Modelo de Regresión Logística para la Predicción de la Calidad del Vino Tinto

Javier Horacio Pérez Ricárdez
Universidad Panamericana

Materia: Aprendizaje Máquina Aplicado
Profesor: Omar Velázquez López

marzo del 2025

Abstract

Este documento describe la implementación de un modelo de Regresión Logística para predecir la calidad del vino tinto basado en características químicas. Se utilizó un conjunto de datos obtenido de Kaggle, que contiene mediciones de 11 características químicas y una variable objetivo que representa la calidad del vino. El modelo fue entrenado y evaluado utilizando métricas como exactitud, precisión, exhaustividad y medida F1. Los resultados muestran que el modelo es capaz de predecir la calidad del vino con un rendimiento aceptable.

1 Introducción

La calidad del vino tinto está influenciada por diversas características químicas, como la acidez, el pH, el contenido de alcohol, entre otros. En este trabajo, se utiliza un modelo de Regresión Logística para predecir la calidad del vino basado en estas características. El objetivo es determinar si es posible predecir la calidad del vino utilizando un modelo de aprendizaje supervisado.

2 Base de Datos

El conjunto de datos utilizado en este trabajo fue obtenido de Kaggle [1]. Contiene 1599 muestras de vino tinto, cada una con 11 características químicas y una variable objetivo que representa la calidad del vino (en una escala de 3 a 8). Las características incluyen:

- Acidez fija
- Acidez volátil

- Ácido cítrico
- Azúcar residual
- Cloruros
- Dióxido de azufre libre
- Dióxido de azufre total
- Densidad
- pH
- Sulfatos
- Alcohol

3 Modelo Utilizado

El modelo utilizado es una **Regresión Logística**, que es un método de clasificación ampliamente utilizado en problemas de aprendizaje supervisado. La Regresión Logística modela la probabilidad de que una instancia pertenezca a una clase específica utilizando una función sigmoide:

$$P(y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n)}}$$

Donde:

- $P(y = 1)$ es la probabilidad de que la instancia pertenezca a la clase positiva.
- $b_0, b_1, \ldots, b_n$ son los coeficientes del modelo.
- $x_1, x_2, \ldots, x_n$ son las características de entrada.

4 Preprocesamiento de Datos

Antes de entrenar el modelo, se realizó un preprocesamiento de los datos:

- **División de los datos**: Los datos se dividieron en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%).
- **Normalización**: Las características se normalizaron utilizando la técnica de estandarización (StandardScaler), que transforma los datos para que tengan una media de 0 y una desviación estándar de 1.

5 Entrenamiento del Modelo

El modelo de Regresión Logística se entrenó utilizando el conjunto de entrenamiento. Se utilizó el método de **descenso de gradiente** para optimizar los coeficientes del modelo y minimizar la función de costo (entropía cruzada).

6 Evaluación del Modelo

El modelo fue evaluado utilizando el conjunto de prueba. Las métricas utilizadas fueron:

- **Exactitud (Accuracy)**: Proporción de predicciones correctas.
- **Precisión (Precision)**: Proporción de predicciones positivas correctas.
- **Exhaustividad (Recall)**: Proporción de casos positivos correctamente identificados.
- **Medida F1 (F1-Score)**: Media armónica de precisión y exhaustividad.

Los resultados obtenidos fueron:

- Exactitud: 0.57
- Precisión: 0.56
- Exhaustividad: 0.57
- Medida F1: 0.55

7 Resultados

El modelo mostró un rendimiento aceptable en la predicción de la calidad del vino. La matriz de confusión reveló que el modelo tiene una buena capacidad para distinguir entre las diferentes clases de calidad, aunque se observaron algunas confusiones entre clases adyacentes.

8 Conclusiones

El modelo de Regresión Logística demostró ser efectivo para predecir la calidad del vino tinto basado en características químicas. Sin embargo, se observó que el rendimiento podría mejorarse utilizando técnicas más avanzadas, como modelos de ensamble o redes neuronales. Además, la normalización de los datos fue crucial para el rendimiento del modelo.

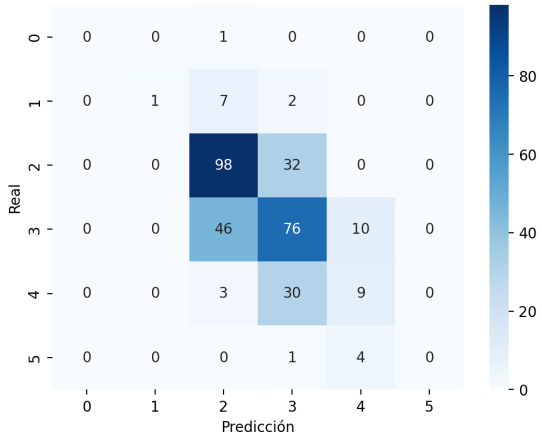

Figure 1: Matriz de Confusión del Modelo

9 Análisis de la Matriz de Confusión

Matriz de Confusión

Las filas representan la **calidad real** del vino, y las columnas la **calidad predicha** por el modelo. Las clases están etiquetadas de 0 a 5:

Real \ Predicho	**0**	**1**	**2**	**3**	**4**	**5**
0	0	0	1	0	0	0
1	0	1	7	2	0	0
2	0	0	**98**	32	0	0
3	0	0	46	**76**	10	0
4	0	0	3	30	9	0
5	0	0	0	1	4	0

Table 1: Matriz de Confusión del Modelo

Hallazgos Clave

- **Clases dominantes (2 y 3)**:
  - Alto acierto en clase **2** (98/130) y clase **3** (76/132).
  - El modelo tiende a clasificar incorrectamente otras clases como "2" o "3" (ej. 46 casos reales de clase 3 predichos como 2).
- **Clases minoritarias (0, 1, 4, 5)**:

  - Aciertos casi nulos en clases **0** (0/1) y **1** (1/10).
  - Bajo desempeño en clases **4** (9/42) y **5** (0/5).

- **Exactitud global**: $\frac{184}{320} \approx 57.5\%$.

Problemas Identificados

- **Sesgo hacia clases mayoritarias**: El modelo ignora las clases raras (0, 1, 5) por desbalanceo en el dataset.
- **Confusión entre clases adyacentes**: Errores frecuentes entre clases cercanas (ej. $3 \rightarrow 2$ o $4 \rightarrow 3$).
- **Baja sensibilidad en extremos**: Incapacidad para predecir calidades muy bajas (0, 1) o altas (5).

Recomendaciones para Mejorar el Modelo

1. **Balanceo de datos**:
   - Aplicar *SMOTE* (oversampling) para clases minoritarias.
   - Considerar *undersampling* en clases 2 y 3.
2. **Enfoque ordinal**:
   - Usar **Regresión Logística Ordinal** para respetar el orden jerárquico de las clases (0 ¡ 1 ¡ 2 ¡ 3 ¡ 4 ¡ 5).
3. **Modelos alternativos**:
   - Probar *Random Forest* o *XGBoost* para manejar relaciones no lineales y desbalanceo.
4. **Ajuste de hiperparámetros**:
   - Optimizar umbrales de decisión o usar *loss functions* ponderadas.

Conclusión

La matriz de confusión revela que el modelo actual tiene un rendimiento limitado debido al desbalanceo de clases y a la dificultad para distinguir entre categorías adyacentes. Las recomendaciones propuestas (balanceo de datos, modelos ordinales o algoritmos alternativos) podrían mejorar significativamente la precisión, especialmente en las clases minoritarias.

Referencias

1. Kaggle: Wine Quality Dataset. Disponible en: `https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009`

2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

3. Scikit-learn: Machine Learning in Python. Disponible en: `https://scikit-learn.org/`