Data-ViLiJ Application:
User Instructions Guide

Intended Audience: These instructions are intended for an audience with some background in computer science. The target audience consists of computer science students starting to learn about algorithms and data visualization. Professors can use this application in classes to teach students about how artificial intelligence algorithms can be used to aid programs in "learning" from data input. People with an interest in computer science may also find this application enlightening.

Application Purpose: Used to show how classification algorithms and clustering algorithms "learn" from a set of data and how these algorithms can be used to alter the data. The application is used to understand data visualization of data run with AI and machine learning algorithms.

Necessary Hardware: Keyboard, Mouse, and Monitor.

Operating Systems include: Windows 10, Mac OS X 10.11 or higher, Ubuntu 16.04 LTS or higher.

Necessary Software Requirements: Java Standard Edition Development Kit (Oracle JDK 8u161 or higher).

1. Run Application: Open in IntelliJ and run DataVisualizer.java class

2. Starting the Application: The application window should appear containing an empty chart and a toolbar with buttons. Some of the buttons will be disabled such as the save button and the run button since there is no data to be saved and there is no data for an algorithm to be run on. Some of the buttons will be enabled such as the exit button for exiting the application, the load button for loading data into the application from a file, and the new button for entering data into a text area inside the application itself. The image at the top of the next page shows what the initial window of the application should look like.
Note: Upon hovering over an enabled button in the toolbar, a tooltip appears describing the purpose of the button.

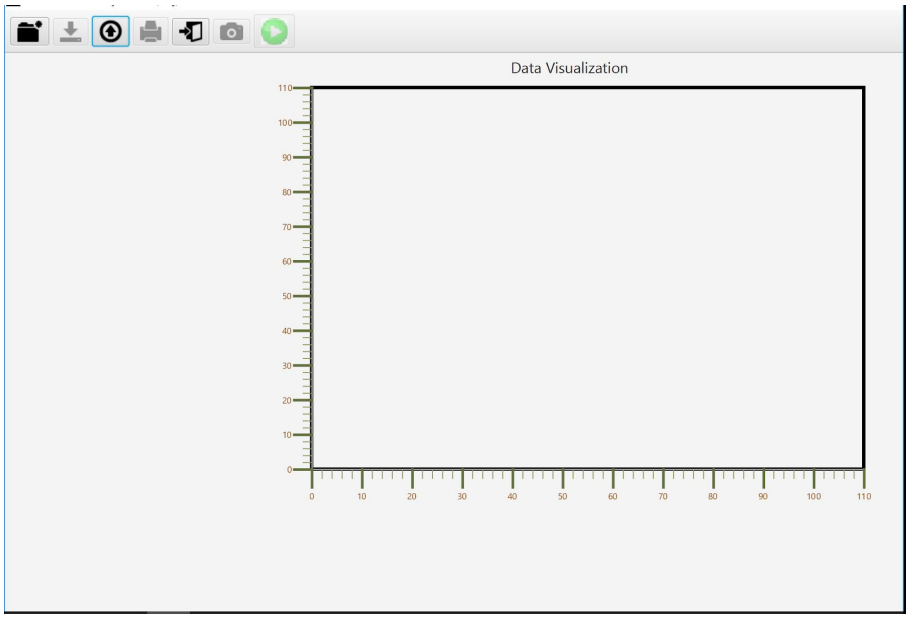

Figure 1: Primary Window

3. Creating Data to be Loaded into the Application:

i. The application only loads tab-separated data files with the file extension (*.tsd).

ii. The application only takes data instances with a certain format in the tab-separated data files.

a. A data instance starts with a string of characters representing the data name. The data name must begin with the symbol "@". There should not be multiple instances with the same name. The data name must then be followed by a tab.

b. The data instance must then contain a string of characters representing the data label. The data label must then be followed by a tab.

c. The data instance must then contain a number (an integer or a double) followed by a comma followed by a number (an integer or a double). The first number represents the x-value of the data point on the chart. The second number represents the y-value of the data point on the chart. Each data instance must end with a new line character.

d. An example of a data instance is:

"@instance1 label1 10,20'\n'".

iii. Open a tab-separated data file and add data instances to the file.

Figure below is opened in notepad. Optimal place to create tab-separated data files is in the IntelliJ IDEA IDE by selecting File in top left corner, selecting new and then selecting file that will be saved as a tab-separated data file.

```
@Instance1      label1  1.5,2.2
@Instance2      label1  1.8,3
@Instance3      null    2.1,2.9
@Instance4      label2  10,9.4
```

Figure 2: Example of a tab-separated data file

iv. In the application, press the load button in the toolbar. A file chooser will appear. Go to the location in the file chooser where the tab-separated file of data you created is saved. Select the file in the file chooser and click open in the bottom right corner.

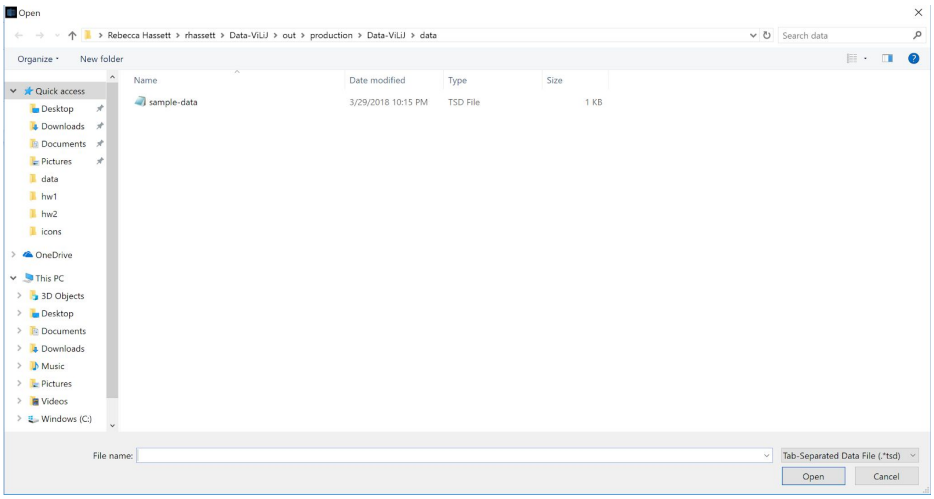

Figure 3: Loading File Chooser

v. If the file selected has any errors such as any instance names not starting with "@" or it is missing two point values separated by a comma for example, then an error dialog will appear indicating the line in the file where the error is or if there are duplicate instance names. If there is an error, then the data will not be loaded into the application and the text area, display button, edit button, done button, and data instance information will not appear in the application.

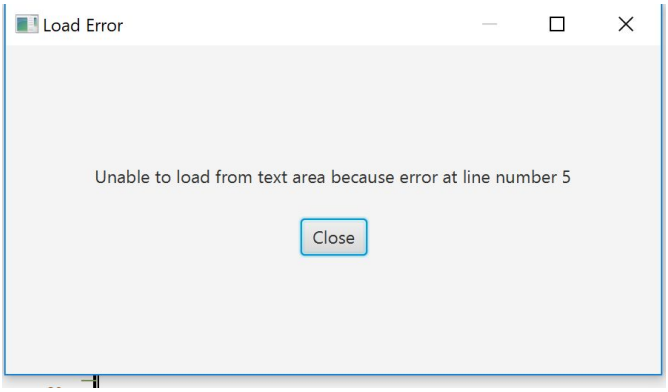

Figure 4: Error Message Loading Files

vi. If there are no errors, then a text area will display the data instances from the loaded file. The display button will appear as well as the edit button, the done button, information about the data instances loaded from the file, and a choicebox allowing a user to select a type of algorithm to run on the data, either classification or clustering. Once the data is loaded, the data cannot be edited in the text area because the text area is disabled. The information about data instances displayed includes the number of instances loaded, the number of instance labels, the names of all of the labels, and the source where the data instances came from. An example of information displayed could be: "4 instances with 3 labels loaded from: sample-data.tsd. The labels are: -null -label1 -label2"

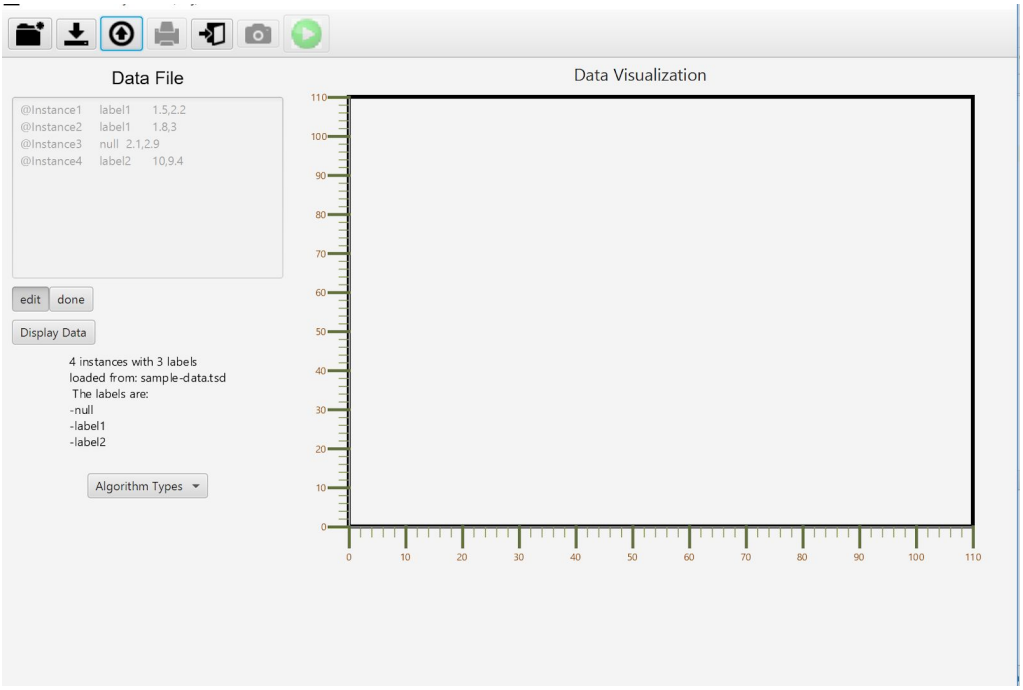

Figure 5: Result from Loading a File

vii. Null is a reserved label, different from all of the other labels. If you are planning on running a classification algorithm on the data, then make sure there are exactly two non-null labels in the data set.

4. Creating New Data in the Text Area:

i. Select the new button in the top left corner of the toolbar at the top of the application.

ii. A text area should appear enabled to enter data into the text area along with a toggle button including the options to either edit the data or complete verification of the data. The toggle will initially have selected the “edit” button. Enter data into the text area in the tab-separated data format as indicated for creating instances of data to be loaded into the text area in the prior step.

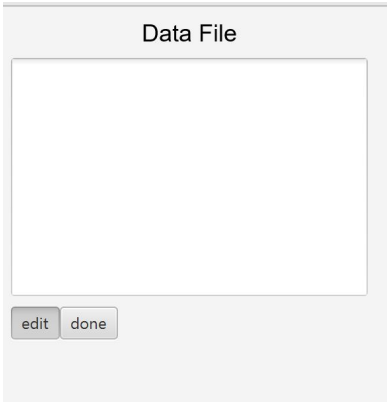

Figure 6: Creating New Data in the Text Area

iii. Once you complete entering data into the text area, select the “done” button to have the data format verified.

iv. If there is an error in the data format or there are duplicate instance names, then an error dialog will appear with the line number of the error or duplicate instance name.

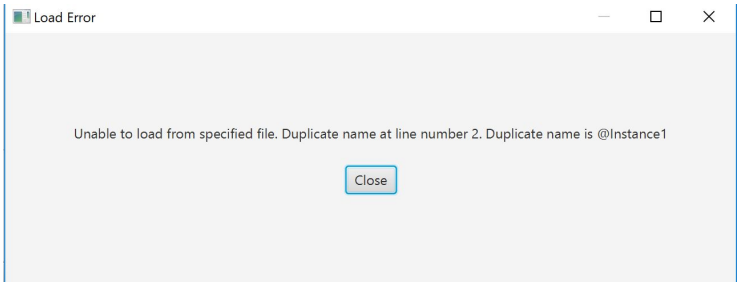

Figure 7: Error Dialog for Duplicate Instance Names

v. If the data format is correct, the text area will be disabled, and a display button will appear as well as a choice box to select the type of algorithm to run on the data. Information about the data instances entered will be displayed as well such as the number of data instances, the number of data labels, the source of the data which is the text area, and the names of the different data labels.

vi. If the user decides to further edit the data after clicking "done", the user can select the "edit" button. The text area will become enabled again, the information displayed about the data will disappear, the display button will disappear, and the choice box for the algorithm type will disappear as well. Once the user has completed entering the data, they can select "done" again to validate the data as indicated above in step three. The user may repeat these steps as many times as they wish.

vii. When the new button is selected, if there is newly created data already in the text area, then a file chooser will appear for the unsaved data before the text area is cleared to start entering completely new data to be saved to a new file. The filechooser will initially open to the data folder of the user's application, but the user can save the file anywhere in their computer. The user will be restricted to saving the data as a tab-separated data file with the (*.tsd) file extension. Once the user locates where they would like to save the data, they should name the data file and then click the "save" button in the bottom right corner of the file chooser.

5. Save Data Created in Text Area:

i. The save button is disabled while there is no created data in the text area (such as the data being loaded into the text area). The save button is also disabled if the data is the same since it was last saved. The save button will be disabled when the "edit" button is selected because the data has not been verified.

ii. If there is newly created data that has not been saved yet and the user has the "done" button selected which has verified the data, and the user has selected the save button in the toolbar, then if the data has already been saved to a location as a tab-separated data file, it will be saved to the same location. Otherwise, a file chooser will appear allowing the user to select where to save the file. The file chooser will initially open in the user's data folder in the application, but the user can save the data anywhere else on their computer. The file chooser will restrict the user to only saving the data as a tab-separated data file by restricting the file extension to (*.tsd). The user should enter a name for the data once they locate where to save the data, and then the user should select the "save" button in the bottom right corner of the file chooser.

6. Choosing the Algorithm Type for the Data:

i. If there is data loaded from a file into the text area or if there is newly created data in the text area verified by selecting the "done" button, then the user will have the ability to select the type of

algorithm to run on the data. The user can make this selection using the choice box in the center of the left panel of the application. The user can choose between “Classification” algorithms or “Clustering” algorithms.

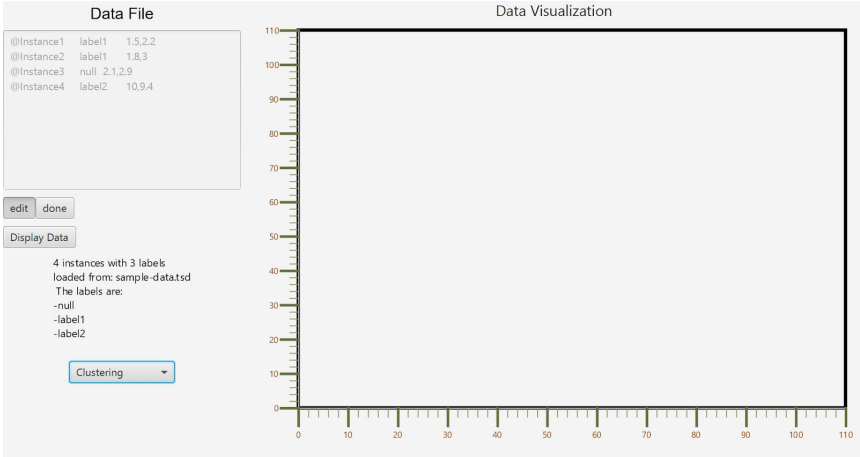

Figure 8: Selecting Algorithm Name in Highlighted Choice Box

ii. Users can choose clustering algorithms for any verified set of data. However, users cannot select classification algorithms for a set of data that does not have exactly two non-null data labels in the set of data. If the user has less than or more than two non-null data labels in the set of the data and the user selects “Classification” algorithms, then an error dialog will pop up telling the user that their data set does not have exactly two non-null labels so classification algorithms cannot be run on the set of data therefore the user must load new data with exactly two non-null labels or create data with exactly two non-null labels to run “Classification” algorithms on the data. The choice box will go back to not having an algorithm type selected in this case.

iii. If the user selects “Clustering” algorithms or “Classification” algorithms for data following the correct criteria, then the choice box will be replaced by a list containing specific algorithms of the chosen algorithm type. Each algorithm will have a configuration button next to it.

7. Choosing the Specific Algorithm for the Data:

i. Once the user determines which specific algorithm to run on they data, they must select it from the list by clicking on the algorithm name.

8. Setting the Run Configuration of the Algorithm:

i. The user will not be able to run the algorithm they have selected for the data until they have set the run configuration for that particular algorithm. The user can set the run configuration for the algorithm by selecting the configuration button directly next to the algorithm name they have chosen or wish to choose in the list.

Figure 9: Algorithm Configuration Button Symbol

ii. Once the configuration button is selected, a window titled “Algorithm Run Configuration” will appear. For classification algorithms, the user will be expected to enter the maximum number of iterations of the algorithm in a text field, the update intervals for the algorithm in a text field, and whether or not the algorithm runs continuously by checking the check box if it runs continuously or not checking the check box if it does not run continuously. For clustering algorithms, the user will be expected to enter the maximum number of iterations of the algorithm in a text field, the update intervals for the algorithm in a text field, the number of labels of the data in a text field, and whether or not the algorithm runs continuously by checking the check box if it runs continuously or not checking the check box if it does not run continuously.

iii. Once the user has entered all of the parameters for the algorithm, the user must select the save button at the bottom of the run configuration window to save the run parameters for that algorithm.

iv. Once the run configuration has been filled in once, this specific run configuration becomes the default run configuration for that algorithm for future runs of the algorithm until the user decides to change the run configuration by selecting the run configuration button for that algorithm, changing the settings, and then selecting the save button at the bottom of the run configuration window again.

v. If the user enters unreasonable values such as a negative value for the maximum number of iterations or update intervals, then the negative value will be changed to a zero and this will be reflected in the run configuration window.

9. Running the Algorithm on the Data:

i. The run button is enabled once a specific algorithm is chosen in the list of algorithms.

Figure 10: Run Button, Rightmost Button in Toolbar

ii. If the run button is clicked without setting up the run configuration for that algorithm, then an error dialog will appear asking the user to set the run configuration of the chosen algorithm before running the algorithm on the data. The algorithm will not run on the data if this error dialog appears.

iii. If the run button is selected and the run configuration settings are set for the algorithm, then the algorithm will run on the data.

iv. If continuous run is chosen for the algorithm, then the algorithm will run for the maximum number of iterations specified. If the maximum number of iterations is more than the number of iterations the algorithm can run for, then the algorithm will run to completion. The chart display will start by displaying the initial data in the text area. The chart display will update with the altered data during each update interval as specified by the user. The maximum iterations will run until completion without user intervention. The only action the user can perform during the algorithm’s running is exiting the application. The run button is disabled during the entire run time of the algorithm.

v. If continuous run is not chosen for the algorithm, then the algorithm will still run for the maximum number of iterations specified whether or not that is the entire number of iterations of the algorithm. The chart display will start by displaying the initial data in the text area. However, during each update interval that was specified by the user, the display will update with changes made to the data and the algorithm will pause running until the user resumes the algorithm running by clicking the run button. The run button is enabled whenever the display chart updates and the algorithm is paused. The run button is disabled while the algorithm is running. When the algorithm is paused, the user has

the option to save data, terminate the running algorithm, take a screenshot of the chart that was updated, or exit the application if the user wishes to. The user can exit the application at any point during the running of the algorithm.
Note: Even though the algorithm updates the data and changes the chart in the application, the data in the text area is not updated so the user will not lose their
original data for future runs of the algorithm.

10. Exporting Chart as Image:

i. The screenshot button in the toolbar is disabled as long as the chart in the application is empty. The screenshot button has the icon of a camera. While the algorithm is running, regardless of whether or not it runs continuously, the screenshot button is disabled. The screenshot button is enabled whenever the chart is not empty and an algorithm is not running. The screenshot button is enabled during pauses between the algorithm running.

Figure 11: Disabled Screenshot Button, Second to Rightmost Button

ii. Once the screenshot button is selected, a filechooser will appear with the default location set as the data folder in the application. The user can save the image of the chart anywhere in their computer. The screenshot of the chart is restricted to being saved as an image file. The user must create a name for the image, and then click the save button at the bottom right corner of the file chooser.

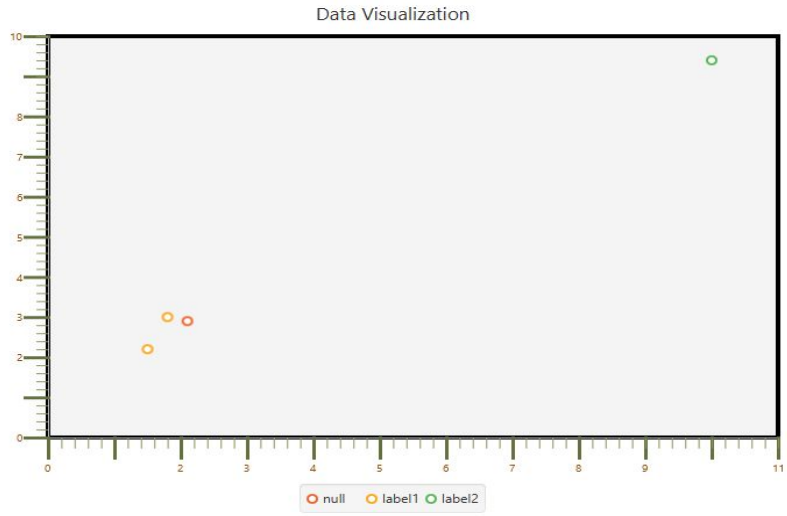

Figure 12: Example Screenshot of Chart Area

11. Displaying Valid Data:

i. The display button appears whenever there is valid data in the text area, and the user is not running an algorithm on the data.

ii. If the user selects the display button. The instances will appear in the chart at their given x value and y value location. The instances will have their labels listed with their given colors below the chart. The colors given for a label may not necessarily match the label name since label names do not

have to be colors. The color of an instance in the chart will match the color of its label. When the mouse is hovered over a data instance, the data instance's name will appear in a tooltip above the mouse and label so the user knows which instance is which in the chart.

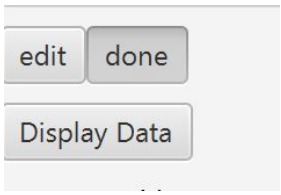

Figure 13: Display Button Underneath Edit and Done Toggle Buttons

12. Selecting a New Algorithm to Run on the Data:

i. If the user has selected an algorithm type or specific algorithm or has completed running an algorithm on a set of data, then the user may select the new algorithm button to choose a new algorithm to run on the data. This button only appears when there is valid data in the text area. It is disabled during algorithm runs.

ii. Once the new algorithm button is selected, the screen will show the choice box with the two different types of algorithms again to be chosen from.

13. Exiting the Application:

i. The user can exit the application at any point of running the application, even if there is an algorithm running on the data or there is no valid data in the text area. The user can exit the application by selecting the exit button in the toolbar of the application.

Figure 14: Exit Button, Third to Rightmost Button Containing Open Door and Arrow

ii. If the user selects the exit button while there is unsaved created data in the text area, a window will appear asking the user whether they want to exit the application without saving the data, return to the application without exiting, or saving the data and exiting. If the user selects to save the data and exit, the file chooser will appear allowing the user to select the location to save the data in if the data has not already been saved to a file. If the data has already been saved to a file, the file chooser will not appear, and the data will be saved to the file it was saved to prior before exiting the application.

iii. If an algorithm is running when the user selects the exit button, then a window will appear asking the user whether they want to return to the application or terminate the running algorithm and exit the application.

iv. If the user has an algorithm running and unsaved data, then the user will be asked whether they want to return to the application or terminate algorithm and exit first. If the user selects terminate and exit, then the user will be asked whether they want to save the data before exiting, exit without saving or return to the application.

Note: Print Button functionality for the application has not been included yet. Resizing the application window has not been implemented yet either.

How Can A Computer Science Student Benefit From This Application:

Take advantage of the functionalities of this application! Instead of running the algorithms continuously and quicker, have the algorithm pause at regular intervals to analyze how each algorithm has affected the data input into the application. Take screenshots of the chart to analyze the changes made to the data closely. Think of algorithms you can write that could be used in this application.

Resources:

Classification Algorithm Information:

Used for categorizing data into one of two categories that are known because these categories are specified by the labels given to the data during their input to the application. A specific classification algorithm can learn by separating the two labels of the data by drawing a straight line through the x-y 2-dimensional plane of data separating the data labels based on their positions. The algorithm attempts to separate the data even if it is not possible. Classification algorithms output a line and update the chart accordingly with a line. The classification algorithms do not alter instance positions nor instance labels.

Clustering Algorithm Information:

Clustering algorithms find patterns about data depending on how they are distributed in space. Clustering algorithms do not need to know specific instance labels. Clustering algorithms look at the position of instances. Clustering algorithms need to know the total number of labels to run. The clustering algorithms update the labels of instances depending on their positions.