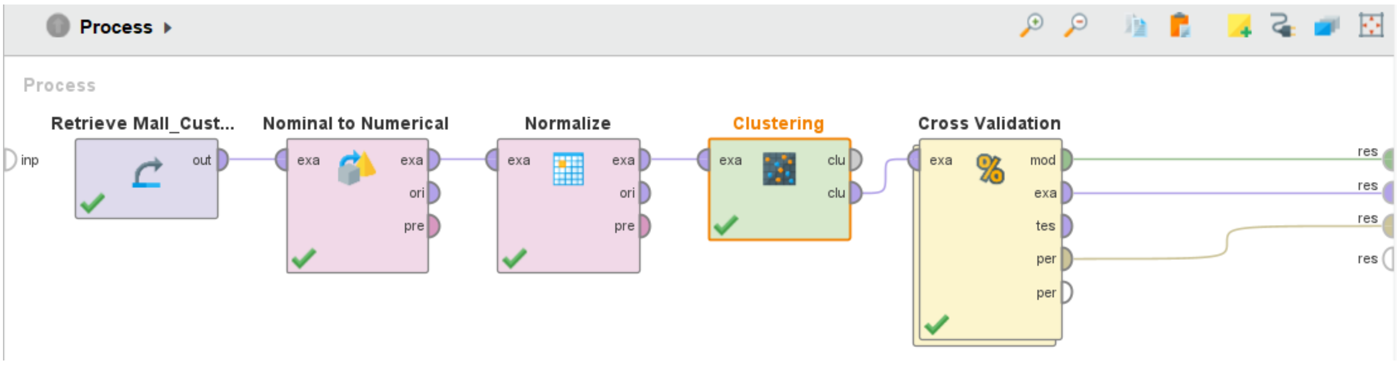
Clustering Analysis

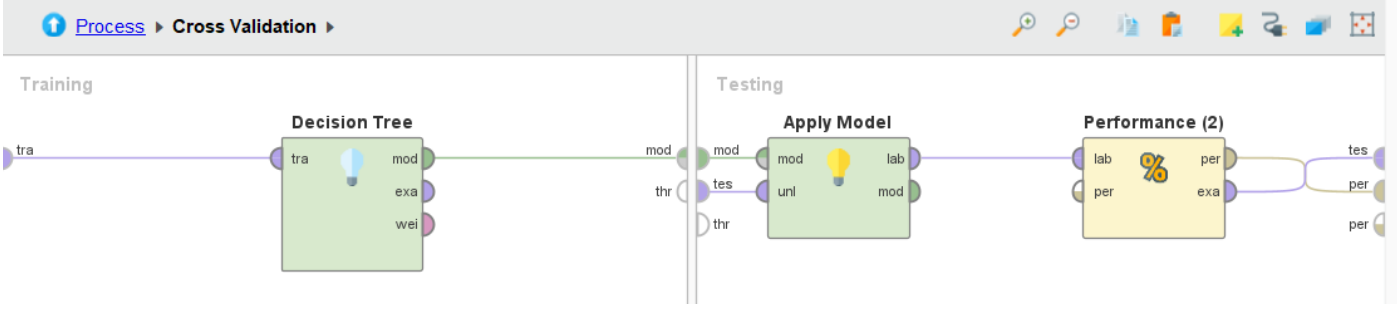
Name: Jahnvi Rameshbhai Patel

Model: Decision Tree

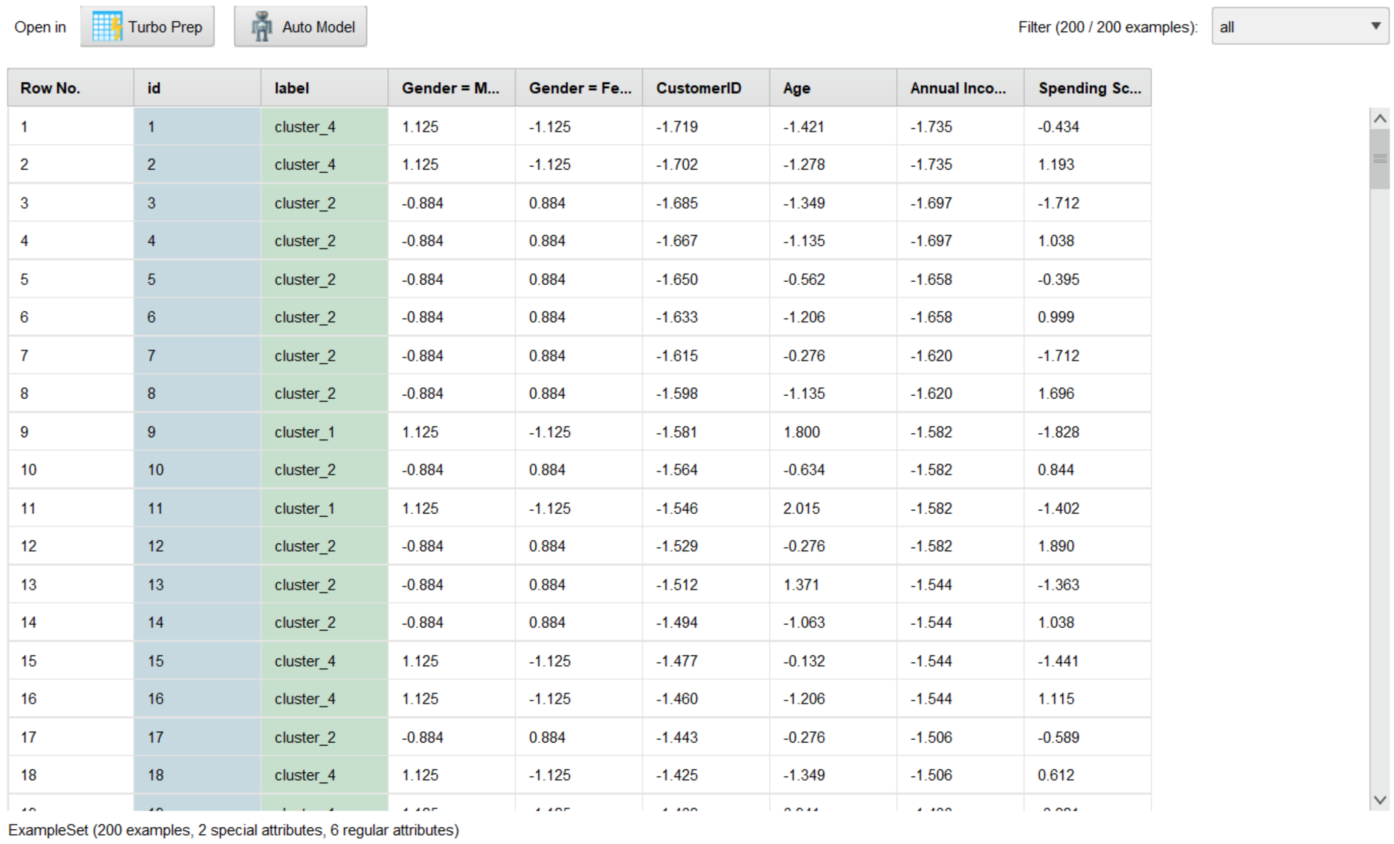
Process:



Subprocess:



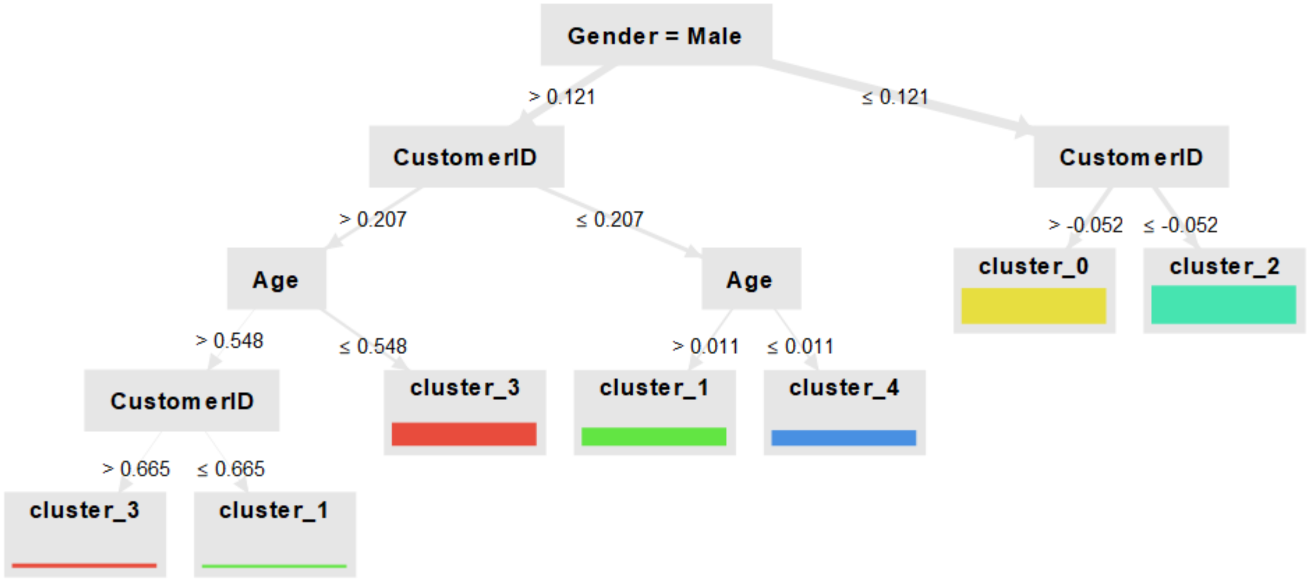
Example Set:



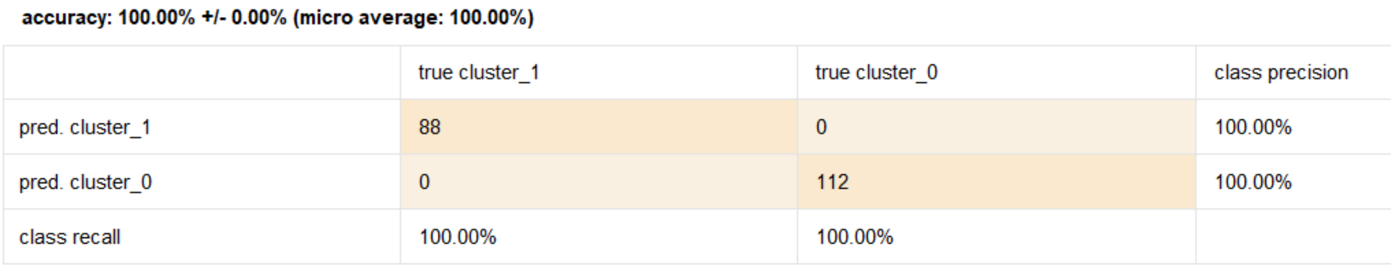
Row No.	id	label	Gender = M...	Gender = Fe...	CustomerID	Age	Annual Inco...	Spending Sc...
1	1	cluster_4	1.125	-1.125	-1.719	-1.421	-1.735	-0.434
2	2	cluster_4	1.125	-1.125	-1.702	-1.278	-1.735	1.193
3	3	cluster_2	-0.884	0.884	-1.685	-1.349	-1.697	-1.712
4	4	cluster_2	-0.884	0.884	-1.667	-1.135	-1.697	1.038
5	5	cluster_2	-0.884	0.884	-1.650	-0.562	-1.658	-0.395
6	6	cluster_2	-0.884	0.884	-1.633	-1.206	-1.658	0.999
7	7	cluster_2	-0.884	0.884	-1.615	-0.276	-1.620	-1.712
8	8	cluster_2	-0.884	0.884	-1.598	-1.135	-1.620	1.696
9	9	cluster_1	1.125	-1.125	-1.581	1.800	-1.582	-1.828
10	10	cluster_2	-0.884	0.884	-1.564	-0.634	-1.582	0.844
11	11	cluster_1	1.125	-1.125	-1.546	2.015	-1.582	-1.402
12	12	cluster_2	-0.884	0.884	-1.529	-0.276	-1.582	1.890
13	13	cluster_2	-0.884	0.884	-1.512	1.371	-1.544	-1.363
14	14	cluster_2	-0.884	0.884	-1.494	-1.063	-1.544	1.038
15	15	cluster_4	1.125	-1.125	-1.477	-0.132	-1.544	-1.441
16	16	cluster_4	1.125	-1.125	-1.460	-1.206	-1.544	1.115
17	17	cluster_2	-0.884	0.884	-1.443	-0.276	-1.506	-0.589
18	18	cluster_4	1.125	-1.125	-1.425	-1.349	-1.506	0.612

ExampleSet (200 examples, 2 special attributes, 6 regular attributes)

Tree:



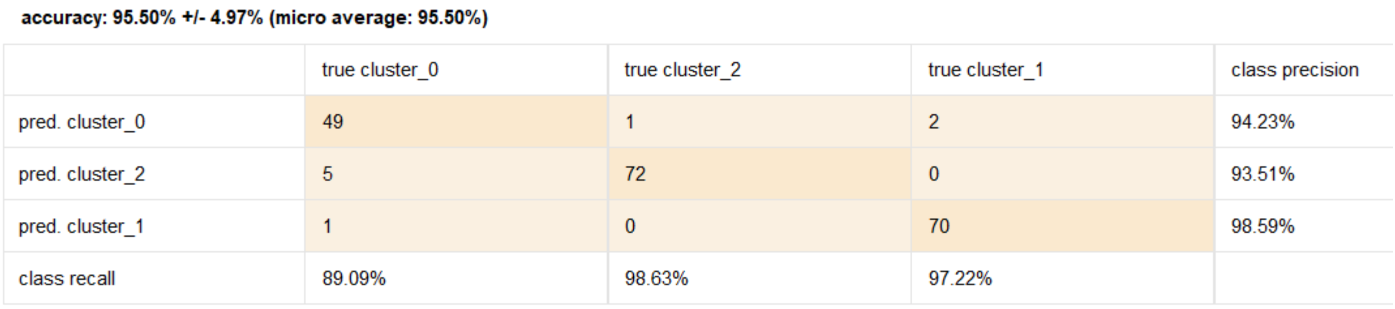
K=2



accuracy: 100.00% +/- 0.00% (micro average: 100.00%)

	true cluster_1	true cluster_0	class precision
pred. cluster_1	88	0	100.00%
pred. cluster_0	0	112	100.00%
class recall	100.00%	100.00%	

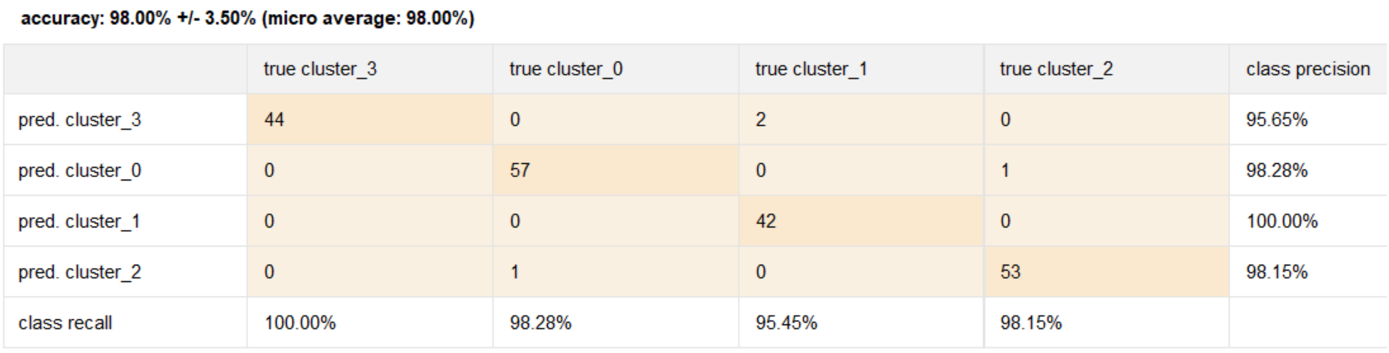
K=3



accuracy: 95.50% +/- 4.97% (micro average: 95.50%)

	true cluster_0	true cluster_2	true cluster_1	class precision
pred. cluster_0	49	1	2	94.23%
pred. cluster_2	5	72	0	93.51%
pred. cluster_1	1	0	70	98.59%
class recall	89.09%	98.63%	97.22%	

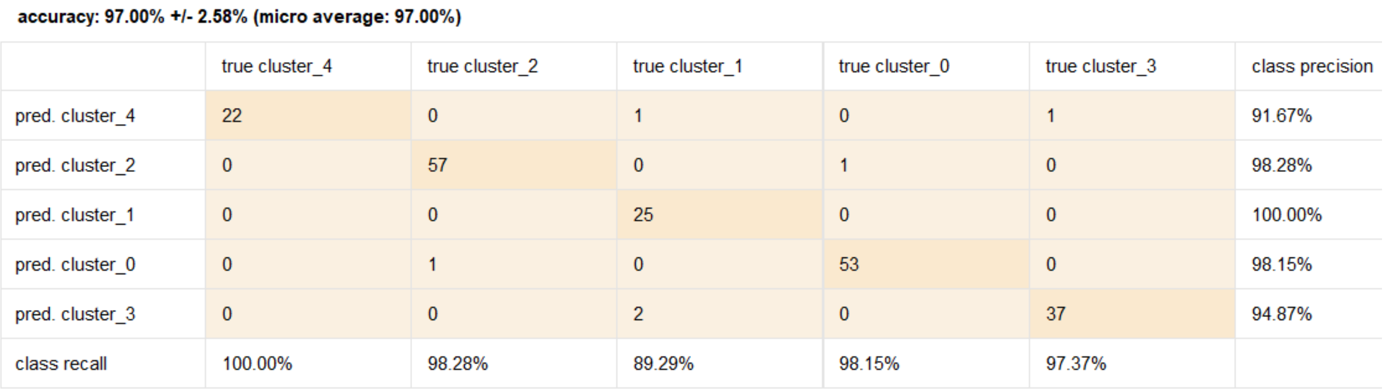
K=4



accuracy: 98.00% +/- 3.50% (micro average: 98.00%)

	true cluster_3	true cluster_0	true cluster_1	true cluster_2	class precision
pred. cluster_3	44	0	2	0	95.65%
pred. cluster_0	0	57	0	1	98.28%
pred. cluster_1	0	0	42	0	100.00%
pred. cluster_2	0	1	0	53	98.15%
class recall	100.00%	98.28%	95.45%	98.15%	

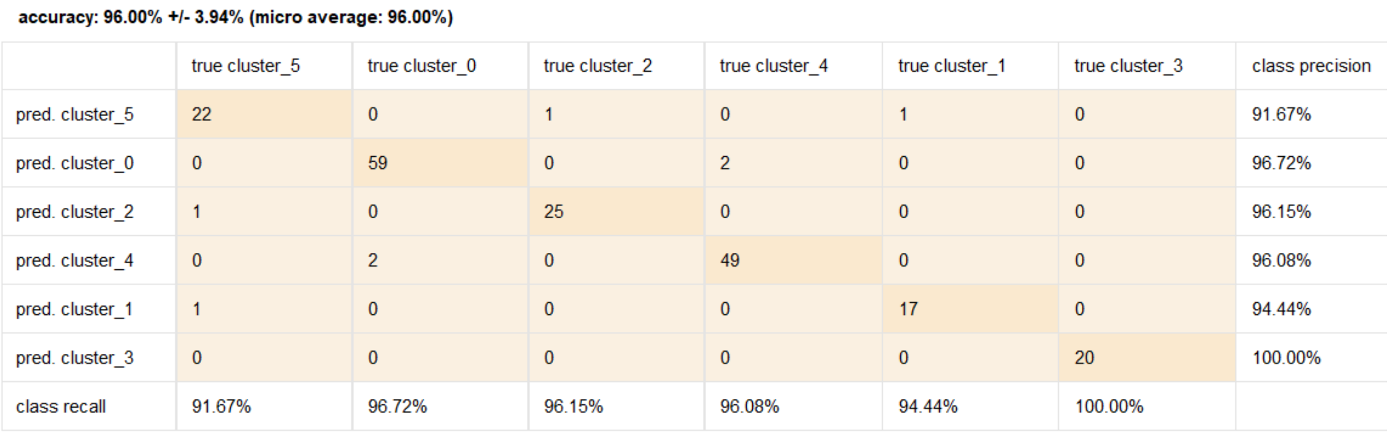
K=5



accuracy: 97.00% +/- 2.58% (micro average: 97.00%)

	true cluster_4	true cluster_2	true cluster_1	true cluster_0	true cluster_3	class precision
pred. cluster_4	22	0	1	0	1	91.67%
pred. cluster_2	0	57	0	1	0	98.28%
pred. cluster_1	0	0	25	0	0	100.00%
pred. cluster_0	0	1	0	53	0	98.15%
pred. cluster_3	0	0	2	0	37	94.87%
class recall	100.00%	98.28%	89.29%	98.15%	97.37%	

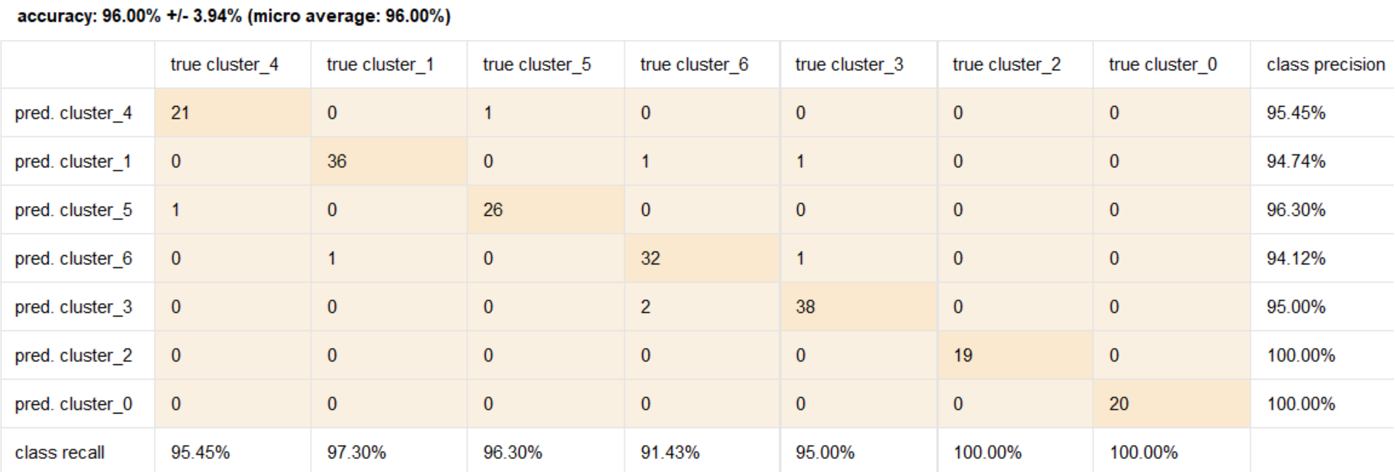
K=6



accuracy: 96.00% +/- 3.94% (micro average: 96.00%)

	true cluster_5	true cluster_0	true cluster_2	true cluster_4	true cluster_1	true cluster_3	class precision
pred. cluster_5	22	0	1	0	1	0	91.67%
pred. cluster_0	0	59	0	2	0	0	96.72%
pred. cluster_2	1	0	25	0	0	0	96.15%
pred. cluster_4	0	2	0	49	0	0	96.08%
pred. cluster_1	1	0	0	0	17	0	94.44%
pred. cluster_3	0	0	0	0	0	20	100.00%
class recall	91.67%	96.72%	96.15%	96.08%	94.44%	100.00%	

K=7



accuracy: 96.00% +/- 3.94% (micro average: 96.00%)

	true cluster_4	true cluster_1	true cluster_5	true cluster_6	true cluster_3	true cluster_2	true cluster_0	class precision
pred. cluster_4	21	0	1	0	0	0	0	95.45%
pred. cluster_1	0	36	0	1	1	0	0	94.74%
pred. cluster_5	1	0	26	0	0	0	0	96.30%
pred. cluster_6	0	1	0	32	1	0	0	94.12%
pred. cluster_3	0	0	0	2	38	0	0	95.00%
pred. cluster_2	0	0	0	0	0	19	0	100.00%
pred. cluster_0	0	0	0	0	0	0	20	100.00%
class recall	95.45%	97.30%	96.30%	91.43%	95.00%	100.00%	100.00%	

Based on the accuracy results above, it seems that K=2 is the optimal value for K in this case. The accuracy is highest for K=2 at 100%, which indicates that this value of K is able to separate the data into distinct clusters that are most representative of the underlying patterns in the data.

It is worth noting that while K=4 also has a high accuracy of 98%, having 4 clusters may not provide enough granularity to fully capture the complexity of the data. On the other hand, as the value of K increases beyond K=2, the accuracy begins to drop, suggesting that the additional clusters may be introducing noise or not capturing meaningful patterns in the data. Therefore, K=2 is likely the best choice for this particular dataset.