



Simple Operation Manual of EasyDS

By: Huijun Hou

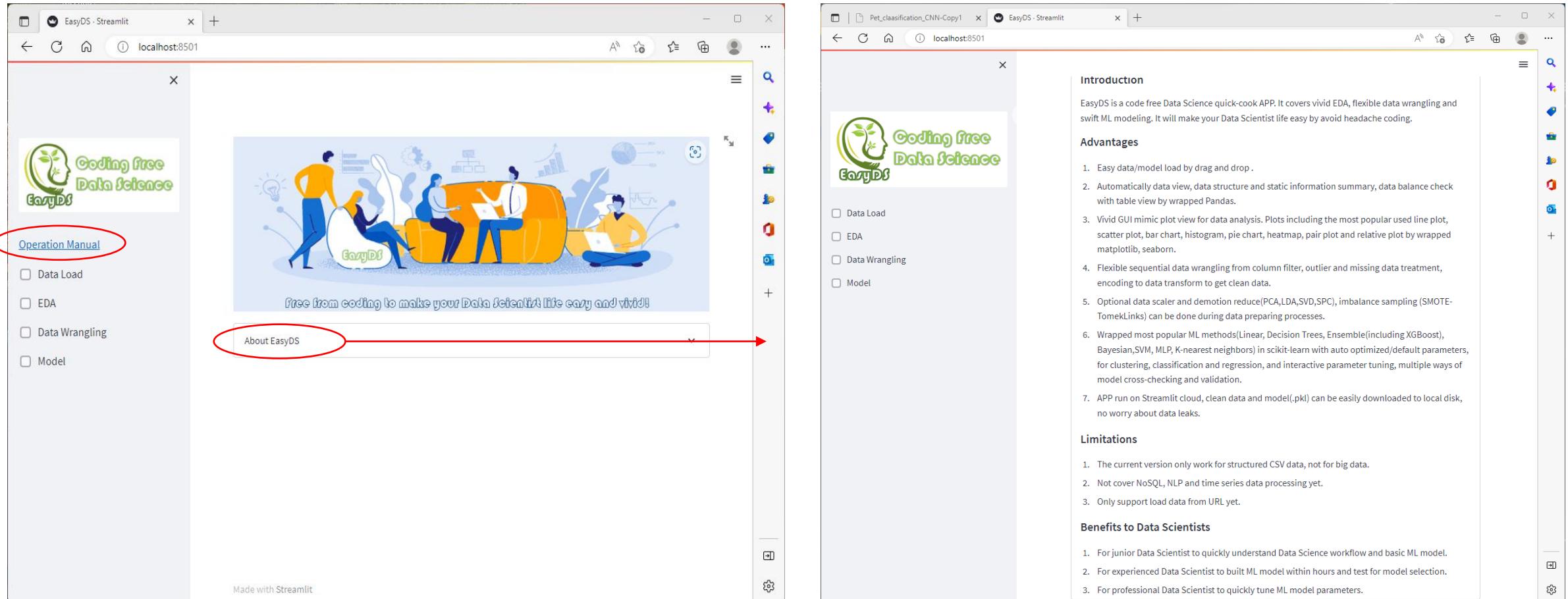


Contents:

- About EasyDS
- Load Data
- Data View in Table
- Data Info Summary
- Data Balance Check
- Plot View
- Data Wrangling
- Data Preparing for Modeling
- Clustering
- Classification
- Regression
- Load Existed Model



About EasyDS



Open EasyDS, left side plot displayed the main menus. Operation manual can be checked by click on “Operation Manual”. A brief introduction is hide under ‘About EasyDS’. You can click it to have a quick look(as displayed in right side plot), and click once more to hide it.

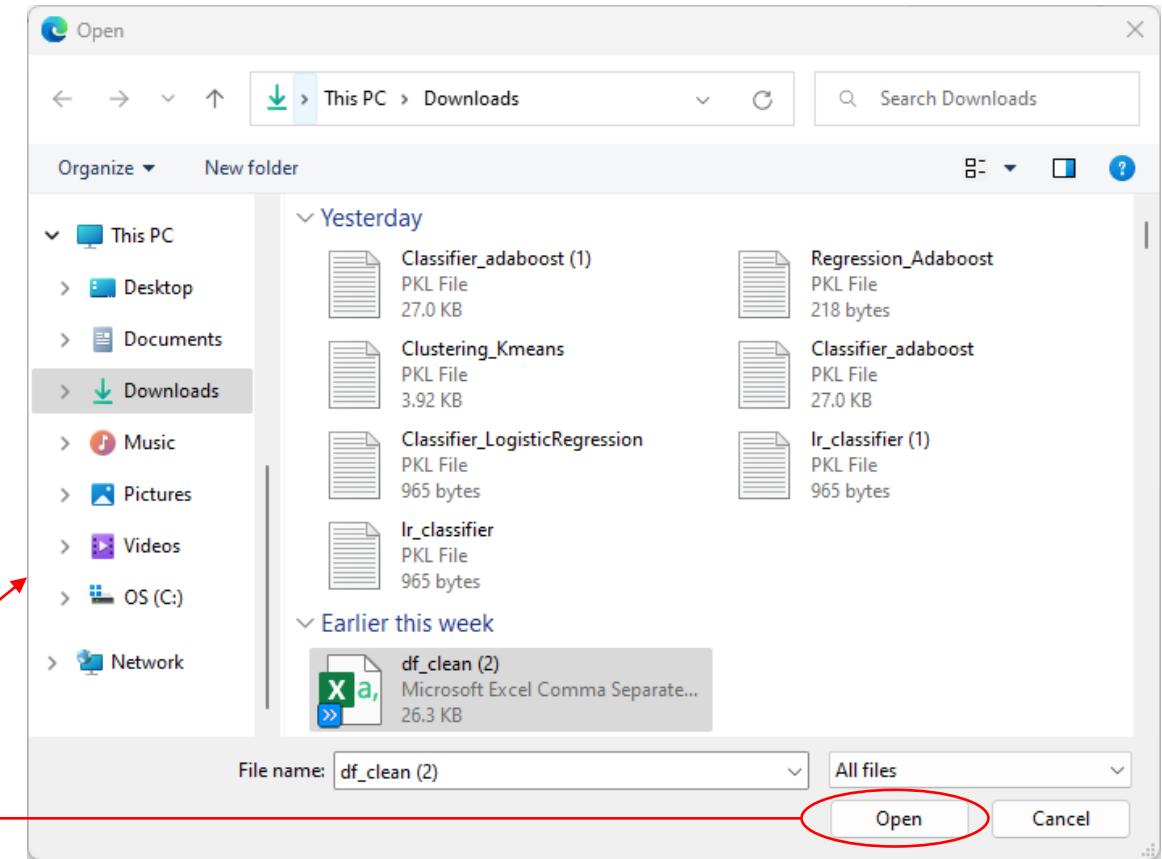
Load Data

This screenshot shows the Streamlit interface for the EasyDS application. On the left, there's a sidebar with the 'EasyDS' logo and a list of options: 'Data Load' (which is checked), 'EDA', 'Data Wrangling', and 'Model'. A red oval highlights the 'Data Load' checkbox. A red arrow points from this oval to a red oval around the 'Load Raw Data' button in the main content area. The main content area features a banner with the text 'Coding free Data Science' and 'Free from coding to make your Data Scientist life easy and vivid!'. Below the banner is a 'Data Load' section with two buttons: 'Load Raw Data' and 'Load Clean Data'. The 'Load Raw Data' button is highlighted with a red oval and has the word 'click' written over it. At the bottom of the page, it says 'Made with Streamlit'.

This screenshot shows the same Streamlit interface after a click on the 'Load Raw Data' button. The 'Data Load' section has expanded to show a detailed form for loading raw data. The form includes a header 'Load Raw Data' and 'Raw data with CSV format', followed by a file upload input field with the placeholder 'Drag and drop file here' and 'Limit 200MB per file', and a 'Browse files' button. A red arrow points from the red oval in the previous screenshot to this expanded form area.

Load Data

The screenshot shows the EasyDS Streamlit application interface. On the left, there's a sidebar with a logo and navigation links: Data Load (selected), EDA, Data Wrangling, and Model. The main area features a banner with the text "free from coding to make your Data Scientist life easy and vivid!" and an illustration of four people working with data. Below the banner is a "Data Load" section with tabs for "Load Raw Data" (selected) and "Load Clean Data". Under "Load Raw Data", there's a "Raw data with CSV format" section with a "Drag and drop file here" input field and a "Browse files" button. A red arrow points from the "df_clean (2).csv" file listed below the input field towards the "Browse files" button.



Data View in Table

This screenshot shows the initial interface of the EasyDS Streamlit application. On the left, there's a sidebar with three main options: "Data Load" (checked), "EDA" (checked), "Data Wrangling" (unchecked), and "Model" (unchecked). A red circle highlights the "EDA" checkbox, and a red arrow points from this circle to the "EDA" tab in the main content area. The "EDA" tab is currently active, displaying sub-options: "Load Raw Data" (selected), "Load Clean Data" (unchecked), and "Table View" (selected). Below these are tabs for "FeatureInfo", "Summary", "BalanceCheck", and "PlotView". The main content area features a banner with the text "Coding free Data Science" and "EasyDS", followed by a cartoon illustration of people working with data. A sub-banner below the illustration says "Free from coding to make your Data Scientist life easy and vivid!". At the bottom of the sidebar, there are several small icons.

This screenshot shows the application after interacting with the "Table View" option in the EDA tab. The "Data Load" tab is now active, showing "Load Raw Data" (selected) and "Load Clean Data" (unchecked). The "EDA" tab is still present but inactive. In the main content area, the "Table View" section is highlighted with a red arrow. It displays a table titled "raw_data" with the "Head" view selected. The table contains the following data:

	Preg	Glucose	Bloo	Skin	Insulin	BMI	Diabete	Age	Outc
0	6	148	72	35	0	33.6000	0.6270	50	1
1	1	85	66	29	0	26.6000	0.3510	31	0
2	8	183	64	0	0	23.3000	0.6720	32	1
3	1	89	66	23	94	28.1000	0.1670	21	0
4	0	137	40	35	168	43.1000	2.2880	33	1

Below the table, a message states "Data Shape(Row,Col) ~ (768, 9);". The sidebar on the right side of the interface is identical to the one in the first screenshot.

Data View in Table

EasyDS · Streamlit

localhost:8501

Data Load

Load Raw Data Load Clean Data

Load Raw Data

EDA

TableView FeatureInfo Summary BalanceCheck PlotView

Table View

raw_data Head

	Preg	Glucose	Blo	Skin	Insulin	BMI
0	6	148	72	35	0	33.6000
1	1	85	66	29	0	26.6000
2	8	183	64	0	0	23.3000
3	1	89	66	23	94	28.1000
4	0	137	40	35	168	43.1000

Data Shape(Row,Col) ~ (768, 9);

select data

EasyDS · Streamlit

localhost:8501

Data Load

Load Raw Data Load Clean Data

Load Raw Data

EDA

TableView FeatureInfo Summary BalanceCheck PlotView

Table View

raw_data Head

Head
Tail
Full

	Preg	Glucose	Blo	Skin	Insulin	BMI
0	6	148	72	35	0	33.6000
1	1	85	66	29	0	26.6000
2	8	183	64	0	0	23.3000
3	1	89	66	23	94	28.1000
4	0	137	40	35	168	43.1000

Data Shape(Row,Col) ~ (768, 9);

select table view method

Data Info Summary

The screenshot shows the Streamlit interface for the EasyDS project. The sidebar on the left has checkboxes for Data Load (checked), EDA (checked), Data Wrangling (unchecked), and Model (unchecked). The main area is titled 'EDA' and has tabs for TableView, FeatureInfo (which is selected), Summary, BalanceCheck, and PlotView. Below the tabs is a table with columns: Col Name, Data Type, Counts, Null counts, and Null ratio(%). The table data is as follows:

Col Name	Data Type	Counts	Null counts	Null ratio(%)
Pregnancies	int64	768	0	0.0
Glucose	int64	768	0	0.0
BloodPressure	int64	768	0	0.0
SkinThickness	int64	768	0	0.0
Insulin	int64	768	0	0.0
BMI	float64	768	0	0.0
DiabetesPedigreeFunction	float64	768	0	0.0
Age	int64	768	0	0.0
Outcome	int64	768	0	0.0

Column information

The screenshot shows the Streamlit interface for the EasyDS project. The sidebar on the left has checkboxes for Data Load (checked), EDA (checked), Data Wrangling (unchecked), and Model (unchecked). The main area is titled 'EDA' and has tabs for TableView, FeatureInfo, Summary (which is selected), BalanceCheck, and PlotView. Below the tabs is a table with columns: count, mean, std, min, 25%, 50%, 75%, and max. The table data is as follows:

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0000	3.8451	3.3696	0.0000	1.0000	3.0000	6.0000	17.0000
Glucose	768.0000	120.8945	31.9726	0.0000	99.0000	117.0000	140.2500	199.0000
BloodPressure	768.0000	69.1055	19.3558	0.0000	62.0000	72.0000	80.0000	122.0000
SkinThickness	768.0000	20.5365	15.9522	0.0000	0.0000	23.0000	32.0000	99.0000
Insulin	768.0000	79.7995	115.2440	0.0000	0.0000	30.5000	127.2500	846.0000
BMI	768.0000	31.9926	7.8842	0.0000	27.3000	32.0000	36.6000	67.1000
DiabetesPedigreeFunction	768.0000	0.4719	0.3313	0.0780	0.2438	0.3725	0.6263	2.4200
Age	768.0000	33.2409	11.7602	21.0000	24.0000	29.0000	41.0000	81.0000
Outcome	768.0000	0.3490	0.4770	0.0000	0.0000	1.0000	1.0000	1.0000

Statistical summary of numerical columns

Data Balance Check

EasyDS - Streamlit

localhost:8501

Drag and drop file here
Limit 200MB per file

Browse files

df_clean (2).csv 26.4KB

EDA

Data Load

EDA

Data Wrangling

Model

BalanceCheck Balance Check Active

Please input data column name (target class/labels)

Outcome

Pregnancies

Glucose

BloodPressure

SkinThickness

Insulin

Age

Outcome

41.1%
58.9%

Select Target for data balance check

EasyDS - Streamlit

localhost:8501

Data Load

Load Raw Data

Load Clean Data

Load Raw Data

EDA

Data Load

EDA

Data Wrangling

Model

BalanceCheck Balance Check Active

Please input data column name (target class/labels)

Outcome

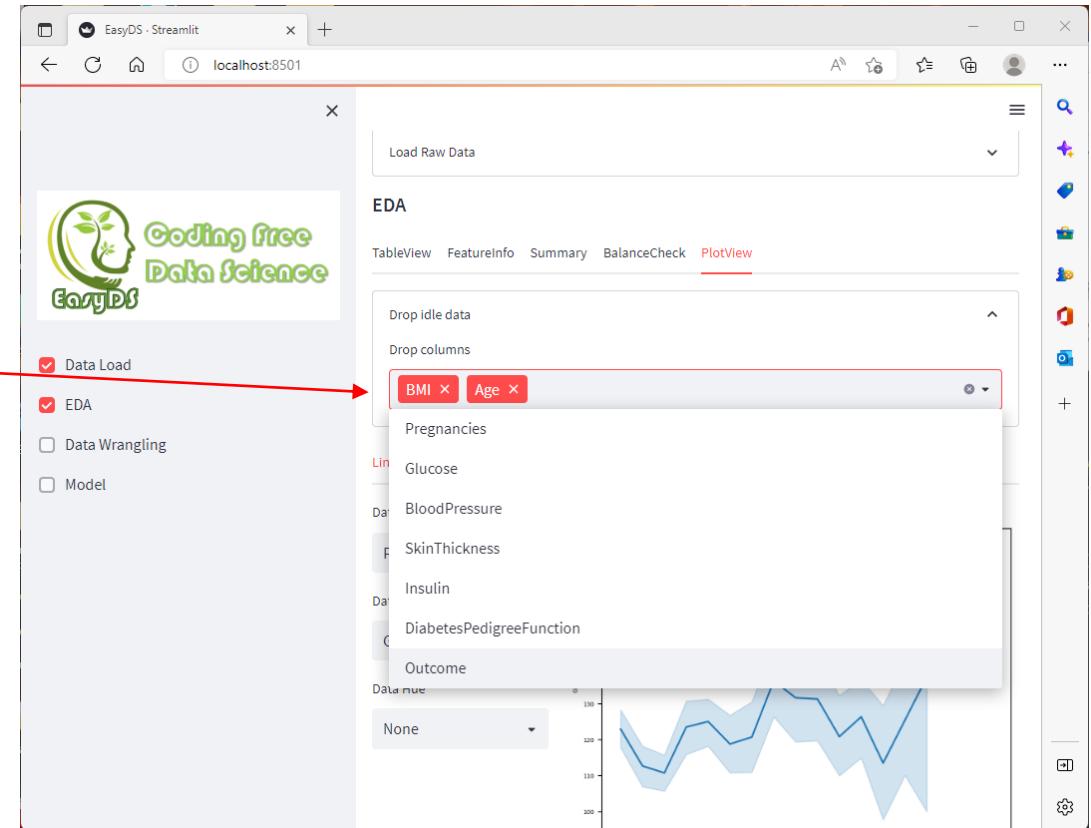
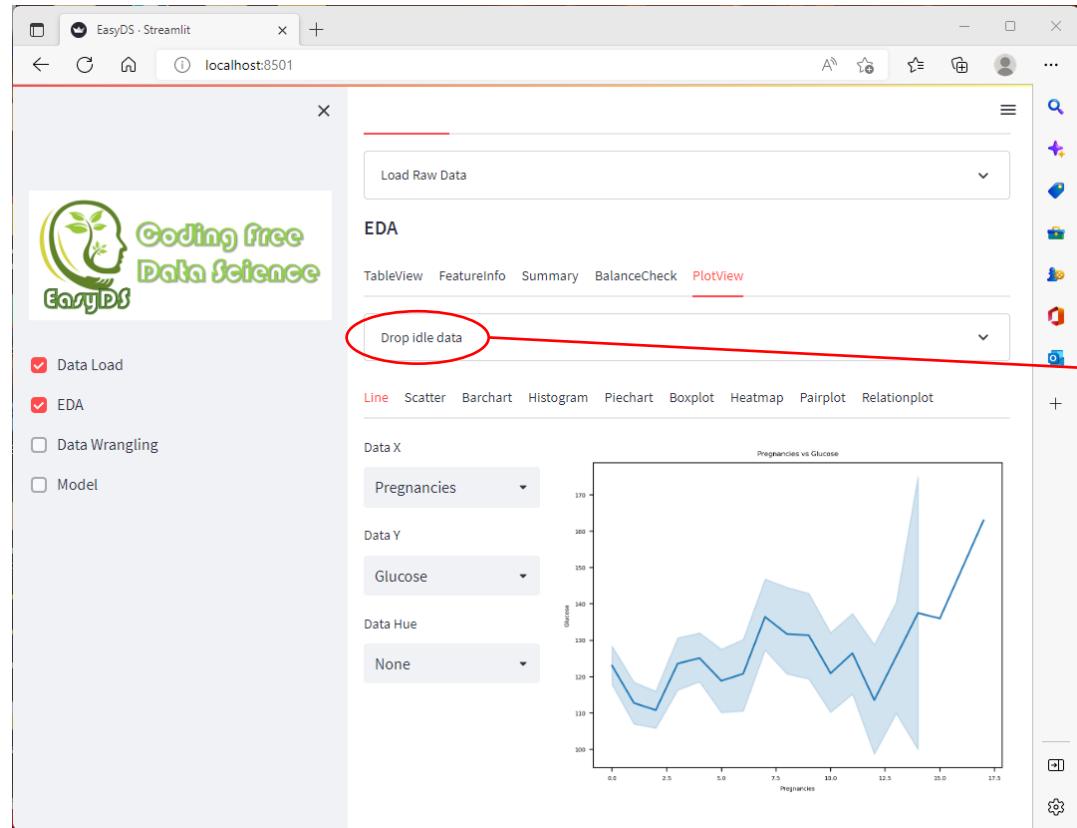
Balance Matrix:

	Class	Counts	Ratio
0	1	500	65.1%
1	0	268	34.9%

41.1%
58.9%

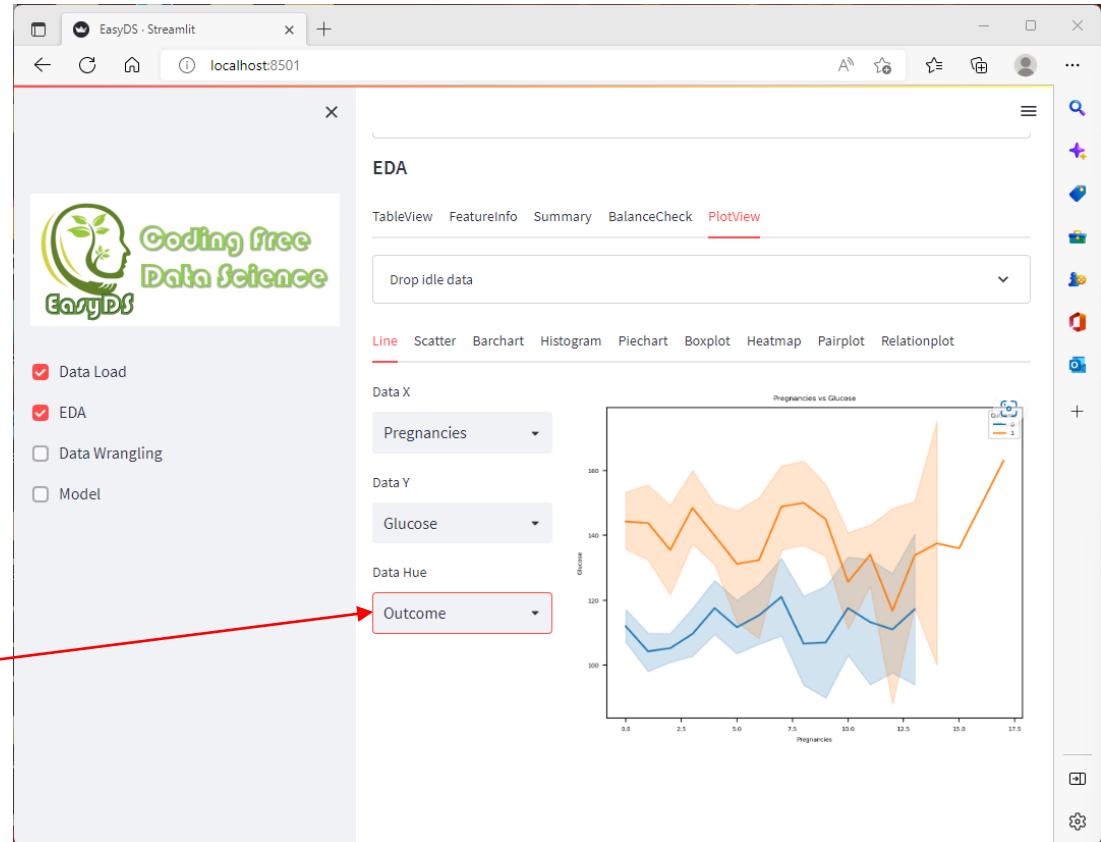
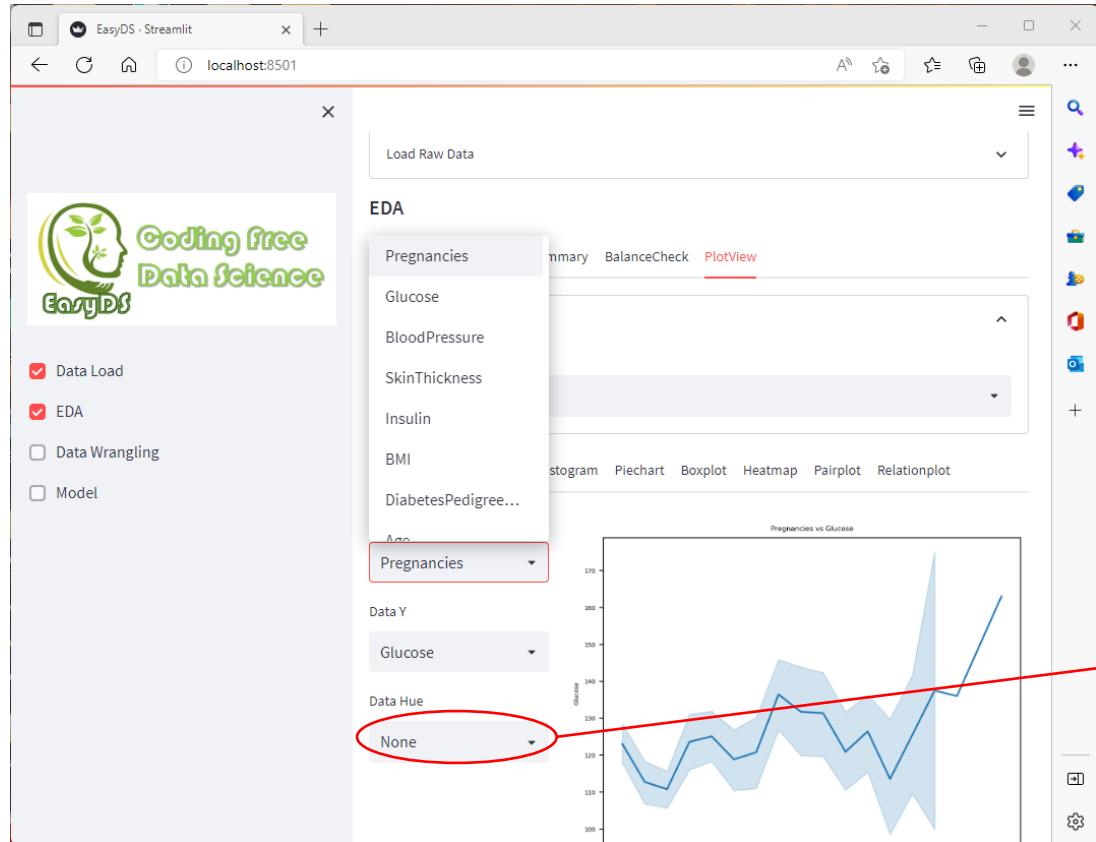
Data balance check display

Plot View – Line Plot



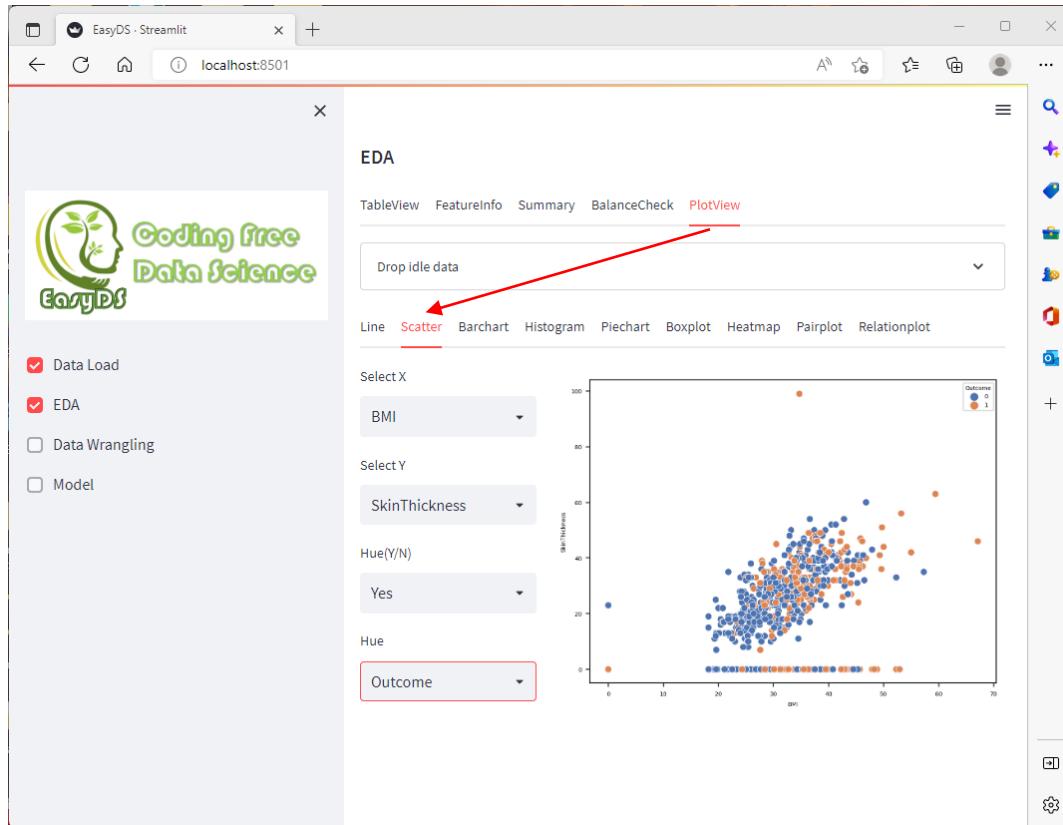
Drop the idle data to processing faster

Plot View – Line Plot

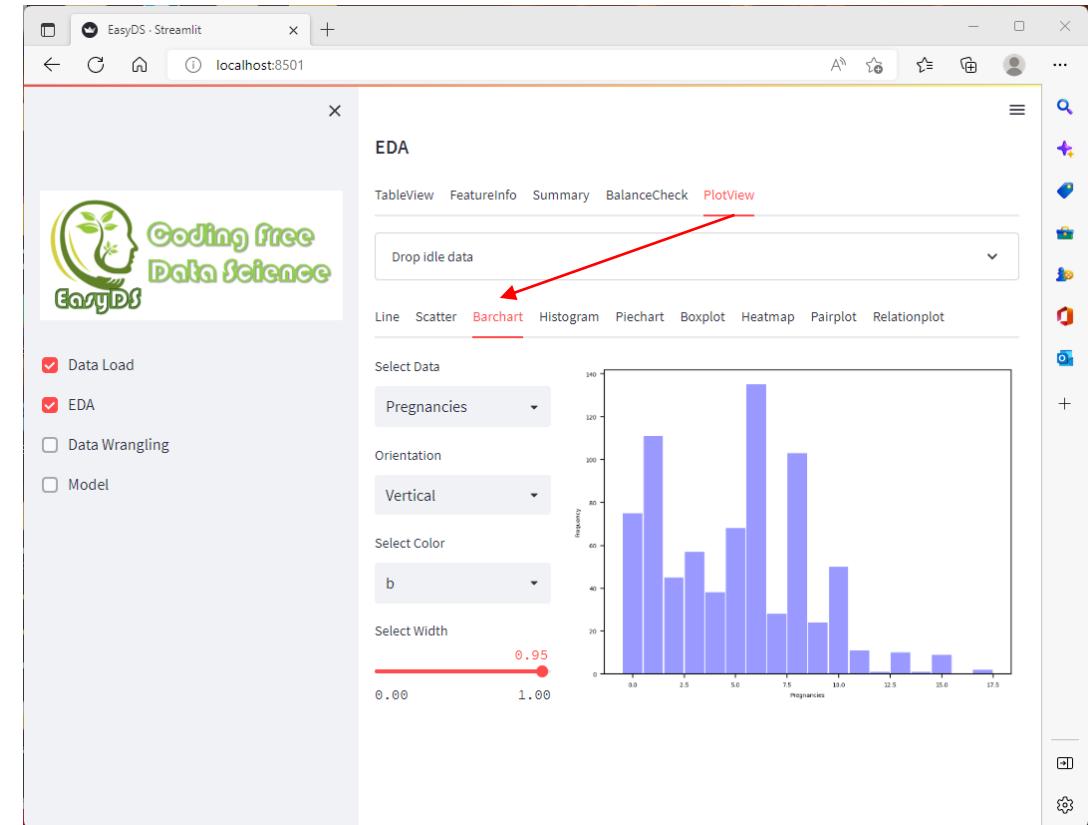


Select plot data and plot styles from drop menus

Plot View – Scatter Plot & Bar Chart

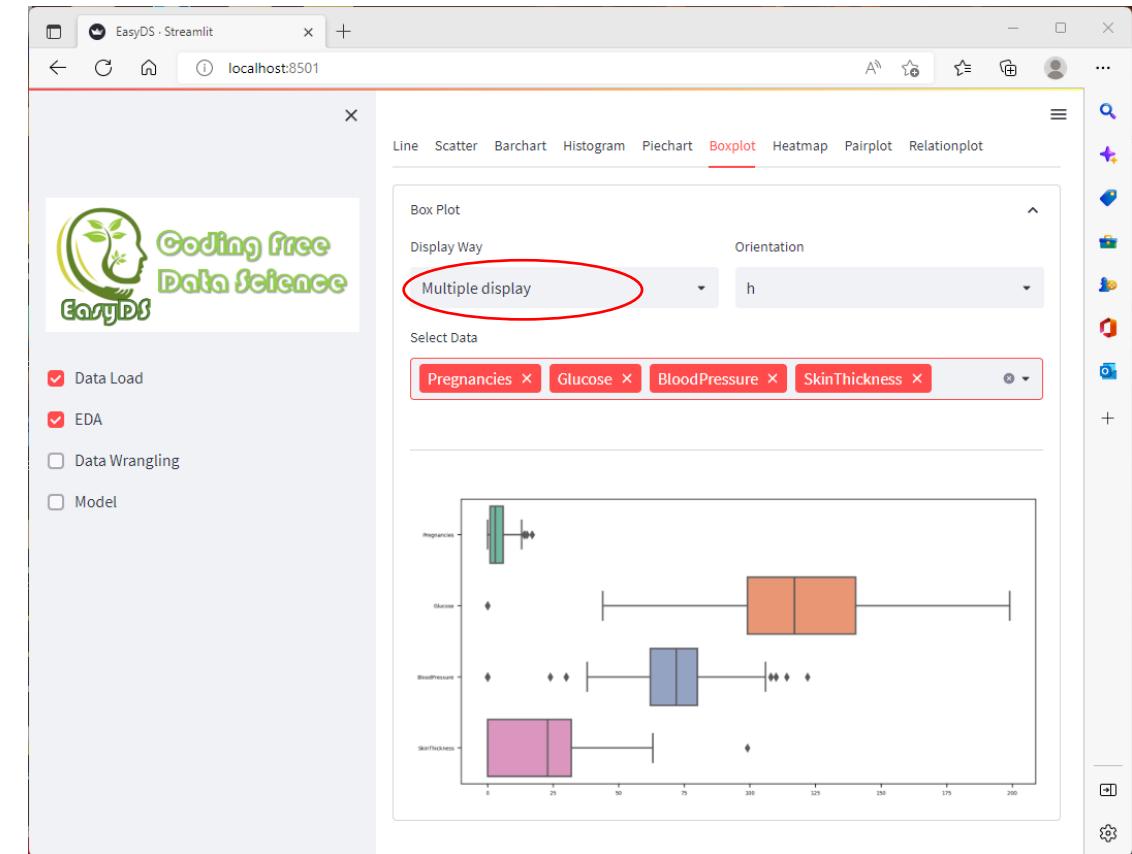
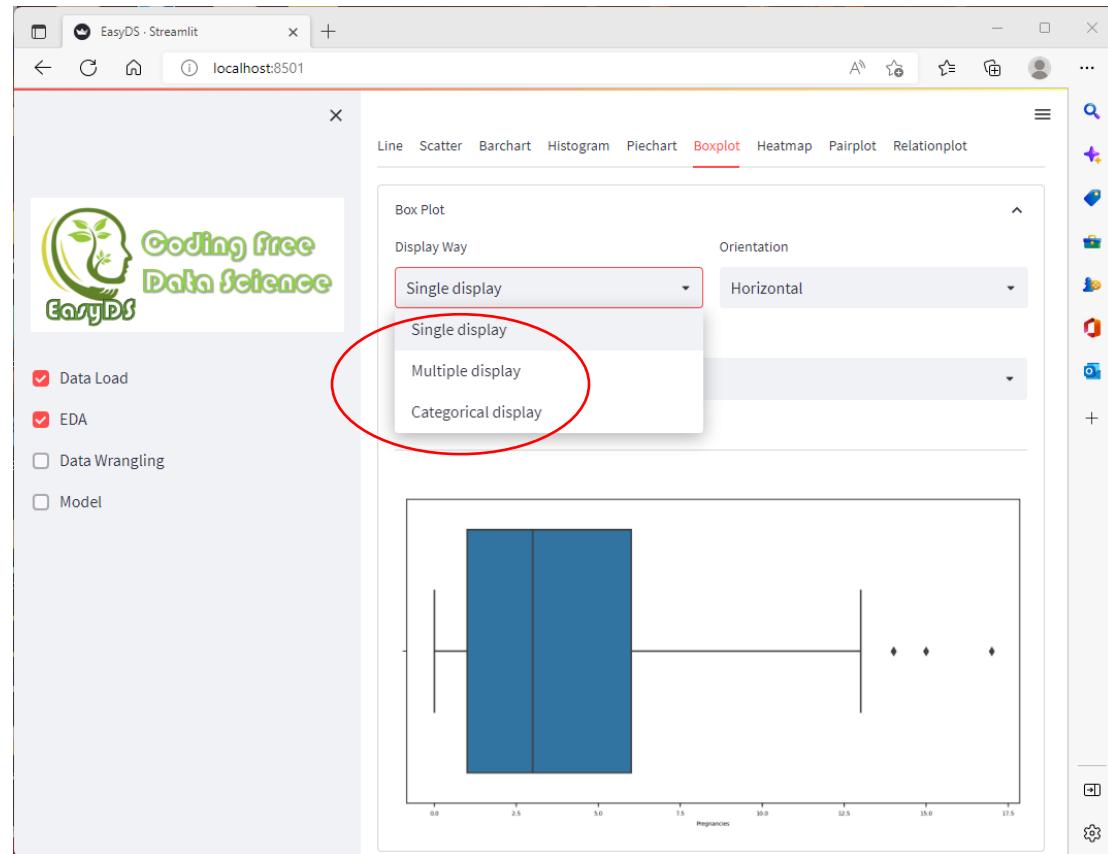


Scatter Plot Example



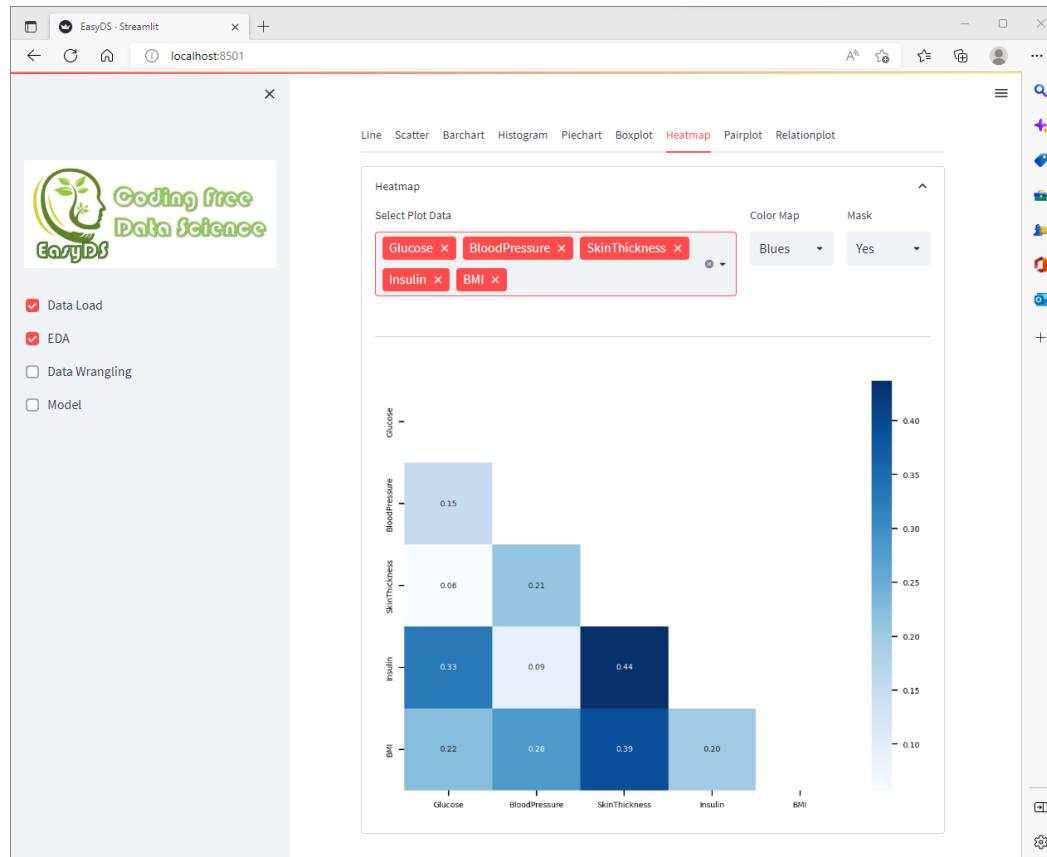
Bar Chart Example

Plot View – Box Plot

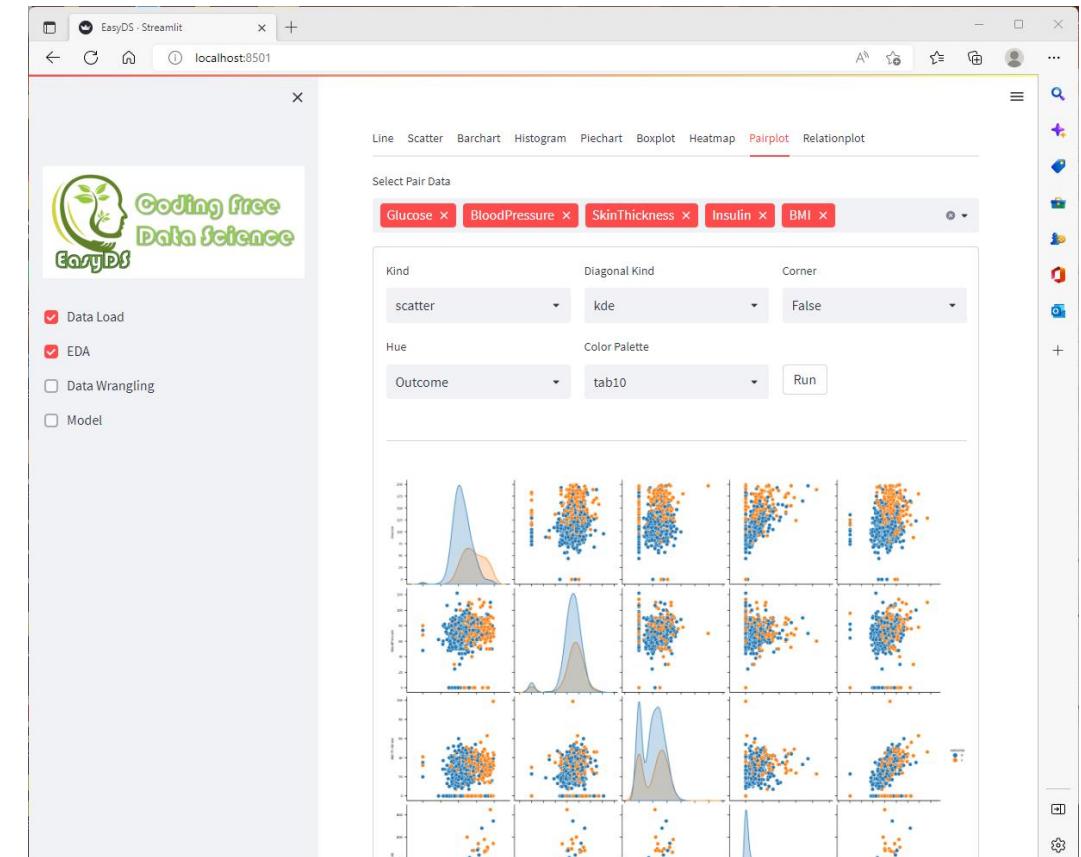


Select plot data and plot styles from drop menus

Plot View – Heat Map & Pair Plot



Heat Map("No" for Mask can display the upper parts)



Pair Plot

Data Wrangling – Filter Columns

The image displays two screenshots of the EasyDS Streamlit application interface, illustrating the process of filtering columns in a dataset.

Left Screenshot: Shows the initial state of the Data Wrangling step. The sidebar on the left has 'Data Wrangling' checked. The main area shows a 'Column Filter' section with 'Select Data' and 'Select method' dropdowns set to 'Drop off' and 'Head'. A preview table shows the first five rows of the dataset. The 'Insulin' and 'BMI' columns are highlighted with a red border. Below the table is a checkbox for 'Define Missing Value Indicator'.

	Preg	Glucose	Bloo	Skin	Insulin	BMI	Diabet	Age	Outc
0	6	148	72	35	0	33.6000	0.6270	50	1
1	1	85	66	29	0	26.6000	0.3510	31	0
2	8	183	64	0	0	23.3000	0.6720	32	1
3	1	89	66	23	94	28.1000	0.1670	21	0
4	0	137	40	35	168	43.1000	2.2880	33	1

Right Screenshot: Shows the result of applying the 'Drop off' method. The 'Select method' dropdown is now set to 'Drop off'. The preview table shows the same five rows, but the 'Insulin' and 'BMI' columns are missing. A red arrow points from the left screenshot to the right one.

	Preg	Glucose	Bloo	Skin	Diabet	Age	Outc
0	6	148	72	35	0.6270	50	1
1	1	85	66	29	0.3510	31	0
2	8	183	64	0	0.6720	32	1
3	1	89	66	23	0.1670	21	0
4	0	137	40	35	2.2880	33	1

Two options for selected data: if “selected”, only selected data keep in the data frame; or, if “Dropoff”, the selected data will be dropped off as this example displayed.

Data Wrangling – Define Null

The image shows two screenshots of the EasyDS Streamlit application interface, illustrating the process of defining missing value indicators for specific columns.

Left Screenshot: The "Data Wrangling" tab is selected. In the "Define Missing Data" section, the "Define Missing Value Indicator" checkbox is checked. A red circle highlights this checkbox, and a red arrow points from it to the "Select data" input field below. The "Select data" field contains "SkinThickness" and "Insulin".

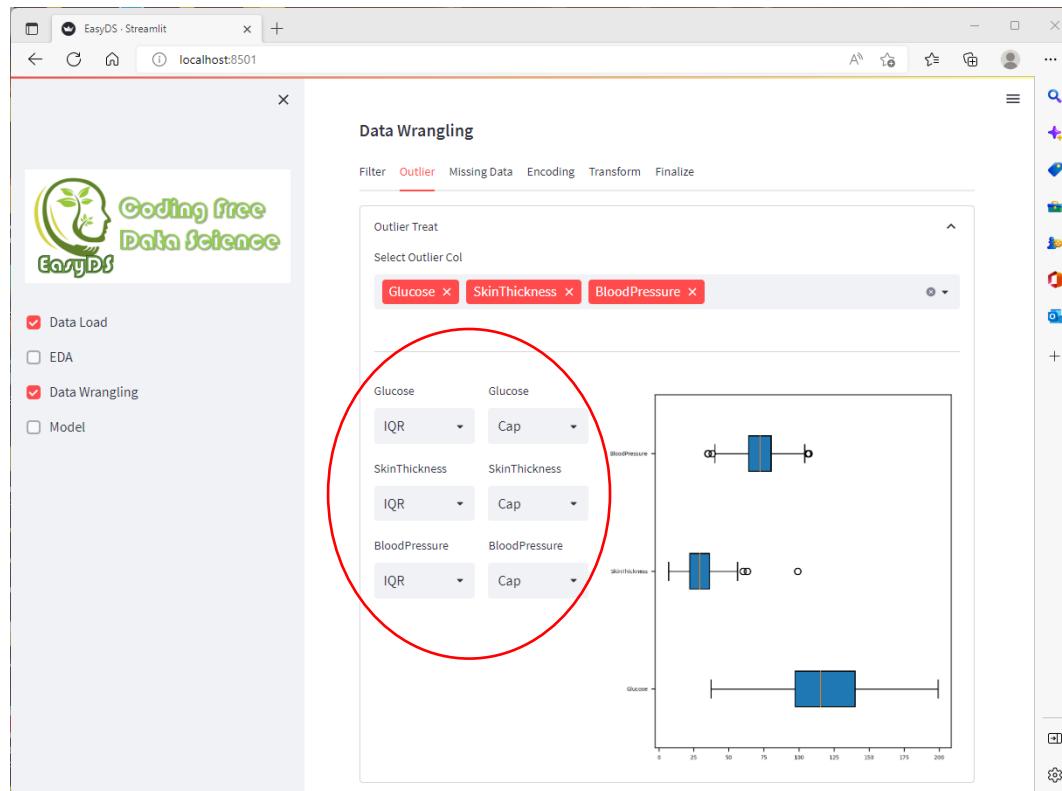
Col Name	Data Type	Counts	Null counts	Null ratio(%)
0 Pregnancies	int64	768	0	0.0
1 Glucose	int64	768	0	0.0
2 BloodPressure	int64	768	0	0.0
3 SkinThickness	int64	768	0	0.0
4 Insulin	int64	768	0	0.0
5 BMI	float64	768	0	0.0
6 DiabetesPedigreeFunction	float64	768	0	0.0
7 Age	int64	768	0	0.0
8 Outcome	int64	768	0	0.0

Right Screenshot: The "Data Wrangling" tab is still selected. The "Select data" field now shows "SkinThickness" and "Insulin" with a red highlight. The "Define Missing Data" table shows updated statistics for these columns.

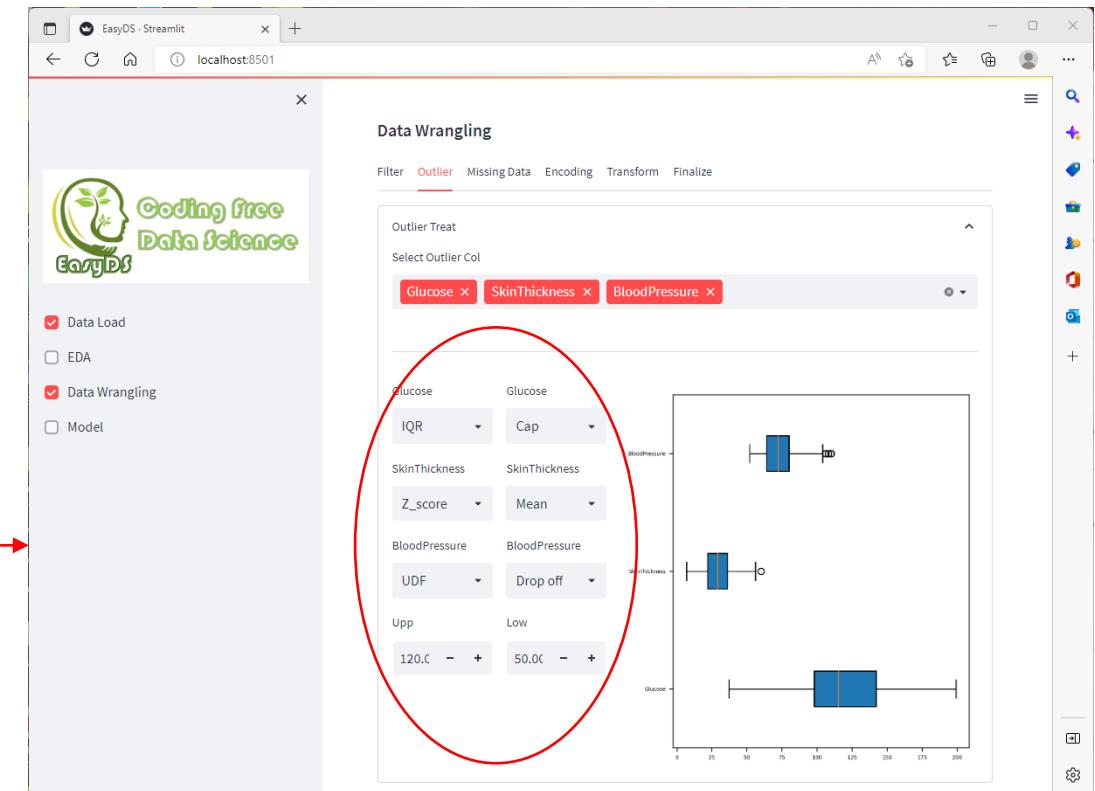
Col Name	Data Type	Counts	Null counts	Null ratio(%)
0 Pregnancies	int64	768	0	0.0
1 Glucose	int64	768	0	0.0
2 BloodPressure	int64	768	0	0.0
3 SkinThickness	float64	768	227	29.6
4 Insulin	float64	768	374	48.7
5 BMI	float64	768	0	0.0
6 DiabetesPedigreeFunction	float64	768	0	0.0
7 Age	int64	768	0	0.0
8 Outcome	int64	768	0	0.0

Sometimes the missed value is saved as '0' or '-999.999' in the dataset. In such case they need to be identified as 'null' for later missing value treatment. In this example '0' were saved for missing values of 'SkinThickness' and 'Insulin'.

Data Wrangling – Treat Outliers



Default outliers treatment is capped by IQR method



Different outliers treatment methods and strategies can be selected, 'UDF' mean user defined function.

Data Wrangling – Treat Missing Values

The screenshot shows the EasyDS Streamlit application interface. On the left, there's a sidebar with a logo and several menu items: Data Load (checked), EDA (unchecked), Data Wrangling (checked), and Model (unchecked). The main area has tabs for Data Load, EDA, Data Wrangling, and Model. The Data Wrangling tab is active, with sub-tabs for Filter, Outlier, Missing Data (which is checked and highlighted in red), Encoding, Transform, and Finalize. Below these tabs is a 'Feature Information' table:

	Col Name	Counts	Null counts
0	SkinThickness	394	0
1	Insulin	394	0

As previously mentioned, there are missing values for 'SkinThickness' and 'Insulin' in this example. The missing data can be dropped off if it is less.

This screenshot shows the same EasyDS Streamlit application after a change in the 'Treat Missing Data' settings. The 'Missing Data' tab is still selected. The 'Feature Information' table now shows different counts for the 'SkinThickness' and 'Insulin' columns:

	Col Name	Counts	Null counts
0	SkinThickness	768	0
1	Insulin	768	0

Or other strategies can be applied, such as 'mean' or 'median' as displayed in this example.

Data Wrangling – Encoding

The screenshot displays two instances of the EasyDS Streamlit application interface, illustrating the process of data encoding. Both instances have a header bar with tabs: Filter, Outlier, Missing Data, Encoding (which is highlighted in red), Transform, and Finalize. The sidebar on the left includes a logo for "Coding free Data Science EasyDS" and checkboxes for Data Load (checked), EDA (unchecked), Data Wrangling (checked), and Model (unchecked).

Left Instance (Raw Data): The "Data Encoding" section shows three dropdown menus: "Item_Fat_Content" (set to "Raw"), "Item_Type" (set to "Raw"), and "Outlet_Size" (set to "Raw"). Below these dropdowns is a table with three columns: "Item_Fat_Content", "Item_Type", and "Outlet_Size". The table contains 13 rows of data:

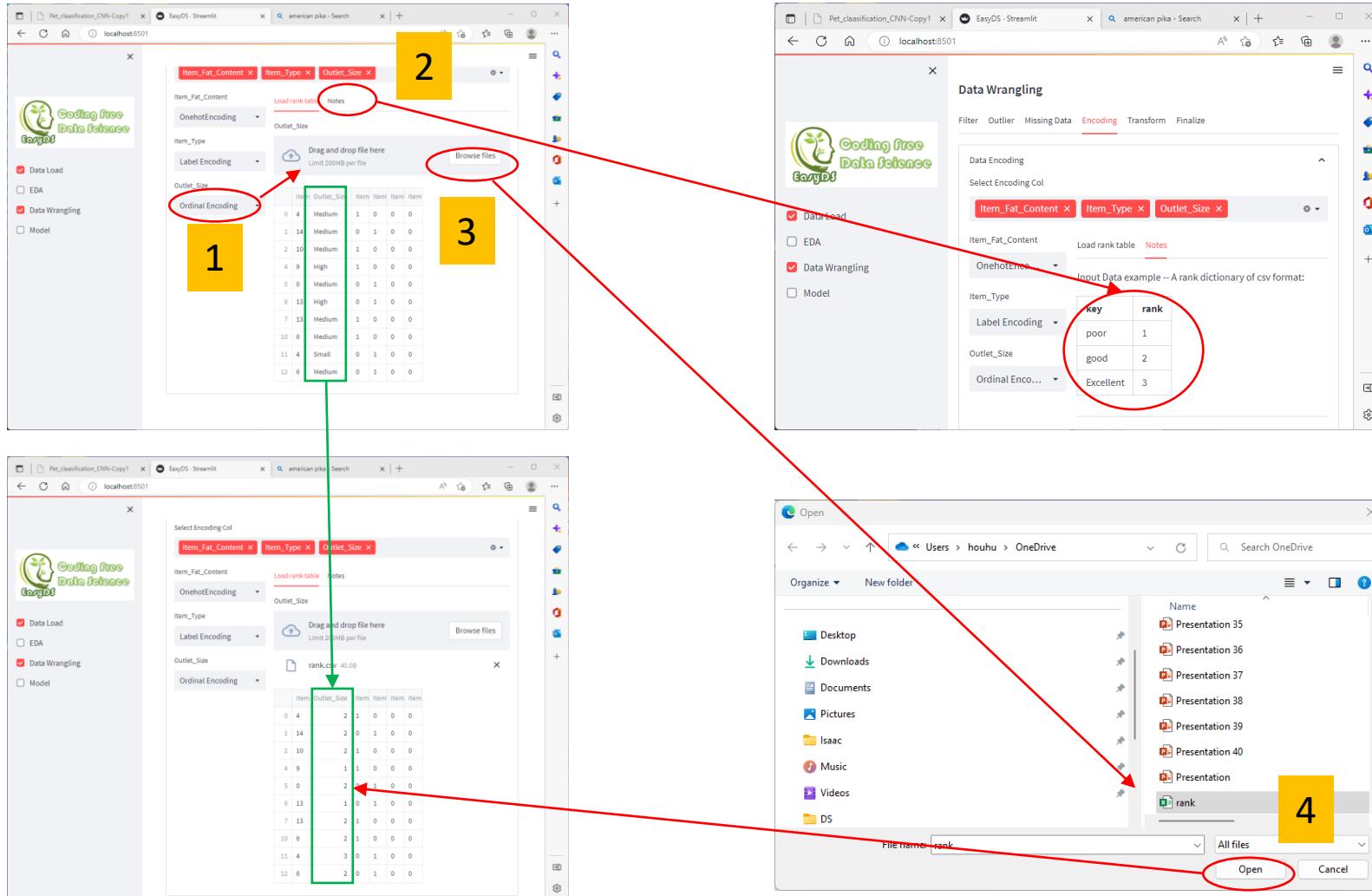
	Item_Fat_Content	Item_Type	Outlet_Size
0	Low Fat	Dairy	Medium
1	Regular	Soft Drinks	Medium
2	Low Fat	Meat	Medium
3	Low Fat	Household	High
4	Regular	Baking Goods	Medium
5	Regular	Snack Foods	High
6	Low Fat	Snack Foods	Medium
7	Low Fat	Fruits and Vegetables	Medium
8	Regular	Dairy	Small
9	Regular	Fruits and Vegetables	Medium

Right Instance (Encoded Data): The "Data Encoding" section shows three dropdown menus: "Item_Fat_Content" (set to "OnehotEncoding"), "Item_Type" (set to "Label Encoding"), and "Outlet_Size" (set to "Raw"). Below these dropdowns is a table with four columns: "Item_Fat_Content", "Item_Type", "Outlet_Size", and "Item_Fat_Content". The table contains 13 rows of data, showing the transformed numerical values for each category:

	Item_Fat_Content	Item_Type	Outlet_Size	Item_Fat_Content
0	4	14	Medium	0 0 0 0
1	14	10	Medium	0 1 0 0
2	10	9	High	1 0 0 0
3	9	0	Medium	0 1 0 0
4	0	13	High	1 0 0 0
5	13	6	Medium	0 1 0 0
6	6	13	Medium	1 0 0 0
7	13	6	Medium	0 1 0 0
8	6	4	Small	0 0 1 0
9	4	6	Medium	0 1 0 0

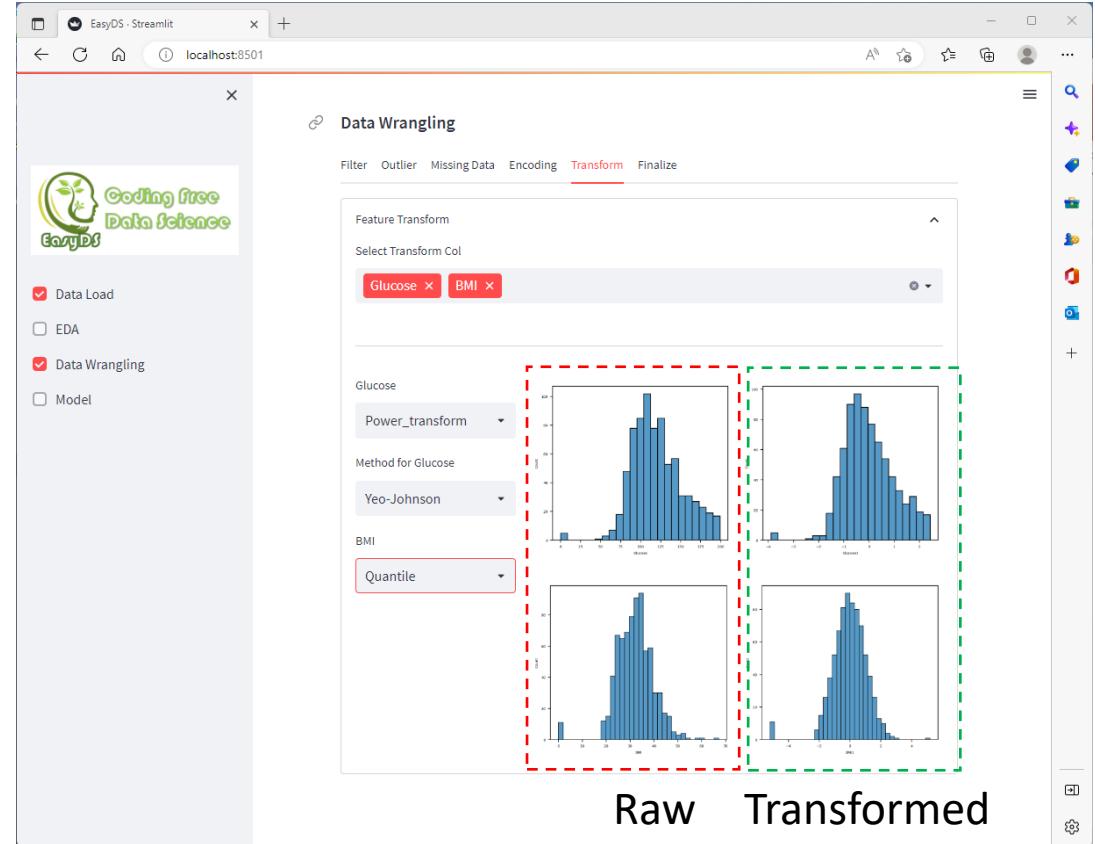
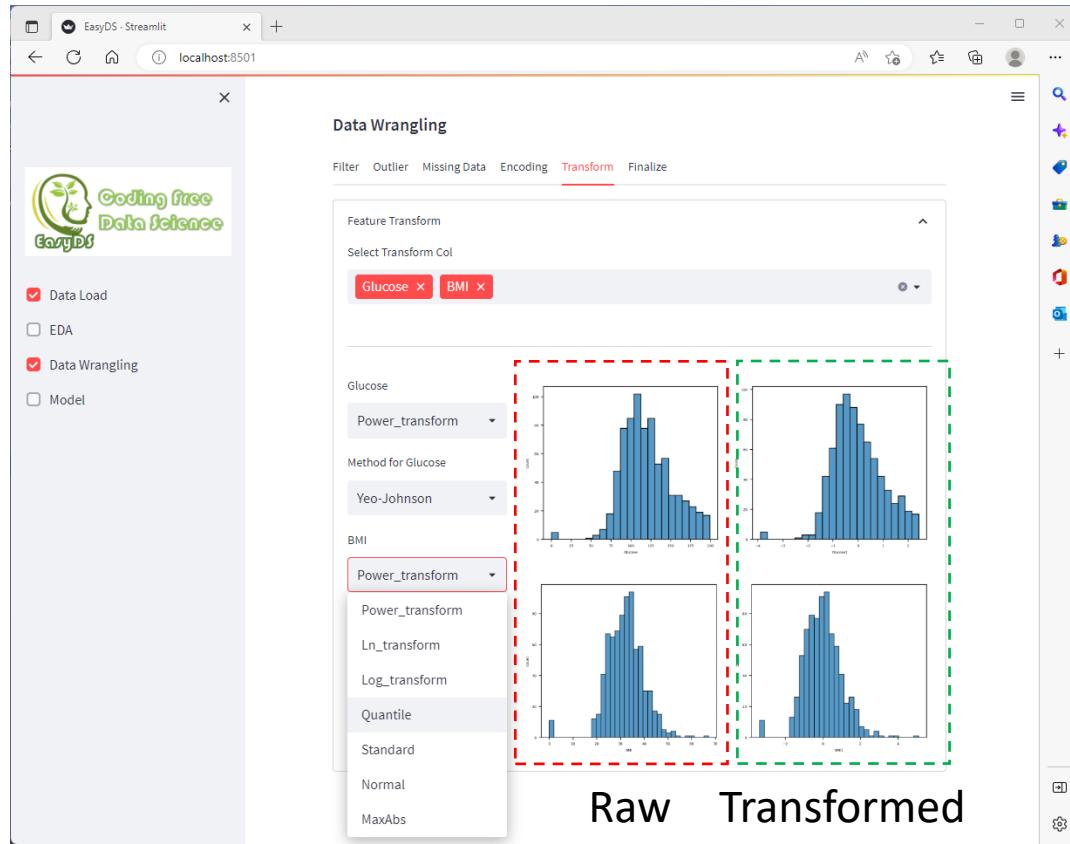
Encoding is transform categorical data into numerical data so that they can be used for modeling. Selecting all the columns to be encoded, the default raw values are displayed(left side plot) . Different methods can be selected to encode the data. This example showed the Label encoding and One Hot Encoding results.

Data Wrangling – Ordinal Encoding



1. Sometimes it needs Ordinal Encoding. An ordinal index file in CSV need be firstly defined and upload.
2. To check the ordinal index file format for reference, please click ‘Notes’.
3. Click browse files to search the file.
4. Open select ordinal index file to encode the data.

Data Wrangling – Transform



To get better model, sometimes it needs transform data to approximate normal distribution. Default method is 'Power Transform' as displayed is left side plot, other methods can be selected such as 'Quantile' in right side cases.

Data Wrangling – Clean & Finalize Data

The screenshot shows the 'Data Wrangling' section of the EasyDS Streamlit application. On the left, there's a sidebar with checkboxes for 'Data Load' (checked), 'EDA' (unchecked), 'Data Wrangling' (checked), and 'Model' (unchecked). The main area has tabs for 'Filter', 'Outlier', 'Missing Data', 'Encoding', 'Transform', and 'Finalize' (which is currently selected). Below these tabs is a table titled 'Cleaning and Save Finalized Data'. The table has columns: Preg, Glucose, BloodPressure, SkinThickness, Insulin, BMI, Diabetic, Age, Outc, and Glucose1. The first and last columns ('Glucose' and 'Glucose1') are highlighted with red boxes. A message at the bottom of the table area says 'Selected [] have been dropped off from the data frame!'. At the very bottom, there's a button to 'Download data named as 'df_clean.csv''.

This screenshot shows the same 'Data Wrangling' interface after some columns have been removed. The 'Glucose1' and 'BMI1' columns are now highlighted with red boxes. The message at the bottom of the table area has changed to 'Selected ['Glucose1', 'BMI1'] have been dropped off from the data frame!'. The 'Download data named as 'df_clean.csv'' button is circled in red.

After data wrangling, a clean data could be saved for modeling. Some repeated data need to be dropped off, for example in this case there are raw and scaled ‘Glucose’ and ‘BMI’ in data set. To save the data to local disk, click the download data button, it will save the data to local download folder and name it as ‘df_clean.csv’.

Data Preparing for Modeling

The screenshot shows the 'Data Prepare' section of the EasyDS Streamlit application. On the left sidebar, under the 'Model' section, the 'Model' checkbox is checked, while 'Clustering', 'Classification', 'Regression', and 'Existing Model' are unchecked. In the main area, the 'Data Define' tab is selected. A dropdown menu titled 'Define_model_feature_target' shows 'Outcome' as the target feature. Below it, a table displays the first five rows of the dataset:

	Preg	Glucose	Bloo	Skin	Insulin	BMI	Diabete	Age	Outc
0	6	148	72	35	0	33.6000	0.6270	50	1
1	1	85	66	29	0	26.6000	0.3510	31	0
2	8	183	64	0	0	23.3000	0.6720	32	1
3	1	89	66	23	94	28.1000	0.1670	21	0
4	0	137	40	35	168	43.1000	2.2880	33	1

A green box at the bottom states: 'Target is "Outcome", and other kept data as features with Data Shape ~ (768, 8)!'.

Define target data and features

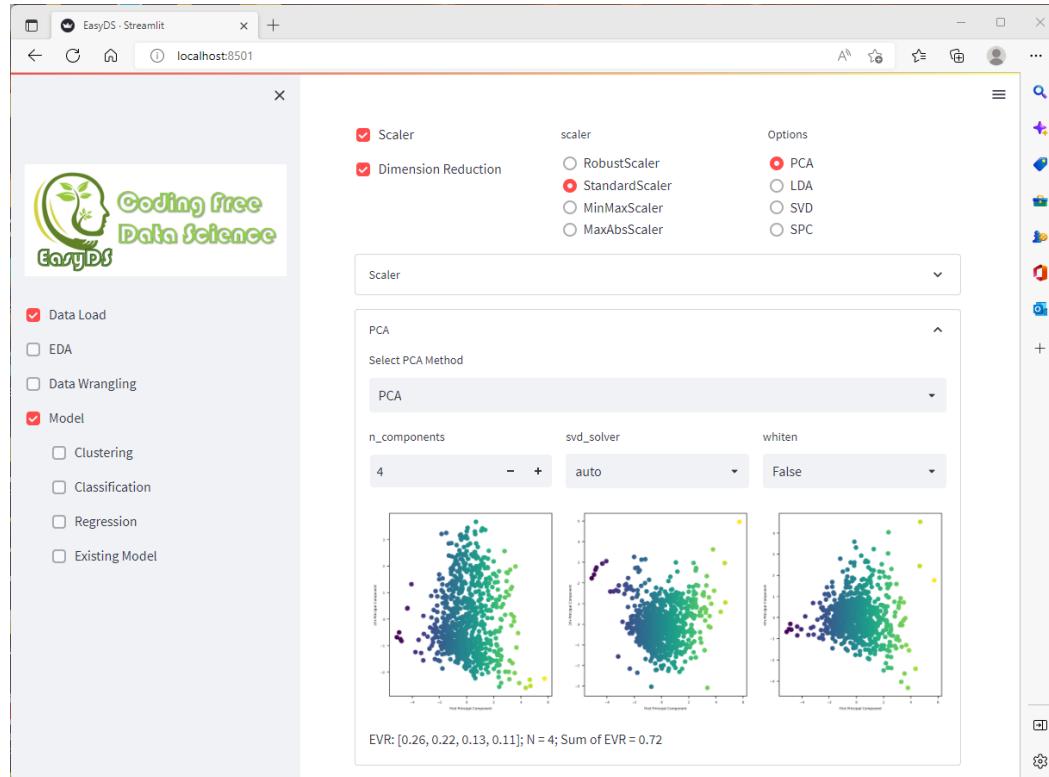
The screenshot shows the 'Scaler/Dimension Reduction' section of the EasyDS Streamlit application. On the left sidebar, under the 'Model' section, the 'Model' checkbox is checked, while 'Clustering', 'Classification', 'Regression', and 'Existing Model' are unchecked. In the main area, the 'Scaler/Dimension Reduction' tab is selected. It shows two checked options: 'Scaler' and 'Dimension Reduction'. Under 'Scaler', there are several radio button options: RobustScaler (unchecked), StandardScaler (checked), MinMaxScaler (unchecked), SVD (unchecked), and MaxAbsScaler (unchecked). Below the scalers, a table titled 'Scaler' provides statistical summary data for each feature:

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0000	0.0000	1.0007	-1.1419	-0.8449	-0.2510	0.6399	3.9066
Glucose	768.0000	0.0000	1.0007	-3.7837	-0.6852	-0.1219	0.6058	2.4445
BloodPressure	768.0000	0.0000	1.0007	-3.5726	-0.3673	0.1496	0.5632	2.7345
SkinThickness	768.0000	0.0000	1.0007	-1.2882	-1.2882	0.1545	0.7191	4.9219
Insulin	768.0000	0.0000	1.0007	-0.6929	-0.6929	-0.4281	0.4120	6.6528
BMI	768.0000	0.0000	1.0007	-4.0605	-0.5956	0.0009	0.5848	4.4558
DiabetesPedigree	768.0000	0.0000	1.0007	-1.1896	-0.6890	-0.3001	0.4662	5.8836
Age	768.0000	0.0000	1.0007	-1.0415	-0.7863	-0.3608	0.6602	4.0637

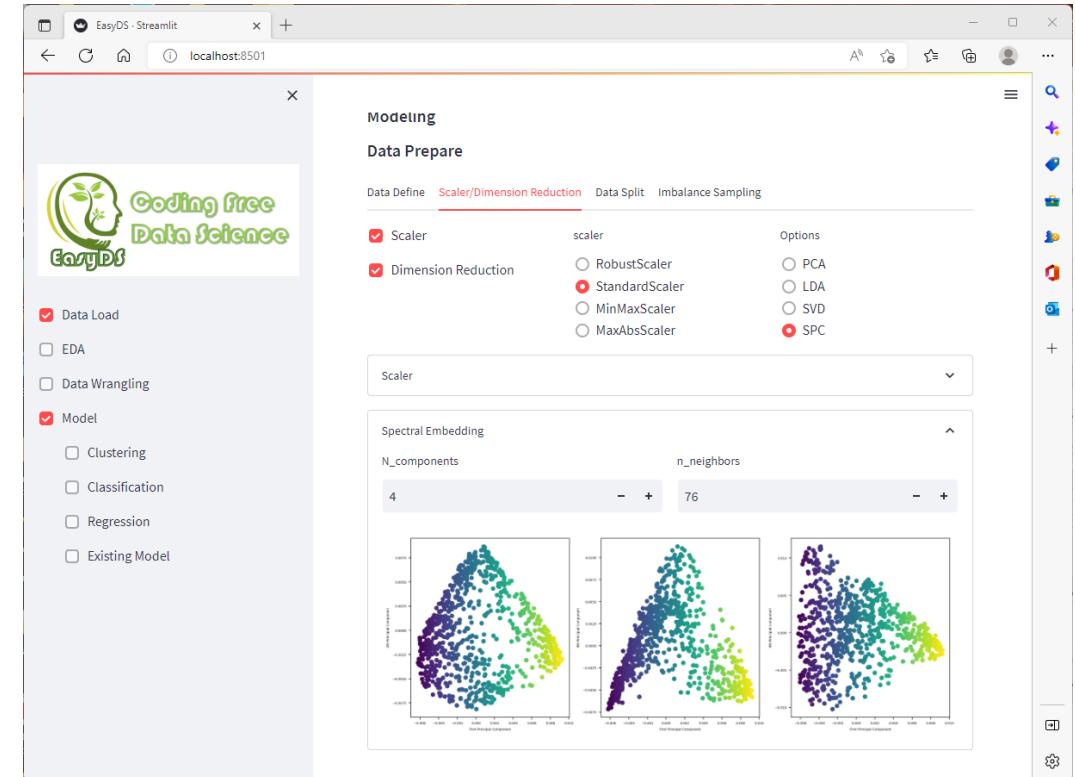
A green box at the bottom states: 'PCA'.

Scaler or reduce dimensions

Data Preparing for Modeling - Reduce dimensions



Example of PCA



Example of SPC

Data preparing for modeling

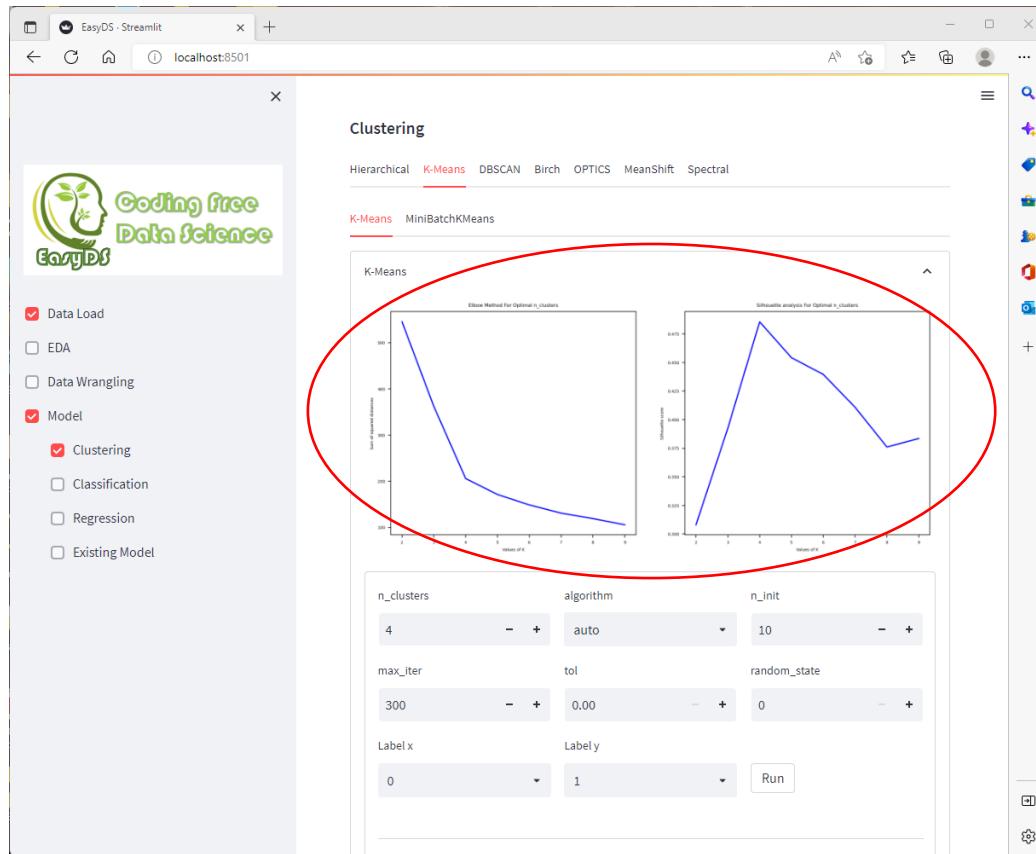
The screenshot shows the EasyDS Streamlit application interface. On the left, a sidebar menu is open with the following options: Data Load (checked), EDA, Data Wrangling, Model (checked), Clustering, Classification, Regression, and Existing Model. The main area is titled "Data Prepare" and contains tabs for Data Define, Scaler/Dimension Reduction, Data Split (which is currently selected), and Imbalance Sampling. Under "Data Split", there are fields for "test_size" (set to 0.20), "random_state" (set to 0), and "Shuffle" (set to True). A green message box at the bottom states: "Post data split: X_train of shape (614, 4) ; X_test of shape (154, 4)".

Data Split

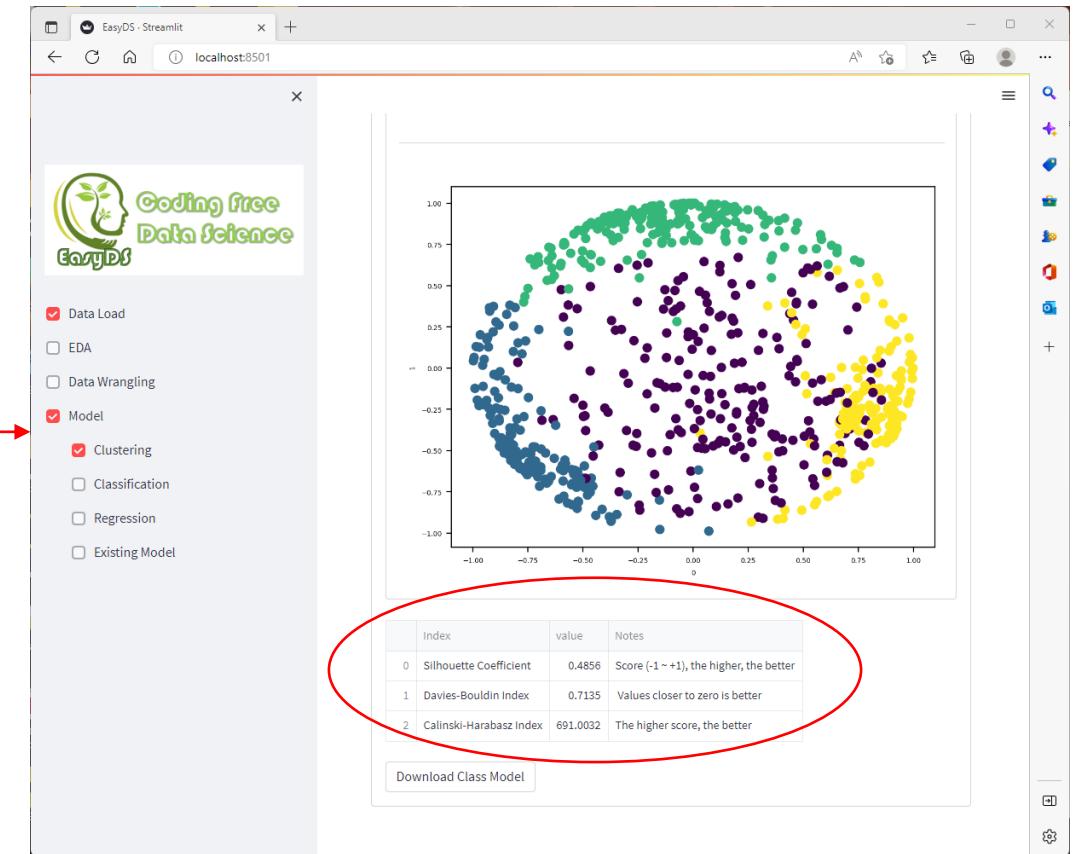
The screenshot shows the same EasyDS Streamlit application interface as the first one, but with different configurations. The sidebar menu is identical. In the main "Data Prepare" area, the "Imbalance Sampling" tab is selected. It includes a section titled "Imbalance Sampling(SMOTE + TomekLinks)" with three tables labeled "Origin Train Data", "Step1: SMOTE", and "Step2: TomekLinks". The "Origin Train Data" table has two rows: Outcome 0 (393) and Outcome 1 (221). The "Step1: SMOTE" table has two rows: Outcome 1 (393) and Outcome 0 (393). The "Step2: TomekLinks" table has two rows: Outcome 1 (393) and Outcome 0 (354). A green message box at the bottom states: "Post imbalance Sampling: X_train of shape (747, 4) ; y_train of shape (747,)".

Optional Imbalance sampling(SMOTE + TomekLinks)

Clustering

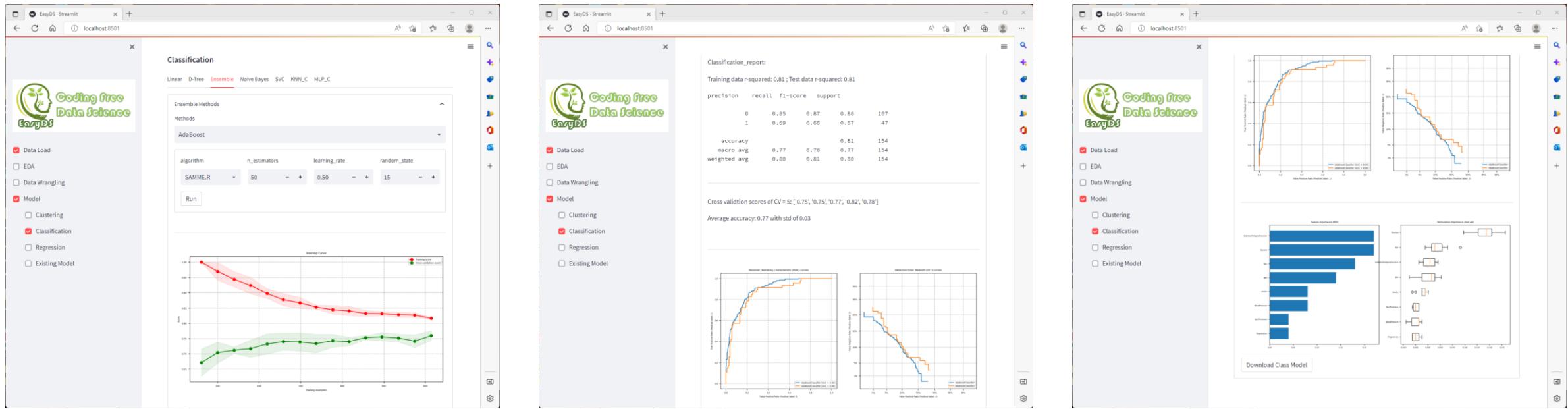


Use different methods as aid for to define n_clusters in K-Means



Use different index to evaluate model

Classification



Wrapped most popular Classification methods, and multiple ways for parameter tuning, model cross-checking and validation

Classification – example of XGBoost

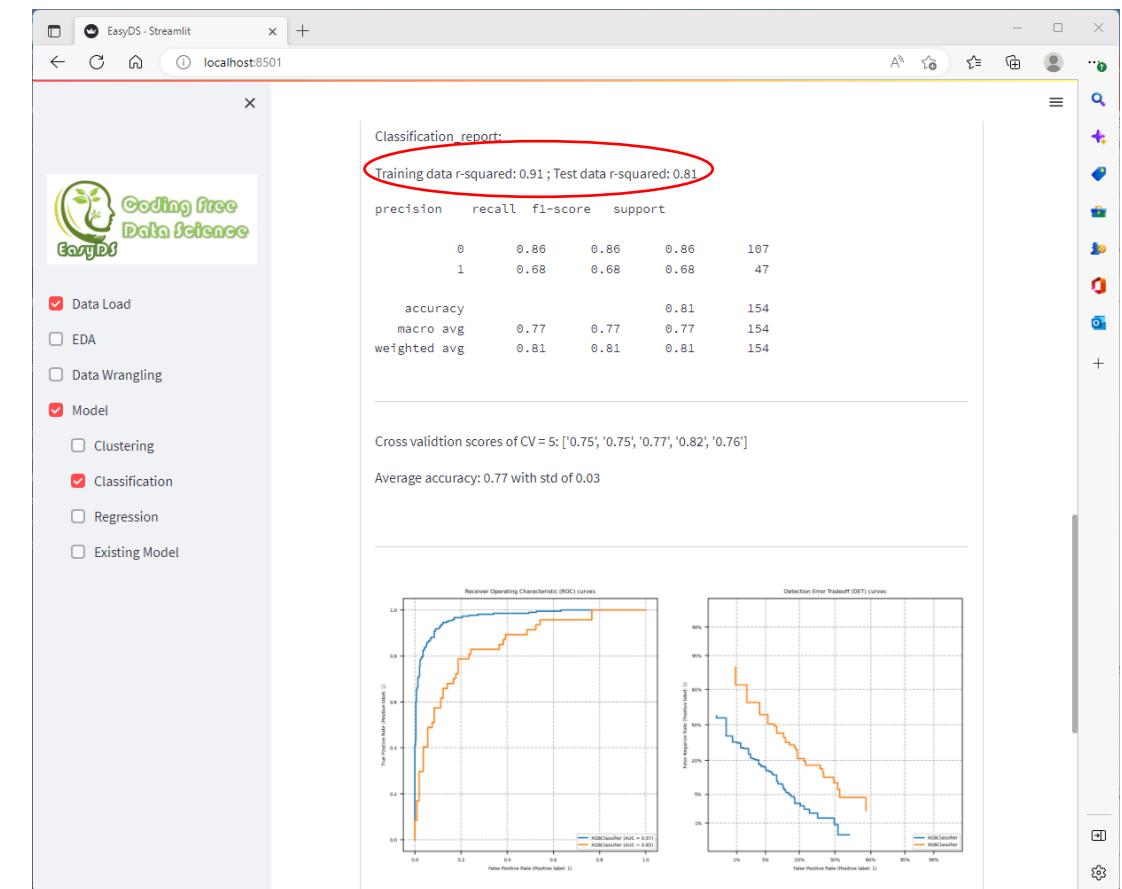
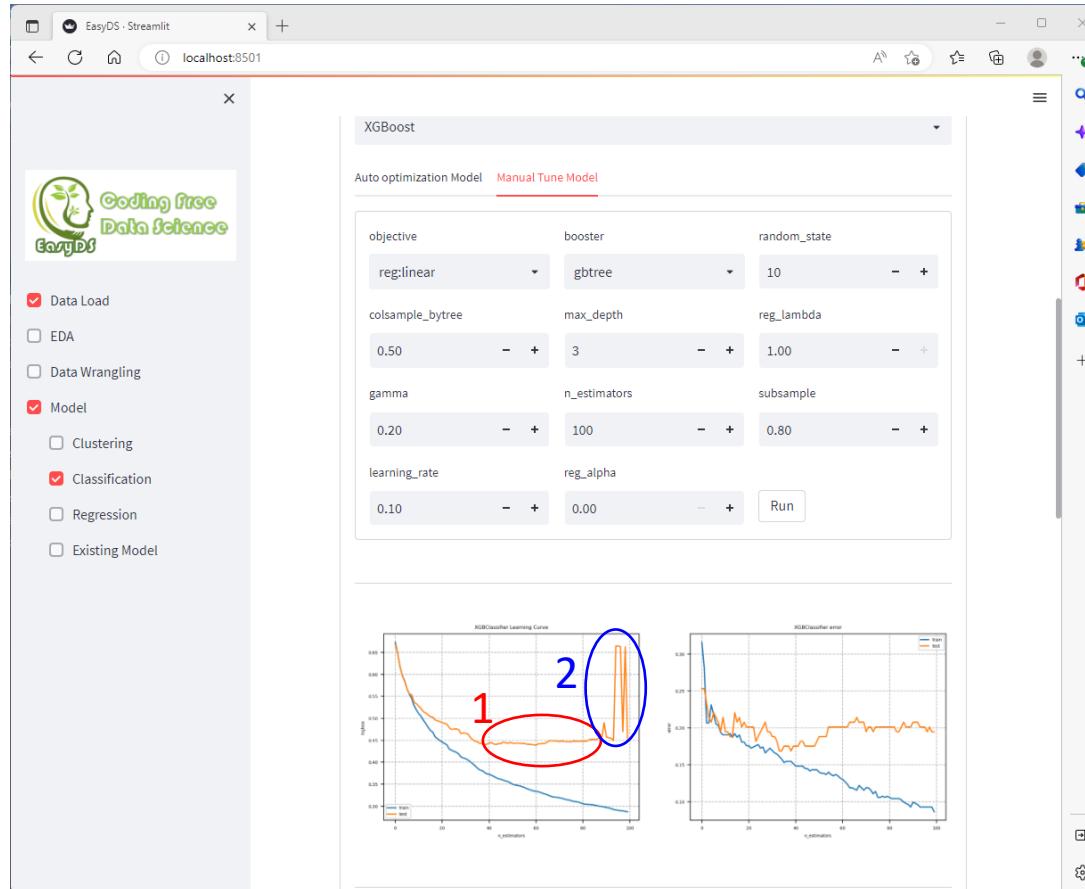
This screenshot shows the 'Modeling' section of the EasyDS Streamlit application. On the left sidebar, under the 'Model' category, 'Classification' is selected. In the main area, 'Ensemble' is chosen from the 'Classification' dropdown. Under 'Ensemble Methods', 'XGBoost' is selected. The 'Auto optimization Model' tab is active, showing parameters: n_iter (50), cv (5), scoring (accuracy), and random_state (10). A 'Run' button is present at the bottom.

This screenshot shows the 'Classification' section of the EasyDS Streamlit application. The 'Manual Tune Model' tab is active for the XGBoost method. It displays detailed tuning parameters:

Parameter	Value	Min	Max							
objective	reg:linear	reg:squarederror	gbtree	gblinear	10	-	+			
colsample_bytree	0.50	-	+	max_depth	3	-	+	1.00	-	+
gamma	0.20	-	+	n_estimators	100	-	+	0.80	-	+
learning_rate	0.10	-	+	subsample	0.00	-	+	reg_alpha	Run	

Some of the methods have both Auto Optimization Model and Manual Tuning Model. Since Auto Optimization Model is still in testing (random search on discrete parameters to save compute times), the result is not really optimized. Therefore, Manual Tuning Model is suggested for the modeling.

Classification – example of XGBoost



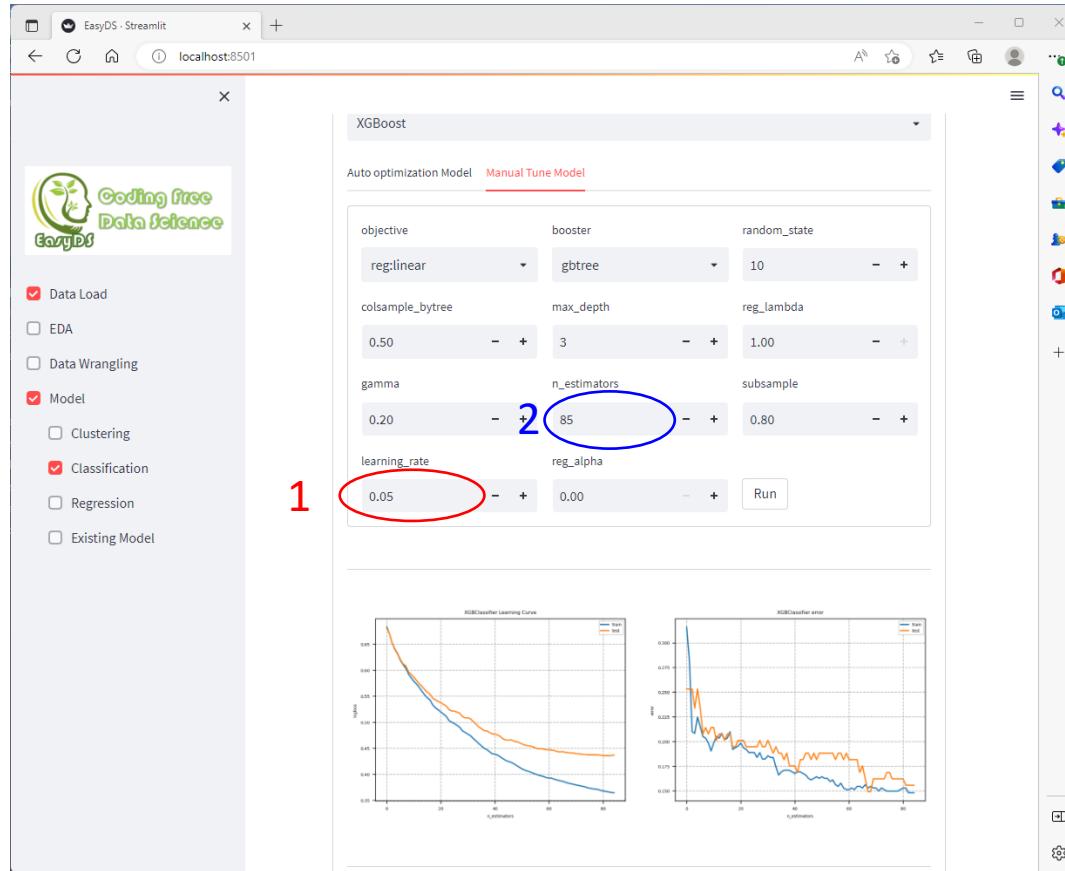
Learning Curves assist for hypo-parameter tuning, it can be seen :

1. It learns too fast as the learning curve quickly get flat
2. Too much learn leading to instable results



Model is over fitted

Classification – example of XGBoost



EasyDS - Streamlit | localhost:8501

XGBoost

Auto optimization Model **Manual Tune Model**

objective booster random_state

reg:linear gbtree 10

colsample_bytree max_depth reg_lambda

0.50 3 1.00

gamma n_estimators subsample

0.20 2 85 0.80

learning_rate reg_alpha

0.05 0.00

Run

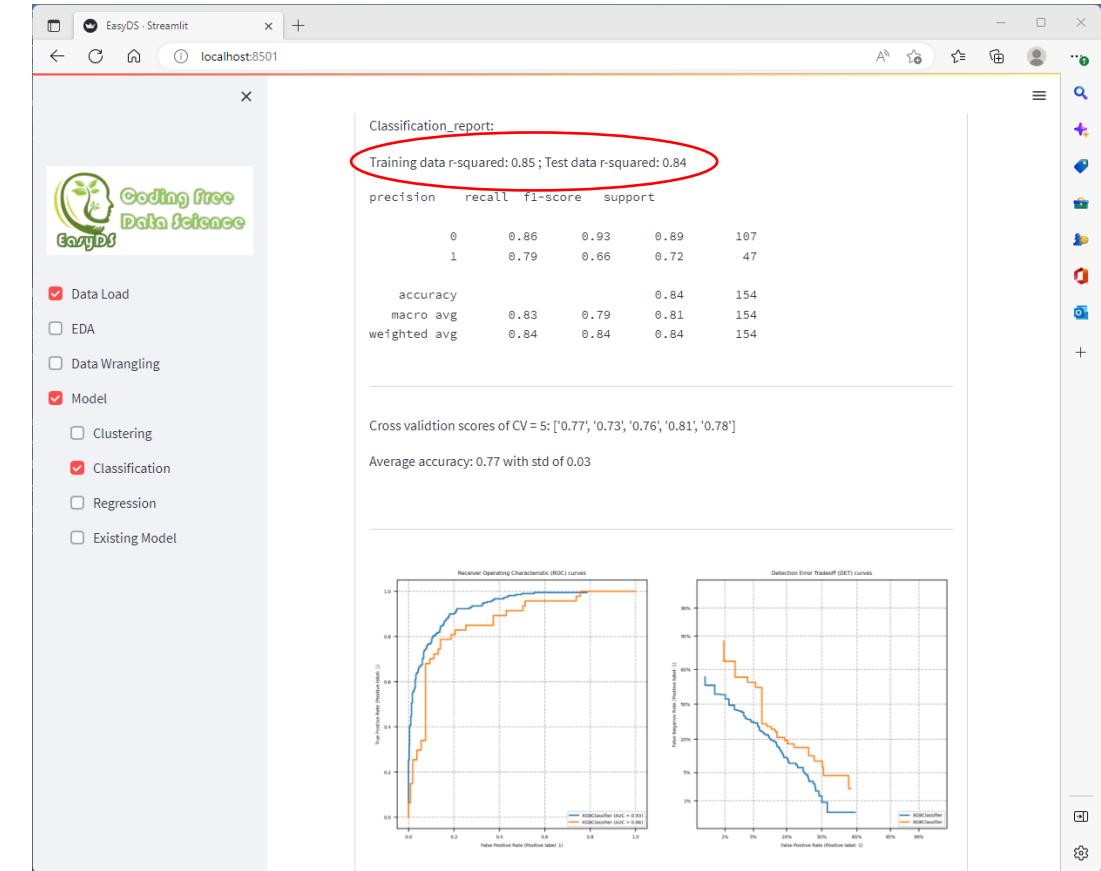
1

Classification free Data Science

Data Load EDA Data Wrangling Model Clustering Classification Regression Existing Model

XGBClassifier Learning Curve XGBClassifier error

The screenshot shows the XGBoost configuration page. The 'Manual Tune Model' tab is selected. The 'n_estimators' parameter is set to 85, and the 'learning_rate' parameter is set to 0.05. Two red circles highlight these parameters. Below the configuration are two line plots: 'XGBClassifier Learning Curve' and 'XGBClassifier error'. The 'XGBClassifier Learning Curve' plot shows training (blue) and testing (orange) accuracy over 85 iterations, with accuracy increasing from approximately 0.45 to 0.85. The 'XGBClassifier error' plot shows training (blue) and testing (orange) error over 85 iterations, with error decreasing from approximately 0.25 to 0.05.



EasyDS - Streamlit | localhost:8501

Classification_report:

Training data r-squared: 0.85 ; Test data r-squared: 0.84

	precision	recall	f1-score	support
0	0.86	0.93	0.89	187
1	0.79	0.66	0.72	47
accuracy				0.84
macro avg	0.83	0.79	0.81	154
weighted avg	0.84	0.84	0.84	154

Cross validation scores of CV = 5: [0.77, 0.73, 0.76, 0.81, 0.78]

Average accuracy: 0.77 with std of 0.03

ROC curves DET curves

The screenshot shows the classification report page. It displays the 'Classification_report:' section with 'Training data r-squared: 0.85 ; Test data r-squared: 0.84'. Below this is a table of precision, recall, f1-score, and support for classes 0 and 1. Further down are 'Cross validation scores of CV = 5' and 'Average accuracy: 0.77 with std of 0.03'. At the bottom are two plots: 'ROC curves' and 'DET curves', both comparing XGBClassifier (blue) and KNNClassifier (orange).

To improve the model referred to previous learning curve:

1. Decrease learning rate to 0.05
2. Decrease n_estimators to 85

Model gets reasonable.

Regression

EasyDS - Streamlit | localhost:8501

Regression

Linear D-Tree Ensemble SVR KNN_R MLP_R

Linear Models

Method: LinearR

Download Class Model

Trained r2 score: 0.429

-- Test Score Matrix --

	score	value
0	mae	3.3666
1	mse	18.2829
2	rmse	4.2758
3	r2	0.4254

EasyDS - Streamlit | localhost:8501

MLPRegressor

Auto optimization Model Manual Tune Model

Solver: adam Beta_1: 0.90 Beta_2: 0.99

hidden_layer_sizes: 100 activation: relu random_state: 20

alpha: 1.00 learning_rate_init: 0.00

Run

Data Load EDA Data Wrangling Model Clustering Classification Regression Existing Model

Trained r2 score: 0.310

-- Test Score Matrix --

	score	value
0	mae	4.9653
1	mse	38.3689
2	rmse	6.1943
3	r2	0.1648

Wrapped most popular Regression methods, and multiple scores are listed for model checking and validation

Load Existed Model

The image shows two screenshots of the EasyDS Streamlit application interface, illustrating the process of loading an existed model.

Left Screenshot: The user has loaded a model named "Clustering". The "Model Type" dropdown menu is open, showing "Clustering" as the selected option. A red oval highlights this selection. Below the dropdown, there is a "Upload Model" section with a "Drag and drop file here" button and a "Browse files" button. A note at the bottom says, "If Loaded Model matched Model Type, check the box for RUN !".

Right Screenshot: The user has changed the "Model Type" from "Clustering" to "Classification". A red arrow points from the "Clustering" option in the left screenshot to the "Classification" option in the right screenshot, indicating the transition. The rest of the interface remains consistent with the first screenshot.

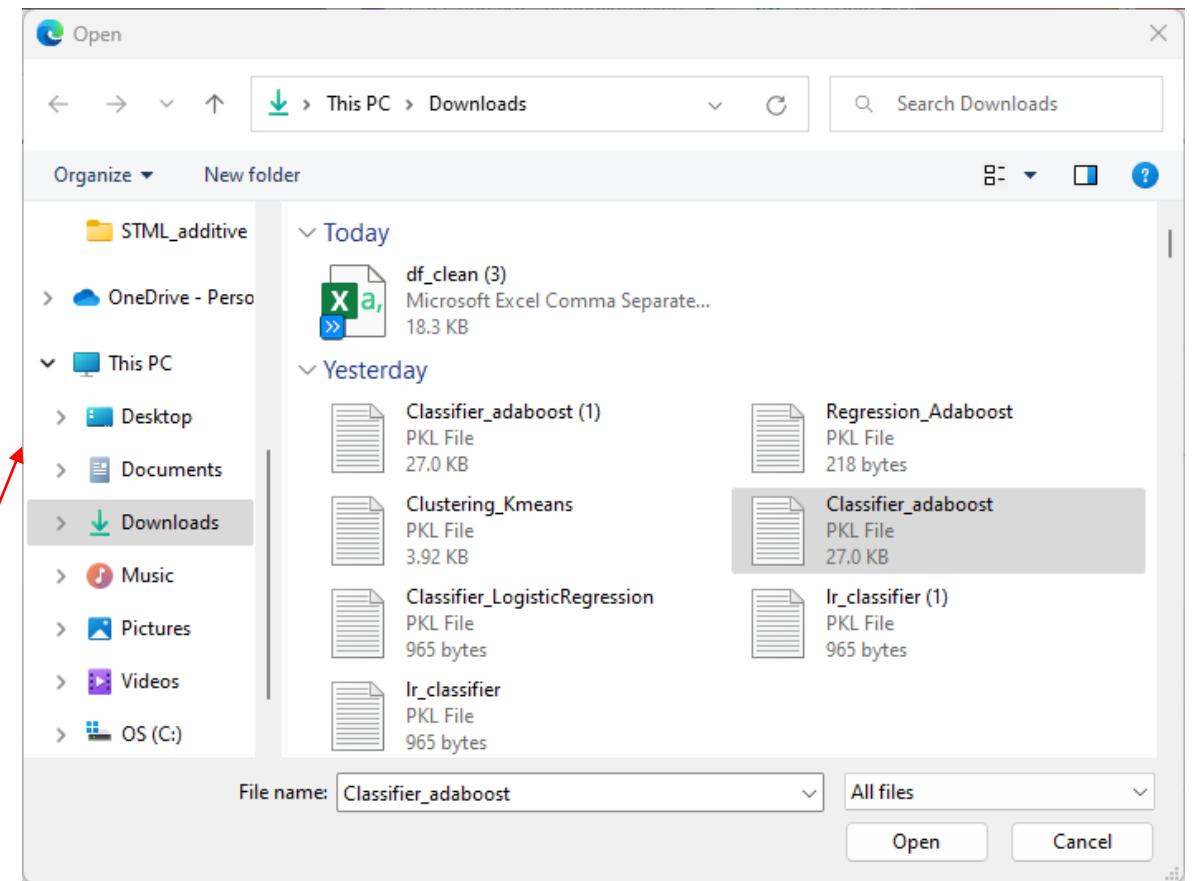
Select the model type to match the model you loaded, which is to make sure your data set is suitable for the model

Load Existed Model

The screenshot shows the 'EasyDS - Streamlit' application running on 'localhost:8501'. On the left, there's a sidebar with a logo and several menu items: Data Load (checked), EDA, Data Wrangling, Model (checked), Clustering, Classification, Regression, and Existing Model (checked). The main area has sections for Data Load (Load Raw Data, Load Clean Data), Modeling, Data Prepare (Data Define, Scaler/Dimension Reduction, Data Split, Imbalance Sampling), and Existing Model. Under Existing Model, there are two sub-sections: Load Model (Model Type: Classification) and Upload Model (a 'Drag and drop file here' field with a limit of 200MB per file). A red arrow points from the 'Click here' text to the 'Browse files' button.

Click here

Browse files



Load Existed Model

Screenshot of the EasyDS Streamlit application interface showing the "Existing Model" section.

The "Model Type" dropdown is set to "Classification".

The uploaded file is "Classifier_adaboost.pkl" (27.0KB).

A red circle highlights the "Classification" model type selection.

A red circle highlights the uploaded file "Classifier_adaboost.pkl".

A learning curve plot is displayed below, showing training and cross-validation scores over 600 training examples.

Screenshot of the EasyDS Streamlit application interface showing the results of the loaded AdaBoost classifier.

Classification report:

	precision	recall	f1-score	support
0	0.85	0.87	0.86	107
1	0.69	0.66	0.67	47
accuracy			0.81	154
macro avg	0.77	0.76	0.77	154
weighted avg	0.80	0.81	0.80	154

Cross validation scores of CV = 5: ['0.75', '0.75', '0.77', '0.82', '0.78']

Average accuracy: 0.77 with std of 0.03

Two plots are shown at the bottom:

- ROC curves: True Positive Rate vs False Positive Rate (Positive label: 1). The AdaBoostClassifier AUC is 0.80.
- DET curves: False Negative Rate vs False Positive Rate (Positive label: 1). The AdaBoostClassifier AUC is 0.80.

Here is the example to load the previous saved AdaBoostClassifier model, it gets the same results in slide 27.



Welcome for testing and give your feed back!

hhou218@gmail.com



Free from coding to make your Data Scientist life easy and vivid!