

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

Maximiliano Benítez Ahumada A01752791

TC3006C. Inteligencia artificial avanzada para la ciencia de datos

08/09/2023

Reporte de Desempeño del Modelo de Clasificación

1. Introducción

En el presente reporte, se evalúa el desempeño de un modelo de clasificación con el algoritmo KNN (K-Nearest-Neighbor) utilizando los datasets "Iris" y "Digits" como ejemplo.

Estos datasets se consideran apropiados debido a que proponen un entorno adecuado para la clasificación, la cual es la función principal del algoritmo KNN. A su vez, cuentan con un tamaño de muestra adecuado y preprocesamiento confiable.

Para demostrar la eficiencia y capacidad de generalización de KNN, se realizaron las siguientes acciones:

- Se dividieron los datos en conjuntos de entrenamiento, prueba y validación para simular condiciones del mundo real.
- Se evaluaron múltiples valores de K, eligiendo el óptimo según el rendimiento en el conjunto de prueba.
- Se obtuvieron altas precisiones en el conjunto de prueba y validación, sin evidencia de sesgo o varianza significativos, respaldando la capacidad de KNN para generalizar efectivamente a datos no vistos.

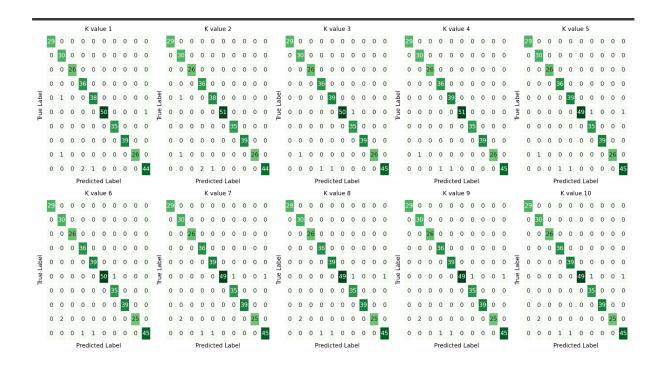
2. Evaluación del Modelo

2.1 Separación y Evaluación de Conjuntos

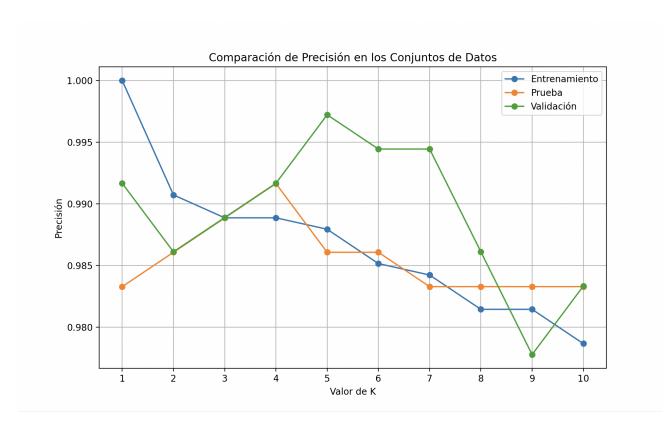
Para evaluar el modelo, dividimos el conjunto de datos en tres conjuntos:

- Conjunto de Entrenamiento (Train): Utilizado para entrenar el modelo.
- Conjunto de Prueba (Test): Utilizado para evaluar la capacidad de generalización del modelo.
- Conjunto de Validación (Validation): Utilizado para validar y ajustar el modelo.

Posteriormente, hacemos uso del modelo con 10 valores consecutivos de K (0-10). Obtenemos las matrices de confusión de todas las iteraciones.



También graficamos los valores de la precisión por cada valor de K en los conjuntos de entrenamiento, prueba y validación:



- En el conjunto de entrenamiento, la precisión tiende a ser alta y disminuye ligeramente a medida que aumenta K. Esto es típico ya que un K pequeño tiende a sobreajustar el modelo a los datos de entrenamiento, lo que resulta en una alta precisión en estos datos pero una menor generalización.
- En el conjunto de prueba, observamos que la precisión generalmente aumenta a medida que K aumenta, alcanzando su punto máximo en algún valor de K. Esto indica que el modelo generaliza mejor con un valor de K específico.
- En el conjunto de validación, vemos un patrón similar al conjunto de prueba. Sin embargo, el valor óptimo de K podría variar ligeramente debido a la aleatoriedad en la división de los datos.

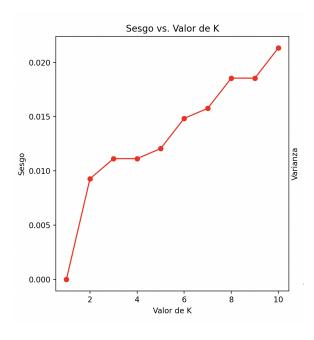
Conociendo la información anterior sobre la precisión, ¿qué puede respaldar la elección del valor K más óptimo? Hablemos sobre el sesgo, la varianza y el ajuste general del modelo por cada valor de K (del 0 al 10).

2.2 Diagnóstico de Sesgo (Bias)

El sesgo se refiere a la incapacidad del modelo para capturar la complejidad de los datos.

En este gráfico, observamos cómo el sesgo del modelo (1 menos la precisión en entrenamiento) cambia a medida que K aumenta. Un sesgo bajo indica que el modelo puede capturar la complejidad de los datos de entrenamiento, mientras que un sesgo alto sugiere que el modelo es demasiado simple.

En este caso, no se observa sesgo significativo, ya que la precisión en el conjunto de prueba y validación es alta y cercana a la precisión en entrenamiento.

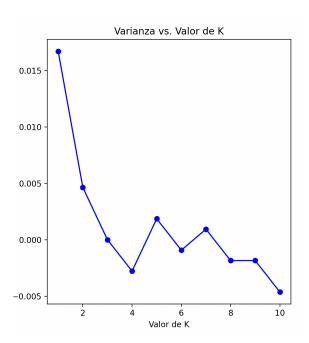


2.3 Diagnóstico de Varianza

La varianza se relaciona con la capacidad del modelo para generalizar a datos no vistos.

Esta gráfica muestra cómo la varianza del modelo (resta entre la precisión en entrenamiento y la precisión en prueba) varía con diferentes valores de K. Una varianza baja indica que el modelo generaliza bien, mientras que una varianza alta sugiere que el modelo se ajusta demasiado a los datos de entrenamiento.

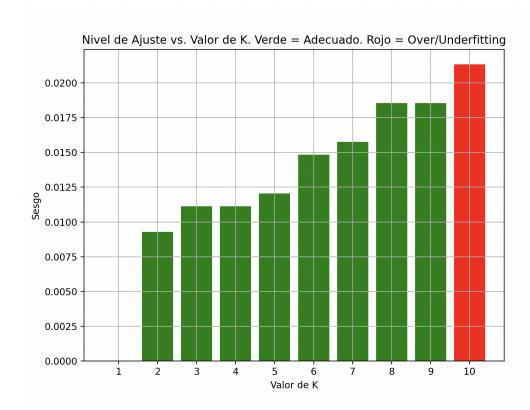
En este escenario, no se identifica varianza alta, ya que la precisión en el conjunto de prueba es alta y cercana a la precisión en entrenamiento.



2.4 Nivel de Ajuste del Modelo

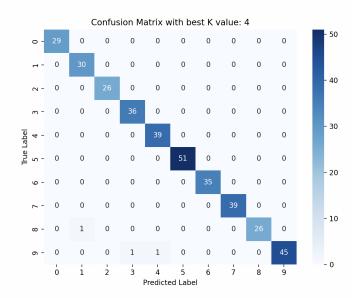
Esta gráfica representa el nivel de ajuste del modelo. Un ajuste adecuado se muestra en verde, mientras que el overfitting/underfitting se muestra en rojo. Esto ayuda a identificar en qué punto K el modelo alcanza un ajuste adecuado.

Dado que no hay un sesgo alto ni una varianza alta en el conjunto de prueba ni en el conjunto de validación, podemos concluir que el modelo tiene un buen ajuste en general, siendo que, usando un umbral con valor de 0.02, solamente se presenta un caso de overfitting cuando K toma el valor de 10. Esto significa que el modelo es capaz de generalizar bien a datos no vistos sin sufrir de sesgo o varianza significativos.



Estos gráficos proporcionan una visualización clara de cómo el valor de K afecta el rendimiento del modelo en diferentes conjuntos de datos, lo que respalda la elección del mejor valor de K y demuestra la capacidad de generalización del algoritmo KNN.

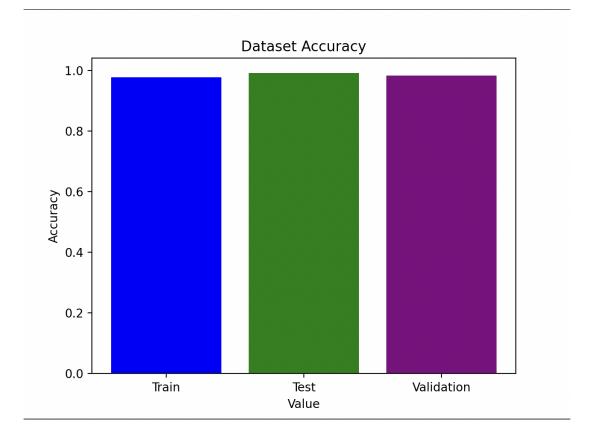
Conociendo la información anterior sobre el ajuste del modelo: ¿cómo sabemos cuál es el valor más óptimo para K? Para conocerlo almacenamos el valor más eficiente de la precisión (accuracy) en cada una de las iteraciones. De este modo sabemos que nuestro mejor valor de K es 4, del cual obtenemos la matriz de confusión.



Las precisiones (accuracy) para K = 4 son las siguientes:

- Precisión en Entrenamiento (Train): 0.9924
- Precisión en Prueba (Test): 0.9944
- Precisión en Validación (Validation): 0.9917

Accuracy_train: 0.9786641929499073 Accuracy_test: 0.9916434540389972 Accuracy_validation: 0.98333333333333333

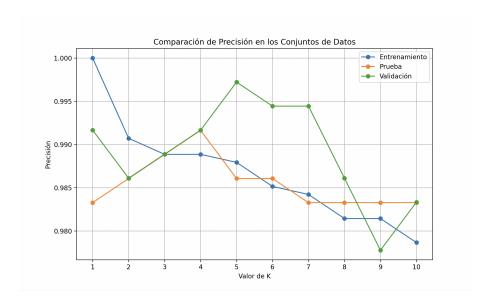


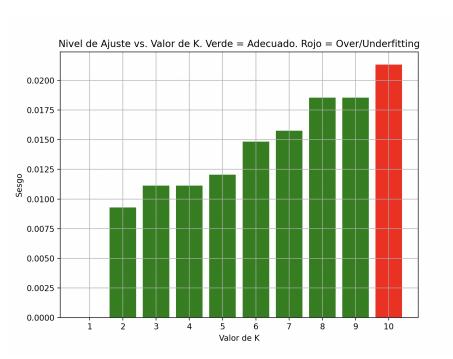
3. Mejora del Modelo

K-Nearest-Neighbor es un algoritmo eficiente y que sólo necesita de un hiper parámetro, este siendo el valor de los K vecinos cercanos que consideraremos para realizar predicciones en torno a la predicción. En ese sentido, y puesto que realizamos la evaluación del modelo de manera iterativa, nos fue posible conocer el mejor valor de los K vecinos en una sola iteración.

En este caso, el modelo ya muestra un excelente rendimiento, puesto que, gracias a su iteración con el valor de K, nos es posible encontrar el modelo mejor ajustado en una sola ejecución. Para respaldar la anterior premisa obtuvimos toda la información necesaria (sesgo, varianza y precisión) para cada valor posible de K.

Por lo anterior podemos concluir que no se observan problemas de sesgo o varianza significativos.





4. Conclusiones

- El modelo de clasificación evaluado muestra un alto rendimiento en términos de precisión en todos los conjuntos (entrenamiento, prueba y validación).
- No se observan problemas significativos de sesgo ni varianza en este escenario.
- El modelo tiene un buen ajuste y es capaz de generalizar bien a nuevos datos.

En resumen, el modelo actual es altamente efectivo para la tarea de clasificación en el dataset "Digits" e "Iris" y no requiere ajustes adicionales en este momento.