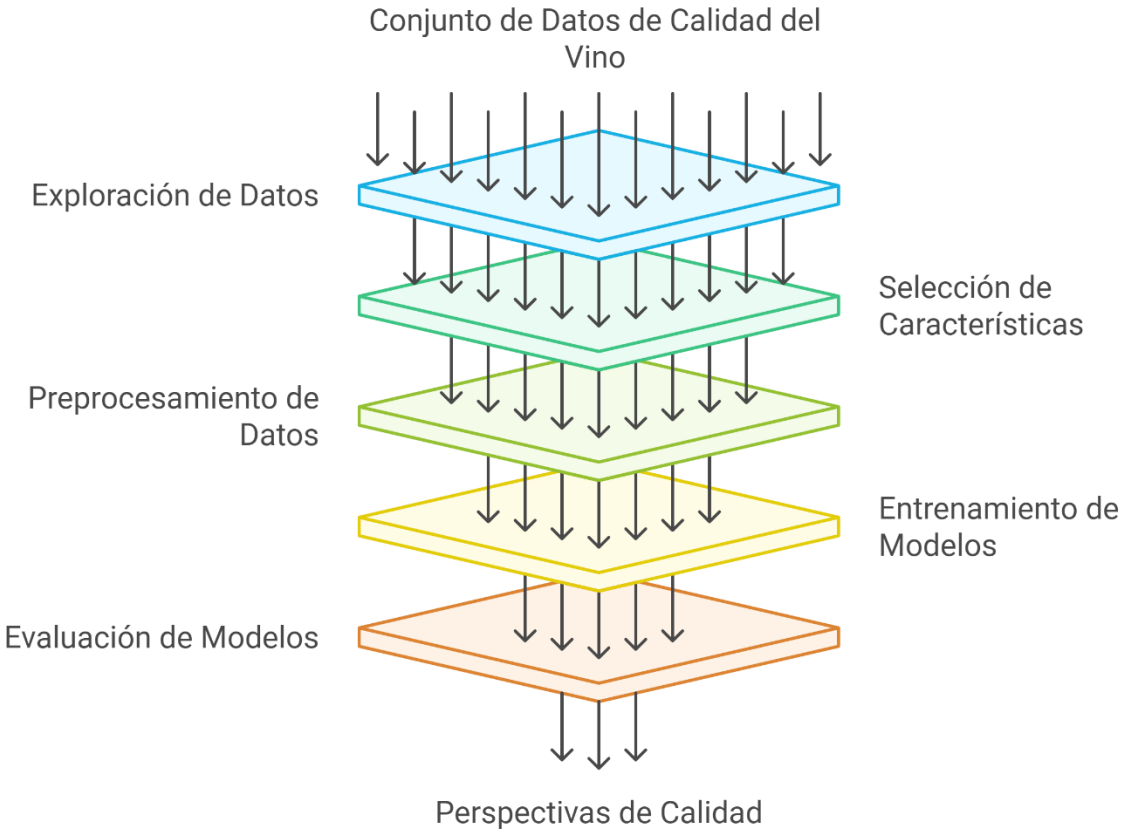
Informe de Análisis de Calidad del Vino

Este informe presenta un análisis detallado sobre la calidad del vino utilizando técnicas de clasificación basadas en características físico-químicas. Se ha utilizado el conjunto de datos de calidad del vino para explorar, preprocesar, entrenar modelos de clasificación y evaluar su rendimiento. A través de este ejercicio, se aplican conceptos fundamentales como la selección de características, el preprocesamiento de datos, el entrenamiento y la evaluación de modelos, así como el análisis de resultados mediante métricas y visualizaciones.



1. Carga y Exploración de Datos

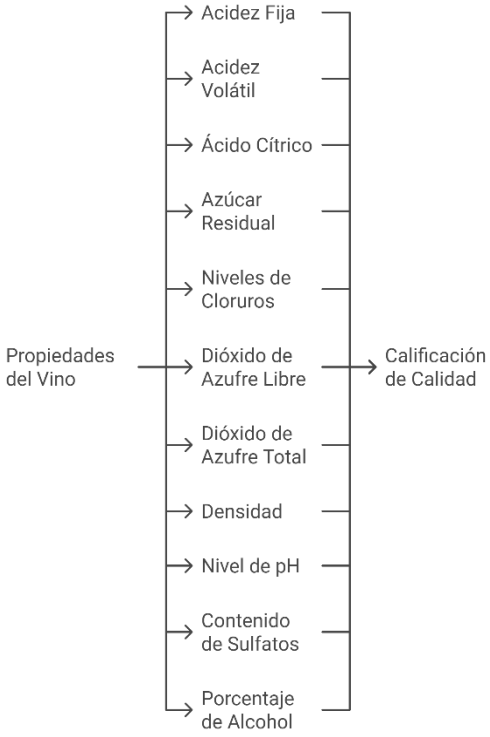
1.1 Carga del Dataset

Se cargó el conjunto de datos de calidad del vino, que contiene información sobre características físico-químicas y su calidad asociada. La estructura básica del dataset fue revisada, encontrando un total de 11 variables.

1.2 Descripción de Variables

Las variables del dataset incluyen:

- **Fixed Acidity**: Acidez fija del vino.
- **Volatile Acidity**: Acidez volátil.
- **Citric Acid**: Contenido de ácido cítrico.
- **Residual Sugar**: Azúcar residual.
- **Chlorides**: Niveles de cloruros.
- **Free Sulfur Dioxide**: Dióxido de azufre libre.
- **Total Sulfur Dioxide**: Dióxido de azufre total.
- **Density**: Densidad del vino.
- **pH**: Nivel de pH.
- **Sulfates**: Contenido de sulfatos.
- **Alcohol**: Porcentaje de alcohol.
- **Quality**: Calificación de calidad del vino (escala de 0 a 10).



1.3 Identificación y Tratamiento de Valores Nulos y Outliers

Se identificaron valores nulos y outliers en el dataset. Tras revisar los diagramas de caja, se decidió no imputar ningún dato, basándose en la naturaleza de los outliers y su posible impacto en el análisis.

2. Preprocesamiento de Datos

2.1 Selección de Características

Se seleccionaron las características más relevantes para la clasificación basándose en su correlación con la variable de calidad.

2.2 Transformación de Variables

Las variables categóricas fueron transformadas en variables numéricas cuando fue necesario.

2.3 División de Datos

Los datos fueron divididos en conjuntos de entrenamiento y prueba, asegurando una representación adecuada de las clases.

2.4 Escalado de Características

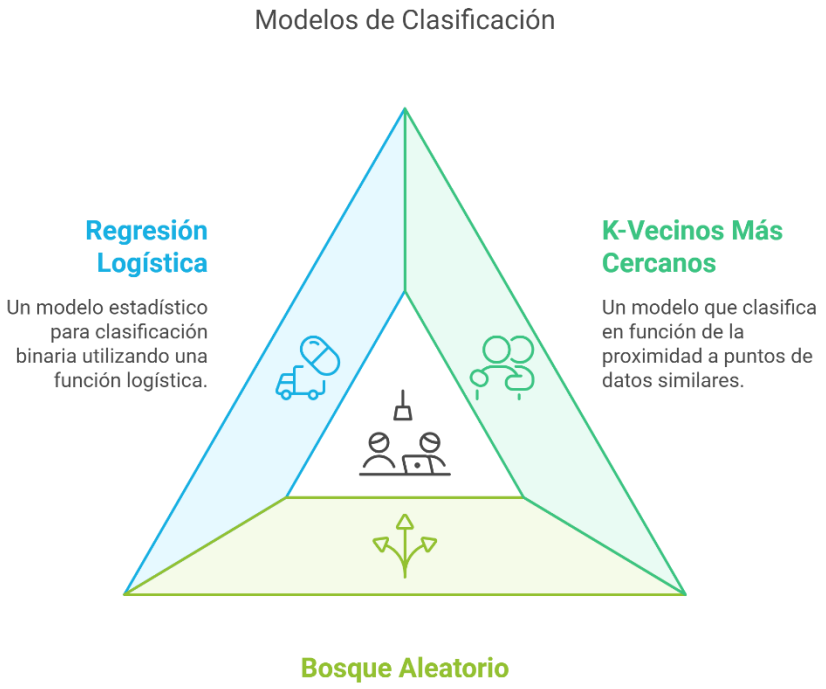
Se aplicó un escalado a las características para normalizar los datos y mejorar el rendimiento de los modelos de clasificación.

3. Entrenamiento de Modelos de Clasificación

3.1 Modelos Entrenados

Se entrenaron tres modelos de clasificación:

- **K-Nearest Neighbors (KNN)**
- **Random Forest**
- **Regresión Logística**



3.2 Validación Cruzada

Se utilizó validación cruzada para seleccionar los mejores hiperparámetros para cada modelo, optimizando así su rendimiento.

4. Evaluación de Modelos

4.1 Métricas de Evaluación

Los modelos fueron evaluados utilizando las siguientes métricas:

- **Exactitud**
- **Precisión**
- **Recall**

- **F1-Score**
- **Matriz de Confusión**

4.2 Informe de Clasificación

Se generó un informe de clasificación para cada modelo, proporcionando un análisis detallado de su rendimiento.

4.3 Curva ROC y AUC

Se creó y visualizó la curva ROC para el mejor modelo, calculando el AUC para evaluar su capacidad de discriminación.

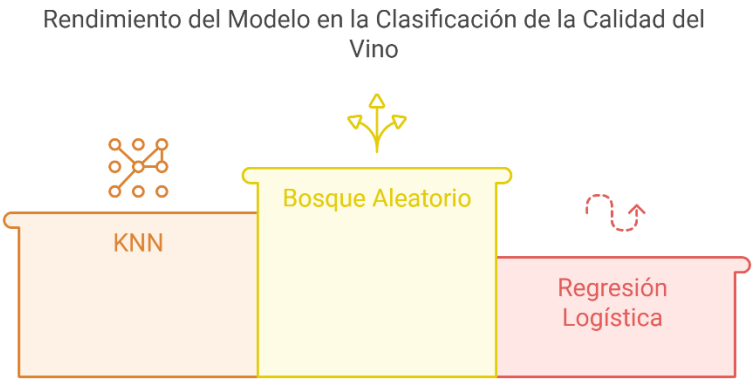
5. Análisis y Comparación de Resultados

5.1 Comparación de Modelos

Se comparó el rendimiento de los diferentes modelos, observando que el Random Forest mostró el mejor rendimiento en términos de exactitud y F1-Score.

5.2 Discusión de Resultados

El modelo Random Forest se destacó por su capacidad para manejar la complejidad de los datos y su robustez frente a overfitting. KNN, aunque efectivo, mostró limitaciones en conjuntos de datos más grandes, mientras que la regresión logística fue menos efectiva en la clasificación de vinos de calidad extrema.



5.3 Fortalezas y Debilidades

- **Random Forest**: Alta precisión y capacidad de manejo de datos no lineales.
- **KNN**: Simple y fácil de interpretar, pero sensible a la escala de los datos.
- **Regresión Logística**: Útil para problemas lineales, pero limitada en su capacidad para capturar relaciones complejas.

Conclusiones

El análisis realizado demuestra que las características físico-químicas del vino son efectivas para predecir su calidad. El modelo Random Forest se identificó como el más adecuado para esta tarea, ofreciendo un equilibrio entre precisión y robustez. Este ejercicio no solo refuerza la comprensión de las técnicas de clasificación, sino que también proporciona una base sólida para futuras investigaciones en la calidad del vino.