SCIKITOOL USER MANUAL

Your Best ML Learning Friend

Group40

Chengxuan Li, Chenwei Yin, Gubin Zhao, Hailin Xie, Haoran Fang, Renyuan Lu

Contents

1. *Introduction* 2

   1) *Basic User* 2

   2) *Advanced User* 3

2. *Installation instruction* 4

   1) *System requirements* 4

   2) *Installation Process* 5

   3) *Pre-installation preparation* 5

   4) *Install steps* 5

   5) *Precautions* 6

3. *Instructions of users* 6

4. Features 12

   1) *My Data* 12
      - i. *Upload Data Set* 12
      - ii. *Data Details* 13

   2) *Data Processing* 13
      - i. *Data Pre-processing* 13
      - ii. *Outlier Handling* 13
      - iii. *Dimension Reduction* 14

   3) *Descriptive Analysis* 14
      - i. *Regression* 14
      - ii. *Classifier* 14

   4) *Export Results* 15
      - i. *Components of results* 15

5. Frequently Asked Questions (FAQs) 17

6. Technical support 18

   1) *Documentation* 18

   2) *Github community Forum* 19

1. Introduction

This user manual provides step-by-step instructions for using the Offline Data Analysis Software, a comprehensive tool designed to simplify data processing and analysis tasks for both basic and advanced users. The software offers a user-friendly interface, powerful data processing tools, and customizable data analysis capabilities, empowering users to efficiently process and analyze their data without any programming knowledge.

1) Basic User

The primary users are individuals with limited experience in data analysis who require an intuitive and user-friendly solution to process and analyze their data. The software caters to the needs of these users by offering the following features:

a) ***My Data:*** With a user-friendly interface, even basic users can effortlessly upload, select, and inspect their datasets. The intuitive design enables users to manage datasets, variables, and labels seamlessly, without requiring any programming expertise：

- Uploading Data Set: Users can effortlessly upload their data sets in CSV
- Selecting Data Set: Users can choose a data set to upload and view the one they wish to work with.
- Data Summary: Users can access an overview of their selected data set, including the number of rows and columns, as well as any missing values.
- Variable Management: Users could rename columns, modify data types, and preview a sample of their data to ensure it is correctly formatted for analysis.
- Label Management: Users can manage labels associated with their data sets, such as categorical variable labels, enhancing the readability of their analysis results.

b) ***Data Processing****:* The software equips basic users with user-friendly yet powerful data processing tools, such as handling missing values, data standardization, and dimension reduction. These tools enhance data quality and prepare it for analysis.

- Missing Value Processing: Users can choose from a variety of methods to handle missing values, including deletion, mean or median imputation, or using a custom value.

- Data Standardization: Basic users can normalise their data using widely used techniques such as min-max normalisation or z-score normalisation, which can improve the performance and accuracy of their analysis.
- Dimension Reduction: Users can apply dimensionality reduction techniques, such as PCA or LDA, to simplify their dataset and concentrate on the most crucial features.

c) ***Data Analysis**: Basic users can perform an array of analyses, including descriptive analysis, statistical modeling, and machine learning classification. The software guides users in selecting suitable algorithms and visualizes the results in an easily comprehensible format.*

Correction — c) ***Data Analysis***: Basic users can perform an array of analyses, including descriptive analysis, statistical modeling, and machine learning classification. The software guides users in selecting suitable algorithms and visualizes the results in an easily comprehensible format.

- Descriptive Analysis: Basic users can generate summary statistics, frequency tables, and cross-tabulations to gain a high-level understanding of their data.
- Statistical Modeling: Users can apply statistical models like linear regression or ANOVA to test hypotheses and identify relationships within their data.
- Machine Learning Classification: The software guides users through selecting and applying classification algorithms, such as decision trees, random forests, or KNN, to predict categorical outcomes.

d) ***Generate Report:*** Once the analysis is complete, basic users can produce a comprehensive report that synthesizes their findings, which can be used for further interpretation and decision-making.

2) Advanced User

Advanced users, who possess a strong background in data analysis, often require more sophisticated tools and customization options to address their unique needs. Our software caters to these individuals by offering the following enhanced features:

a) Advanced Data Processing:

In addition to the basic data processing features, advanced users can fine-tune the data processing parameters, such as selecting specific normalization methods or customizing dimension reduction techniques.

- ***Custom Normalization:*** Advanced users can implement their own normalization formulas or modify the software's built-in methods for a more tailored data preprocessing experience.

• Dimension Reduction Customization: Users can fine-tune dimensionality reduction techniques by adjusting parameters, such as the number of principal components in PCA or the number of target dimensions in LDA.

b) Customizable Data Analysis:

Advanced users could use more advanced statistical modelling and machine learning techniques, as well as customising settings to suit their analytical requirements. This allows them to dig deeper into the data and discover more complex patterns and relationships.

• ***Advanced Statistical Modeling:*** Advanced users can employ more complex statistical models, such as logistic regression, mixed-effects models, or time series analysis, to address specific research questions.

• Custom Machine Learning Algorithms: Users can integrate their own machine learning algorithms or leverage advanced techniques available in the software, such as neural networks, ensemble methods, or deep learning.

c) Data Visualization:

Advanced users can use the software's data visualisation tools to create interactive charts, graphs and heatmaps, enabling them to better understand and communicate their findings.

2. Installation instruction

1) System requirements

To have better experiences, before installing the software, please ensure your computer meets the following minimum system requirements:

Operating System: Windows 10, macOS X+

Processor: Intel Core i5 or equivalent

Memory: 8GB

Display: 1280*720 resolution or higher

Internet connection for installation and updates(optional)

2) Installation Process

The following sections will guide you through the pre-installation preparation and the installation steps required to set up SciKiTools on your computer.

3) Pre-installation preparation

Before you begin the installation process, make sure to meet the following requirements:

a) Ensure your computer meets the minimum system requirements to run the software, such as the necessary hardware specifications, operating system, and available disk space.

b) Check if you have the required software dependencies installed on your computer, such as Python, required Python libraries and any other relevant software.

c) Download the latest version of SciKiTools from the official GitHub repository website.

d) Unzip the downloaded file to a suitable location on your computer.

4) Install steps

Follow these steps to install SciKiTools

a) Open the official GitHub repository website:

https://github.com/COMP208TEAM40/SciKiTool_v0.1.6

b) Download the latest application version

c) Open the file and run QT_Design.py

d) Congratulations! You have successfully installed SciKiTools!

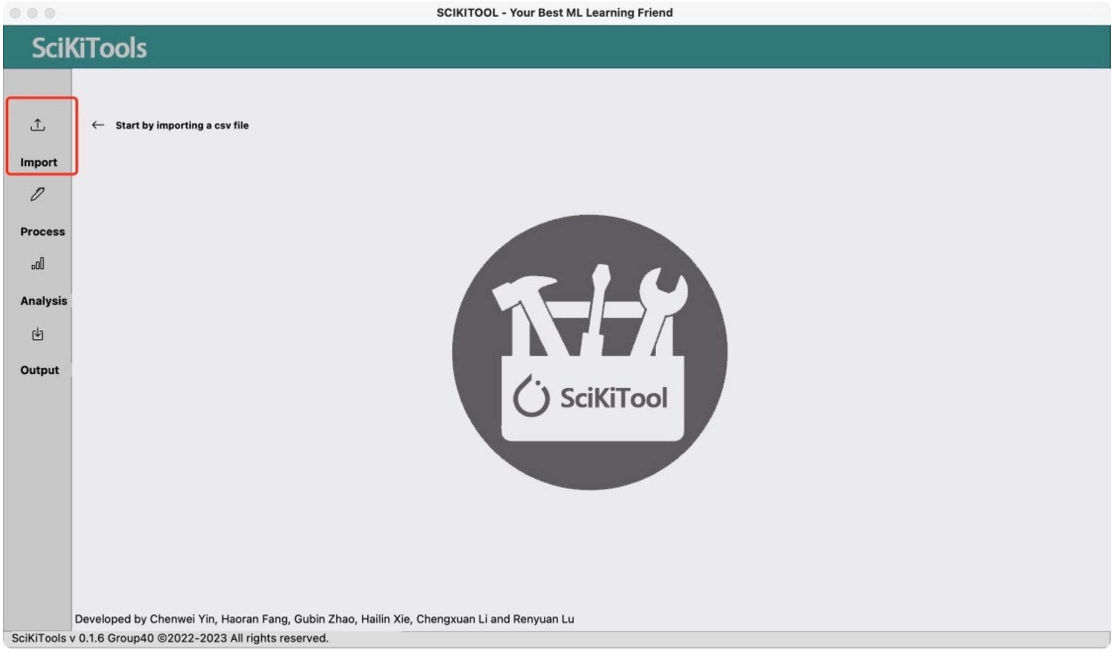
By following these pre-installation preparation steps and installation instructions, you can easily set up SciKiTools on your computer and start taking advantage of its powerful data processing and analysis capabilities.

5) Precautions

Please check that the system version meets the minimum requirements for the software to run. When downloading software, please take care to select the appropriate version. Usually, the latest stable version is the first choice, but in some cases you may need to use an older version. You can find different versions on the "Releases" page in GitHub (this only applies after the second iteration has been released).

3. Instructions of users

Step1: When you open the program, the main page of the program will appear as follows, click on "Import" to start uploading csv files.



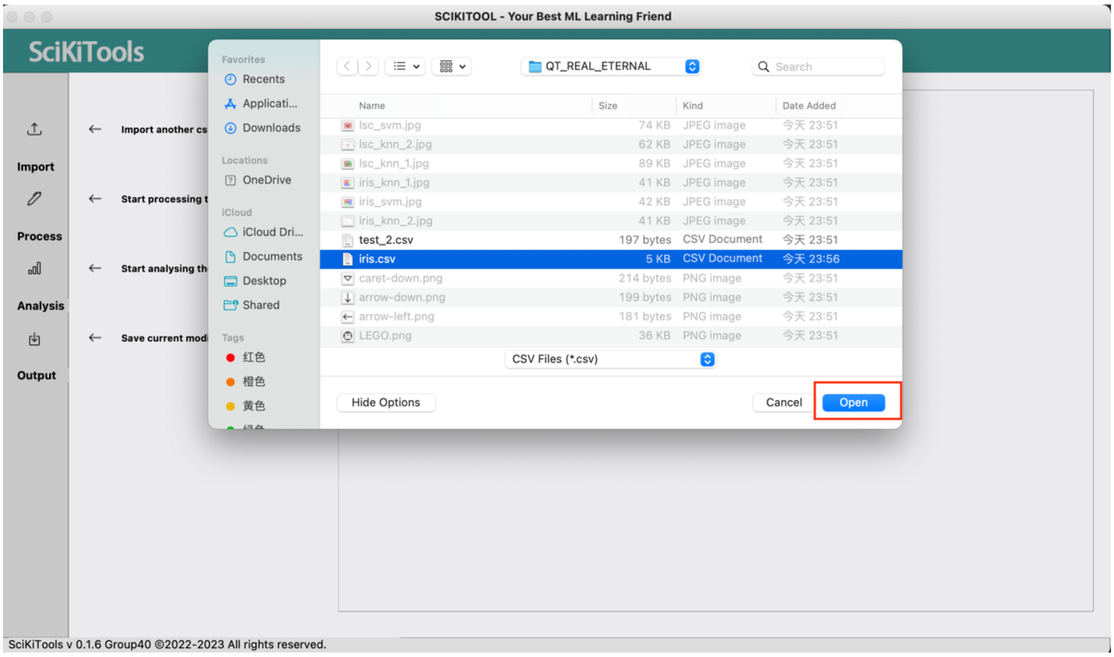
SCIKITOOL - Your Best ML Learning Friend

← Start by importing a csv file

Developed by Chenwei Yin, Haoran Fang, Gubin Zhao, Hailin Xie, Chengxuan Li and Renyuan Lu

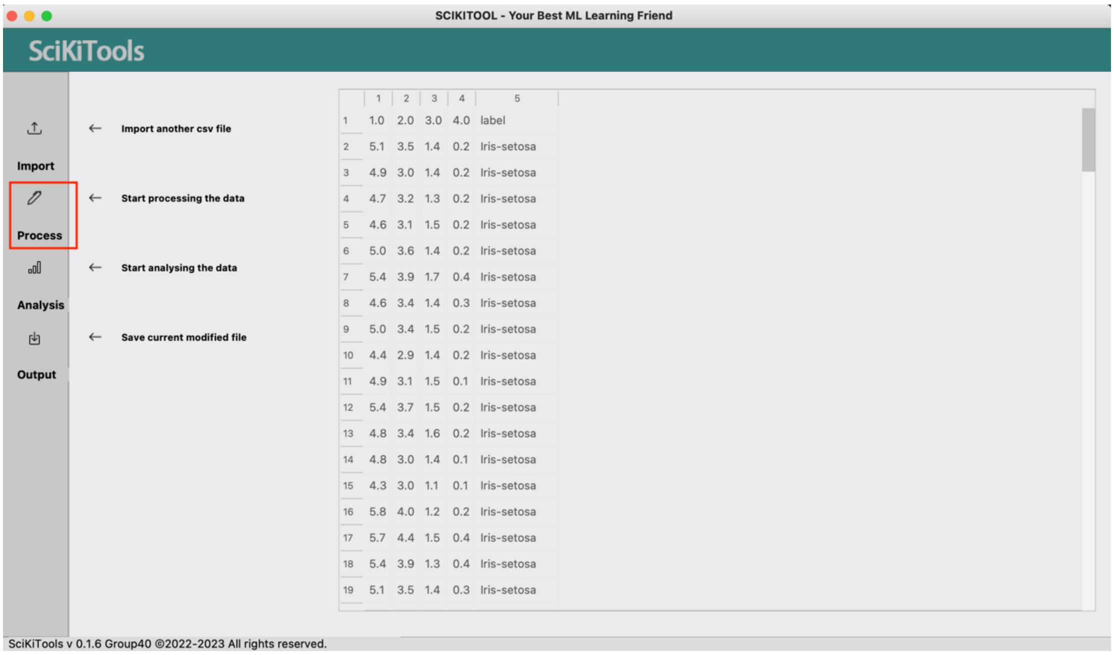
SciKiTools v 0.1.6 Group40 ©2022-2023 All rights reserved.

Step2: Select the csv file you need to upload and click open to upload.



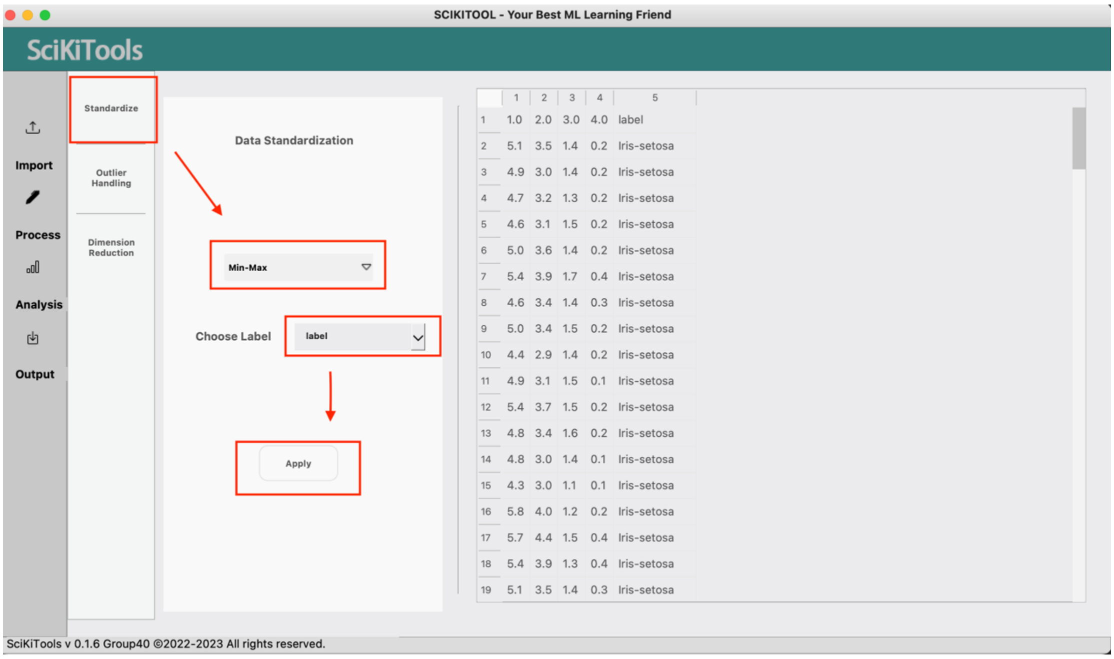
Name	Size	Kind	Date Added
lsc_svm.jpg	74 KB	JPEG image	今天 23:51
lsc_knn_2.jpg	62 KB	JPEG image	今天 23:51
lsc_knn_1.jpg	89 KB	JPEG image	今天 23:51
iris_knn_1.jpg	41 KB	JPEG image	今天 23:51
iris_svm.jpg	42 KB	JPEG image	今天 23:51
iris_knn_2.jpg	41 KB	JPEG image	今天 23:51
test_2.csv	197 bytes	CSV Document	今天 23:51
iris.csv	5 KB	CSV Document	今天 23:56
caret-down.png	214 bytes	PNG image	今天 23:51
arrow-down.png	199 bytes	PNG image	今天 23:51
arrow-left.png	181 bytes	PNG image	今天 23:51
LEGO.png	36 KB	PNG image	今天 23:51

Step3: After that, all the data are shown in app. Click “Process” to start processing the data.


SCIKITOOL - Your Best ML Learning Friend
SciKiTools
Import
Process
Analysis
Output
Import another csv file
Start processing the data
Start analysing the data
Save current modified file
SciKiTools v 0.1.6 Group40 ©2022-2023 All rights reserved.

	1	2	3	4	5
1	1.0	2.0	3.0	4.0	label
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3.0	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5.0	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5.0	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3.0	1.4	0.1	Iris-setosa
15	4.3	3.0	1.1	0.1	Iris-setosa
16	5.8	4.0	1.2	0.2	Iris-setosa
17	5.7	4.4	1.5	0.4	Iris-setosa
18	5.4	3.9	1.3	0.4	Iris-setosa
19	5.1	3.5	1.4	0.3	Iris-setosa

Step4: After click “Process”, you will see the new page blew. Choose the function you want here. For example, we choose the data standardization first. Click the “Standardize” and choose the method you want, we choose “Min-Max” and choose label as label here.


SCIKITOOL - Your Best ML Learning Friend
SciKiTools
Import
Process
Analysis
Output
Standardize
Outlier Handling
Dimension Reduction
Apply
SciKiTools v 0.1.6 Group40 ©2022-2023 All rights reserved.

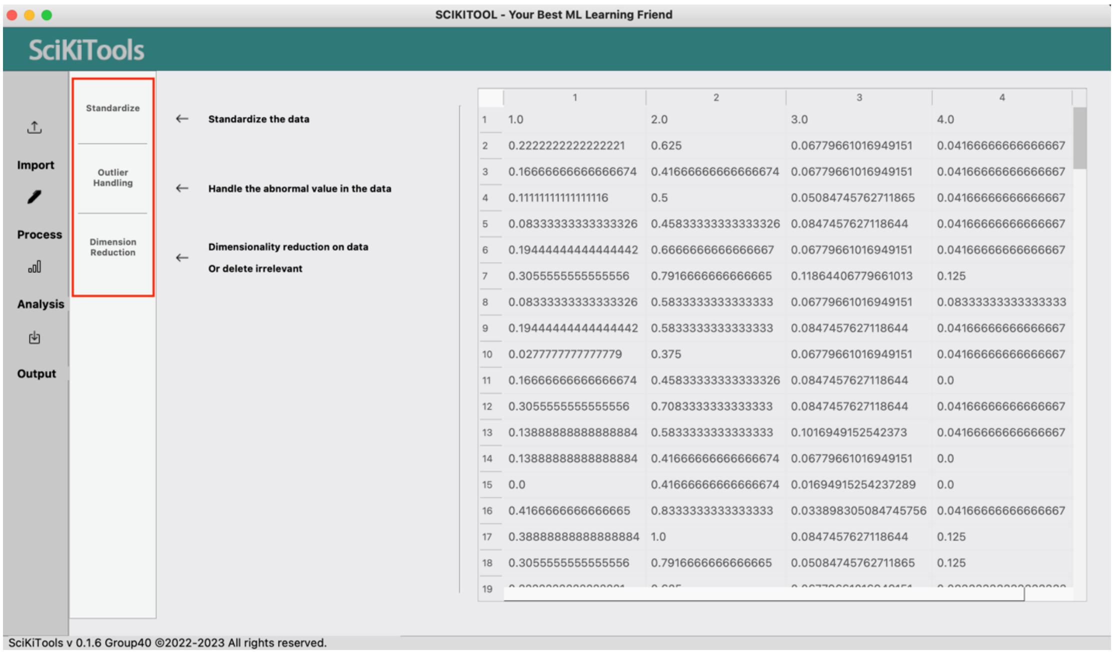
Data Standardization

Min-Max

Choose Label: label

	1	2	3	4	5
1	1.0	2.0	3.0	4.0	label
2	5.1	3.5	1.4	0.2	Iris-setosa
3	4.9	3.0	1.4	0.2	Iris-setosa
4	4.7	3.2	1.3	0.2	Iris-setosa
5	4.6	3.1	1.5	0.2	Iris-setosa
6	5.0	3.6	1.4	0.2	Iris-setosa
7	5.4	3.9	1.7	0.4	Iris-setosa
8	4.6	3.4	1.4	0.3	Iris-setosa
9	5.0	3.4	1.5	0.2	Iris-setosa
10	4.4	2.9	1.4	0.2	Iris-setosa
11	4.9	3.1	1.5	0.1	Iris-setosa
12	5.4	3.7	1.5	0.2	Iris-setosa
13	4.8	3.4	1.6	0.2	Iris-setosa
14	4.8	3.0	1.4	0.1	Iris-setosa
15	4.3	3.0	1.1	0.1	Iris-setosa
16	5.8	4.0	1.2	0.2	Iris-setosa
17	5.7	4.4	1.5	0.4	Iris-setosa
18	5.4	3.9	1.3	0.4	Iris-setosa
19	5.1	3.5	1.4	0.3	Iris-setosa

After clicking on the apply, the processed data is displayed on the left side of the page. You can continue your data processing according to your needs.


SCIKITOOL - Your Best ML Learning Friend
SciKiTools
Import
Process
Analysis
Output
Standardize
Outlier Handling
Dimension Reduction
SciKiTools v 0.1.6 Group40 ©2022-2023 All rights reserved.

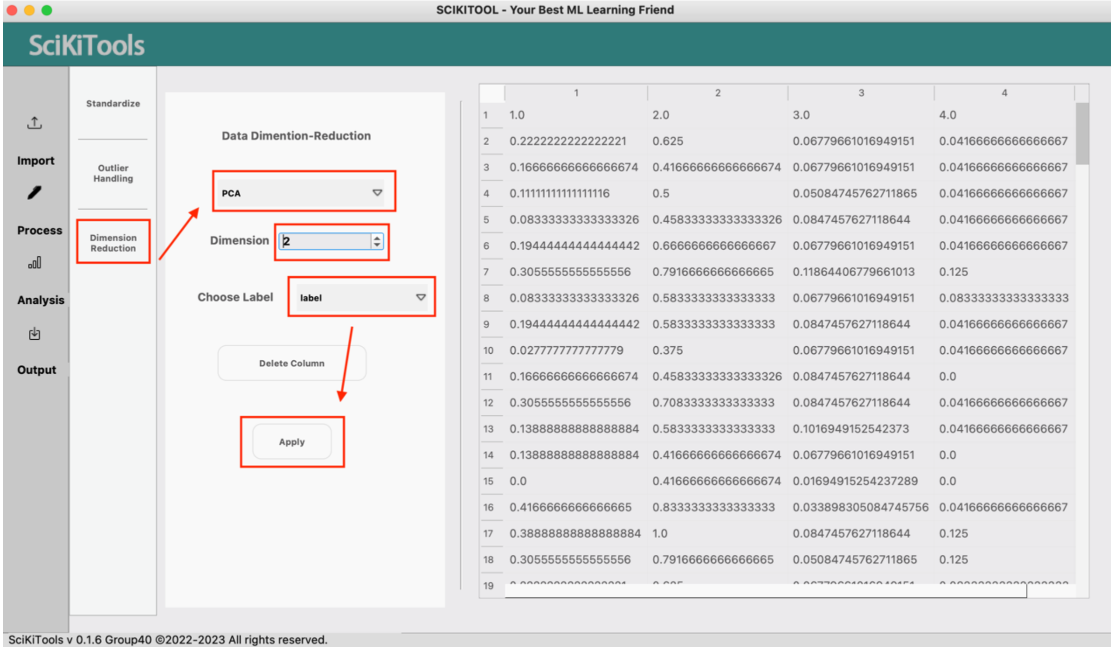
Standardize the data

Handle the abnormal value in the data

Dimensionality reduction on data
Or delete irrelevant

	1	2	3	4
1	1.0	2.0	3.0	4.0
2	0.2222222222222221	0.625	0.06779661016949151	0.0416666666666667
3	0.16666666666666674	0.41666666666666674	0.06779661016949151	0.0416666666666667
4	0.1111111111111116	0.5	0.05084745762711865	0.0416666666666667
5	0.08333333333333326	0.45833333333333326	0.0847457627118644	0.0416666666666667
6	0.19444444444444442	0.6666666666666667	0.06779661016949151	0.0416666666666667
7	0.3055555555555556	0.7916666666666665	0.11864406779661013	0.125
8	0.08333333333333326	0.5833333333333333	0.06779661016949151	0.08333333333333333
9	0.19444444444444442	0.5833333333333333	0.0847457627118644	0.0416666666666667
10	0.02777777777777779	0.375	0.06779661016949151	0.0416666666666667
11	0.16666666666666674	0.45833333333333326	0.0847457627118644	0.0
12	0.3055555555555556	0.7083333333333333	0.0847457627118644	0.0416666666666667
13	0.13888888888888884	0.5833333333333333	0.1016949152542373	0.0416666666666667
14	0.13888888888888884	0.41666666666666674	0.06779661016949151	0.0
15	0.0	0.41666666666666674	0.01694915254237289	0.0
16	0.4166666666666665	0.8333333333333333	0.033898305084745756	0.0416666666666667
17	0.38888888888888884	1.0	0.0847457627118644	0.125
18	0.3055555555555556	0.7916666666666665	0.05084745762711865	0.125
19	[illegible]	0.625	[illegible]	[illegible]

Step5: In order to better help users understand the operational process, we have opted for data reduction. First click on "Dimension Reduction" for Data dimention-reduction, select the method you want and choose the dimension you want. If you want to delete a column, you can select the corresponding column on the right and click on "Delete Column". Click on "Apply" when finished.



Data Dimention-Reduction

PCA

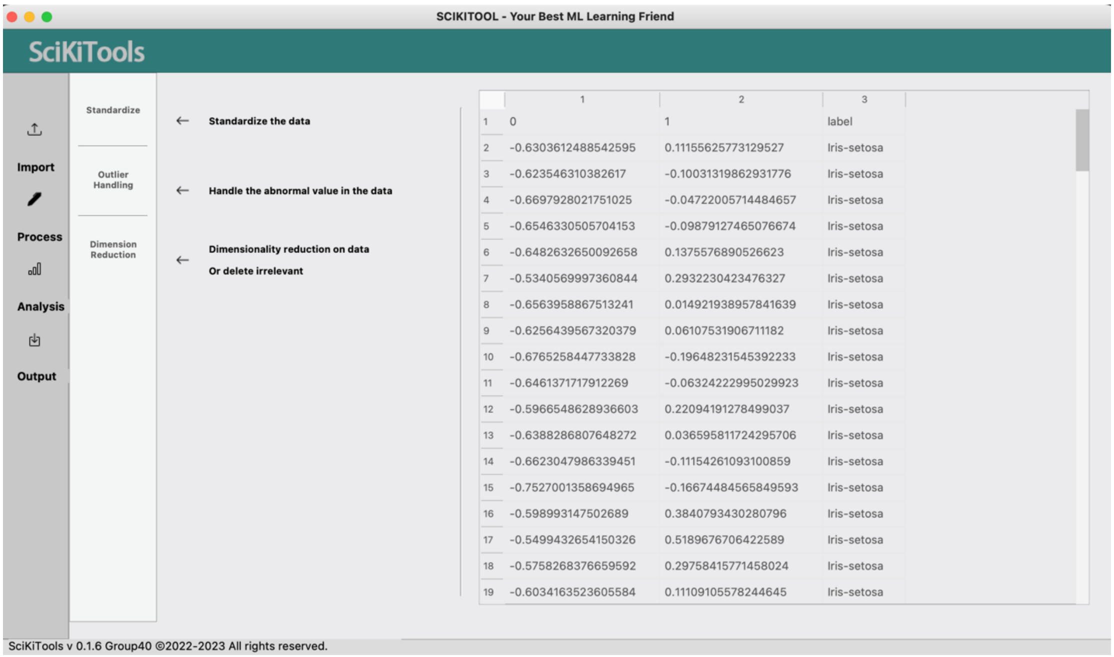
Dimension 2

Choose Label label

	1	2	3	4
1	1.0	2.0	3.0	4.0
2	0.2222222222222221	0.625	0.06779661016949151	0.0416666666666667
3	0.16666666666666674	0.41666666666666674	0.06779661016949151	0.0416666666666667
4	0.1111111111111116	0.5	0.05084745762711865	0.0416666666666667
5	0.08333333333333326	0.45833333333333326	0.0847457627118644	0.0416666666666667
6	0.19444444444444442	0.6666666666666667	0.06779661016949151	0.0416666666666667
7	0.3055555555555556	0.7916666666666665	0.11864406779661013	0.125
8	0.08333333333333326	0.5833333333333333	0.06779661016949151	0.0833333333333333
9	0.19444444444444442	0.5833333333333333	0.0847457627118644	0.0416666666666667
10	0.0277777777777779	0.375	0.06779661016949151	0.0416666666666667
11	0.16666666666666674	0.45833333333333326	0.0847457627118644	0.0
12	0.3055555555555556	0.7083333333333333	0.0847457627118644	0.0416666666666667
13	0.13888888888888884	0.5833333333333333	0.1016949152542373	0.0416666666666667
14	0.13888888888888884	0.41666666666666674	0.06779661016949151	0.0
15	0.0	0.41666666666666674	0.01694915254237289	0.0
16	0.4166666666666665	0.8333333333333333	0.033898305084745756	0.0416666666666667
17	0.38888888888888884	1.0	0.0847457627118644	0.125
18	0.3055555555555556	0.7916666666666665	0.05084745762711865	0.125
19	[illegible]	[illegible]	[illegible]	[illegible]

SciKiTools v 0.1.6 Group40 ©2022-2023 All rights reserved.

After click “Apply”, you can get the result displayed on the left side of the page.



← Standardize the data

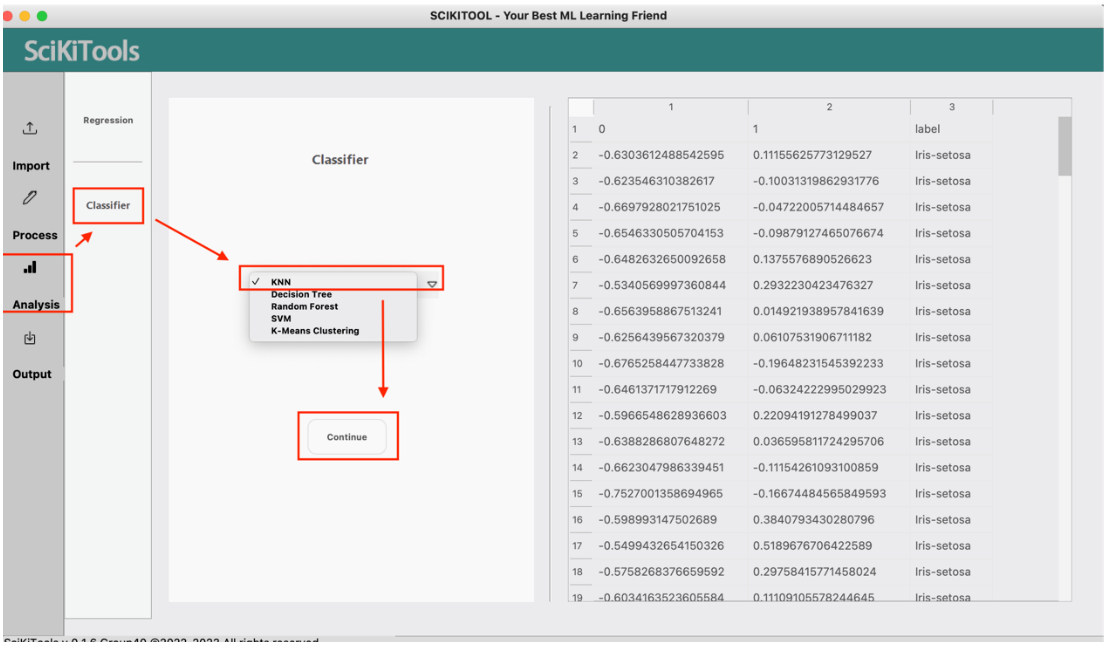
← Handle the abnormal value in the data

← Dimensionality reduction on data
Or delete irrelevant

	1	2	3
1	0	1	label
2	-0.6303612488542595	0.11155625773129527	Iris-setosa
3	-0.623546310382617	-0.10031319862931776	Iris-setosa
4	-0.6697928021751025	-0.04722005714484657	Iris-setosa
5	-0.6546330505704153	-0.09879127465076674	Iris-setosa
6	-0.6482632650092658	0.1375576890526623	Iris-setosa
7	-0.5340569997360844	0.2932230423476327	Iris-setosa
8	-0.6563958867513241	0.014921938957841639	Iris-setosa
9	-0.6256439567320379	0.06107531906711182	Iris-setosa
10	-0.6765258447733828	-0.19648231545392233	Iris-setosa
11	-0.6461371717912269	-0.06324222995029923	Iris-setosa
12	-0.5966548628936603	0.22094191278499037	Iris-setosa
13	-0.6388286807648272	0.036595811724295706	Iris-setosa
14	-0.6623047986339451	-0.11154261093100859	Iris-setosa
15	-0.7527001358694965	-0.16674484565849593	Iris-setosa
16	-0.598993147502689	0.3840793430280796	Iris-setosa
17	-0.5499432654150326	0.5189676706422589	Iris-setosa
18	-0.5758268376659592	0.29758415771458024	Iris-setosa
19	-0.6034163523605584	0.11109105578244645	Iris-setosa

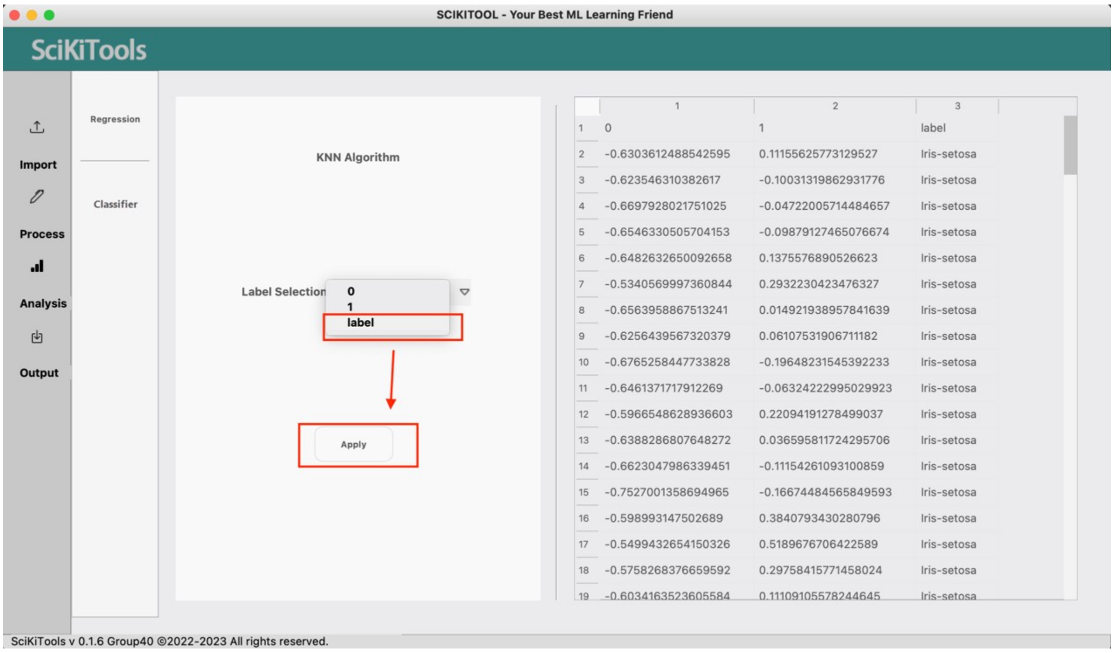
SciKiTools v 0.1.6 Group40 ©2022-2023 All rights reserved.

Step6: After this, we can analyse the data by clicking on "Analysis" and selecting the analysis method you wish to use. Here we have chosen KNN in Machine Learning for our demonstration. First select the machine learning algorithm you wish to use and click on the "Continue" button to proceed.



	1	2	3
1	0	1	label
2	-0.6303612488542595	0.11155625773129527	Iris-setosa
3	-0.623546310382617	-0.10031319862931776	Iris-setosa
4	-0.6697928021751025	-0.04722005714484657	Iris-setosa
5	-0.6546330505704153	-0.09879127465076674	Iris-setosa
6	-0.6482632650092658	0.1375576890526623	Iris-setosa
7	-0.5340569997360844	0.2932230423476327	Iris-setosa
8	-0.6563958867513241	0.014921938957841639	Iris-setosa
9	-0.6256439567320379	0.0610753190671182	Iris-setosa
10	-0.6765258447733828	-0.19648231545392233	Iris-setosa
11	-0.646137171791269	-0.06324222995029923	Iris-setosa
12	-0.5966548628936603	0.22094191278499037	Iris-setosa
13	-0.6388286807648272	0.036595811724295706	Iris-setosa
14	-0.6623047986339451	-0.11154261093100859	Iris-setosa
15	-0.7527001358694965	-0.16674484565849593	Iris-setosa
16	-0.598993147502689	0.3840793430280796	Iris-setosa
17	-0.5499432654150326	0.5189676706422589	Iris-setosa
18	-0.5758268376659592	0.29758415771458024	Iris-setosa
19	-0.6034163523605584	0.11109105578244645	Iris-setosa

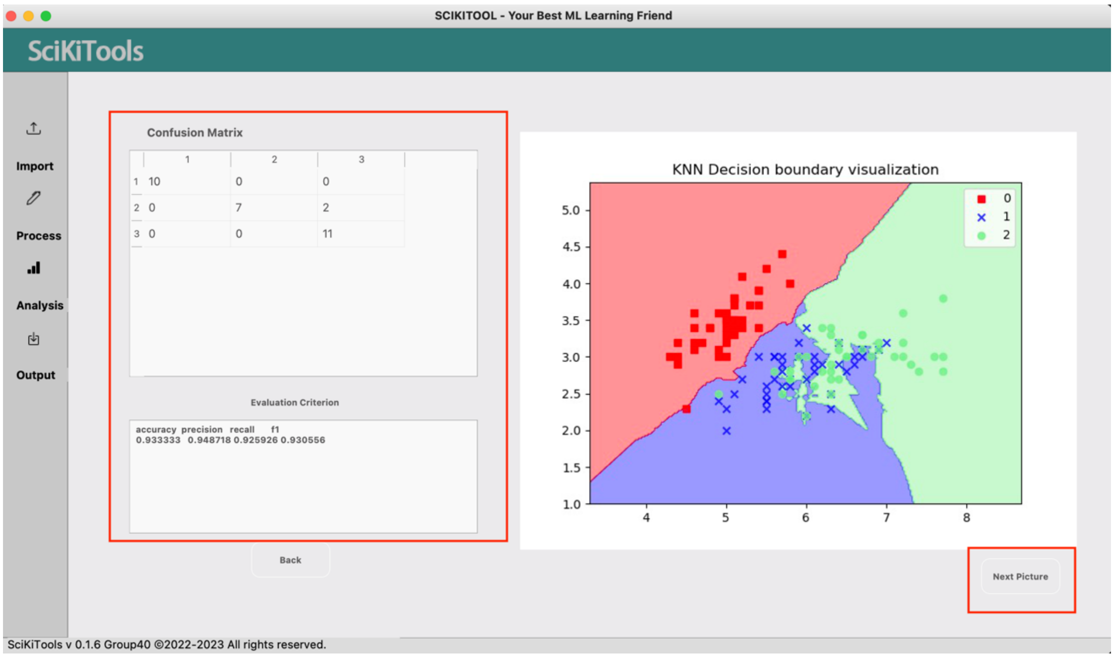
Step7: Select the label you want to select. For demonstration, we choose label here.



	1	2	3
1	0	1	label
2	-0.6303612488542595	0.11155625773129527	Iris-setosa
3	-0.623546310382617	-0.10031319862931776	Iris-setosa
4	-0.6697928021751025	-0.04722005714484657	Iris-setosa
5	-0.6546330505704153	-0.09879127465076674	Iris-setosa
6	-0.6482632650092658	0.1375576890526623	Iris-setosa
7	-0.5340569997360844	0.2932230423476327	Iris-setosa
8	-0.6563958867513241	0.014921938957841639	Iris-setosa
9	-0.6256439567320379	0.0610753190671182	Iris-setosa
10	-0.6765258447733828	-0.19648231545392233	Iris-setosa
11	-0.646137171791269	-0.06324222995029923	Iris-setosa
12	-0.5966548628936603	0.22094191278499037	Iris-setosa
13	-0.6388286807648272	0.036595811724295706	Iris-setosa
14	-0.6623047986339451	-0.11154261093100859	Iris-setosa
15	-0.7527001358694965	-0.16674484565849593	Iris-setosa
16	-0.598993147502689	0.3840793430280796	Iris-setosa
17	-0.5499432654150326	0.5189676706422589	Iris-setosa
18	-0.5758268376659592	0.29758415771458024	Iris-setosa
19	-0.6034163523605584	0.11109105578244645	Iris-setosa

SciKiTools v 0.1.6 Group40 ©2022-2023 All rights reserved.

Step8: After click “Apply”, you can get the result displayed below. For the visualisation section, you can click on "Next Picture" to browse other visualisations, if multiple images are applicable.


SCIKITOOL - Your Best ML Learning Friend
SciKiTools
Import
Process
Analysis
Output
KNN Decision boundary visualization
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
4
5
6
7
8
0
1
2
Back
Next Picture
SciKiTools v 0.1.6 Group40 ©2022-2023 All rights reserved.

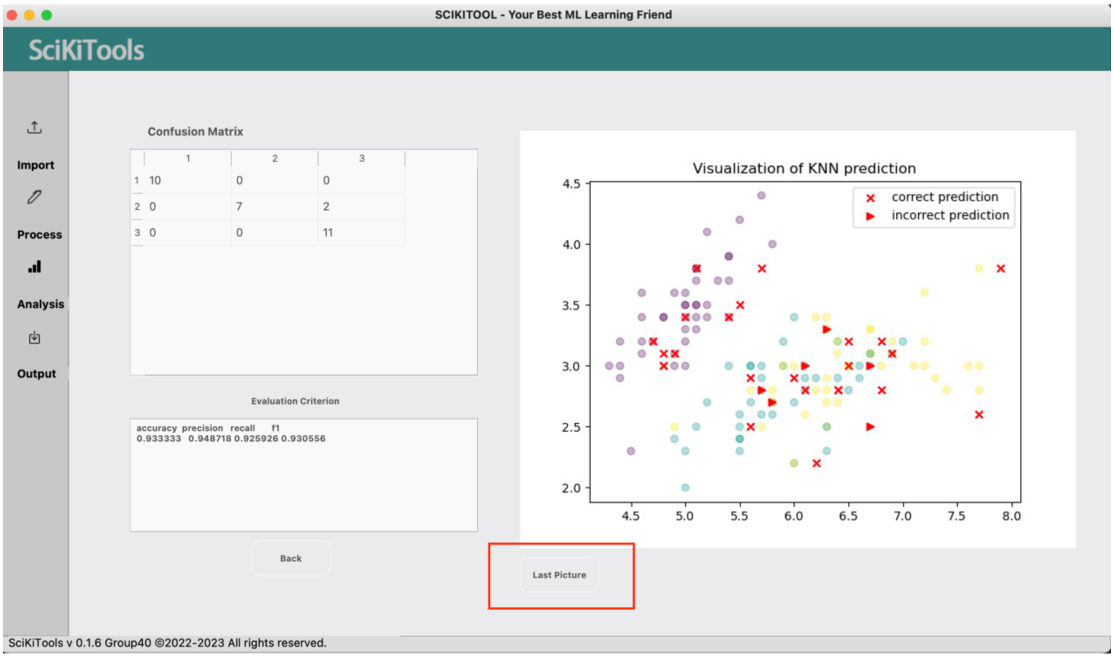
Confusion Matrix

	1	2	3
1	10	0	0
2	0	7	2
3	0	0	11

Evaluation Criterion

```
accuracy precision  recall     f1
0.933333  0.948718 0.925926 0.930556
```

If you want to view the last visualisation, you can click on the "Last Picture" button.



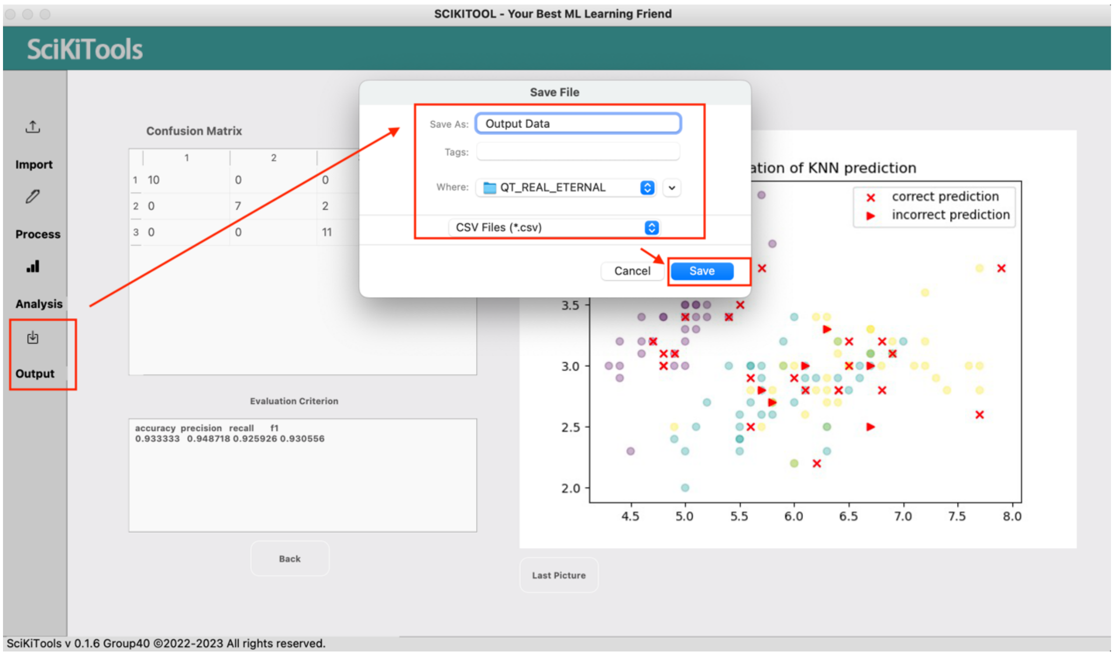
Confusion Matrix

	1	2	3
1	10	0	0
2	0	7	2
3	0	0	11

Evaluation Criterion

```
accuracy precision  recall     f1
0.933333  0.948718 0.925926 0.930556
```

Step9: If you wish to save your analysis data, simply click on the 'Output' button on the right hand side, edit your file name and save address and the format you wish to save it in, then click on the 'Save' button.



Save As: Output Data

Tags:

Where: QT_REAL_ETERNAL

CSV Files (*.csv)

Confusion Matrix

	1	2	
1	10	0	0
2	0	7	2
3	0	0	11

Evaluation Criterion

```
accuracy  precision  recall     f1
0.933333  0.948718 0.925926 0.930556
```

SciKiTools v 0.1.6 Group40 ©2022-2023 All rights reserved.

When using the software, the user can select the required steps for the above operations. They can be saved at any time by clicking on the "Output" button. If you wish to analyse more than one piece of data you can repeat the above actions.

4. Features

This section of the user manual describes in detail the features of the offline data analysis software SciKiTools, including function descriptions, usage, parameter settings, etc.

1) My Data

i. Upload Data Set

Function: Upload data sets in CSV.

Usage: Click on “Import” in the main window, then browse to the file you want to upload and select the appropriate format.

Parameters: None

ii. Data Details

Function: Display information about the selected data set, such as the number of rows, columns, and data types.

Usage: After selecting a data set, the “Data Details" will show on the right part of the main window.

Parameters: None

2) Data Processing

i. Data Pre-processing

Function: Standardize data using min-max or z-score normalization techniques.

Usage: In the "Data Processing" section, click on "Standardization." Choose between min-max or z-score normalization and apply it to the selected data set.

Parameters:

- Columns: Select the columns to apply the normalization method to.
- New Variable: Choose whether to create a new variable for the normalized data or overwrite the existing data.

ii. Outlier Handling

Function: Remove duplicate data points from the selected data set.

Usage: In the "Data Processing" section, click on "Outlier Handling." Choose the "Remove Duplicate" option to process the data.

Parameters:

- Columns: Select the columns to apply the outlier handling method to.
- Method: Choose the outlier detection method, such as IQR, Z-score, or Tukey fences.
- Action: Select the action to perform on detected outliers, such as remove or replace with a specified value.

iii. Dimension Reduction

Function: Reduce the dimensionality of the data using techniques such as PCA, NMF, TSNE, or Isomap.

Usage: In the "Data Processing" section, click on "Dimension Reduction." Select a processing technique and adjust its parameters accordingly. The default value for the label column is "None," but you can select and delete a column using the display bar.

Parameters (specific to each technique):

- PCA: Number of components, scaling (standard or none).
- NMF: Number of components, initialization method (random, nndsvd, nndsvda, or nndsvdar), and random state.
- TSNE: Number of components, perplexity, early exaggeration, learning rate, and random state.
- Isomap: Number of components and neighbors.

3) Descriptive Analysis

i. Regression

Function: Perform logistic and linear regression on the selected data set.

Usage: In the "Data Analysis" section, click on "Regression." Choose between logistic or linear regression and adjust the parameters accordingly.

Parameters (specific to each regression type):

- Logistic Regression: Penalty (L1 or L2), C (inverse of regularization strength), and random state.
- Linear Regression: Fit intercept (True or False) and normalize (True or False).

ii. Classifier

Function: Apply various machine learning techniques, including KNN, decision tree, random forest, SVM, and K-means clustering. Generate analysis tables and images based on the chosen algorithms.

Usage: In the "Data Analysis" section, click on "Machine Learning." Select a machine learning technique and adjust its parameters accordingly. The software will generate analysis tables and images based on the chosen algorithm.

Parameters (specific to each machine learning technique):

- KNN: Number of neighbors, weights (uniform or distance), and distance metric (e.g., euclidean, manhattan, minkowski).
- Decision Tree: Criterion (gini or entropy), max depth, and random state.
- Random Forest: Number of estimators, criterion (gini or entropy), max depth, and random state.
- SVM: Kernel (linear, poly, rbf, or sigmoid), degree (for poly kernel), gamma (for rbf, poly, or sigmoid kernels), and C (regularization parameter).
- K-means Clustering: Number of clusters, initialization method (k-means++ or random), and random state.

4) Export Results

Function: Export the results of the data analysis in csv and jpg format.

Usage: After completing the data analysis, click on the "Export Results" button and save the file to users' computer.

i. Components of results

Confusion Matrix

This matrix helps users understand how well a model is performing by comparing its predicted outcomes with the actual outcomes. The matrix consists of four components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). To evaluate model's performance, we use the following criteria derived from the Confusion Matrix, shown as evaluation criterion:

Evaluation criterion

a) Accuracy

This is the most common evaluation metric, which measures the overall performance of a model. It is calculated by dividing the sum of correct predictions (TP and TN) by the total number of predictions. The formula is:

$$\boldsymbol{Accuracy} = \frac{(\boldsymbol{TP} + \boldsymbol{TN})}{(\boldsymbol{TP} + \boldsymbol{TN} + \boldsymbol{FP} + \boldsymbol{FN})}$$

b) Precision

Precision is the ratio of correctly predicted positive instances to the total predicted positive instances. This metric is useful when the cost of false positives is high. The formula for precision is:

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FP})}$$

c) Recall

Also known as sensitivity, recall is the ratio of correctly predicted positive instances to the actual positive instances. This metric is particularly useful when the cost of false negatives is high. The formula for recall is:

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{(\boldsymbol{TP} + \boldsymbol{FN})}$$

d) F1

The F1 Score is another evaluation criterion that considers both precision and recall to provide a more balanced assessment of a model's performance. It is particularly useful when dealing with imbalanced datasets or when both false positives and false negatives are important to consider. The F1 Score is the harmonic mean of precision and recall and ranges from 0 to 1, with 1 being the best possible score. The formula for the F1 Score is:

$$\boldsymbol{F1\,Score} = \frac{\boldsymbol{2} * (\boldsymbol{Precision} * \boldsymbol{Recall})}{(\boldsymbol{Precision} + \boldsymbol{Recall})}$$

By using these evaluation criteria, we can assess the performance of the classification model in terms of its ability to accurately predict results, minimise false positives and minimise false negatives. This allows us to better understand the strengths and weaknesses of the model and to make informed decisions about its deployment and further development.

By following this updated feature section of the user manual, you will be able to fully utilize the software's data processing and analysis capabilities to generate insightful reports and visualizations.

5. Frequently Asked Questions (FAQs)

Q1: What file formats are supported for uploading data sets?

A: The software is CSV files only. Please ensure that your data is in CSV format before attempting to upload.

Q2: Can I standardize my data using the software?

A: Yes, SciKiTools allows you to standardize your data using either min-max or z-score normalization methods. You can find this feature in the Data Processing section under Standardization.

Q3: Can I perform both regression and machine learning analyses using the software?

A: SciKiTools offers both regression (logistic and linear) and machine learning techniques such as KNN, decision trees, random forests, SVM and K-means clustering. and these options can be found in the Data Analysis section. However, due to the iterative nature of the software and the order in which features are developed, only regression and machine learning analysis of datasets are currently supported separately.

Q4: I encountered an error while using the software. What should I do?

A: First, ensure that your data set is properly formatted and uploaded. Check the parameters for the specific feature you are using and make sure they are correctly set. If the issue persists, consult the user manual for further guidance or seek assistance from the SciKiTools's support team.

Q5: How do I export the results of my data analysis?

A: After completing the data analysis, you can export the results in csv or as images. Click on the “Export Results” button and save the file to your computer.

Q6: Can I see the data I've processed before?

A: The processed data can be conveniently saved locally by clicking the "Save" button. To access the saved data, simply navigate to the corresponding local directory. We have no authority to access users’ data. If you need to re-process the data, you can easily re-import it into the software for further analysis.

The FAQ's provide a quick reference for users who need help in using the offline data analysis software. Users are encouraged to consult the user manual or seek assistance from the support team if additional questions arise.

6. Technical support

Our team is committed to providing exceptional support for the offline data analysis software SciKiTools. If you encounter any issues or have questions regarding the software, you can access technical assistance through the following channels:

1) Documentation

Before reaching out to our support team, we recommend consulting the software documentation, which provides comprehensive information on features, functions, usage, and troubleshooting. The documentation is available on our website and can be accessed directly within the software.

2) Github community Forum

We have active community forums available on GitHub where users can ask questions, share their experiences, and help each other with software-related issues. You can search the forum for existing answers or post a new question to get help from the community. We developers also respond to questions on a regular basis.