

# **INFORME DE ANÁLISIS CLASIFICACIÓN DE CALIDAD DEL VINO**

Aplicación de Machine Learning para Predicción de Calidad del Vino

**Alumna: Carolina Melero**

Fecha: 08 de July de 2025

# RESUMEN DEL ANÁLISIS

## OBJETIVO:

Desarrollar modelos de clasificación para predecir la calidad del vino usando características físico-químicas.

## DATASET:

- 1,143 muestras de vino tinto
- 11 características físico-químicas
- Variable objetivo: calidad del vino (escala 3-8)

## MODELOS IMPLEMENTADOS:

1. Random Forest Classifier
2. K-Nearest Neighbors (KNN)
3. Regresión Logística

## METODOLOGÍA:

- División de datos: 80% entrenamiento, 20% prueba
- Optimización de hiperparámetros con GridSearchCV
- Validación cruzada de 5 pliegues
- Métricas: Accuracy, Precision, Recall, F1-Score

## PREPROCESAMIENTO:

- Limpieza de nombres de columnas
- Verificación de valores nulos: 0 encontrados
- Detección de duplicados: 0 encontrados
- Escalado de características para KNN y Regresión Logística

# RESULTADOS OBTENIDOS

## MÉTRICAS DE EVALUACIÓN:

Modelo	Accuracy	Precision	Recall	F1-Score
Random Forest	0.657	0.649	0.657	0.648
KNN	0.584	0.576	0.584	0.573
Logistic Regression	0.540	0.532	0.540	0.528

## DISTRIBUCIÓN DE CALIDAD:

- Calidad 3: 6 muestras (0.5%)
- Calidad 4: 33 muestras (2.9%)
- Calidad 5: 483 muestras (42.3%)
- Calidad 6: 462 muestras (40.4%)
- Calidad 7: 143 muestras (12.5%)
- Calidad 8: 16 muestras (1.4%)

# HALLAZGOS PRINCIPALES

## 1. MEJOR MODELO:

- Random Forest obtuvo el mejor rendimiento
- Accuracy: 65.7%
- F1-Score: 64.8%

## 2. CARACTERÍSTICAS MÁS IMPORTANTES:

- Alcohol: Factor más influyente en la calidad
- Sulfatos: Importante para conservación
- Acidez volátil: Impacto negativo en exceso
- Dióxido de azufre: Balance crítico

## 3. DISTRIBUCIÓN DE DATOS:

- Clases desbalanceadas con predominio de calidades 5 y 6
- Pocas muestras en calidades extremas (3, 8)
- Total: 1143 muestras analizadas

## 4. RENDIMIENTO DE MODELOS:

- Random Forest: Mejor para problemas no lineales
- KNN: Rendimiento moderado, sensible a escala
- Regresión Logística: Limitada por asunciones lineales

## 5. DIFICULTADES IDENTIFICADAS:

- Distinguir entre calidades adyacentes (5-6, 6-7)
- Predicción de calidades extremas
- Desbalance de clases afecta rendimiento

## 6. CORRELACIONES CLAVE:

- Alcohol: correlación positiva con calidad
- Acidez volátil: correlación negativa
- Sulfatos: correlación positiva moderada
- Densidad: correlación negativa

# CONCLUSIONES Y RECOMENDACIONES

## CONCLUSIONES PRINCIPALES:

### 1. EFECTIVIDAD DE LOS MODELOS:

- Los modelos de Machine Learning pueden predecir la calidad del vino con precisión moderada
- Random Forest demuestra ser el más efectivo para este problema
- La precisión de 65.7% es útil pero mejorable

### 2. FACTORES DETERMINANTES:

- El contenido de alcohol es el predictor más fuerte de calidad
- La química del vino es compleja y multifactorial
- No existe un único factor que determine la calidad

### 3. LIMITACIONES DEL ANÁLISIS:

- Dataset limitado a vinos de una región específica
- Posible subjetividad en las evaluaciones de calidad
- Clases desbalanceadas afectan el rendimiento

### 4. NATURALEZA DEL PROBLEMA:

- La calidad del vino es inherentemente subjetiva
- Múltiples factores químicos interactúan
- Modelos no lineales son más apropiados

## RECOMENDACIONES:

### 1. MEJORAS TÉCNICAS:

- Implementar técnicas de balanceo de clases (SMOTE)
- Explorar feature engineering para crear nuevas variables
- Probar modelos ensemble más avanzados (XGBoost)
- Considerar agrupación de calidades similares

### 2. AMPLIACIÓN DEL ESTUDIO:

- Incluir vinos de diferentes regiones
- Ampliar el tamaño del dataset
- Incorporar factores de proceso de vinificación
- Validar con expertos en cata

### **3. APLICACIONES PRÁCTICAS:**

- Herramienta de apoyo para productores
- Control de calidad en procesos industriales
- Educación en enología
- Investigación en ciencia de alimentos

### **4. TRABAJO FUTURO:**

- Desarrollo de interfaces de usuario
- Integración con sistemas de producción
- Análisis temporal de calidad
- Comparación con evaluaciones humanas

### **REFLEXIÓN FINAL:**

Este análisis demuestra que el Machine Learning puede aportar valor significativo a la industria vitivinícola, proporcionando herramientas objetivas para evaluar la calidad del vino basándose en características medibles.