Coding free Data Science

EasyDS

# Simple Operation Manual of EasyDS

By: Huijun Hou

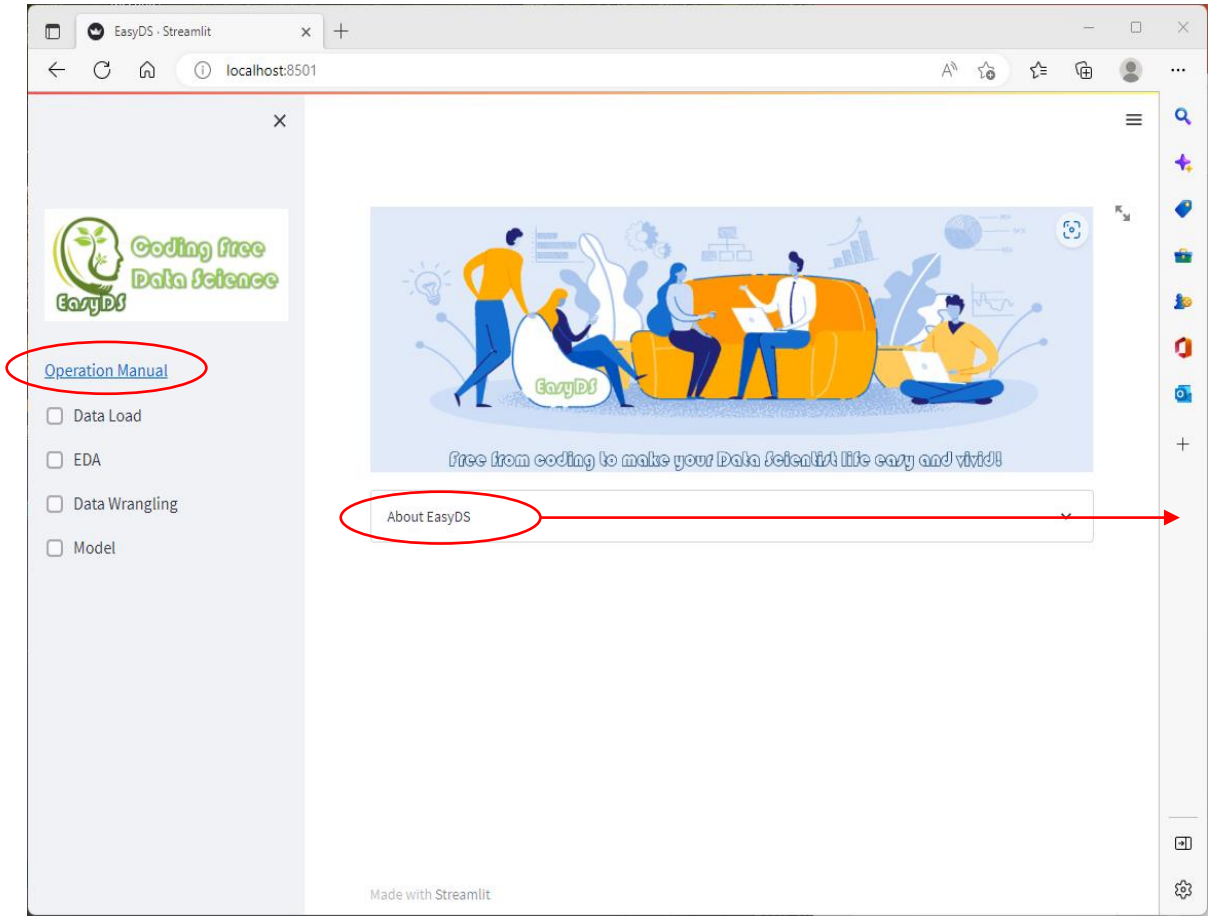
Coding free Data Science

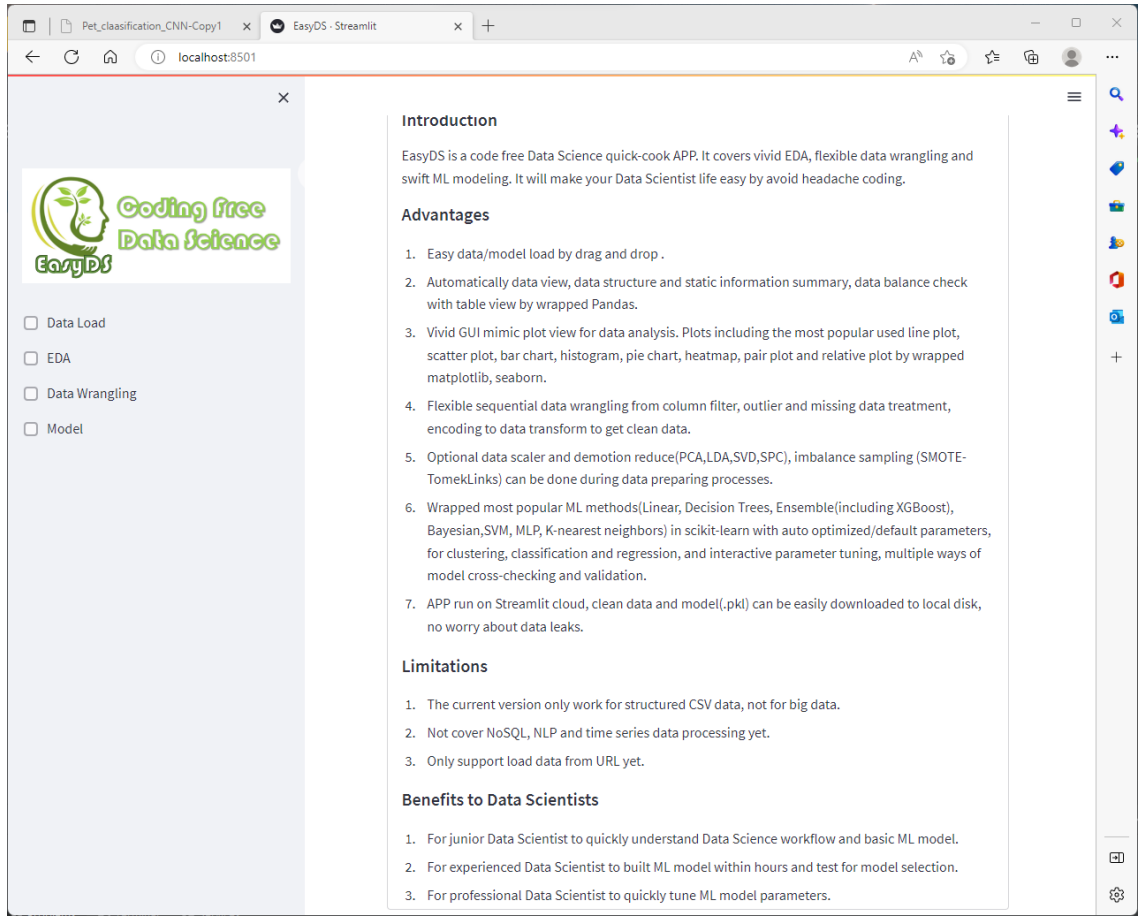
EasyDS

# Contents:



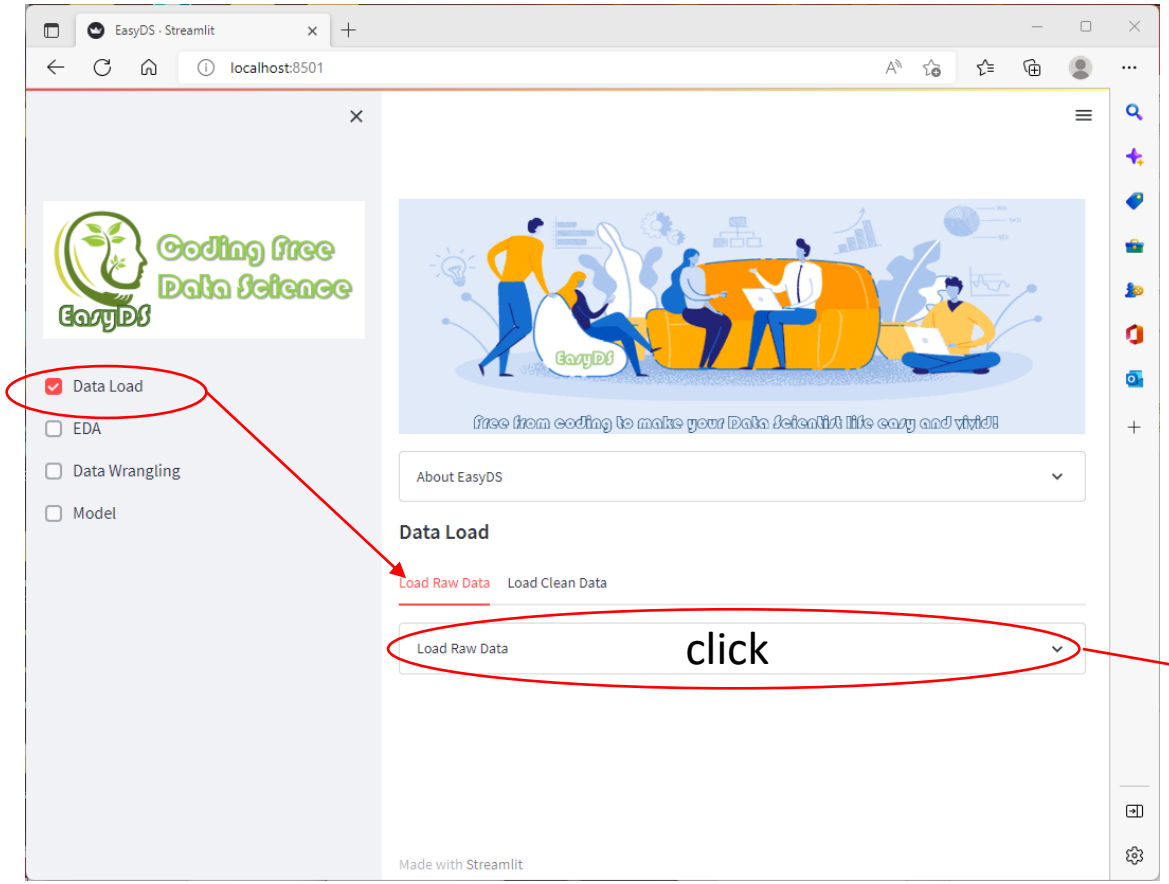
# About EasyDS

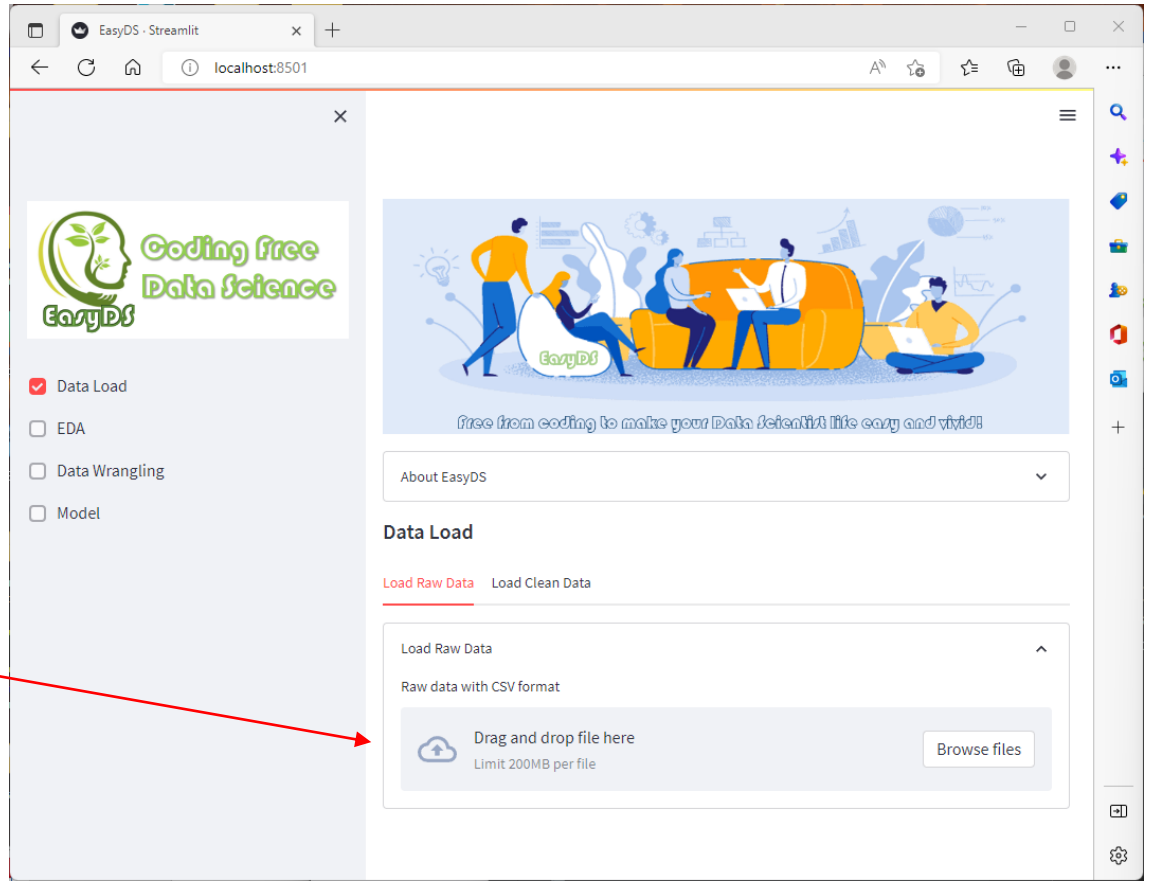


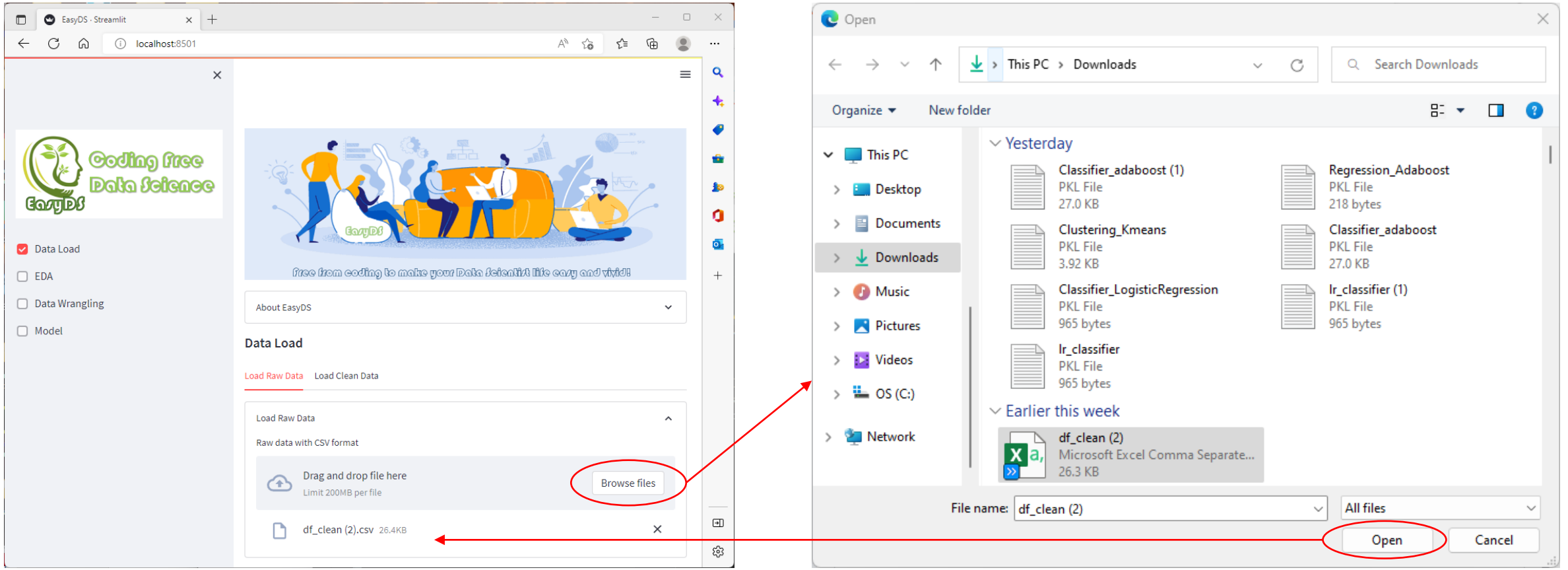

Open EasyDS, left side plot displayed the main menus.  Operation manual can be checked by click on “Operation Manual” . A brief introduction is hide under ‘About EasyDS’. You can click it to have a quick look( as displayed in right side plot), and click once more to hide it.

# Load Data

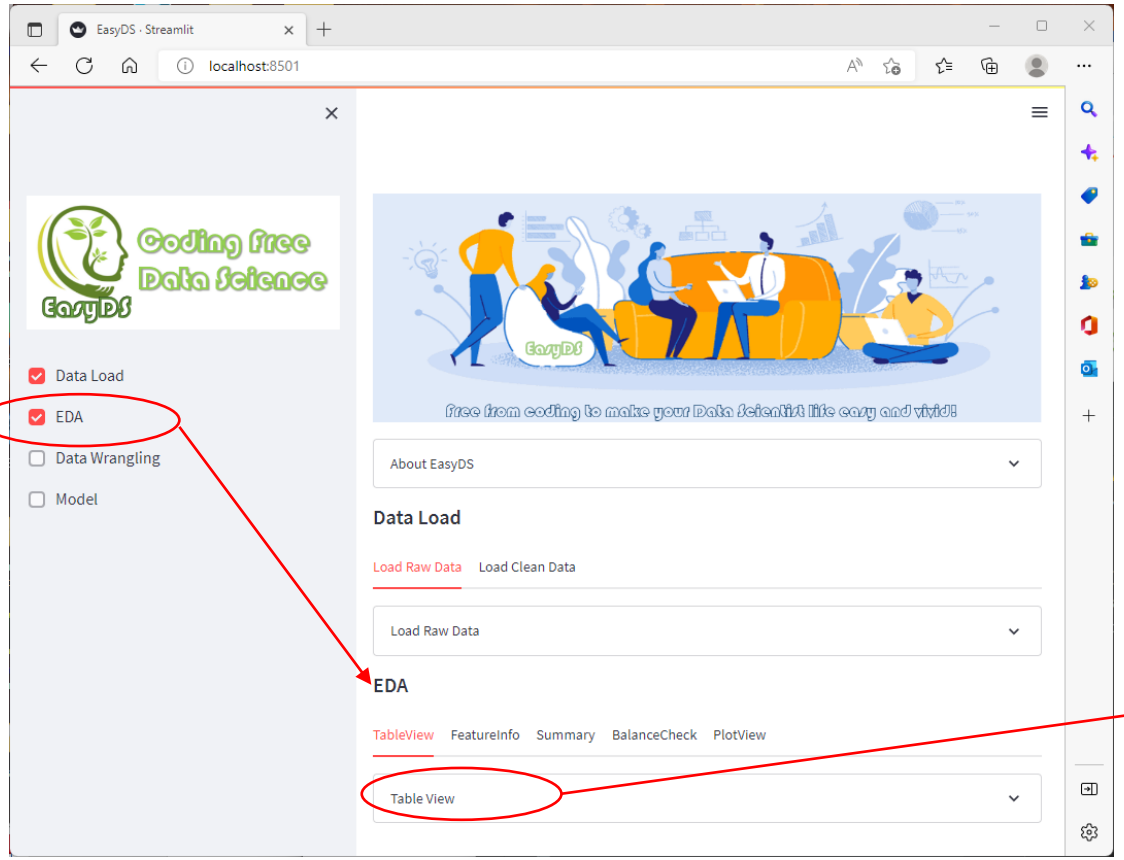

EasyDS · Streamlit
localhost:8501
Coding free Data Science
EasyDS
Data Load
EDA
Data Wrangling
Model
free from coding to make your Data Scientist life easy and vivid!
About EasyDS
Data Load
Load Raw Data
Load Clean Data
Load Raw Data
click
Made with Streamlit

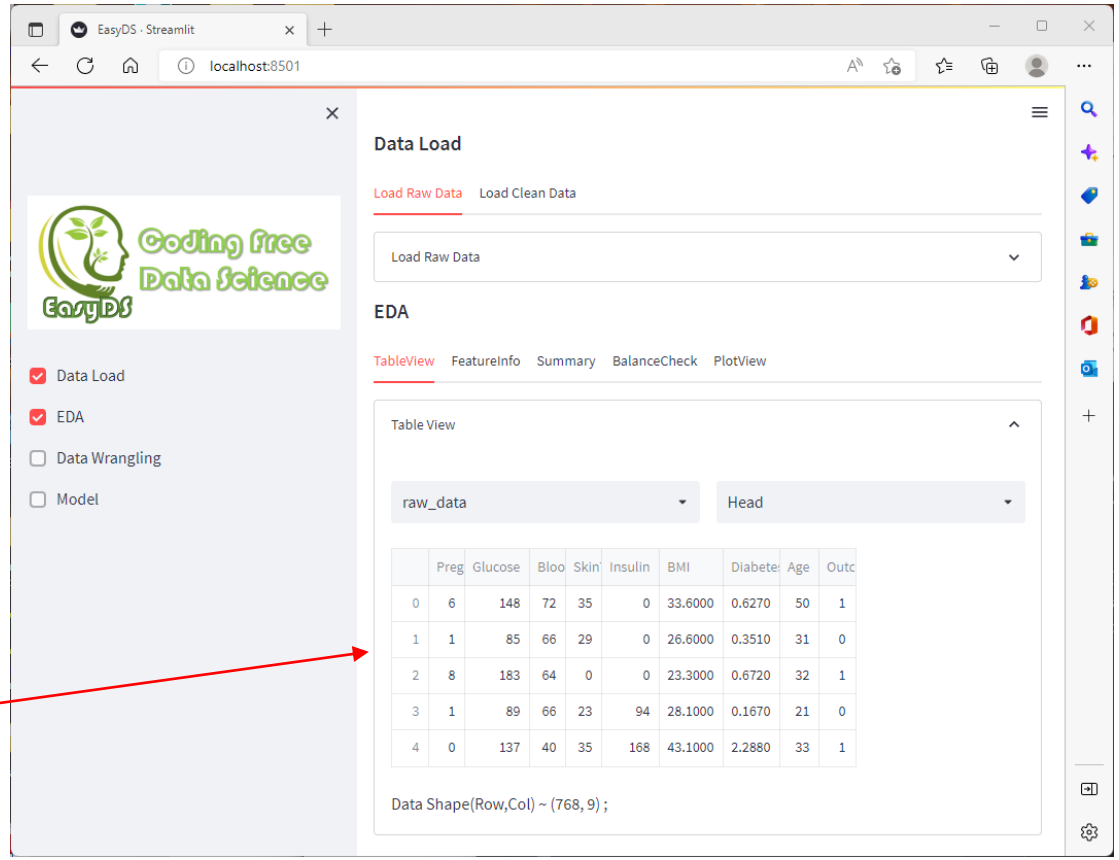

EasyDS · Streamlit
localhost:8501
Coding free Data Science
EasyDS
Data Load
EDA
Data Wrangling
Model
free from coding to make your Data Scientist life easy and vivid!
About EasyDS
Data Load
Load Raw Data
Load Clean Data
Load Raw Data
Raw data with CSV format
Drag and drop file here
Limit 200MB per file
Browse files

# Load Data

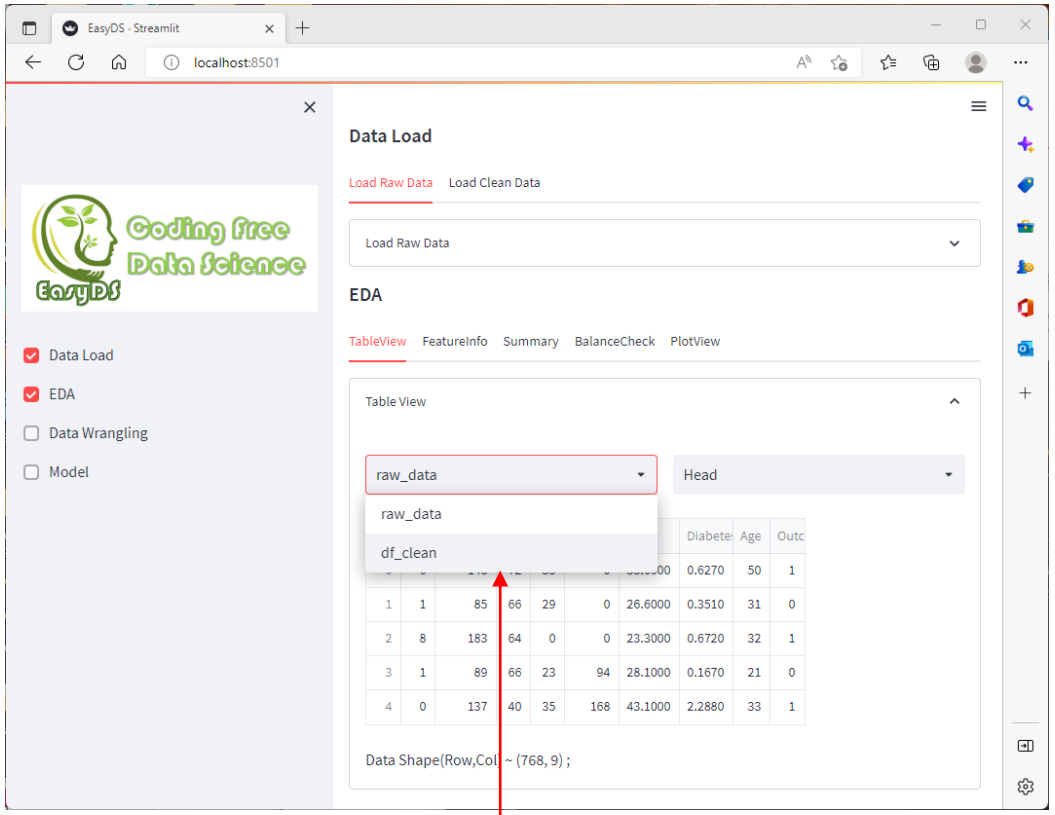

EasyDS · Streamlit
localhost:8501
Coding free Data Science
EasyDS
Data Load
EDA
Data Wrangling
Model
About EasyDS
Data Load
Load Raw Data
Load Clean Data
Load Raw Data
Raw data with CSV format
Drag and drop file here
Limit 200MB per file
Browse files
df_clean (2).csv 26.4KB
Open
This PC › Downloads
Search Downloads
Organize
New folder
This PC
Desktop
Documents
Downloads
Music
Pictures
Videos
OS (C:)
Network
Yesterday
Classifier_adaboost (1)
PKL File
27.0 KB
Regression_Adaboost
PKL File
218 bytes
Clustering_Kmeans
PKL File
3.92 KB
Classifier_adaboost
PKL File
27.0 KB
Classifier_LogisticRegression
PKL File
965 bytes
lr_classifier (1)
PKL File
965 bytes
lr_classifier
PKL File
965 bytes
Earlier this week
df_clean (2)
Microsoft Excel Comma Separate...
26.3 KB
File name: df_clean (2)
All files
Open
Cancel

# Data View in Table

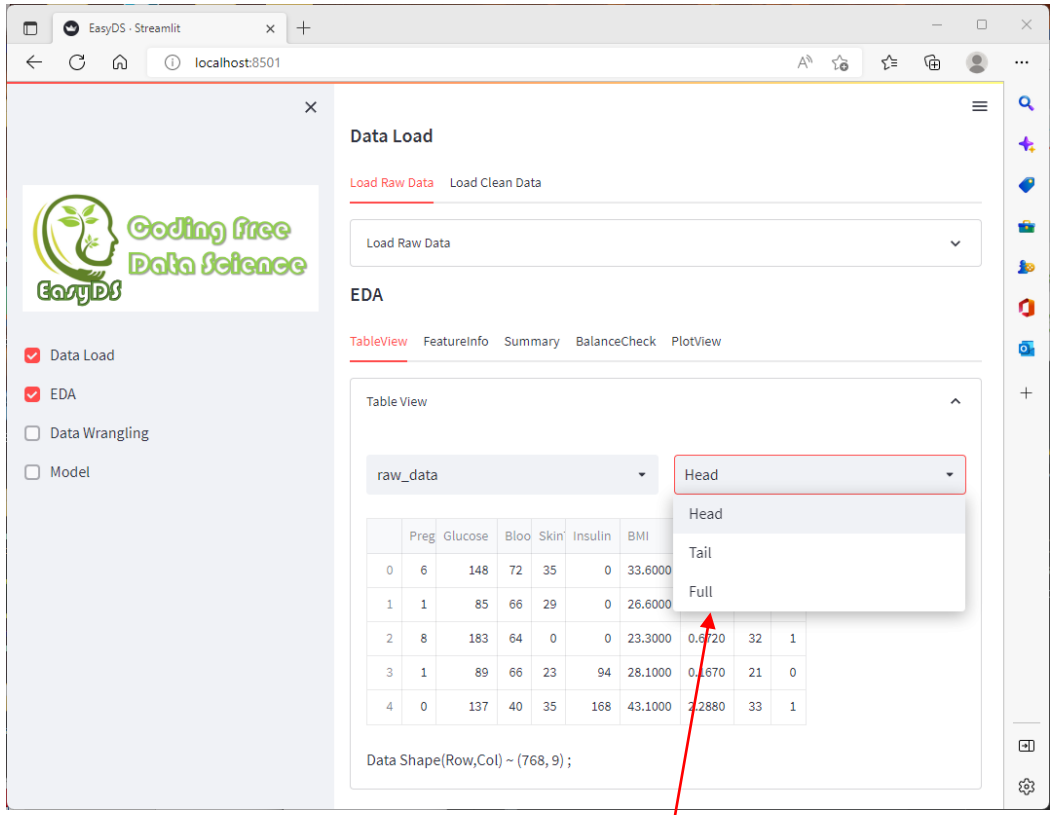


# Data View in Table

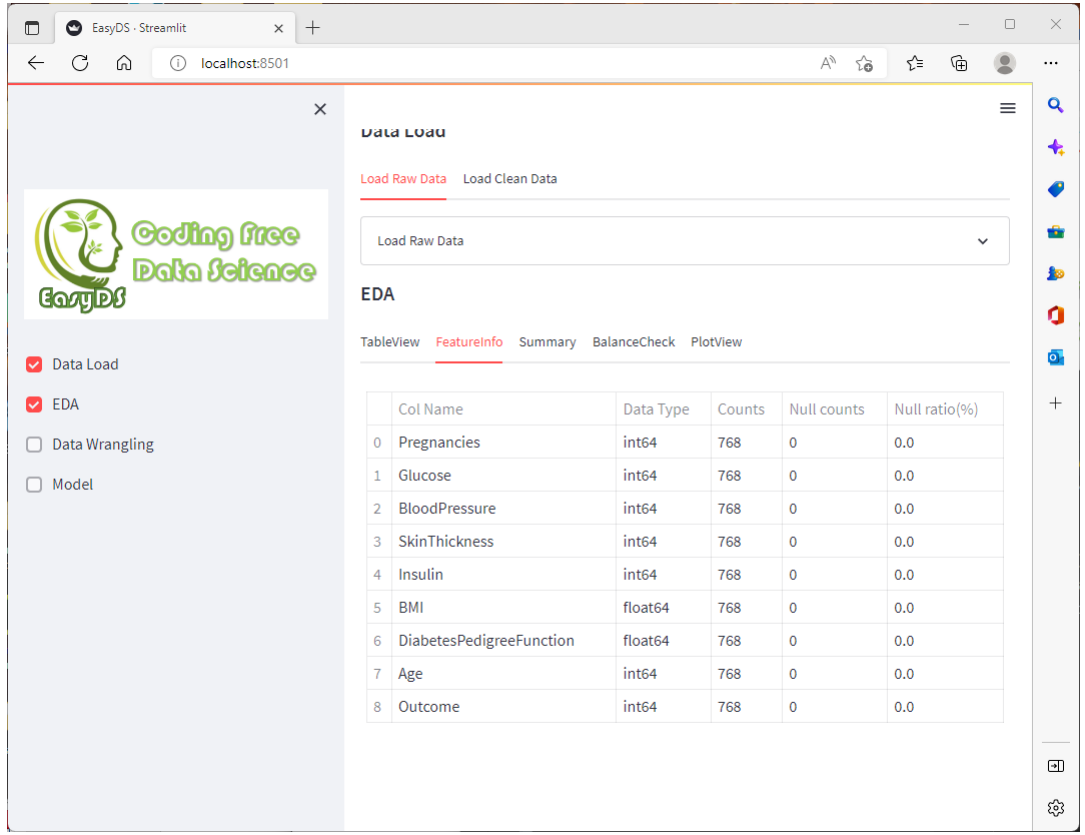



select data

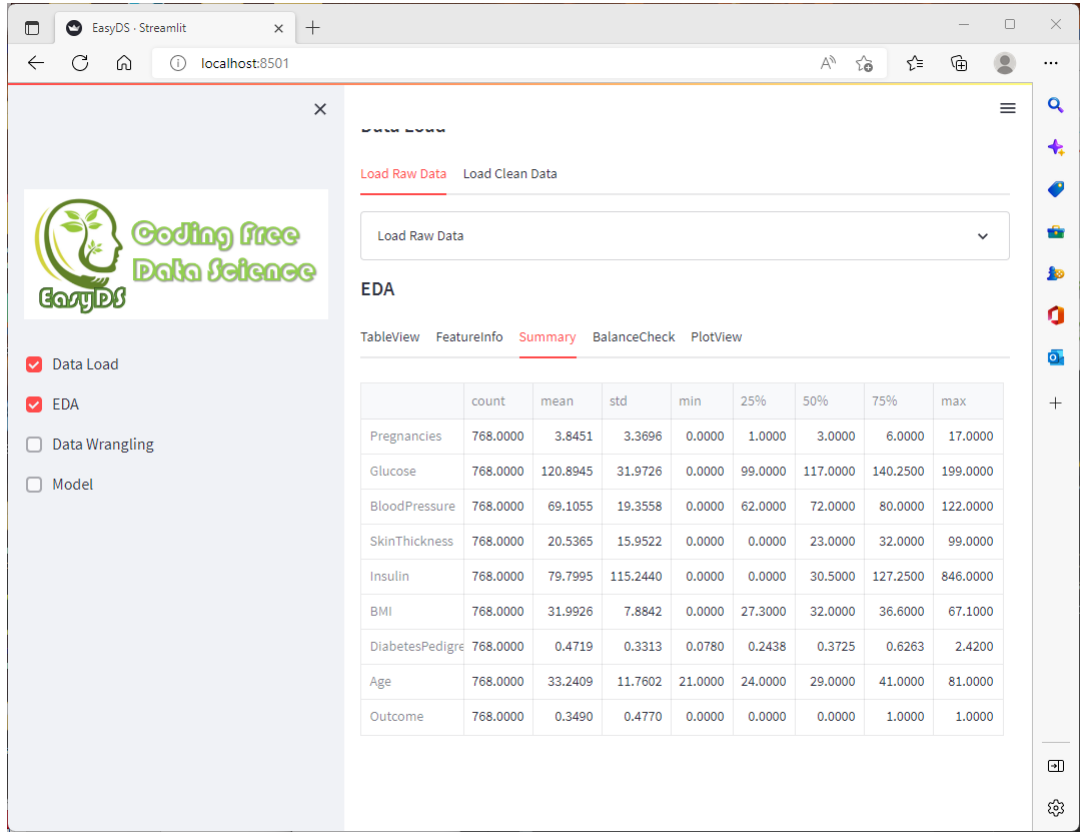



select table view method

# Data Info Summary

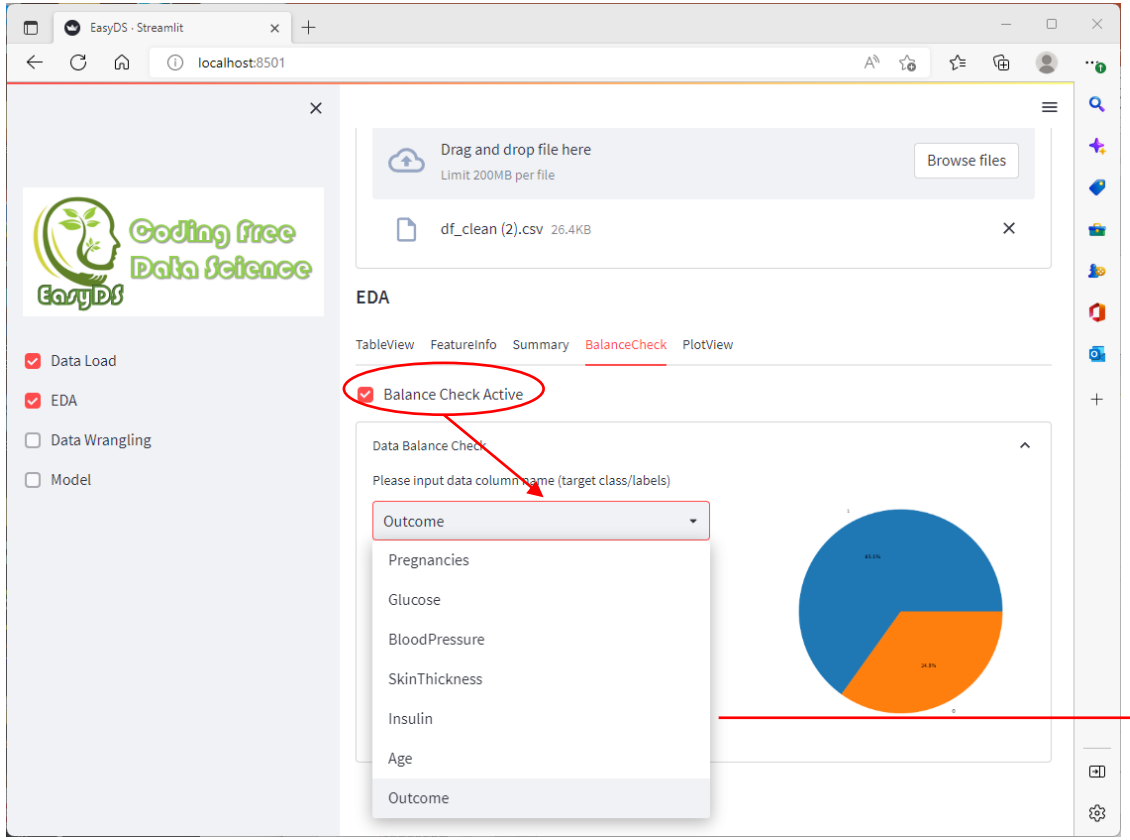



Column information

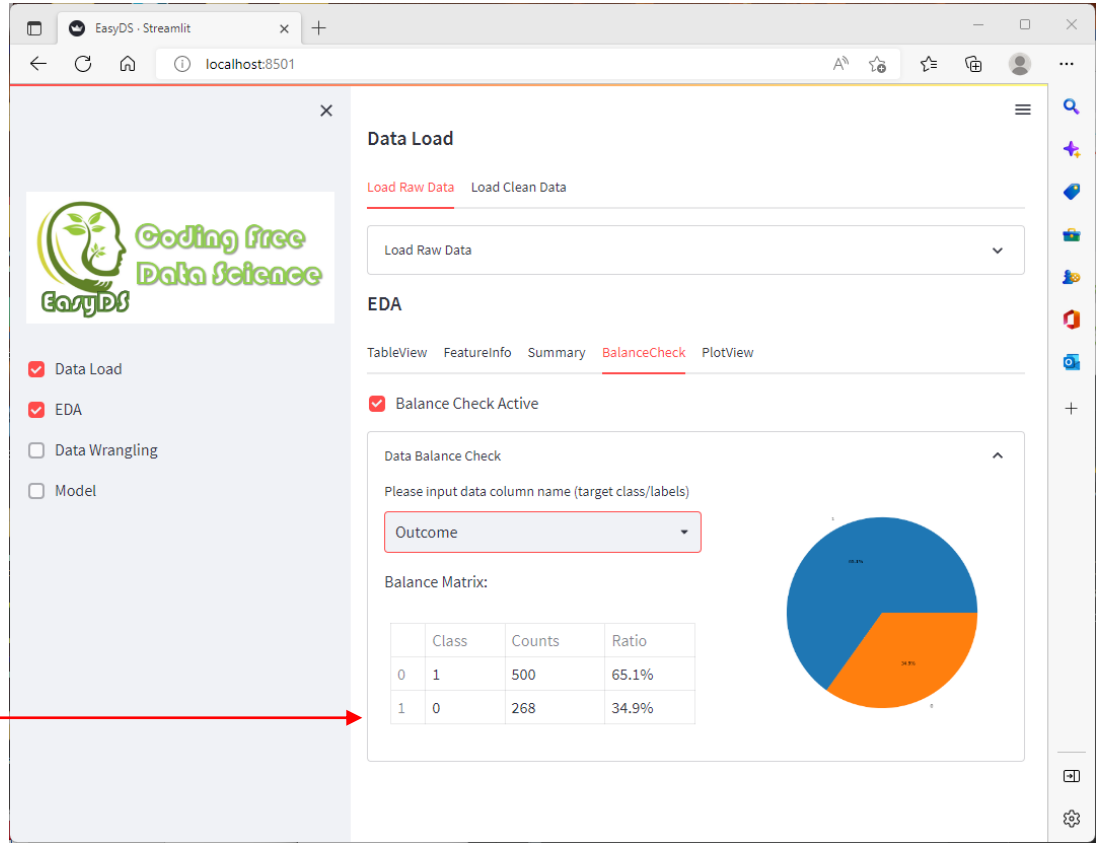



Statistical summary of numerical columns

# Data Balance Check

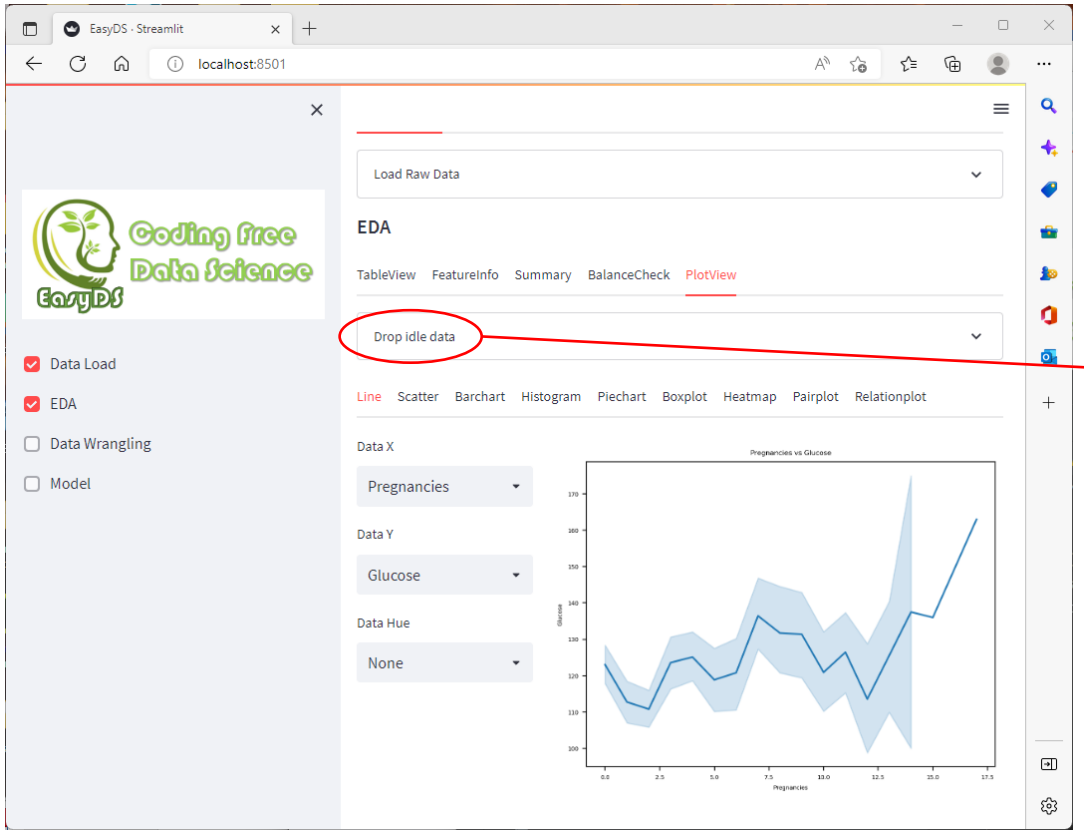


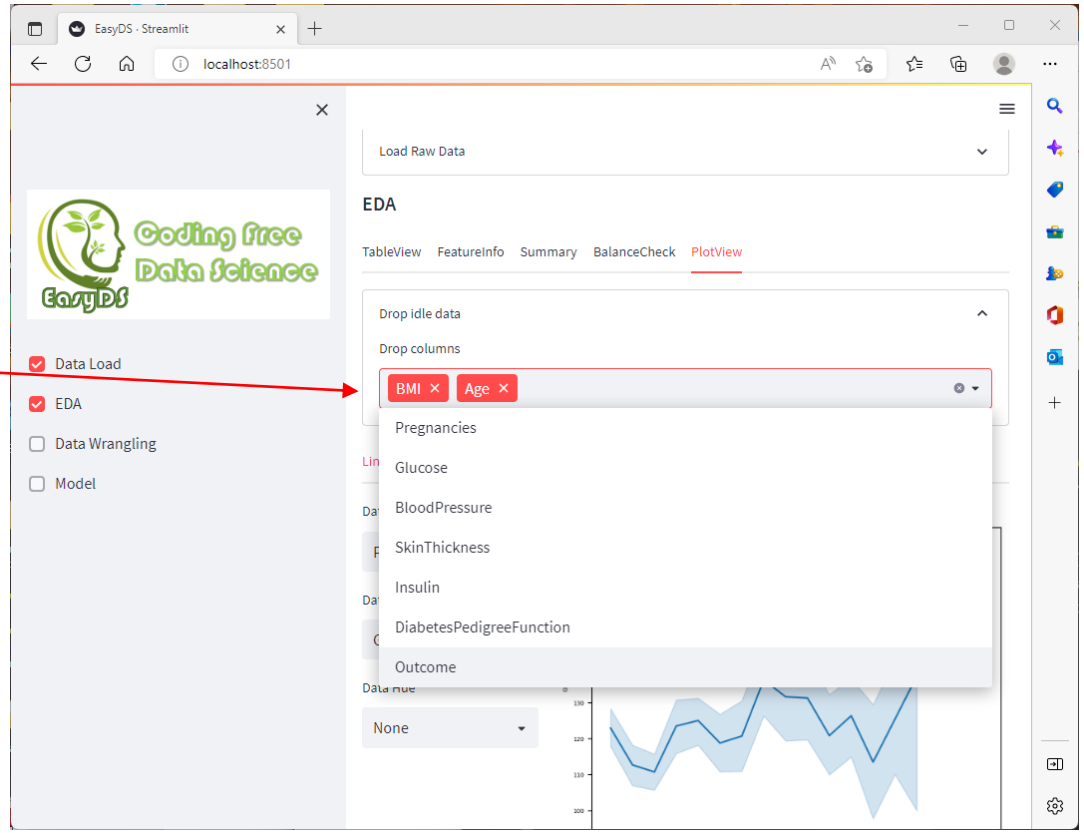

Select Target for data balance check




Data balance check display

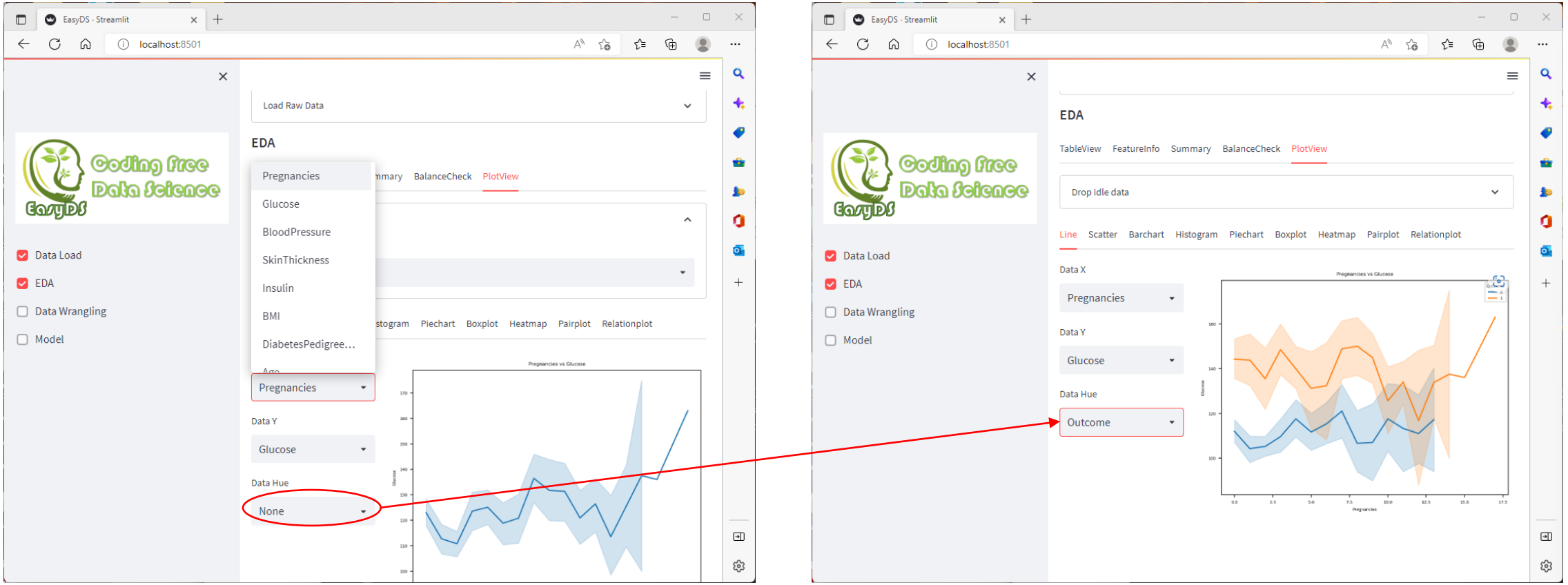
# Plot View – Line Plot



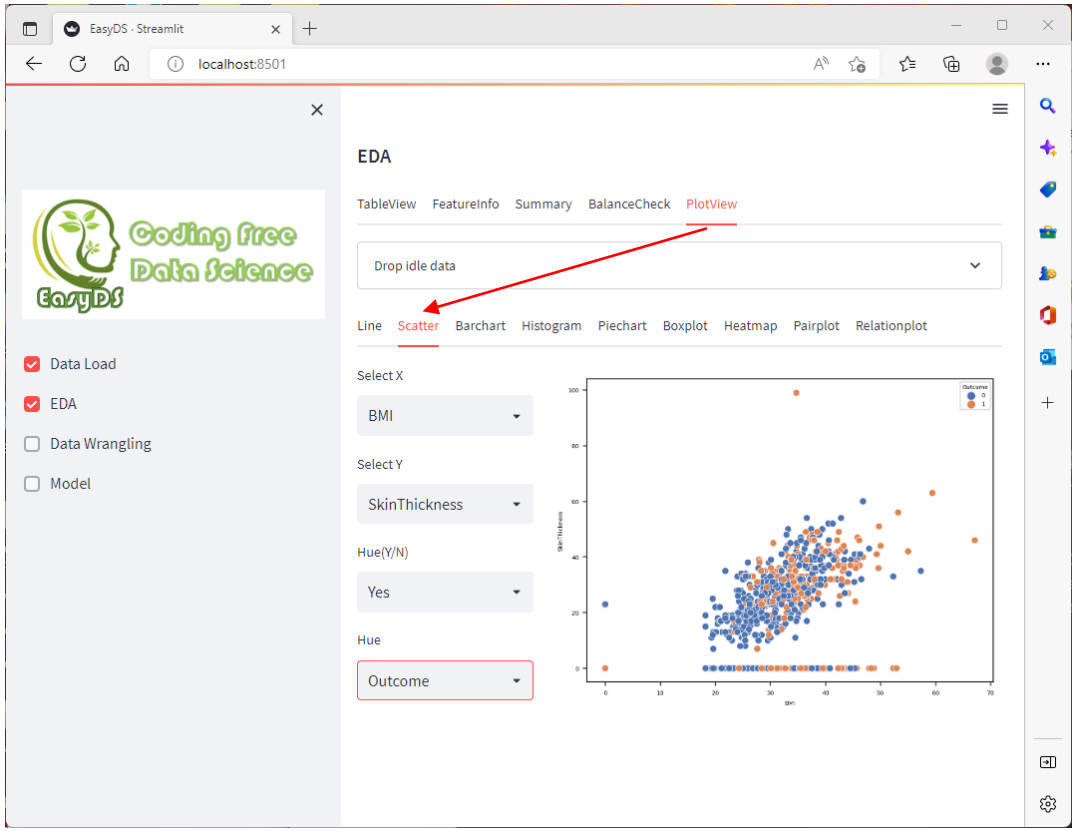

Drop the idle data to processing faster

# Plot View – Line Plot

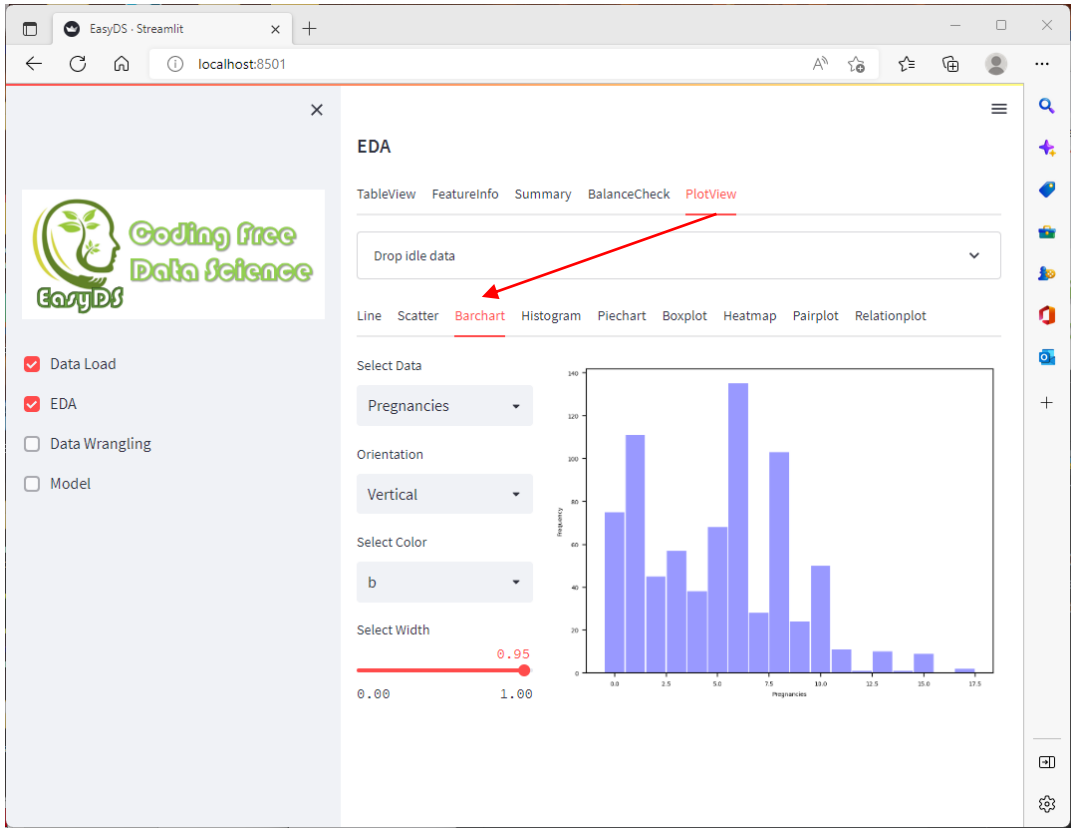



Select plot data and plot styles from drop menus

# Plot View – Scatter Plot & Bar Chart

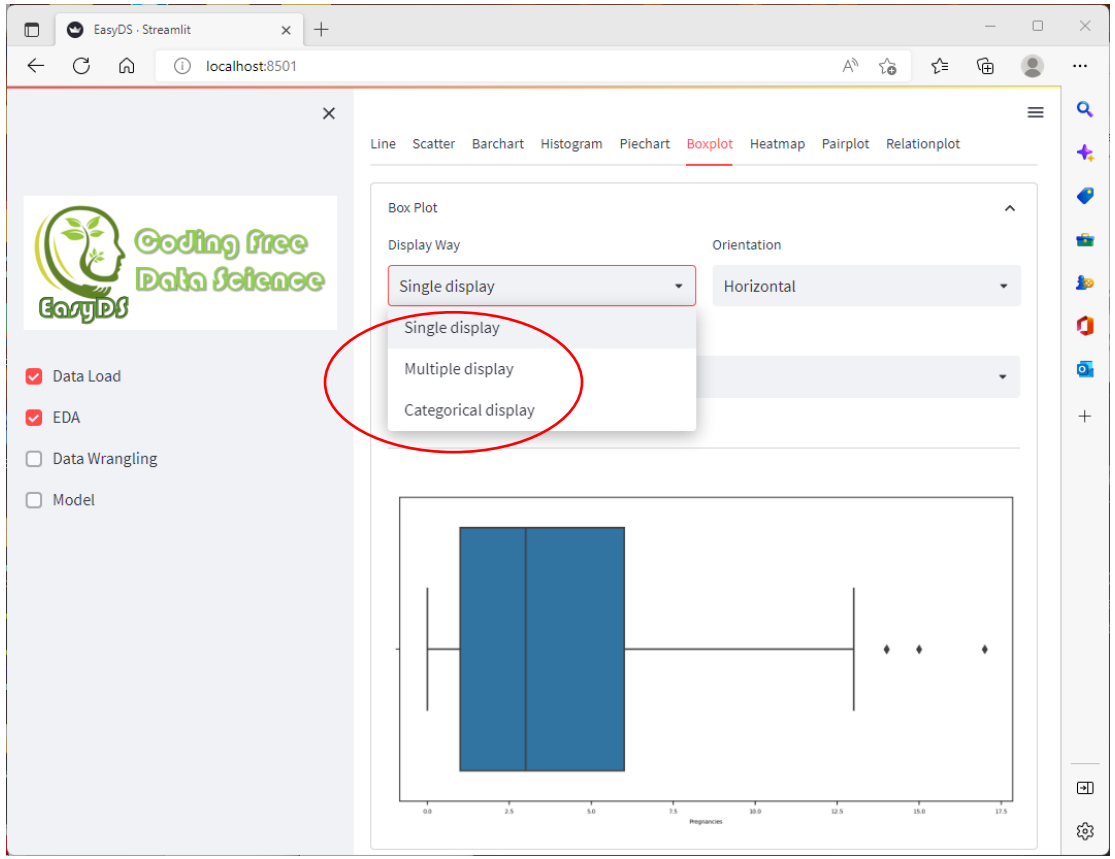


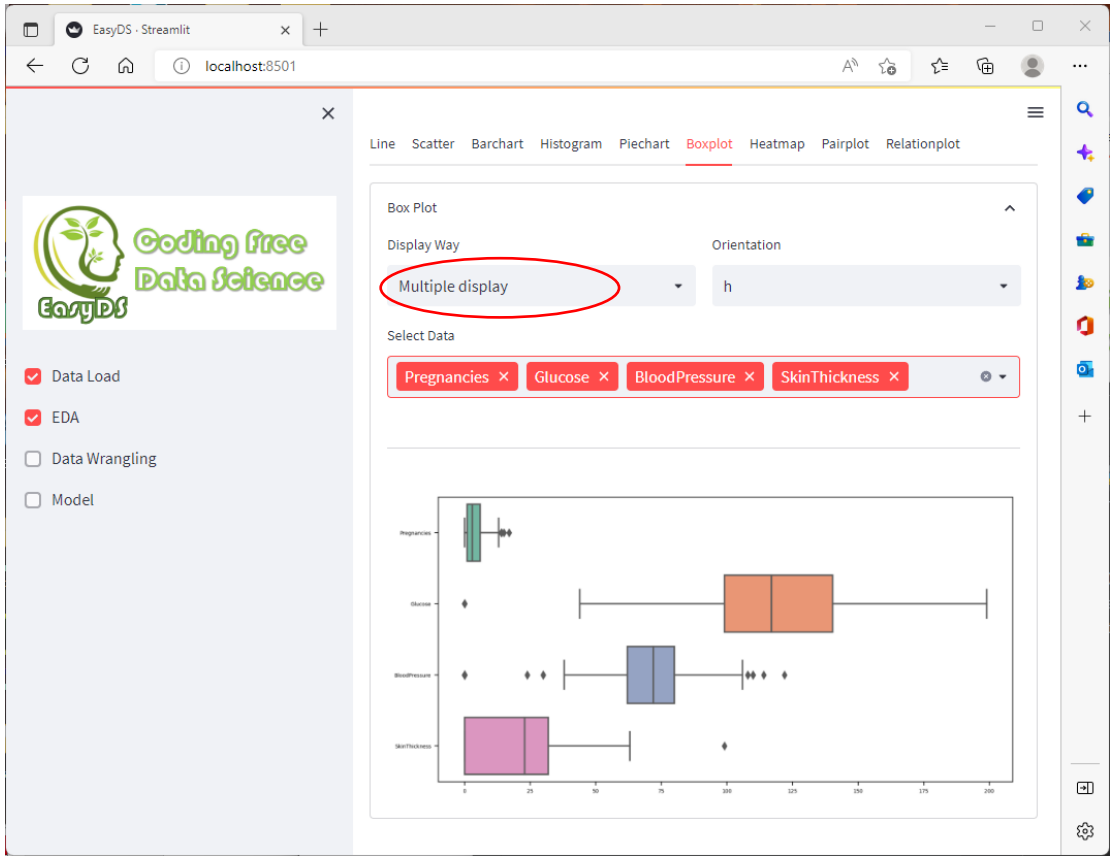

Scatter Plot Example




Bar Chart Example

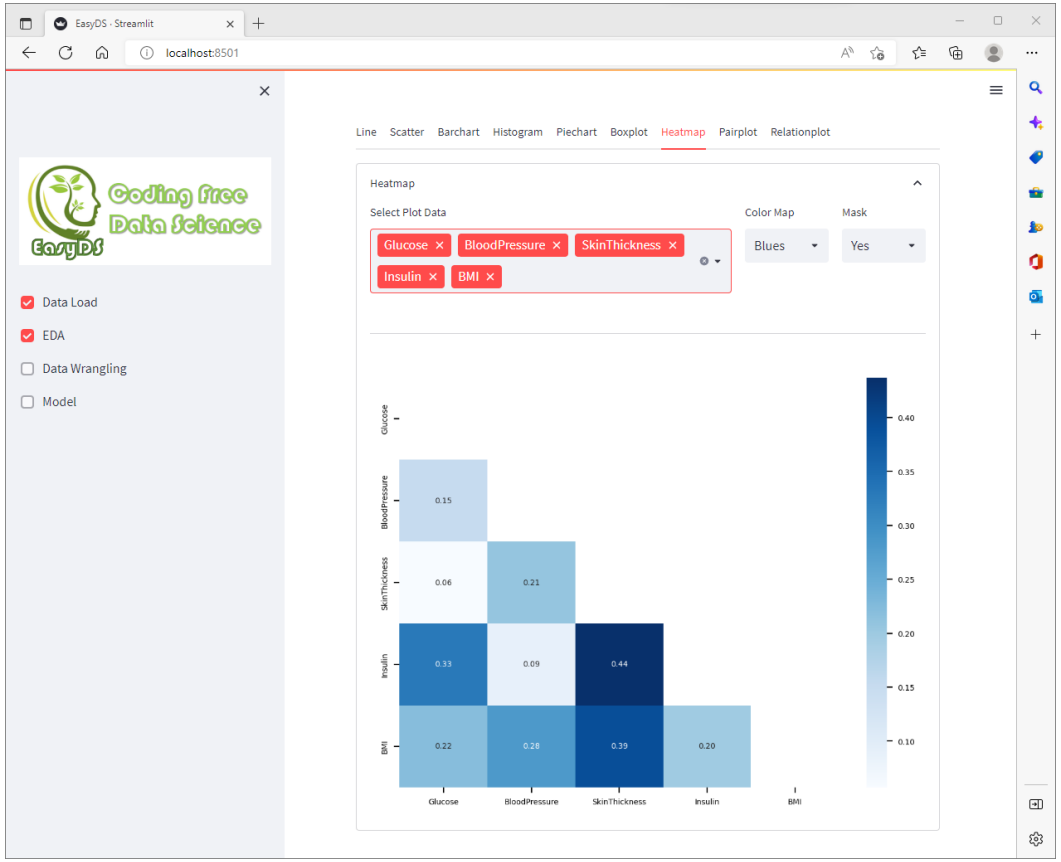
# Plot View – Box Plot



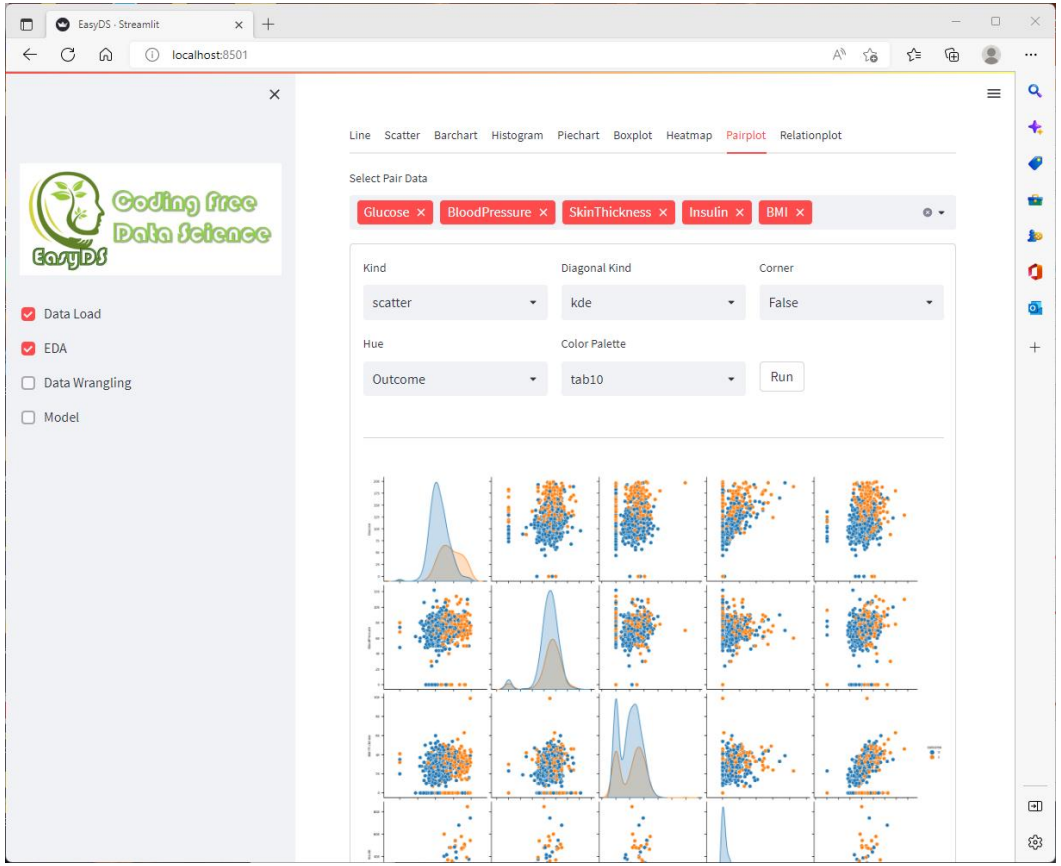




Select plot data and plot styles from drop menus

# Plot View – Heat Map & Pair Plot

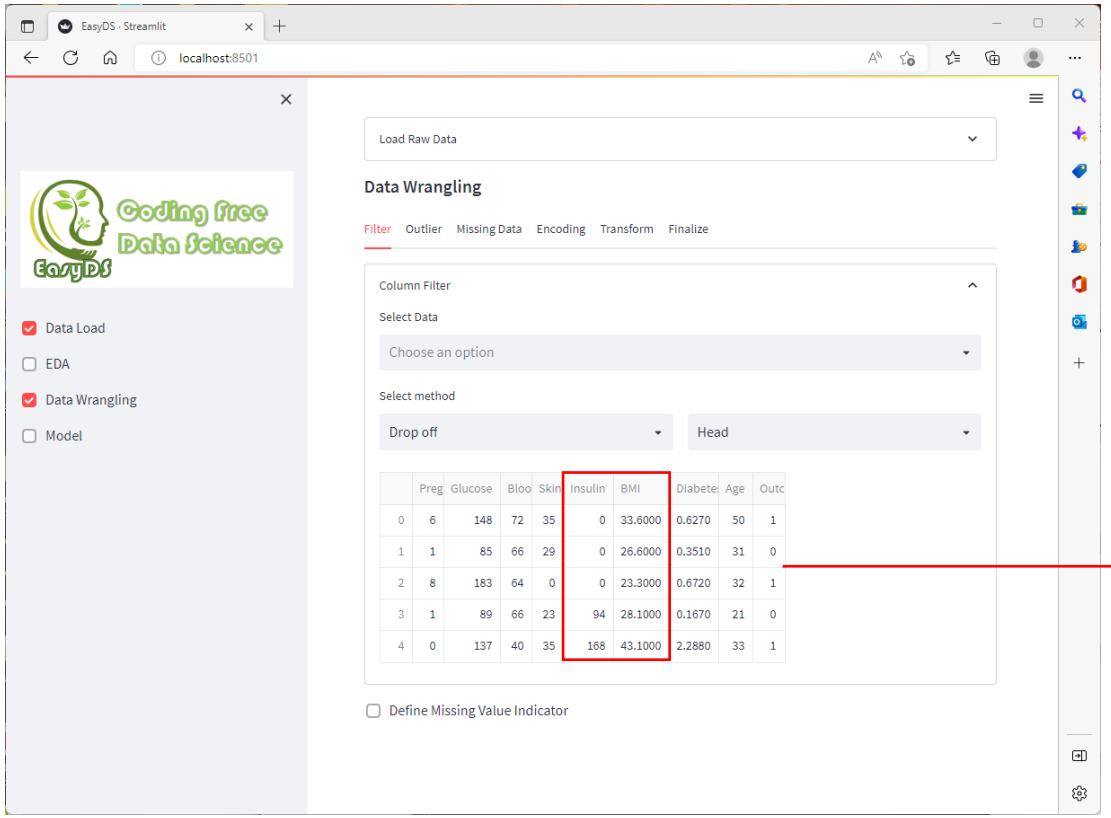


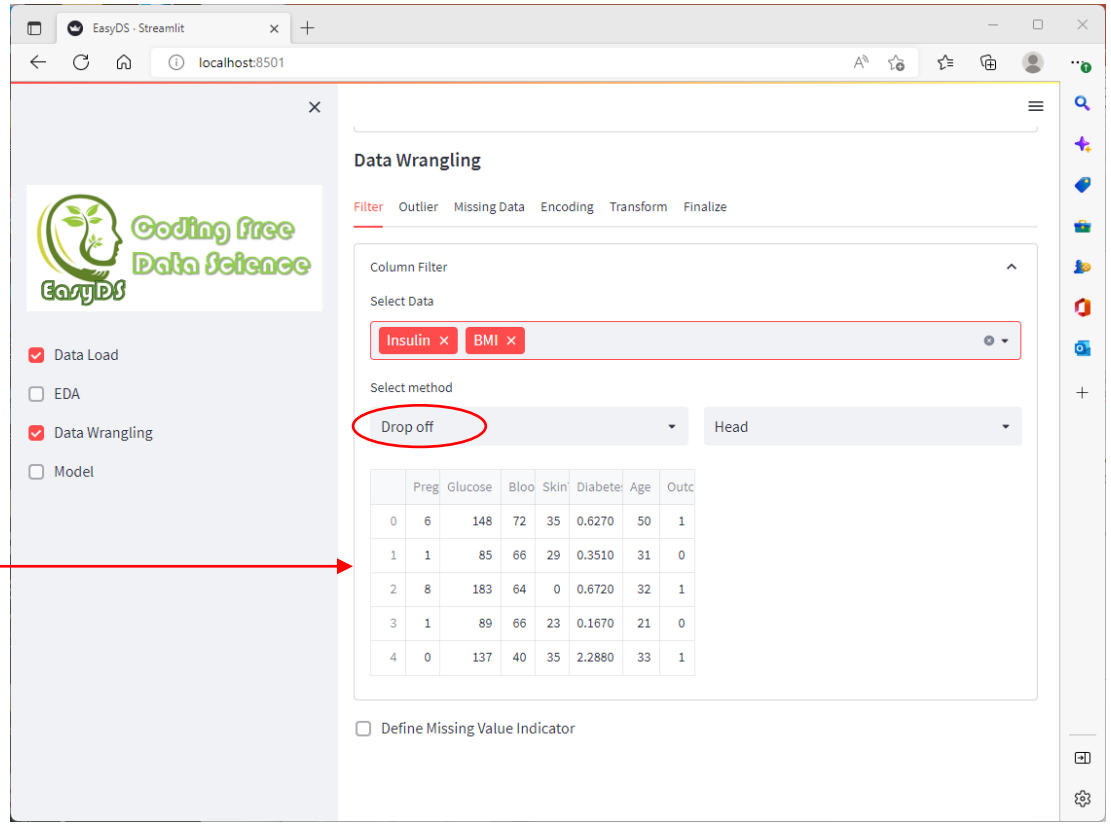

Heat Map("No" for Mask can display the upper parts)




Pair Plot

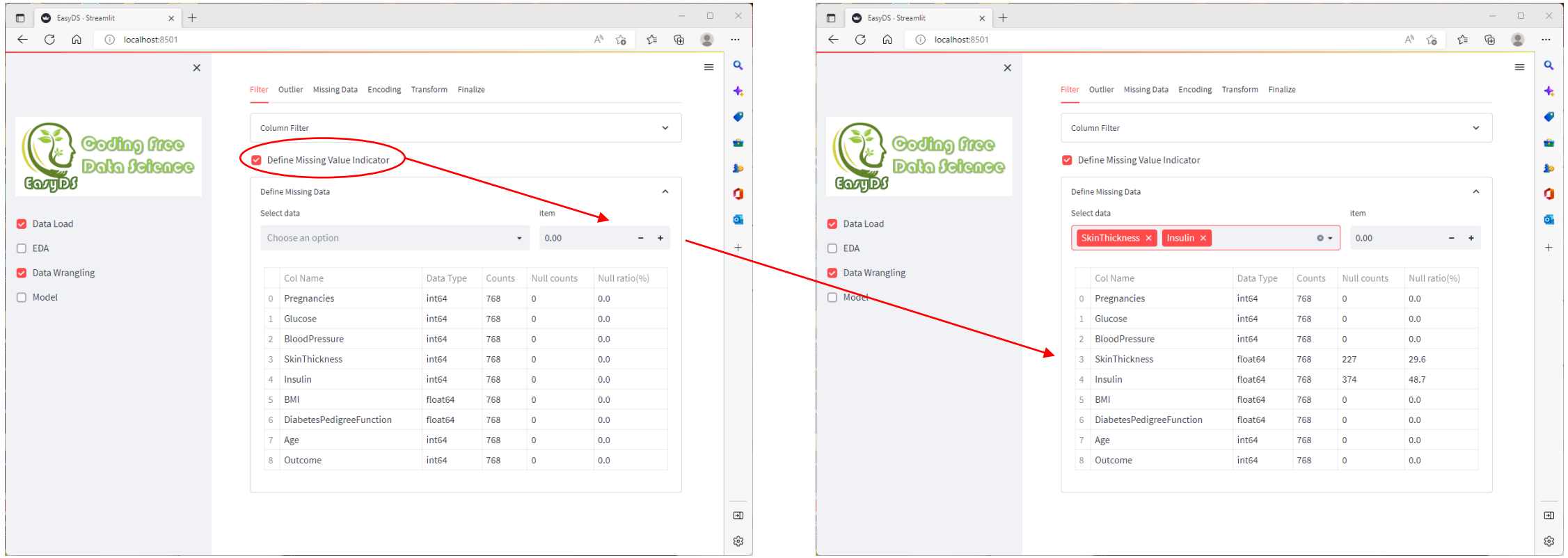
# Data Wrangling – Filter Columns



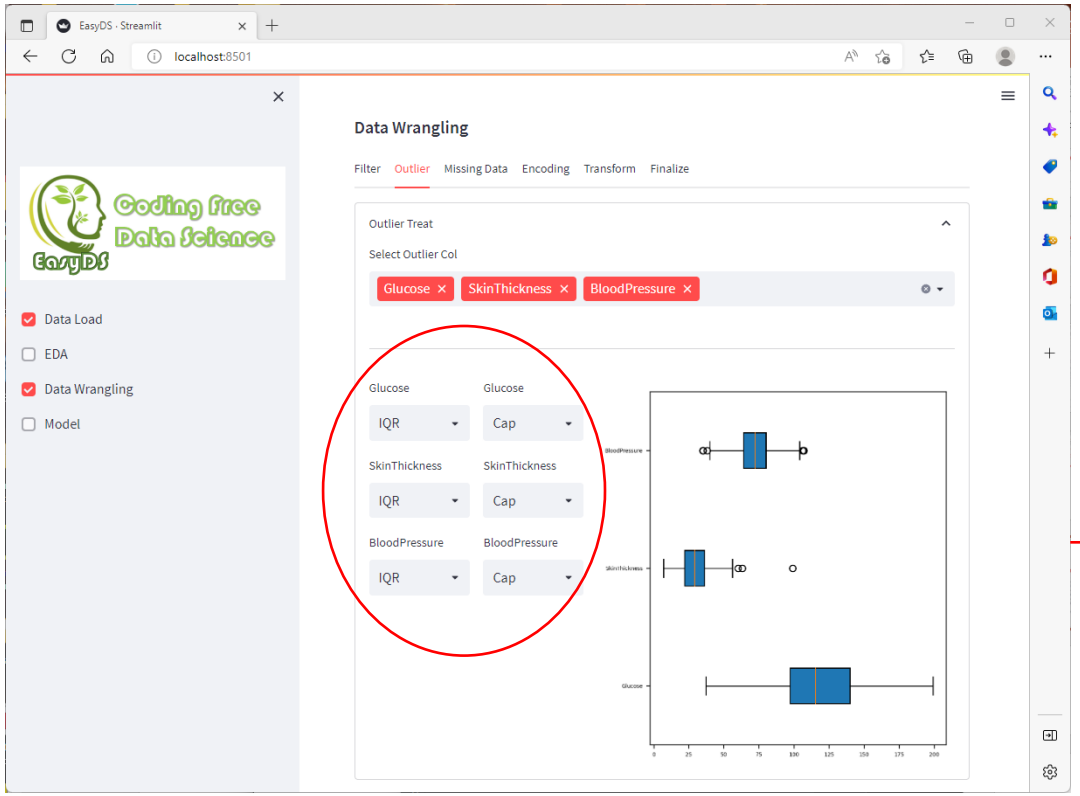

Two options for selected data: if “selected”, only selected data keep in the data frame; or, if “Dropoff”, the selected data will be dropped off as this example displayed.

# Data Wrangling – Define Null

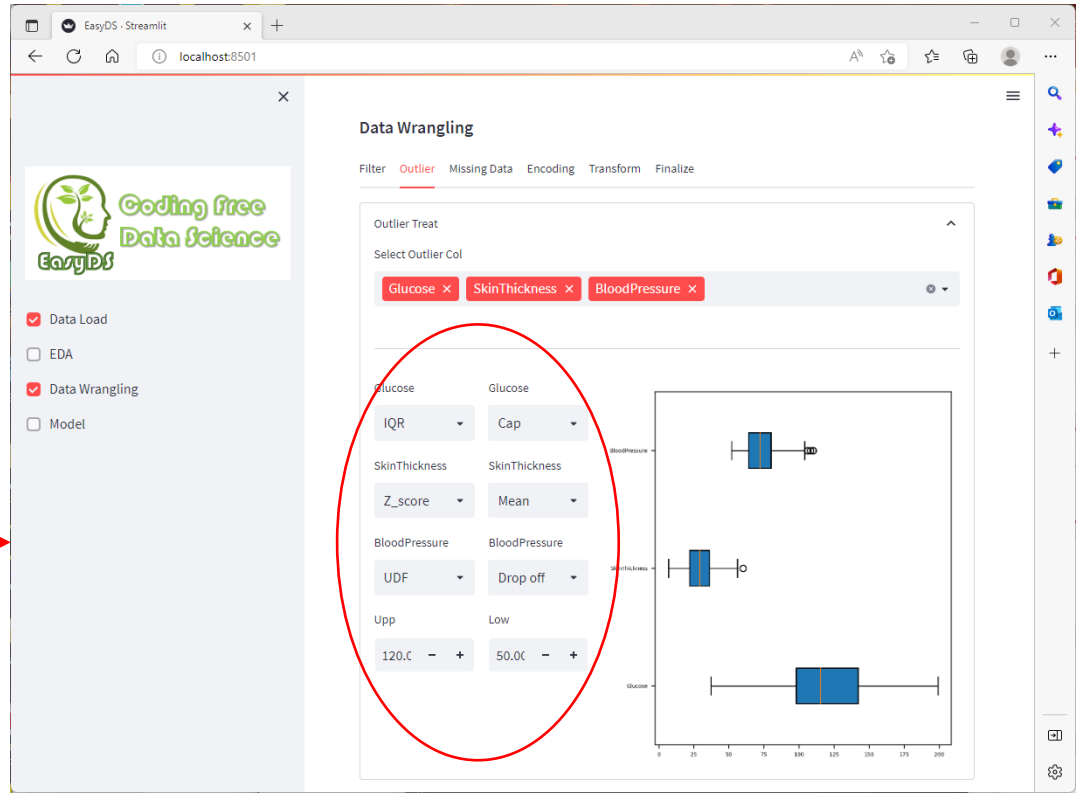



Sometimes the missed value is saved as ’0’ or ‘-999.999’ in the dataset. In such case they need to be identified as ‘null’ for later missing value treatment. In this example ‘0’ were saved for missing values of ‘SkinThickness’ and ‘Insulin’.

# Data Wrangling – Treat Outliers

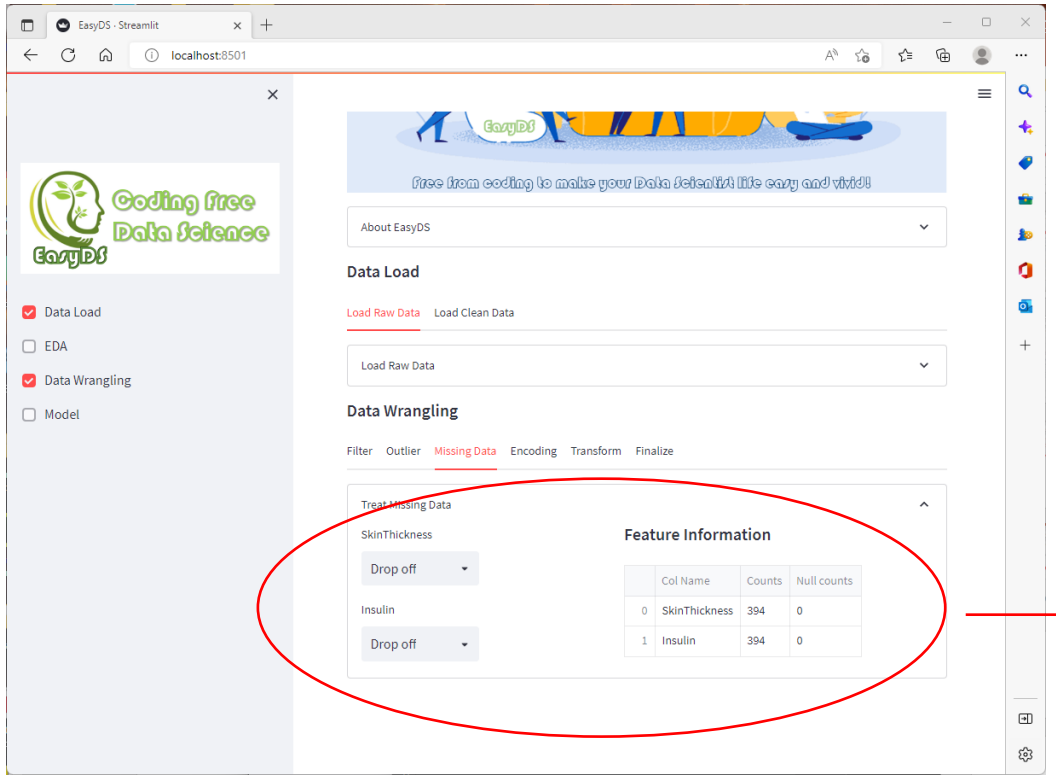


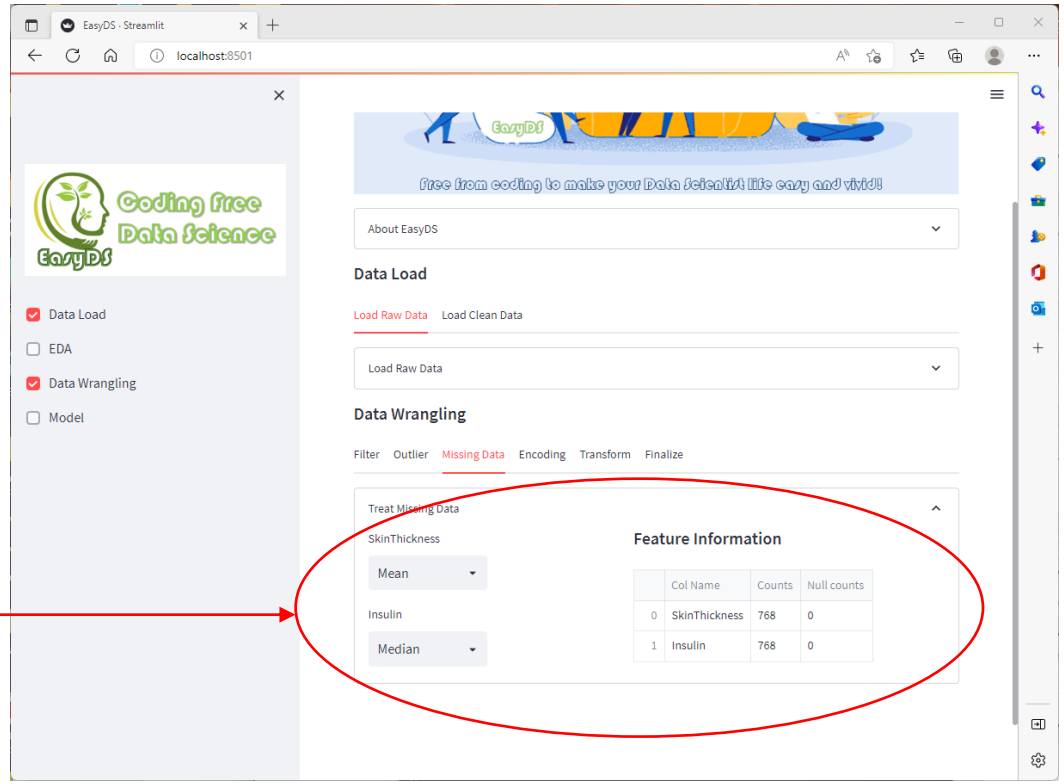

Default outliers treatment is capped by IQR method




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

# Data Wrangling – Treat Missing Values

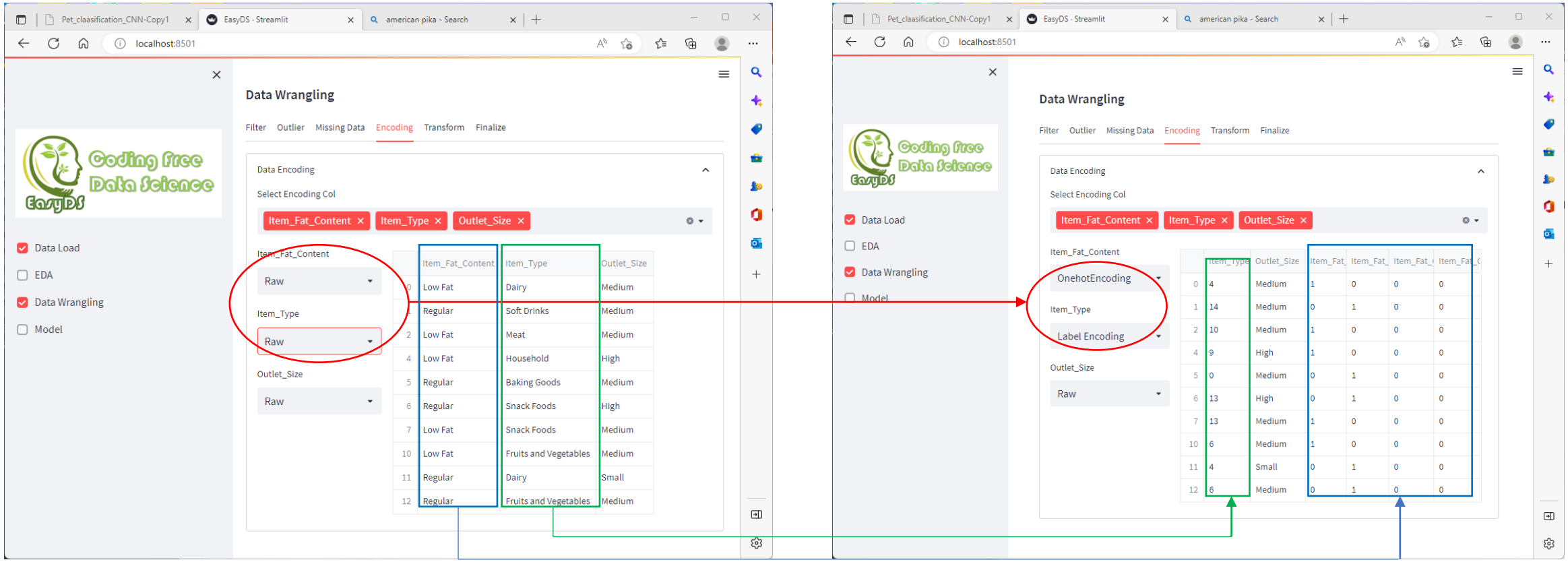



As previous mentioned, there are missing values for ‘SkinThickness’ and ‘Insulin’ in this example . The missing data can be dropped off if it is less.

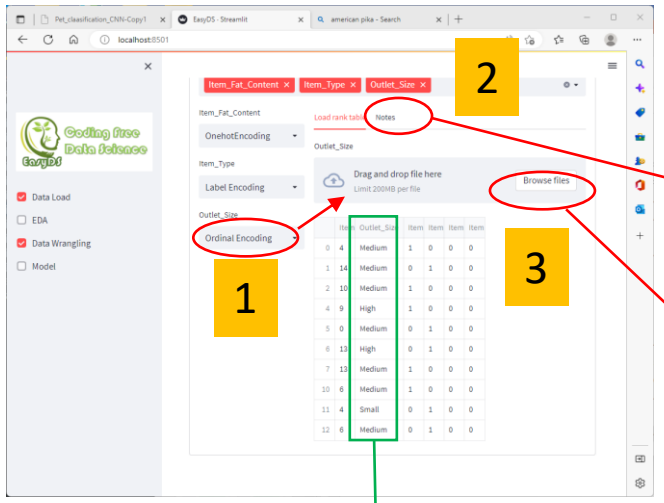
Or other strategies can be applied, such as ‘mean’ or ‘median’ as displayed in this example

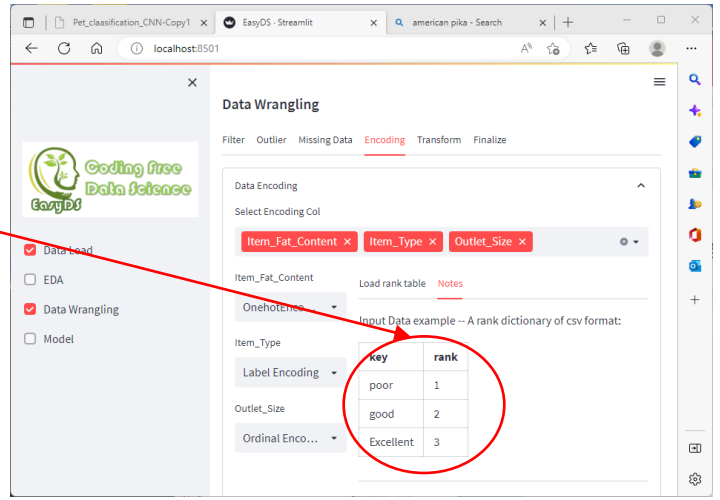
# Data Wrangling – Encoding

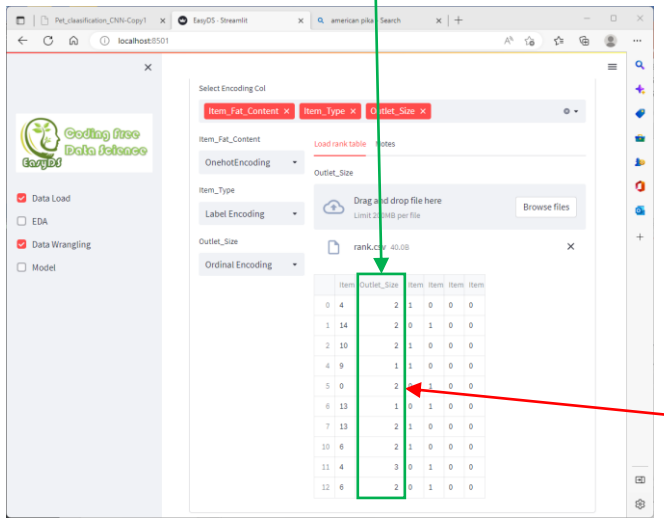


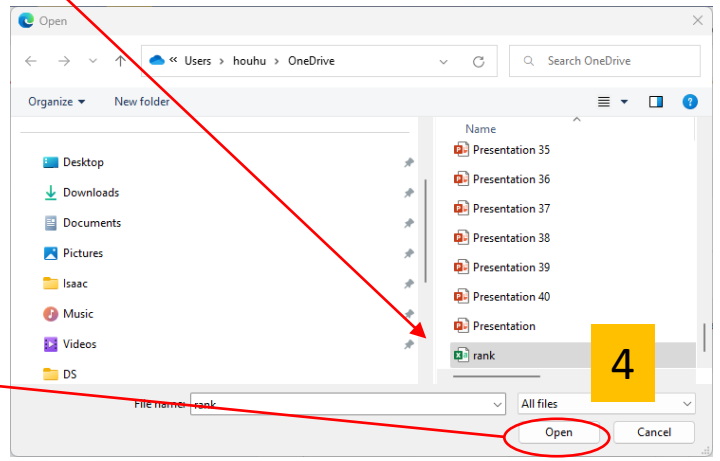

Encoding is transform categorical data into numerical data so that they can be used for modeling. Selecting all the columns to be encoded, the default raw values are displayed(left side plot) . Different methods can be selected to encode the data. This example showed the Label  encoding and One Hot Encoding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

# Data Wrangling – Ordinal Encoding





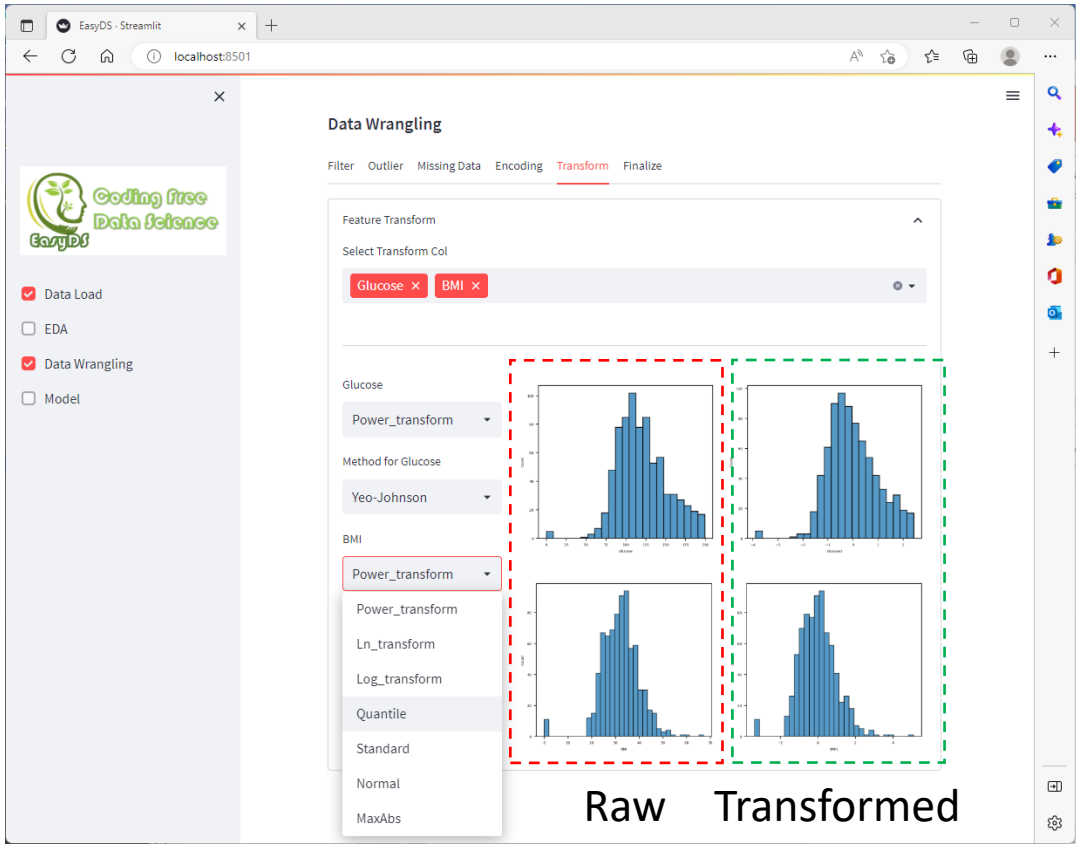

1. Sometimes it needs Ordinal Encoding. An ordinal index file in CSV need be firstly defined and upload.

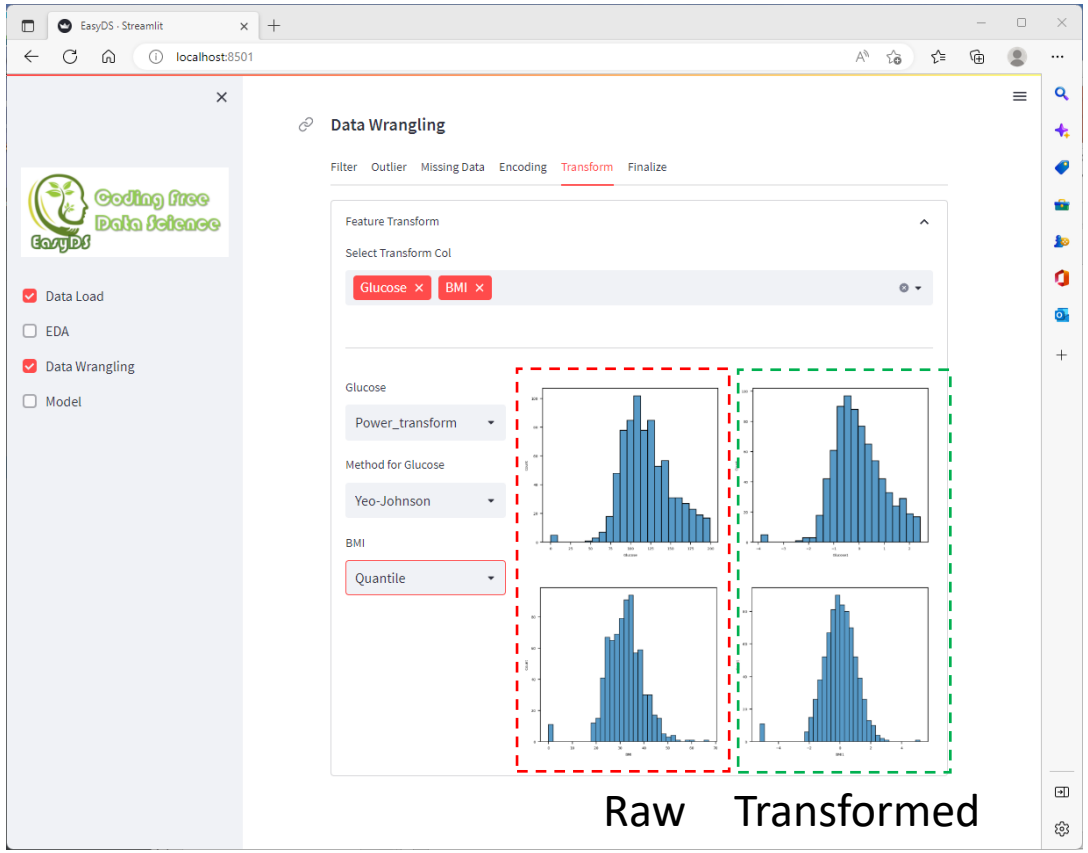
2. To check the ordinal index file format for reference, please click ‘Notes’.

3. Click browse files to search the file.

4. Open select ordinal index file to encode the data.

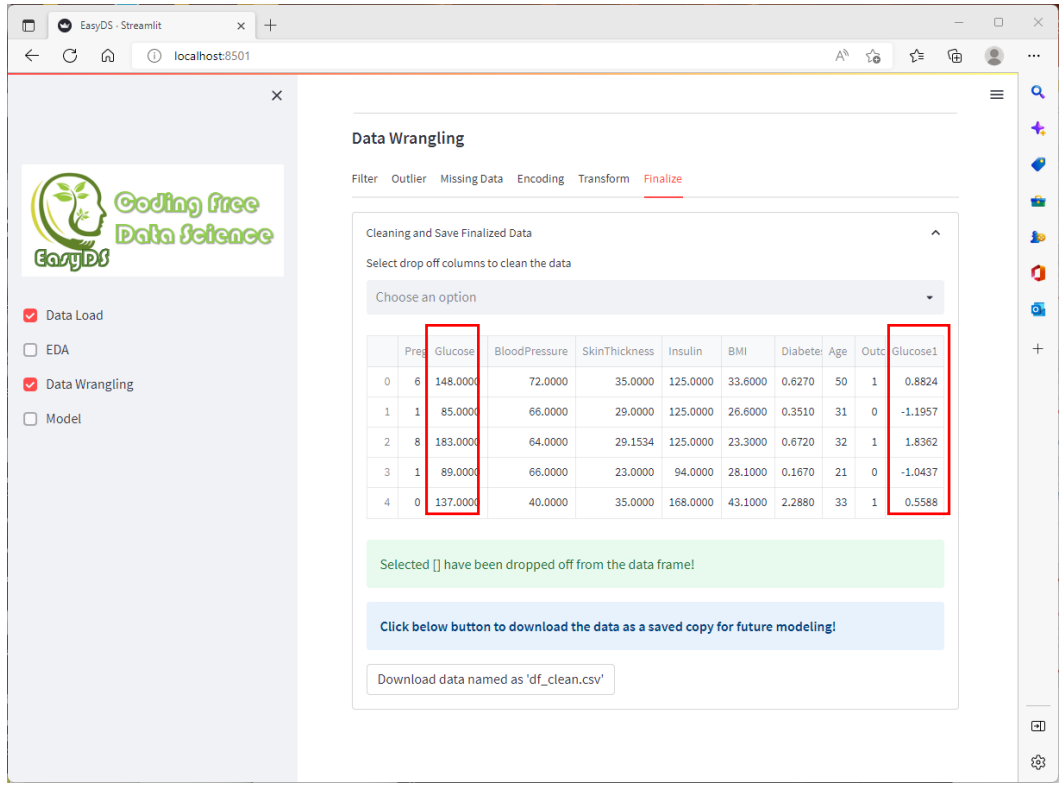
# Data Wrangling – Transform

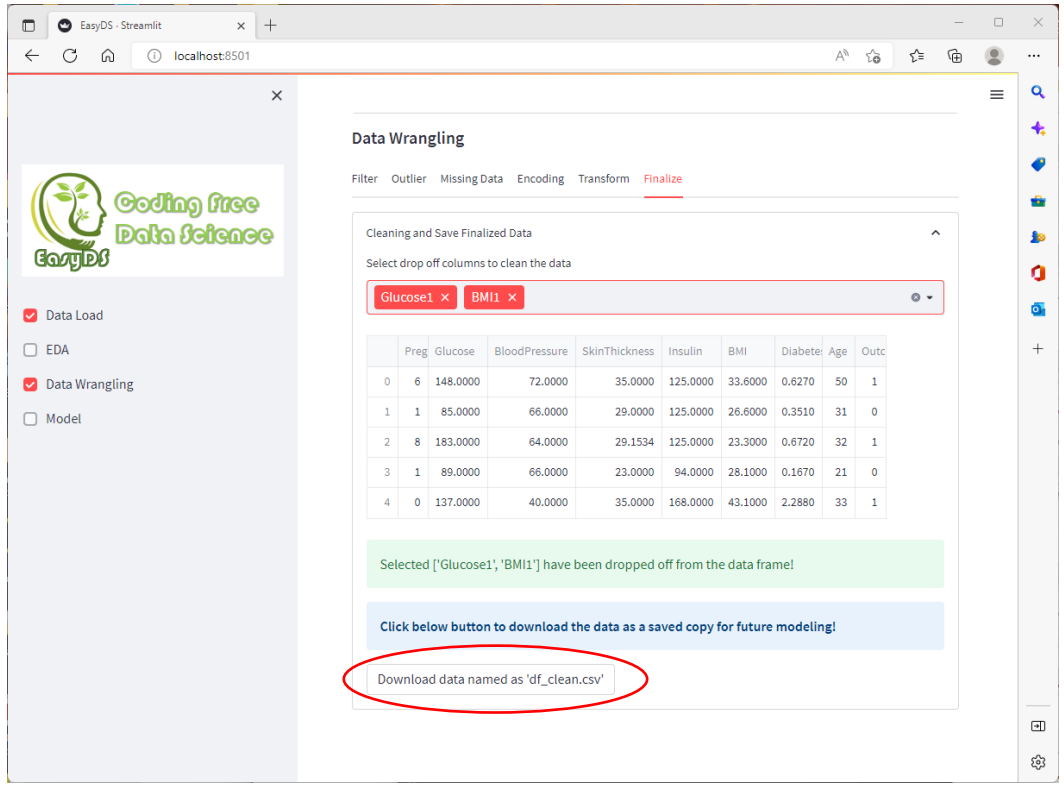



To get better model, sometimes it needs transform data to approximate normal distribution. Default method is ‘Power Transform’ as displayed is left side plot, other methods can be selected such as ‘Quantile’ in right side cases.

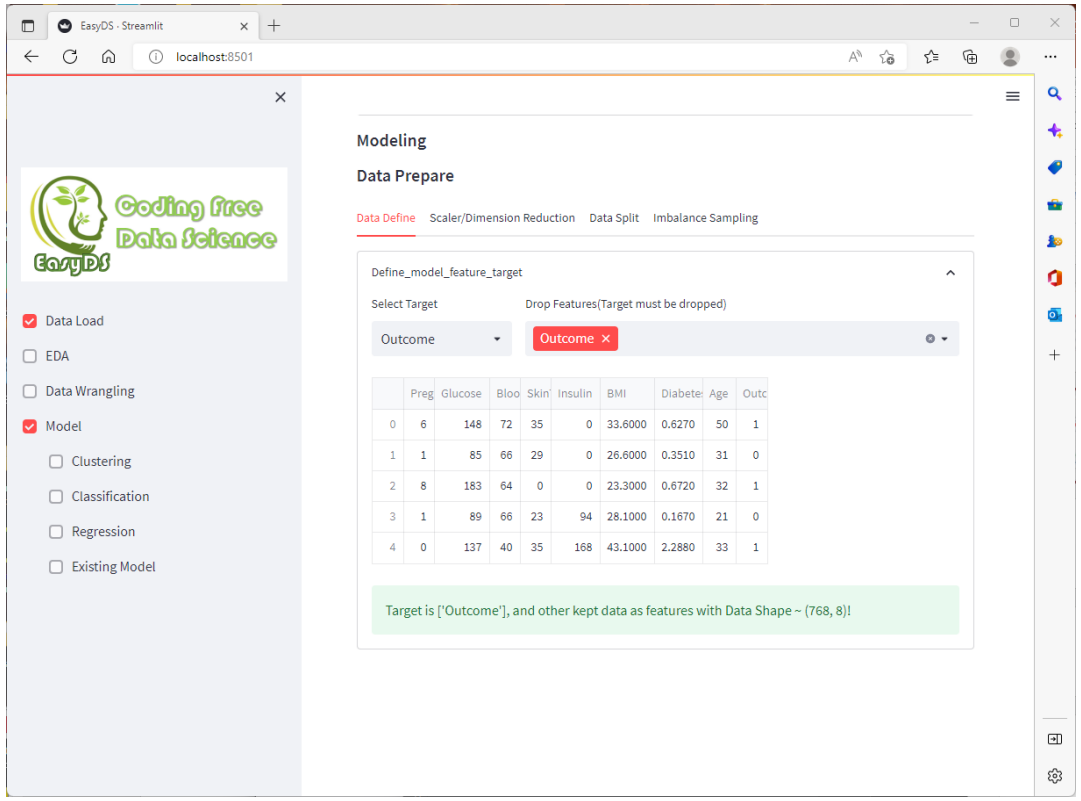
# Data Wrangling – Clean & Finalize Data



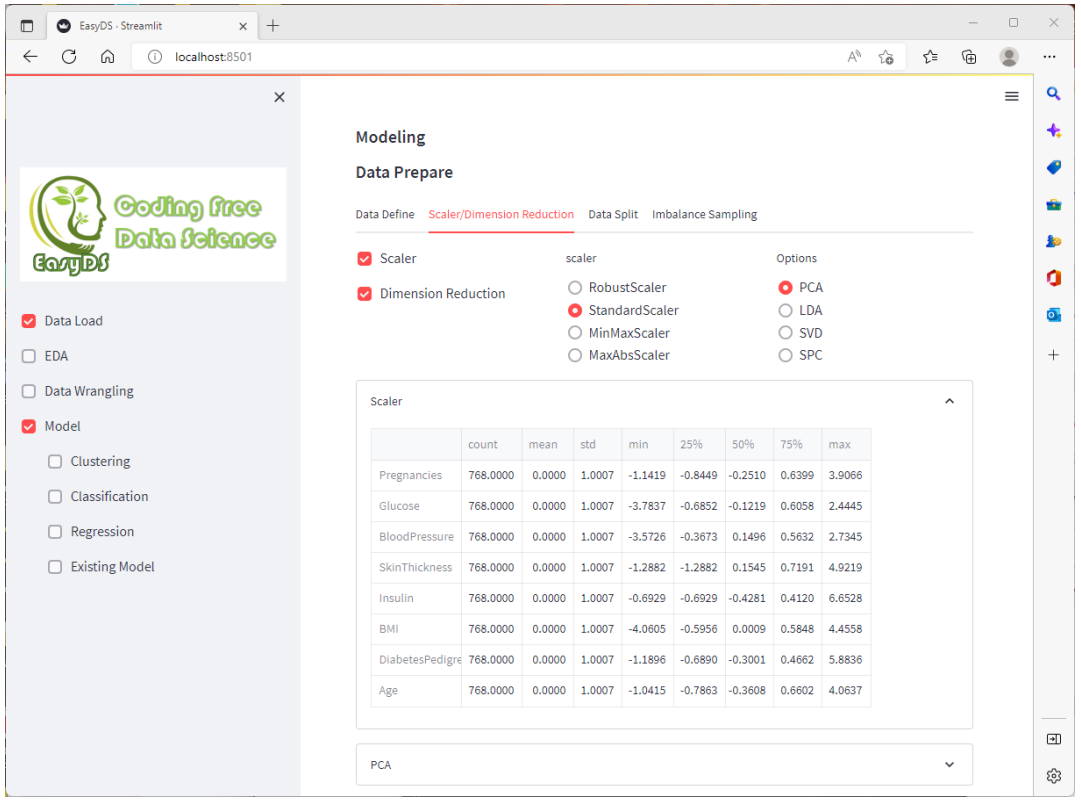




After data wrangling, a clean data could be saved for modeling. Some repeated data need to be dropped off, for example in this case there are raw and scaled ‘Glucose’ and ‘BMI’ in data set. To save the data to local disk, click the download data button, it will save the data  to local download folder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

# Data Preparing for Modeling

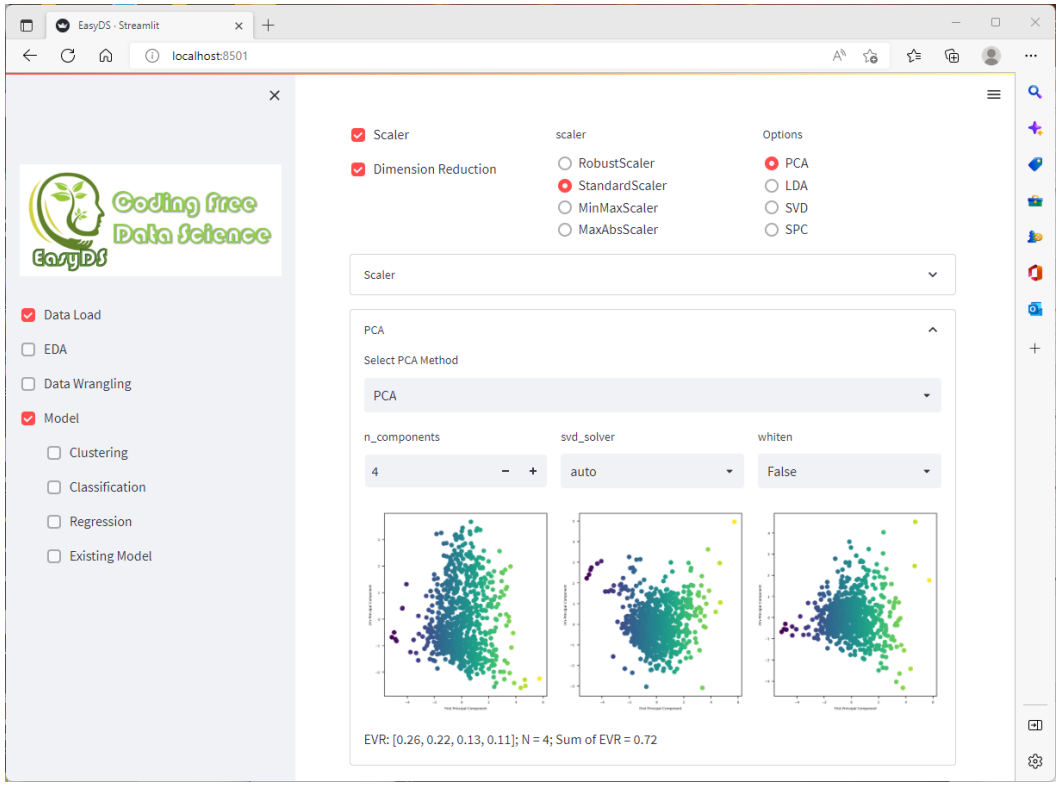



Define target data and features

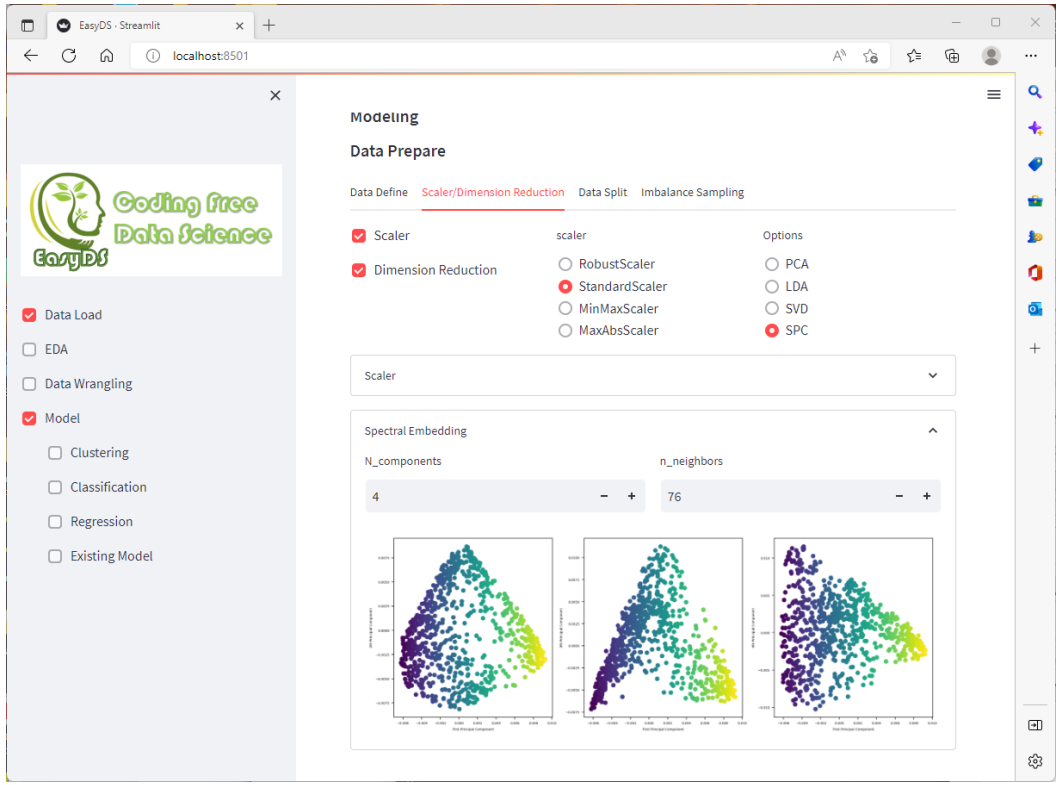



Scaler or reduce dimensions

# Data Preparing for Modeling - Reduce dimensions

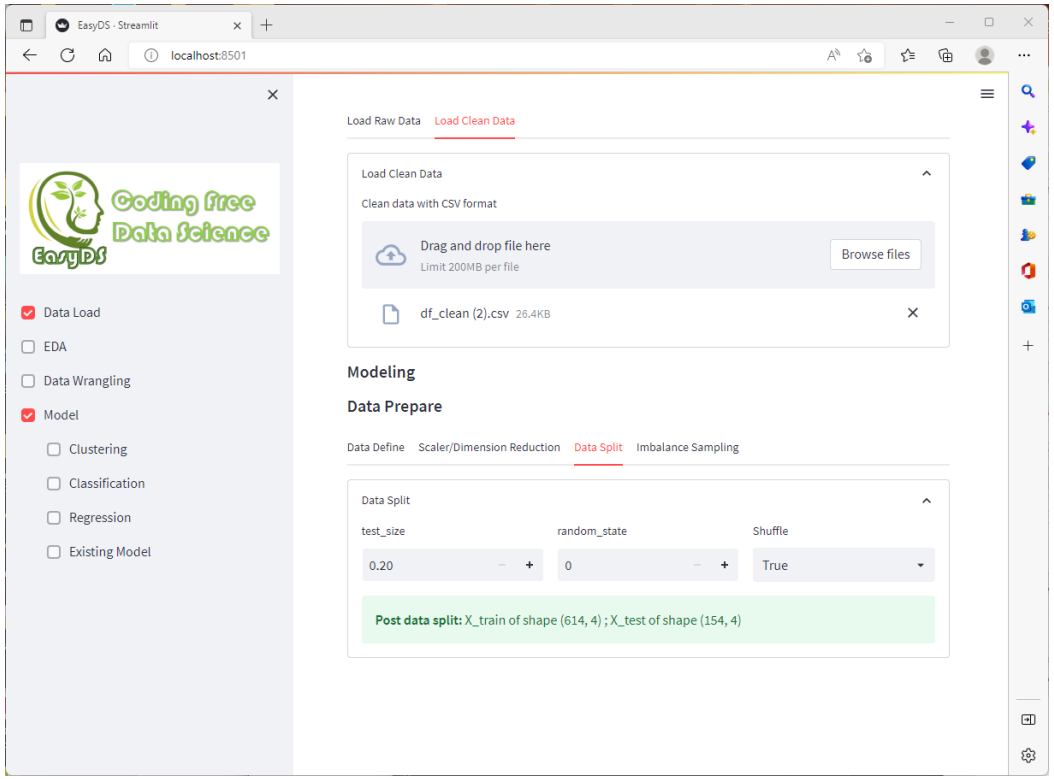



Example of PCA

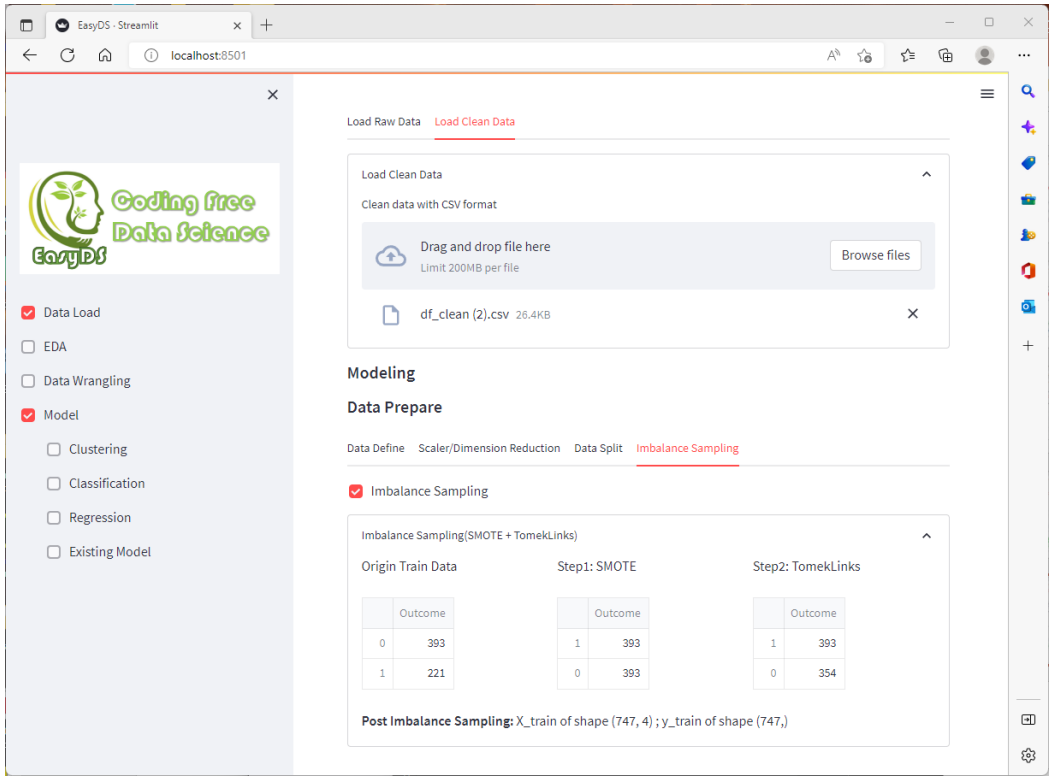



Example of SPC

# Data preparing for modeling

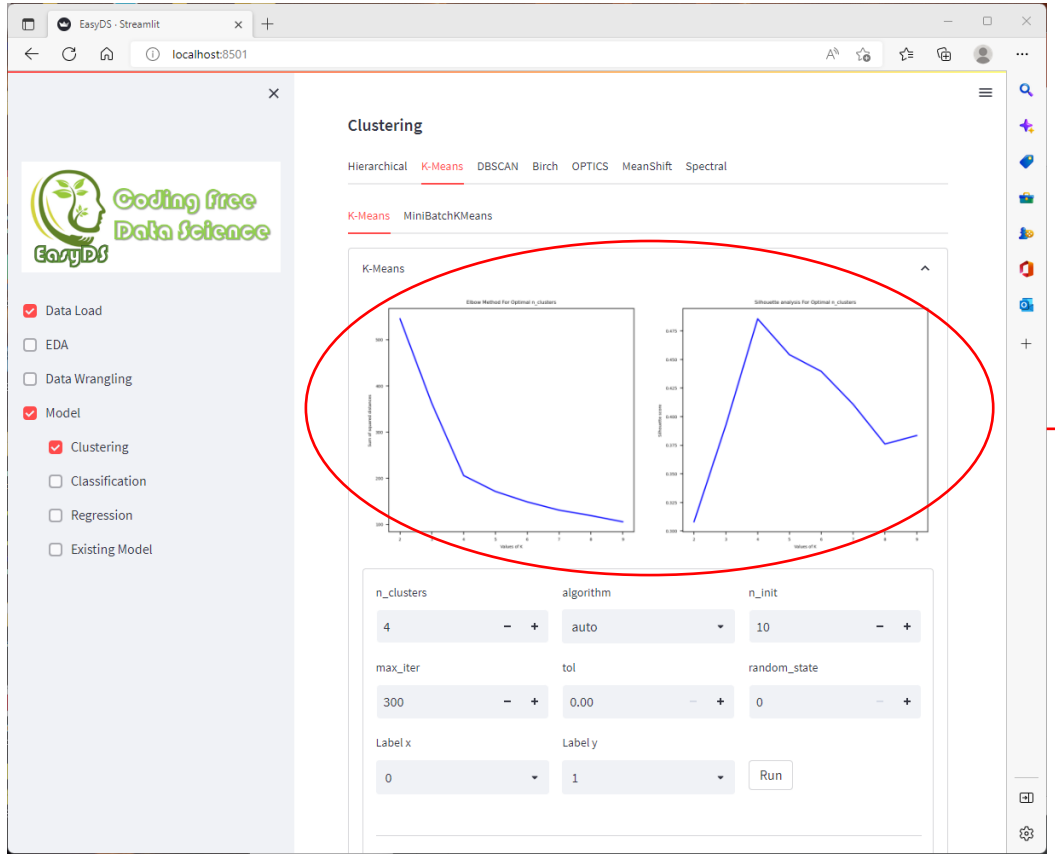


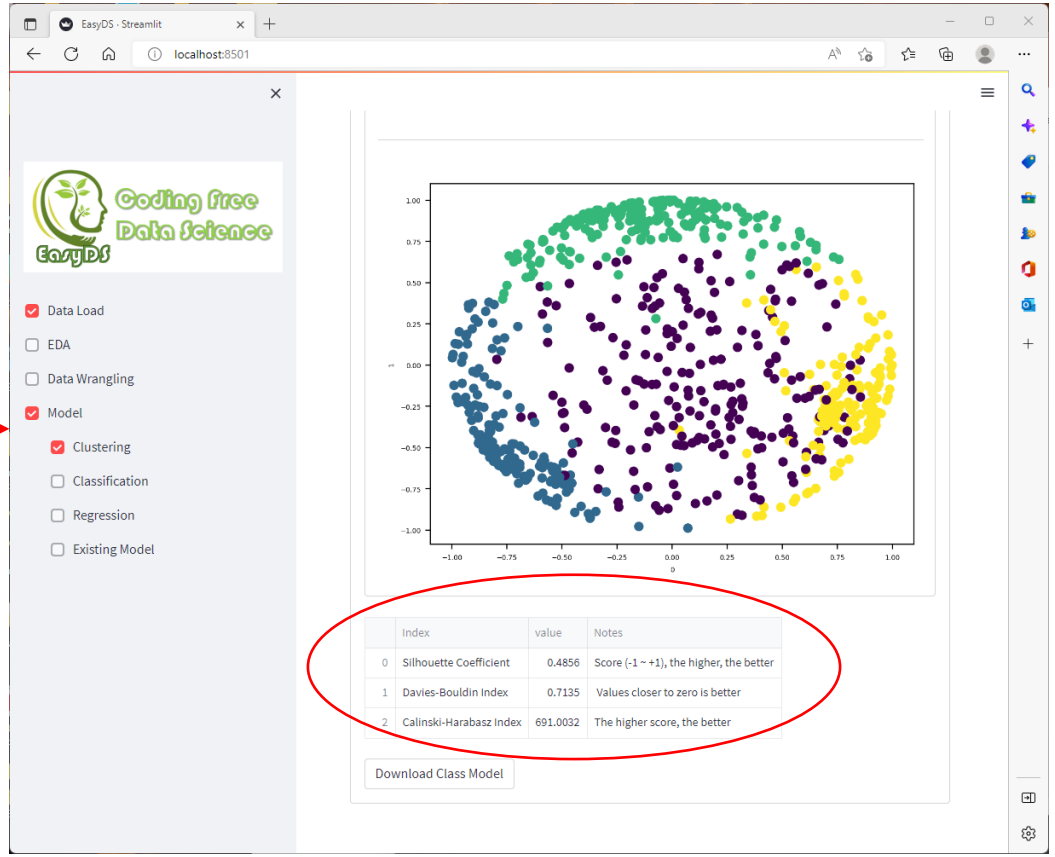

Data Split




Optional Imbalance sampling(SMOTE + TomekLinks)

# Clustering

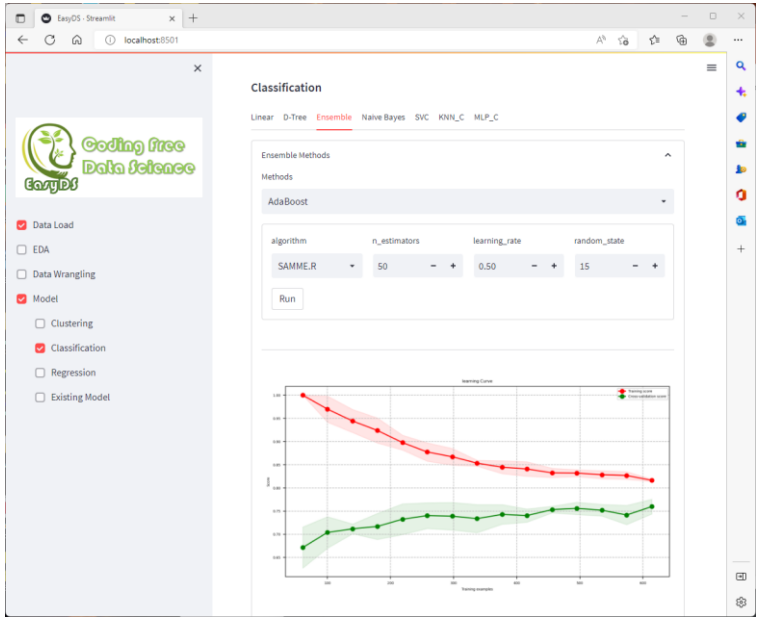


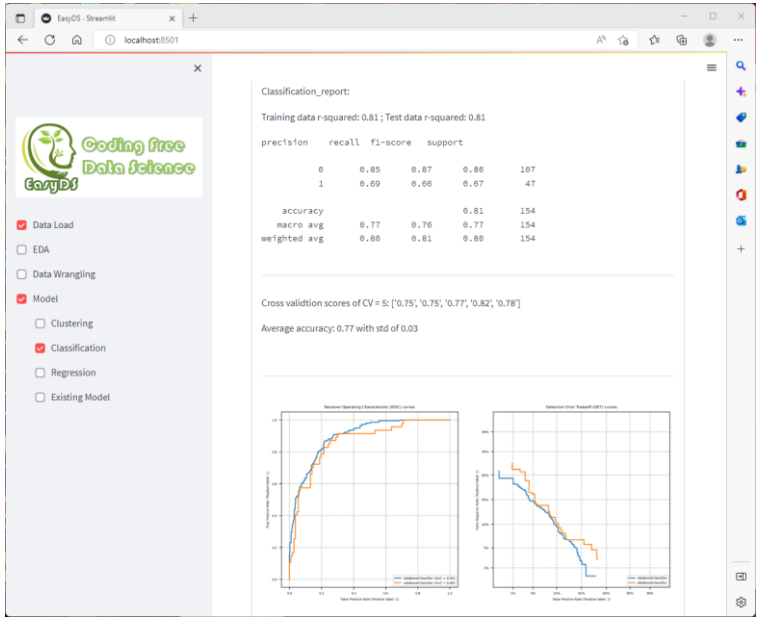



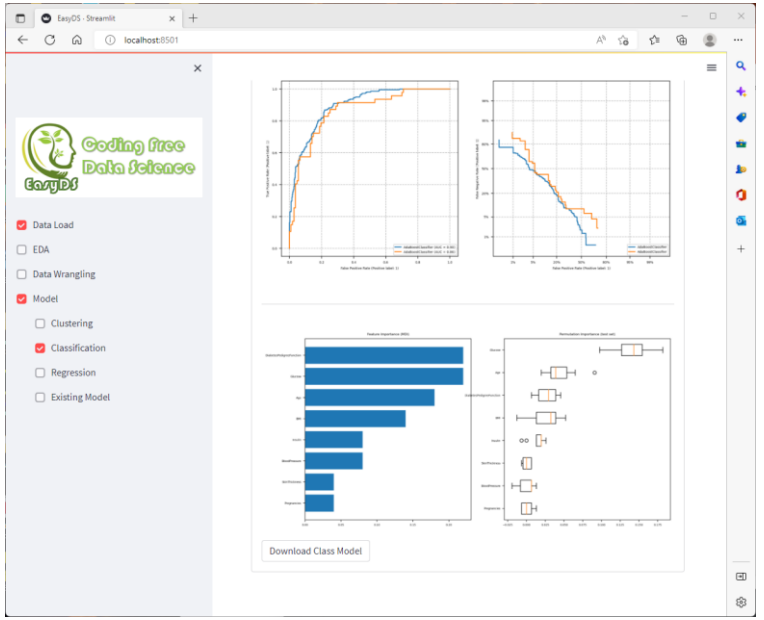

Use different methods as aid for to define n_clusters in K-Means

Use different index to evaluate model

# Classification

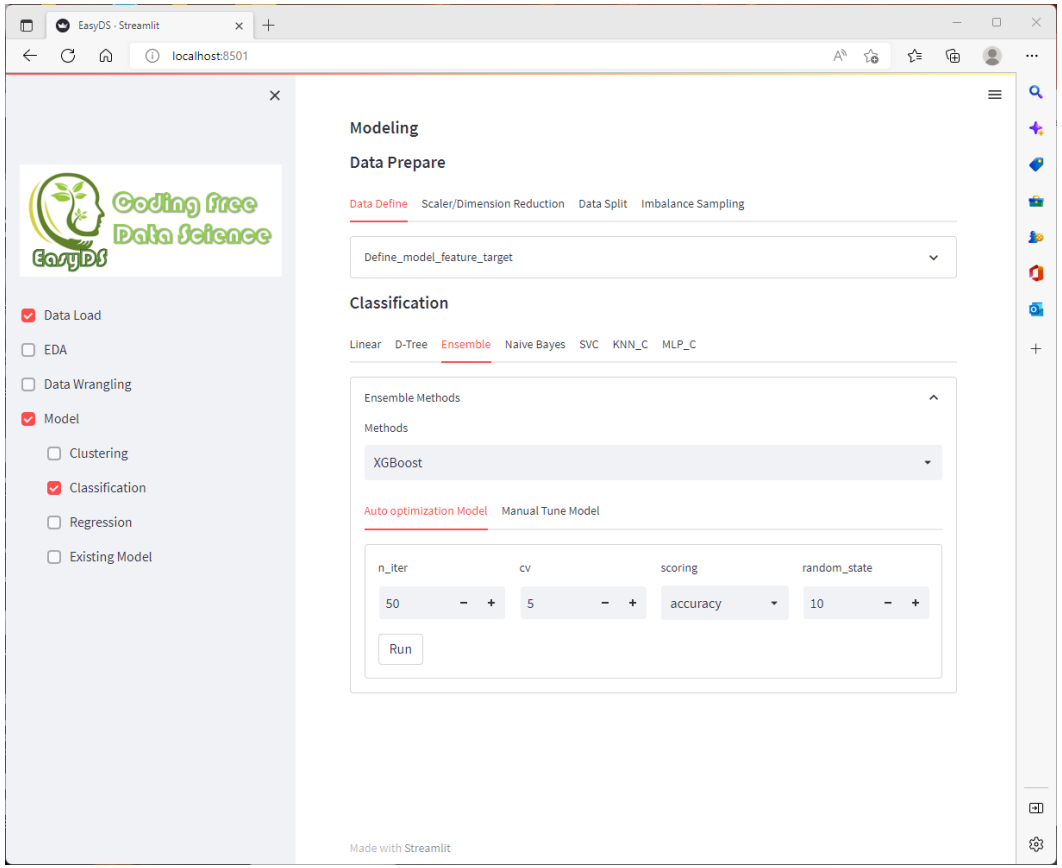


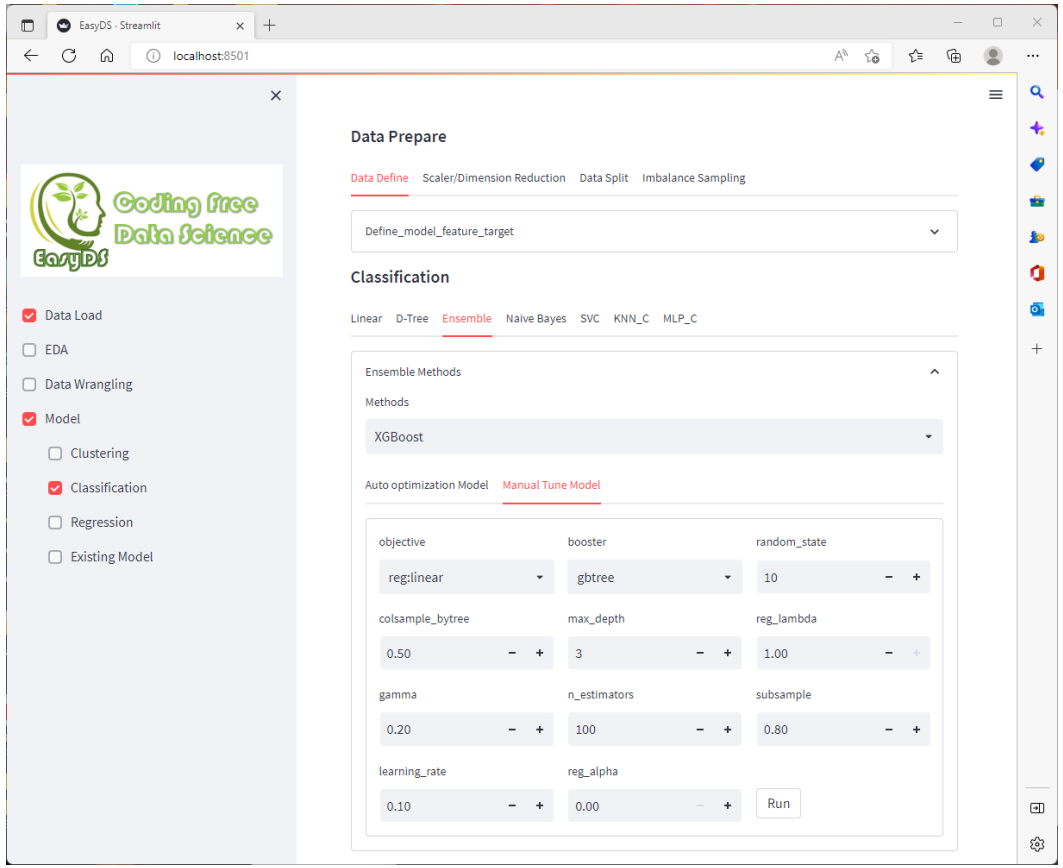




Wrapped most popular Classification methods, and multiple ways for parameter tuning, model cross-checking and validation

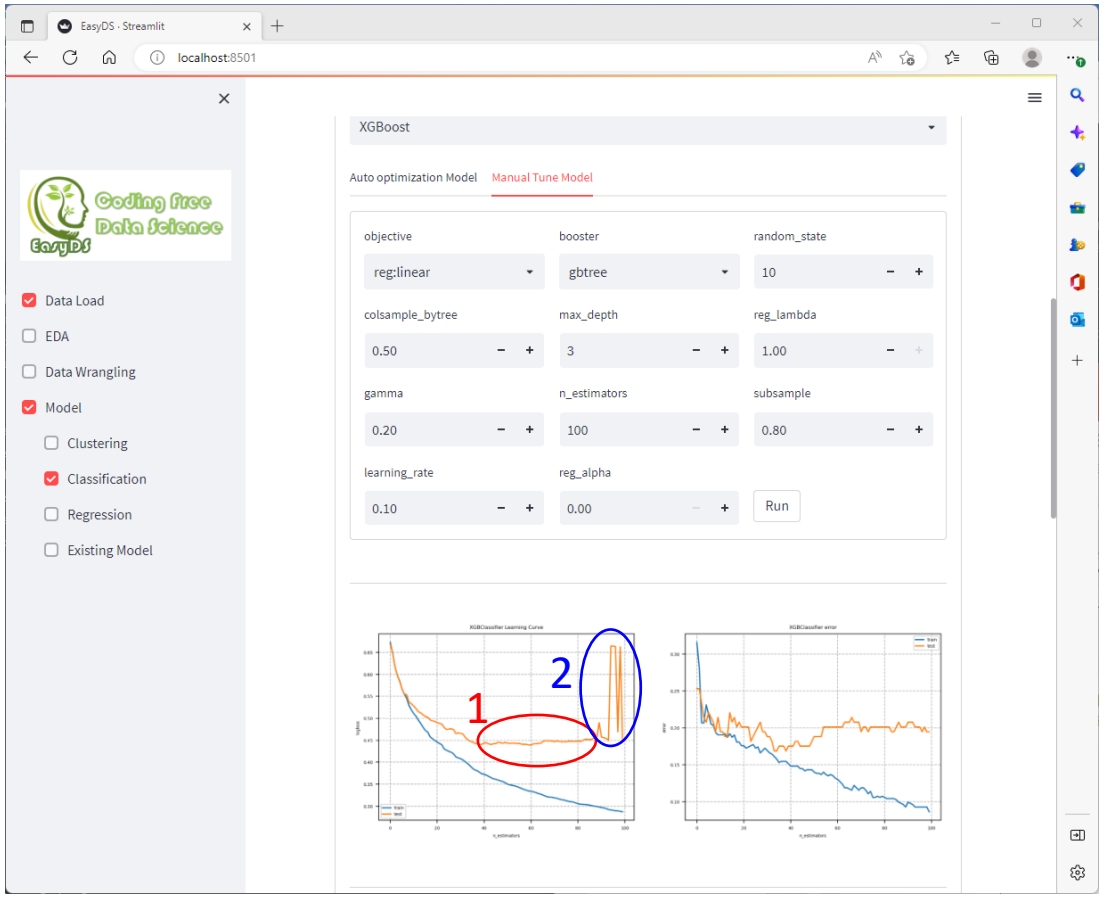
# Classification – example of XGBoost

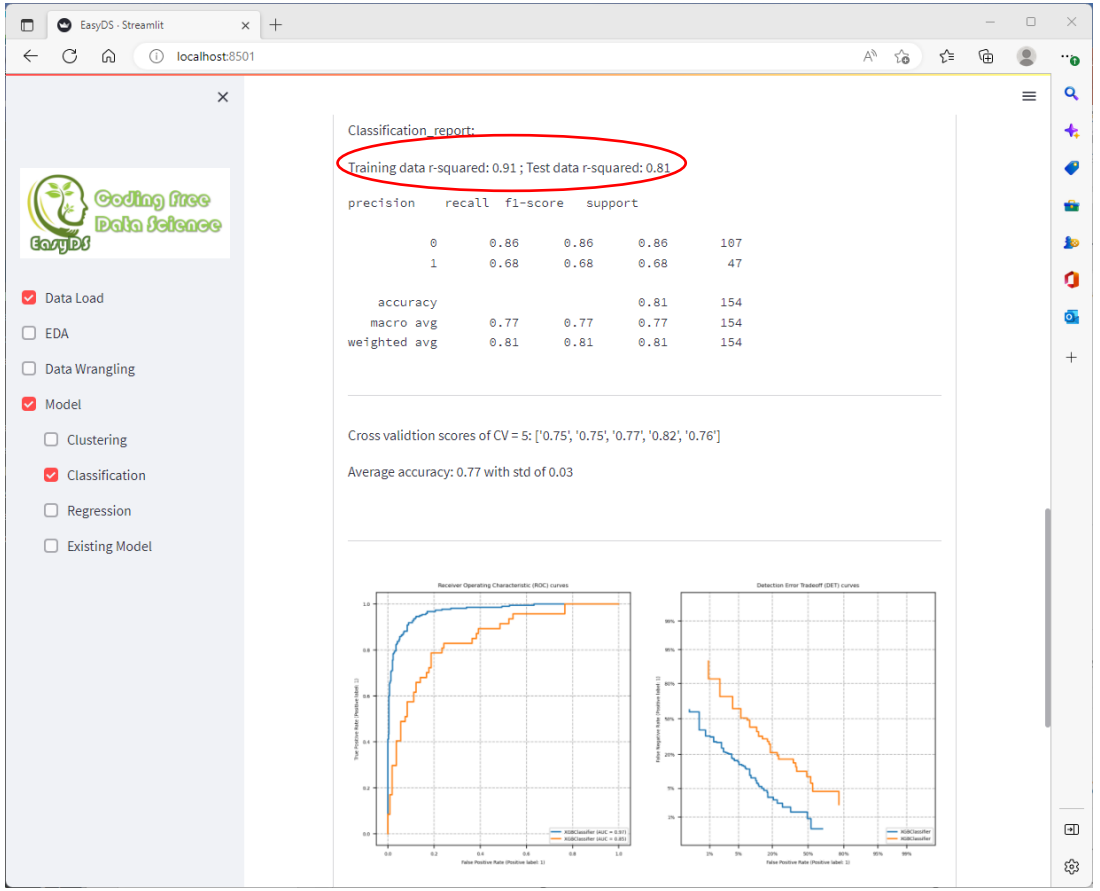



Some of the methods have both Auto Optimization Model and Manual Tuning Model. Since Auto Optimization Model is still in testing (random search on discrete parameters to save compute times), the result is not really optimized. Therefore, Manual Tuning Model is suggested for the modeling.

# Classification – example of XGBoost



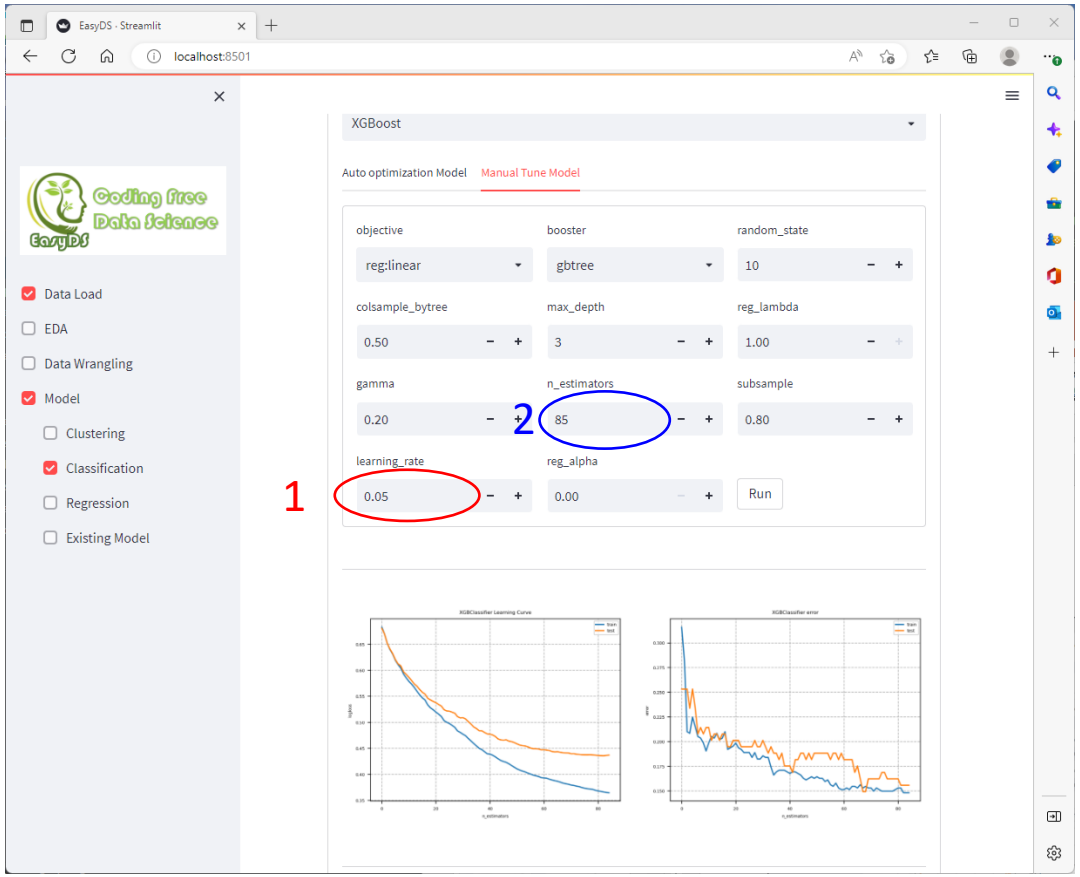



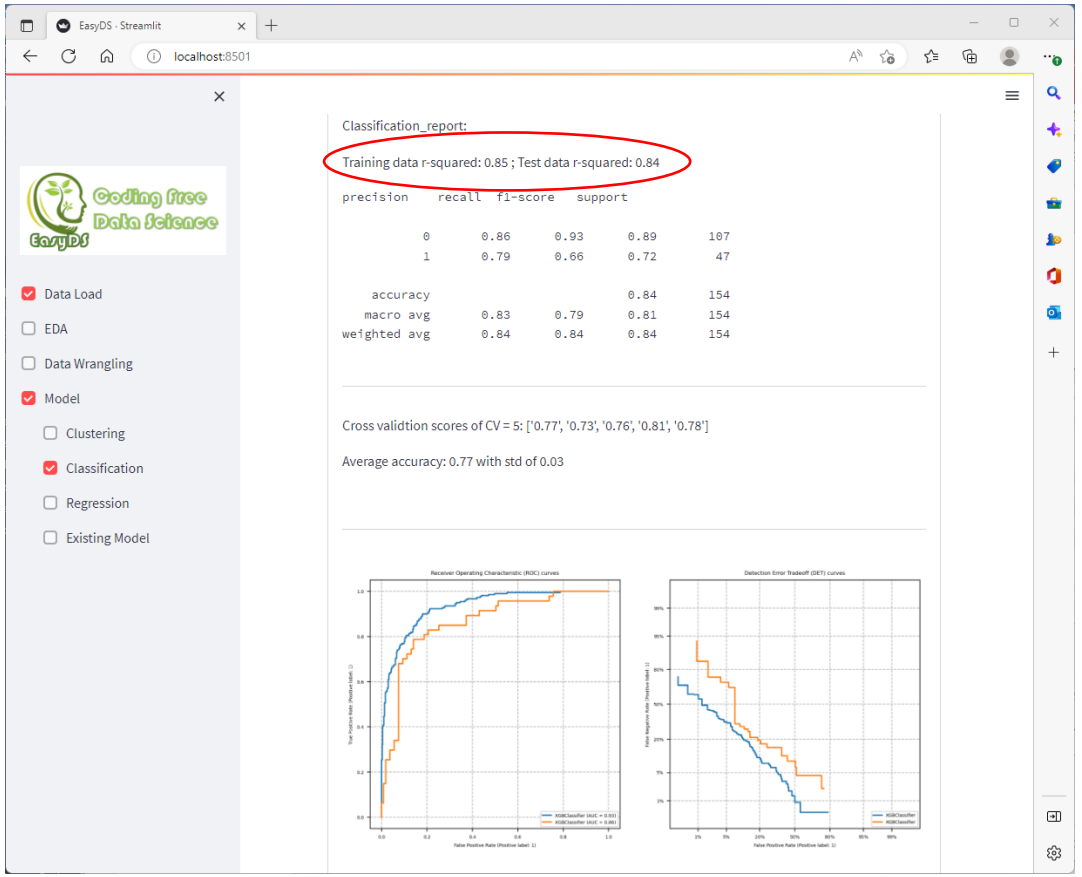

Learning Curves assist for hypo-parameter tuning, it can be seen :

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



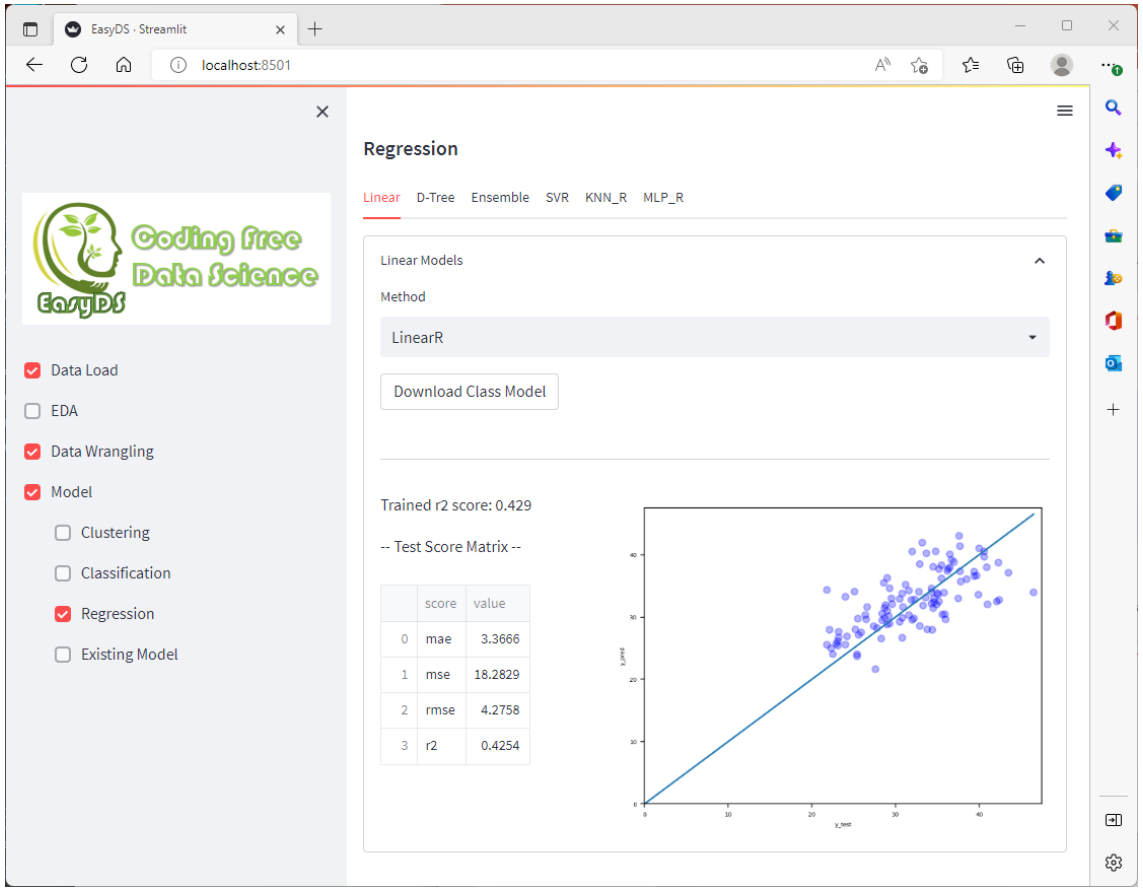



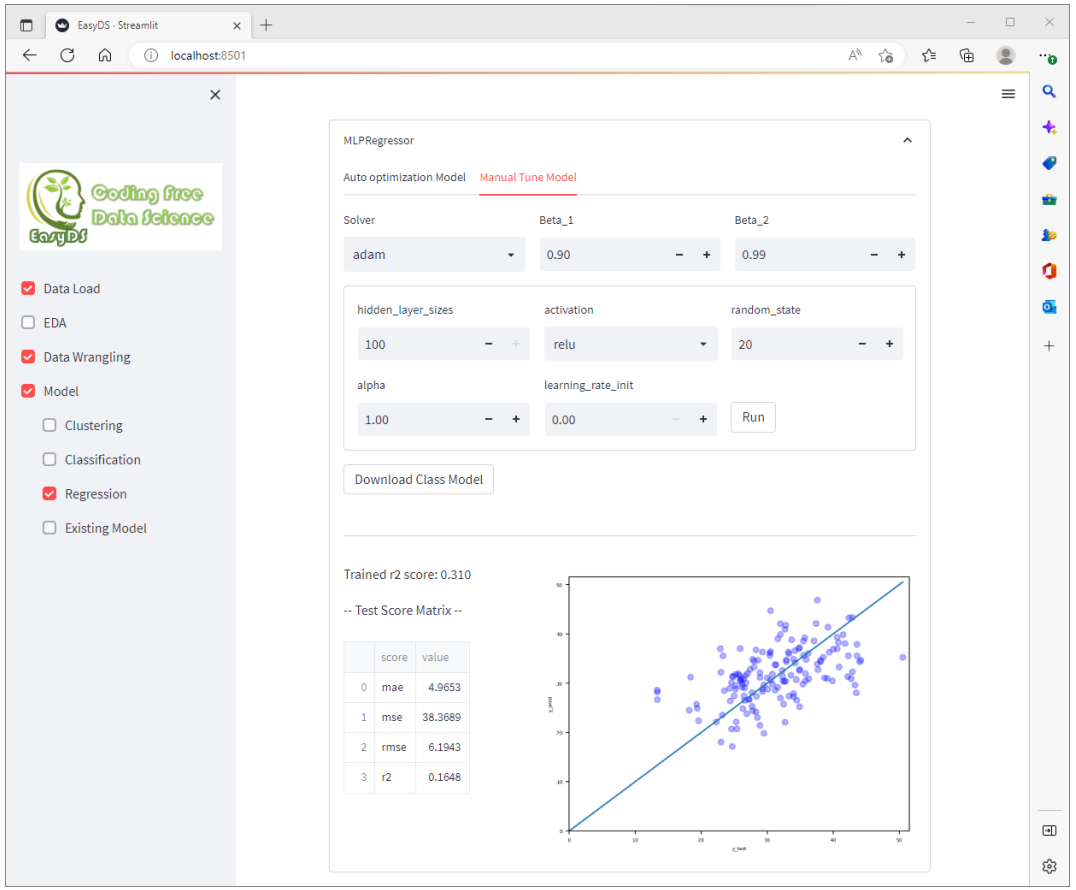

To improve the model referred to previous learning curve:

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

→ Model gets reasonable.

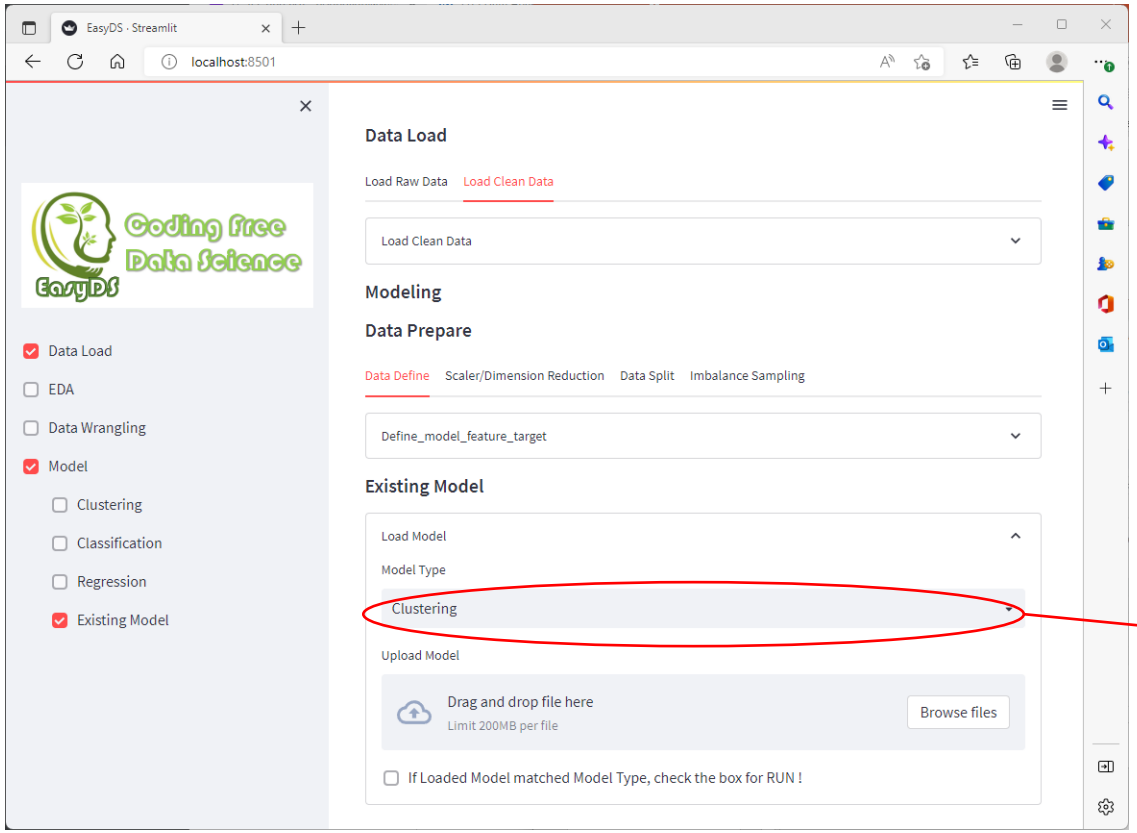
# Regression

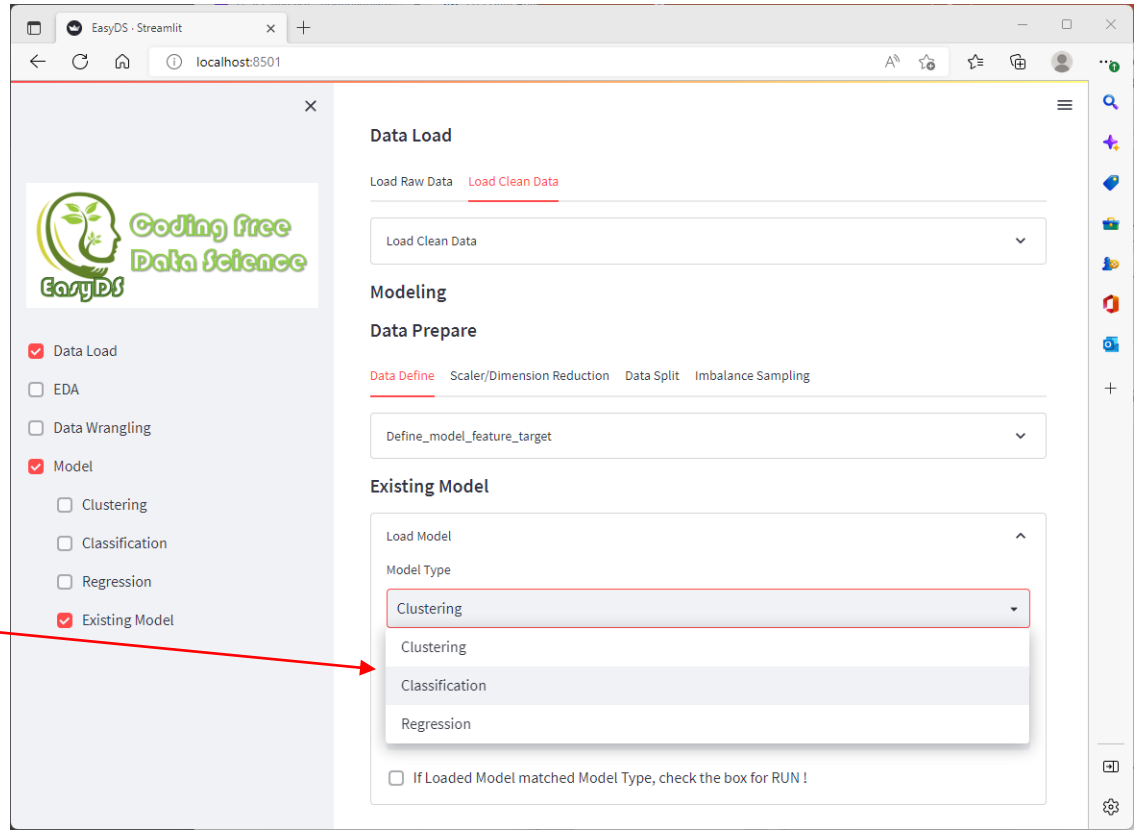



Wrapped most popular Regression methods, and multiple scores are list  for model checking and validation

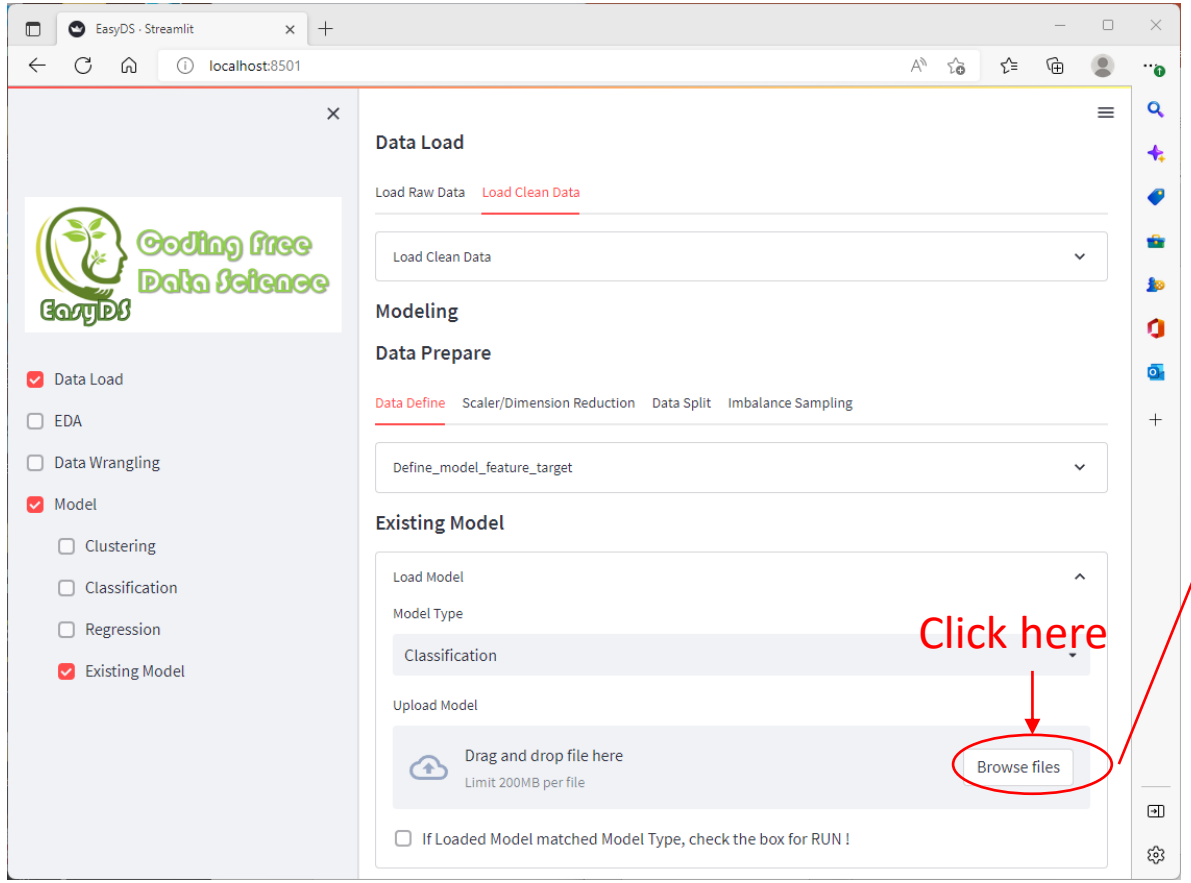
# Load Existed Model

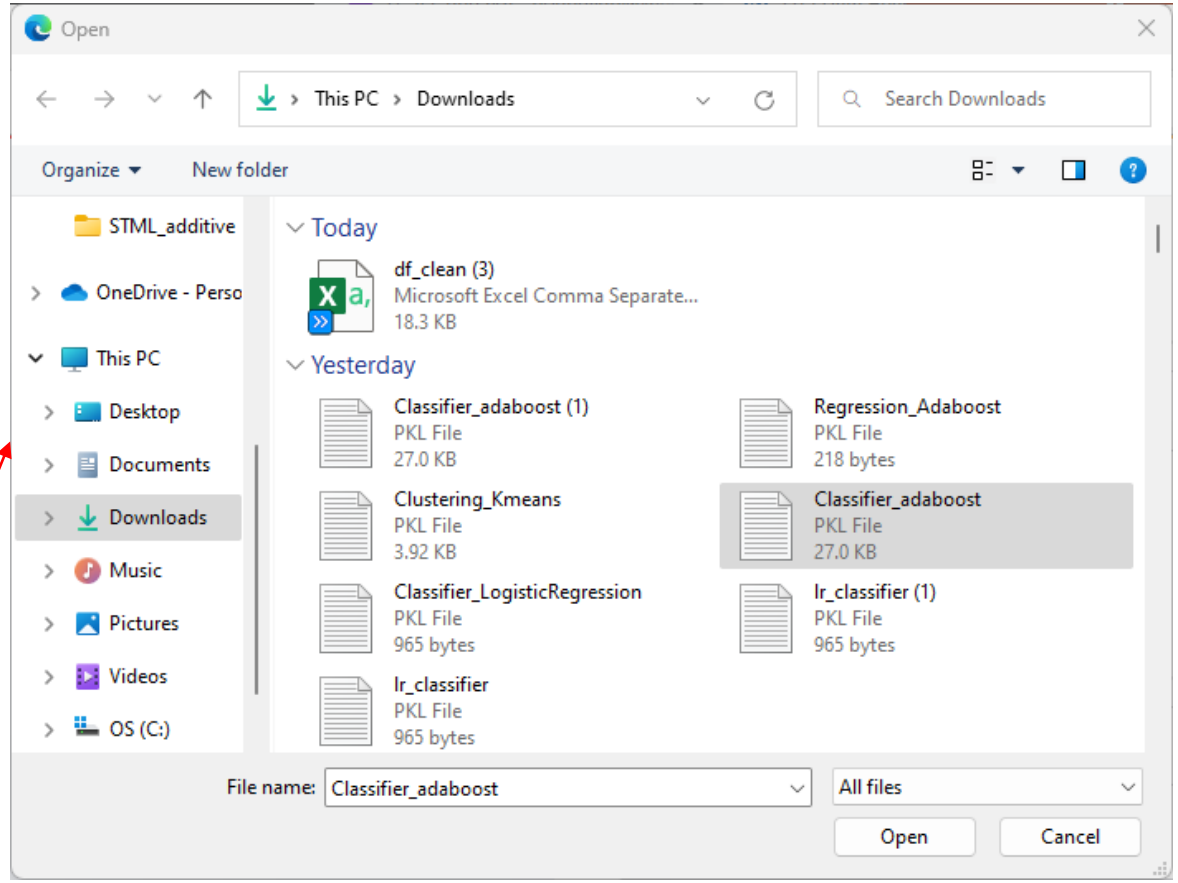


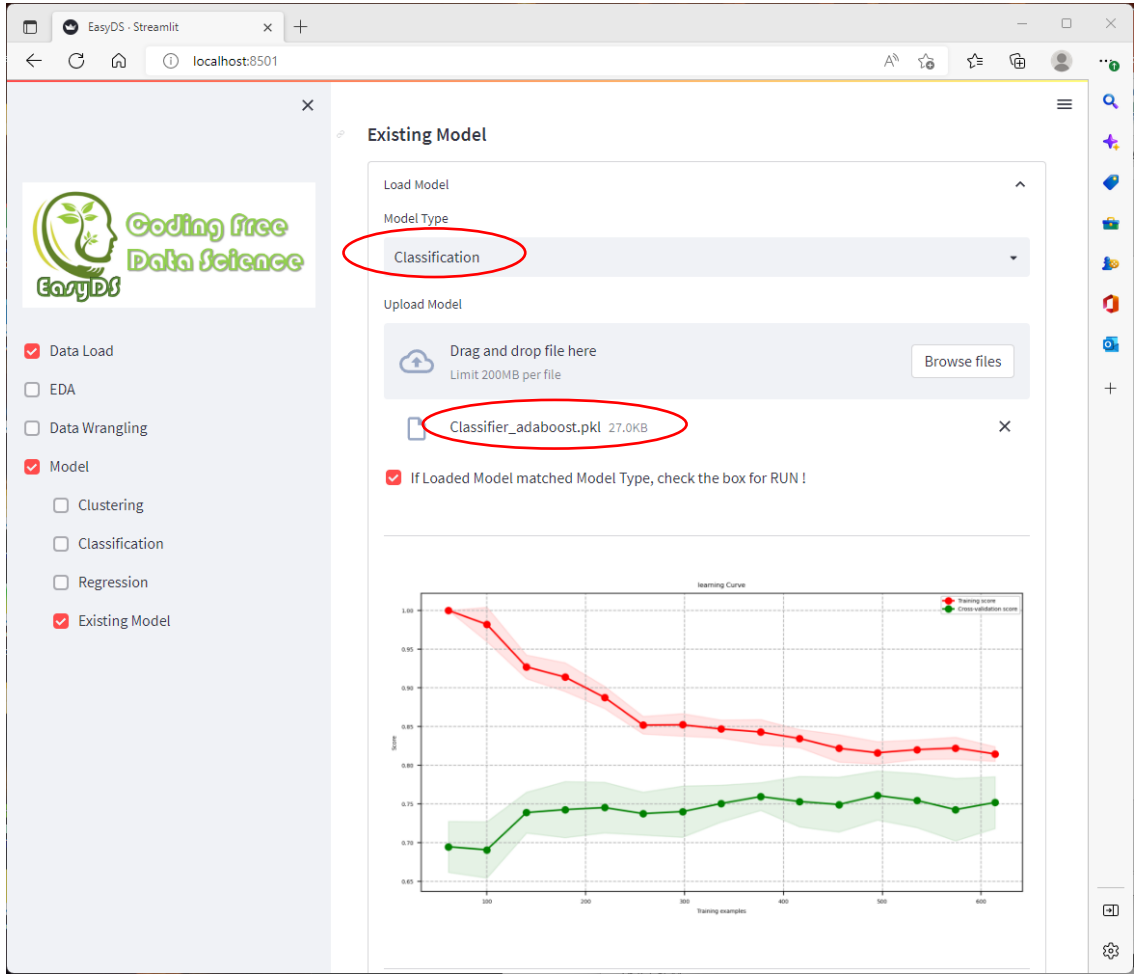

Select the model type to match the model you loaded, which is to make sure your data set is suitable for the model

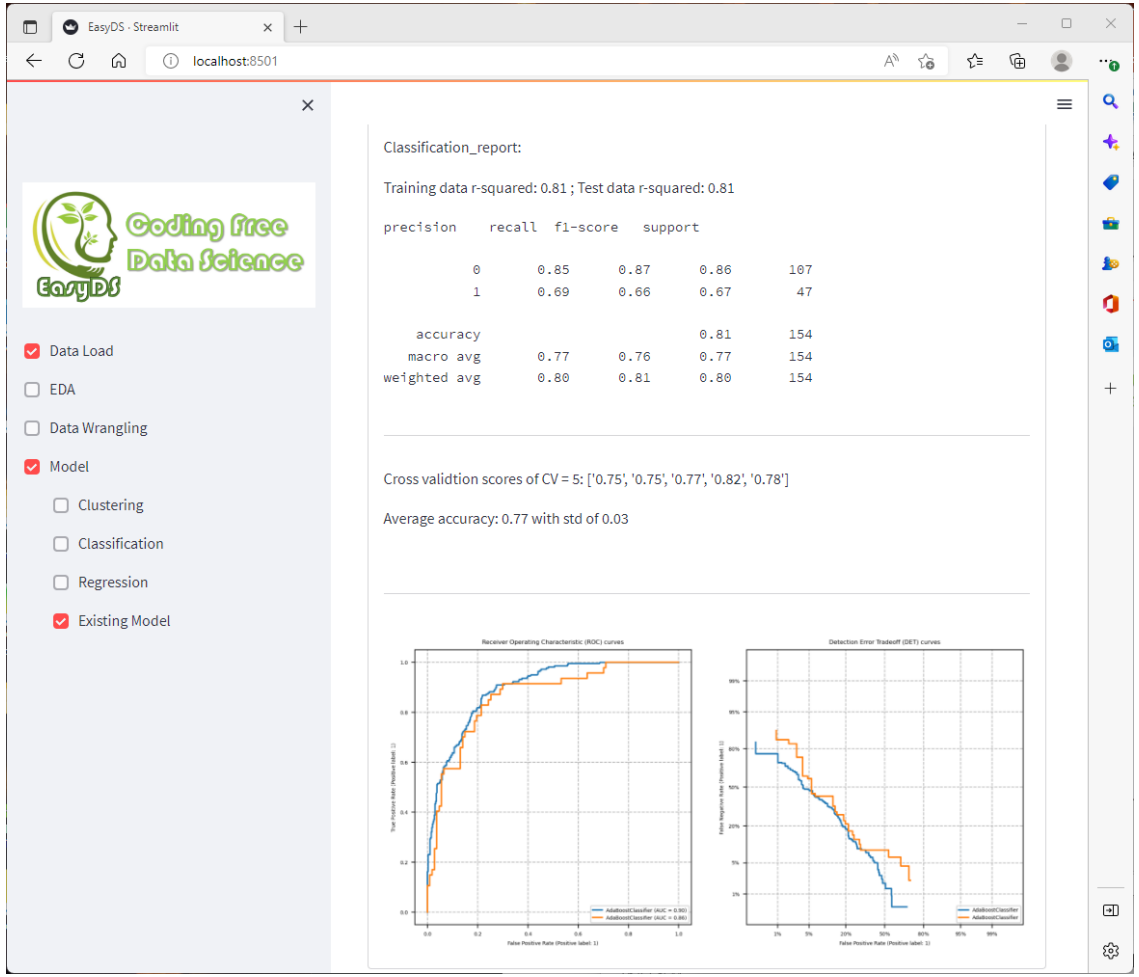
# Load Existed Model



# Load Existed Model




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

Coding free Data Science

EasyDS

# Welcome for testing and give your feed back!

hhou218@gmail.com