



# Graduate Admission

The goal of this study is to train machine learning models in order to predict the probability of admission of students applying to graduate programs and to evaluate and compare the performance of different machine learning algorithms, specifically k-Nearest Neighbors (kNN), Linear Stochastic Gradient Descent (Linear SGD), Random Forest, XGBoost, MLP and RBF.

The dataset, which can be found at <https://www.kaggle.com/datasets/mohansacharya/graduate-admissions>, contains 7 predictor variables and 500 observations. It includes information on several academic and profile-related factors that are considered important during the graduate admission process, such as GRE scores, TOEFL scores, university rating, statement of purpose (SOP) strength, letter of recommendation (LOR) strength, undergraduate GPA (CGPA), and whether the applicant has prior research experience. The target variable of the dataset is the Chance of Admit, which represents the estimated probability of admission and ranges between 0 and 1. The dataset contains no missing values and all variables are numerical, making it suitable for regression-based machine learning models.

*Isalos version used: 2.0.2*

## 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 Admission Predict Ver.1.1 data.

A screenshot of a spreadsheet application interface. The spreadsheet has 8 columns labeled Col1 through Col8 and 16 rows. The first row is a header row with "User Header" in the first cell and "User Row ID" in the second cell. The remaining rows are numbered 1 through 16. A right-click context menu is open over the cell at row 6, column 4. The menu items are: Show Spreadsheet Toolbar, Import from File, Import from Spreadsheet, Import from Multiple Spreadsheets, Adjust Spreadsheet Precision, Export Spreadsheet Data, and Clear Spreadsheet. The "Import from File" option is highlighted in blue.

	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
User Header	User Row ID							
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								

The data will appear on the left spreadsheet.

	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (I)	Col9 (D)
User Header	User Row ID	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	1	337	118	4	4.5	4.5	9.65	1	0.92
2	2	324	107	4	4	4.5	8.87	1	0.76
3	3	316	104	3	3	3.5	8	1	0.72
4	4	322	110	3	3.5	2.5	8.67	1	0.8
5	5	314	103	2	2	3	8.21	0	0.65
6	6	330	115	5	4.5	3	9.34	1	0.9
7	7	321	109	3	3	4	8.2	1	0.75
8	8	308	101	2	3	4	7.9	0	0.68
9	9	302	102	1	2	1.5	8	0	0.5
10	10	323	108	3	3.5	3	8.6	0	0.45
11	11	325	106	3	3.5	4	8.4	1	0.52
12	12	327	111	4	4	4.5	9	1	0.84
13	13	328	112	4	4	4.5	9.1	1	0.78
14	14	307	109	3	4	3	8	1	0.62
15	15	311	104	3	3.5	2	8.2	1	0.61
16	16	314	105	3	3.5	2.5	8.3	0	0.54
17	17	317	107	3	4	3	8.7	0	0.66
18	18	319	106	3	4	3	8	1	0.65
19	19	318	110	3	4	3	8.8	0	0.63
20	20	303	102	3	3.5	3	8.5	0	0.62

## Step 2: Manipulate data

In this dataset there are not any empty values or categorical features, so we can select all the columns to be used.

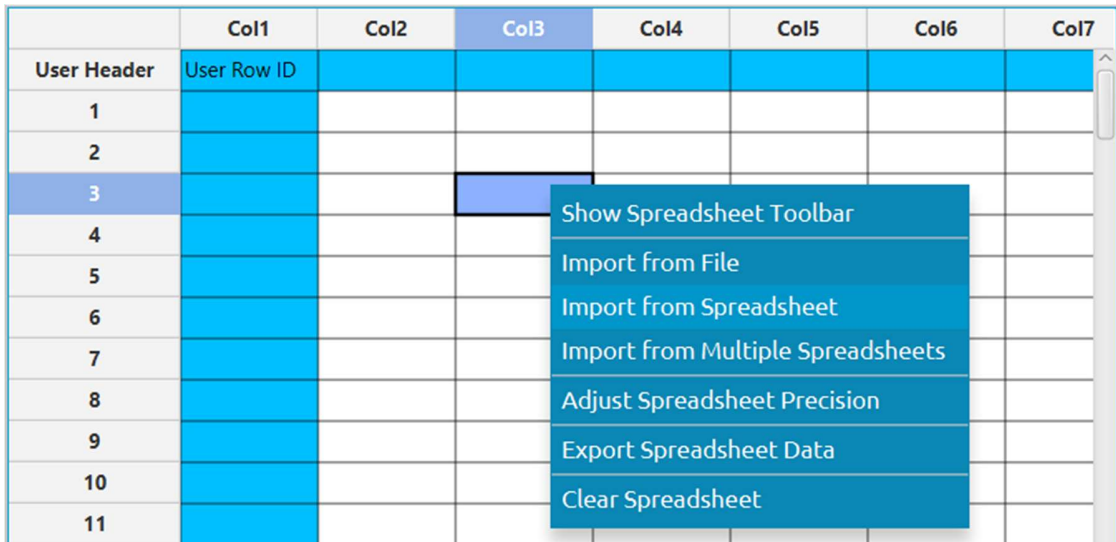
On the menu click on *Data Transformation* → *Data Manipulation* → *Select Column(s)* and select all columns.

All 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.

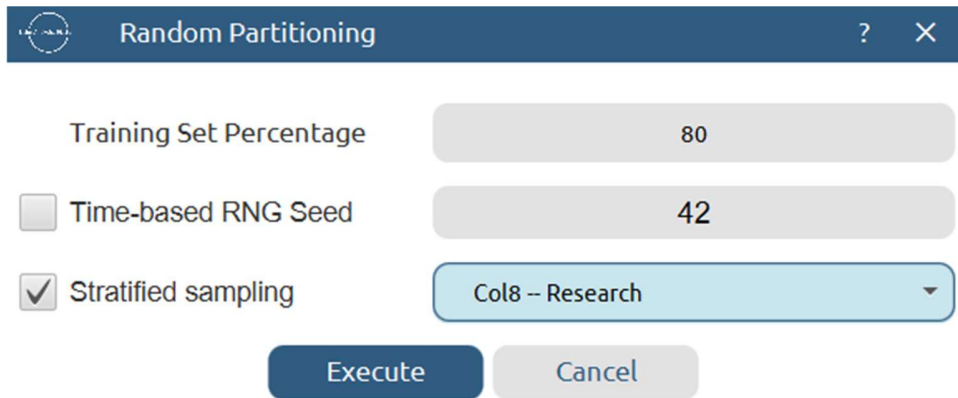
### Step 3: Split data

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 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:



In regression problems, stratified sampling is generally not required because the target variable is continuous and does not consist of discrete classes whose proportions need to be preserved. However, if the dataset contains categorical predictor variables and it is important to maintain the distribution of their categories in both the training and test sets, stratified sampling can be applied based on those specific features. This ensures that the relative proportions of the categories are preserved across the data partitions.

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.

## Step 4: Feature selection for training set

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

Import data into the input spreadsheet of the “FEATURE\_SELECTION\_TRAINING\_SET” tab 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”.

	Col1	Col2 (I)	Col3 (I)	Col4 (I)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (I)	Col9 (D)	Co
User Header	User Row ID	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit	
1	1	337	118	4	4.5	4.5	9.65	1	0.92	
2	3	316	104	3	3	3.5	8	1	0.72	
3	4	322	110	3	3.5	2.5	8.67	1	0.8	
4	5	314	103	2	2	3	8.21	0	0.65	
5	6	330	115	5	4.5	3	9.34	1	0.9	
6	7	321	109	3	3	4	8.2	1	0.75	
7	8	308	101	2	3	4	7.9	0	0.68	
8	10	323	108	3	3.5	3	8.6	0	0.45	
9	11	325	106	3	3.5	4	8.4	1	0.52	
10	12	327	111	4	4	4.5	9	1	0.84	
11	14	307	109	3	4	3	8	1	0.62	
12	15	311	104	3	3.5	2	8.2	1	0.61	
13	17	317	107	3	4	3	8.7	0	0.66	
14	18	319	106	3	4	3	8	1	0.65	
15	19	318	110	3	4	3	8.8	0	0.63	
16	24	334	119	5	5	4.5	9.7	1	0.95	
17	25	336	119	5	4	3.5	9.8	1	0.97	
18	27	322	109	5	4.5	3.5	8.8	0	0.76	

Then do Stepwise with the “Forward Selection” as the method: *Data Transformation* → *Variable Selection* → *Stepwise*

The results will appear on the right spreadsheet.

	Col1	Col2 (D)	Col3 (I)	Col4 (D)	Col5 (I)	Col6 (I)	Col7 (D)	Col8 (D)
User Header	User Row ID	CGPA	GRE Score	LOR	Research	TOEFL Score	SOP	Chance of Admit
1		9.65	337	4.5	1	118	4.5	0.92
2		8	316	3.5	1	104	3	0.72
3		8.67	322	2.5	1	110	3.5	0.8
4		8.21	314	3	0	103	2	0.65
5		9.34	330	3	1	115	4.5	0.9
6		8.2	321	4	1	109	3	0.75
7		7.9	308	4	0	101	3	0.68
8		8.6	323	3	0	108	3.5	0.45
9		8.4	325	4	1	106	3.5	0.52
10		9	327	4.5	1	111	4	0.84
11		8	307	3	1	109	4	0.62
12		8.2	311	2	1	104	3.5	0.61
13		8.7	317	3	0	107	4	0.66
14		8	319	3	1	106	4	0.65
15		8.8	318	3	0	110	4	0.63
16		9.7	334	4.5	1	119	5	0.95

## Step 5: Feature selection for 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 “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”

Knowing which features were selected for the training set, the same features should be selected for the test set as well, without applying any feature selection algorithm again. This is done simply by selecting the corresponding columns in the test dataset. On the menu click on *Data Transformation* → *Data Manipulation* → *Select Column(s)*

The image shows two parts of the software interface. On the left, a menu is open under 'Data Transformation' with 'Data Manipulation' selected, and a sub-menu is visible containing 'Remove Column(s)', 'Select Column(s)', 'Matrix Transpose', 'Wide to Long Format', 'Sort by Column', and 'Fill Missing Column(s) Values'. On the right, the 'Select Column(s)' dialog box is shown. It has a title bar with a refresh icon, a question mark, and a close icon. The dialog is divided into two panes: 'Excluded Columns' (orange border) containing 'Col4 -- University Rating' and 'Included Columns' (blue border) containing 'Col2 -- GRE Score', 'Col3 -- TOEFL Score', 'Col5 -- SOP', 'Col6 -- LOR', 'Col7 -- CGPA', 'Col8 -- Research', and 'Col9 -- Chance of Admit'. Between the panes are navigation buttons: '>>', '>', '<', and '<<'. At the bottom are 'Execute' and 'Cancel' buttons.

The results will appear on the right spreadsheet.

	Col1	Col2 (I)	Col3 (I)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (I)	Col8 (D)
User Header	User Row ID	GRE Score	TOEFL Score	SOP	LOR	CGPA	Research	Chance of Admit
1	2	324	107	4	4.5	8.87	1	0.76
2	9	302	102	2	1.5	8	0	0.5
3	13	328	112	4	4.5	9.1	1	0.78
4	16	314	105	3.5	2.5	8.3	0	0.54
5	20	303	102	3.5	3	8.5	0	0.62
6	21	312	107	3	2	7.9	1	0.64
7	22	325	114	3	2	8.4	0	0.7
8	23	328	116	5	5	9.5	1	0.94
9	26	340	120	4.5	4.5	9.6	1	0.94
10	42	316	105	2.5	2.5	8.2	1	0.49
11	51	313	98	2.5	4.5	8.3	1	0.76
12	52	312	100	1.5	3.5	7.9	1	0.56
13	54	324	112	4	2.5	8.1	1	0.72
14	59	300	99	3	2	6.8	1	0.36
15	61	309	100	3	3	8.1	0	0.48
16	62	307	101	4	3	8.2	0	0.47
17	78	301	99	3	2	8.22	0	0.64

## Step 6: 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 into the input spreadsheet of the “NORMALIZE\_TRAIN\_SET” tab the train set from the output of the “FEATURE\_SELECTION\_TRAINING\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Normalize the data using Z-score: Data Transformation → Normalizers → Z Score and select all columns except the “Chance of Admit” target column.

The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
User Header	User Row ID	CGPA	GRE Score	LOR	Research	TOEFL Score	SOP	Chance of Admit
1		1.7217882	1.7859708	1.0590905	0.8852966	1.7705357	1.1191968	0.92
2		-0.9596819	-0.0532047	-0.0188642	0.8852966	-0.5224800	-0.3664627	0.72
3		0.1291575	0.4722740	-1.0968190	0.8852966	0.4602410	0.1287572	0.8
4		-0.6184039	-0.2283643	-0.5578416	-1.1267411	-0.6862668	-1.3569023	0.65
5		1.2179968	1.1729123	-0.5578416	0.8852966	1.2791752	1.1191968	0.9
6		-0.6346552	0.3846942	0.5201132	0.8852966	0.2964542	-0.3664627	0.75
7		-1.1221952	-0.7538430	0.5201132	-1.1267411	-1.0138405	-0.3664627	0.68
8		0.0153981	0.5598538	-0.5578416	-1.1267411	0.1326673	0.1287572	0.45
9		-0.3096285	0.7350133	0.5201132	0.8852966	-0.1949063	0.1287572	0.52
10		0.6654515	0.9101729	1.0590905	0.8852966	0.6240278	0.6239770	0.84
11		-0.9596819	-0.8414228	-0.5578416	0.8852966	0.2964542	0.6239770	0.62
12		-0.6346552	-0.4911036	-1.6357963	0.8852966	-0.5224800	0.1287572	0.61
13		0.1779115	0.0343751	-0.5578416	-1.1267411	-0.0311195	0.6239770	0.66
14		-0.9596819	0.2095346	-0.5578416	0.8852966	-0.1949063	0.6239770	0.65
15		0.3404248	0.1219548	-0.5578416	-1.1267411	0.4602410	0.6239770	0.63
16		1.8030449	1.5232314	1.0590905	0.8852966	1.9343225	1.6144166	0.95

## Step 7: 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 train set from the output of the “FEATURE\_SELECTION\_TEST\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

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

The image shows two parts of the software interface. On the left, a menu is open with 'Existing Model Utilization' selected. On the right, the 'Existing Model Execution' dialog box is displayed. The dialog has a title bar with a refresh icon, a question mark, and a close button. It contains the following fields and options:

- Model:** (from Tab:)NORMALIZE...
- Type:** Z Score Normalizer Model
- Description:** (empty text box)
- Model Input:**
  - Header -> Datatype
  - CGPA -> Double
  - GRE Score -> Double
  - LOR -> Double
  - Research -> Double
  - TOEFL Score -> Double
  - SOP -> Double
- Transfer Column(s) to Output
- Execute** button
- Cancel** button

The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)
User Header	User Row ID	GRE Score	TOEFL Score	SOP	LOR	CGPA	Research	Chance of Admit
1	2	0.6474336	-0.0311195	0.6239770	1.0590905	0.4541842	0.8852966	0.76
2	9	-1.2793217	-0.8500537	-1.3569023	-2.1747737	-0.9596819	-1.1267411	0.5
3	13	0.9977527	0.7878147	0.6239770	1.0590905	0.8279648	0.8852966	0.78
4	16	-0.2283643	-0.3586932	0.1287572	-1.0968190	-0.4721419	-1.1267411	0.54
5	20	-1.1917419	-0.8500537	0.1287572	-0.5578416	-0.1471152	-1.1267411	0.62
6	21	-0.4035239	-0.0311195	-0.3664627	-1.6357963	-1.1221952	0.8852966	0.64
7	22	0.7350133	1.1153883	-0.3664627	-1.6357963	-0.3096285	-1.1267411	0.7
8	23	0.9977527	1.4429620	1.6144166	1.5980679	1.4780182	0.8852966	0.94
9	26	2.0487101	2.0981093	1.1191968	1.0590905	1.6405315	0.8852966	0.94
10	42	-0.0532047	-0.3586932	-0.8616825	-1.0968190	-0.6346552	0.8852966	0.49
11	51	-0.3159441	-1.5052010	-0.8616825	1.0590905	-0.4721419	0.8852966	0.76
12	52	-0.4035239	-1.1776273	-1.8521221	-0.0188642	-1.1221952	0.8852966	0.56
13	54	0.6474336	0.7878147	0.6239770	-1.0968190	-0.7971686	0.8852966	0.72
14	59	-1.4544813	-1.3414142	-0.3664627	-1.6357963	-2.9098419	0.8852966	0.36
15	61	-0.6662632	-1.1776273	-0.3664627	-0.5578416	-0.7971686	-1.1267411	0.48
16	62	-0.8414228	-1.0138405	0.6239770	-0.5578416	-0.6346552	-1.1267411	0.47

## Step 8: Train the Model-kNN

Create a new tab by pressing the “+” button on the bottom of the page with the name “kNN”.

Import data into the input spreadsheet of the “kNN” 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 k-Nearest Neighbors (kNN) method to train and fit the model: *Analytics* → *Regression* → *kNearest Neighbors (kNN)*.

The screenshot shows the 'kNN Regression Model' configuration window. On the left, a menu is open with 'Analytics' selected, and 'Regression' is expanded to show 'k-Nearest Neighbors (kNN)'. The main window has a title bar 'kNN Regression Model' with a help icon and a close button. Below the title bar, there are two configuration fields: 'Target Column' with a dropdown menu showing 'Col8 -- Chance of Admit', and 'Number of Neighbors' with a text input field containing '10'. At the bottom of the window are two buttons: 'Execute' and 'Cancel'.

The predictions will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (S)	Col5 (D)	Col6 (S)	Col7 (D)	Col8 (S)	Col9 (D)	Col10 (S)	Col11 (D)
User Header	User Row ID	Chance of Admit	kNN Prediction	Closest NN1	Distance from NN1	Closest NN2	Distance from NN2	Closest NN3	Distance from NN3	Closest NN4	Distance from NN4
1		0.92	0.9216461	Entry 1	0.0	Entry 95	0.0942231	Entry 198	0.1186058	Entry 287	0.1323571
2		0.72	0.7197125	Entry 2	0.0	Entry 268	0.1542698	Entry 140	0.1859756	Entry 385	0.1969072
3		0.8	0.7987870	Entry 3	0.0	Entry 207	0.1425476	Entry 326	0.1484742	Entry 382	0.1823461
4		0.65	0.6512105	Entry 4	0.0	Entry 137	0.0475219	Entry 302	0.1308373	Entry 126	0.1542603
5		0.9	0.8993721	Entry 5	0.0	Entry 281	0.0726498	Entry 150	0.1268033	Entry 111	0.1300340
6		0.75	0.7480760	Entry 6	0.0	Entry 319	0.1415554	Entry 40	0.1590394	Entry 268	0.1592930
7		0.68	0.6779105	Entry 7	0.0	Entry 204	0.1981164	Entry 356	0.2020108	Entry 185	0.2071907
8		0.45	0.4608048	Entry 8	0.0	Entry 156	0.1563072	Entry 133	0.1642009	Entry 132	0.1683698
9		0.52	0.5312534	Entry 9	0.0	Entry 77	0.1452010	Entry 248	0.1514207	Entry 319	0.1693688
10		0.84	0.8400480	Entry 10	0.0	Entry 258	0.0568814	Entry 188	0.0829337	Entry 240	0.1363699
11		0.62	0.6210225	Entry 11	0.0	Entry 32	0.1318967	Entry 121	0.2085852	Entry 255	0.2235093
12		0.61	0.6125937	Entry 12	0.0	Entry 254	0.1336480	Entry 255	0.1572434	Entry 375	0.2113516
13		0.66	0.6612722	Entry 13	0.0	Entry 15	0.1187321	Entry 156	0.1483191	Entry 133	0.1541467
14		0.65	0.6525605	Entry 14	0.0	Entry 385	0.1373335	Entry 48	0.2038696	Entry 183	0.2144048
15		0.63	0.6339694	Entry 15	0.0	Entry 13	0.1187321	Entry 156	0.1371605	Entry 181	0.1529583
16		0.95	0.9497661	Entry 16	0.0	Entry 172	0.0355676	Entry 55	0.1381424	Entry 1	0.1446881
17		0.97	0.9683750	Entry 17	0.0	Entry 120	0.1262762	Entry 278	0.1519302	Entry 149	0.1868530
18		0.76	0.7585443	Entry 18	0.0	Entry 135	0.1445787	Entry 89	0.1624260	Entry 181	0.1914548
19		0.44	0.4437674	Entry 19	0.0	Entry 284	0.2698475	Entry 312	0.3134005	Entry 363	0.3966354

## Step 9: Validate the model-kNN

Create a new tab by pressing the “+” button on the bottom of the page with the name “VALIDATE\_MODEL\_kNN”.

Import data into the input spreadsheet of the “VALIDATE\_MODEL\_kNN” 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:) kNN(.fit)*. Choose the column “Chance of Admit” to be transferred to the output spreadsheet.

The screenshot shows the 'Existing Model Execution' window. The 'Model' dropdown is set to '(from Tab:)kNN' and the 'Type' is 'kNN Model'. Under 'Model Input', several variables are listed with their datatypes. The 'Transfer Column(s) to Output' section is active, showing a list of 'Excluded Columns' (Col2 - GRE Score, Col3 - TOEFL Score, Col4 - SOP, Col5 - LOR, Col6 - CGPA, Col7 - Research) and one 'Included Column' (Col8 - Chance of Admit). Navigation arrows between the lists are visible. 'Execute' and 'Cancel' buttons are at the bottom.

The predictions will appear on the output spreadsheet.

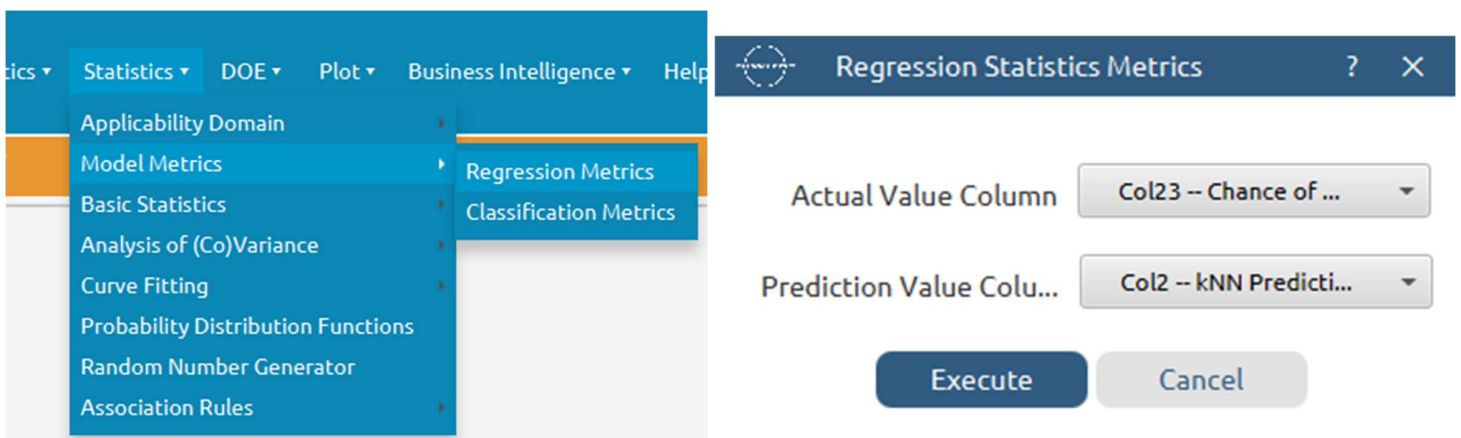
	Col1	Col2 (D)	Col3 (S)	Col4 (D)	Col5 (S)	Col6 (D)	Col7 (S)	Col8 (D)	Col9 (S)	Col10 (D)
User Header	User Row ID	kNN Prediction	Closest NN1	Distance from NN1	Closest NN2	Distance from NN2	Closest NN3	Distance from NN3	Closest NN4	Distance from NN4
1	2	0.8024019	Entry 320	0.0831752	Entry 260	0.1402891	Entry 362	0.1464504	Entry 10	0.1668297
2	9	0.5531868	Entry 251	0.1573723	Entry 358	0.1632738	Entry 342	0.1751363	Entry 377	0.2007071
3	13	0.8577740	Entry 10	0.0558873	Entry 188	0.0652863	Entry 258	0.0706064	Entry 142	0.1168590
4	16	0.6851984	Entry 159	0.0888448	Entry 321	0.1510077	Entry 378	0.1660463	Entry 133	0.1684105
5	20	0.6562741	Entry 337	0.1278244	Entry 309	0.1283895	Entry 145	0.1572004	Entry 144	0.1708075
6	21	0.6583305	Entry 254	0.1114851	Entry 375	0.1668828	Entry 12	0.2013342	Entry 255	0.2280593
7	22	0.6370622	Entry 259	0.2693165	Entry 51	0.3406717	Entry 132	0.3431402	Entry 8	0.3667619
8	23	0.9225639	Entry 84	0.0744739	Entry 348	0.0989607	Entry 122	0.1137627	Entry 55	0.1205661
9	26	0.9461179	Entry 1	0.0970818	Entry 164	0.1139706	Entry 95	0.1331094	Entry 287	0.1531366
10	42	0.6524470	Entry 33	0.1920332	Entry 162	0.1927095	Entry 64	0.2050105	Entry 254	0.2052195
11	51	0.6754427	Entry 205	0.2457746	Entry 386	0.2768789	Entry 140	0.2904872	Entry 391	0.2932485
12	52	0.6438278	Entry 323	0.1684548	Entry 110	0.1765279	Entry 140	0.2579166	Entry 64	0.2597882
13	54	0.7399180	Entry 3	0.2581218	Entry 383	0.2616238	Entry 14	0.2763320	Entry 41	0.2809351
14	59	0.5884604	Entry 62	0.3199436	Entry 19	0.4748182	Entry 312	0.4768479	Entry 363	0.5171682
15	61	0.6599327	Entry 144	0.1048486	Entry 139	0.1433287	Entry 309	0.1660463	Entry 337	0.1822256
16	62	0.6309445	Entry 309	0.1437039	Entry 337	0.1467157	Entry 243	0.1867002	Entry 145	0.2186022
17	78	0.6307954	Entry 271	0.2105554	Entry 180	0.2147776	Entry 376	0.2296997	Entry 252	0.2302808
18	82	0.9429441	Entry 171	0.0711351	Entry 94	0.1511035	Entry 155	0.1662336	Entry 102	0.1677145
19	88	0.6685602	Entry 133	0.1268221	Entry 241	0.1407818	Entry 310	0.1523107	Entry 132	0.1674931

## Step 10: Statistics calculation-kNN

Create a new tab by pressing the “+” button on the bottom of the page with the name “STATISTICS\_ACCURACIES\_kNN”.

Import data into the input spreadsheet of the “STATISTICS\_ACCURACIES\_kNN” tab from the output of the “VALIDATE\_MODEL\_kNN” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Calculate the statistical metrics for the regression: *Statistics → Model Metrics → Regression Metrics*



The results will appear on the output spreadsheet.

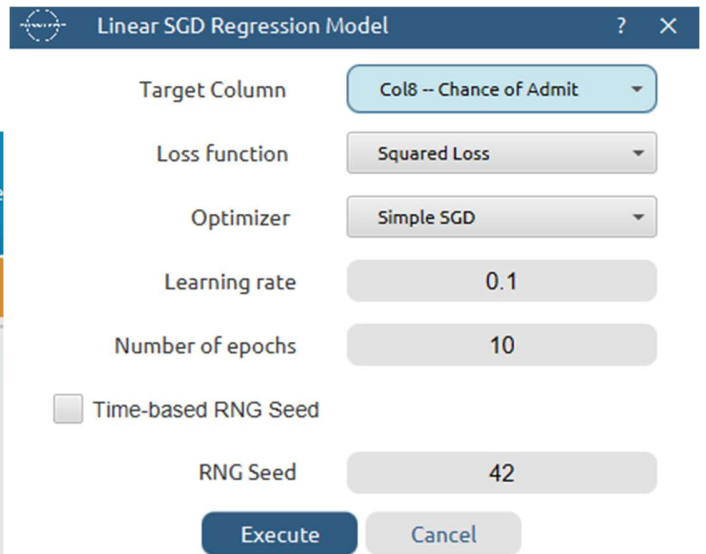
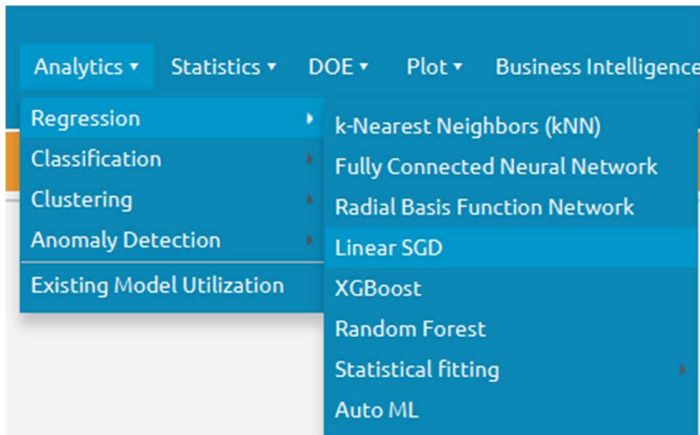
User Header	User Row ID	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R Squared
1		0.0051844	0.0720025	0.0509998	0.7537962

## Step 11: Train the Model – Linear SGD

Create a new tab by pressing the “+” button on the bottom of the page with the name “Linear SGD”.

Import data into the input spreadsheet of the “Linear SGD” 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 Linear SGD method to train and fit the model: *Analytics → Regression → Linear SGD.*



The predictions will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Chance of Admit	Prediction
1		0.92	0.9727237
2		0.72	0.6973000
3		0.8	0.7732449
4		0.65	0.6615049
5		0.9	0.8897950
6		0.75	0.7567868
7		0.68	0.6625871
8		0.45	0.7518404
9		0.52	0.7671405
10		0.84	0.8660627
11		0.62	0.7245354
12		0.61	0.6825182
13		0.66	0.7593282
14		0.65	0.7147787
15		0.63	0.7848823
16		0.95	0.9890421
17		0.97	0.9572203
18		0.76	0.8022262
19		0.44	0.5626646
20		0.46	0.4872419

## Step 12: Validate the model - Linear SGD

Create a new tab by pressing the “+” button on the bottom of the page with the name “VALIDATE\_MODEL\_Linear SGD”.

Import data into the input spreadsheet of the “VALIDATE\_MODEL\_Linear SGD” 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:) Linear SGD (.fit)*. Choose the column “Chance of Admit” to be transferred to the output spreadsheet.

The image shows two parts of the software interface. On the left is a partial view of the 'Analytics' menu, which includes options like Regression, Classification, Clustering, Anomaly Detection, and Existing Model Utilization. On the right is the 'Existing Model Execution' dialog box. The dialog has a title bar with a refresh icon, a question mark, and a close button. It contains the following fields and controls:

- Model:** A dropdown menu set to '(from Tab:)Linear SGD'.
- Type:** A dropdown menu set to 'Linear SGD Regression'.
- Description:** A text box containing 'It is a Linear SGD Regression model'.
- Model Input:** A list of input variables with their datatypes: CGPA (Double), GRE Score (Double), LOR (Double), Research (Double), TOEFL Score (Double), and SOP (Double).
- Transfer Column(s) to Output:** A checked checkbox.
- Excluded Columns:** A list box containing Col2 -- GRE Score, Col3 -- TOEFL Score, Col4 -- SOP, Col5 -- LOR, Col6 -- CGPA, and Col7 -- Research.
- Included Columns:** A list box containing Col8 -- Chance of Admit.
- Navigation:** Four arrow buttons (>>, >, <, <<) between the Excluded and Included columns lists.
- Buttons:** 'Execute' and 'Cancel' buttons at the bottom.

The predictions will appear on the output spreadsheet.

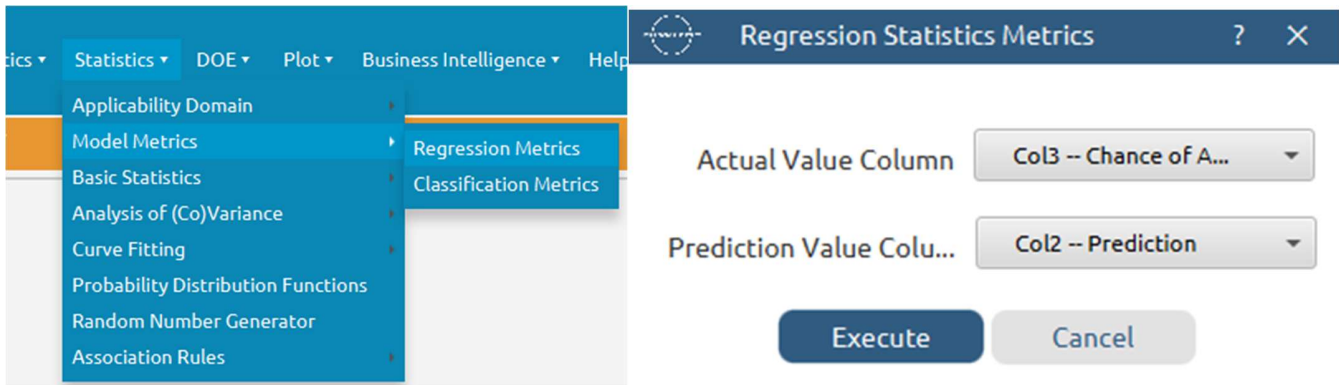
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Prediction	Chance of Admit
1	2	0.8312901	0.76
2	9	0.5950354	0.5
3	13	0.8804381	0.78
4	16	0.6929934	0.54
5	20	0.6984169	0.62
6	21	0.6668841	0.64
7	22	0.7373304	0.7
8	23	0.9645798	0.94
9	26	0.9815543	0.94
10	42	0.6864818	0.49
11	51	0.7026654	0.76
12	52	0.6392605	0.56
13	54	0.7472230	0.72
14	59	0.5249382	0.36
15	61	0.6495556	0.48
16	62	0.6789403	0.47
17	78	0.6247041	0.64
18	82	0.9939486	0.96
19	88	0.7164996	0.66
20	89	0.7378308	0.64

## Step 13: Statistics calculation – Linear SGD

Create a new tab by pressing the “+” button on the bottom of the page with the name “STATISTICS\_ACCURACIES\_ Linear SGD”.

Import data into the input spreadsheet of the “STATISTICS\_ACCURACIES\_ Linear SGD” tab from the output of the “VALIDATE\_MODEL\_ Linear SGD” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Calculate the statistical metrics for the regression: *Statistics → Model Metrics → Regression Metrics*



The results will appear on the output spreadsheet.

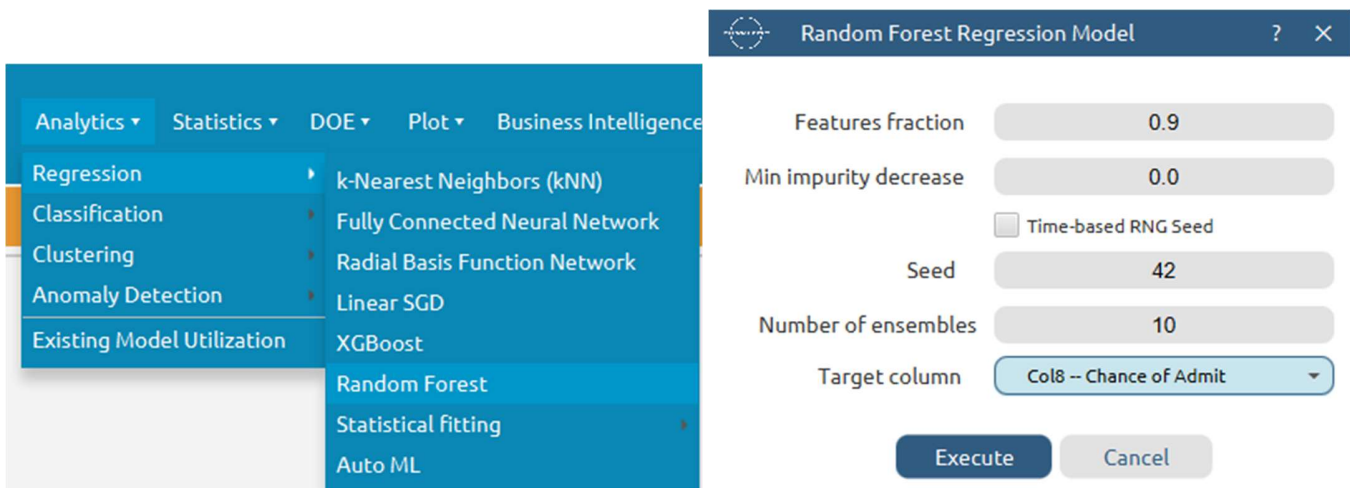
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)
User Header	User Row ID	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R Squared
1		0.0061736	0.0785722	0.0575766	0.8106150

## Step 14: Train the Model – Random Forest

Create a new tab by pressing the “+” button on the bottom of the page with the name “Random Forest”.

Import data into the input spreadsheet of the “Random Forest” tab from the output of the “FEATURE\_SELECTION\_TRAINING\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Use the Linear SGD method to train and fit the model: *Analytics* → *Regression* → *Random Forest*.



The predictions will appear on the output spreadsheet.

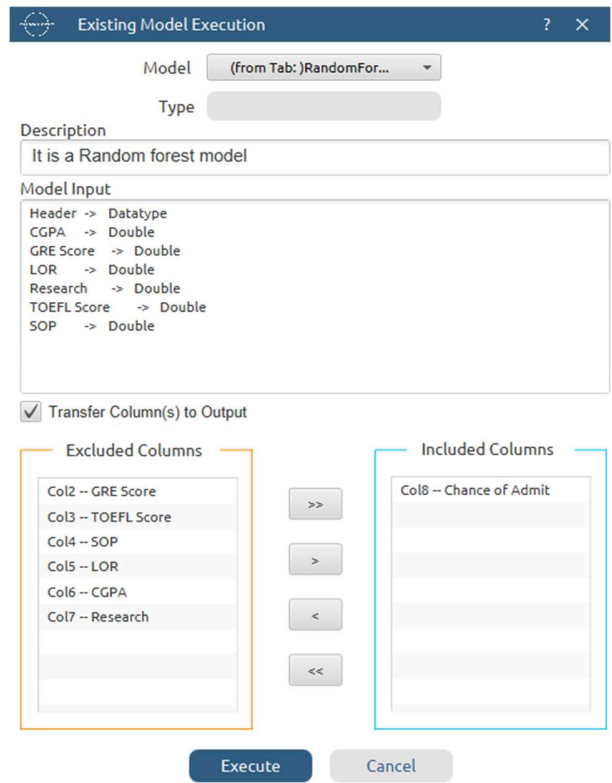
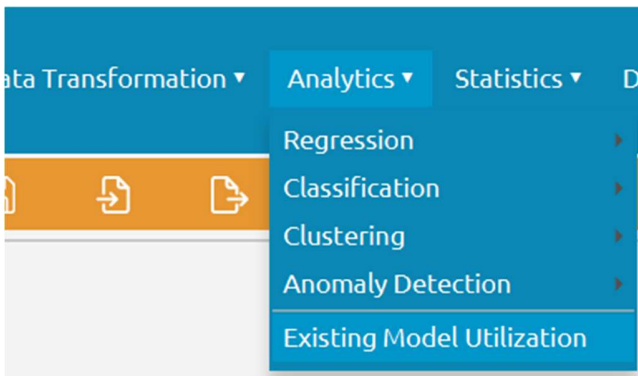
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Chance of Admit	Prediction
1		0.92	0.9287500
2		0.72	0.7168334
3		0.8	0.7910000
4		0.65	0.6641666
5		0.9	0.9055000
6		0.75	0.7335833
7		0.68	0.6745000
8		0.45	0.5680000
9		0.52	0.7175000
10		0.84	0.8295000
11		0.62	0.5733333
12		0.61	0.6010000
13		0.66	0.6970000
14		0.65	0.6673333
15		0.63	0.6503333
16		0.95	0.9490000
17		0.97	0.9660000
18		0.76	0.7443333
19		0.44	0.4650000
20		0.46	0.4485833

## Step 15: Validate the Model – Random Forest

Create a new tab by pressing the “+” button on the bottom of the page with the name “VALIDATE\_MODEL\_Random Forest”.

Import data into the input spreadsheet of the “VALIDATE\_MODEL\_Random Forest” 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: *Analytics* → *Existing Model Utilization* → *Model (from Tab:) Random Forest (.fit)*.  
Choose the column “Chance of Admit” to be transferred to the output spreadsheet.



The predictions will appear on the output spreadsheet.

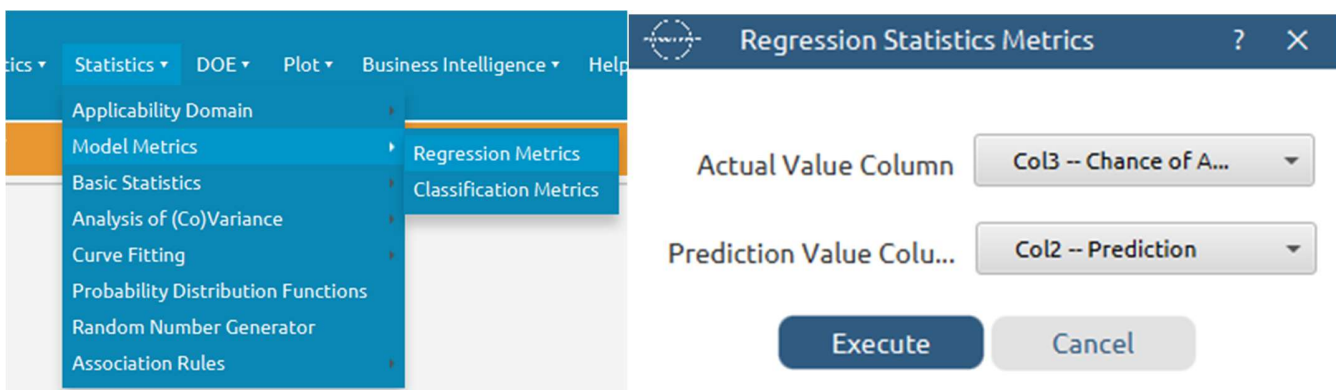
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Prediction	Chance of Admit
1	2	0.7788333	0.76
2	9	0.5603333	0.5
3	13	0.8426667	0.78
4	16	0.6865000	0.54
5	20	0.6705000	0.62
6	21	0.6180000	0.64
7	22	0.5915000	0.7
8	23	0.9250000	0.94
9	26	0.9359167	0.94
10	42	0.6860000	0.49
11	51	0.6875	0.76
12	52	0.5792500	0.56
13	54	0.7300000	0.72
14	59	0.4545000	0.36
15	61	0.6443333	0.48
16	62	0.6530000	0.47
17	78	0.6390000	0.64
18	82	0.9372500	0.96
19	88	0.6982500	0.66
20	89	0.6680000	0.64

## Step 16: Statistics calculation – Random Forest

Create a new tab by pressing the “+” button on the bottom of the page with the name “STATISTICS\_ACCURACIES\_Random Forest”.

Import data into the input spreadsheet of the “STATISTICS\_ACCURACIES\_Random Forest” tab from the output of the “VALIDATE\_MODEL\_Random Forest” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Calculate the statistical metrics for the regression: *Statistics → Model Metrics → Regression Metrics*



The results will appear on the output spreadsheet.

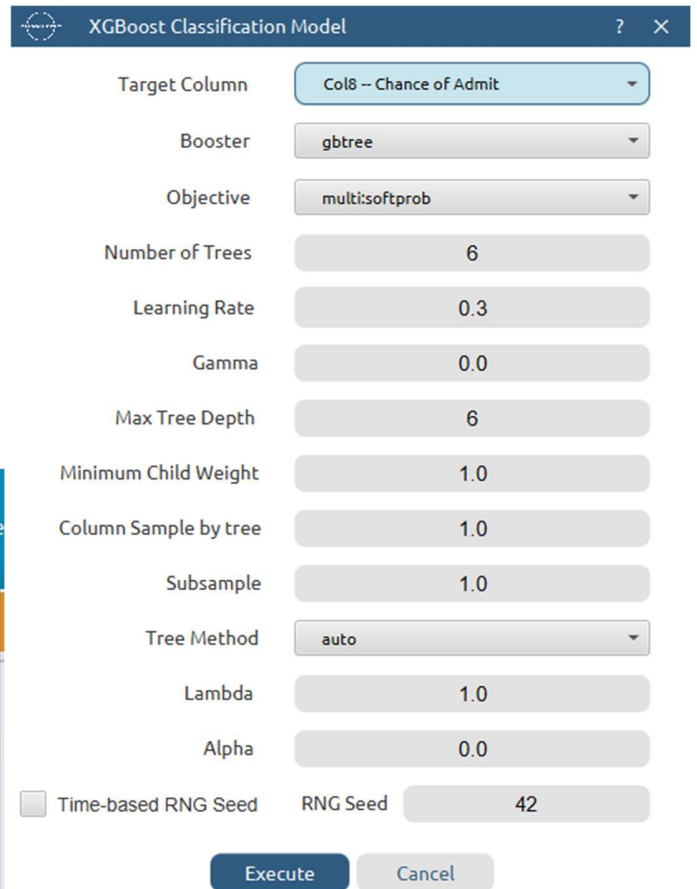
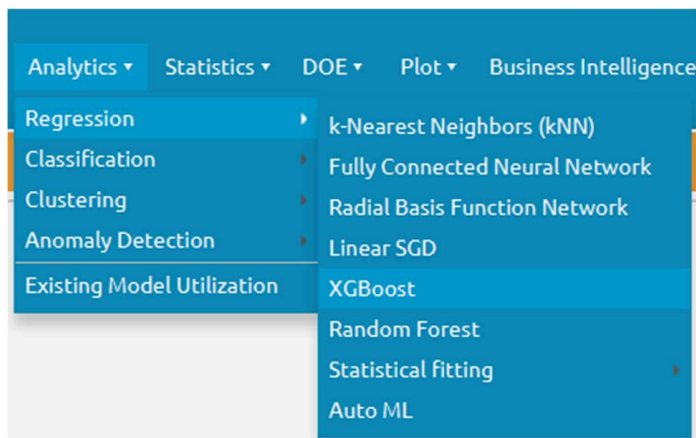
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)
User Header	User Row ID	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R Squared
1		0.0049985	0.0707001	0.0521033	0.7523433

## Step 17: Train the Model – XGBoost

Create a new tab by pressing the “+” button on the bottom of the page with the name “XGBoost”.

Import data into the input spreadsheet of the “XGBoost” tab from the output of the “FEATURE\_SELECTION\_TRAINING\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Use the Linear SGD method to train and fit the model: *Analytics → Regression → XGBoost*.



The predictions will appear on the output spreadsheet.

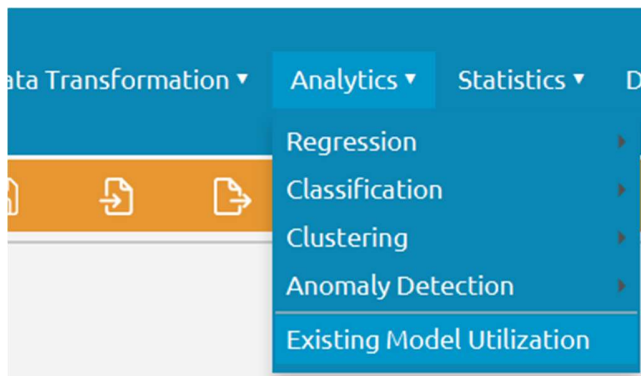
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Chance of Admit	Prediction
1		0.92	0.94
2		0.72	0.71
3		0.8	0.8
4		0.65	0.65
5		0.9	0.93
6		0.75	0.75
7		0.68	0.68
8		0.45	0.79
9		0.52	0.78
10		0.84	0.84
11		0.62	0.62
12		0.61	0.65
13		0.66	0.66
14		0.65	0.65
15		0.63	0.78
16		0.95	0.95
17		0.97	0.97
18		0.76	0.76
19		0.44	0.46
20		0.46	0.47

## Step 18: Validate the Model – XGBoost

Create a new tab by pressing the “+” button on the bottom of the page with the name “VALIDATE\_MODEL\_Random Forest”.

Import data into the input spreadsheet of the “VALIDATE\_MODEL\_XGBoost” 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: *Analytics* → *Existing Model Utilization* → *Model (from Tab:) XGBoost (.fit)*. Choose the column “Chance of Admit” to be transferred to the output spreadsheet.



The 'Existing Model Execution' dialog box is shown. At the top, the 'Model' is set to '(from Tab:)XGBoost' and the 'Type' is 'XGBoost classification mod'. Below this is a 'Description' field. The 'Model Input' section lists the following columns and datatypes: Header -> Datatype, CGPA -> Double, GRE Score -> Integer, LOR -> Double, Research -> Integer, TOEFL Score -> Integer, and SOP -> Double. A checkbox labeled 'Transfer Column(s) to Output' is checked. Below this are two lists: 'Excluded Columns' (containing Col2 -- GRE Score, Col3 -- TOEFL Score, Col4 -- SOP, Col5 -- LOR, Col6 -- CGPA, and Col7 -- Research) and 'Included Columns' (containing Col8 -- Chance of Admit). Navigation buttons (>>, >, <, <<) are positioned between the lists. At the bottom are 'Execute' and 'Cancel' buttons.

The predictions will appear on the output spreadsheet.

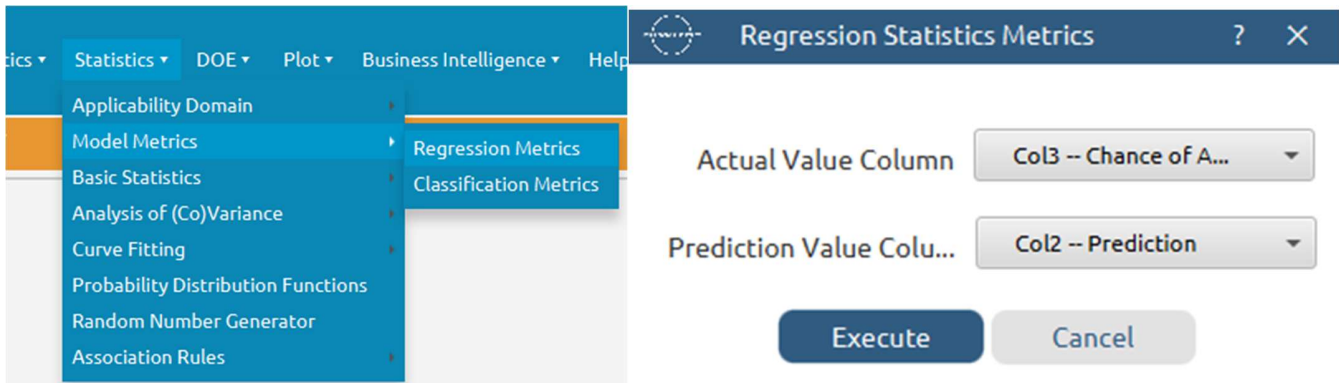
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Prediction	Chance of Admit
1	2	0.79	0.76
2	9	0.58	0.5
3	13	0.84	0.78
4	16	0.73	0.54
5	20	0.63	0.62
6	21	0.52	0.64
7	22	0.75	0.7
8	23	0.93	0.94
9	26	0.94	0.94
10	42	0.65	0.49
11	51	0.74	0.76
12	52	0.58	0.56
13	54	0.65	0.72
14	59	0.46	0.36
15	61	0.65	0.48
16	62	0.64	0.47
17	78	0.62	0.64
18	82	0.94	0.96
19	88	0.74	0.66
20	89	0.72	0.64

## Step 19: Statistics calculation – XGBoost

Create a new tab by pressing the “+” button on the bottom of the page with the name “STATISTICS\_ACCURACIES\_XGBoost”.

Import data into the input spreadsheet of the “STATISTICS\_ACCURACIES\_XGBoost” tab from the output of the “VALIDATE\_MODEL\_XGBoost” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Calculate the statistical metrics for the regression: *Statistics → Model Metrics → Regression Metrics*



The results will appear on the output spreadsheet.

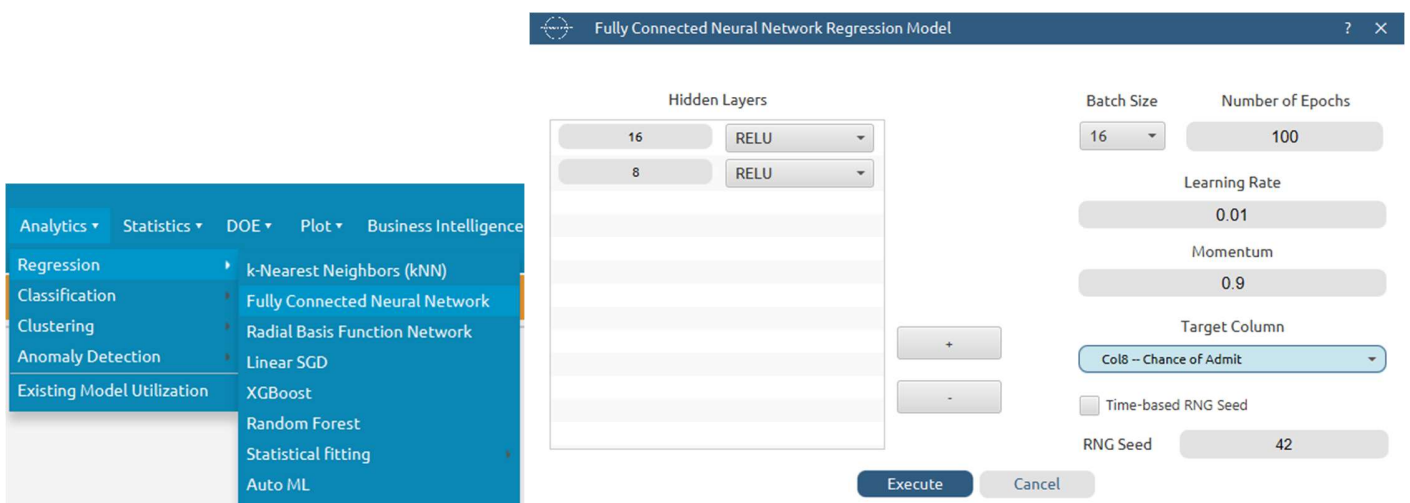
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)
<b>User Header</b>	<b>User Row ID</b>	<b>Mean Squared Error</b>	<b>Root Mean Squared Error</b>	<b>Mean Absolute Error</b>	<b>R Squared</b>
<b>1</b>		0.0056860	0.0754056	0.0562000	0.7316792

## Step 20: Train the Model- MLP

Create a new tab by pressing the “+” button on the bottom of the page with the name “MLP”.

Import data into the input spreadsheet of the “MLP” tab from the output of the “NORMALIZE\_TRAINING\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Use the Linear SGD method to train and fit the model: *Analytics* → *Regression* → *Fully Connected Neural Network*



The predictions will appear on the output spreadsheet.

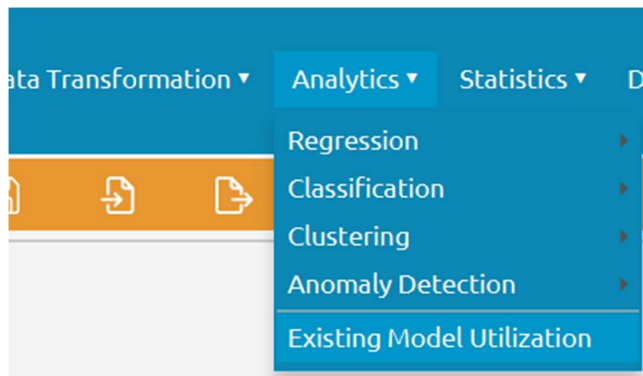
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Chance of Admit	Prediction
1		0.92	0.9476
2		0.72	0.6747
3		0.8	0.7079
4		0.65	0.6651
5		0.9	0.8696
6		0.75	0.7225
7		0.68	0.6374
8		0.45	0.6961
9		0.52	0.7292
10		0.84	0.8271
11		0.62	0.6688
12		0.61	0.6775
13		0.66	0.7106
14		0.65	0.6848
15		0.63	0.7009
16		0.95	0.9703
17		0.97	0.9365
18		0.76	0.7214
19		0.44	0.4852
20		0.46	0.4516

## Step 21: Validate the Model – MLP

Create a new tab by pressing the “+” button on the bottom of the page with the name “VALIDATE\_MODEL\_MLP”.

Import data into the input spreadsheet of the “VALIDATE\_MODEL\_MLP” 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:) MLP (.fit)*. Choose the column “Chance of Admit” to be transferred to the output spreadsheet.



The 'Existing Model Execution' dialog box is shown. It has a title bar with a question mark and a close button. The 'Model' dropdown is set to '(from Tab: )MLP' and the 'Type' is 'MultiLayerPerceptron Mode'. The 'Description' field contains the text 'It is a Fully Connected Neural Network Model'. The 'Model Input' section lists the following variables and their datatypes: Header -> Datatype, CGPA -> Double, GRE Score -> Double, LOR -> Double, Research -> Double, TOEFL Score -> Double, and SOP -> Double. A checkbox labeled 'Transfer Column(s) to Output' is checked. Below this are two columns: 'Excluded Columns' and 'Included Columns'. The 'Excluded Columns' list includes: Col2 -- GRE Score, Col3 -- TOEFL Score, Col4 -- SOP, Col5 -- LOR, Col6 -- CGPA, and Col7 -- Research. The 'Included Columns' list includes: Col8 -- Chance of Admit. Between the columns are four arrow buttons: '>>', '>', '<', and '<<'. At the bottom are 'Execute' and 'Cancel' buttons.

The predictions will appear on the output spreadsheet.

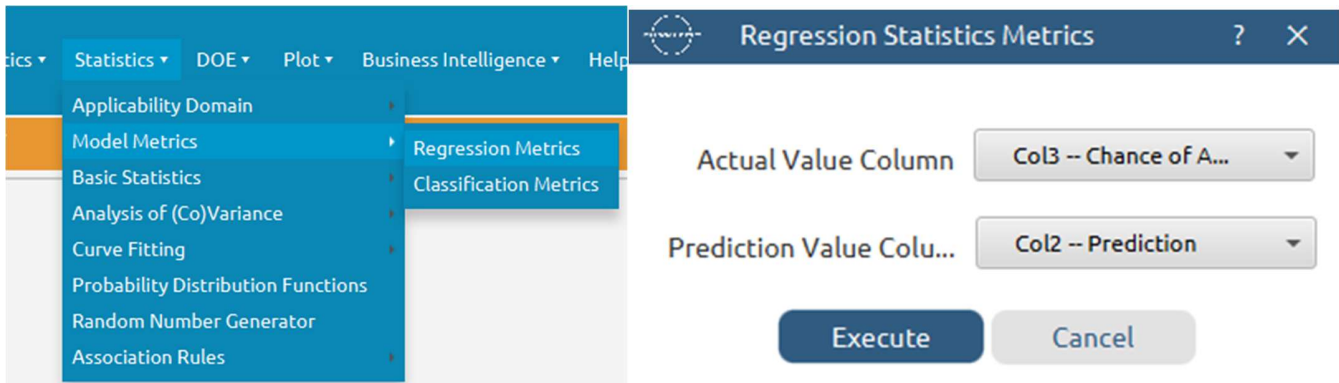
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Prediction	Chance of Admit
1		0.7898	0.76
2		0.5447	0.5
3		0.8424	0.78
4		0.6511	0.54
5		0.6395	0.62
6		0.582	0.64
7		0.6742	0.7
8		0.9393	0.94
9		0.9627	0.94
10		0.6097	0.49
11		0.6774	0.76
12		0.5875	0.56
13		0.6654	0.72
14		0.4899	0.36
15		0.6268	0.48
16		0.6375	0.47
17		0.5712	0.64
18		0.9798	0.96
19		0.6699	0.66
20		0.6619	0.64

## Step 22: Statistics calculation – MLP

Create a new tab by pressing the “+” button on the bottom of the page with the name “STATISTICS\_ACCURACIES\_MLP”.

Import data into the input spreadsheet of the “STATISTICS\_ACCURACIES\_MLP” tab from the output of the “VALIDATE\_MODEL\_MLP” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Calculate the statistical metrics for the regression: *Statistics → Model Metrics → Regression Metrics*



The results will appear on the output spreadsheet.

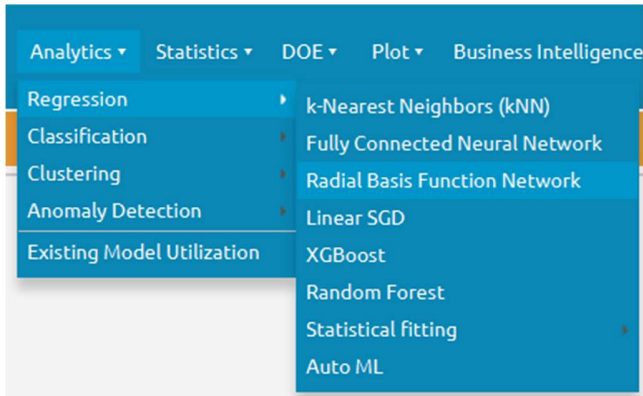
	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)
User Header	User Row ID	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R Squared
1		0.0042195	0.0649576	0.0466730	0.7894110

## Step 23: Train the Model – RBF

Create a new tab by pressing the “+” button on the bottom of the page with the name “RBF”.

Import data into the input spreadsheet of the “RBF” tab from the output of the “NORMALIZE\_TRAINING\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Use the Linear SGD method to train and fit the model: *Analytics* → *Regression* → *Radial Basis Function Network*



RBF Network Regression Model
?
✕

Hidden Neurons

RBK Kernel

Epsilon

$$\varphi(r) = e^{-(\epsilon r)^2}$$

Point Selection  Random Points from Training Set  Use KMeans

Time-based RNG Seed

RNG Seed

Target Column

The predictions will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Chance of Admit	Prediction
1		0.92	0.9233205
2		0.72	0.6638773
3		0.8	0.7065851
4		0.65	0.6460793
5		0.9	0.8560105
6		0.75	0.6774333
7		0.68	0.6521541
8		0.45	0.7138320
9		0.52	0.6871424
10		0.84	0.8542375
11		0.62	0.6541967
12		0.61	0.6564722
13		0.66	0.6970250
14		0.65	0.6468953
15		0.63	0.7016721
16		0.95	0.9172216
17		0.97	0.8746523
18		0.76	0.7313897
19		0.44	0.6209212
20		0.46	0.4704350

## Step 24: Validate the Model – RBF

Create a new tab by pressing the “+” button on the bottom of the page with the name “VALIDATE\_MODEL\_RBF”.

Import data into the input spreadsheet of the “VALIDATE\_MODEL\_RBF” 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:) RBF (.fit)*. Choose the column “Chance of Admit” to be transferred to the output spreadsheet.

The image shows two parts of the software interface. On the left is a portion of the 'Analytics' menu, which includes options like 'Regression', 'Classification', 'Clustering', 'Anomaly Detection', and 'Existing Model Utilization'. On the right is the 'Existing Model Execution' dialog box. The 'Model' dropdown is set to '(from Tab:) RBF'. The 'Description' field contains the text: 'It is a Radial Basis Function model for regression'. The 'Model Input' section lists several variables: Header (Datatype), CGPA (Double), GRE Score (Double), LOR (Double), Research (Double), TOEFL Score (Double), and SOP (Double). A checkbox labeled 'Transfer Column(s) to Output' is checked. Below this, there are two columns: 'Excluded Columns' and 'Included Columns'. The 'Excluded Columns' list includes 'Col2 -- GRE Score', 'Col3 -- TOEFL Score', 'Col4 -- SOP', 'Col5 -- LOR', 'Col6 -- CGPA', and 'Col7 -- Research'. The 'Included Columns' list includes 'Col8 -- Chance of Admit'. Navigation buttons (>>, >, <, <<) are positioned between the two columns. At the bottom of the dialog are 'Execute' and 'Cancel' buttons.

The predictions will appear on the output spreadsheet.

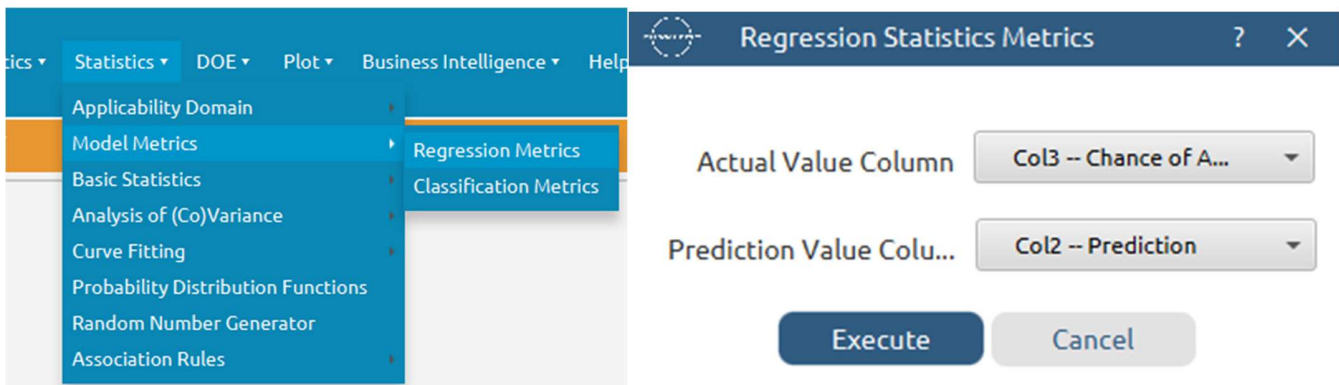
	Col1	Col2 (D)	Col3 (D)
User Header	User Row ID	Prediction	Chance of Admit
1	2	0.7833434	0.76
2	9	0.5349428	0.5
3	13	0.8853141	0.78
4	16	0.6502177	0.54
5	20	0.6449347	0.62
6	21	0.6569140	0.64
7	22	0.6969693	0.7
8	23	0.9223697	0.94
9	26	0.8749672	0.94
10	42	0.6802303	0.49
11	51	0.7139152	0.76
12	52	0.6812115	0.56
13	54	0.6484079	0.72
14	59	0.6309022	0.36
15	61	0.5942574	0.48
16	62	0.6423573	0.47
17	78	0.5488275	0.64
18	82	0.8503617	0.96
19	88	0.6774053	0.66
20	89	0.6850462	0.64

## Step 25: Statistics calculation – RBF

Create a new tab by pressing the “+” button on the bottom of the page with the name “STATISTICS\_ACCURACIES\_RBF”.

Import data into the input spreadsheet of the “STATISTICS\_ACCURACIES\_RBF” tab from the output of the “VALIDATE\_MODEL\_RBF” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet”.

Calculate the statistical metrics for the regression: *Statistics* → *Model Metrics* → *Regression Metrics*



The results will appear on the output spreadsheet.

	Col1	Col2 (D)	Col3 (D)	Col4 (D)	Col5 (D)
User Header	User Row ID	Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R Squared
1		0.0060015	0.0774696	0.0572547	0.7039536

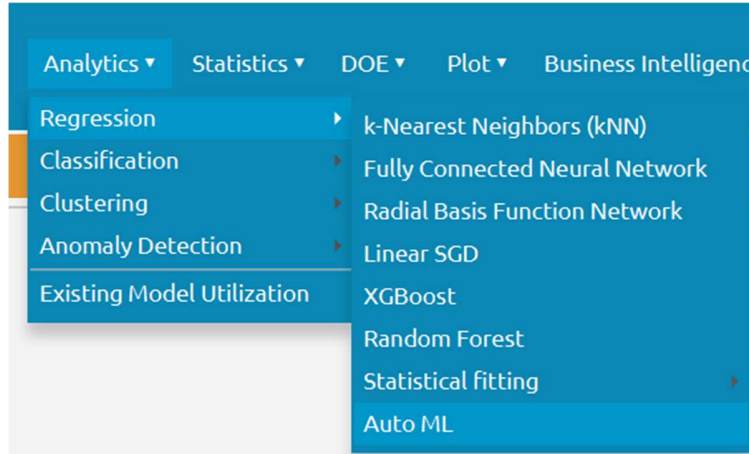
## Step 26: Automated ML optimization

We will compare four machine learning algorithms commonly employed for classification applications: the kNN, Linear SGD, Random Forest, and Radial Basis Function Network models. This procedure can be performed automatically with the Auto ML option of Isalos Analytics Platform, and it is beneficial when optimizing a predictive model.

Create a new tab by pressing the “+” button on the bottom of the page with the name “AutoML.”

Import into the input spreadsheet of the “AutoML” tab the output of the “FEATURE\_SELECTION\_TRAINING\_SET” tab by right-clicking on the input spreadsheet and then choosing “Import from Spreadsheet.”

Perform the algorithm optimisation by choosing: *Analytics* → *Classification* → *Auto ML*.



Select the kNN, Linear SGD, Random Forest, and Radial Basis Function Network models to be used inside the Auto ML configuration. Define the search range of all hyperparameters for each algorithm with the values written in Table 1, by double-clicking on them inside the “Selected Models” box. It should be noted that several parameters are kept constant to decrease computational costs.

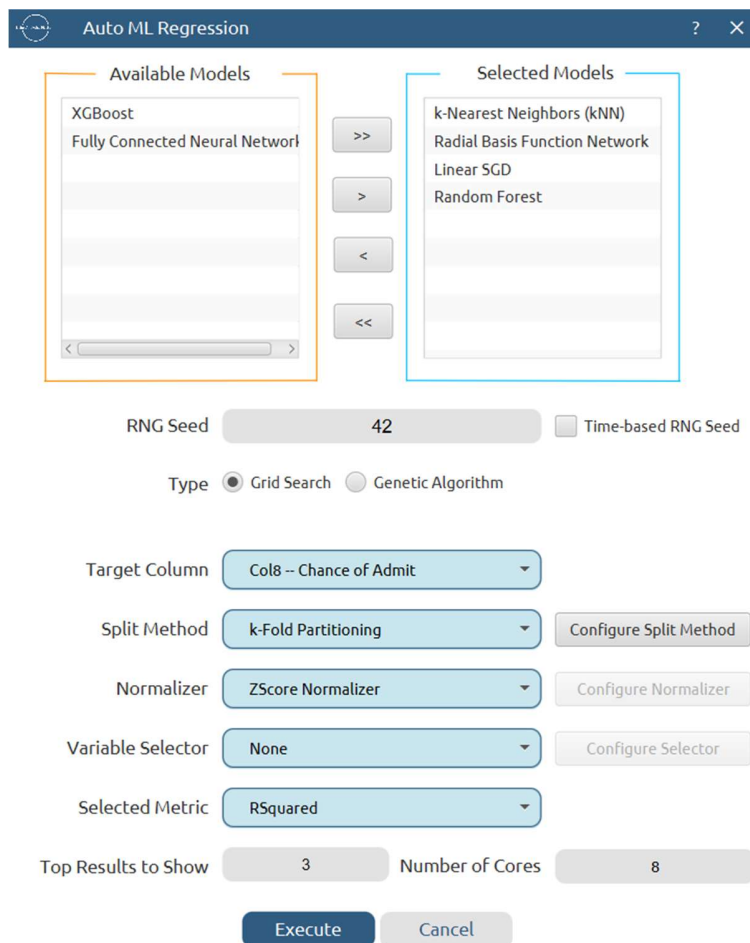
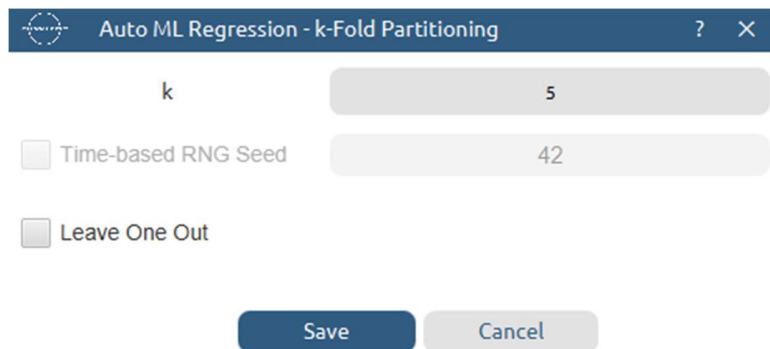


Table 1: Hyperparameter search ranges for model optimization inside the Isalos AutoML scheme

ML method	Hyperparameter	Search range [min,max;step]
kNN	Number of nearest neighbours, k	[5,15,5]
Linear SGD	Loss Function	Squared Loss
	Optimizer	Simple SGD
	Learning rate	[0.005, 0.01, 0.005]
	Number of epochs	[10,15,5]
Random forest	Features fraction	[0.7, 0.9, 0.2]
	Min impurity decrease	[0.0, 0.1, 0.1]
	Number of ensembles	[10,15,5]
Radial Basis Function Network	Hidden Neurons	[4,8,4]
	Point Selection	Random Points from Training
	RBF Kernel	GAUSSIAN
	Epsilon	[0.3, 0.5, 0.2]

Afterwards, employ the grid search method for the exploration of the hyperparameter space, and define the target column, “Chance of Admit.” Choose the split method “k-Fold Partitioning,” and click on the “Configure Split Method” button to select 5 folds.



Choose the z-score normalizer to maintain consistency with the previous preprocessing steps. There is no need for variable selection, so choose the option “None.” Finally, for the selected metric choose “R squared.” The fine-tuned model with the hyperparameters yielding the highest average R squared across the five folds are selected as optimal.

The results will appear on the output spreadsheet. The algorithm indicated as optimal is Linear SGD with Squared Loss as Loss Function , Simple SGD as Optimizer , Learning rate 0.005 and 10 as the number of epochs.

	Col1	Col2 (I)	Col3 (S)	Col4 (S)	Col5 (S)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)	Col10 (D)
User Header	User Row ID	Rank	Model	Description	Value	MAE	RSquared	MAPE	RMSE	MSE
1		1	Linear SGD Regression	Selected Metric	RSquared = 0.8142169961488064	0.0432843	0.8142170	6.9733523	0.0610253	0.0037747
2				Feature Headers	CGPA, GRE Score, LOR, Research, TOEFL Score, SOP					
3				Loss Function	Squared Loss					
4				Optimizer	Simple SGD					
5				Learning Rate	0.005					
6				Number of Epochs	10					
7										
8		2	Linear SGD Regression	Selected Metric	RSquared = 0.8134213371914278	0.0438782	0.8134213	7.0504193	0.0614193	0.0038234
9				Feature Headers	CGPA, GRE Score, LOR, Research, TOEFL Score, SOP					
10				Loss Function	Squared Loss					
11				Optimizer	Simple SGD					

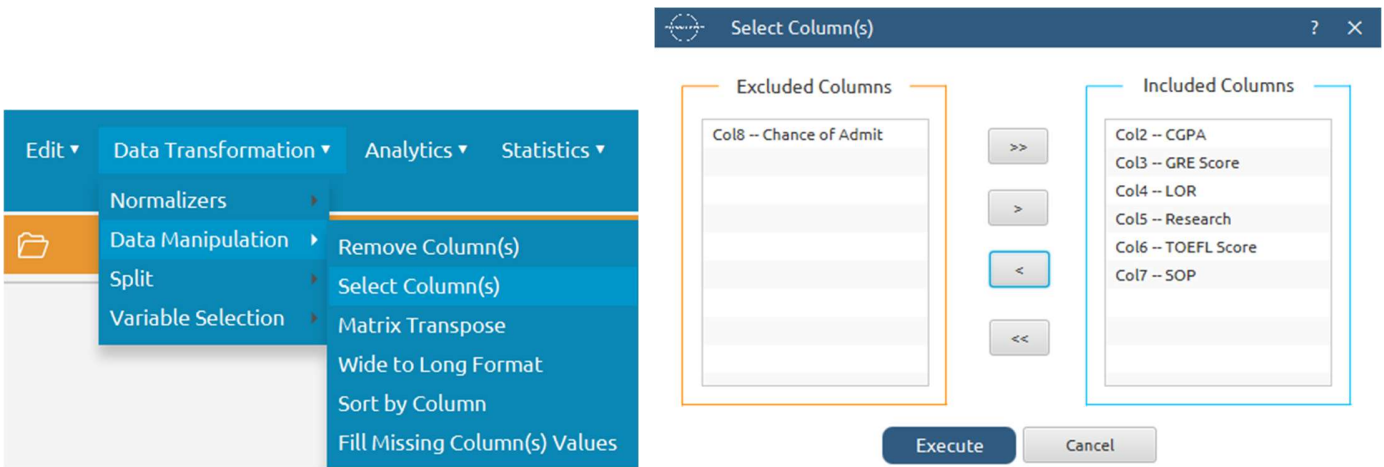
## Step 27: Reliability check for each record of the test set

### Step 27.a: Create the domain

Create a new tab by pressing the “+” button on the bottom of the page with the name “EXCLUDE\_SCORES”.

Import data into the input spreadsheet of the “EXCLUDE\_SCORES” 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 “Performance Index”: *Data Transformation → Data Manipulation → Select Column(s)*

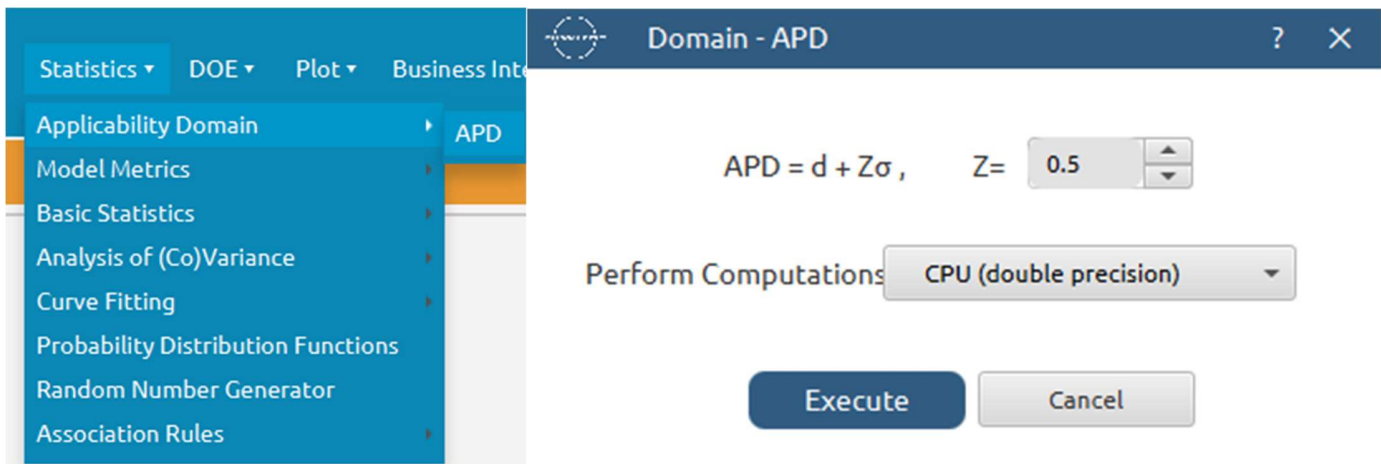


The results will appear on the output spreadsheet.

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\_SCORES” 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.

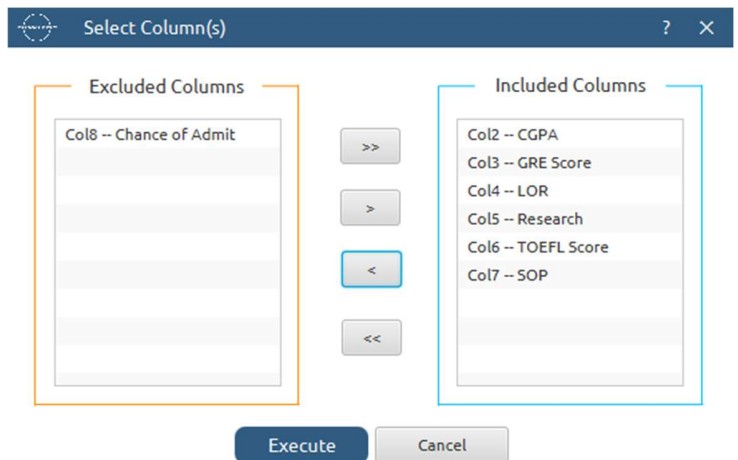
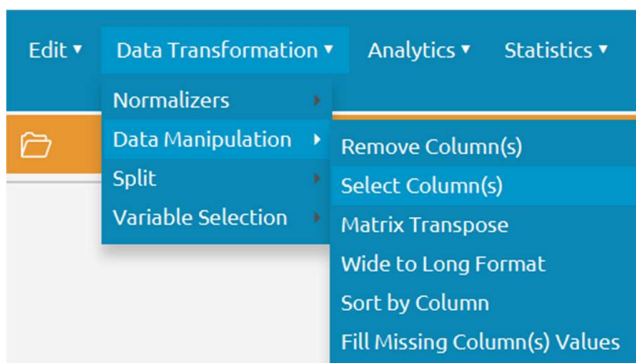
	Col1	Col2 (D)	Col3 (D)	Col4 (S)
User Header	User Row ID	Domain	APD	Prediction
1		0.0	2.5463097	reliable
2		0.0	2.5463097	reliable
3		0.0	2.5463097	reliable
4		0.0	2.5463097	reliable
5		0.0	2.5463097	reliable
6		0.0	2.5463097	reliable
7		0.0	2.5463097	reliable
8		0.0	2.5463097	reliable
9		0.0	2.5463097	reliable
10		0.0	2.5463097	reliable
11		0.0	2.5463097	reliable
12		0.0	2.5463097	reliable
13		0.0	2.5463097	reliable
14		0.0	2.5463097	reliable
15		0.0	2.5463097	reliable
16		0.0	2.5463097	reliable
17		0.0	2.5463097	reliable
18		0.0	2.5463097	reliable
19		0.0	2.5463097	reliable

### Step 27.b: Check the test set reliability

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

Import data into the input spreadsheet of the “EXCLUDE\_SCORES\_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 “Performance Index”: *Data Transformation* → *Data Manipulation* → *Select Column(s)*

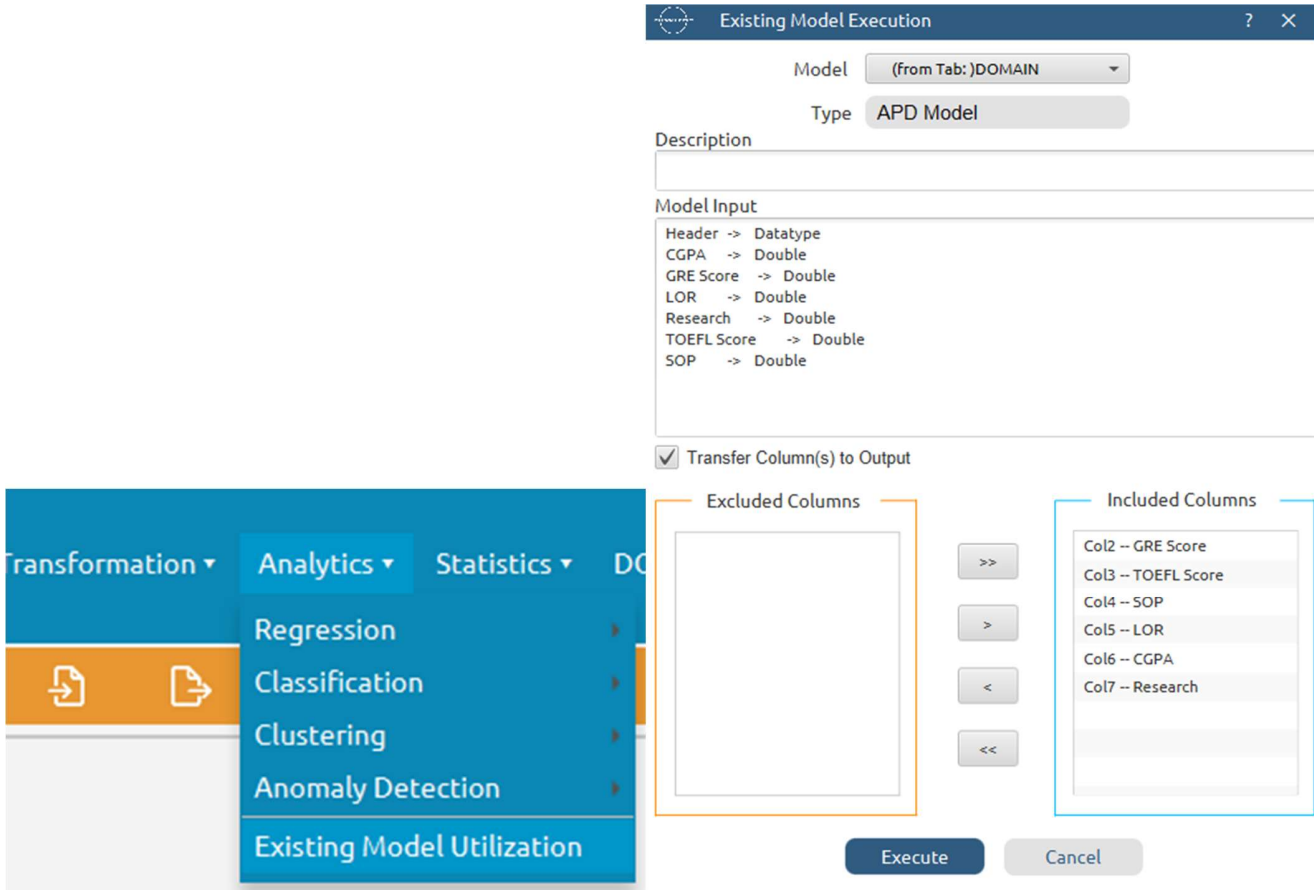


The results will appear on the output spreadsheet.

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\_SCORES\_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. There are no unreliable samples in the test set.

	Col1	Col2 (D)	Col3 (D)	Col4 (S)	Col5 (D)	Col6 (D)	Col7 (D)	Col8 (D)	Col9 (D)
User Header	User Row ID	Domain	APD	Prediction	GRE Score	TOEFL Score	SOP	LOR	CGPA
1	2	0.3659089	2.5463097	reliable	0.6474336	-0.0311195	0.6239770	1.0590905	0.4541842
2	9	0.6840245	2.5463097	reliable	-1.2793217	-0.8500537	-1.3569023	-2.1747737	-0.9596819
3	13	0.2467933	2.5463097	reliable	0.9977527	0.7878147	0.6239770	1.0590905	0.8279648
4	16	0.3897776	2.5463097	reliable	-0.2283643	-0.3586932	0.1287572	-1.0968190	-0.4721419
5	20	0.5638893	2.5463097	reliable	-1.1917419	-0.8500537	0.1287572	-0.5578416	-0.1471152
6	21	0.4927494	2.5463097	reliable	-0.4035239	-0.0311195	-0.3664627	-1.6357963	-1.1221952
7	22	1.1821791	2.5463097	reliable	0.7350133	1.1153883	-0.3664627	-1.6357963	-0.3096285
8	23	0.3272052	2.5463097	reliable	0.9977527	1.4429620	1.6144166	1.5980679	1.4780182
9	26	0.4277138	2.5463097	reliable	2.0487101	2.0981093	1.1191968	1.0590905	1.6405315
10	42	0.8393619	2.5463097	reliable	-0.0532047	-0.3586932	-0.8616825	-1.0968190	-0.6346552
11	51	1.0781526	2.5463097	reliable	-0.3159441	-1.5052010	-0.8616825	1.0590905	-0.4721419
12	52	0.7015507	2.5463097	reliable	-0.4035239	-1.1776273	-1.8521221	-0.0188642	-1.1221952
13	54	1.1141400	2.5463097	reliable	0.6474336	0.7878147	0.6239770	-1.0968190	-0.7971686
14	59	1.4135718	2.5463097	reliable	-1.4544813	-1.3414142	-0.3664627	-1.6357963	-2.9098419
15	61	0.4629384	2.5463097	reliable	-0.6662632	-1.1776273	-0.3664627	-0.5578416	-0.7971686
16	62	0.5849420	2.5463097	reliable	-0.8414228	-1.0138405	0.6239770	-0.5578416	-0.6346552
17	78	0.8896332	2.5463097	reliable	-1.3669015	-1.3414142	-0.3664627	-1.6357963	-0.6021525
18	82	0.3135155	2.5463097	reliable	2.0487101	2.0981093	1.6144166	1.5980679	1.4780182
19	88	0.5606813	2.5463097	reliable	0.0343751	-0.0311195	0.1287572	-0.5578416	-0.5046445
20	89	0.5674801	2.5463097	reliable	-0.2283643	0.1326673	1.1191968	-0.0188642	-0.7321632
21	91	1.0065795	2.5463097	reliable	0.1219548	-0.1949063	0.6239770	0.5201132	-1.0896926

## Final Isalos Workflow

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

