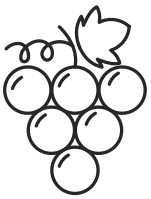


Abstract

Este proyecto se enfoca en analizar cómo las propiedades fisicoquímicas del vino influyen en su calidad. Utilizando un dataset descargado de Kaggle, se examinan las relaciones entre variables como acidez fija, acidez volátil, contenido de azúcar residual, cloruros, entre otras, y la calidad del vino, la cual se clasificará en tres categorías: malo, regular y bueno. La hipótesis principal es que ciertas propiedades fisicoquímicas están significativamente correlacionadas con la calidad del vino. A través de visualizaciones y análisis numéricos, se buscará identificar patrones que puedan ser utilizados para desarrollar un modelo predictivo capaz de estimar la calidad del vino. Para explorar y presentar los hallazgos, se emplearán herramientas de visualización como Seaborn y Matplotlib, facilitando así una interpretación clara y comprensible de los datos.

Hipótesis de Trabajo

La hipótesis central de este proyecto es que existe una relación entre la calidad del vino y ciertas propiedades fisicoquímicas. Específicamente, se investigará si las variables como la acidez fija, la acidez volátil, el contenido de azúcar residual, los cloruros y otros factores influyen en la calidad del vino. Además, se analizará cómo se distribuyen estas propiedades fisicoquímicas en función de la calidad del vino.

- ¿Qué propiedades fisicoquímicas están más fuertemente correlacionadas con la calidad del vino?
- ¿Existe una combinación específica de variables que prediga mejor la calidad del vino?
- ¿Cómo se distribuyen las diferentes calidades del vino en función de sus propiedades fisicoquímicas?

Objetivo

El objetivo principal de este proyecto es comprender en profundidad los factores que influyen en la calidad del vino mediante el análisis de propiedades fisicoquímicas. Además, se busca identificar patrones significativos en estos datos con el fin de desarrollar un modelo predictivo que pueda estimar la calidad del vino en función de estas características.

Considerando las características del dataframe, así como la naturaleza y distribución de sus variables, se ha determinado que la aplicación de un modelo supervisado de clasificación es la metodología de Machine Learning más adecuada.

La variable objetivo en este análisis es la calidad del vino, la cual se clasifica en tres categorías: malo, regular y bueno, basándose en una escala del 1 al 10. Este enfoque permitirá predecir la categoría de calidad de los vinos a partir de los patrones identificados en los datos de entrenamiento, optimizando así la precisión y la eficiencia del análisis predictivo.

Dataset

La base de datos utilizada en el proyecto fue descargada desde **Kaggle**.

El dataset seleccionado posee 12 columnas y 1599 filas y se compone de las siguientes variables (traducidas al español, ya que el idioma original es el inglés) :

Variables de entrada (basadas en pruebas fisicoquímicas):

1 - Acidez fija

2 - Acidez volátil

3 - Ácido cítrico

4 - Azúcar residual

5 - Cloruros

6 - Dióxido de azufre libre

7 - Dióxido de azufre total

8 - Densidad

9 - PH

10 - Sulfatos

11 - Alcohol

Variable de salida (basada en datos sensoriales):

12 - Calidad (puntuación entre 0 y 10)

Datos

A partir del método .info(), se identificó que no existen valores nulos en los datos, por lo que no fue necesario realizar tareas de limpieza ni de relleno. Además, se observó que todas las columnas, excepto la de "quality" que contiene valores de tipo entero, presentan datos de tipo flotante.

```
[ ] # Detección de valores nulos
    df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64  
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
```

Variables

En el conjunto de datos seleccionado, todas las variables son de tipo cuantitativo. Se considera que la variable de respuesta es 'quality' (calidad).

Estadística Descriptiva

En el marco del análisis realizado, se tomó la decisión de crear un diccionario denominado statistics con el propósito de almacenar las medidas descriptivas calculadas para cada variable del conjunto de datos. La implementación de esta estructura de datos responde a la necesidad de organizar y gestionar de manera eficiente la información estadística obtenida, facilitando su posterior consulta y análisis.

El diccionario statistics presenta una estructura clave-valor, donde cada clave corresponde al nombre de la variable y el valor asociado es un subdiccionario que alberga las medidas de tendencia central calculadas para dicha variable. Específicamente, este subdiccionario contiene las siguientes entradas:

- Media: Representa el valor promedio de todos los datos en la variable correspondiente.

- Mediana: Define el valor que divide al conjunto de datos en dos mitades de igual tamaño, de manera que la mitad de los valores son inferiores a la mediana y la otra mitad superiores.
- Moda: Se refiere al valor que presenta mayor frecuencia de aparición dentro del conjunto de datos.

```python
# Creación de un diccionario para almacenar las estadísticas
statistics = {}

# Calcular media, mediana y moda para cada variable
for column in df.columns:
    mean_value = np.mean(df[column])
    median_value = np.median(df[column])
    mode_value = df[column].mode()[0] if not df[column].mode().empty else np.nan  # Verificar si la moda no está vacía

    statistics[column] = {
        'Media': mean_value,
        'Mediana': median_value,
        'Moda': mode_value
    }

# Imprimir las estadísticas
for column, stats in statistics.items():
    print(f"{column}:")
    print(f"  Media: {stats['Media']}")
    print(f"  Mediana: {stats['Mediana']}")
    print(f"  Moda: {stats['Moda']}\n")
```

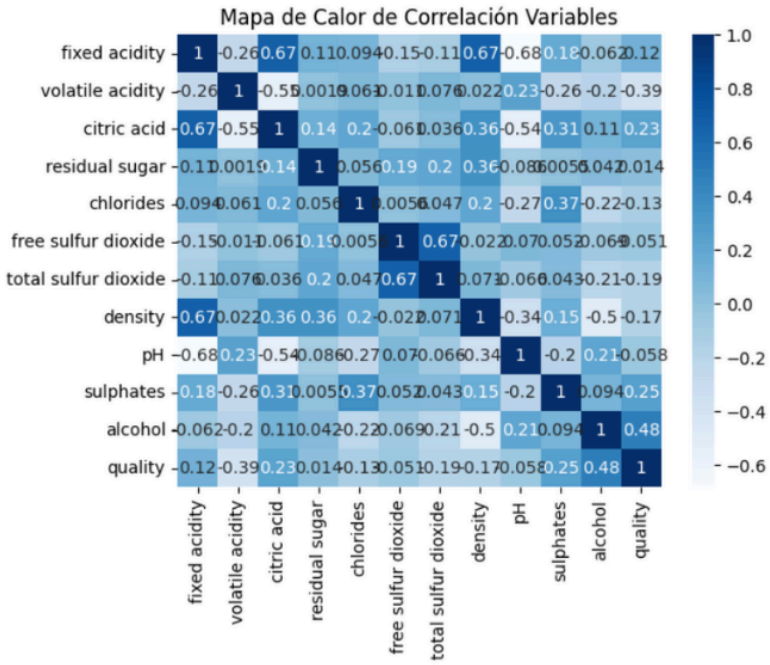
Visualizaciones

```python
#Importación de paquetes
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Seaborn

Mapa de Calor de Correlación Variables

```python
# Heatmap: Correlación Variables
correlation = df.corr()
sns.heatmap(correlation, annot=True, cmap="Blues")
plt.title("Mapa de Calor de Correlación Variables")
plt.show()
```

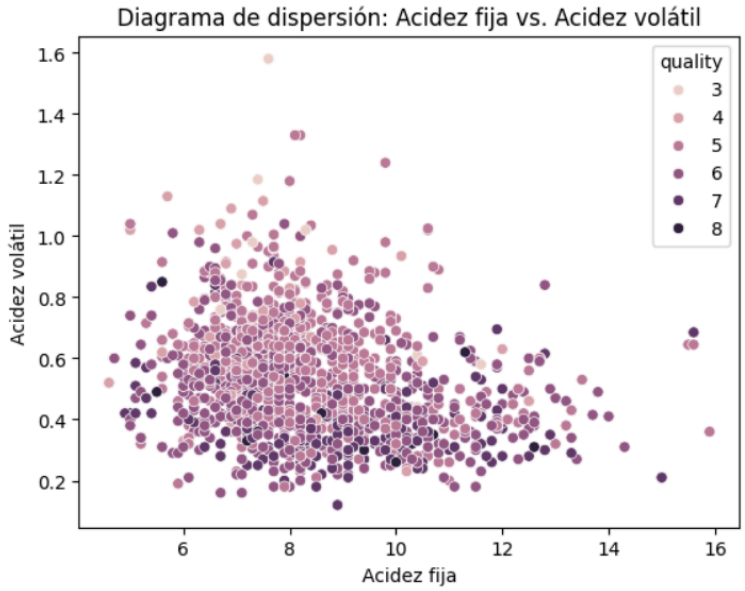


- **Acidez fija (fixed acidity):** Esta variable tiene una correlación positiva con la calidad del vino. A medida que aumenta la acidez fija, es más probable que la calidad del vino también aumente.
- **Acidez volátil** (volatile acidity): La acidez volátil muestra una correlación negativa con la calidad del vino, es decir, tienen una relación inversamente proporcional. Cuanto mayor sea la acidez volátil, es más probable que la calidad del vino disminuya. Esto puede deberse a que niveles elevados de acidez volátil pueden afectar negativamente el sabor y la estabilidad del vino.
- **Ácido cítrico** (citric acid): El ácido cítrico tiene una correlación positiva con la calidad del vino. Un mayor contenido de ácido cítrico tiende a estar asociado con una mejor calidad.
- **Azúcar residual** (residual sugar): No se identifica una correlación fuerte entre el azúcar residual y la calidad del vino.
- **Cloruros** (chlorides): Los cloruros tienen una correlación negativa con la calidad del vino. Cuanto mayor sea la concentración de cloruros, es más probable que la calidad disminuya.
- **Dióxido de azufre libre** (free sulfur dioxide): No muestra una correlación significativa con la calidad del vino.
- **Dióxido de azufre total** (total sulfur dioxide): No hay una correlación fuerte con la calidad.
- **Densidad (density)**: La densidad tiene una correlación negativa con la calidad. Valores más bajos de densidad suelen estar asociados con vinos de mejor calidad.
- **PH:** El PH tampoco muestra una correlación significativa con la calidad del vino.
- **Sulfatos** (sulphates): Los sulfatos tienen una correlación positiva moderada con la calidad. Mayor contenido de sulfatos se relaciona con vinos de mejor calidad.

En síntesis, las variables que se correlacionan positivamente con la calidad del vino son la acidez fija, el ácido cítrico y los sulfatos. Mientras tanto, la acidez volátil, los cloruros y la densidad tienen una relación inversa con la calidad del vino, lo que implica que a medida que aumentan, la calidad tiende a disminuir. En cuanto al azúcar residual, el dióxido de azufre libre, el dióxido de azufre total y el pH, no muestran una correlación significativa con la calidad.

Diagrama de dispersión: Acidez fija vs. Acidez volátil

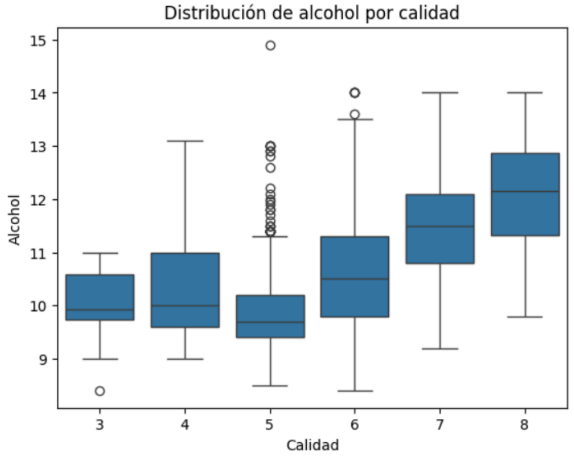
```
# Diagrama de dispersión: Acidez fija vs. Acidez volátil
sns.scatterplot(data=df, x='fixed acidity', y='volatile acidity', hue='quality')
plt.xlabel('Acidez fija')
plt.ylabel('Acidez volátil')
plt.title('Diagrama de dispersión: Acidez fija vs. Acidez volátil')
plt.show()
```



El gráfico de dispersión muestra que la acidez fija y la acidez volátil están débilmente correlacionadas positivamente. Los vinos de mayor calidad tienden a tener una acidez fija y volátil más bajas.

Boxplot: Distribución de contenido de alcohol por calidad del vino

```
# Boxplot: Distribución de contenido de alcohol por calidad del vino
sns.boxplot(x='quality', y='alcohol', data=df)
plt.xlabel('Calidad')
plt.ylabel('Alcohol')
plt.title('Distribución de alcohol por calidad')
plt.show()
```



La mediana del alcohol por calidad se encuentra en el centro de la caja, lo que indica que la distribución de los datos es simétrica. El IQR, representado por la altura de la caja, es relativamente pequeño, lo que sugiere que los datos están agrupados alrededor de la mediana.

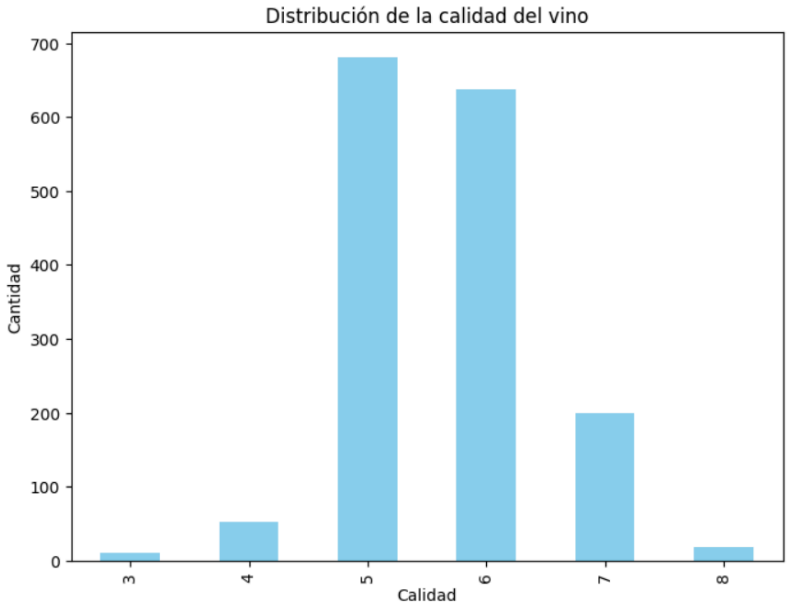
La distribución de los datos parece ser simétrica, ya que la mediana se encuentra en el centro de la caja y los bigotes tienen una longitud similar a ambos lados.

El gráfico de distribución de alcohol por calidad muestra que los datos están distribuidos de manera simétrica y agrupados alrededor de la mediana. No se observan valores atípicos en el conjunto de datos.

Matplotlib

Gráfico de barras: Distribución de la calidad del vino

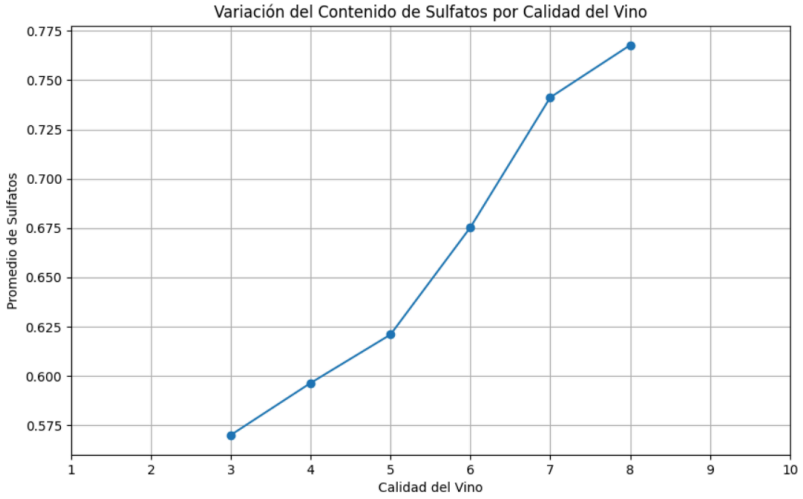
```
# Gráfico de barras: Distribución de la calidad del vino
plt.figure(figsize=(8, 6))
df['quality'].value_counts().sort_index().plot(kind='bar', color='skyblue')
plt.title('Distribución de la calidad del vino')
plt.xlabel('Calidad')
plt.ylabel('Cantidad')
plt.show()
```



El gráfico presenta la distribución de muestras de vino según diferentes puntuaciones de calidad. Se puede observar que la mayoría de las muestras se concentran en las calificaciones de 5 y 6. En contraste, las puntuaciones más bajas (4) y más altas (7 y 8) son menos frecuentes. Esto indica que la calidad del vino tiende a situarse alrededor de las puntuaciones medias.

Gráfico de Líneas: Contenido de Sulfatos por Calidad del Vino

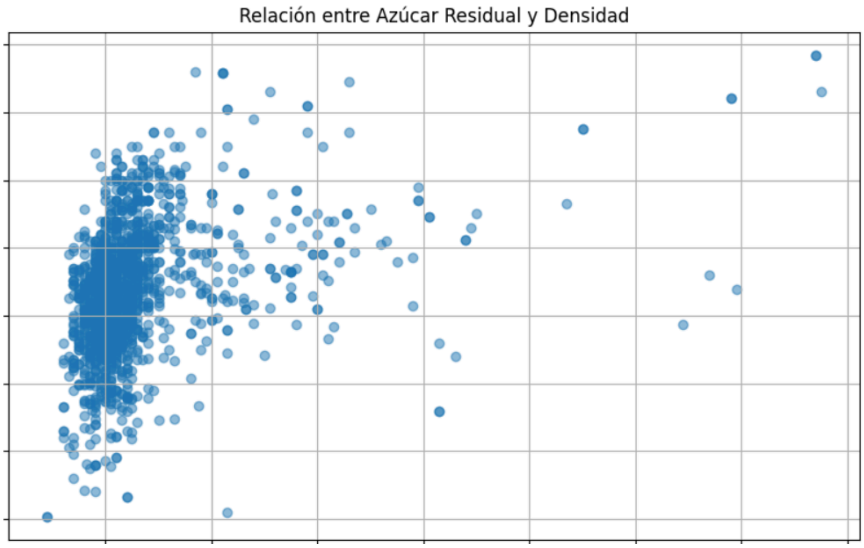
```
#Gráfico de Líneas: Contenido de Sulfatos por Calidad del Vino
plt.figure(figsize=(10, 6))
plt.plot(sulfatos.index, sulfatos.values, marker='o', linestyle='-')
plt.xlabel('Calidad del Vino')
plt.ylabel('Promedio de Sulfatos')
plt.title('Variación del Contenido de Sulfatos por Calidad del Vino')
plt.xticks(range(1, 11))
plt.grid(True)
plt.show()
```



Se identifica una marcada relación positiva entre las variables analizadas. A medida que aumenta la calidad del vino, también aumenta el porcentaje de sulfatos presente en él. Esto sugiere que los vinos de mayor calidad pueden requerir una cantidad específica de sulfatos para lograr su perfil sensorial deseado. Los sulfatos, como componentes químicos, pueden influir en el sabor y la calidad del vino.

Diagrama de Dispersión: Relación entre Azúcar Residual y Densidad

```
#Diagrama de Dispersión: Relación entre Azúcar Residual y Densidad
plt.figure(figsize=(10, 6))
plt.scatter(df['residual sugar'], df['density'], alpha=0.5)
plt.xlabel('Azúcar Residual')
plt.ylabel('Densidad')
plt.title('Relación entre Azúcar Residual y Densidad')
plt.grid(True)
plt.show()
```



El gráfico muestra una concentración densa de puntos alrededor de los valores más bajos de azúcar residual (cerca de 0 en el eje horizontal). La mayoría de estos puntos tienen densidades que oscilan entre aproximadamente 0.992 y 1.002. A medida que aumenta el azúcar residual , los puntos tienden a dispersarse más. Sin embargo, hay una ligera tendencia que sugiere que, a medida que aumenta el azúcar residual, la densidad tiende a disminuir. Esta distribución podría indicar una relación entre el contenido de azúcar residual y la densidad en los vinos, posiblemente influenciada por la cantidad de azúcar no fermentado presente