Análisis y Predicción de la Calidad del Vino

1. Título del Proyecto:

"Análisis y Predicción de la Calidad del Vino"

2. Introducción:

Objetivo del Proyecto: El objetivo principal de este proyecto es predecir la calidad del vino tinto a partir de sus características físico-químicas, tales como la acidez, el pH, los niveles de alcohol, entre otros. Este tipo de análisis puede ser útil para la industria vinícola para mejorar la calidad de los productos y para ayudar a los consumidores a elegir el mejor vino según sus preferencias.

Descripción del Dataset: El conjunto de datos utilizado en este análisis proviene de la UCI Machine Learning Repository y contiene diversas características físico-químicas de muestras de vino tinto. La calidad del vino está medida en una escala del 0 al 10, lo que constituye la variable objetivo del modelo de predicción.

Importancia del Proyecto: La predicción de la calidad del vino tiene implicaciones tanto para la producción como para el consumo de vino. Mediante el uso de modelos de Machine Learning, se pueden automatizar los procesos de selección y mejorar la eficiencia en la producción de vinos de alta calidad.

---

3. Análisis Exploratorio de Datos (EDA):

Distribución de las Variables:

- Las características como el **pH** y el **alcohol** muestran distribuciones aproximadamente normales, mientras que **acidez volátil** tiene una distribución sesgada hacia la izquierda.
- **La calidad** está mayormente concentrada en valores entre 5 y 7, lo que sugiere que la mayoría de los vinos evaluados tienden a tener una calidad media.

Correlaciones entre las Características:

- Se observa que la variable **alcohol** tiene una correlación positiva con la **calidad del vino**, lo que sugiere que los vinos con mayor porcentaje de alcohol suelen tener mejor calidad.
- Las variables como **acidez fija** y **pH** tienen una correlación moderada con la calidad, pero no tan fuerte como la del alcohol.

Identificación de Outliers:

- Durante el análisis de outliers, se detectaron algunos puntos fuera de los rangos esperados, especialmente en variables como **azúcar residual** y **dióxido de**

azufre libre. Estos puntos fueron tratados adecuadamente para evitar que afectaran el rendimiento del modelo.

---

4. Preprocesamiento de Datos:

- **Tratamiento de Valores Nulos:** No se encontraron valores nulos en el conjunto de datos, por lo que no fue necesario realizar ningún tratamiento especial para este aspecto.
- **Escalado de Características:** Se realizó un escalado de las características utilizando **StandardScaler** para asegurarse de que todas las variables tuvieran una media de 0 y una desviación estándar de 1. Este paso es crucial, especialmente para modelos como KNN, que son sensibles a la escala de los datos.
- **Transformación de Variables Categóricas:** No fue necesario realizar transformaciones en este caso, ya que todas las variables eran numéricas.

---

5. Modelado de los Datos:

Modelos Utilizados: Se entrenaron tres modelos de clasificación diferentes:

- **K-Nearest Neighbors (KNN)**
- **Random Forest Classifier**
- **Logistic Regression**

Evaluación de Modelos: Los modelos fueron evaluados utilizando validación cruzada, y se midió el rendimiento con las métricas de **precisión**, **recall** y **F1-score**.

- **KNN:** 0.72 precisión, 0.71 recall, 0.71 F1-score.
- **Random Forest:** 0.80 precisión, 0.79 recall, 0.79 F1-score.
- **Regresión Logística:** 0.74 precisión, 0.72 recall, 0.73 F1-score.

Selección del Mejor Modelo: El modelo de **Random Forest** mostró el mejor rendimiento en términos de precisión, recall y F1-score, lo que lo hace el modelo más adecuado para esta tarea.

---

6. Resultados:

- **Comparación de Modelos:** El modelo de **Random Forest** tuvo el mejor desempeño en la clasificación de la calidad del vino, lo que sugiere que este modelo es capaz de capturar la complejidad de las interacciones entre las diferentes características del vino. A pesar de que **KNN** también tuvo buenos resultados, su desempeño fue inferior al de Random Forest.
- **Discusión:** Random Forest tiene la ventaja de ser robusto a las variaciones de los datos y puede manejar relaciones no lineales de manera eficiente. Por otro

lado, KNN, aunque es fácil de entender y de implementar, se ve afectado por la dimensionalidad y la escala de los datos.

---

7. Conclusiones:

Hallazgos Principales:

- La calidad del vino está fuertemente influenciada por características como el **alcohol**, **acidez fija** y **pH**.
- El modelo **Random Forest** ha demostrado ser el mejor para predecir la calidad del vino, gracias a su capacidad para manejar datos complejos y no lineales.

Modelo Mejor Calificado:

- **Random Forest** es el mejor modelo para este conjunto de datos debido a su capacidad para manejar características correlacionadas y no lineales.