows a spreadsheet with the following data:

	Col1	Col2 (S)	Col3 (S)	Col4 (S)	Col5 (I)
User Header	User Row ID	GENDER	Car_Owner	Propert_Owner	CHILDREN
1	5008827	M			
2	5009744	F			
3	5009746	F			
4	5009749	F	Y	N	0

A context menu is open over the spreadsheet with the following options:

- Adjust Spreadsheet Precision
- Export Spreadsheet Data
- Clear Spreadsheet

The 'File preferences' dialog box is open on the right with the following settings:

- File Extension: CSV
- Select Separator: ,
- User Header: ☒
- User Row ID: ☒
- Buttons: Execute, Cancel

Afterwards, apply one-hot encoding on the categorical features of the dataset, which are: "GENDER", "Car\_Owner", "Propert\_Owner", "Type\_Income", "EDUCATION", "Marital\_status", "Housing\_type" and "Type\_Occupation".

## Step 4: 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 the train and test set.

Import data into the input spreadsheet of the “TRAIN\_TEST\_SPLIT” tab from the one-hot encoded dataset by right-clicking on the input spreadsheet and then choosing “Import from File”. Then navigate through your files to load the one with the encoded dataset.

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 'Random Partitioning' dialog box shows the following configuration:

- Training Set Percentage: 75
- Time-based RNG Seed: ☐ (unchecked)
- Stratified sampling: ☒ (checked)
- Column for sampling: Col11 – label
- Buttons: Execute, Cancel

The results will be two separate spreadsheets, “TRAIN\_TEST\_SPLIT: Training Set” and “TRAIN\_TEST\_SPLIT: Test Set”, which will be available to import into the next tabs.

Create a new tab by pressing the “+” button on the bottom of the page with the name “NORMALIZE\_TRAIN\_SET”.

[illegible]

The screenshot shows the 'ZScore Normalizer' dialog box in the Orange3 software. The dialog has two main sections: 'Excluded Columns' and 'Included Columns'. The 'Excluded Columns' list contains the following items: 'Col6 -- Mobile\_phone', 'Col11 -- label', 'Col12 -- GENDER\_F', 'Col13 -- GENDER\_M', 'Col14 -- Car\_Owner\_N', 'Col15 -- Car\_Owner\_Y', 'Col16 -- Propert\_Owner\_N', 'Col17 -- Propert\_Owner\_Y', and 'Col18 -- Tyme\_Income\_Com'. The 'Included Columns' list contains the following items: 'Col2 -- CHILDREN', 'Col3 -- Annual\_income', 'Col4 -- Birthday\_count', 'Col5 -- Employed\_days', 'Col7 -- Work\_Phone', 'Col8 -- Phone', 'Col9 -- EMAIL\_ID', and 'Col10 -- Family\_Members'. Between the two lists are four buttons: '>>', '>', '<', and '<<'. At the bottom of the dialog are two buttons: 'Execute' and 'Cancel'. The 'Execute' button is highlighted with a blue border.

User Header	ID	Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7	Cat8	Cat9	Cat10	Cat11	Cat12	Cat13	Cat14	Cat15	Cat16	Cat17	Cat18	Cat19	Cat20	Cat21	Cat22	Cat23	Cat24	Cat25	Cat26	Cat27	Cat28	Cat29	Cat30	
		Row ID	Cat1(Cat2)	Virtualization/Storage	Enterprise/Cloud	Mobile/Healthcare	Web/Phone	Finance/Log	IT/Security	Energy/Transport	Label	GENESYS	GENESYS	Cat_Demo1	Cat_Demo2	Open-Source/Custom	Open-Source/Custom	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account	Types:Account
1	000012	11183000	03783010	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
2	000013	11183000	03783010	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
3	000014	11183000	03783010	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
4	000015	11183000	03783010	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
5	000016	03783020	03783020	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
6	000017	03783020	03783020	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
7	000018	03783020	03783020	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
8	000019	03783020	03783020	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
9	000020	03783020	03783020	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
10	000021	03783020	03783020	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
11	000022	03783020	03783020	02628130	1.0	1.603201	1.603000	03711800	-0.160729	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0
12	000023	03783020	03783020	02628130	1.0	1.603201	1.603000</																									



## Step 6: 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 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”.

User Header	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12	Col13	Col14	Col15	Col16	Col17	Col18	Col19	Col20	Col21	Col22	Col23	Col24	Col25	Col26	Col27	Col28	Col29	Col30			
	User Key (0)	Col1 (0)	Annual_Income	BirthDay_Count	Employed_Days	Phone	Work_Phone	Phone	EMAIL_ID	Family_Members	Label	Col10 (0)	Col11 (0)	Col12 (0)	Col13 (0)	Col14 (0)	Col15 (0)	Col16 (0)	Col17 (0)	Col18 (0)	Col19 (0)	Col20 (0)	Col21 (0)	Col22 (0)	Col23 (0)	Col24 (0)	Col25 (0)	Col26 (0)	Col27 (0)	Col28 (0)	Col29 (0)	Col30 (0)	
1	500021	0	18000.0	-18772.0	265243	1	0	0	0	2	1	0	1	0	1	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	
2	500044	0	316000.0	-13857.0	-588	1	1	1	0	2	1	1	0	0	1	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	0	
3	500046	0	316000.0	-14046.3421	-588	1	1	1	0	2	1	1	0	0	1	1	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	
4	500049	0	191399.3282	-15527.0	-588	1	1	1	0	2	1	1	0	0	1	1	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	
5	500089	1	48000.0	-18173.0	-478	1	0	1	1	3	1	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0
6	500866	0	90000.0	-18950.0	-1002	1	1	1	0	2	1	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	
7	500871	0	191399.3282	-18950.0	-1002	1	1	1	0	2	1	1	0	0	1	0	1	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	
8	500895	0	90000.0	-18950.0	-1002	1	1	1	0	2	1	1	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0	1	0	1	0	0	
9	500899	0	540000.0	-15996.0	-491	1	0	0	0	2	1	0	1	0	1	1	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	
10	503451	0	135000.0	-10748.0	-2384	1	0	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	
11	503514	1	135000.0	-15992.0	-384	1	1	1	0	0	3	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	
12	503669	0	112500.0	-20550.0	-1546	1	0	0	0	2	1	0	1	0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	1	
13	503684	2	270000.0	-19553.0	-7569	1	1	1	0	4	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	
14	503694	0	225000.0	-23268.0	365243	1	0	0	0	1	1	1	1	0	1	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	
15	503695	0	225000.0	-23268.0	365243	1	0	0	0	1	1	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	

Normalize the test set using the existing normalizer of the training set: *Analytics → Existing Model Utilization → Model (from Tab:) NORMALIZE\_TRAIN\_SET*

Data Transformation ▾

Analytics ▾

Statistics ▾

Regression

Classification

Clustering

Anomaly Detection

Existing Model Utilization

Model (from Tab:) NORMALIZE...

Type Z Score Normalizer Model

Description

Model In...

Header -> Datatype

CHILDREN -> Double

Annual\_income -> Double

BirthDay\_Count -> Double

Employed\_days -> Double

Work\_Phone -> Double

Phone -> Double

EMAIL\_ID -> Double

Family\_Members -> Double

Transfer Column(s) to Output

Execute

Cancel

The results will appear on the output spreadsheet.

	Col1	Col2 (0)	Col3 (0)	Col4 (0)	Col5 (0)	Col6 (0)	Col7 (0)	Col8 (0)	Col9 (0)	Col10 (0)	Col11 (0)	Col12 (0)	Col13 (0)	Col14 (0)	Col15 (0)	Col16 (0)	Col17 (0)	Col18 (0)	Col19 (0)	Col20 (0)	Col21 (0)	Col22 (0)	Col23 (0)	Col24 (0)	Col25 (0)	Col26 (0)	Col27 (0)	Col28 (0)	Col29 (0)	Col30 (0)		
User Header	User Key (0)	CHILDREN	Annual_Income	BirthDay_Count	Employed_Days	Phone	Work_Phone	Phone	EMAIL_ID	Family_Members	Label	Col10 (0)	Col11 (0)	Col12 (0)	Col13 (0)	Col14 (0)	Col15 (0)	Col16 (0)	Col17 (0)	Col18 (0)	Col19 (0)	Col20 (0)	Col21 (0)	Col22 (0)	Col23 (0)	Col24 (0)	Col25 (0)	Col26 (0)	Col27 (0)	Col28 (0)	Col29 (0)	Col30 (0)
1	500827	-0.5165499	-0.5822318	-0.6599024	2.3232083	1.0	-0.5089183	-0.6714621	-0.3118808	-0.1880729	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
2	500914	-0.5165499	1.1130308	0.1735010	-0.4281130	1.0	1.8632671	1.4880048	-0.3118808	-0.1880729	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
3	500746	-0.5165499	1.1130308	-0.0185070	-0.4281130	1.0	1.8632671	1.4880048	-0.3118808	-0.1880729	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
4	500746	-0.5165499	0.0186226	0.1735010	-0.4281130	1.0	1.8632671	1.4880048	-0.3118808	-0.1880729	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
5	500089	0.1727382	2.3114852	-0.5270821	-0.4280872	1.0	-0.5089183	1.4880048	3.2039913	0.8683792	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
6	500869	-0.5165499	-0.8827192	-0.1723205	-0.4311613	1.0	1.8632671	1.4880048	-0.3118808	-0.1880729	1.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
7	500851	-0.5165499	0.0186226	-0.1723205	-0.4311613	1.0	1.8632671	1.4880048	-0.3118808	-0.1880729	1.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
8	500895	-0.5165499	-0.8827192	-0.1723205	-0.4311613	1.0	1.8632671	1.4880048	-0.3118808	-0.1880729	1.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
9	503595	-0.5165499	3.1100548	-0.9617362	-0.4280824	1.0	-0.5089183	-0.6714621	-0.3118808	-0.1880729	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
10	503451	-0.5165499	-0.4035064	1.2403084	-0.4415912	1.0	-0.5089183	-0.6714621	-0.3118808	-0.1880729	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
11	503514	0.1727382	-0.4831004	0.0020482	-0.4280829	1.0	1.8632671	-0.6714621	-0.3118808	0.8683792	1.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
12	503669	-0.5165499	-0.8827192	-0.1723205	-0.4311613	1.0	-0.5089183	-0.6714621	-0.3118808	-0.1880729	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
13	504254	1.8712183	0.1743140	-0.0218150	-0.4781016	1.0	1.8632671	1.4880048	-0.3118808	1.8886229	1.0	1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
14	504394	-0.5165499	0.3130812	-1.7481859	2.3232083	1.0	-0.5089183	-0.6714621	-0.3118808	-0.1880729	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
15	504395	-0.5165499	0.3130812	-1.7481859	2.3232083	1.0	-0.5089183	-0.6714621	-0.3118808	-0.1880729	1.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0

## 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 “NORMALIZE\_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*

The image shows two parts of the software interface. On the left is a menu titled 'Analytics' with a dropdown arrow. The 'Classification' option is selected, opening a submenu with the following items: 'k-Nearest Neighbors (kNN)', 'Fully Connected Neural Network', 'Radial Basis Function Network', 'XGBoost', 'J48 Decision Tree', 'Random Forest', 'Statistical fitting', and 'Auto ML'. On the right is the 'XGBoost Classification Model' configuration window. It contains the following settings:

- Target Column: Col11 - label
- Booster: gbtrees
- Objective: multisoftprob
- Number of Trees: 210
- Learning Rate: 0.3
- Gamma: 0.0
- Max Tree Depth: 6
- Minimum Child Weight: 1.0
- Column Sample by tree: 1.0
- Subsample: 1.0
- Tree Method: auto
- Lambda: 1.0
- Alpha: 0.0
- ☐ Time-based RNG Seed
- RNG Seed: 1765809404695

At the bottom of the configuration window are 'Execute' and 'Cancel' buttons.

The predictions will appear on the output spreadsheet.

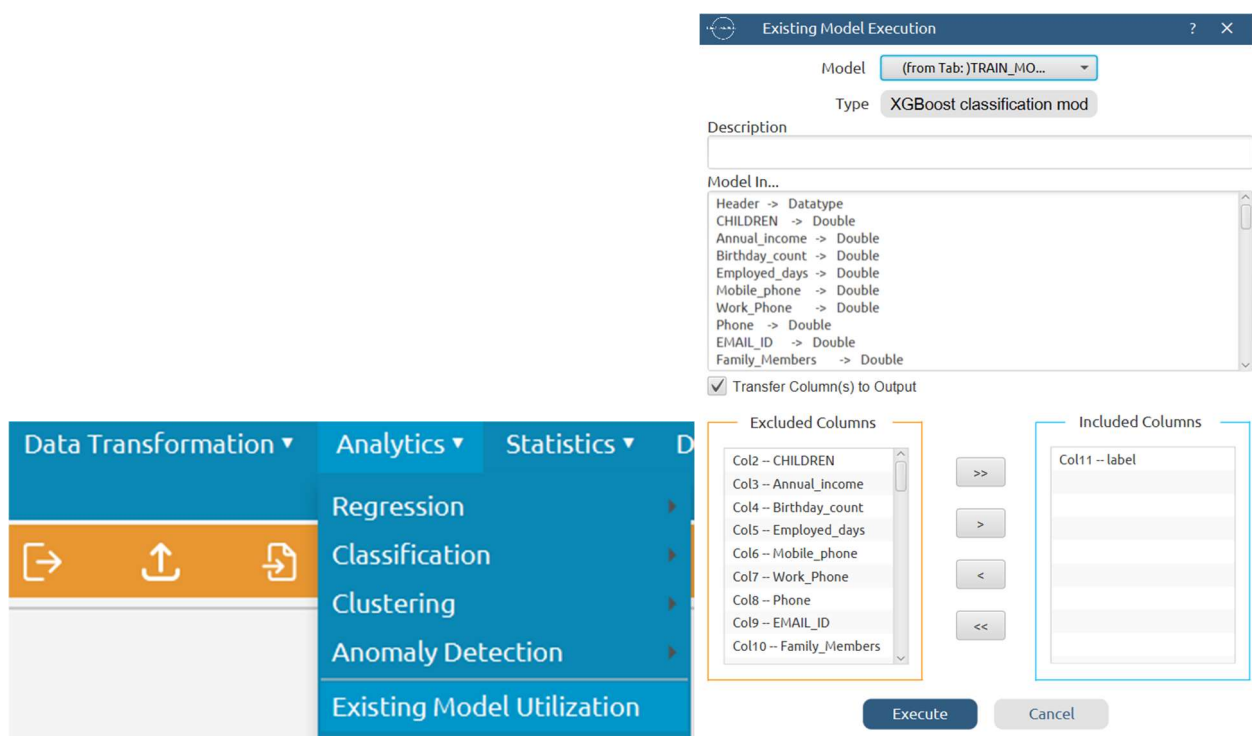
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	label	Prediction
1	5009752	1.0	1.0
2	5009753	1.0	1.0
3	5009754	1.0	1.0
4	5009894	1.0	1.0
5	5010864	1.0	0.0
6	5010869	1.0	1.0
7	5021303	1.0	1.0
8	5021310	1.0	1.0
9	5021314	1.0	1.0
10	5021430	1.0	1.0
11	5021431	1.0	1.0
12	5021998	1.0	1.0
13	5022053	1.0	1.0
14	5022617	1.0	1.0
15	5023781	1.0	1.0

## 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 “NORMALIZE\_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

To validate the model: *Analytics* → *Existing Model Utilization* → *Model (from Tab:) TRAIN\_MODEL(.fit)*. Choose the column “label” to be transferred to the output spreadsheet.



The predictions will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Prediction	label
1	5008827	0.0	1.0
2	5009744	1.0	1.0
3	5009746	0.0	1.0
4	5009749	0.0	1.0
5	5010868	1.0	1.0
6	5018498	0.0	1.0
7	5018501	0.0	1.0
8	5018503	0.0	1.0
9	5028580	0.0	1.0
10	5033453	0.0	1.0
11	5033514	0.0	1.0
12	5036469	0.0	1.0
13	5042064	0.0	1.0
14	5045894	1.0	1.0
15	5045895	1.0	1.0

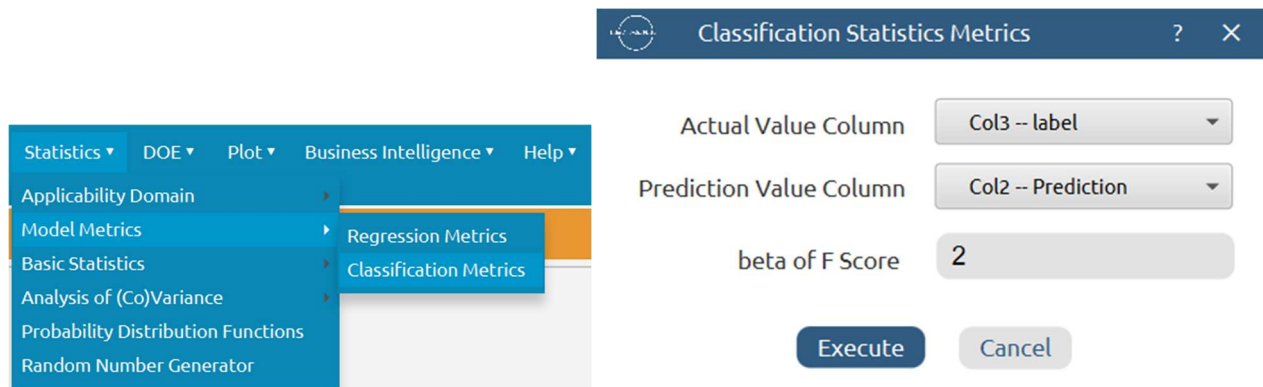


## Step 9: Statistics calculation

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 image shows the 'Classification Statistics Metrics' dialog box and the 'Statistics' menu. The dialog box has the following fields:

- Actual Value Column:** Col3 -- label
- Prediction Value Column:** Col2 -- Prediction
- beta of F Score:** 2
- Buttons:** Execute, Cancel

The 'Statistics' menu is open, showing the following options:

- Statistics ▾
- DOE ▾
- Plot ▾
- Business Intelligence ▾
- Help ▾
- Applicability Domain ▾
- Model Metrics ▾
  - Regression Metrics
  - Classification Metrics
- Basic Statistics ▾
- Analysis of (Co)Variance ▾
- Probability Distribution Functions
- Random Number Generator

The results will appear on the output spreadsheet.

	Col1 (S)	Col2 (D)	Col3 (S)	Col4 (S)
User Header	User Row ID			
1			Predicted Class	Predicted Class
2			1.0	0.0
3	Actual Class	1.0	16	28
4	Actual Class	0.0	13	330
5				
6				
7	Classification Accuracy	0.8940568		
8				
9	Precision		0.5517241	0.9217877
10				
11	Recall/Sensitivity		0.3636364	0.9620991
12				
13	Specificity		0.9620991	0.3636364
14				
15	F1 Score		0.4383562	0.9415121
16				
17	F (beta=2)		0.3902439	0.9537572
18				
19	MCC	0.3927335		

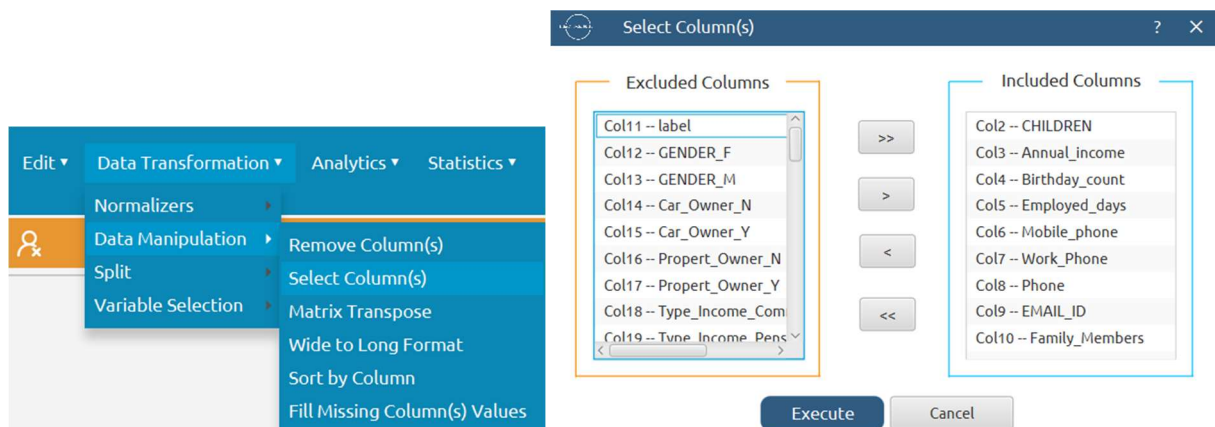
## Step 10: Reliability check for each record of the test set

### 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 “NORMALIZE\_TRAIN\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Manipulate the data to exclude the target column “label” and the one-hot encoded columns: *Data Transformation → Data Manipulation → Select Column(s)*

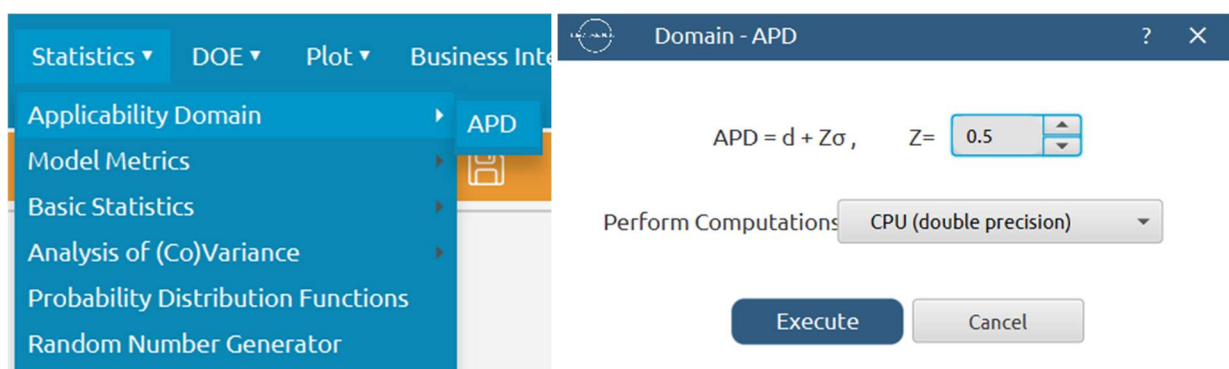


The results will appear on the output spreadsheet.

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 → Applicability Domain → APD*



The results will appear on the output spreadsheet.

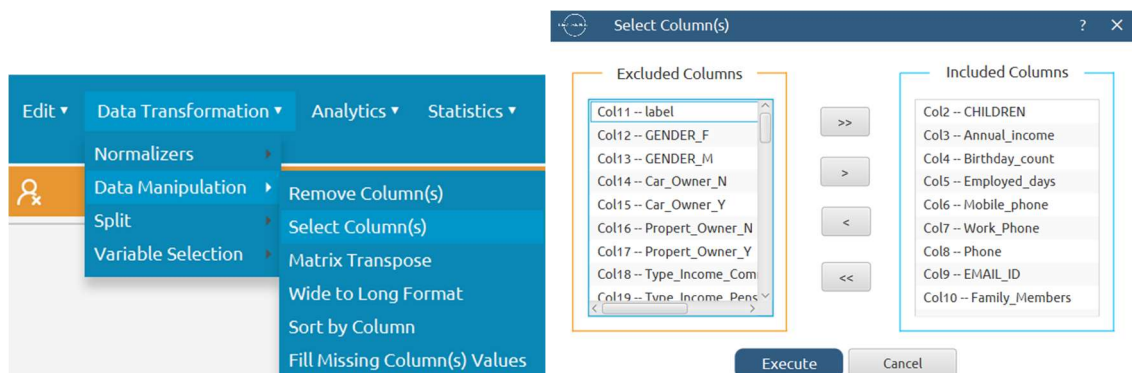
	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1	5009752	0.0	2.9798623	reliable
2	5009753	0.0	2.9798623	reliable
3	5009754	0.0	2.9798623	reliable
4	5009894	0.0	2.9798623	reliable
5	5010864	0.0	2.9798623	reliable
6	5010869	0.0	2.9798623	reliable
7	5021303	0.0	2.9798623	reliable
8	5021310	0.0	2.9798623	reliable
9	5021314	0.0	2.9798623	reliable
10	5021430	0.0	2.9798623	reliable
11	5021431	0.0	2.9798623	reliable
12	5021998	0.0	2.9798623	reliable
13	5022053	0.0	2.9798623	reliable
14	5022617	0.0	2.9798623	reliable
15	5023781	0.0	2.9798623	reliable

## 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 “NORMALIZE\_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Manipulate the data to exclude the target column “label” and the one-hot encoded columns: *Data Transformation* → *Data Manipulation* → *Select Column(s)*

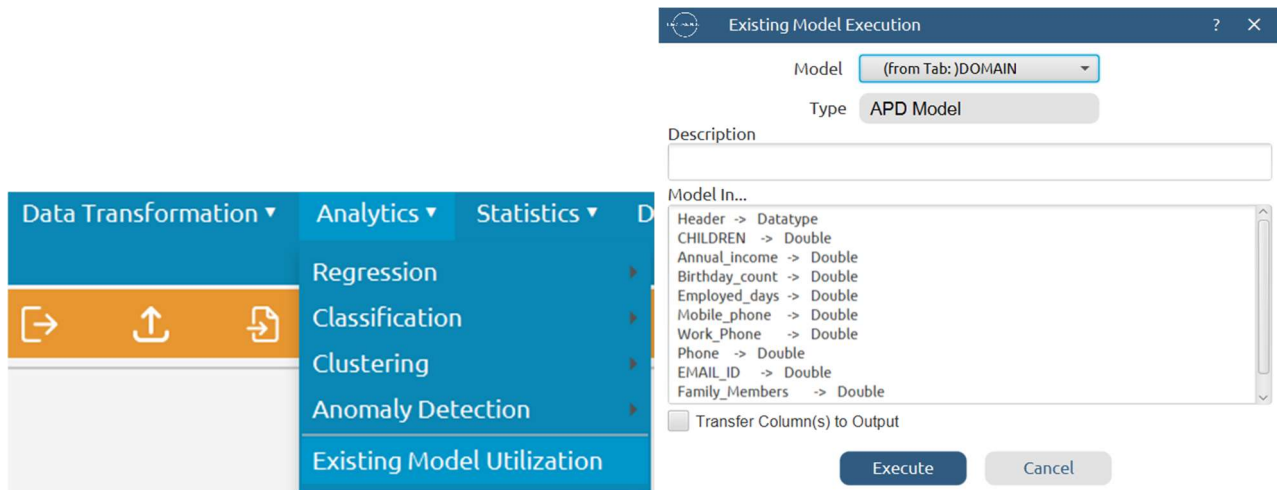


The results will appear on the output spreadsheet.

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 → Model (from Tab:) DOMAIN*



The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1	5008827	0.2945236	2.9798623	reliable
2	5009744	0.0	2.9798623	reliable
3	5009746	0.0562452	2.9798623	reliable
4	5009749	0.1035060	2.9798623	reliable
5	5010868	0.0	2.9798623	reliable
6	5018498	0.1401839	2.9798623	reliable
7	5018501	0.0472154	2.9798623	reliable
8	5018503	0.1401839	2.9798623	reliable
9	5028580	1.3404827	2.9798623	reliable
10	5033453	0.0214850	2.9798623	reliable
11	5033514	0.1765022	2.9798623	reliable
12	5036469	0.0553171	2.9798623	reliable
13	5042064	0.4014907	2.9798623	reliable
14	5045894	0.0	2.9798623	reliable
15	5045895	0.0	2.9798623	reliable

## Final Isalos Workflow

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

