Práctica 2: PCA

Román Alberto Vélez Jiménez

CU: 165462

Fecha: 27 Oct 23

Problema: de la base de datos red wine (!https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009/) de kaggle, se hará un análisis PCA. El objetivo es poder observar si es posible reducir la dimensionalidad de los datos y poder clasificarlos en base a la calidad del vino.

Red Wine Quality

Los Datos

Se observan 12 columnas y 1599 filas, donde la columna de calidad es la variable objetivo. Ninguna variable contiene vacíos. Las variables son las siguientes:

- fixed acidity: acidez fija
- volatile acidity: acidez volátil
- citric acid: ácido cítrico
- residual sugar: azúcar residual
- chlorides: cloruros
- free sulfur dioxide: dióxido de azufre libre
- total sulfur dioxide: dióxido de azufre total
- density: densidad
- pH: pH
- sulphates: sulfatos
- alcohol: alcohol
- quality: calidad (puntuación entre 3 y 8)

Un ejemplo de las primeras 5 observaciones se puede ver en la siguiente tabla

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

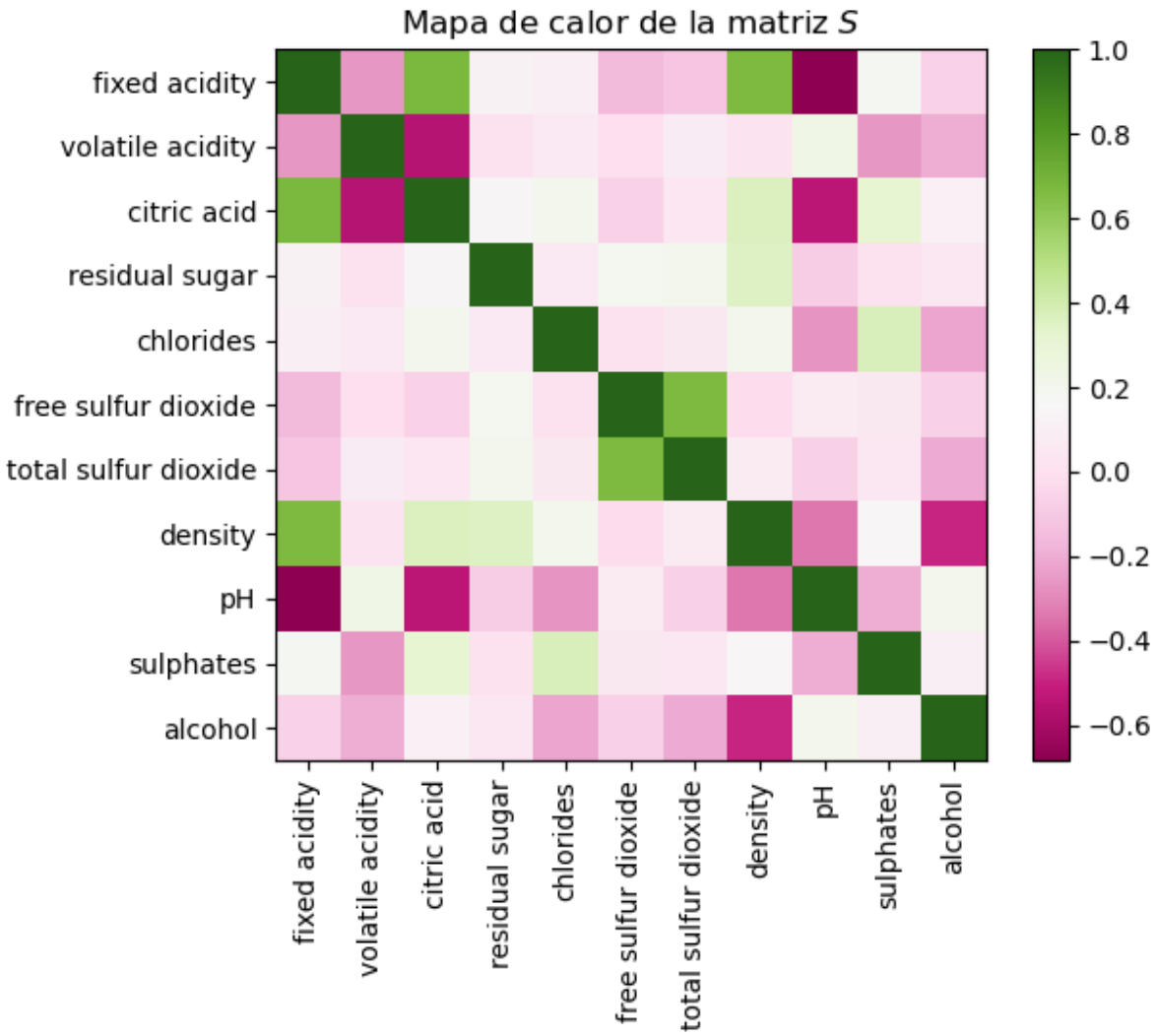
Preparando los Datos

Para hacer el análisis PCA, primero se estandarizó la data para que tenga media 0 y desviación estándar 1. Esto se hace para que las variables tengan el mismo peso en el análisis. Obtenemos la matriz de covarianza usando el sesgo de Besel:

$$\frac{1}{n-1}(X-\mu)^T(X-\mu).$$

Mapa de Calor

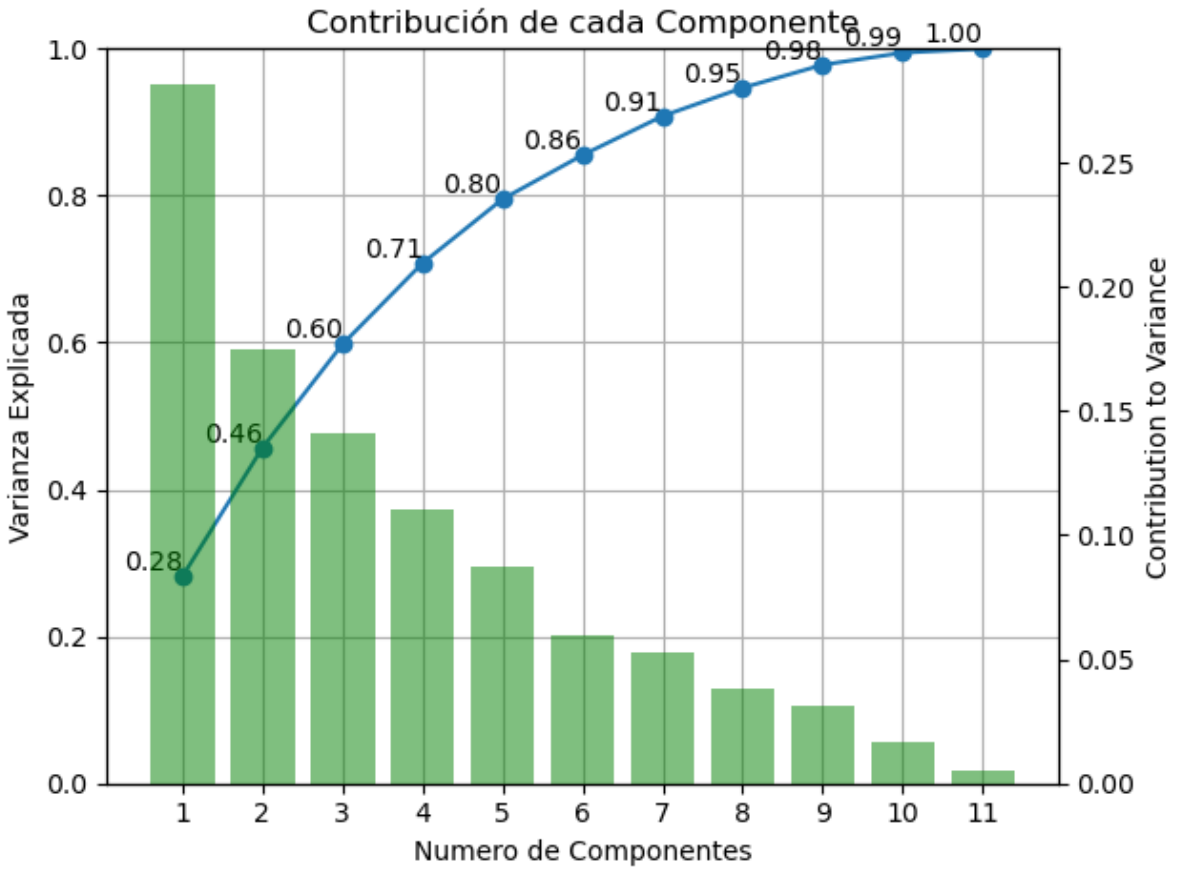
Para entender las relaciones entre variables, se muestra en la figura de abajo la matriz de varianzas y covarianzas. Se observa una fuerte correlación entre `free sulfure dioxide` y `total sulfure dioxide` , por lo que es muy probable que están correlacionadas, por lo que seguramente una componente principal tendrá muy poca varianza explciada gracias a la colinealidad de estas dos variables. De igual forma vemos la fuerte relación entre `volatile acidity` y `citric acid` .



PCA

Cálculo de los valores y vectores propios

Pimero se hizo el cálculo de eigenvectores a partir de la matriz de varianzas y covarianzas S. Se obtuvieron 12 eigenvectores y eigenvalores. Observamos que la varianza explicada para las variables 1 y 2 es moderadamente buena, pues con solo dos variables se logra explicar casi el 50% de la varianza de 12 variables. Por lo que podemos decir que estas variables son las que más explican la varianza de los datos. De todos modos, observamos como se ve con 3 componentes, las cuales ya explican el 60% de la varianza total. Esto se puede observar en la figura de abajo.



Contribución de las Variables

La primera componente podría explicar el tradeoff entre acidez y alcanilidad, mientras que la segunda componente podría explicar el tradeoff entre alcohol y carbonatado. En la tabla de acontinuación se puede observar la contribución de cada variable a cada componente.

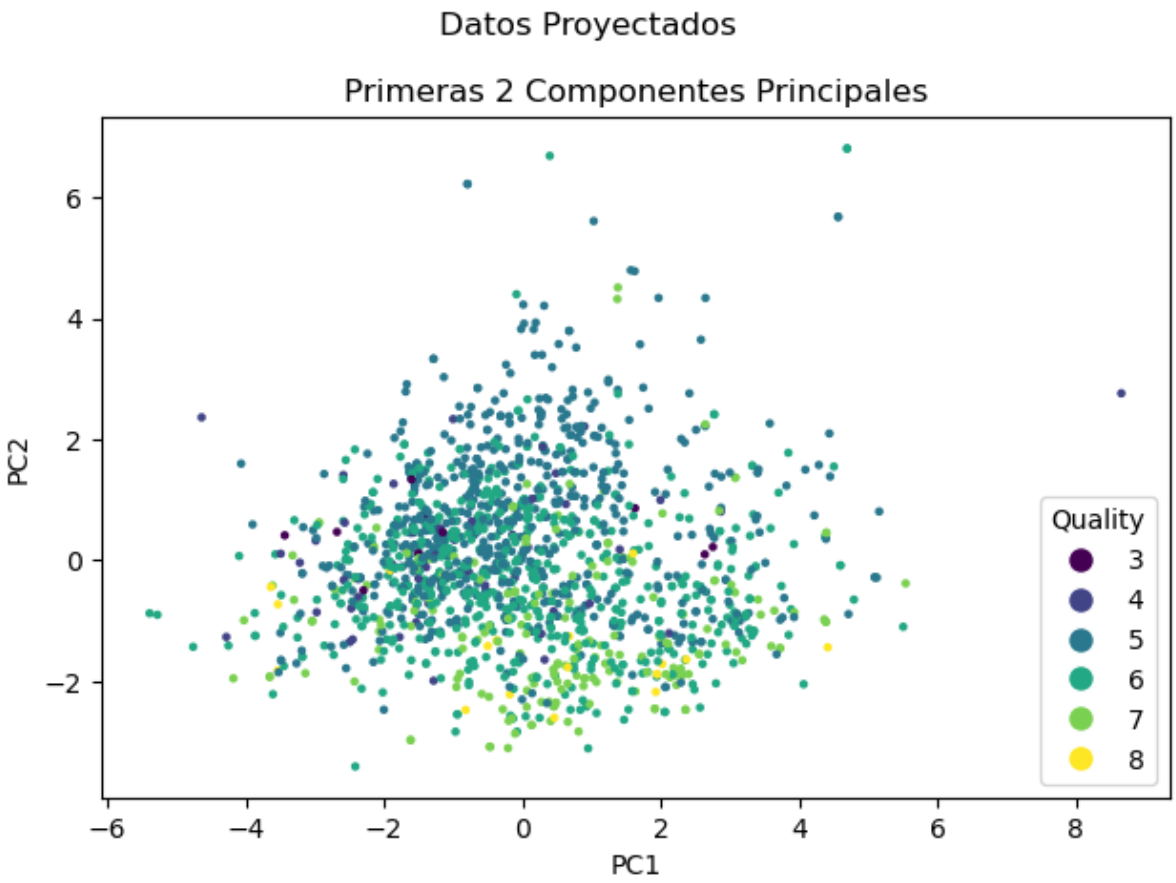
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
fixed acidity	0.4893	-0.1105	-0.1233	-0.2296	-0.0826	-0.6397	-0.2495	0.1940	-0.1776	-0.3502	0.1015
volatile acidity	-0.2386	0.2749	-0.4500	0.0790	0.2187	-0.0024	0.3659	-0.1291	-0.0788	-0.5337	0.4114
citric acid	0.4636	-0.1518	0.2382	-0.0794	-0.0586	0.0709	0.6217	-0.3814	-0.3775	0.1055	0.0696
residual sugar	0.1461	0.2721	0.1013	-0.3728	0.7321	-0.1840	0.0929	0.0075	0.2998	0.2907	0.0492
chlorides	0.2122	0.1481	-0.0926	0.6662	0.2465	-0.0531	-0.2177	0.1113	-0.3570	0.3704	0.3043
free sulfur dioxide	-0.0362	0.5136	0.4288	-0.0435	-0.1592	0.0514	0.2485	0.6354	-0.2048	-0.1166	-0.0140
total sulfur dioxide	0.0236	0.5695	0.3224	-0.0346	-0.2225	-0.0687	-0.3708	-0.5921	0.0190	-0.0937	0.1363
density	0.3954	0.2336	-0.3389	-0.1745	0.1571	0.5673	-0.2400	0.0207	-0.2392	-0.1705	-0.3912
pH	-0.4385	0.0067	0.0577	-0.0038	0.2675	-0.3407	-0.0110	-0.1677	-0.5614	-0.0251	-0.5221
sulphates	0.2429	-0.0376	0.2798	0.5509	0.2260	-0.0696	0.1123	-0.0584	0.3746	-0.4475	-0.3813
alcohol	-0.1132	-0.3862	0.4717	-0.1222	0.3507	0.3145	-0.3030	0.0376	-0.2176	-0.3277	0.3616

Vemos que en la primera componente, la `fixed acidity` es la variable que más aporta información. Por otro lado, la que más aporta en la segunda componente es `total sulfur dioxide` . Esto nos indica que estas dos variables son las que más variación aportan al modelo.

Pareciera, por la matriz de varianzas y covarianzas S, que la variable `fixed acidity` es la que tiene mayor correlación con las demás variables, por lo que es la que más contribuye a la varianza de los datos.

Valores Proyectados

Al graficar las primeras 2 componentes principales, se puede observar que las componentes principales no son capaces de separar del todo las clases de vino por calidad. Si bien se ve un poco de orden creciente de arriba hacia abajo, la separación no es evidente. Esto se puede observar en la última figura.



Referencias

- Analisis Numerico, Erick Palacios, 2022 (!https://itam-ds.github.io/analisis-numerico-computo-cientifico/README.html)