

30 DE MARZO DE 2023

# PRÁCTICA 6

Análisis Exploratorio de Datos y Análisis de Componentes Principales



## Objetivo de la práctica

Por Angel Damian Monroy Mendoza

No. de Cuenta: 316040707

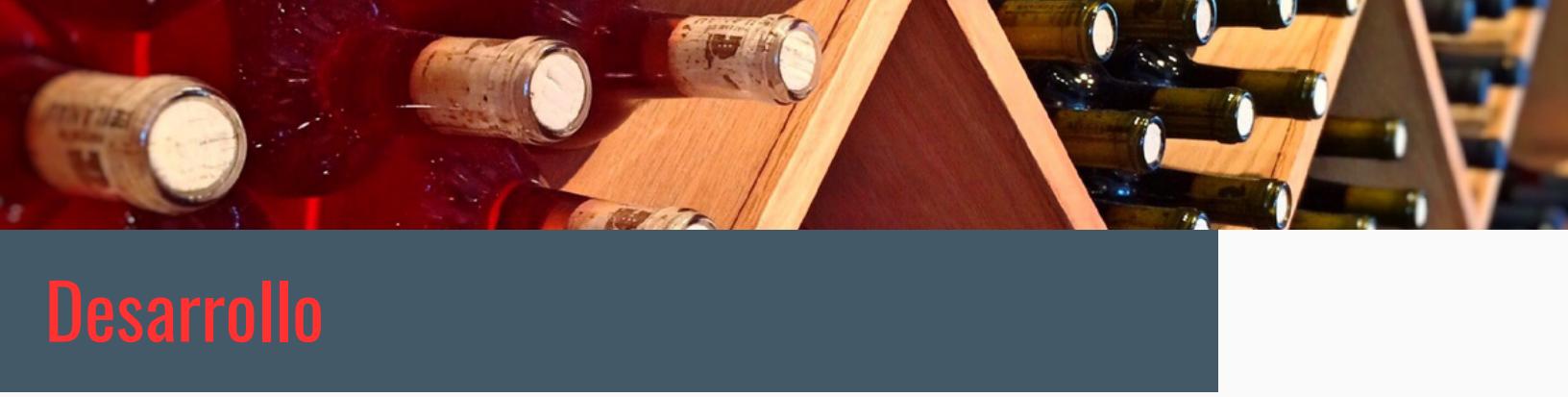
identificar la variable objetivo y posibles técnicas de modelado. Realizar un análisis de los datos donde podamos identificar su estructura, datos faltantes, datos atípicos; realizar una preparación de los datos y realizar una selección de características a través del Análisis de Componentes Principales (PCA).



## Características:

- Existen dos conjuntos de datos están relacionados con variantes rojas y blancas del vino portugués "Vinho Verde". Hoy en día, el vino es disfrutado cada vez más por una gama más amplia de consumidores. El "Vinho Verde" de Portugal ha tenido crecimiento en sus exportaciones, por lo que se vuelve de interés el garantizar la calidad el mismo.
- A continuación, se muestra un diccionario de las variables relevantes para esta práctica

1. - fixed acidity (tartaric acid - g / dm<sup>3</sup>)
2. - volatile acidity (acetic acid - g / dm<sup>3</sup>)
3. - citric acid (g / dm<sup>3</sup>)
4. - residual sugar (g / dm<sup>3</sup>)
5. - chlorides (sodium chloride - g / dm<sup>3</sup>)
6. - free sulfur dioxide (mg / dm<sup>3</sup>)
7. - total sulfur dioxide (mg / dm<sup>3</sup>)
8. - density (g / cm<sup>3</sup>)
9. - pH
10. - sulphates (potassium sulphate - g / dm<sup>3</sup>)
11. - alcohol (% by volume)
12. - quality (score between 0 and 10)



# Desarrollo

Para el desarrollo de esta práctica dividimos el proceso en cinco secciones:

1. Importación de las bibliotecas y datos.
2. Paso 1: Descripción de la estructura de los datos
3. Paso 2: Identificación de datos faltantes.
4. Paso 3: Detección de valores atípicos.
5. Paso 4: Identificación de relaciones entre pares variables.
6. Análisis de Componentes Principales y Conclusiones.

## 1. Bibliotecas y Datos

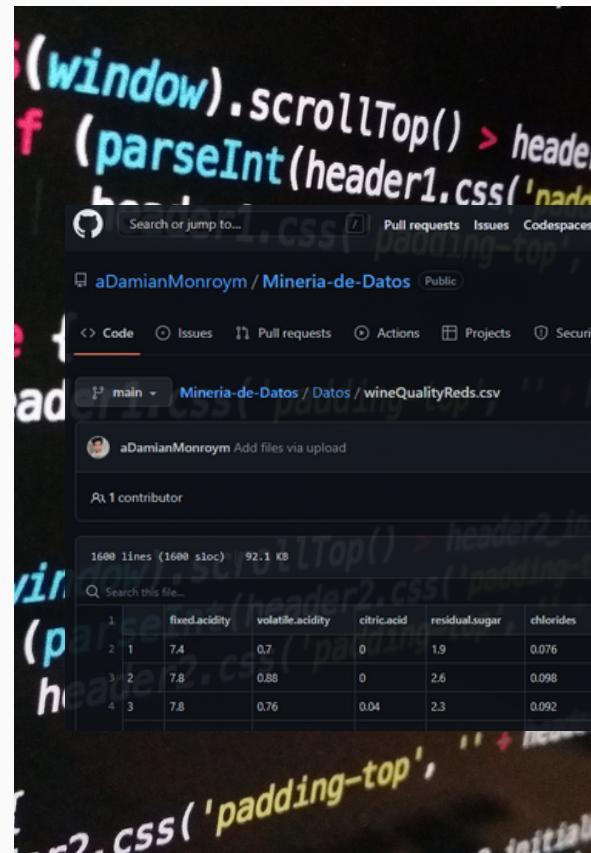
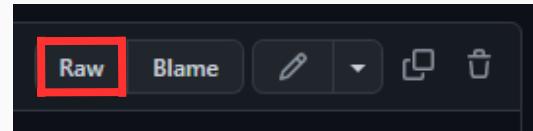
En este primer apartado se importó la biblioteca de pandas para la manipulación y análisis de los datos, la biblioteca de numpy para crear vectores y matrices de n dimensiones, la biblioteca de matplotlib.pyplot para la generación de gráficas a partir de los datos y la biblioteca de seaborn que, de igual forma, corresponde a la visualización de los datos.

Posteriormente, para la parte de cargar los datos se tuvo como requerimiento el realizar una conexión entre una plataforma Git y python. En este sentido, como la plataforma que se eligió fue GitHub, se creó un repositorio de GitHub que contendrá todos los archivos de datos con extensión “csv”, así como las prácticas que se realicen durante la materia. Así pues, a este repositorio se le añadió el conjunto de datos correspondientes a “Wine Quality” de Kaggle,

Estos datos se leyeron mediante el siguiente link que se obtiene desde el archivo de GitHub, el cuál se instanció en la variable 'url' y se le pasó como argumento a la función de pandas para leer archivos con extensión csv.

## 2. Paso Uno: Descripción de la Estructura de los Datos

En este primer paso, para poder describir los datos de una forma correcta, necesitamos conocer la dimensión de nuestro conjunto, así como los tipos de datos con los que se cuenta, pues de esta forma podemos establecer un punto de partida que nos ayudará a comprender el comportamiento de los datos.



# Desarrollo

## 2. Paso Uno: Descripción de la Estructura de los Datos

Ahora bien, para saber la dimensión de nuestro conjunto de datos bastó con utilizar el atributo shape de Pandas, que proporciona la estructura general de los datos debido a que devuelve la cantidad de filas y columnas que tiene el conjunto de datos, que en este caso fue de 1599 filas por 12 columnas.

En el caso de los tipos de datos que se tienen, el atributo dtypes muestra los tipos de datos de las variables, en donde se observó que el conjunto de datos tiene una combinación de variables categóricas (objeto) y numéricas (flotante e int).

## 3. Paso Dos: Identificación de Datos faltantes

Para este paso también se tienen dos formas para poder realizarlo. La primera de ellas es mediante la función de Pandas isnull().sum() que regresa la suma de todos los valores nulos (faltantes) de cada variable. Aquí podemos observar que ninguna de las columnas tuvo suma de valores nulos diferentes de cero.

La otra forma es a través de la función info() para obtener el tipo de datos y la suma de valores nulos. En pocas palabras, con esta función podemos saber el tipo de datos que tenemos (Paso Uno) y si es que existen valores nulos (Paso Dos), sin embargo, la información que nos aporta es poco directa, pues no nos dice la cantidad de datos faltantes, sino la cantidad de datos no nulos, lo cuál puede dificultar el análisis de los datos.

En este ejemplo no encontramos valores faltantes que pudiéramos especular si podemos deshacernos de ellos o no, pues si tuviéramos una cantidad de valores nulos considerable (20%, 30% o 50%) de los datos, los modelos resultarían sesgados por la falta de información; encontrándonos con uno de los tres grandes retos de la minería de datos en la actualidad.

```
Vinos.shape  
(1599, 12)  
Vinos.dtypes  
fixed.acidity    float64  
volatile.acidity float64  
citric.acid     float64  
residual.sugar   float64  
chlorides        float64
```

```
Vinos.isnull().sum()  
fixed.acidity      0  
volatile.acidity   0  
citric.acid       0  
residual.sugar    0  
chlorides          0  
free.sulfur.dioxide 0  
total.sulfur.dioxide 0  
density            0  
pH                 0  
sulphates          0  
alcohol             0  
quality             0  
dtype: int64
```

# Desarrollo

## 4. Paso Tres: Detección de Valores Atípicos

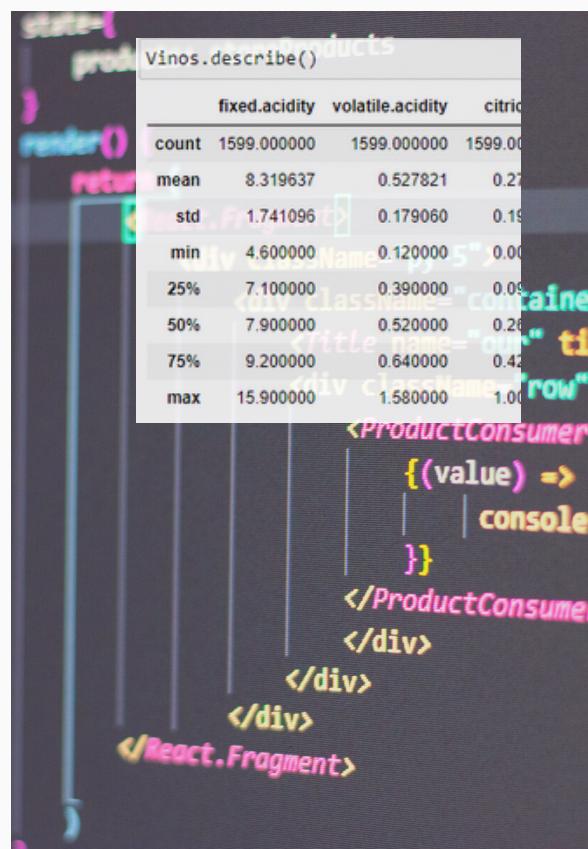
En este paso, tenemos varias formas de identificar valores atípicos, pues podemos hacerlo mediante gráficas de distribución, gráficas de bigote y mediante resúmenes estadísticos, tanto para las variables numéricas, como para las categóricas.

La razón por la que se utilizan las gráficas de distribución es para tener una idea de cómo se están comportando los datos o con qué frecuencia ocurren. Para las variables numéricas, se observa cuántas veces aparecen grupos de números en una columna. Mientras que, para las variables categóricas, son las clases de cada columna y su frecuencia.

Específicamente, se utilizan histogramas porque agrupan los números en rangos acotados que permiten realizar un análisis previo; aquí la altura de una barra muestra cuántos números caen en ese rango, en donde se emplea la función `hist()` para trazar el histograma de las variables numéricas.

Aquí logramos observar que '`citric.acid`', '`residual.sugar`', '`chlorides`' y '`total.sulfur.dioxide`' son variables que están sesgadas a la izquierda, lo cual podría indicar que hay valores atípicos en estas características, además de que '`total.sulfur.dioxide`' tiene la particularidad de contar con un límite en 250, cuando el extremo derecho de la distribución de los datos acaba entre 150 y 200, lo cual es extraño. Por lo que, lo más probable es que ese dato sea un dato erróneo y que si se deja, afectará considerablemente los resultados.

Una forma de poder investigar estos datos atípicos es mediante un resumen estadístico que se obtiene mediante la función `'.describe()'`. Este resumen se muestra mediante una tabla en donde se muestra un recuento, la media, la desviación, el valor mínimo, el valor máximo y los percentiles 25%, 50% (mediana) y 75% y aquí observamos que nos ayuda a identificar las variables con valores nulos, aunque en este caso no contemos con alguna.



# Desarrollo

## 4. Paso Tres: Detección de Valores Atípicos

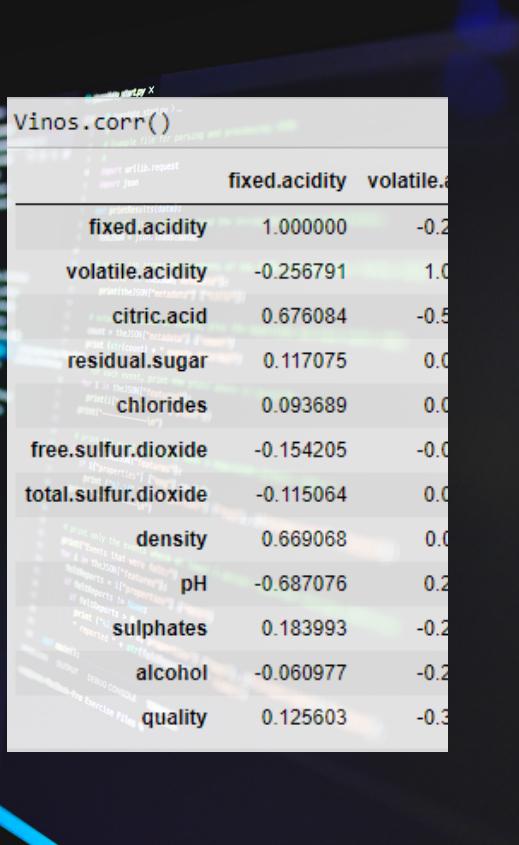
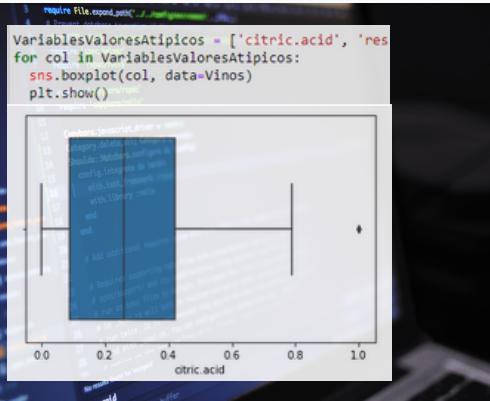
Otra herramienta que nos permite observar con más detalle nuestras sospechas de valores atípicos son los diagramas de cajas o de bigote, pues aquí se muestran explícitamente todos los valores atípicos que salen de la varianza esperada de la variable, por lo que podemos confirmar que existen valores atípicos en las variables de 'citric.acid', 'total.sulfur.dioxide', no obstante, como en las variables 'residual.sugar' y 'chlorides' no se tiene una separación notable de los supuestos valores atípicos, se toman como valores que son coherentes y significativos para nuestro conjunto de datos.

Por último, aunque no se tengan variables categóricas, cabe mencionar que este proceso también se hace para dichas variables en las cuales, los resúmenes estadísticos muestran números reales que ayudan a obtener información puntual de estas variables, mientras que las gráficas sí nos ayuda a tener una idea general de las distribuciones categóricas.

## 5. Paso Cuatro: Identificación de Relaciones entre Variables

En este paso se analiza la correlación que existe entre las variables. Esta correlación es un valor que está entre -1 y 1 y explica qué tanto aumenta (o disminuye) una variable mientras que la otra aumenta (o disminuye). En dado caso que el valor de la correlación sea cercano a 0, es que no varían de la misma forma dichas variables.

En este sentido, una forma de saber las relaciones que hay es mediante una matriz de correlación que se basa en el método de Pearson y se obtiene con la función 'corr()'. A pesar de que esta matriz nos ayuda a saber las relaciones que hay entre variables numéricas, se vuelve confuso identificar estas relaciones cuando se tienen muchas variables, por lo que se auxilia del mapa de calor de Seaborn, mostrando la diagonal inferior de la matriz con la función 'sns.heatmap()' y pasando como parámetros la matriz de correlación de nuestro conjunto de datos, los colores que se quieren utilizar, la matriz triangular que se quiere mostrar y los valores de la correlación en cada celda.



# Desarrollo

## 6. Análisis de Componentes Principales

Como parte final del desarrollo de la práctica se realizó un Análisis de Componentes Principales. Recordando la teoría, este método cuenta con cinco pasos: 1.- Evidencia de correlaciones entre las variables, 2.- Estandarizar los datos, 3.- Calcular matriz de covarianza o correlaciones, 4.- Calcular eigen-valores y eigen-vectores a partir de la matriz anterior y 5.- Decidir el número de componentes principales. En este sentido, lo que se hizo fue lo siguiente:

- Paso uno: Como ya habíamos visto en el mapa de calor, comprobamos que sí existen correlaciones entre pares de variables, por lo que se toma como cumplido este paso.
- Paso dos: En este paso se importaron las funciones correspondientes al ACP y a la estandarización de los datos. Asimismo, se instanció el objeto StandardScaler (*Escalado*) y se ajustó este objeto a la matriz de Vinos, guardándose en una variable de nombre MEstandarizada.
- Paso tres y cuatro: Aquí se instanció el objeto PCA con el parámetro de '*n\_components=None*' que significa que agarra como componentes todas las variables de nuestro conjunto. Después, este objeto se ajusta a la matriz estandarizada para obtener los componentes y se imprimen.
- Paso cinco: Se instancia el objeto PCA con su atributo que muestra la proporción de varianza y se hace la suma de la varianza de 6 componentes, pues se tiene el 87% de varianza acumulada y con 7 el 91%. Posteriormente, se grafica la varianza acumulada de los componentes para visualizar mejor nuestra decisión.

## Conclusiones

De esta práctica podemos concluir que el *Análisis Exploratorio de Datos* es una parte fundamental tanto en la minería de datos, como en la aplicación de algoritmos de inteligencia artificial, por las razones vistas en la primera práctica. A su vez, al implementar el proceso para el Análisis de Componentes Principales, podemos concluir que es una herramienta que ayuda a tener un fundamento matemático para suprimir variables que no representan una alta "Carga" de información. Pues en esta práctica tuvimos que hacer una poda del 48% de Carga de las variables, para quedarnos con la mitad de variables en nuestro conjunto de datos (6). Además de servir para discriminar variables que, aunque entran en el rango de la poda, puedes quitar la de menor carga, si es que éstas estaban correlacionadas.



	fixed.acidity	residual.sugar	chlorides	free.sulfur.dioxide	sulphates	alcohol
1	7.4	1.9	0.076	11.0	0.56	9.4
2	7.8	2.6	0.098	25.0	0.68	9.6
3	7.8	2.3	0.092	15.0	0.65	9.8
4	11.2	1.9	0.075	17.0	0.58	9.8
5	7.4	1.9	0.076	11.0	0.56	9.4
...	...	...	...	...	...	...
1595	6.2	2.0	0.090	32.0	0.58	10.5
1600	6.0	2.2	0.062	30.0	0.76	11.2