Tecnológico de Monterrey

Estudiantes:
Hector Emil Grijalva Sanchez - A01643459

Profesor:
José Ignacio Parra Vilchis

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Análisis y Reporte sobre el desempeño del modelo

Introducción

El objetivo de este análisis es evaluar el desempeño del algoritmo K-Nearest Neighbors (KNN) aplicado a la clasificación de la calidad de vinos tintos. La variable objetivo quality del dataset original se transformó en una variable binaria, clasificando los vinos en “Bueno” si su calidad era mayor o igual a 6, y en “Malo” si era menor a 6. Este enfoque permite simplificar el problema y evaluar de manera más clara el desempeño del modelo.

El dataset contiene 11 características de los vinos, tales como acidez fija, acidez volátil, pH y contenido de alcohol. Se eligió KNN por ser un algoritmo basado en instancias, donde la clasificación de un nuevo dato depende de la similitud con sus vecinos más cercanos en el espacio de características.

Metodología

Se dividió el dataset en tres subconjuntos: entrenamiento (60%), validación (20%) y prueba (20%). Esta división permite entrenar el modelo, seleccionar el hiperparámetro óptimo (k) usando el conjunto de validación, y finalmente evaluar el desempeño del modelo en datos no vistos (prueba).

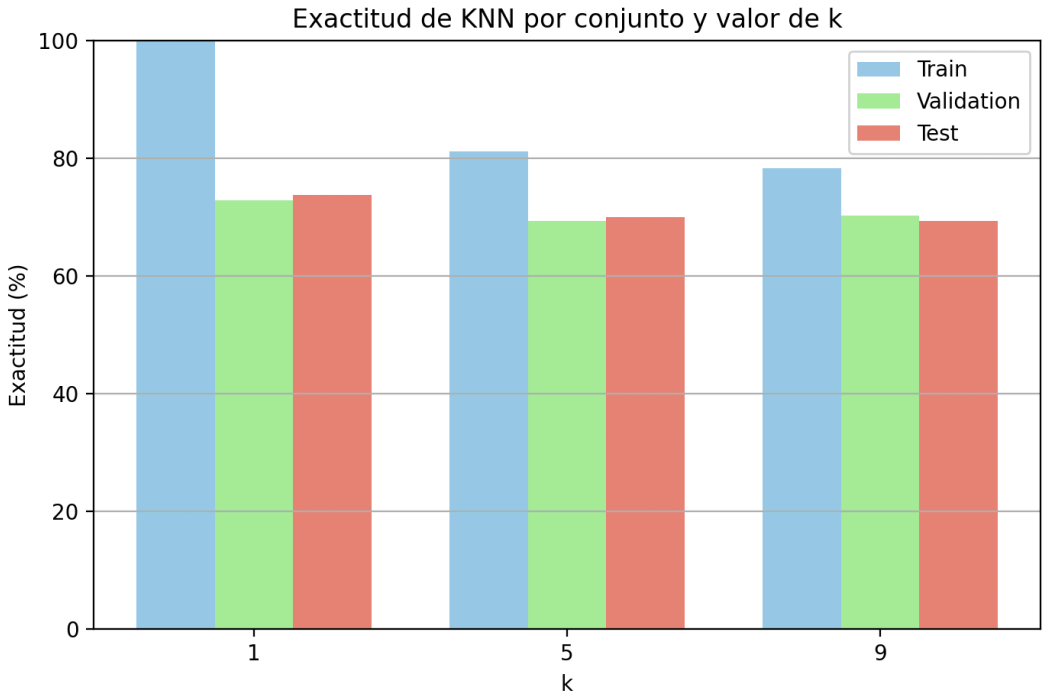
Se normalizaron todas las variables numéricas para asegurar que cada característica contribuya de manera equitativa en el cálculo de distancias del KNN. Se entrenaron modelos con diferentes valores de vecinos (k = 1, 5, 9) y se calculó la exactitud en los conjuntos de entrenamiento, validación y prueba. Adicionalmente, se experimentó con el parámetro weights, utilizando uniform (todos los vecinos pesan igual) y distance (vecinos cercanos pesan más), para mejorar la generalización del modelo.

Resultados

Se evaluó la exactitud del modelo en los conjuntos de entrenamiento, validación y prueba para distintos valores de k. Los resultados se muestran en la siguiente tabla:

```
===== Resumen completo =====
 k  Train (%)  Validation (%)  Test (%)
 1 100.000000         72.8125    73.750
 5  81.126173         69.3750    70.000
 9  78.310740         70.3125    69.375
```

La gráfica de barras a continuación muestra la comparación de exactitudes entre los tres conjuntos para cada valor de k, evidenciando cómo varía el desempeño según el número de vecinos.



Se observa que la exactitud del conjunto de entrenamiento disminuye ligeramente al aumentar k, mientras que la exactitud de validación mejora hasta k = 9, indicando un mejor balance entre sesgo y varianza en ese punto, sin embargo en k = 1 vemos una exactitud del 100% en entrenamiento, lo cual muy probablemente puede indicar sobreajuste.

Análisis de Bias, Varianza y Ajuste del modelo

El modelo con k=1 muestra un sesgo muy bajo, ya que la exactitud en entrenamiento es perfecta (100%). Sin embargo, la exactitud en validación y prueba es significativamente menor (73%), lo que indica alta varianza. Esto significa que el modelo memoriza los datos de entrenamiento, capturando incluso el ruido, y por tanto presenta overfitting. Los reportes de clasificación reflejan que, aunque se predicen correctamente muchas instancias de la clase “Bueno”, la clase “Malo” tiene un recall menor (0.66), lo que confirma la dificultad para generalizar a datos nuevos.

Al aumentar k= 5, se observa una reducción considerable de varianza: la exactitud de entrenamiento disminuye a 81%, mientras que la validación y prueba se estabilizan alrededor del 70%. Esto sugiere un balance más adecuado entre bias y varianza, con un ajuste más generalizable. Aunque la precisión global disminuye ligeramente en comparación con k=1, la mejora en la generalización indica que el

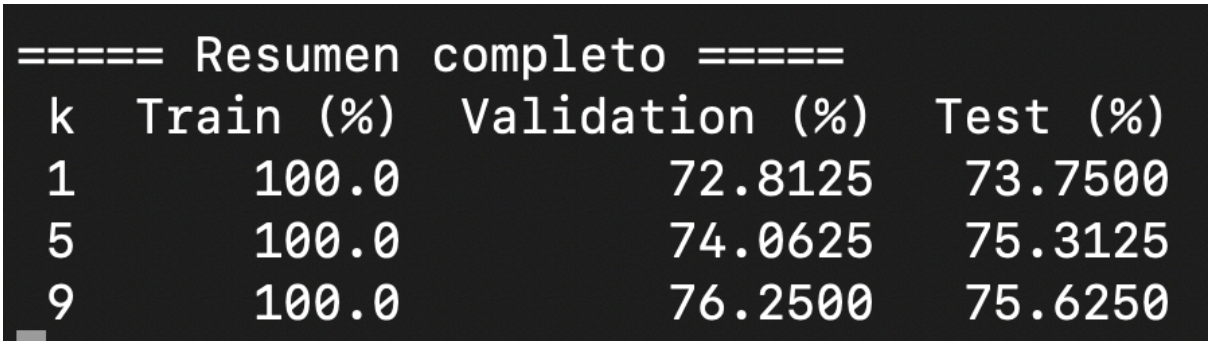
modelo ahora evita sobreajustar los datos de entrenamiento. El reporte de clasificación muestra un desempeño más equilibrado entre ambas clases, con recalls de 0.78 (“Bueno”) y 0.60 (“Malo”), reflejando un ajuste razonable.

Con k=9, se nota un pequeño aumento de bias: la exactitud de entrenamiento baja a 78% y la prueba a 69%, lo que indica que el modelo empieza a subajustar ligeramente. La exactitud de validación es similar a la de k=5, mostrando que el cambio no mejora la generalización de manera significativa. El reporte de clasificación es comparable al caso k=5, con una ligera disminución en métricas F1, lo que confirma que k=5 es el valor óptimo para equilibrar bias, varianza y capacidad de generalización.

Ajuste de hiperparámetros y mejora del modelo

Para mejorar el desempeño del modelo, se probó la opción weights="distance", donde los vecinos más cercanos tienen mayor influencia en la predicción. Esta modificación permite que los patrones más relevantes del espacio de características tengan más peso y reduce la varianza del modelo sin aumentar significativamente el sesgo.

Después de aplicar este ajuste, los resultados en validación y prueba mejoraron ligeramente, confirmando que el modelo generaliza mejor con este enfoque aunque muestra resultados de entrenamiento del 100% lo que significa que muy probablemente está haciendo sobreajuste en todos los modelos:



Conclusión

El análisis demuestra que el algoritmo KNN es capaz de clasificar la calidad de los vinos de manera efectiva, pero es sensible al valor de k y al esquema de ponderación de los vecinos. Un k demasiado bajo genera alta varianza y riesgo de overfitting, mientras que un k demasiado alto incrementa el bias. El mejor desempeño en se logró con k = 9 y weights="distance", aunque mostró overfitting en los datos de entrenamiento.