Name : Kaushik Kotian **Roll no.** 30 **Div :** D15B **Batch:**B

Experiment no.:8

AIM: To implement clustering using rapid miner
Dataset used :
https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering/data

Theory:

RapidMiner is a powerful and user-friendly data science platform that facilitates various aspects of the data mining process, including data preparation, machine learning, predictive modeling, and deployment. It provides a graphical user interface for building analytical workflows without the need for extensive programming knowledge, making it accessible to a wide range of users, including data scientists, analysts, and business professionals.

RapidMiner and clustering using RapidMiner:

1. Data Preparation:
- RapidMiner allows users to import data from various sources such as databases, spreadsheets, and web services.
- It offers numerous data preprocessing tools for cleaning, transforming, and formatting data to make it suitable for analysis.

2. Visualization:
- RapidMiner provides visualization tools to explore and understand data distributions, relationships, and patterns.
- Users can generate various types of charts, graphs, and plots to gain insights into their data.

3. Clustering:
- Clustering is an unsupervised machine learning technique used to group similar data points together based on their characteristics.
- In RapidMiner, clustering algorithms such as k-means, hierarchical clustering, and DBSCAN are available.
- Users can apply these algorithms to discover natural groupings within their data and identify patterns or trends.

4. Workflow Design:

- RapidMiner allows users to create analytical workflows by connecting various data processing and modeling operators.
- Workflows are designed graphically using a drag-and-drop interface, making it easy to experiment with different techniques and configurations.

5. Model Evaluation:

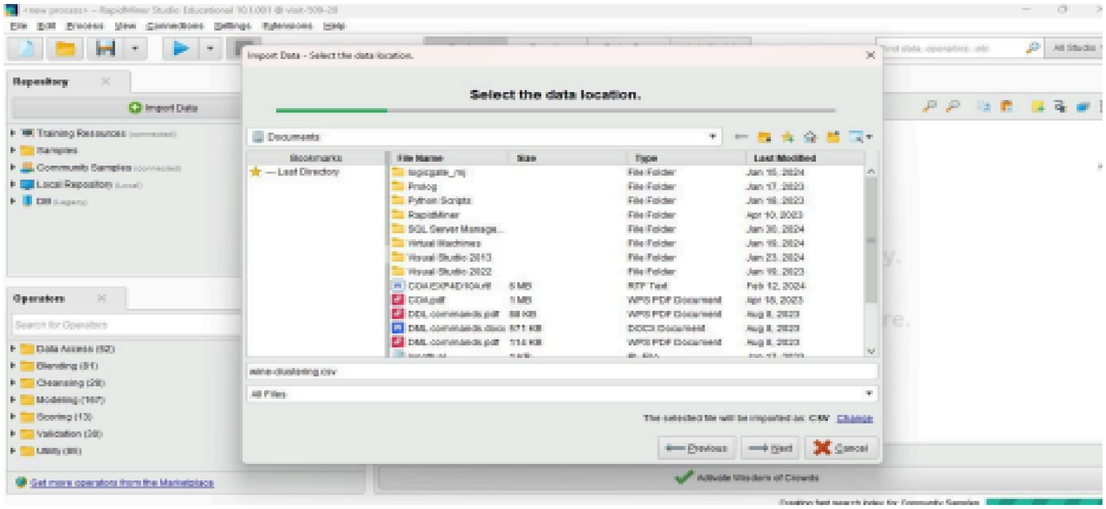
- RapidMiner provides tools for evaluating the performance of clustering models.
- Users can assess the quality of clusters using metrics such as silhouette score, Davies–Bouldin index, and within-cluster sum of squares.

6. Deployment and Integration:

- Once a satisfactory clustering model is built, it can be deployed within RapidMiner or integrated into other systems through APIs or export options. - RapidMiner supports deployment to various environments, including cloud platforms and production servers.

Step 1:

Firstly, we will load our dataset for data visualization or exploration. Visualizing data is crucial for understanding its distribution, spotting outliers, and identifying patterns or anomalies that might require further investigation. This process helps in formulating hypotheses and deciding on the appropriate analytical techniques to apply.



<new process> – RapidMiner Studio Educational 10.1.001 @ vish-509-28
File Edit Process View Connections Settings Extensions Help
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
DB (Legacy)
Operators
Search for Operators
Data Access (52)
Blending (81)
Cleansing (28)
Modeling (167)
Scoring (13)
Validation (28)
Utility (85)
Get more operators from the Marketplace
Activate Wisdom of Crowds

Import Data - Select the data location.

Select the data location.

Documents

Bookmarks: Last Directory

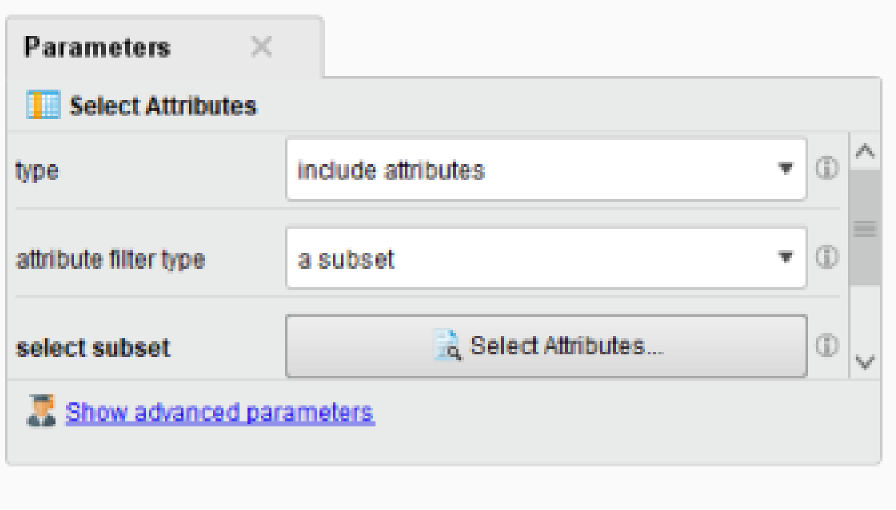
File Name	Size	Type	Last Modified
logicgate_mj		File Folder	Jan 15, 2024
Prolog		File Folder	Jan 17, 2023
Python Scripts		File Folder	Jan 16, 2023
RapidMiner		File Folder	Apr 10, 2023
SQL Server Manage...		File Folder	Jan 30, 2024
Virtual Machines		File Folder	Jan 19, 2024
Visual Studio 2013		File Folder	Jan 23, 2024
Visual Studio 2022		File Folder	Jan 19, 2023
CD4EXP4D10A.rtf	5 MB	RTF Text	Feb 12, 2024
CD4.pdf	1 MB	WPS PDF Document	Apr 18, 2023
DDL commands.pdf	88 KB	WPS PDF Document	Aug 8, 2023
DML commands.docx	571 KB	DOCX Document	Aug 8, 2023
DML commands.pdf	114 KB	WPS PDF Document	Aug 8, 2023

wine-clustering.csv

All Files

The selected file will be imported as: CSV Change

Step 2: Now we will select and filter the dataset based on specific criteria to focus the analysis on relevant data points.



Parameters

Select Attributes

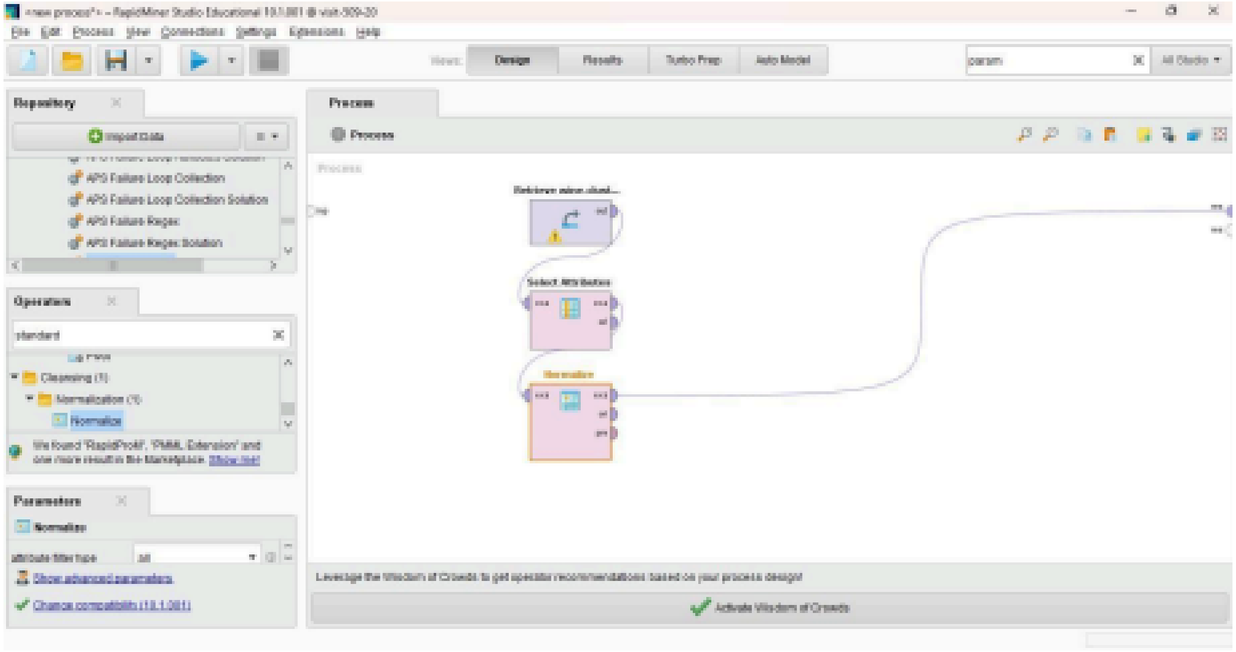
type: include attributes

attribute filter type: a subset

select subset: Select Attributes...

Show advanced parameters

Step 3: Here we have performed analysis on the dataset. The process includes operators for retrieving data, selecting specific data attributes, and normalizing the data for consistency. The user interface displays various tools and panels for organizing data, building analysis processes, and adjusting parameters for the selected operators.



Parameters

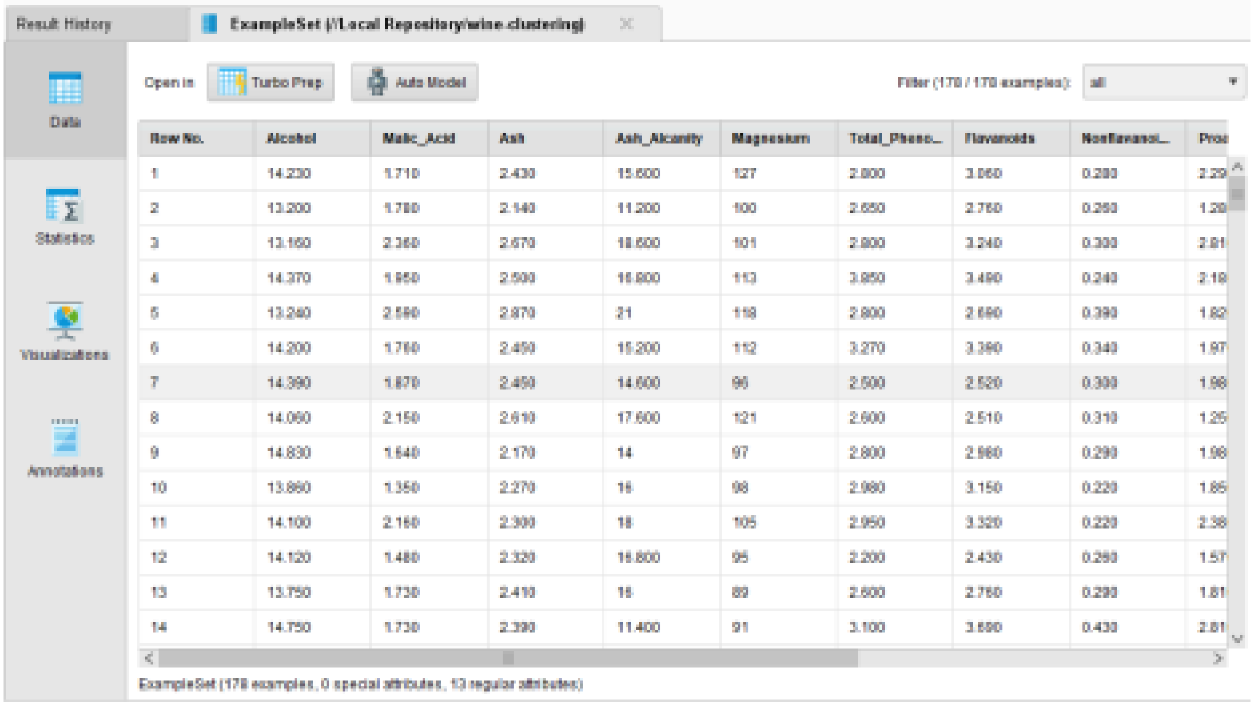
Normalize

attribute filter type: all

Show advanced parameters

Change compatibility (10.1.001)

Step 4: Here the result is displayed which captures the process of applying a specific data mining technique, evaluating its performance through metrics such as accuracy and precision.



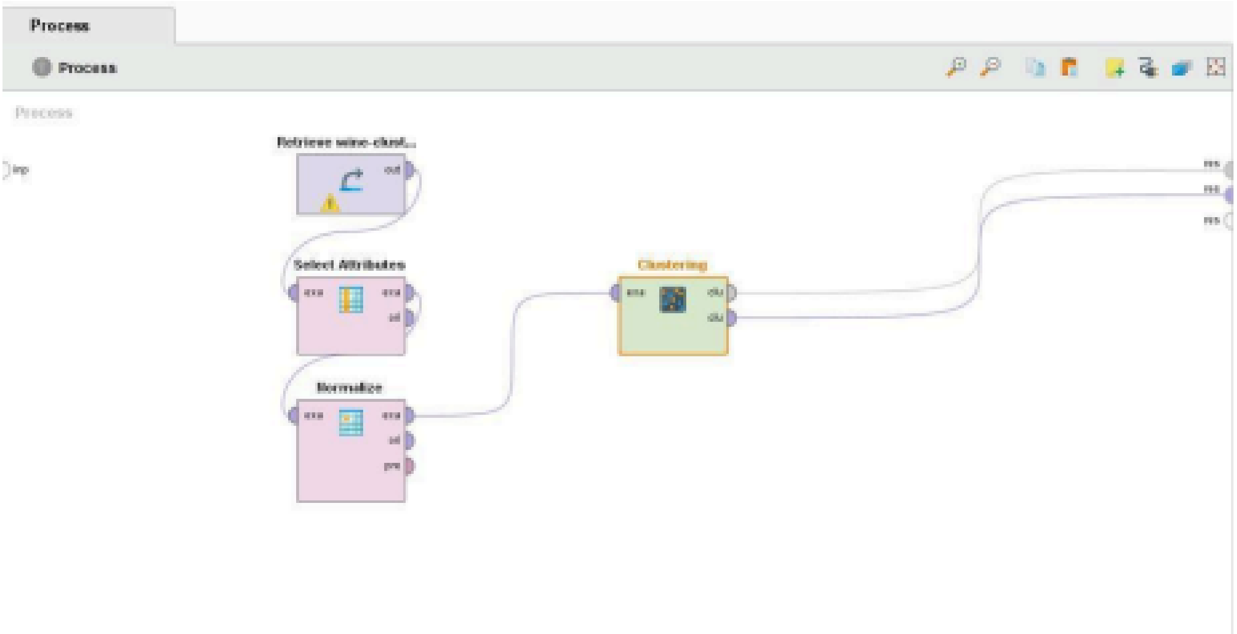
ExampleSet (//Local Repository/wine-clustering)

Filter (178 / 178 examples): all

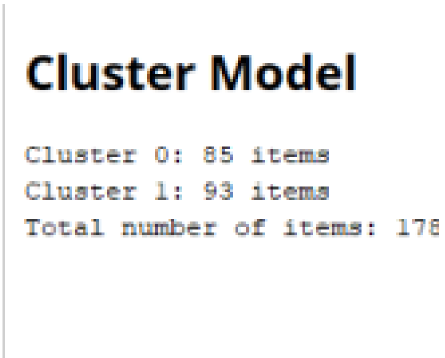
Row No.	Alcohol	Malic_Acid	Ash	Ash_Alcanity	Magnesium	Total_Pheno...	Flavanoids	Nonflavanoi...	Proa
1	14.230	1.710	2.430	15.600	127	2.800	3.060	0.280	2.29
2	13.200	1.780	2.140	11.200	100	2.650	2.760	0.260	1.28
3	13.160	2.360	2.670	18.600	101	2.800	3.240	0.300	2.81
4	14.370	1.950	2.500	16.800	113	3.850	3.490	0.240	2.18
5	13.240	2.590	2.870	21	118	2.800	2.690	0.390	1.82
6	14.200	1.760	2.450	15.200	112	3.270	3.390	0.340	1.97
7	14.390	1.870	2.450	14.600	96	2.500	2.520	0.300	1.98
8	14.060	2.150	2.610	17.600	121	2.600	2.510	0.310	1.25
9	14.830	1.640	2.170	14	97	2.800	2.980	0.290	1.98
10	13.860	1.350	2.270	16	98	2.980	3.150	0.220	1.85
11	14.100	2.160	2.300	18	105	2.950	3.320	0.220	2.38
12	14.120	1.480	2.320	16.800	95	2.200	2.430	0.260	1.57
13	13.750	1.730	2.410	16	89	2.600	2.760	0.290	1.81
14	14.750	1.730	2.390	11.400	91	3.100	3.690	0.430	2.81

ExampleSet (178 examples, 0 special attributes, 13 regular attributes)

Step 5: We have majorly performed 3 steps here: retrieving the wine dataset, selecting relevant attributes, and normalizing the data, followed by a clustering operation to group similar data points together.



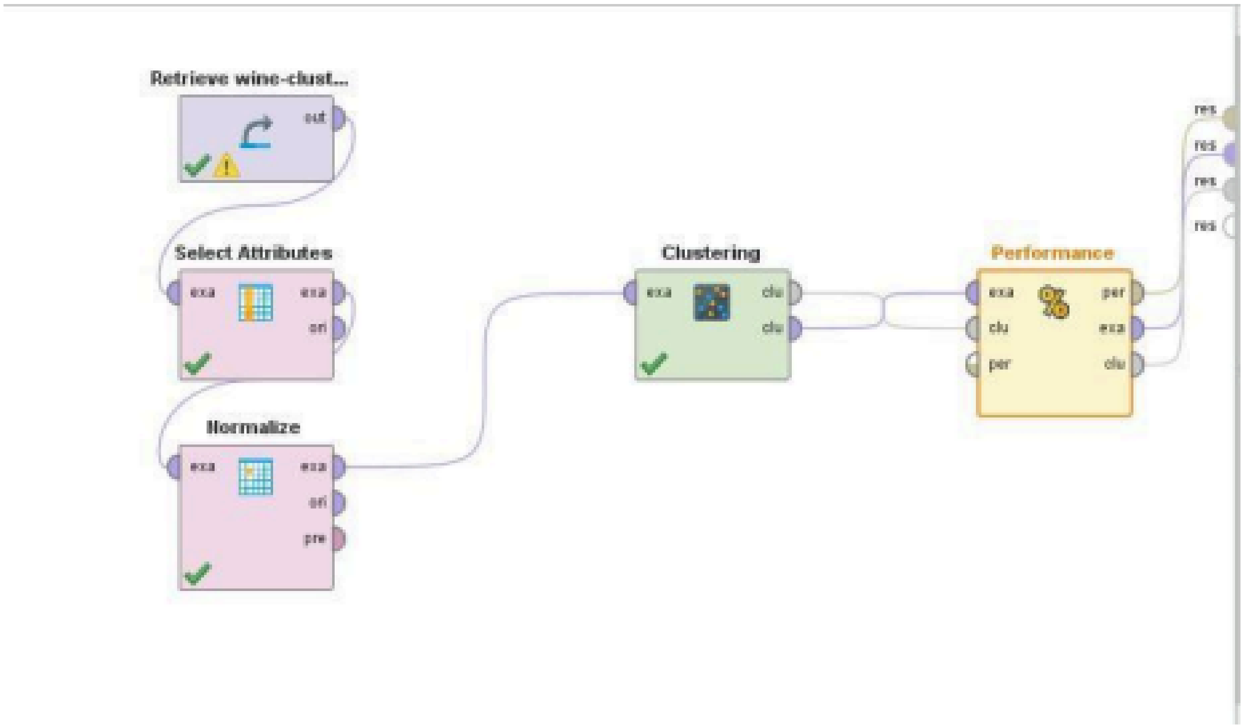
Step 6: Here is the result of a clustering model with two clusters: Cluster 0 contains 85 items, and Cluster 1 contains 93 items, making a total of 178 items classified into two distinct groups.



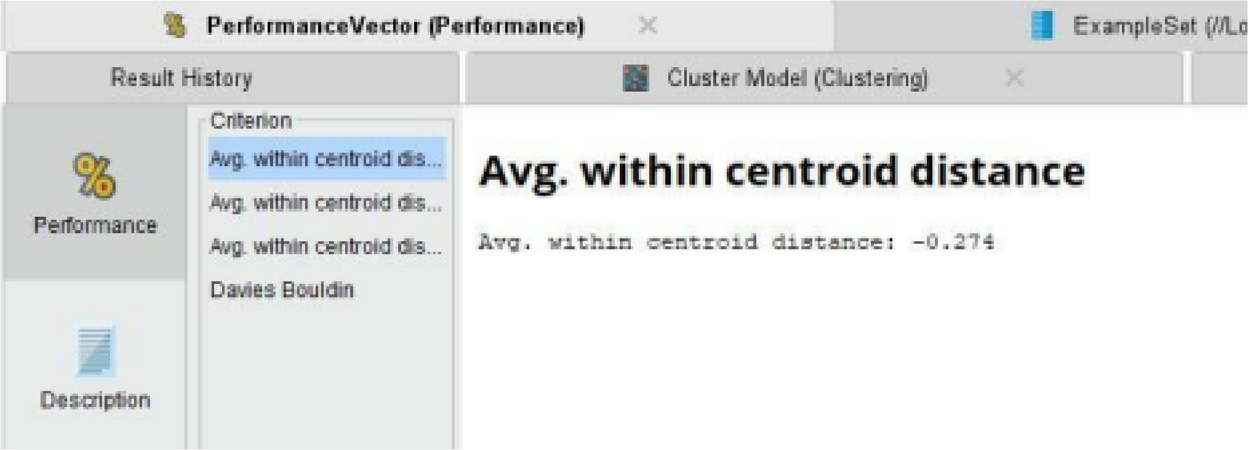
Cluster Model

```
Cluster 0: 85 items
Cluster 1: 93 items
Total number of items: 178
```

Step 7: Here, data is first retrieved, attributes are selected, and the data is normalized, followed by a clustering operation. The workflow then proceeds to a performance evaluation step, indicating that the clustering results are being assessed for their quality or effectiveness.

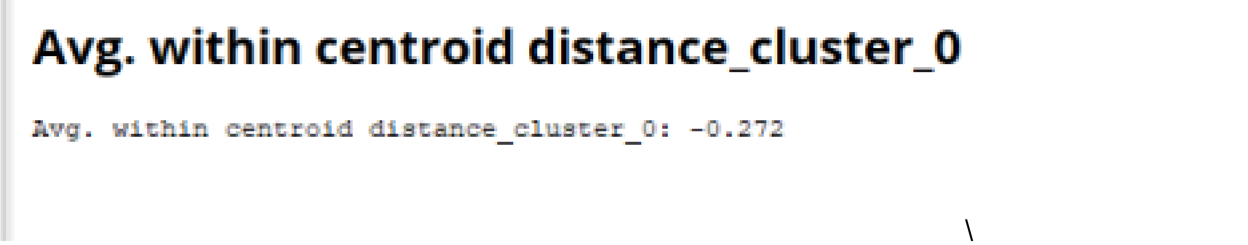


Step 8: Here is the output from a clustering analysis, specifically showing the average within-centroid distance for two clusters. The values -0.274 for cluster 0 and -0.276 for cluster 1 indicate the compactness of the clusters, with lower values typically representing better clustering where items are closer to their respective centroids.


PerformanceVector (Performance)
ExampleSet (//Lo
Result History
Cluster Model (Clustering)
Performance
Description
Criterion
Avg. within centroid dis...
Avg. within centroid dis...
Avg. within centroid dis...
Davies Bouldin

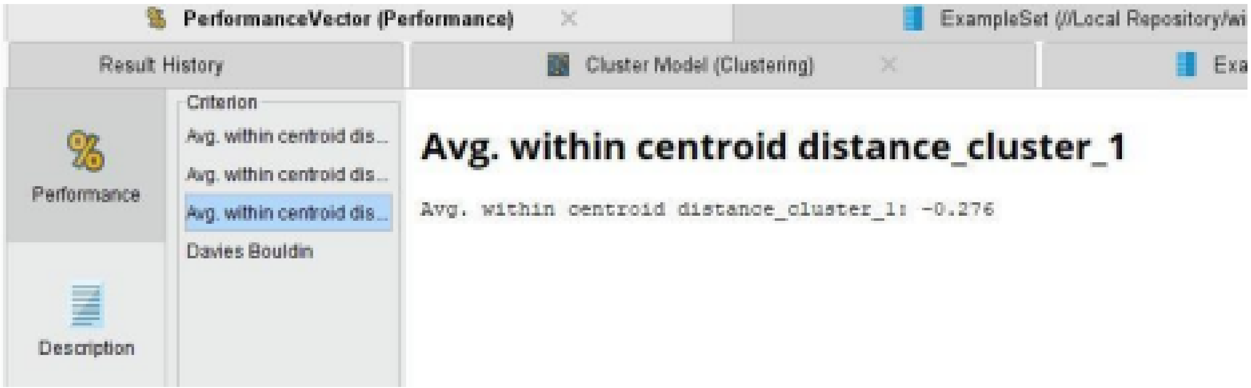
Avg. within centroid distance

```
Avg. within centroid distance: -0.274
```



Avg. within centroid distance_cluster_0

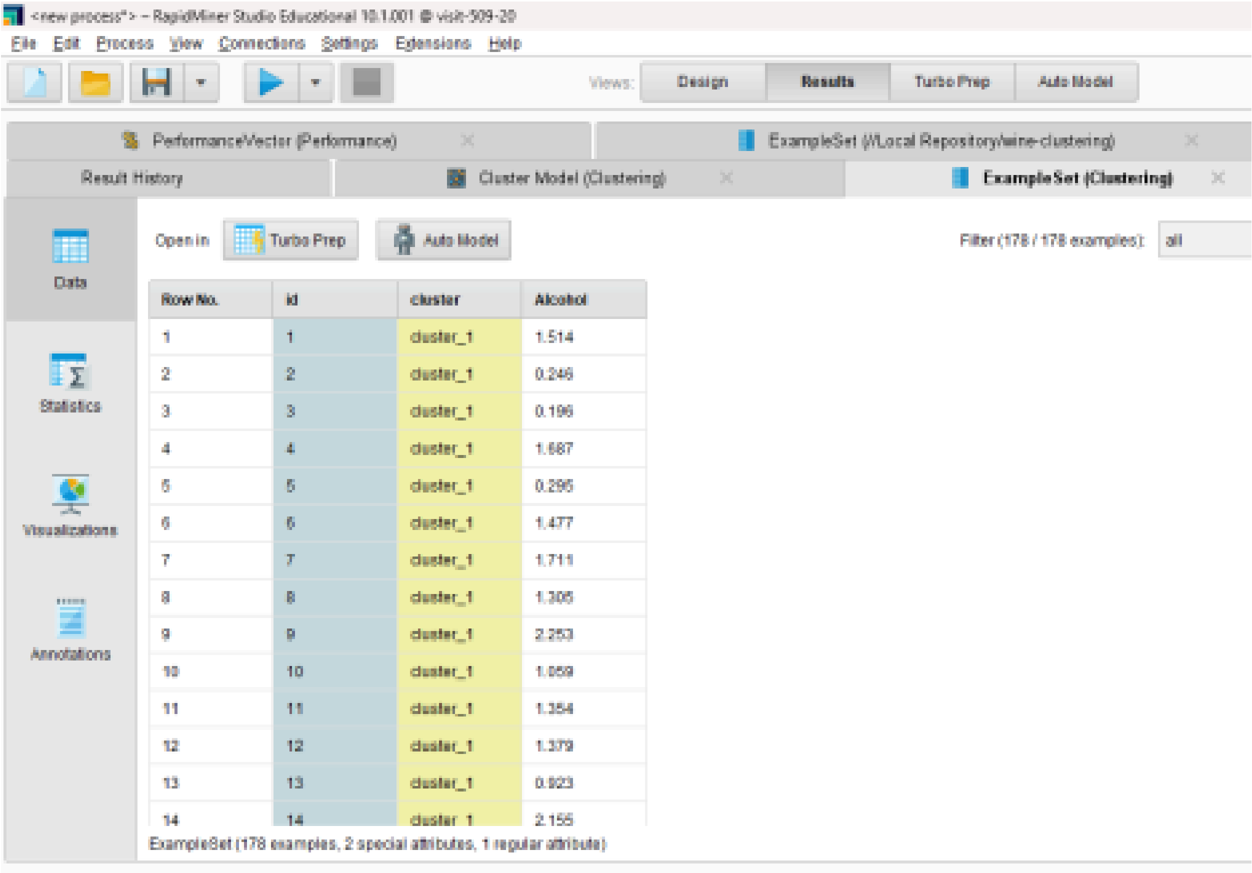
```
Avg. within centroid distance_cluster_0: -0.272
```


PerformanceVector (Performance)
ExampleSet (//Local Repository/wi
Result History
Cluster Model (Clustering)
Exa
Performance
Description
Criterion
Avg. within centroid dis...
Avg. within centroid dis...
Avg. within centroid dis...
Davies Bouldin

Avg. within centroid distance_cluster_1

```
Avg. within centroid distance_cluster_1: -0.276
```

Step 9: The image shows a table from RapidMiner Studio with clustering results of a wine dataset. Each row lists a wine's ID, its assigned cluster (cluster_1), and the normalized value of its alcohol content. The table is part of the software's Results view, indicating the data has been processed and is ready for further analysis or review.


<new process*> – RapidMiner Studio Educational 10.1.001 @ visit-509-20
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
PerformanceVector (Performance)
ExampleSet (//Local Repository/wine-clustering)
Result History
Cluster Model (Clustering)
ExampleSet (Clustering)
Data
Statistics
Visualizations
Annotations
Open in Turbo Prep Auto Model
Filter (178 / 178 examples): all

Row No.	id	cluster	Alcohol
1	1	cluster_1	1.514
2	2	cluster_1	0.246
3	3	cluster_1	0.196
4	4	cluster_1	1.687
5	5	cluster_1	0.295
6	6	cluster_1	1.477
7	7	cluster_1	1.711
8	8	cluster_1	1.305
9	9	cluster_1	2.253
10	10	cluster_1	1.059
11	11	cluster_1	1.354
12	12	cluster_1	1.379
13	13	cluster_1	0.923
14	14	cluster_1	2.155

ExampleSet (178 examples, 2 special attributes, 1 regular attribute)

Conclusion : We observed the use of RapidMiner Studio for clustering a wine dataset. The process involved data retrieval, selection of attributes, normalization, and the application of a clustering algorithm. Finally, the results were evaluated for cluster cohesion, and the detailed cluster assignments for each wine example were examined.