CODER HOUSE

Proyecto Final Data Science I:
Fundamentos para la Ciencia de Datos

"Predicción de la Calidad de Vino Tinto utilizando algoritmos de machine learning"

Docente: Jorge Ruiz

Tutor: Abril Noguera

Alumno: Carla Salamone

Año: 2024

Indice

Resumen

Este estudio tiene como objetivo predecir la calidad del vino utilizando modelos de aprendizaje automático, tales como Random Forest, Decision Tree y XGBoost. Se realizó un análisis de datos exploratorio (EDA) para comprender y preparar el conjunto de datos para el modelado. Se aplicaron técnicas como el manejo de valores atípicos y la normalización de datos. Para abordar el desequilibrio de clases, se empleó la técnica de sobremuestreo de minorías sintéticas (SMOTE). Los modelos se evaluaron en función de métricas que incluyen exactitud, precisión, recuperación y puntuación f1. Los resultados indicaron que XGBoost logró la mayor precisión con un 96 %, seguido de Random Forest con un 95 % y Decision Tree con un 92 %. XGBoost fue identificado como el modelo más eficaz debido a su alta precisión y rendimiento equilibrado. Random Forest también fue muy eficaz y robusto, y sirvió como una alternativa sólida. A pesar de la menor precisión, Decision Tree proporcionó una interpretabilidad valiosa. Este estudio proporciona un enfoque integral para predecir la calidad del vino, destacando las fortalezas y la efectividad de cada modelo evaluado.

Introducción

La calidad del vino es un aspecto crucial para su aceptación en el mercado y su valor comercial, es por ello que la Bodega "XXXX", situada en Valle de Uco, Mendoza; nos presentó su problemática para poder mejorar y predecir la calidad de sus vinos tintos.

Cada uno de los parámetros químicos del vino aporta información de relevancia. Los azúcares suelen medirse al finalizar la fermentación alcohólica para tener una prueba de que las levaduras han consumido la totalidad (o casi) de los azúcares naturalmente presentes en el mosto.

Las medidas de sulfuroso, en combinación con el pH del vino, sirven para conocer qué cantidad de sulfitos están protegiendo el vino durante su elaboración en bodega. Así, en determinados momentos críticos (especialmente cuando se va a embotellar y cuando se hacen trasiegos) es habitual tener que sulfitar para proteger el vino (siempre con dosis calculadas) de oxidaciones y contaminaciones con microorganismos como las bacterias acéticas. Estas bacterias darían lugar a un incremento de la acidez volátil, es decir, de olores a ácido acético (vinagre) que estropearían el vino.

Los parámetros de acidez total y grado alcohólico son especialmente relevantes en los vinos que se van a someter a crianza en barrica u otros recipientes, ya que ambos (acidez y alcohol) han de ser elevados para que el vino evolucione favorablemente en el tiempo. También durante la fermentación maloláctica de los tintos (principalmente), la disminución de la acidez total es uno de los parámetros que sirven para supervisar que ésta transcurre de forma adecuada.

La predicción de la calidad del vino mediante técnicas de machine learning puede proporcionar a los productores una herramienta valiosa para mejorar sus procesos de producción y asegurar un producto de alta calidad.

Objetivo

Este proyecto tiene como objetivo explorar y comparar tres modelos de machine learning para predecir la calidad del vino basado en sus características químicas, físicas y sensoriales.

Metodología

Descripción de los datos

Para este estudio se utilizó el conjunto de datos del vino tinto disponible en Kaggle. El dataset contiene 1599 muestras de vino tinto con 11 variables independientes que describen las características químicas y físicas tales como:

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

y una variable dependiente basada en datos sensoriales que representa la calidad del vino en una escala de 0 a 10.

Preprocesamiento de datos

En primer lugar, se hace la carga del dataset en estudio; el mismo puede observarse aquí: "Dataset del vino tinto". Allí se obtienen un grupo de datos compuesto por 1599 filas y 12 columnas.

En segundo lugar, se estudia la posible existencia de datos nulos y/o faltantes junto con la tipificación de las variables.

En este sentido, se detectan valores faltantes en la variable densidad, y una incorrecta tipificación de las variables alcohol y densidad, ya que las mismas se encuentran como "Object" y corresponden al tipo "Float".

A continuación, se detalla lo observado:

```
#   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   object
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   object
 11  quality               1599 non-null   int64
```

Figura 1: Conteo y tipo de cada una de las variables.

```
fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 1
pH                      0
sulphates               0
alcohol                 0
quality                 0
```

Figura 2: Verificación de datos faltantes en cada variable

Seguidamente se realiza la corrección de la tipificación correspondiente y la eliminación de los datos faltantes de la variable independiente densidad. En la figura 3 y 4 se muestran los resultados:

```
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
```

Figura 3: Verificación de la eliminación de datos faltantes de densidad

```
#   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1592 non-null   float64
 1   volatile acidity      1592 non-null   float64
 2   citric acid           1592 non-null   float64
 3   residual sugar        1592 non-null   float64
 4   chlorides             1592 non-null   float64
 5   free sulfur dioxide   1592 non-null   float64
 6   total sulfur dioxide  1592 non-null   float64
 7   density               1592 non-null   float64
 8   pH                    1592 non-null   float64
 9   sulphates             1592 non-null   float64
 10  alcohol               1592 non-null   float64
 11  quality               1592 non-null   int64
```

Figura 4: Verificación de la tipificación corregida de las variables correspondientes.

Luego de la limpieza de datos, el dataframe queda con 1592 filas y 12 columnas.

Análisis Exploratorio de Datos (EDA)

En esta parte del proyecto el objetivo es conocer la distribución de las variables individuales, identificar valores atípicos, analizar las relaciones entre las variables independientes y la variable dependiente o correlación entre las ellas.

Con esta información es posible preparar adecuadamente los datos para aplicar los diferentes modelos de machine learning más adelante.

Para ello se analiza:

1- El comportamiento del conteo total de valores en función de la variable calidad.

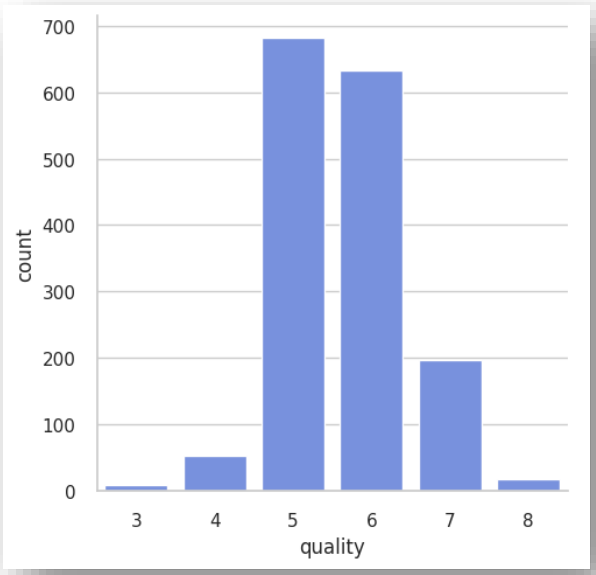

Figura 5: Conteo vs Calidad

En la figura 5 se puede observar que la distribución no es simétrica y los valores más frecuentes corresponden a la calidad 5 y 6.

2- Se realiza diagrama de cajas de cada una de las variables para observar posible presencia de valores atípicos.

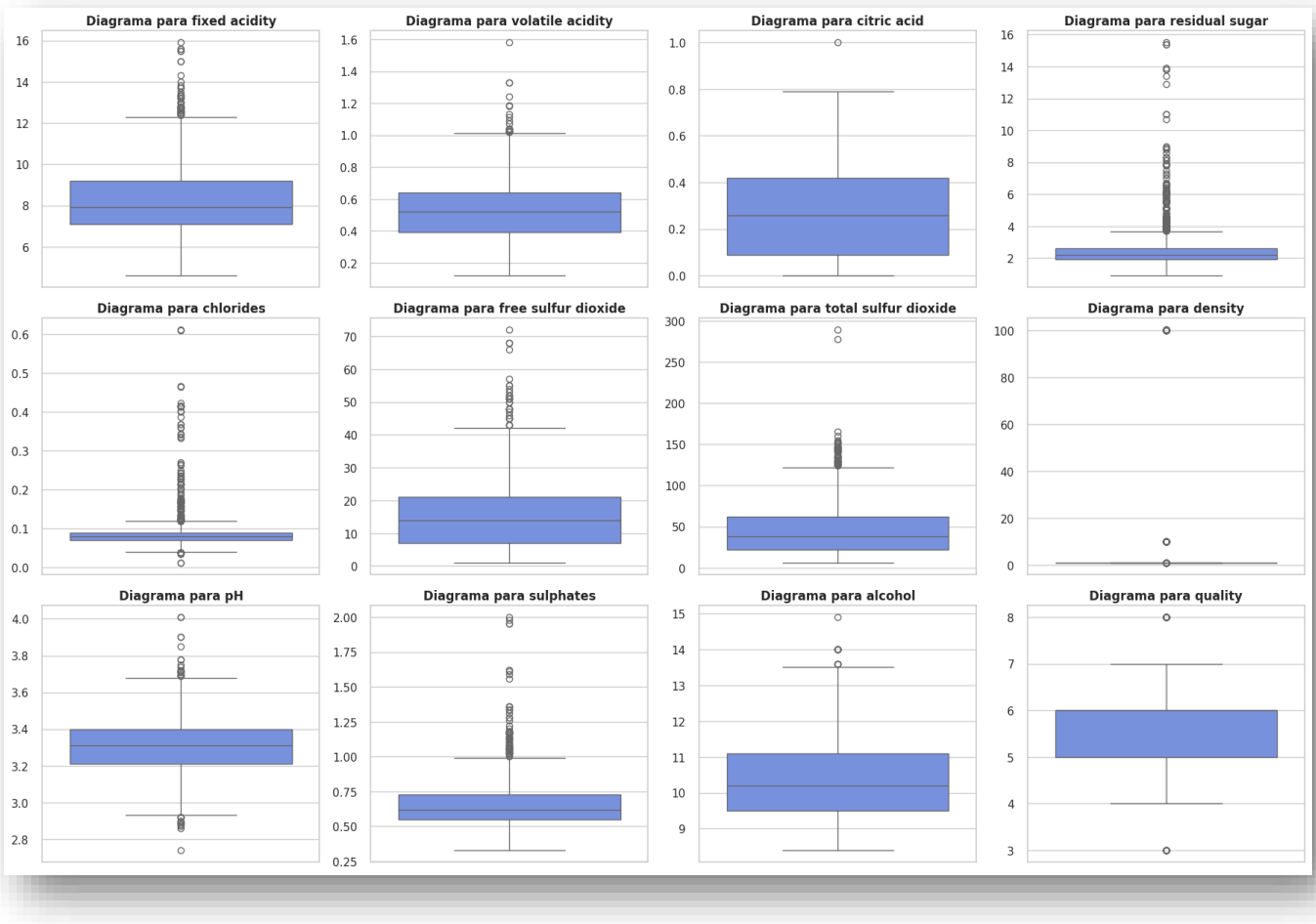

Figura 6: Diagrama de Cajas de todas las variables

De la figura 6, de pueden observar el diagrama de cajas de las diferentes variables. La mayoría presenta valores atípicos. Por ejemplo; la variable densidad tiene un IQR muy estrecho (entre 0.99 y 1) y valores atípicos muy altos (100), el dióxido de azufre total tiene un IQR entre 25 y 75 teniendo muchos valores atípicos por encima de 100.

De acuerdo a lo expuesto más arriba se decide hacer un tratamiento de los outliers utilizando el método intercuartil (IQR).

Una vez aplicada la eliminación de valores atípicos, el dataframe queda con 1152 filas y 12 columnas.

Con los nuevos datos, se realizan los histogramas de todas las variables ya que los mismos nos proporcionan una visión de la distribución de cada característica.

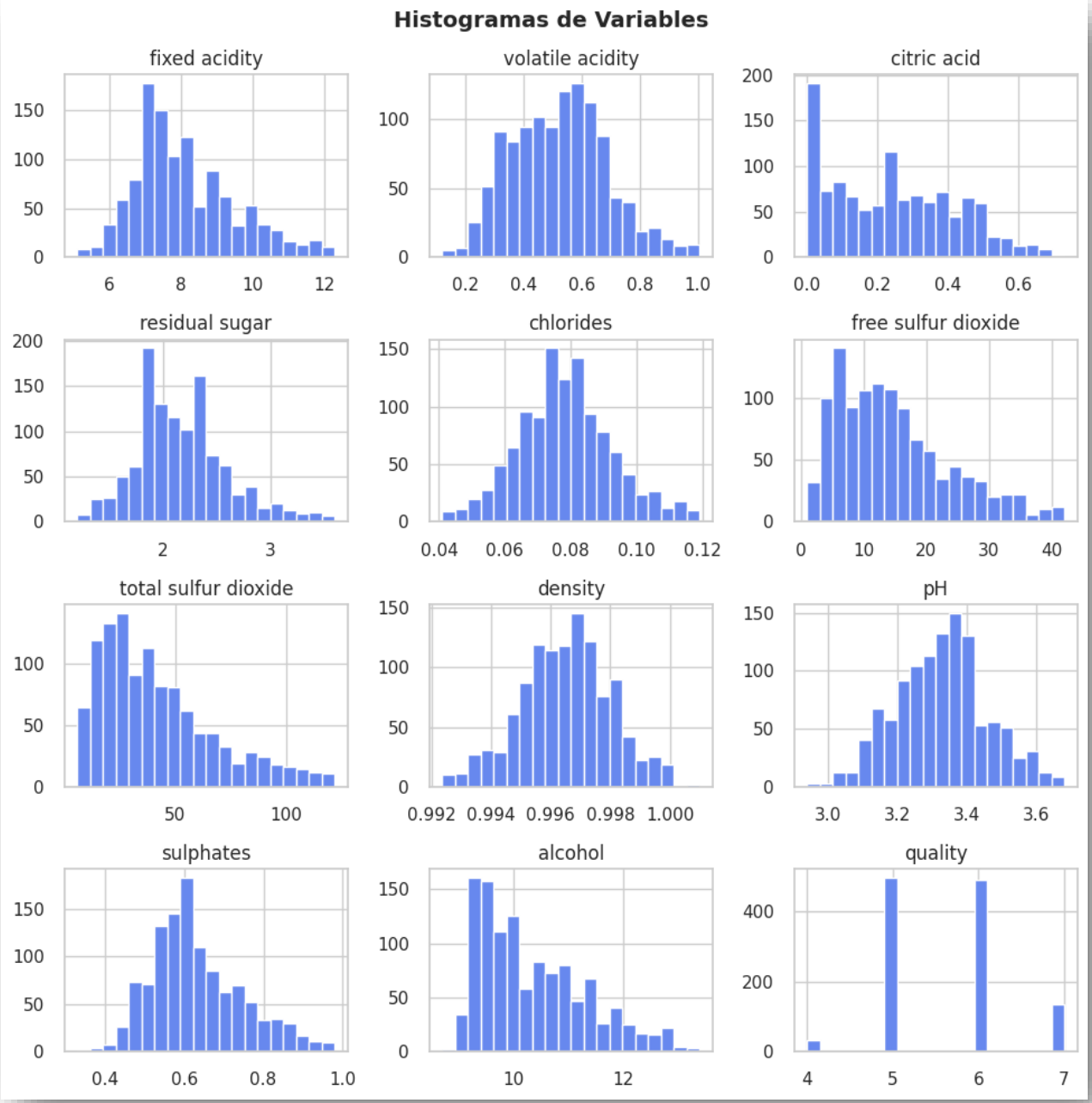

Figura 7: Histograma de todas las variables

En la figura 7 se puede observar una leve distribución asimétrica de acidez fija, acidez volátil, azúcar residual, dióxido de azufre libre, dióxido de azufre total, sulfatos, alcohol. El ácido cítrico muestra dos agrupaciones principales de datos. Las variables cloruros, densidad y pH se comportan aproximadamente con una distribución normal. La variable calidad muestra una distribución discreta.

3- Con los resultados obtenidos en el punto anterior se realiza mapa de calor que mostrará la correlación entre las variables. Este análisis nos permitirá identificar las relaciones fuertes o débiles entre variables.

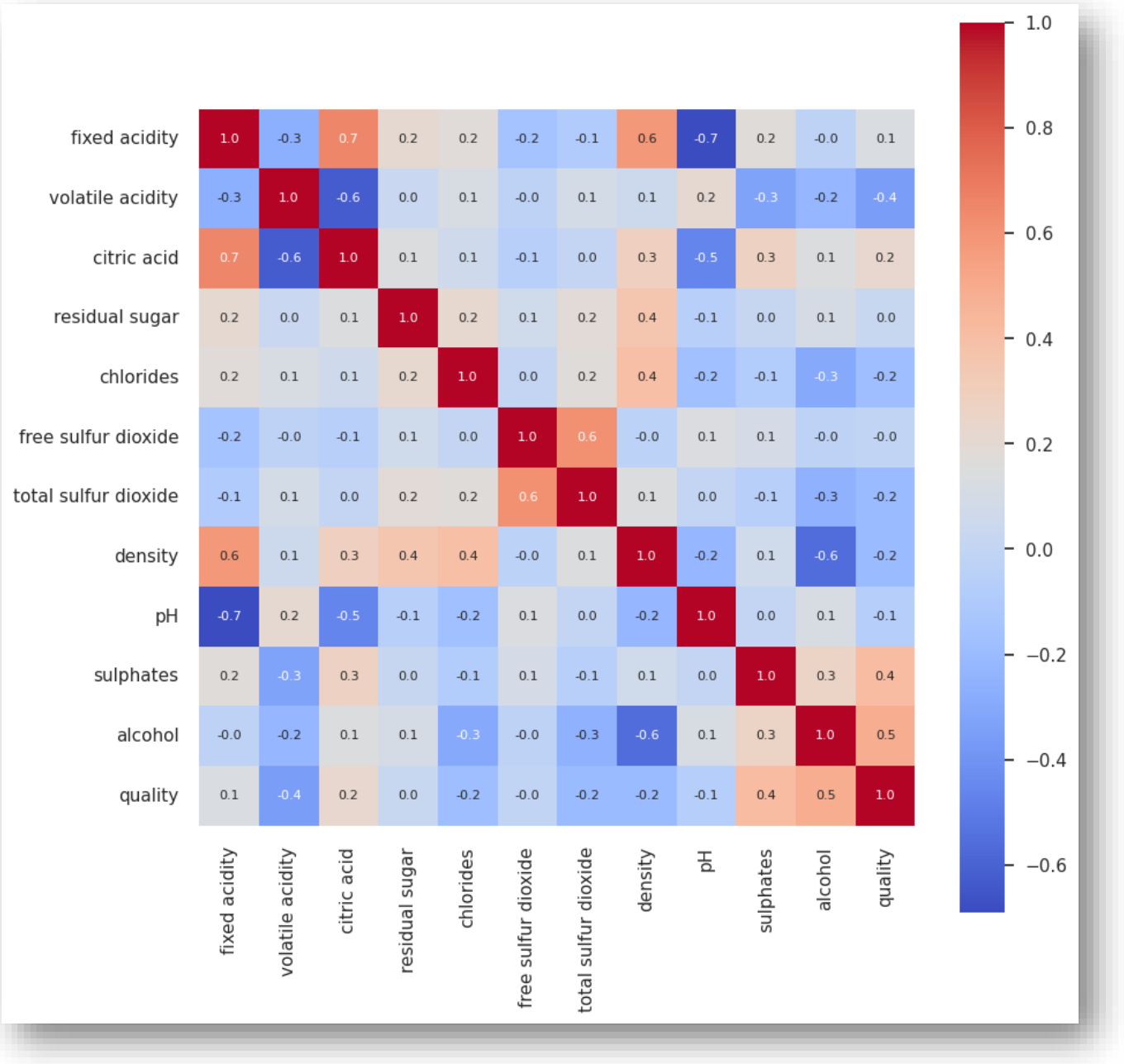

Figura 8: Mapa de Calor de todas las variables

En la figura 8 se muestra el mapa de calor de todas las variables, se puede apreciar que existe una correlación positiva entre la variable alcohol vs calidad, y sulfatos vs calidad. Hay una correlación negativa entre la variable acidez volátil vs calidad.

Modelos de machine learning

Antes de desafiar los modelos, se modificó a variable calidad que es discreta a dependiente y además se modificaron los valores de respuesta de manera binomial para poder dividir la calidad de vino en buena calidad y mala calidad.

Los valores obtenidos en el dataframe: 3, 4, 5 y 6 pasaron a 0 (Mala Calidad), valores ≥ a 7 (Buena Calidad).

Luego se utilizó la técnica SMOTE para balancear la distribución de clases, la misma, genera nuevas muestras sintéticas en lugar de duplicar las existentes, lo que ayuda a evitar el sobreajuste. SMOKE permite mejorar el rendimiento del modelo a utilizar.

Seguidamente se realizó la división de datos de entrenamiento y prueba para evaluar el rendimiento del modelo en datos no vistos, lo que proporciona una estimación más realista de su capacidad de generalización.

El 20% del conjunto de datos se asignó al conjunto de prueba, mientras que el 80% restante se utilizó para el entrenamiento del modelo.

-Análisis de los Modelos Seleccionados de Machine Learning:

1- Regresión Lineal:

De acuerdo al mapa de calor mostrado en la figura 8, se realizó el análisis del modelo de regresión lineal suponiendo que las variables alcohol y sulfatos tienen una alta correlación con la calidad del vino.

Los resultados obtenidos fueron los siguientes:

```
Mean Squared Error: 0.12197570802265173

R-squared: 0.5120854437034076
```

El modelo de regresión lineal obtenido expresa aproximadamente el 51.2% (R-Squared: 0.512) de la variabilidad en la calidad del vino. Este porcentaje es moderado a bajo, lo que sugiere que el modelo tiene una parte significativa de la variabilidad que no logra capturar.

Este valor de R relativamente bajo, requiere el uso de modelos no lineales tales como Random Forest, Diagrama de Árbol de Decisiones, y XGBoost Classifier**.**

2- Random Forest:

Se procedió al análisis del modelo Random Forest y arrojó los siguientes resultados:

```
Confusion Matrix:
 [[190  13]
 [  6 199]]

Classification Report: Random Forest Classifier
              precision    recall  f1-score   support

           0       0.97      0.94      0.95       203
           1       0.94      0.97      0.95       205

    accuracy                           0.95       408
   macro avg       0.95      0.95      0.95       408
weighted avg       0.95      0.95      0.95       408
```

Figura 9: Reporte Random Forest

El modelo de Random Forest muestra una precisión global del 95%, lo que sugiere que predice correctamente la mayoría de los casos.

Ambos f1-scores para las clases 0 y 1 son altos (0.95), indicando un buen balance entre precisión y recall.

Aunque hay algunos falsos positivos (13) y falsos negativos (6), el número es relativamente bajo en comparación con el total de predicciones, lo que sugiere que el modelo es bastante fiable.

3- Diagrama de Árbol de Decisiones:

El modelo de Decision Tree arrojó los siguientes resultados:

```
Confusion Matrix:
[[188  15]
[ 16 189]]

Classification Report: Decision Tree Classifier
              precision    recall  f1-score   support

           0       0.92      0.93      0.92       203
           1       0.93      0.92      0.92       205

    accuracy                           0.92       408
   macro avg       0.92      0.92      0.92       408

weighted avg       0.92      0.92      0.92       408
```

Figura 10: Reporte de Árbol de Decisiones

El modelo de Decision Tree muestra una precisión global del 92%, lo que sugiere que predice correctamente la mayoría de los casos.

Ambos f1-scores para las clases 0 y 1 son altos (0.92), indicando un buen balance entre precisión y recall.

Aunque hay algunos falsos positivos (15) y falsos negativos (16), el número es relativamente bajo en comparación con el total de predicciones, lo que sugiere que el modelo es bastante fiable.

4- XGBoost Classifier:

El modelo XGBoost Classifier arrojó los siguientes resultados:

```
XGBClassifier Model Evaluation:
Accuracy: 0.9583333333333334
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.94      0.96       203
           1       0.94      0.98      0.96       205

    accuracy                           0.96       408
   macro avg       0.96      0.96      0.96       408
weighted avg       0.96      0.96      0.96       408
```

Figura 11: Reporte de XGB Classifier

El modelo XGBClassifier muestra una precisión global del 96%, lo que sugiere que predice correctamente la gran mayoría de los casos.

Ambos f1-scores para las clases 0 y 1 son altos (0.96), indicando un excelente balance entre precisión y recall.

Aunque hay algunos falsos positivos y falsos negativos, el número es relativamente bajo en comparación con el total de predicciones, lo que sugiere que el modelo es extremadamente fiable.

Análisis Comparativo entre los modelos

1. Precisión Global (Accuracy)**:**

XGBoost mostró la mayor precisión (96%), seguido por Random Forest (95%) y Decision Tree (92%). Esto indica que XGBoost es ligeramente superior en términos de precisión general.

2. Balance entre Precisión y Recall**:**

Todos los modelos mostraron un buen balance entre precisión y recall. Sin embargo, XGBoost y Random Forest tuvieron f1-scores superiores (0.96 y 0.95, respectivamente) en comparación con Decision Tree (0.92). Esto sugiere que XGBoost y Random Forest son más efectivos en manejar tanto falsos positivos como falsos negativos.

Los modelos de Random Forest y XGBoost tienden a ser más complejos y robustos que los modelos de árbol de decisión simples. Esto se refleja en su mayor precisión y mejor manejo de datos con características más complejas.

XGBoost, debido a su naturaleza de boosting, generalmente es más robusto y menos propenso a sobreajuste comparado con los modelos de árbol de decisión simples. Random Forest también es menos propenso a sobreajuste debido a la combinación de múltiples árboles.

Conclusiones y Recomendaciones

Basado en la precisión y el balance de métricas, XGBoost se destaca como el modelo más eficaz para predecir la calidad del vino en este conjunto de datos. Su alta precisión y capacidad para manejar tanto falsos positivos como falsos negativos lo hacen ideal para esta tarea.

El modelo de Random Forest también mostró un rendimiento muy sólido y sería una excelente alternativa si la implementación de XGBoost no es factible por alguna razón. Su capacidad para manejar datos complejos y su robustez general lo hacen una opción viable.