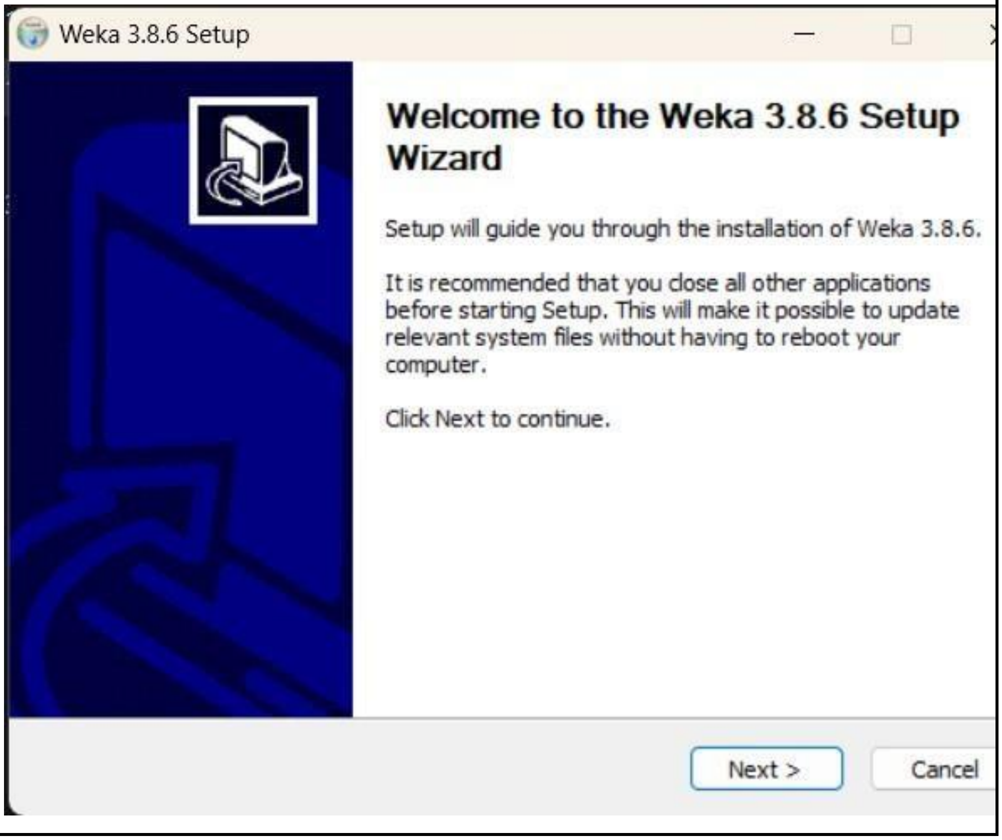
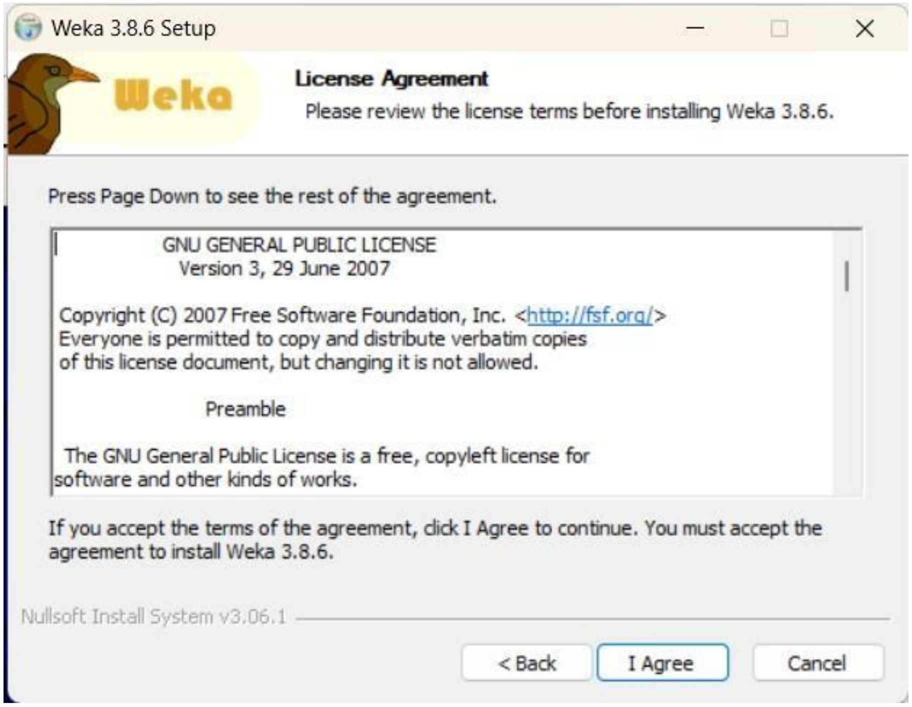
|  |  |
|---|---|
| **Name** | Shubham Golwal |
| **UID no.** | 2020300015 |
| **Experiment No.** | 07 |

|  |  |
|---|---|
| **AIM:** | Build a Data Warehouse and Explore WEKA Implementation |
| **Program 1** |  |
| **PROBLEM STATEMENT :** | The objective of this exp is to build a data warehouse using breast cancer datasets and explore WEKA implementation for data analysis. The expt involves data preprocessing, integration, and transformation to build the data warehouse and applying various data mining algorithms using WEKA. The expt is expected to generate insights for breast cancer diagnosis and treatment and contribute to the improvement of healthcare outcomes. |
| **Theory :** | **What is Weka Tool?**<br>Weka is a popular open-source machine learning software tool that provides a wide range of algorithms for data preprocessing, classification, clustering, and regression. It is widely used in both academia and industry for data analysis, data mining, and machine learning tasks.<br>In addition to its powerful features, Weka also provides extensive documentation to help users learn how to use the tool effectively. The documentation includes a user manual, a developer guide, and a set of online tutorials that cover various aspects of Weka's functionality.<br>The user manual provides an overview of Weka's features and explains how to install and use the tool. It also includes detailed descriptions of Weka's algorithms, including their underlying mathematical models and their use cases.<br>The developer guide provides information on Weka's architecture, including the various modules and components that make up the tool. It also includes information on how to extend Weka's functionality by adding new algorithms or modules.<br>The online tutorials cover various topics, including data preprocessing, classification, clustering, and regression. Each tutorial provides step-by-step |

|  |  |
|---|---|
|  | instructions on how to use Weka's features to solve a particular problem, and includes sample data and code snippets to help users get started.<br>Overall, Weka's extensive documentation makes it an accessible and user-friendly tool for data analysis and machine learning tasks. |
| **Installation** | <br> <br>**Welcome to the Weka 3.8.6 Setup Wizard**<br>Setup will guide you through the installation of Weka 3.8.6.<br>It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.<br>Click Next to continue.<br> |



Weka 3.8.6 Setup

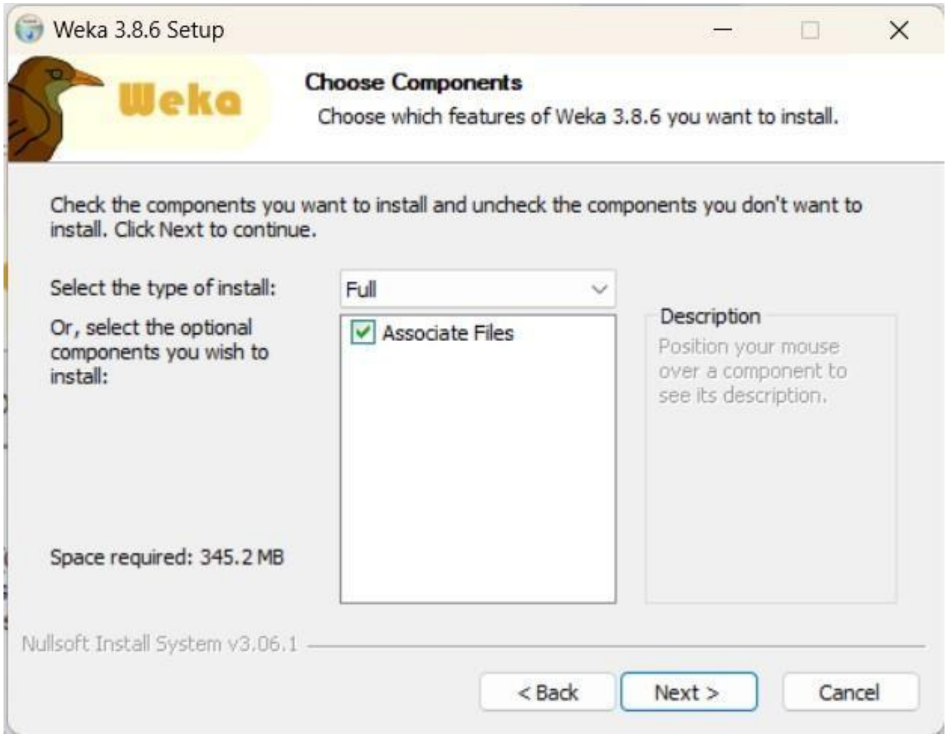
**License Agreement**

Please review the license terms before installing Weka 3.8.6.

Press Page Down to see the rest of the agreement.

```
                GNU GENERAL PUBLIC LICENSE
                  Version 3, 29 June 2007

 Copyright (C) 2007 Free Software Foundation, Inc. <http://fsf.org/>
 Everyone is permitted to copy and distribute verbatim copies
 of this license document, but changing it is not allowed.

                       Preamble

  The GNU General Public License is a free, copyleft license for
software and other kinds of works.
```

If you accept the terms of the agreement, click I Agree to continue. You must accept the agreement to install Weka 3.8.6.



Weka 3.8.6 Setup

**Choose Components**

Choose which features of Weka 3.8.6 you want to install.

Check the components you want to install and uncheck the components you don't want to install. Click Next to continue.

Select the type of install: Full

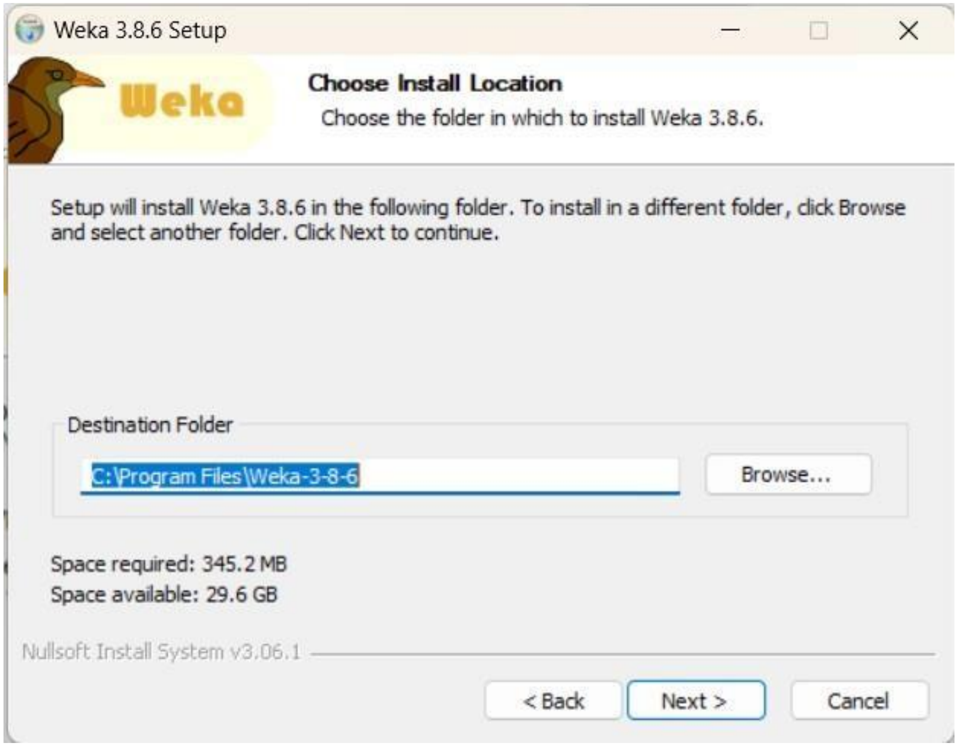
Or, select the optional components you wish to install:

☑ Associate Files

Space required: 345.2 MB

Description

Position your mouse over a component to see its description.



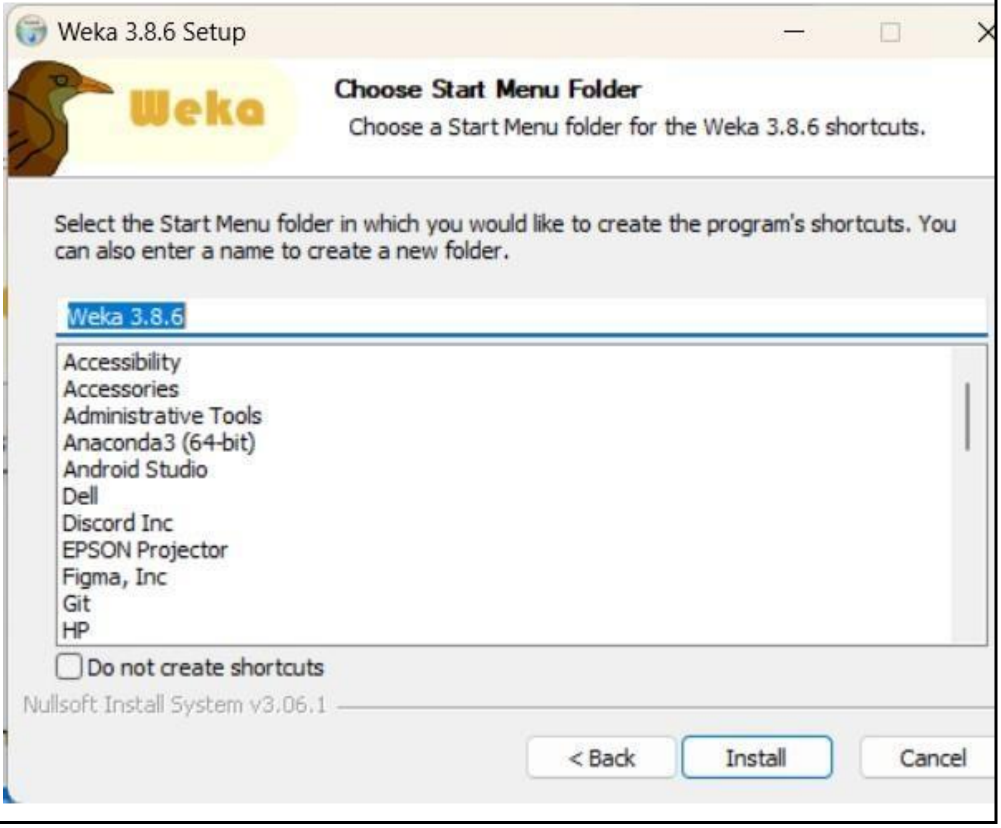
**Choose Install Location**

Choose the folder in which to install Weka 3.8.6.

Setup will install Weka 3.8.6 in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder: C:\Program Files\Weka-3-8-6

Space required: 345.2 MB
Space available: 29.6 GB



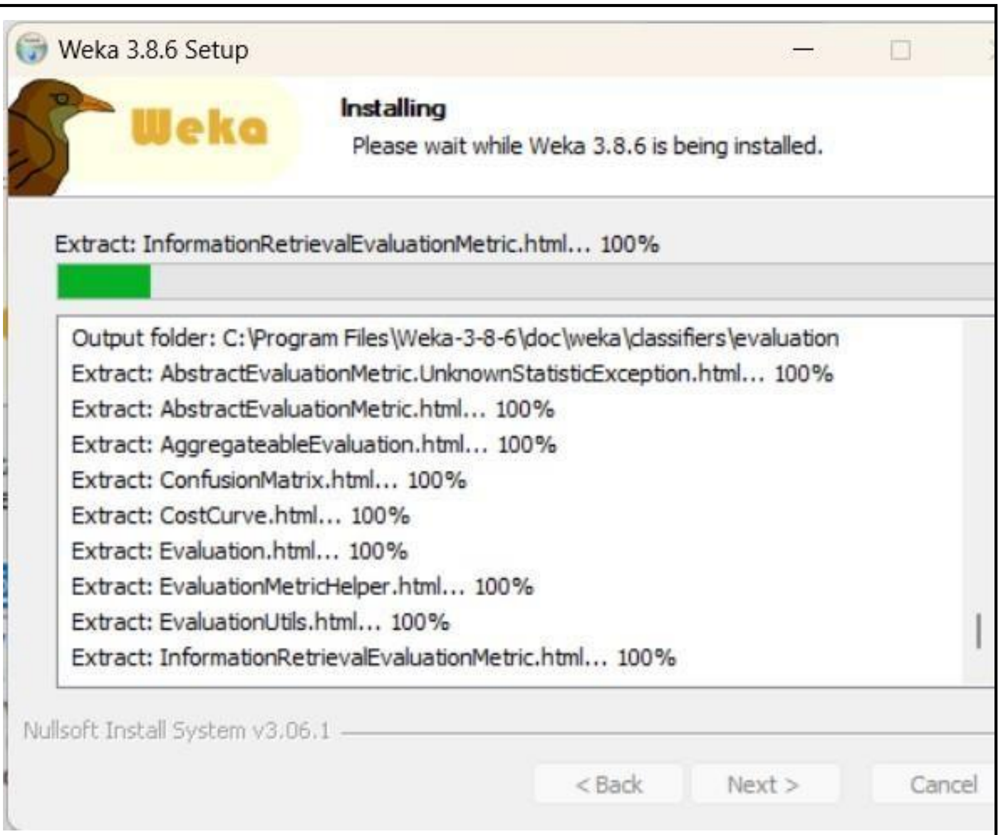
**Choose Start Menu Folder**

Choose a Start Menu folder for the Weka 3.8.6 shortcuts.

Select the Start Menu folder in which you would like to create the program's shortcuts. You can also enter a name to create a new folder.

Weka 3.8.6

☐ Do not create shortcuts



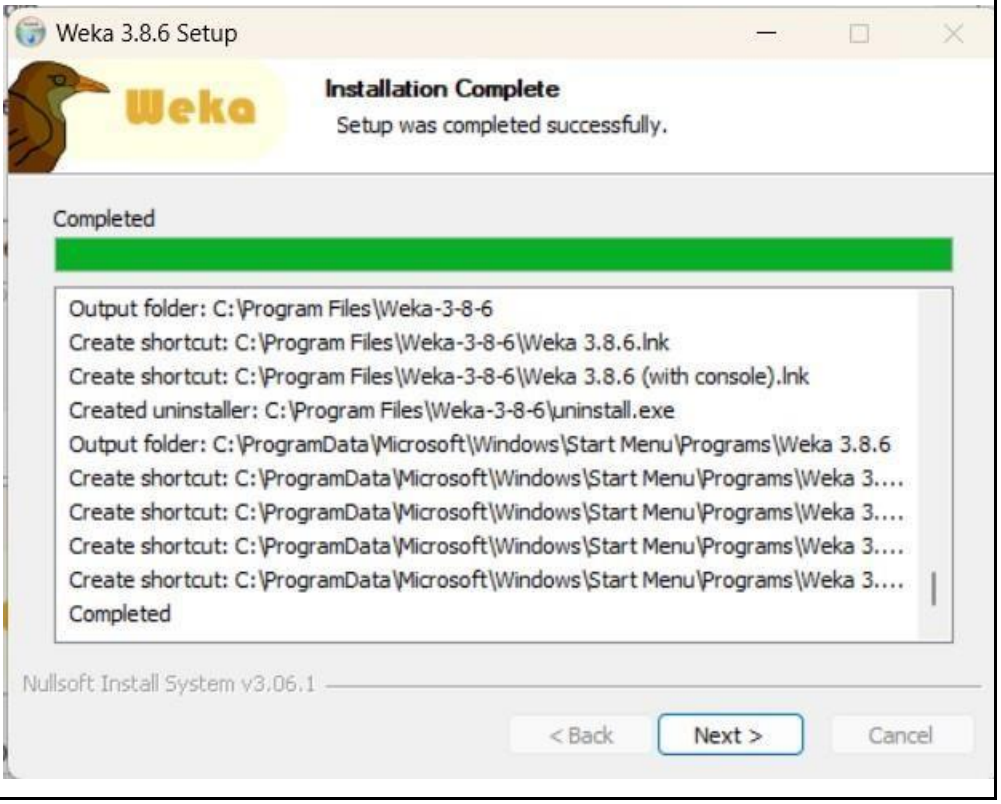
Weka 3.8.6 Setup

**Installing**

Please wait while Weka 3.8.6 is being installed.

```
Extract: InformationRetrievalEvaluationMetric.html... 100%
```

```
Output folder: C:\Program Files\Weka-3-8-6\doc\weka\classifiers\evaluation
Extract: AbstractEvaluationMetric.UnknownStatisticException.html... 100%
Extract: AbstractEvaluationMetric.html... 100%
Extract: AggregateableEvaluation.html... 100%
Extract: ConfusionMatrix.html... 100%
Extract: CostCurve.html... 100%
Extract: Evaluation.html... 100%
Extract: EvaluationMetricHelper.html... 100%
Extract: EvaluationUtils.html... 100%
Extract: InformationRetrievalEvaluationMetric.html... 100%
```



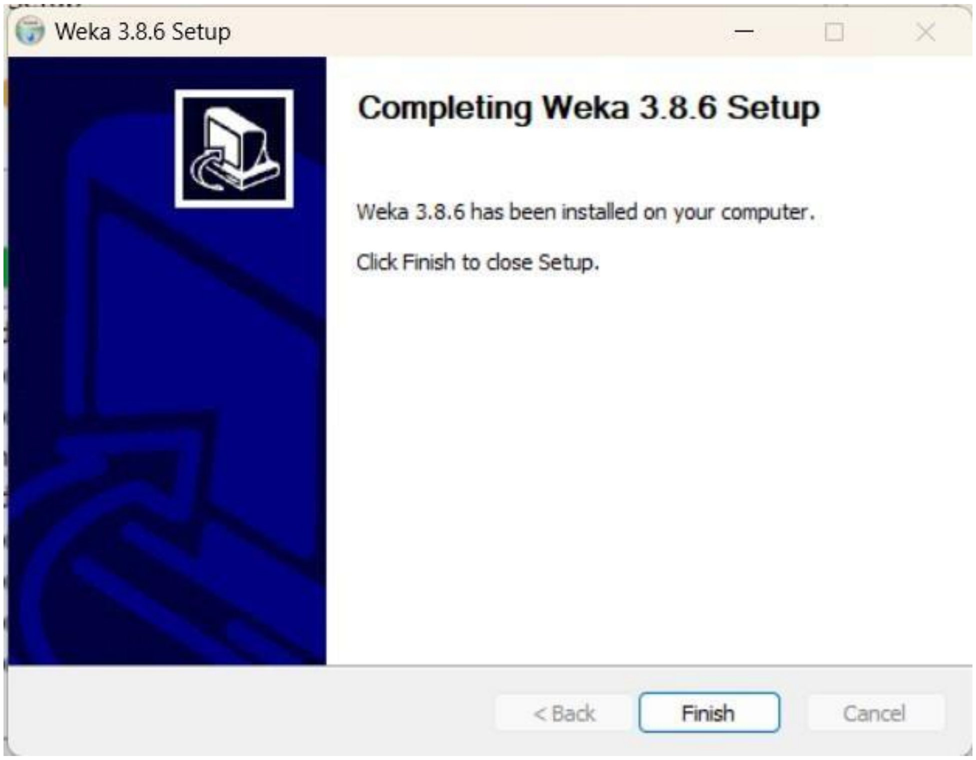
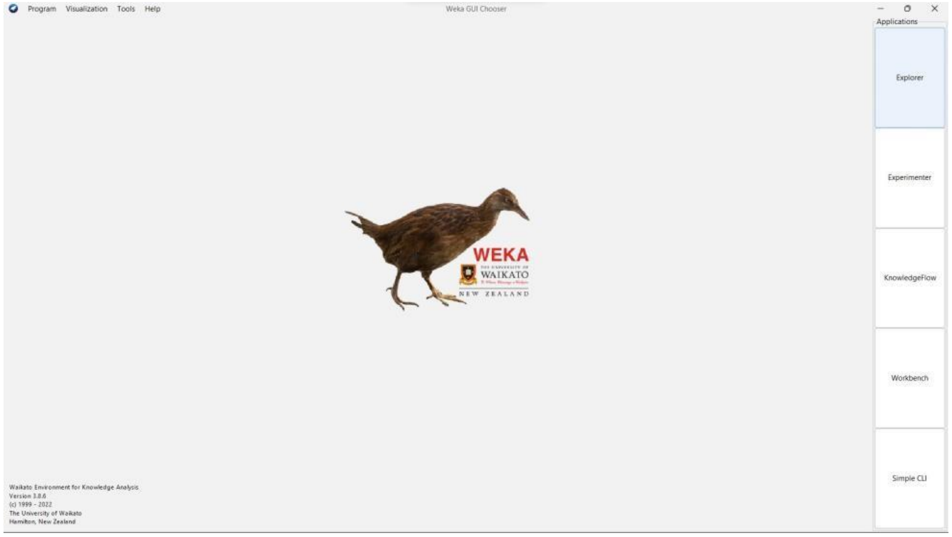
Weka 3.8.6 Setup

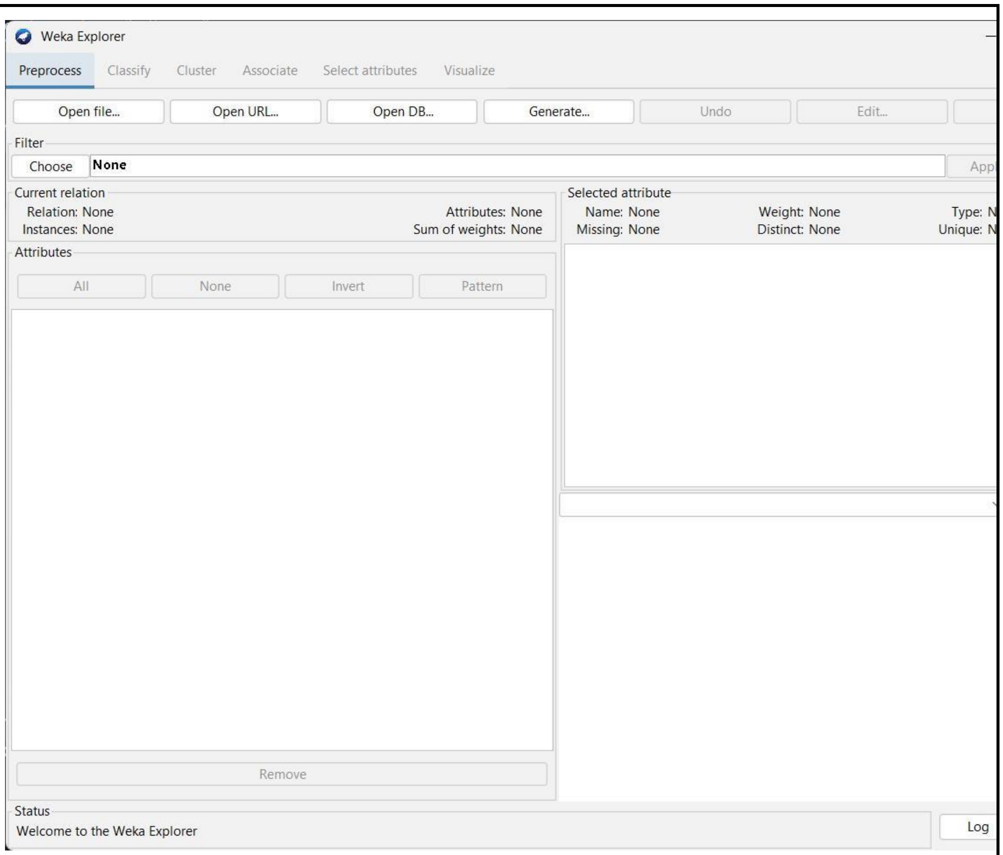
**Installation Complete**

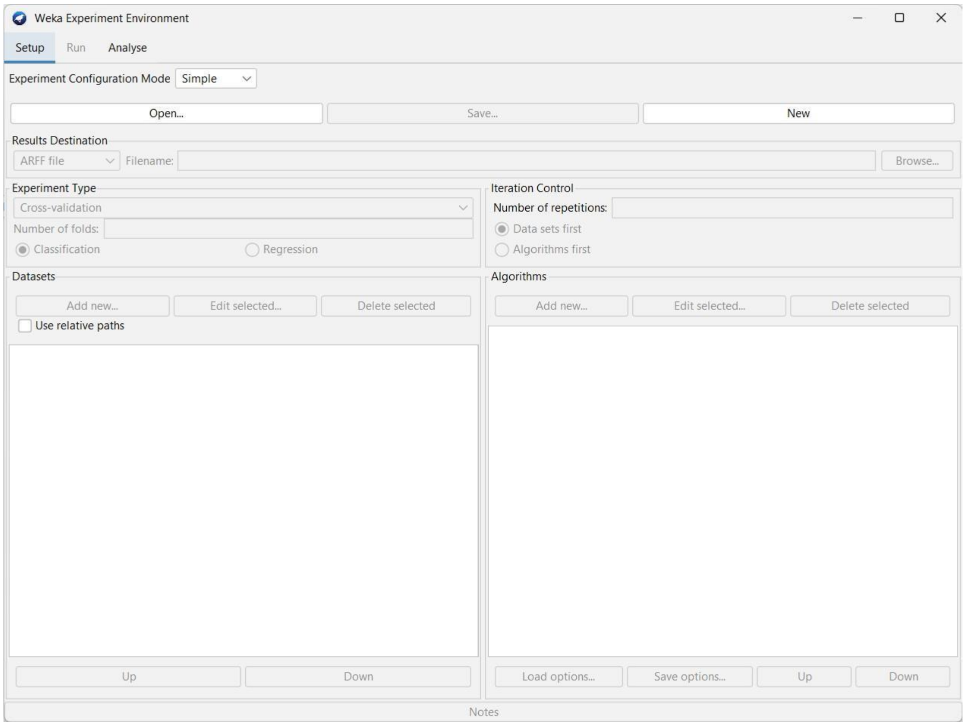
Setup was completed successfully.

Completed

```
Output folder: C:\Program Files\Weka-3-8-6
Create shortcut: C:\Program Files\Weka-3-8-6\Weka 3.8.6.lnk
Create shortcut: C:\Program Files\Weka-3-8-6\Weka 3.8.6 (with console).lnk
Created uninstaller: C:\Program Files\Weka-3-8-6\uninstall.exe
Output folder: C:\ProgramData\Microsoft\Windows\Start Menu\Programs\Weka 3.8.6
Create shortcut: C:\ProgramData\Microsoft\Windows\Start Menu\Programs\Weka 3....
Create shortcut: C:\ProgramData\Microsoft\Windows\Start Menu\Programs\Weka 3....
Create shortcut: C:\ProgramData\Microsoft\Windows\Start Menu\Programs\Weka 3....
Create shortcut: C:\ProgramData\Microsoft\Windows\Start Menu\Programs\Weka 3....
Completed
```

| | |
|---|---|
| | <br>Weka 3.8.6 Setup<br>Completing Weka 3.8.6 Setup<br>Weka 3.8.6 has been installed on your computer.<br>Click Finish to close Setup.<br> |
| **Output** | **Weka Opened: The landing page of Weka software**<br><br>Waikato Environment for Knowledge Analysis<br>Version 3.8.6<br>(c) 1999 - 2022<br>The University of Waikato<br>Hamilton, New Zealand<br> Program<br>Visualization<br>Tools<br>Help<br>Weka GUI Chooser<br>Applications<br>Explorer<br>Experimenter<br>KnowledgeFlow<br>Workbench<br>Simple CLI<br>WEKA<br>THE UNIVERSITY OF WAIKATO<br>NEW ZEALAND<br>**Explorer tab: It allows datasets and the predictions of Classifiers and Clutterers to be visualized in two dimensions.** |

| | |
|---|---|
| | <br> Weka Explorer<br>Preprocess<br>Classify<br>Cluster<br>Associate<br>Select attributes<br>Visualize<br>Open file...<br>Open URL...<br>Open DB...<br>Generate...<br>Undo<br>Edit...<br>Filter<br>Choose<br>App<br>Current relation<br>Selected attribute<br>Attributes<br>All<br>None<br>Invert<br>Pattern<br>Remove<br>Status<br>Log<br>None<br>Relation: None<br>Instances: None<br>Attributes: None<br>Sum of weights: None<br>Name: None<br>Missing: None<br>Weight: None<br>Distinct: None<br>Type: N<br>Unique: N<br>Welcome to the Weka Explorer<br>**Experimenter tab: The experimenter configures the test options for you with sensible defaults.** |



Experiment Configuration Mode: Simple

Results Destination: ARFF file Filename:

Experiment Type: Cross-validation

Number of folds:

● Classification ○ Regression

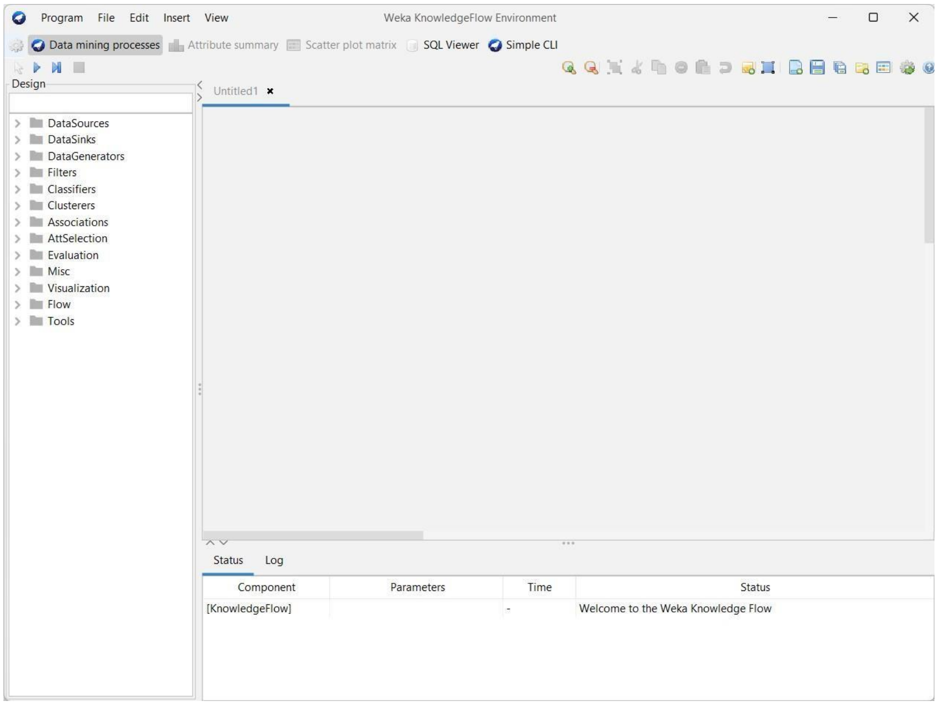
Number of repetitions:

● Data sets first

○ Algorithms first

□ Use relative paths

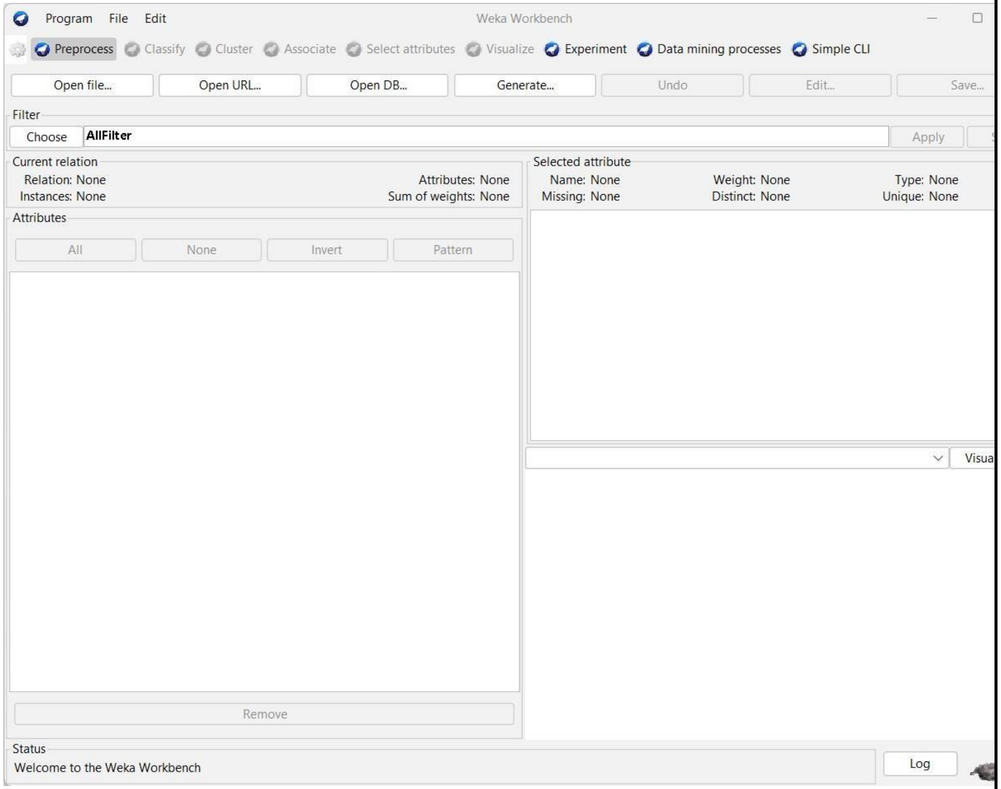
**KnowledgeFlow tab: It presents a "data-flow" inspired interface to Weka**



| Component | Parameters | Time | Status |
|---|---|---|---|
| [KnowledgeFlow] |  | - | Welcome to the Weka Knowledge Flow |

**Workbench tab It contains a collection of data pre-processing tools and machine learning algorithms wrapped in an easy-**

**to-use graphical interface.**



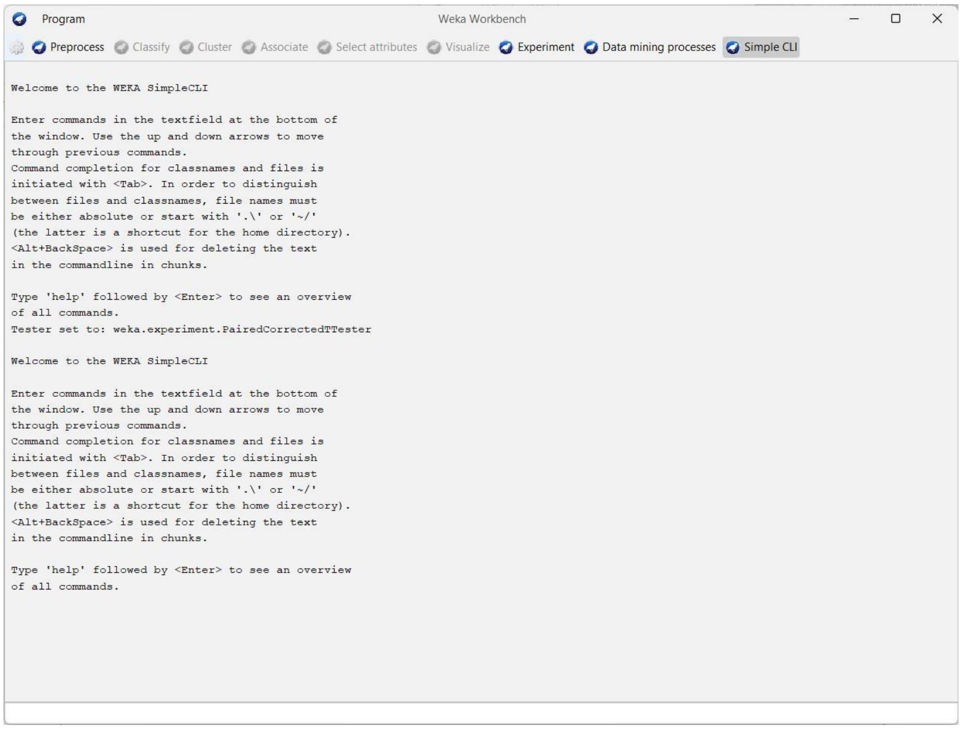
Filter: AllFilter

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

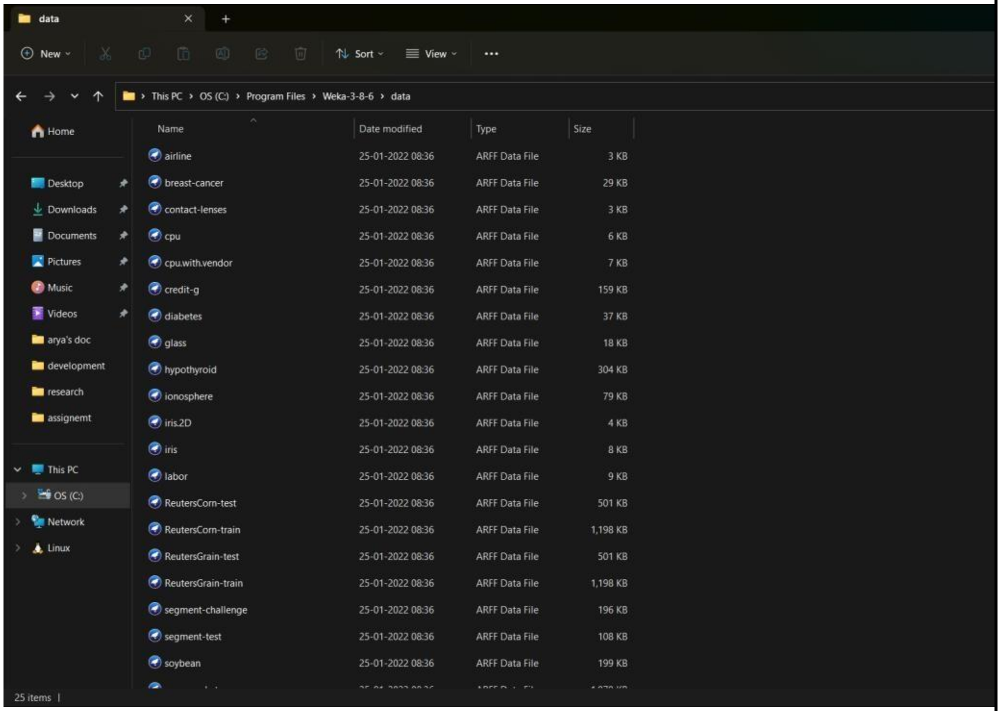
Welcome to the Weka Workbench

## Simple CLI



```
Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.\' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

Type 'help' followed by <Enter> to see an overview
of all commands.
Tester set to: weka.experiment.PairedCorrectedTTester

Welcome to the WEKA SimpleCLI

Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with '.\' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.

Type 'help' followed by <Enter> to see an overview
of all commands.
```

## WEKA DATASETS:



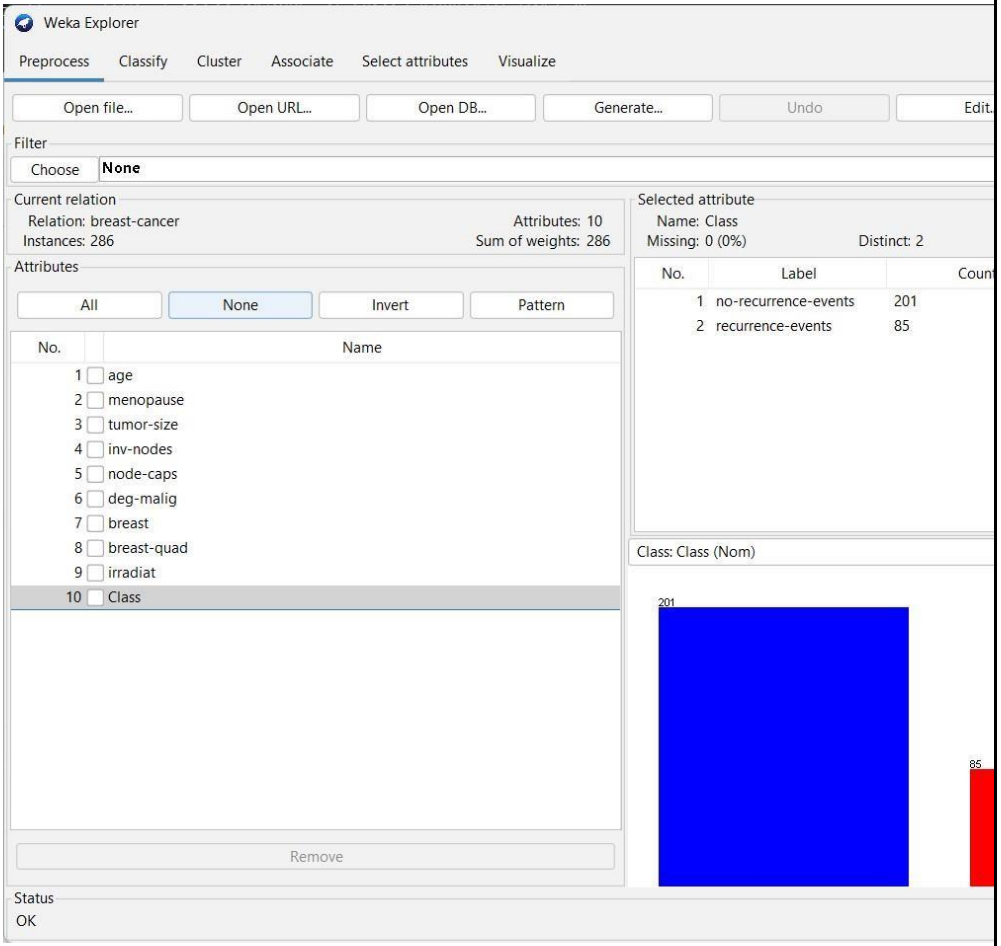
| Name | Date modified | Type | Size |
|---|---|---|---|
| airline | 25-01-2022 08:36 | ARFF Data File | 3 KB |
| breast-cancer | 25-01-2022 08:36 | ARFF Data File | 29 KB |
| contact-lenses | 25-01-2022 08:36 | ARFF Data File | 3 KB |
| cpu | 25-01-2022 08:36 | ARFF Data File | 6 KB |
| cpu.with.vendor | 25-01-2022 08:36 | ARFF Data File | 7 KB |
| credit-g | 25-01-2022 08:36 | ARFF Data File | 159 KB |
| diabetes | 25-01-2022 08:36 | ARFF Data File | 37 KB |
| glass | 25-01-2022 08:36 | ARFF Data File | 18 KB |
| hypothyroid | 25-01-2022 08:36 | ARFF Data File | 304 KB |
| ionosphere | 25-01-2022 08:36 | ARFF Data File | 79 KB |
| iris.2D | 25-01-2022 08:36 | ARFF Data File | 4 KB |
| iris | 25-01-2022 08:36 | ARFF Data File | 8 KB |
| labor | 25-01-2022 08:36 | ARFF Data File | 9 KB |
| ReutersCorn-test | 25-01-2022 08:36 | ARFF Data File | 501 KB |
| ReutersCorn-train | 25-01-2022 08:36 | ARFF Data File | 1,198 KB |
| ReutersGrain-test | 25-01-2022 08:36 | ARFF Data File | 501 KB |
| ReutersGrain-train | 25-01-2022 08:36 | ARFF Data File | 1,198 KB |
| segment-challenge | 25-01-2022 08:36 | ARFF Data File | 196 KB |
| segment-test | 25-01-2022 08:36 | ARFF Data File | 108 KB |
| soybean | 25-01-2022 08:36 | ARFF Data File | 199 KB |

Choosing dataset - We have chosen a breast cancer.arff dataset which contains 3 attributes with yes or no questions and 7 continuous regression matrix.

The breast cancer dataset contains ten attributes that describe various characteristics of breast cancer patients and their tumors. Here's a brief description of each attribute:

1. age: The age of the patient at the time of diagnosis.
2. menopause: The menopausal status of the patient at the time of diagnosis.
3. tumor-size: The size of the tumor, measured in centimeters.
4. inv-nodes: The number of axillary lymph nodes involved in the cancer.
5. node-caps: Whether or not the cancer has spread to the lymph node capsule.
6. deg-malig: The degree of malignancy of the tumor, ranging from 1 (low) to 3 (high).
7. breast: The affected breast (left or right).
8. breast-quad: The quadrant of the breast where the tumor is located.
9. irradiat: Whether or not the patient received radiation therapy.
10. Class: The classification of the tumor as benign or malignant.

These attributes provide important information for assessing the severity and treatment options for breast cancer. Age, menopause, and tumor size can provide insights into the progression of the disease, while inv-nodes, node-caps, and deg-malig can help determine the stage and aggressiveness of the cancer. The breast and breast-quad attributes can help locate the tumor, while irradiat can indicate the type of treatment received. Finally, the Class attribute is the target variable that is used to classify the tumor as either benign or malignant.



Filter

Choose None

Current relation

Relation: breast-cancer Attributes: 10

Instances: 286 Sum of weights: 286

Selected attribute

Name: Class

Missing: 0 (0%) Distinct: 2

| No. | Label | Count |
|---|---|---|
| 1 | no-recurrence-events | 201 |
| 2 | recurrence-events | 85 |

Class: Class (Nom)

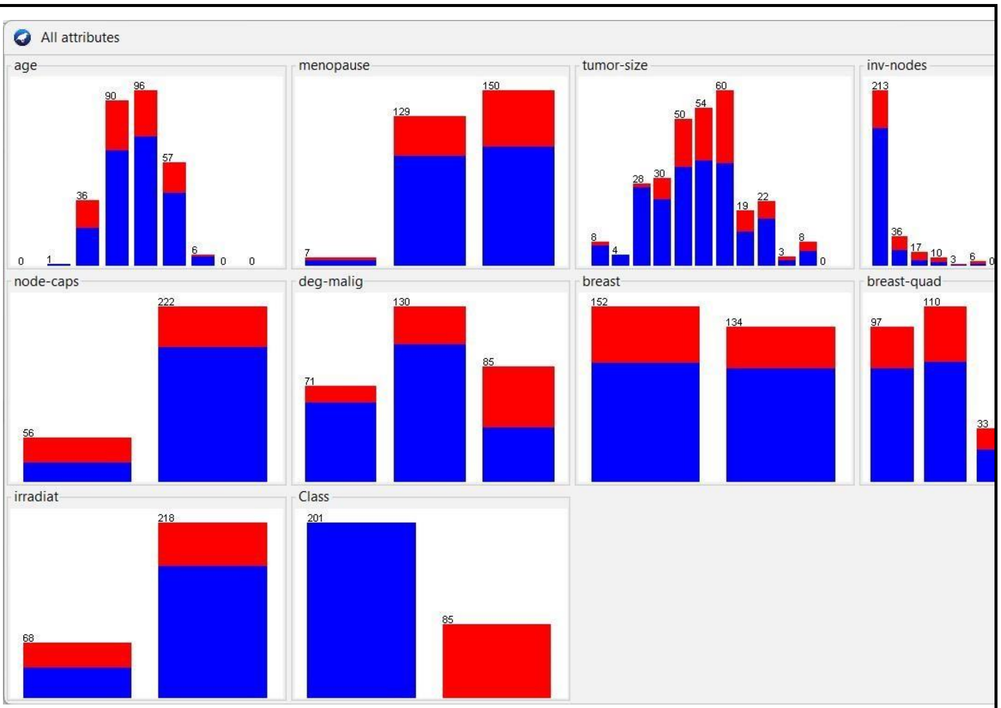
Attributes

| No. | | Name |
|---|---|---|
| 1 | □ | age |
| 2 | □ | menopause |
| 3 | □ | tumor-size |
| 4 | □ | inv-nodes |
| 5 | □ | node-caps |
| 6 | □ | deg-malig |
| 7 | □ | breast |
| 8 | □ | breast-quad |
| 9 | □ | irradiat |
| 10 | □ | Class |

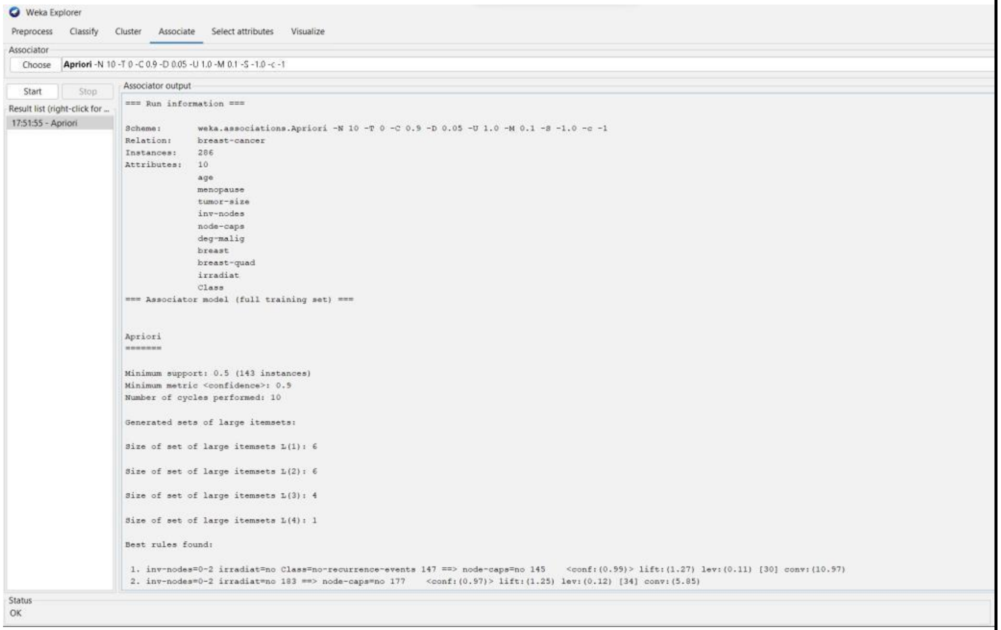
Status

OK

## PREPROCESSING DATA:



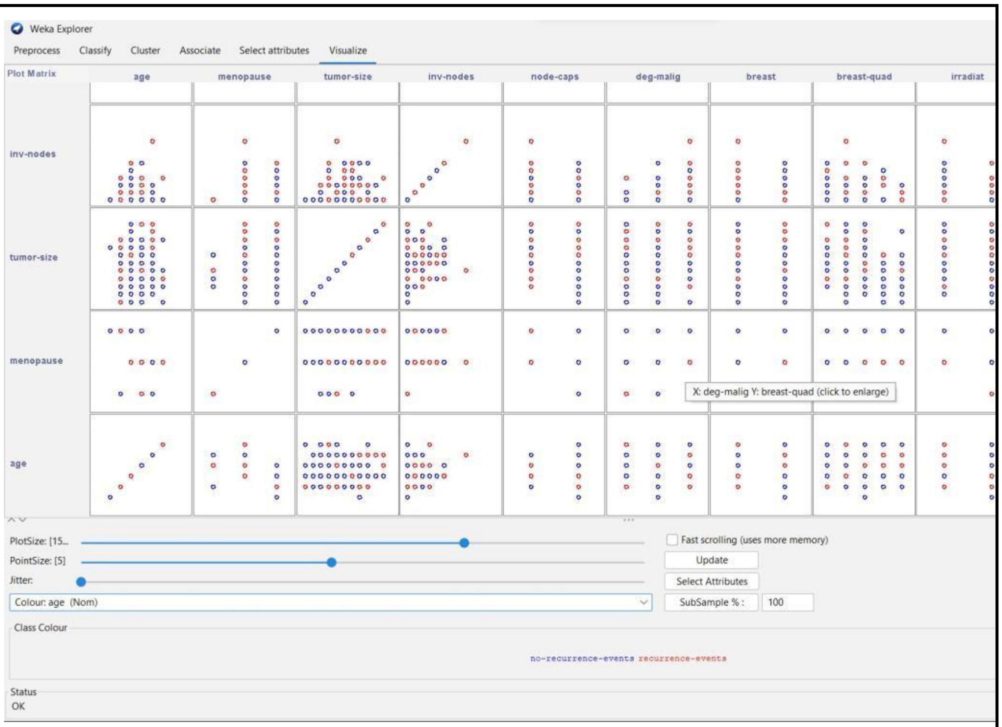
## Applying apriori algorithm:



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

```
=== Run information ===

Scheme:       weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     breast-cancer
Instances:    286
Attributes:   10
              age
              menopause
              tumor-size
              inv-nodes
              node-caps
              deg-malig
              breast
              breast-quad
              irradiat
              Class
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.5 (143 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 6

Size of set of large itemsets L(2): 6

Size of set of large itemsets L(3): 4

Size of set of large itemsets L(4): 1

Best rules found:

 1. inv-nodes=0-2 irradiat=no Class=no-recurrence-events 147 ==> node-caps=no 145    <conf:(0.99)> lift:(1.27) lev:(0.11) [30] conv:(10.97)
 2. inv-nodes=0-2 irradiat=no 183 ==> node-caps=no 177    <conf:(0.97)> lift:(1.25) lev:(0.12) [34] conv:(5.85)
```

## VISUALISING THE DATSET:

<table>
<tr><td></td><td>

PlotSize: [15...<br>PointSize: [5]<br>Jitter:<br>Colour: age (Nom)<br>□ Fast scrolling (uses more memory)<br>SubSample % : 100<br>Status<br>OK</td></tr>
<tr><td colspan="2"><b>Conclusion:</b> Through this experiment, I have gained knowledge and skills in using Weka for data visualization and association. I have become familiar with the Filtered Associator algorithm, which enables efficient exploration of large datasets. I have also explored the command-line interface (CLI) of Weka, which provides more flexibility for advanced users. Overall, I have gained a better understanding of data analysis and exploration using Weka.</td></tr>
</table>