Materia: Aprendizaje Automatico

Actividad: Clase 5

Informe de los modelos:

El dataset utilizado corresponde a la base de datos Wine Quality (Red) de UCI Machine Learning Repository. Contiene 1599 vinos tintos, cada uno descrito mediante 11 variables fisicoquímicas numéricas (por ejemplo: acidez fija, acidez volátil, ácido cítrico, azúcares residuales, cloruros, dióxido de azufre, densidad, pH, sulfatos y alcohol). La variable objetivo es la calidad del vino, originalmente en una escala de 0 a 10, que en este trabajo se reclasificó en tres categorías:

- Baja (0): calificaciones 3 y 4,
- Media (1): calificaciones 5 y 6,
- Alta (2): calificaciones 7 y 8.

La distribución de clases resultó muy desbalanceada, con 1319 vinos en la clase Media, 217 en Alta y solo 63 en Baja. Este desbalance condicionó fuertemente el desempeño de los modelos.

Con el fin de comparar el desempeño de diferentes técnicas de clasificación, se entrenaron dos modelos:

- K-Nearest Neighbors (KNN)
- Árbol de Decisión

En ambos casos se utilizó una división de los datos en 80% para entrenamiento y 20% para prueba, y se evaluaron las métricas de accuracy y F1-score, además de analizar en detalle la matriz de confusión y los reportes de clasificación.

En el caso del modelo KNN, se obtuvo un accuracy de 0.8094 y un F1-macro de 0.3952. Estos valores reflejan que, si bien el modelo logra una buena precisión global, su rendimiento no es homogéneo en todas las clases. En particular, la clase media fue la mejor representada, alcanzando un recall de 0.94 y un f1-score cercano a 0.90, lo que significa que identifica muy bien a los vinos de calidad intermedia.
Sin embargo, el desempeño en la clase alta fue mucho más bajo, con un f1-score de apenas 0.29, y directamente no logró predecir ningún caso de la clase baja (f1=0).
La matriz de confusión mostró que casi todos los vinos fueron clasificados como “media”, evidenciando las dificultades de KNN para manejar el desbalance de clases.

Estudiante: Mariana Lopez

Materia: Aprendizaje Automatico

Actividad: Clase 5

Por su parte, el Árbol de Decisión alcanzó un accuracy de 0.8469 y un F1-macro de 0.5084, superando al KNN en ambas métricas.
El árbol también clasificó con gran éxito los vinos de calidad media, con un recall de 0.91 y un f1-score de 0.90, pero además mejoró el rendimiento sobre la clase alta, alcanzando un f1-score de 0.61.

Al igual que KNN, no logró reconocer la clase baja, debido a la escasa cantidad de ejemplos disponibles.No obstante, las matrices de confusión muestran que el árbol sí consiguió diferenciar un mayor número de vinos de calidad alta, lo que explica su mejor promedio en el F1-macro.

A pesar de los buenos resultados del Árbol de Decisión, es importante señalar sus limitaciones. Este modelo toma decisiones a partir de cortes binarios en variables numéricas, evaluando si un valor es mayor o menor que un umbral, lo que lo hace menos flexible frente a datos continuos como los del vino. En cambio, aunque el KNN obtuvo métricas más bajas, su lógica de clasificación basada en la proximidad entre observaciones lo convierte en un modelo conceptualmente más adecuado para este tipo de datos, ya que aprovecha la continuidad y similitud entre las variables fisicoquímicas.

En conclusión, ambos modelos demostraron ser útiles para identificar los vinos de calidad media, pero presentaron grandes dificultades con la clase baja y un desempeño intermedio con la clase alta. El Árbol de Decisión logró mejores valores globales, con un F1-macro de 0.50 frente al 0.39 de KNN, pero su interpretación sobre variables numéricas es rígida. Por su parte, KNN, aunque más débil en las métricas, ofrece una mayor coherencia con la naturaleza de los datos.

Este análisis sugiere que, con un ajuste más cuidadoso de parámetros y un tratamiento del desbalance de clases, el KNN podría convertirse en una mejor alternativa para el problema estudiado.

Estudiante: Mariana Lopez