UT5 TA03

El Data set contiene información nutricional de 77 marcas y/o variedades de cereales a la venta en el mercado obtenidos de sus etiquetas nutricionales. El objetivo es reducir estos atributos de 13 (numéricos) a la menor cantidad posible que expliquen en mayor proporción el problema.

Parámetros del operador PCA

Indica que tipo de reducción de dimensiones realizar.

- None: si se selecciona esta opción ningun atributo es removido del data set.
- keep_variance: si se selecciona esta opción todos los atributos con una variación acumulada mayor al umbral definido son removidos del data set. El umbral es un parámetro configurable.
- fixed_number: Si se selecciona esta opción un número fijo de componentes se mantiene, este número se fija como parámetro.

Paso 3. Ejecución e Interpretación

Contribución de cada componente a la variación total

Component	Standard Deviation	Proportion of Variance	Cumulative Variance
PC 1	84.178	0.565	0.565
PC 2	71.202	0.405	0.970
PC 3	18.697	0.028	0.998
PC 4	4.758	0.002	1.000
PC 5	1.712	0.000	1.000
PC 6	0.948	0.000	1.000
PC 7	0.866	0.000	1.000
PC 8	0.711	0.000	1.000
PC 9	0.456	0.000	1.000
PC 10	0.392	0.000	1.000
PC 11	0.000	0.000	1.000
PC 12	?	-0.000	1.000

Se observa en los primeros tres componentes principales el peso fuerte de 3 atributos

Attribute	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12
calories	-0.074	0.009	-0.986	0.080	-0.090	0.013	0.036	0.019	-0.003	0.072	0.000	0.000
sugars	-0.005	-0.002	-0.137	-0.719	0.588	-0.003	-0.005	-0.133	-0.025	-0.316	0.000	-0.000
carbo	-0.014	0.019	-0.029	0.682	0.647	-0.063	-0.036	-0.022	0.037	-0.329	-0.000	-0.000
protein	0.001	-0.008	-0.004	0.087	-0.251	0.322	0.047	-0.849	0.061	-0.315	-0.000	0.000
fat	0.000	-0.003	-0.029	-0.029	-0.391	-0.089	-0.242	0.366	0.055	-0.801	-0.000	-0.000
sodium	-0.992	-0.102	0.073	-0.010	-0.005	-0.001	-0.001	-0.001	0.001	0.000	-0.000	0.000
fiber	0.004	-0.030	0.029	0.009	-0.044	0.070	0.960	0.168	-0.004	-0.201	0.000	-0.000
potass	0.101	-0.994	-0.017	0.015	0.016	-0.001	-0.029	0.003	0.002	0.006	-0.000	0.000
shelf	0.001	-0.004	-0.007	-0.009	-0.079	-0.765	0.093	-0.253	0.052	-0.016	0.139	0.560
shelf_1	-0.001	0.001	0.005	0.024	0.026	0.363	-0.050	0.106	-0.432	-0.020	0.699	0.421
shelf_2	0.000	0.002	-0.003	-0.038	0.028	0.039	0.007	0.041	0.811	0.056	0.560	-0.139
shelf_3	0.000	-0.003	-0.002	0.014	-0.053	-0.402	0.043	-0.147	-0.380	-0.036	0.421	-0.699

Significando cada atributo casi todo el peso del componente principal.

Lo que demuesta sodio para el componente principal 1, potasio para el 2 y calorías para el 3 como los atributos determinantes del modelo para explicar gran parte de la variación.

En cuanto a mejora en el procesamiento, poder reducir de una gran número de atributos a unos pocos resulta en una optimización del modelo que puede implicar grandes ventajas, no sólo en rapidez de procesamiento sino también en interpretabilidad del modelo y facilidad de análisis.

Riesgos a considerar cuando se utiliza PCA:

• ¿qué características estadísticas tienen los atributos reales que fueron identificados como más significativos?

Los rangos de los atributos identificados son notoriamente más elevados que los de los demás atributos, esto sin dudas afecta la sensibilidad del modelo, una variación en un atributo con un rango muy amplio puede ser menor porcentualmente pero provocar un cambio en el modelo si no está normalizado.

• ¿Cuál sería el efecto de otro posible atributo, “volumen de ventas”, cuyo rango estuviera en los millones (de $ o cajas)?

Sin dudas cualquier atributo cuya escala fuera sensiblemente superior a la de los otros alteraría los resultados ya que el modelo lo identificaría como un atributo de peso cuando podría no serlo.

Luego de efectuar la normalización se observa que disminuye el peso de los componentes principales, en el modelo anterior los dos primeros componentes explicaban el 97% de la variación. En este caso los 5 primeros explican el 92,6%.

Al observar los vectores dentro de cada componente se observa el cambio del peso de cada atributo.

Attribute	PC 1 ↑	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8	PC 9	PC 10	PC 11	PC 12
sodium	-0.020	-0.057	0.477	0.376	0.613	-0.392	-0.292	-0.112	-0.015	-0.006	0.000	0.000
carbo	-0.057	-0.130	-0.075	0.532	0.184	0.280	0.470	0.329	-0.087	0.492	-0.000	-0.000
shelf_2	-0.213	0.723	-0.195	-0.001	0.243	0.030	0.011	0.032	-0.006	-0.003	-0.459	-0.350
shelf_1	-0.438	-0.511	0.149	-0.105	-0.118	-0.001	-0.010	0.009	0.003	-0.002	-0.707	-0.025
sugars	0.011	0.292	0.672	-0.237	-0.295	-0.121	0.338	-0.170	-0.046	0.400	-0.000	-0.000
calories	0.018	0.064	0.409	0.115	0.066	0.422	0.314	0.151	0.276	-0.660	0.000	0.000
protein	0.073	-0.153	-0.090	-0.296	0.473	0.414	0.182	-0.657	-0.006	0.126	-0.000	-0.000
fat	0.076	0.052	0.261	-0.168	0.040	0.577	-0.620	0.311	-0.019	0.281	-0.000	-0.000
fiber	0.093	-0.080	-0.072	-0.362	0.263	-0.223	0.113	0.316	0.756	0.213	-0.000	0.000
potass	0.137	-0.086	0.042	-0.475	0.341	-0.136	0.224	0.443	-0.585	-0.150	0.000	-0.000
shelf	0.544	0.150	-0.051	0.105	-0.004	-0.014	0.005	-0.024	0.000	0.003	-0.495	0.649
shelf_3	0.651	-0.212	0.046	0.105	-0.125	-0.030	-0.000	-0.040	0.003	0.004	-0.212	-0.675

En los primeros 5 componentes se observa la importancia del atributo shelf (estante del supermercado donde es exhibido el producto). Así como azúcares y sodio, carbohidratos, proteínas y potasio. Pero su peso, al estar normalizado, disminuye en relación al primer análisis efectuado. Lo que reafirma la importancia de la normalización en el modelo.