Reducción de Dimensionalidad en el conjunto de datos de la Calidad del Vino a través de PCA

Alejandro Villazón Gutiérrez
Análisis Estadístico Multivariado MAT269
Universidad Técnica Federico Santa María

Abstract

Este reporte presenta un análisis de la calidad del vino utilizando el método Análisis de Componentes Principales (PCA, por sus siglas en inglés) aplicado a un conjunto de datos de vinos blancos y tintos de la región Vinho Verde en Portugal. El conjunto de datos incluye diversas características químicas y sensoriales de los vinos, como acidez, contenido de azúcar residual y alcohol. Al aplicar PCA, nuestro objetivo fue identificar los componentes químicos clave que contribuyen a la calidad del vino y comprender sus relaciones. Los resultados revelaron como era esperado que los primeros componentes principales capturaron una parte significativa de la varianza, brindando información valiosa sobre los factores clave que influyen en la calidad del vino. Este análisis contribuye al campo de la enología al mejorar nuestra comprensión de las características químicas que definen la calidad del vino y tiene implicancias para los productores y consumidores de vino en su búsqueda de una producción y selección de vinos mejorada.

1 Introducción

La calidad del vino es un aspecto fundamental tanto para los productores como para los consumidores, ya que influye en su sabor, aroma y disfrute general. La evaluación de la calidad del vino se basa en numerosos factores, incluyendo propiedades químicas como el contenido de alcohol, la acidez volátil, el pH y los niveles de azúcar residual. La comprensión de la relación entre estas características químicas y la calidad del vino es crucial para mejorar los procesos de producción y ofrecer productos de mayor calidad.

En este reporte, se aplica el análisis de componentes principales (PCA) al conjunto de datos de calidad del vino de la UCI Machine Learning Repository proporcionado en Kaggle [2]. El PCA es una técnica de reducción de dimensionalidad que permite resumir la información contenida en múltiples variables en un número menor de componentes principales. Al aplicar PCA a este conjunto de datos, se busca identificar las combinaciones lineales de características químicas que mejor explican la varianza en las características que definen la calidad del vino.

Este informe está organizado de la siguiente manera: se proporcionará una breve descripción del conjunto de datos. Luego, en la sección de Metodología, se describirán los pasos seguidos para llevar a cabo el análisis de PCA, incluyendo el preprocesamiento de los datos. A continuación, en la sección de Resultados, se presentarán los resultados obtenidos después de aplicar PCA al conjunto de datos de calidad del vino, se incluirán gráficos y tablas que resuman la varianza explicada, y las influencias en las componentes principales. Finalmente, en la sección de Conclusiones, se analizarán los resultados encontrados para seleccionar la cantidad adecuada de componentes principales y se realizará su interpretación correspondiente.

El análisis de PCA aplicado al conjunto de datos de calidad del vino tiene el potencial de proporcionar una comprensión más profunda de las características químicas que determinan la calidad del vino según nuestro conjunto de datos. Esto puede ser valioso tanto para los productores en la mejora de sus procesos de producción, como para los consumidores en la selección de vinos de mayor calidad. A través de este estudio, se espera obtener conocimientos significativos que contribuyan como ejemplo básico al campo de la enología y al análisis de datos en general.

2 Conjunto de datos

El conjunto de datos recopila información química y sensorial de vinos blancos y tintos de la región de Vinho Verde en Portugal. Fue creado con el propósito de analizar y comprender las propiedades y características que influyen en la calidad de estos vinos [1].

La versión del conjunto de datos utilizado [2] combina la información de ambos tipos de vinos. No se distingue entre vino tinto o vino blanco, pues nuestro objetivo es comprender de forma general las características que mejor explican la calidad del vino.

Este conjunto de datos contiene 6.497 observaciones y 12 atributos que describen diferentes aspectos de los vinos. A continuación, se detallan las variables incluidas:

1. *fixed acidity*: la cantidad de ácido tartárico en el vino (g/dm^3).
2. *volatile acidity*: la cantidad de ácido acético en el vino (g/dm^3).
3. *citric acid*: la cantidad de ácido cítrico en el vino (g/dm^3).
4. *residual sugar*: la cantidad de azúcar residual en el vino (g/dm^3).
5. *chlorides*: la cantidad de sales minerales en el vino (g/dm^3).
6. *free sulfur dioxide*: la cantidad de dióxido de azufre libre en el vino (mg/dm^3).
7. *total sulfur dioxide*: la cantidad total de dióxido de azufre en el vino (mg/dm^3).
8. *density*: la densidad del vino (g/cm^3).
9. *pH*: el nivel de acidez o alcalinidad del vino.
10. *sulphates*: la cantidad de sulfatos en el vino (g/dm^3).
11. *alcohol*: el contenido de alcohol en el vino ($\%vol.$).
12. *quality*: la calidad del vino, evaluada en una escala del 0 al 10.

3 Metodología

Para llevar a cabo el trabajo descrito en este reporte, se utilizó el lenguaje de programación Python, en particular, se emplearon las bibliotecas Scikit-Learn, Pandas y Matplotlib para el procesamiento de datos y la visualización de resultados. En el anexo, se encuentra un enlace al cuaderno de códigos desarrollado, donde se encuentran detalladas las implementaciones realizadas.

A continuación, se describe el procedimiento utilizado para aplicar PCA al conjunto de datos:

En primer lugar, se llevó a cabo el procesamiento de los datos. Se identificó la presencia de valores faltantes en el conjunto de datos. Ante este problema, se consideraron diversas estrategias para abordar el problema, como la eliminación de las observaciones afectadas o el relleno de los valores faltantes. Con el objetivo de preservar la integridad de las observaciones, se optó por la última opción y se decidió rellenar los valores faltantes utilizando la media de la variable correspondiente.

Posteriormente, se procedió al cálculo y visualización de las correlaciones entre las variables del conjunto de datos (ver Figura 1). Durante este análisis, se observó que existen relaciones significativas de alta correlación entre algunas variables. Esta observación refuerza nuestra decisión de aplicar PCA, ya que indica la presencia de información superpuesta o redundante en los datos. Al implementar PCA, buscamos capturar la variabilidad esencial de las características químicas del vino, eliminando la redundancia causada por la alta correlación y permitiendo una representación más compacta y significativa de estas características. La Figura 1 ofrece una visualización clara de las relaciones de correlación entre las variables. Los coeficientes de correlación se representan mediante una escala de colores, donde los valores más altos se presentan con tonos más cálidos.

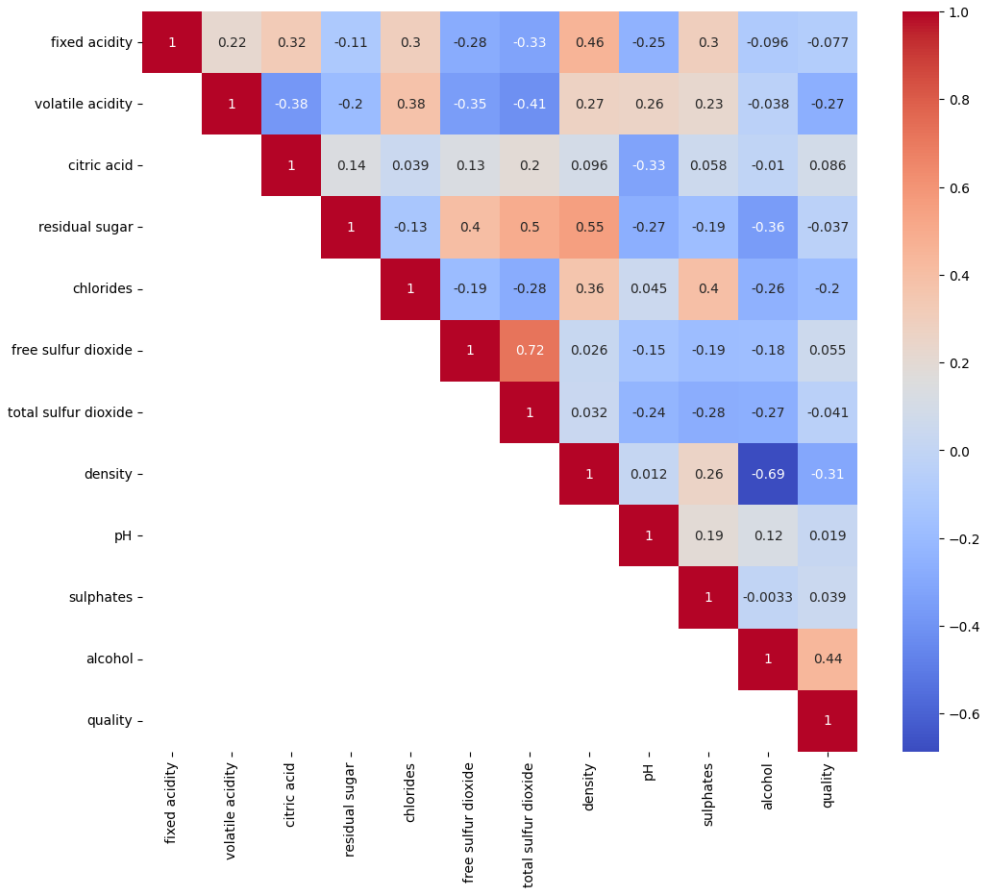

Figure 1: Correlación entre variables.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates
Media	7.22	0.34	0.32	5.44	0.06	30.53	115.74	0.99	3.22	0.53
Varianza	1.68	0.03	0.02	22.63	0.0	315.04	3194.72	0.0	0.03	0.02

Table 1: Estadísticas descriptivas aproximadas de las variables.

Continuando con el desarrollo de nuestro análisis, se calculó las medias y desviaciones estándar de las variables (ver Tabla 1), notando la necesidad de realizar una estandarización de las variables, dada la gran diferencia de escalas. Este paso es crucial antes de aplicar PCA, ya que asegura que todas las variables tengan la misma escala y evita que alguna variable tenga un peso desproporcionado en el análisis debido a diferencias en las unidades de medida. Al estandarizar las variables, se logra una comparabilidad adecuada y se garantiza que la variabilidad total de los datos sea correctamente representada por las componentes principales extraídas.

Una vez que las variables fueron estandarizadas, se procedió a aplicar PCA al conjunto de datos.

4 Resultados

En esta sección, presentamos los resultados obtenidos al aplicar PCA al conjunto de datos y determinar la cantidad óptima de componentes a retener que capturen la mayor variabilidad de los datos.

En primer lugar, se presenta en la Tabla 2 la varianza explicada por componente, donde se detalla el porcentaje de varianza capturada por cada componente. Como se puede observar en la tabla, los componentes están ordenados de forma ascendente según su contribución a la varianza total. Esta información nos permite evaluar la importancia relativa de cada componente en la explicación de la variabilidad de los datos.

Componente	1	2	3	4	5	6	7	8	9	10	11
% de Varianza Explicada	27.54	22.67	14.13	8.83	6.55	5.52	4.76	4.56	3.07	2.07	0.30

Table 2: Porcentaje de Varianza explicada por componente

Observamos que los primeros componentes principales explican una gran proporción de la varianza total, las primeras dos componentes explican aproximadamente el 50% de la varianza total, mientras que la suma de componentes adicionales proporciona una mejora pequeña, pero aun considerable en la explicación de la varianza.

Además, se presenta un gráfico de varianza acumulada en la Figura 2, el cual muestra cómo se acumula la varianza a medida que se añaden más componentes. Este gráfico es útil para determinar el número óptimo de componentes a retener, considerando un umbral de varianza acumulada preestablecido. Observando el gráfico, podremos identificar el punto en el que se alcanza o supera el umbral establecido, lo que nos permitirá seleccionar un número adecuado de componentes para continuar con nuestro análisis.

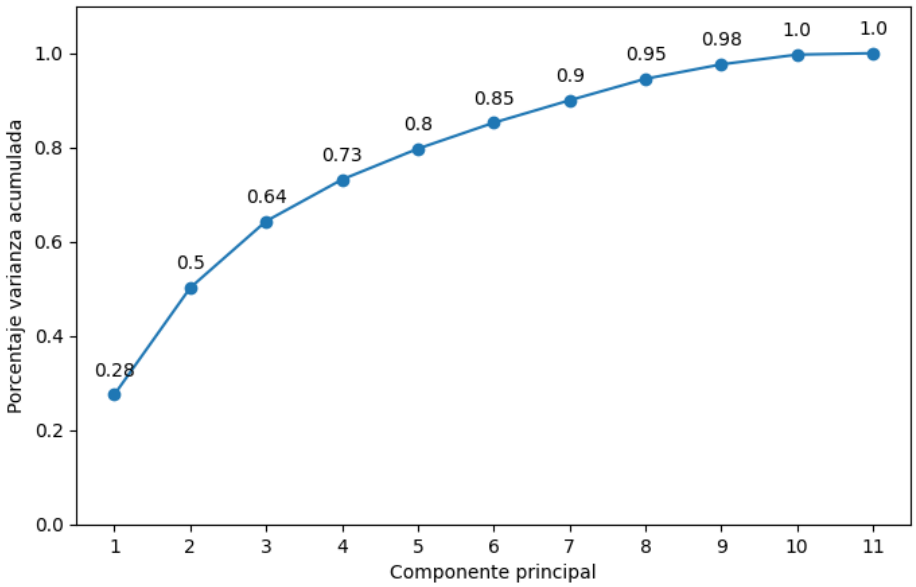

Figure 2: Varianza explicada acumulada.

Por último, se presenta en la Figura 3 un gráfico tipo "heatmap" que muestra las influencias de cada variable en las componentes. Este gráfico nos permite visualizar de manera intuitiva cómo cada variable contribuye a la formación de las componentes principales a través de los pesos asociados. Mediante intensidades, el "heatmap" resalta las relaciones entre las variables y las componentes, revelando patrones de correlación y ayudándonos a comprender mejor la estructura de los datos.

En conjunto, estos elementos proporcionan una visión completa de los resultados obtenidos al aplicar PCA al conjunto de datos, permitiéndonos comprender la importancia de cada componente, la varianza acumulada y las influencias de las variables en las componentes.

5 Conclusiones

Luego de analizar los resultados obtenidos mediante el PCA en el conjunto de datos de vinos, se procedió a seleccionar el umbral de varianza acumulada con el objetivo de determinar el número óptimo de componentes principales.

Después de revisar detalladamente los porcentajes de varianza acumulada por cada componente, se concluyó que seleccionar un umbral del 80% de varianza acumulada consigue un equilibrio óptimo entre la reducción significativa de la dimensionalidad del conjunto de datos y la retención de una cantidad considerable de varianza explicada.

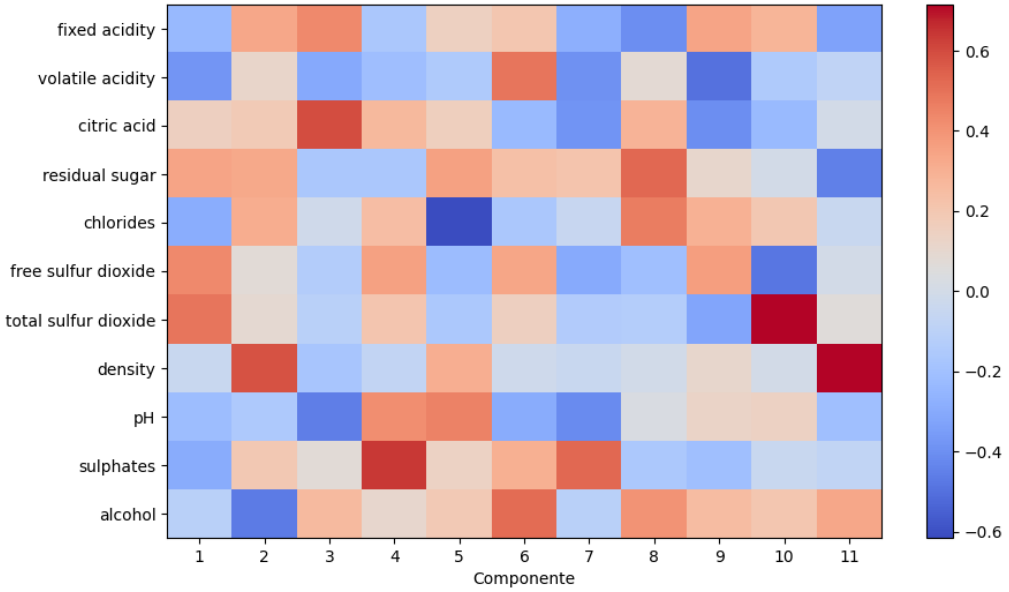

Figure 3: Influencia de las variables en las componentes.

De la Figura 2 observamos que dado el umbral escogido debemos retener las primeras cinco componentes principales.

Analizando el gráfico de influencias presentado en la Figura 3 se ha observado que ciertas variables tienen una mayor influencia en cada una de las componentes principales retenidas. Estas observaciones sugieren relaciones significativas entre las variables y proporcionan información clave sobre las características más relevantes de cada componente.

- En la primera componente principal, las variables *volatile acidity*, *free sulfur dioxide* y *total sulfur dioxide* han demostrado ser las más influyentes. Esto indica una fuerte relación entre la acidez volátil y las concentraciones de dióxido de azufre en los vinos analizados, lo cual puede tener implicaciones importantes en términos de la calidad y las propiedades sensoriales de los vinos.
- La segunda componente principal está fuertemente influenciada por las variables *density* y *alcohol*. Estas variables desempeñan un papel importante en la determinación de la densidad y el contenido de alcohol en los vinos, lo cual puede tener un impacto significativo en la percepción del consumidor.
- En la tercera componente principal, se observa un mayor aporte de las variables *citric acid* y *pH*. Estas variables están relacionadas con aspectos como la acidez cítrica y el nivel de alcalinidad en los vinos, lo cual puede ser relevante para la evaluación de la frescura y el equilibrio de los mismos.
- En la cuarta componente principal, la variable *sulphates* predomina, lo cual mantiene información sobre los niveles de sulfatos y ciertas características de los vinos, como su capacidad antioxidante o estabilidad química.
- Por último, la quinta componente principal muestra una mayor influencia de la variable *chlorides*. Esto indica que los niveles de cloruros pueden ser un factor importante a considerar en el análisis de la salinidad o el sabor salado en los vinos.

En conjunto, estas cinco componentes principales capturan una gran cantidad de información de la mayoría de las variables en el conjunto de datos de vinos, lo cual nos permite comprender mejor las principales fuentes de variabilidad y resaltar las variables más relevantes en el análisis.

En conclusión, el análisis de PCA ha permitido identificar las combinaciones lineales de características químicas más importantes para explicar la variabilidad en la calidad del vino. Estos hallazgos podrían tener implicancias tanto para los productores de vino, al mejorar sus procesos de producción, como para los consumidores, al seleccionar vinos de mayor calidad, centrándose en las características más influyentes mencionadas en este reporte.

6 Anexo

Con el objetivo de garantizar la reproducibilidad de los resultados presentados en este reporte, adjuntamos el enlace al cuaderno que contiene los códigos utilizados.

7 Referencias

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47*(4), 547-553.

[2] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J (2009). Wine Quality. UCI Machine Learning Repository. https://doi.org/10.24432/C56S3T. Recuperado de Kaggle.