Universidad Tecnológica de Bolívar
CARTAGENA DE INDIAS

Taller 6: PCA

Caso: implementación de PCA

Análisis de datos – Prof. David Porta

Maestría en Estadística Aplicada

TRANSFORMAMOS VIDAS
CON EDUCACIÓN DE EXCELENCIA

Presentado por:

DIEGO HERRERA MALAMBO - malambod@utb.edu.co

Implementación de PCA

El dataset usado para el ejercicio de PCA corresponde a datos antropométricos y de hábitos de personas orientados a la clasificación de sobrepeso. Link de dataset

El dataset consta de 15 variables independientes y 1 dependiente, las cuales no están normalizadas, se aplican técnicas de separación de variables categóricas a dicotómicas, a fin de implementar técnica de componentes principales.

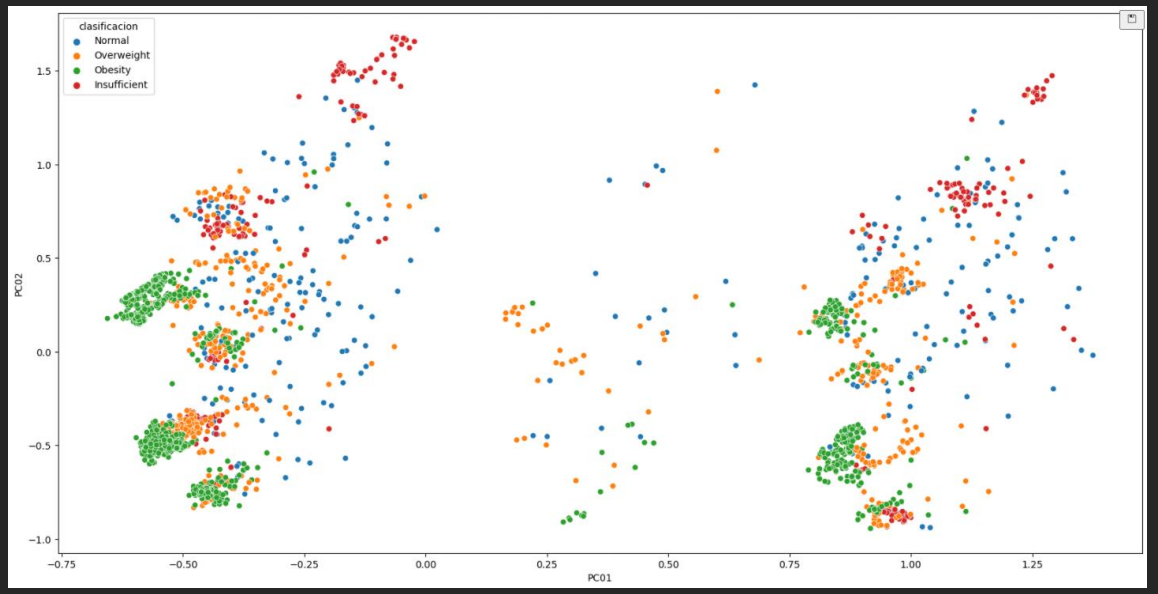
Al implementar PCA se pudo encontrar que en 5 variables se podía concentrar el 65% del dataset, al usar 10 componentes llegamos al 86.5%.

Distribución de variables con mayor importancia sobre los componentes (>=0.25 o <=-0.25):

- PC01:consumo_alcohol__Sometimes,consumo_alcohol__no
- PC02:alimentos_entre_comidas__Frequently,alimentos_entre_comidas__Sometimes,historia_fam,medio_de_transporte__Public_Transportation,sexo
- PC03:alimentos_entre_comidas__Frequently,alimentos_entre_comidas__Sometimes,historia_fam,sexo
- PC04:medio_de_transporte__Public_Transportation,sexo
- PC05:alimentos_entre_comidas__Frequently,alimentos_entre_comidas__Sometimes,historia_fam
- PC06:actividad_fisica,consumo_alto_calorias,consumo_de_agua
- PC07:uso_de_tecnologia
- PC08:consumo_alto_calorias,consumo_de_agua,historia_fam
- PC09:actividad_fisica,consumo_de_agua,consumo_vegetales,numero_de_comidas
- PC10:consumo_vegetales,numero_de_comidas

A continuación muestro dos graficas:
1. un heatmap que nos muestra las variables que mas impactan sobre cada uno de los componentes .
2. Scatterplot de los componentes que tienen mayor concentración (PC01, PC02) y los cluster según los datos originales.

Los resultados no son concluyentes ya que los gráficos con las dos componentes no separan los datos



A fin de validar los resultados obtenidos con los PCA, si implementaron modelos de regresión Lineal(OLS), realizando diferentes combinaciones de variables además de implementación con los PCA.

El mejor modelo tomando solamente las variables con un P<|t| <= 0.05, el resultado es el siguiente:

	coef	std err	t	P>|t|	[0.025	0.975]
sexo	-0.1886	0.045	-4.194	0.000	-0.277	-0.100
edad	1.0460	0.127	8.251	0.000	0.797	1.295
altura	-1.0758	0.141	-7.622	0.000	-1.353	-0.799
peso	1.7834	0.122	14.607	0.000	1.544	2.023
historia_fam	0.4532	0.051	8.929	0.000	0.354	0.553
consumo_alto_calorias	-0.1794	0.054	-3.339	0.001	-0.285	-0.074
consumo_vegetales	-0.5172	0.067	-7.749	0.000	-0.648	-0.386
numero_de_comidas	-0.4865	0.066	-7.404	0.000	-0.615	-0.358
monitoreo_calorias	0.3390	0.081	4.166	0.000	0.179	0.499
alimentos_entre_comidas__Frequently	-0.3258	0.112	-2.903	0.004	-0.546	-0.106
alimentos_entre_comidas__Sometimes	0.3477	0.105	3.323	0.001	0.142	0.553
alimentos_entre_comidas__no	0.8713	0.148	5.879	0.000	0.581	1.162
consumo_alcohol__Frequently	2.1920	0.163	13.422	0.000	1.872	2.512
consumo_alcohol__Sometimes	1.7643	0.142	12.408	0.000	1.485	2.043
consumo_alcohol__no	1.7066	0.140	12.195	0.000	1.432	1.981

Este tiene un R2 de 87,7%, al implementar el modelo con los componentes PCA, el R2 es del 40% aproximadamente, los que nos da indicios de que los componentes no son adecuados para el problema planteado.

Con la siguiente grafica deberíamos poder indicar que las variables influyen y hacen separación de los datos, pero visualmente no es consistente.