Exercise : PCA + k-NN

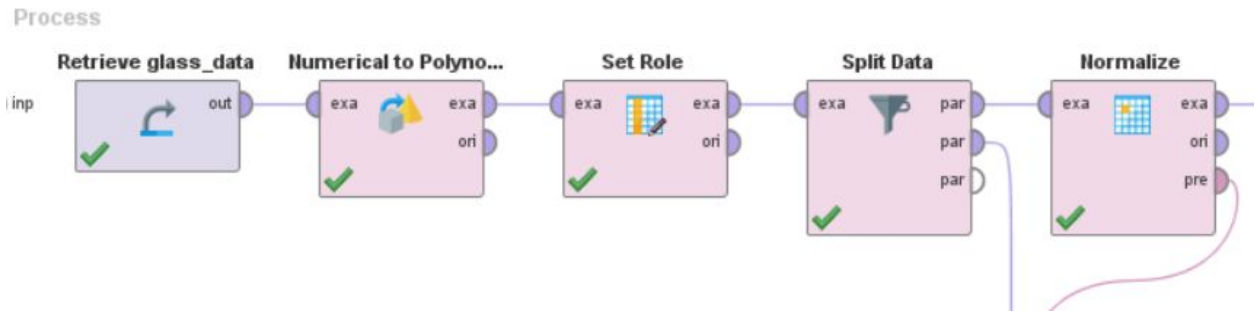
From Rapidminer

1) Construct the classifier to predict type of glass without the use of PCA

   - Pre-processing:

	Id *integer*	RI *real*	Na *real*	att4 *real*	att5 *real*	att6 *real*	att7 *real*	att8 *real*
1	1	1.521	13.640	4.490	1.100	71.780	0.060	8.750
2	2	1.518	13.890	3.600	1.360	72.730	0.480	7.830
3	3	1.516	13.530	3.550	1.540	72.990	0.390	7.780
4	4	1.518	13.210	3.690	1.290	72.610	0.570	8.220
5	5	1.517	13.270	3.620	1.240	73.080	0.550	8.070
6	6	1.516	12.790	3.610	1.620	72.970	0.640	8.070
7	7	1.517	13.300	3.600	1.140	73.090	0.580	8.170
8	8	1.518	13.150	3.610	1.050	73.240	0.570	8.240
9	9	1.519	14.040	3.580	1.370	72.080	0.560	8.300
10	10	1.518	13.000	3.600	1.360	72.990	0.570	8.400
11	11	1.516	12.720	3.460	1.560	73.200	0.670	8.090
12	12	1.518	12.800	3.660	1.270	73.010	0.600	8.560
13	13	1.516	12.880	3.430	1.400	73.280	0.690	8.050
14	14	1.517	12.860	3.560	1.270	73.210	0.540	8.380
15	15	1.518	12.610	3.590	1.310	73.290	0.580	8.500
16	16	1.518	12.810	3.540	1.230	73.240	0.580	8.390
17	17	1.518	12.680	3.670	1.160	73.110	0.610	8.700

Rename Header



- Classification:
    - Show results on different values of K

iteration	k-NN.k	accuracy
1	1	0.662
6	6	0.708
3	3	0.677
2	2	0.662
4	4	0.692
5	5	0.723
7	7	0.708
8	8	0.708
10	10	0.662
9	9	0.692

Ans best K to obtain best accuracy is 5

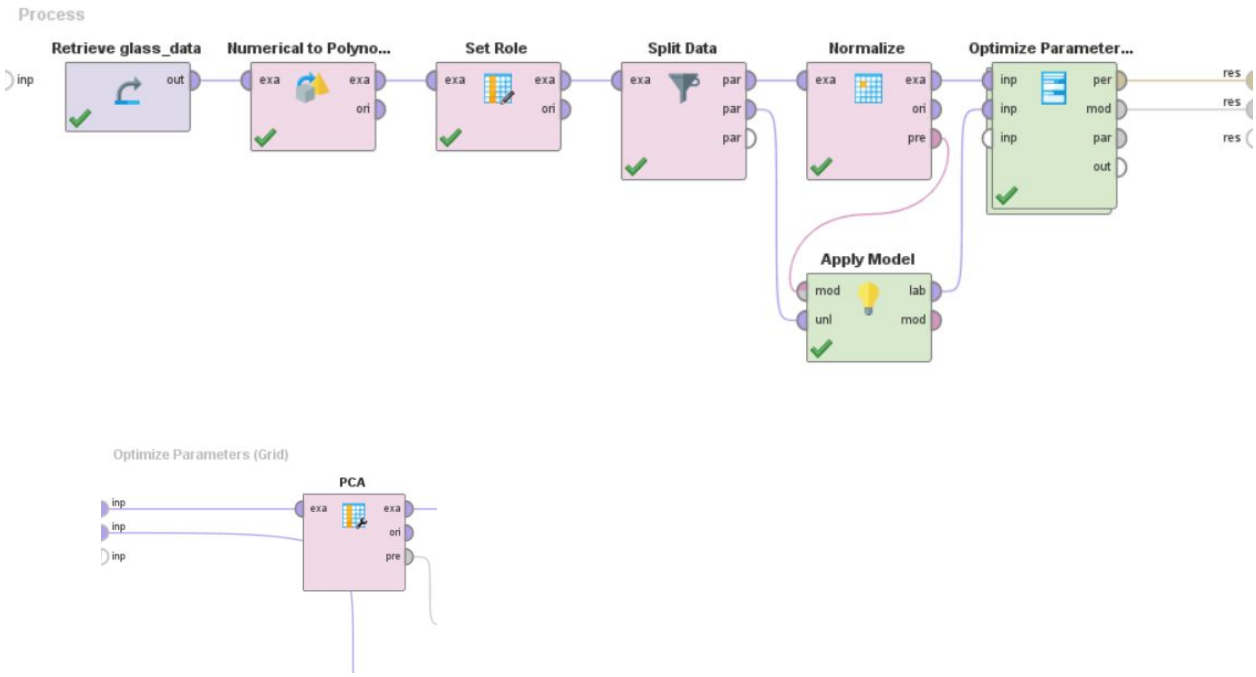
- Show confusion-matrix for each class

accuracy: 72.31%

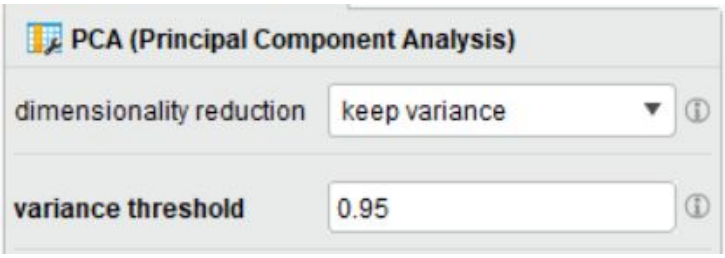
	true 1	true 2	true 3	true 5	true 6	true 7	class precision
pred. 1	19	5	4	0	1	0	65.52%
pred. 2	2	17	1	1	1	0	77.27%
pred. 3	0	0	0	0	0	0	0.00%
pred. 5	0	1	0	1	0	0	50.00%
pred. 6	0	0	0	0	1	0	100.00%
pred. 7	0	0	0	2	0	9	81.82%
class recall	90.48%	73.91%	0.00%	25.00%	33.33%	100.00%	

2) Construct the classifier to predict type of glass with the use of PCA

- Pre-processing:
  - Show each step of your pre-processing (include PCA)



  - Show & explain setting of PCA parameters



Ans ปรับ variance threshold ไว้ที่ 0.95 เพื่อให้คุณลักษณะที่แตกต่างกันระหว่างจุดข้อมูลถึงที่ขอบเขตกำหนด เนื่องจากใน component ที่ 1 จะมีค่า variance สูงสุดและลดลงตามลำดับ

- Explain output of PCA in terms of
    - Principal components

Row No.	Id	Type of glass	pc_1	pc_2	pc_3	pc_4	pc_5	pc_6	pc_7
1	3	1	-0.932	-0.807	0.612	0.159	-0.354	-0.249	0.404
2	7	1	-0.291	-0.984	0.575	0.496	-0.409	0.103	-0.269
3	8	1	-0.188	-1.058	0.606	0.763	-0.329	0.226	-0.354
4	11	1	-0.506	-1.437	-0.406	-0.074	2.042	-0.032	0.176
5	12	1	-0.069	-1.073	0.088	0.677	-0.375	0.313	0.129
6	13	1	-0.441	-1.424	-0.182	-0.035	2.031	-0.043	-0.157
7	14	1	0.027	-1.182	0.075	0.228	1.324	0.167	-0.082
8	15	1	-0.189	-1.120	0.025	1.047	-0.210	0.391	0.210
9	16	1	-0.197	-1.071	0.204	0.910	-0.265	0.315	0.016
10	17	1	0.092	-1.127	0.107	0.890	-0.331	0.427	-0.005
11	18	1	1.611	0.149	1.248	-1.607	-1.383	-0.080	-0.252
12	20	1	-0.331	-0.842	-0.236	-0.096	0.189	-0.051	0.650
13	22	1	1.309	-0.047	2.371	-0.884	-1.072	-0.230	-1.273
14	24	1	-0.169	-1.014	-0.019	0.671	-0.363	0.246	0.202
15	25	1	-0.156	-0.816	0.592	0.290	-0.485	0.000	-0.168
16	27	1	-0.144	-0.761	0.089	0.046	-0.587	0.036	0.209

    - Accumulate variance

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	1.637	0.298	0.298
PC 2	1.433	0.228	0.526
PC 3	1.096	0.133	0.660
PC 4	1.054	0.123	0.783
PC 5	0.966	0.104	0.887
PC 6	0.697	0.054	0.941
PC 7	0.678	0.051	0.992

- Classification:
    - Show format of the dataset in terms of new dimensions

Row No.	Id	Type of glass	pc_1	pc_2	pc_3	pc_4	pc_5	pc_6	pc_7
1	3	1	-0.932	-0.807	0.612	0.159	-0.354	-0.249	0.404
2	7	1	-0.291	-0.984	0.575	0.496	-0.409	0.103	-0.269
3	8	1	-0.188	-1.058	0.606	0.763	-0.329	0.226	-0.354
4	11	1	-0.506	-1.437	-0.406	-0.074	2.042	-0.032	0.176
5	12	1	-0.069	-1.073	0.088	0.677	-0.375	0.313	0.129
6	13	1	-0.441	-1.424	-0.182	-0.035	2.031	-0.043	-0.157
7	14	1	0.027	-1.182	0.075	0.228	1.324	0.167	-0.082
8	15	1	-0.189	-1.120	0.025	1.047	-0.210	0.391	0.210
9	16	1	-0.197	-1.071	0.204	0.910	-0.265	0.315	0.016
10	17	1	0.092	-1.127	0.107	0.890	-0.331	0.427	-0.005
11	18	1	1.611	0.149	1.248	-1.607	-1.383	-0.080	-0.252
12	20	1	-0.331	-0.842	-0.236	-0.096	0.189	-0.051	0.650
13	22	1	1.309	-0.047	2.371	-0.884	-1.072	-0.230	-1.273
14	24	1	-0.169	-1.014	-0.019	0.671	-0.363	0.246	0.202
15	25	1	-0.156	-0.816	0.592	0.290	-0.485	0.000	-0.168
16	27	1	-0.144	-0.761	0.089	0.046	-0.587	0.036	0.209

- What is best combination of PCA (variance) & K to obtain best accuracy

Optimize Parameters (Grid) (200 rows, 4 columns)

iteration	PCA.variance_threshold	k-NN.k	acc... ↓
95	0.700	5	0.738
96	0.750	5	0.738
120	0.950	6	0.723
81	0	5	0.723
101	0	6	0.708
112	0.550	6	0.708
113	0.600	6	0.708
114	0.650	6	0.708
115	0.700	6	0.708
116	0.750	6	0.708
121	0	7	0.708
137	0.800	7	0.708

Ans best PCA (variance) is 0.700 and k is 5

- Show confusion matrix for each class

accuracy: 73.85%

	true 1	true 2	true 3	true 5	true 6	true 7	class precision
pred. 1	18	5	2	0	0	0	72.00%
pred. 2	3	17	3	1	1	0	68.00%
pred. 3	0	0	0	0	0	0	0.00%
pred. 5	0	1	0	2	0	0	66.67%
pred. 6	0	0	0	0	2	0	100.00%
pred. 7	0	0	0	1	0	9	90.00%
class recall	85.71%	73.91%	0.00%	50.00%	66.67%	100.00%	

3) Discuss and Compare accuracy obtained between with PCA and without PCA

Ans การใช้ PCA ทำให้ dimension ของข้อมูลลดลง ส่งผลให้โอกาศที่จะเกิด Over fitting น้อยลงตามไปด้วย ทำให้ accuracy เพิ่มสูงขึ้นจากการทำ k-NN เพียงอย่างเดียว