MODULE 10

CLUSTERING : k-Means

Sort by :

Hamzah L200154013

\

Informatics Study Program

Faculty of Communication and Informatics

Muhammadiyah University of Surakarta

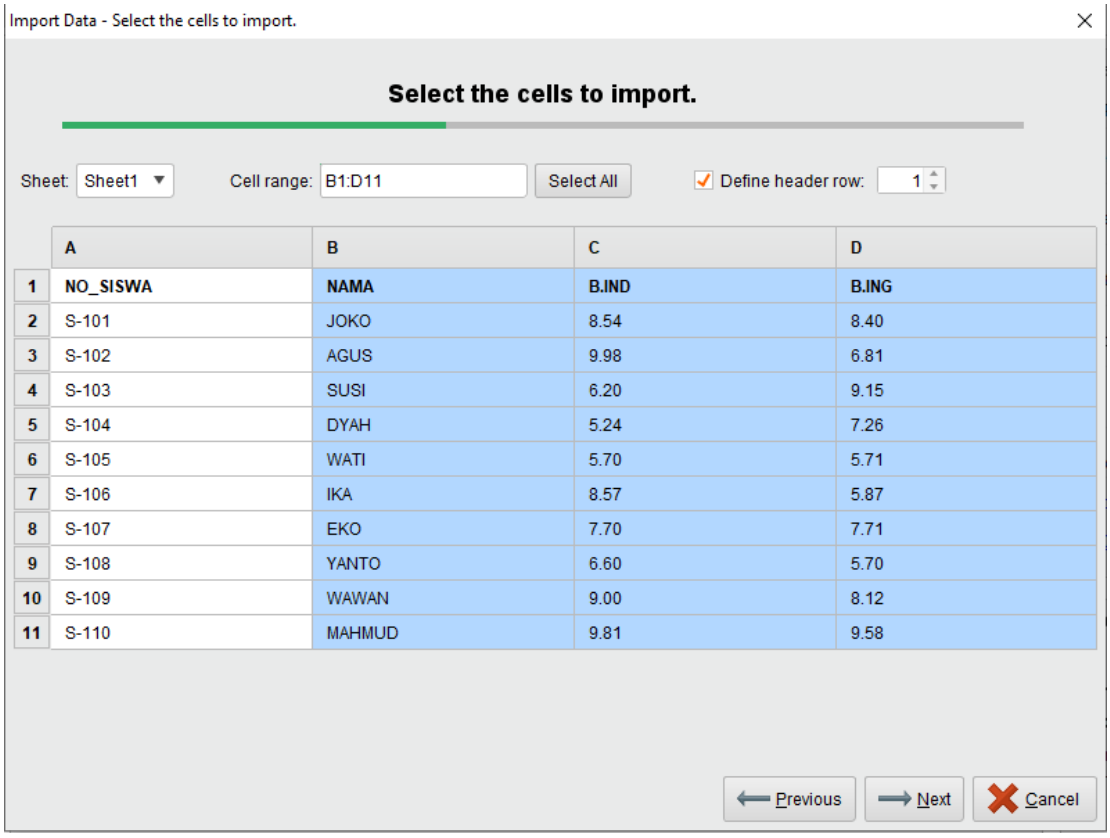
Praktikum Steps

K-Means algorithm using RapidMiner.

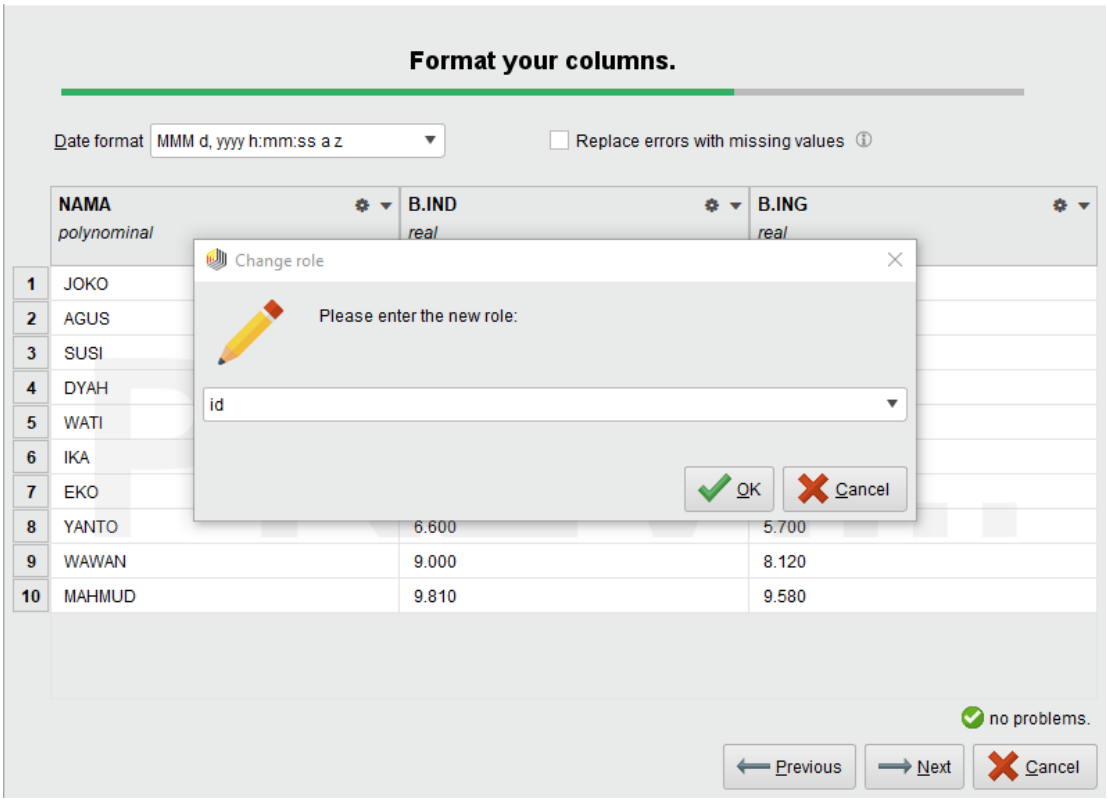
1. Open MS. Excel, and make table data score of students. Save with the name **Tabel_Ujian.xls** (Excel format 2013 *xls).

	A	B	C	D
1	NO_SISWA	NAMA	B.IND	B.ING
2	S-101	JOKO	8.54	8.40
3	S-102	AGUS	9.98	6.81
4	S-103	SUSI	6.20	9.15
5	S-104	DYAH	5.24	7.26
6	S-105	WATI	5.70	5.71
7	S-106	IKA	8.57	5.87
8	S-107	EKO	7.70	7.71
9	S-108	YANTO	6.60	5.70
10	S-109	WAWAN	9.00	8.12
11	S-110	MAHMUD	9.81	9.58

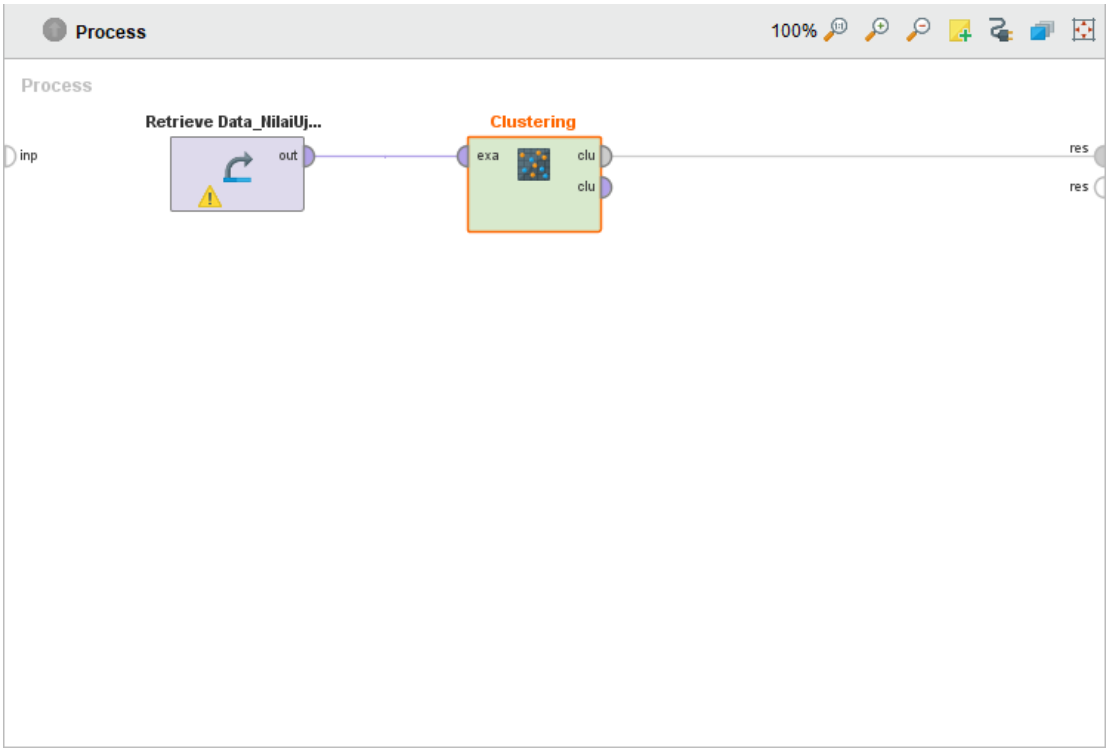
2. Open RapidMiner application and import the excel file, in cells selection only select 3 columns (name of students, score of Indonesian language and English language).

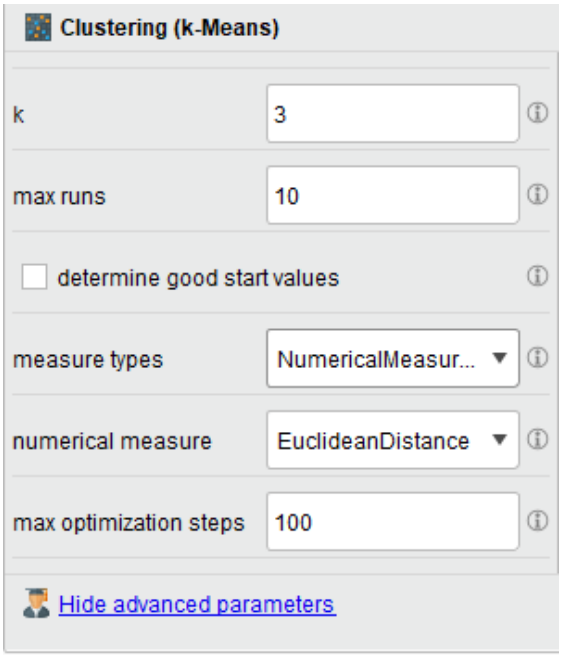


3. Change the role of column name of student with **id**, and save the file with name **Data_NilaiUjian**.

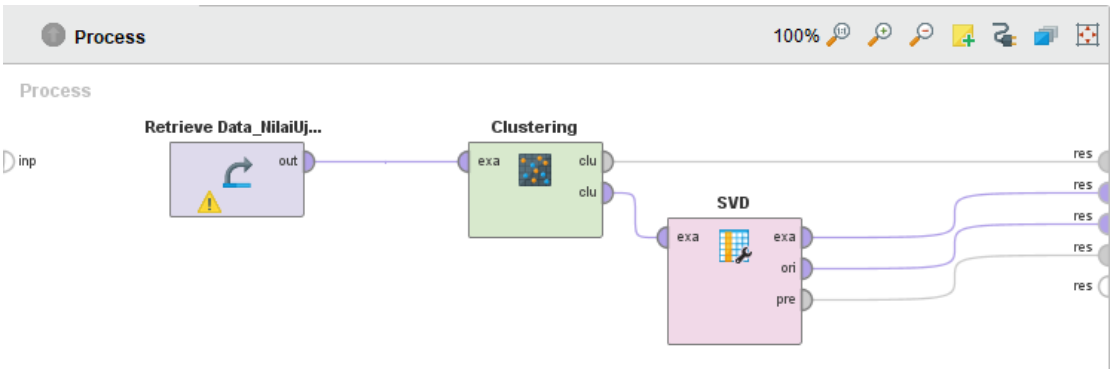


4. Drag Data_NilaiUjian into the process area, then add k-Means operator and connect the port with the retrieve data and res panel. Double click in operator k-Means and change the parameter value into 3 and measures type to NumericalMeasures.





5. Add another operator called SVD (Singular Value Decomposition) and connect operator k-Means output into SVD, then 3 SVD output into the res panel.



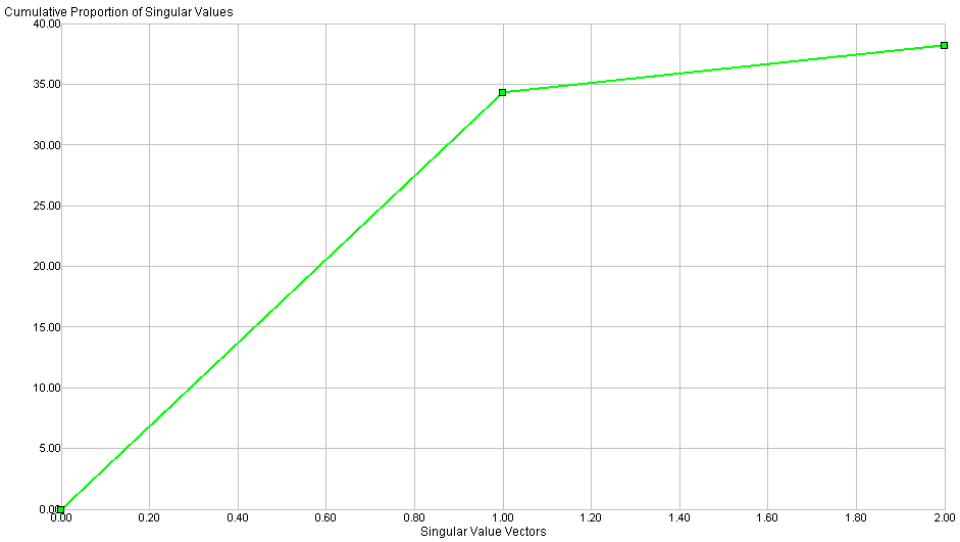
6. Run the process, and this is the result of clustering using k-Means algorithm.
   a) SVD (Singular Value Decomposition)
      i. Eigenvalue

Component	Singular Value	Proportion of Singular Values	Cumulative Singular Values	Cumulative Proportion of Sin...
SVD 1	34.340	0.898	34.340	0.898
SVD 2	3.906	0.102	38.246	1.000

      ii. SVD vectors value

Attribute	SVD Vector 1
B.IND	0.723
B.ING	0.690

iii. Cumulative variance value

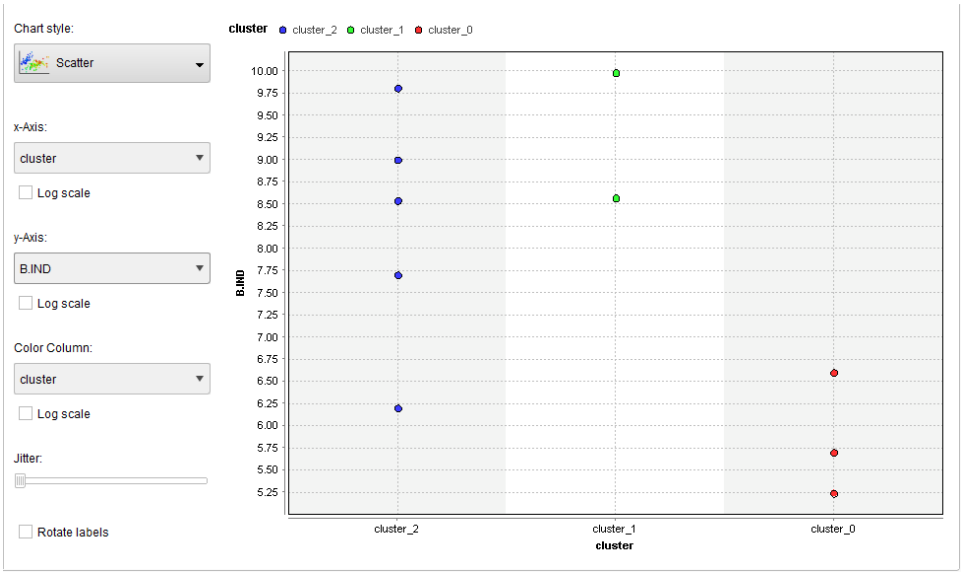


b) ExampleSet (K-Means)

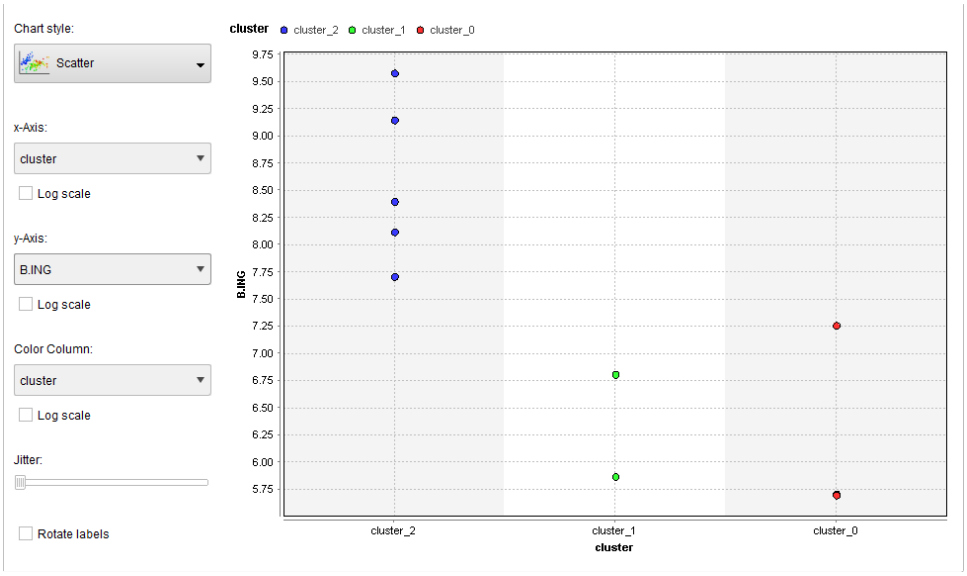
This result we see with Plot View using Scatter model graphic to decided what cluster of student that will be the candidate for Olympic objects based on their highest score in their exam.

x-Axis = cluster
y-Axis = B.IND, B.ING
Color Column = cluster
Jitter = can be change to see more detailed distribution of the data

i. Group of Student with Indonesian Language Subject



ii. Group of Student with English Language Subject



c) ExampleSet (SVD)

See with data Data View, Click on header of the cluster column to sort the data based on their cluster group.

Row No.	NAMA	cluster	svd_1
4	DYAH	cluster_0	0.256
5	WATI	cluster_0	0.235
8	YANTO	cluster_0	0.254
2	AGUS	cluster_1	0.347
6	IKA	cluster_1	0.299
1	JOKO	cluster_2	0.349
3	SUSI	cluster_2	0.315
7	EKO	cluster_2	0.317
9	WAWAN	cluster_2	0.353
10	MAHMUD	cluster_2	0.399

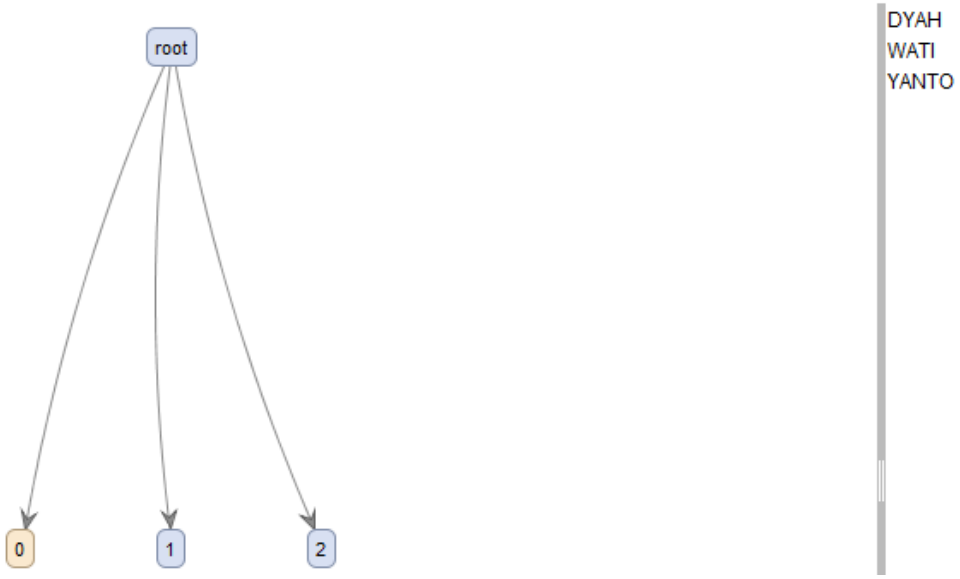
d) Cluster Model (Clustering)

i. Description

Cluster Model

```
Cluster 0: 3 items
Cluster 1: 2 items
Cluster 2: 5 items
Total number of items: 10
```

ii. Graph



Interpretation of the k-Means Algorithm Result

Based on the above activities it can be concluded the division of student groups that will be submitted for the Indonesian and English Olympiad as follows:

Row No.	NAMA	cluster	B.IND	B.ING
1	JOKO	cluster_2	8.540	8.400
2	AGUS	cluster_1	9.980	6.810
3	SUSI	cluster_2	6.200	9.150
4	DYAH	cluster_0	5.240	7.260
5	WATI	cluster_0	5.700	5.710
6	IKA	cluster_1	8.570	5.870
7	EKO	cluster_2	7.700	7.710
8	YANTO	cluster_0	6.600	5.700
9	WAWAN	cluster_2	9	8.120
10	MAHMUD	cluster_2	9.810	9.580

Division of groups submitted for the Olympic competition:

1. Cluster_1 submitted for the Indonesian Language Olympics competition.
2. Cluster_2 submitted for the English Language Olympics competition.

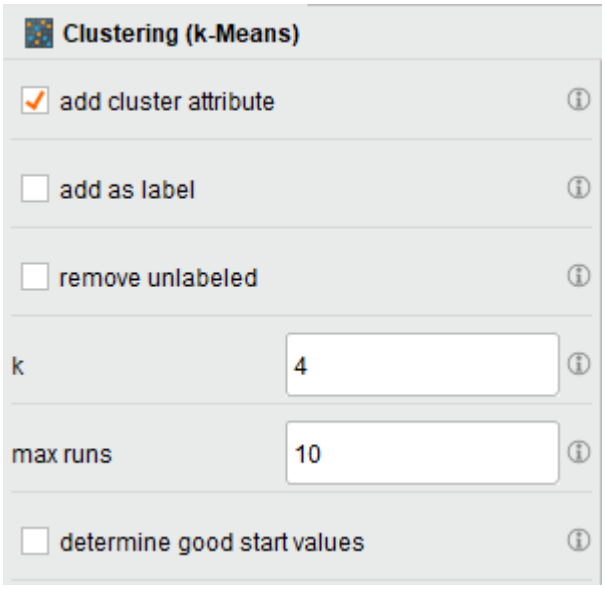
Task

There are 30 students that already participated in 4 subject of exams (Indonesian Language, English Language, Mathematics and Science).

1. Use formula = **5 + RAND () * 5** to determined the value of the score of each student in every subject they participated.

NO_SISWA	NAMA	B.IND	B.ING	MTK	IPA
S-101	JOKO	8,828306	8,360408	6,494591	5,355725
S-102	AGUS	5,755186	6,372207	8,023927	7,258384
S-103	SUSI	8,926182	5,193639	6,229738	7,855382
S-104	DYAH	9,760329	9,965872	6,682567	7,806478
S-105	WATI	6,849591	7,259882	7,502125	5,059896
S-106	IKA	6,277814	7,694837	7,227369	8,854705
S-107	EKO	8,555671	7,237916	9,851829	7,914521
S-108	YANTO	9,453641	7,765326	5,206221	5,073491
S-109	WAWAN	5,512191	6,679037	8,507526	7,800654
S-110	MAHMUD	5,277028	5,810813	7,316658	6,945107
S-111	BUDI	9,831588	7,637738	7,585503	9,283161
S-112	SANTI	7,629082	5,654027	8,588582	6,613303
S-113	DIAN	5,858248	5,711518	6,317297	9,601754
S-114	DANI	5,856101	6,199075	5,608205	8,714048
S-115	AHMAD	6,803392	6,976948	5,670149	5,534259
S-116	BAYU	8,487961	5,133832	5,890456	6,068852
S-117	RISA	5,114542	9,202735	6,470355	5,529113
S-118	RANI	9,414261	9,128862	8,485589	5,89367
S-119	YANI	7,213685	7,367937	5,418838	7,396049
S-120	RATIH	6,844012	7,619553	8,957801	9,966827
S-121	INDAH	7,286627	6,276063	6,387379	9,203564
S-122	JONO	8,410654	9,245861	9,877901	9,100181
S-123	SARAH	5,370875	7,211999	5,615049	7,925657
S-124	RAMA	9,764665	7,53025	6,7816	8,807563
S-125	BAMBANG	5,403232	6,24252	6,971314	8,941572
S-126	HADI	8,338096	6,183913	6,693759	5,188805
S-127	HANA	9,719172	6,02104	7,089418	7,032261
S-128	FEBRI	5,925135	8,037754	6,596817	9,817236
S-129	DENI	9,748079	5,451981	9,477778	8,616619
S-130	TONI	5,292724	6,297933	6,208291	9,616139

2. We do praktikum steps that we did earlier using this data as resource, and in k-Means operator we set the parameter value into 4.



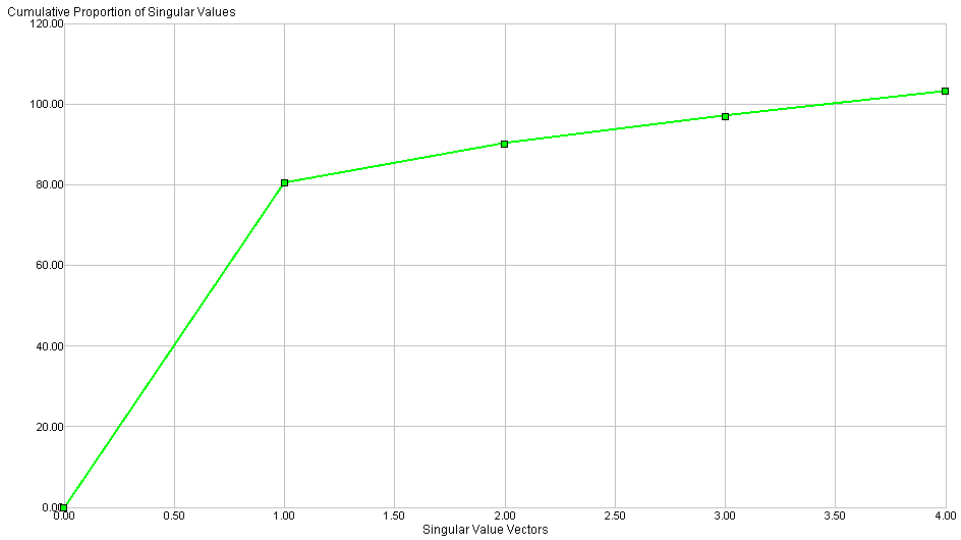
3. The result of the clustering using k-Means algorithm are :
   a) SVD (Singular Value Decomposition)
      i. Eigenvalue

Component	Singular Value	Proportion of Singular Values	Cumulative Singular Values	Cumulative Proportion of Sin...
SVD 1	80.606	0.780	80.606	0.780
SVD 2	9.709	0.094	90.315	0.874
SVD 3	6.884	0.067	97.199	0.941
SVD 4	6.103	0.059	103.302	1.000

      ii. SVD vectors value

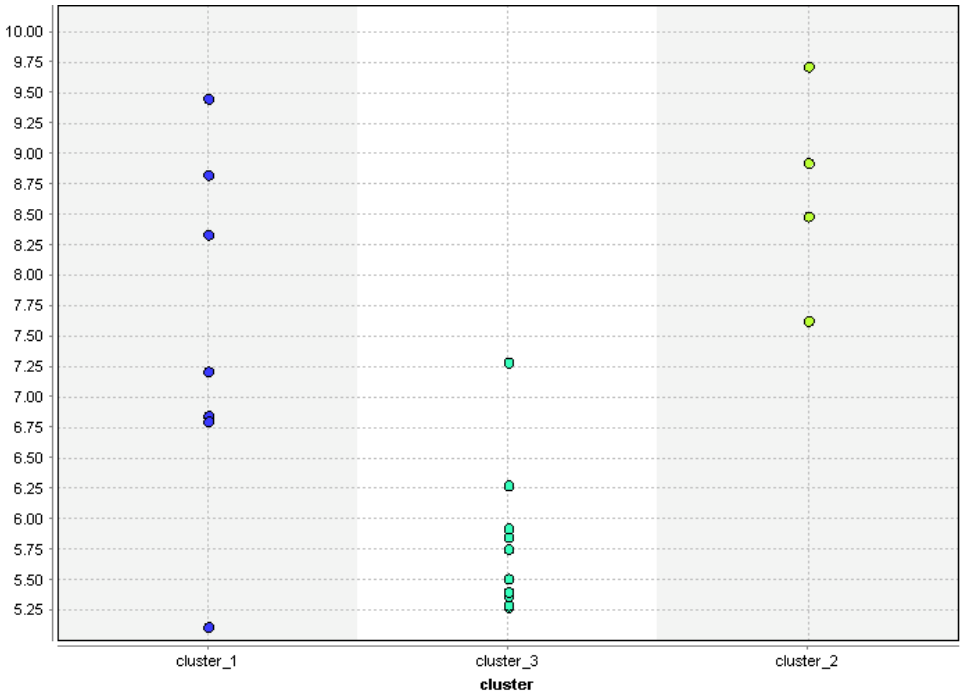
Attribute	SVD Vector 1	SVD Vector 2	SVD Vector 3
B.IND	0.510	-0.674	0.463
B.ING	0.481	-0.144	-0.845
MTK	0.488	0.077	0.074
IPA	0.520	0.721	0.256

iii. Cumulative variance value

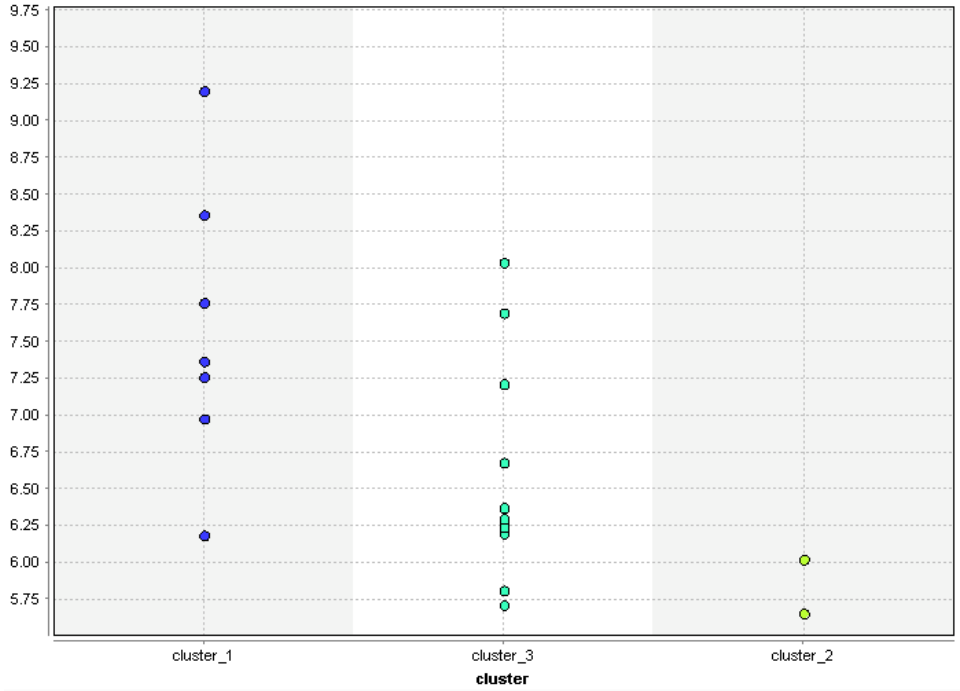


b) ExampleSet (K-Means)

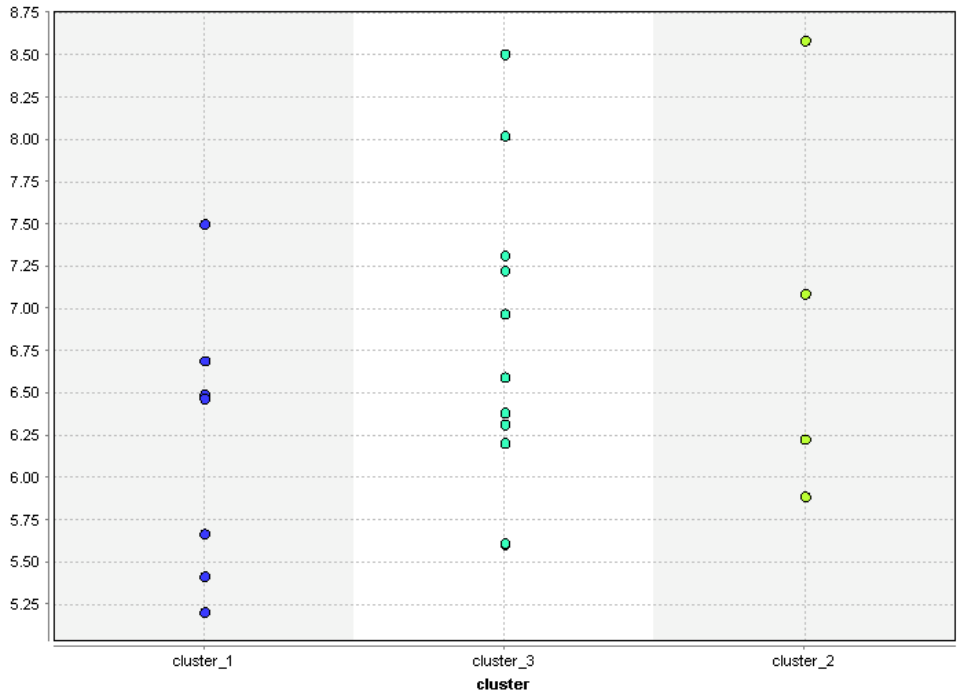
i. Group of Student with Indonesian Language Subject.



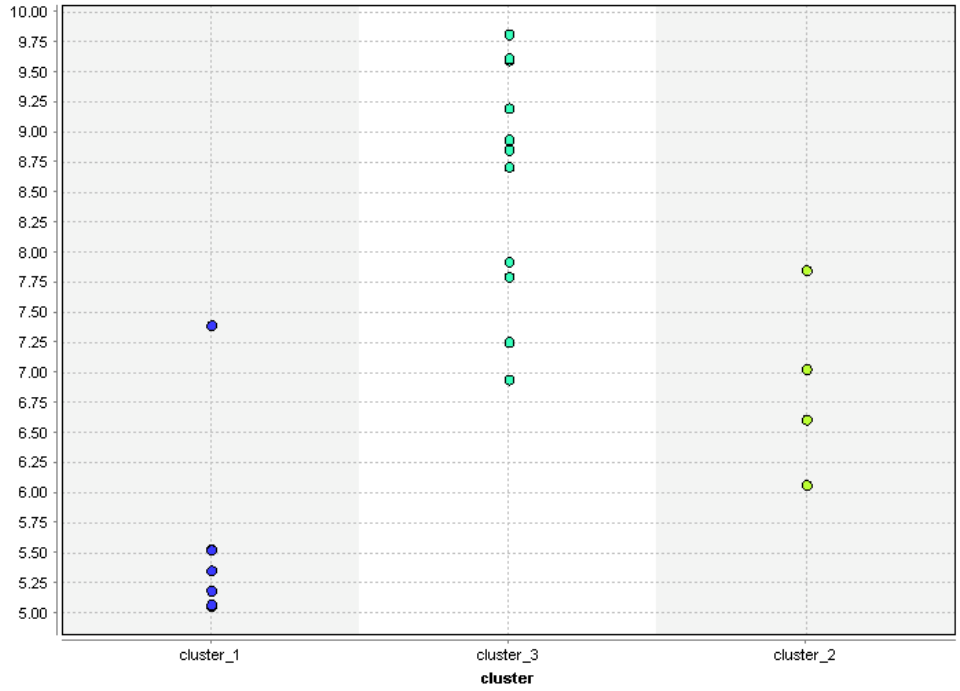
ii. Group of Student with English Language Subject.



iii. Group of Student with Mathematic Subject.



iv. Group of Student with Science Subject.



c) ExampleSet (SVD)

Row No.	NAMA	cluster	svd_1
4	DYAH	cluster_0	0.212
7	EKO	cluster_0	0.208
11	BUDI	cluster_0	0.214
18	RANI	cluster_0	0.203
20	RATIH	cluster_0	0.207
22	JONO	cluster_0	0.227
24	RAMA	cluster_0	0.205
29	DENI	cluster_0	0.207
1	JOKO	cluster_1	0.180
5	WATI	cluster_1	0.165
8	YANTO	cluster_1	0.170
15	AHMAD	cluster_1	0.155
17	RISA	cluster_1	0.162
19	YANI	cluster_1	0.170
26	HADI	cluster_1	0.164

3	SUSI	cluster_2	0.176
12	SANTI	cluster_2	0.177
16	BAYU	cluster_2	0.159
27	HANA	cluster_2	0.186
2	AGUS	cluster_3	0.170
6	IKA	cluster_3	0.187
9	WAWAN	cluster_3	0.177
10	MAHMUD	cluster_3	0.157
13	DIAN	cluster_3	0.171
14	DANI	cluster_3	0.164
21	INDAH	cluster_3	0.182
23	SARAH	cluster_3	0.162
25	BAMBANG	cluster_3	0.171
28	FEBRI	cluster_3	0.189
30	TONI	cluster_3	0.171

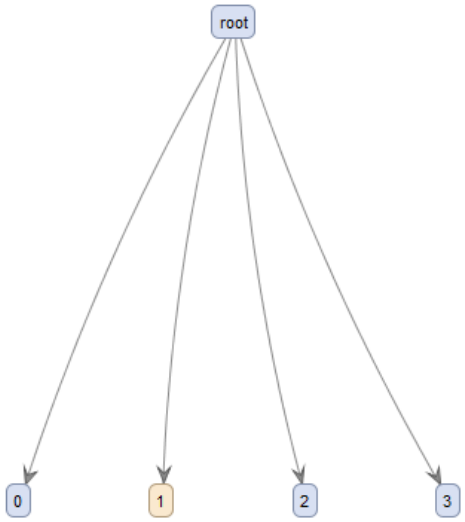
d) Cluster Model (Clustering)

i. Description

Cluster Model

```
Cluster 0: 8 items
Cluster 1: 7 items
Cluster 2: 4 items
Cluster 3: 11 items
Total number of items: 30
```

ii. Graph



JOKO
WATI
YANTO
AHMAD
RISA
YANI
HADI

4. This is the name of the student based on their clustering group.
   i. Cluster 0

DYAH	cluster_0	0.212
EKO	cluster_0	0.208
BUDI	cluster_0	0.214
RANI	cluster_0	0.203
RATIH	cluster_0	0.207
JONO	cluster_0	0.227
RAMA	cluster_0	0.205
DENI	cluster_0	0.207

ii. Cluster 1

5	WATI	cluster_1	0.165
8	YANTO	cluster_1	0.170
15	AHMAD	cluster_1	0.155
17	RISA	cluster_1	0.162
19	YANI	cluster_1	0.170
26	HADI	cluster_1	0.164

iii. Cluster 2

3	SUSI	cluster_2	0.176
12	SANTI	cluster_2	0.177
16	BAYU	cluster_2	0.159
27	HANA	cluster_2	0.186

iv. Cluster 3

2	AGUS	cluster_3	0.170
6	IKA	cluster_3	0.187
9	WAWAN	cluster_3	0.177
10	MAHMUD	cluster_3	0.157
13	DIAN	cluster_3	0.171
14	DANI	cluster_3	0.164
21	INDAH	cluster_3	0.182
23	SARAH	cluster_3	0.162
25	BAMBANG	cluster_3	0.171
28	FEBRI	cluster_3	0.189
30	TONI	cluster_3	0.171