Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

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

Reporte sobre el desempeño del modelo

Yaritzi Itzayana Nicio Nicolás

A01745837

Profesor: Jorge Alfredo Ramírez Uresti

Reporte sobre el desempeño del modelo

A lo largo de este documento se estará hablando sobre el desempeño del modelo KNN con el uso de framework. En primera instancia se hizo la selección del dataset a utilizar, el cual fue el de Iris. Dicho dataset se puede encontrar en diferentes plataformas como Kaggle o incluso (como se ve en el código) se puede descargar desde la librería de *sklearn*. La razón de este dataset se debe a que ya se había trabajado con él en un modelo anterior de KNN sin el uso de un framework, por lo que permitiría comparar las diferencias entre el uso de un algoritmo desarrollado desde cero y un algoritmo con uso de framework. De la misma manera, el dataset de Iris es uno de los más populares para el uso de modelos de regresión . Así mismo, aunque es una base relativamente pequeña, tiene los datos suficientes para el desarrollo del entrenamiento del modelo así como obtener predicciones bastante precisas.

1. **Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación (Train/Test/Validation).**

El manejo de los datos inició con la división de los mismos en datos de entrenamiento, prueba y validación. Cada uno de ellos cumple con una función diferente al momento de implementar el modelo. Los datos de entrenamiento ayudan a confirmar si es necesario modificar parámetros; los datos de prueba ayudan a evaluar el modelo final mientras que los datos de validación ayudan a evaluar los ajustes vistos en los datos de entrenamiento. Así, se dividieron los datos en 3 con las siguientes líneas de código.

```
# Se dividen los datos en train, test y validation
#Se genera la semilla con el stándar de 42 para que el modelo sea reproducible
#Primero se divide en un conjunto temporal y en prueba
X_temp, X_valid, y_temp, y_valid = train_test_split(X, y, test_size=0.2, random_state=42)
#Se divide de nuevo, ahora en un conjunto de entrenamiento y validación con el conjunto temporal anterior
X_train_c, X_test_c, y_train_c, y_test_c = train_test_split(X_temp, y_temp, test_size=0.5, random_state=42)
```

Figura 1. Código donde se separan los datos

Primero, se divide en entrenamiento y prueba, teniendo 80% entrenamiento, 20% prueba, así obtenemos el conjunto de datos de validación. Se tienen las variables de X_temp y y_temp que son variables temporales utilizadas en la siguiente línea de código donde se vuelve a dividir en entrenamiento y prueba, 50% entrenamiento, 50% prueba, que dividen el dataset en los conjuntos de datos de entrenamiento y de prueba.

2. **Diagnóstico y explicación el grado de bias o sesgo**

Para el sesgo de los datos es necesario observar las matrices de confusión. Enfocándonos en los datos de prueba, se podrá observar como cuando hay un valor menor de k, hay más errores de clasificación. Sin embargo, cuando el valor de k llega a 6, se establece el número de errores para la clasificación y obtenemos una precisión mayor.

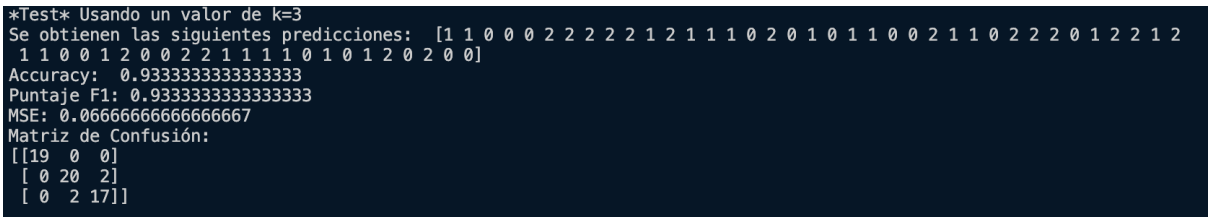

Figura 2. Datos de prueba con k = 3

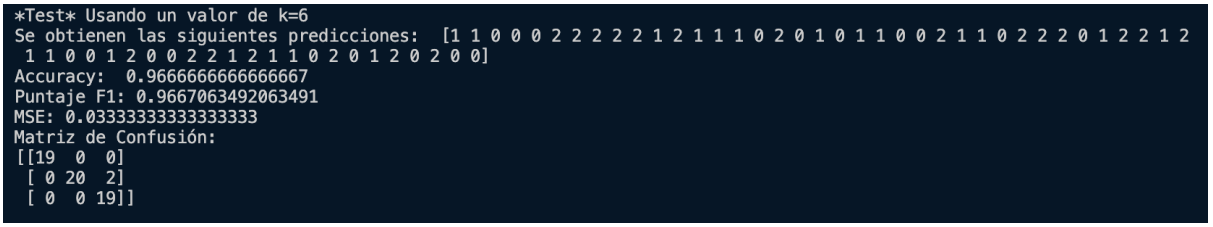

Figura 3. Datos de prueba con k = 6

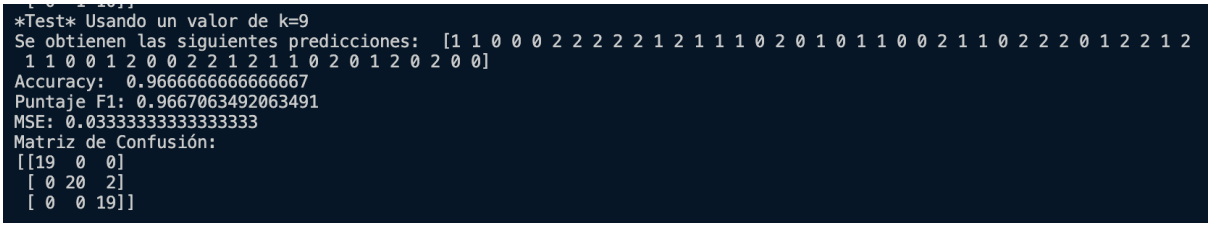

Figura 4. Datos de prueba con k = 9

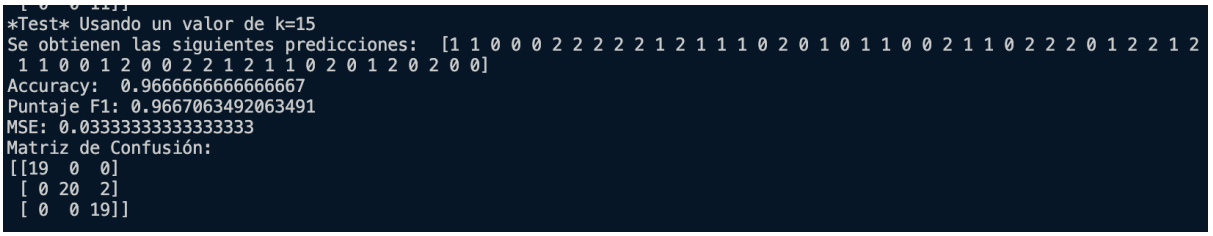

Figura 5. Datos de prueba con k=15

No obstante, no se muestra una diferencia significativa entre los valores de k entre 6, 9 y 15, por lo que se optó por ver valores mayores y se observaron los siguientes resultados.

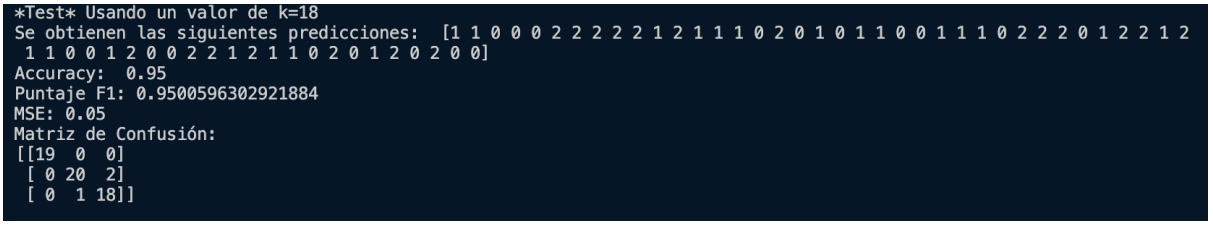

Figura 6. Datos de prueba con k= 18

```
*Test* Usando un valor de k=21
Se obtienen las siguientes predicciones:  [1 1 0 0 0 2 2 2 2 2 1 2 1 1 1 0 2 0 1 0 1 1 0 0 2 2 1 0 2 2 2 0 1 2 2 1 2
 1 1 0 0 1 2 0 0 2 2 1 2 1 1 0 2 0 1 2 0 2 0 0]
Accuracy:  0.95
Puntaje F1: 0.95
MSE: 0.05
Matriz de Confusión:
[[19  0  0]
 [ 0 19  3]
 [ 0  0 19]]
```

Figura 7. Datos de prueba con k=21

Se observa un incremento en el número de errores (a pesar de no incrementar demasiado), por lo que se puede considerar que el sesgo para valores de k mayores de 15 es mayor al igual que el sesgo para valores de k menores a 6.

3. Diagnóstico y explicación el grado de varianza

De la misma forma que pasó con el sesgo, la varianza no se vió tan modificada con valores de k entre 6 y 15. Sin embargo, por fines de demostrar el cambio con diferentes valores de k, se evaluó nuevamente el modelo en los valores de k de 18 y 21, donde se observan los siguientes resultados.

```
*Validación* Usando un valor de k=18
Se obtienen las siguientes predicciones:  [1 0 2 1 1 0 1 2 1 1 1 0 0 0 0 1 2 1 1 2 0 1 0 2 2 2 2 2 0 0]
Accuracy:  0.9333333333333333
Puntaje F1: 0.9333333333333333
MSE: 0.06666666666666667
Matriz de Confusión:
[[10  0  0]
 [ 0  9  0]
 [ 0  2  9]]
*Test* Usando un valor de k=18
Se obtienen las siguientes predicciones:  [1 1 0 0 0 2 2 2 2 2 1 2 1 1 1 0 2 0 1 0 1 1 0 0 1 1 1 0 2 2 2 0 1 2 2 1 2
 1 1 0 0 1 2 0 0 2 2 1 2 1 1 0 2 0 1 2 0 2 0 0]
Accuracy:  0.95
Puntaje F1: 0.9500596302921884
MSE: 0.05
Matriz de Confusión:
[[19  0  0]
 [ 0 20  2]
 [ 0  1 18]]
```

Figura 8. Datos de validación y prueba con k=18

```
*Validación* Usando un valor de k=21
Se obtienen las siguientes predicciones:  [2 0 2 2 2 0 1 2 2 1 2 0 0 0 0 2 2 1 2 2 0 2 0 2 2 2 2 2 0 0]
Accuracy:  0.8
Puntaje F1: 0.7714285714285716
MSE: 0.2
Matriz de Confusión:
[[10  0  0]
 [ 0  3  6]
 [ 0  0 11]]
*Test* Usando un valor de k=21
Se obtienen las siguientes predicciones:  [1 1 0 0 0 2 2 2 2 2 1 2 1 1 1 0 2 0 1 0 1 1 0 0 2 2 1 0 2 2 2 0 1 2 2 1 2
 1 1 0 0 1 2 0 0 2 2 1 2 1 1 0 2 0 1 2 0 2 0 0]
Accuracy:  0.95
Puntaje F1: 0.95
MSE: 0.05
Matriz de Confusión:
[[19  0  0]
 [ 0 19  3]
 [ 0  0 19]]
```

Figura 8. Datos de validación y prueba con k=21

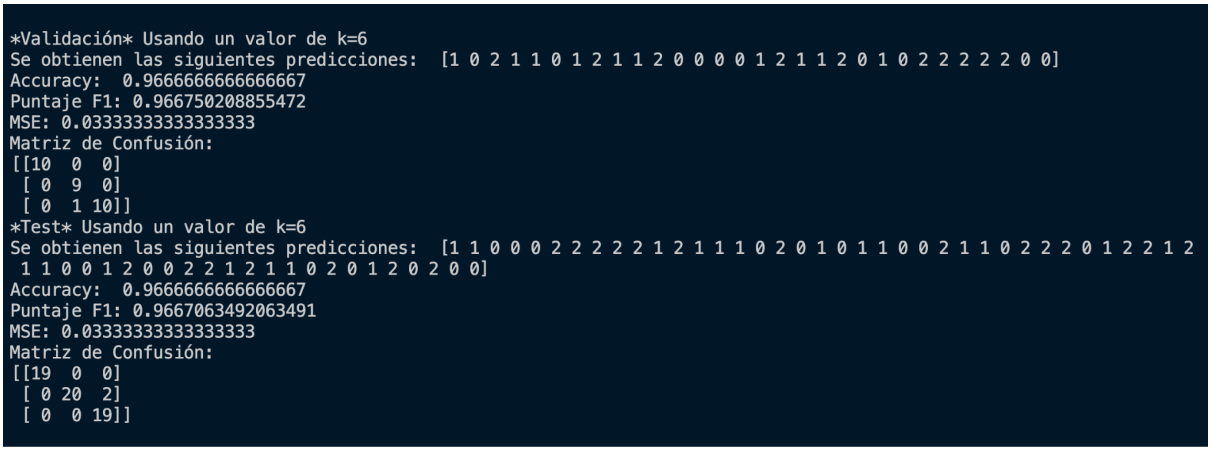

Figura 9. Datos de validación y prueba con k= 6

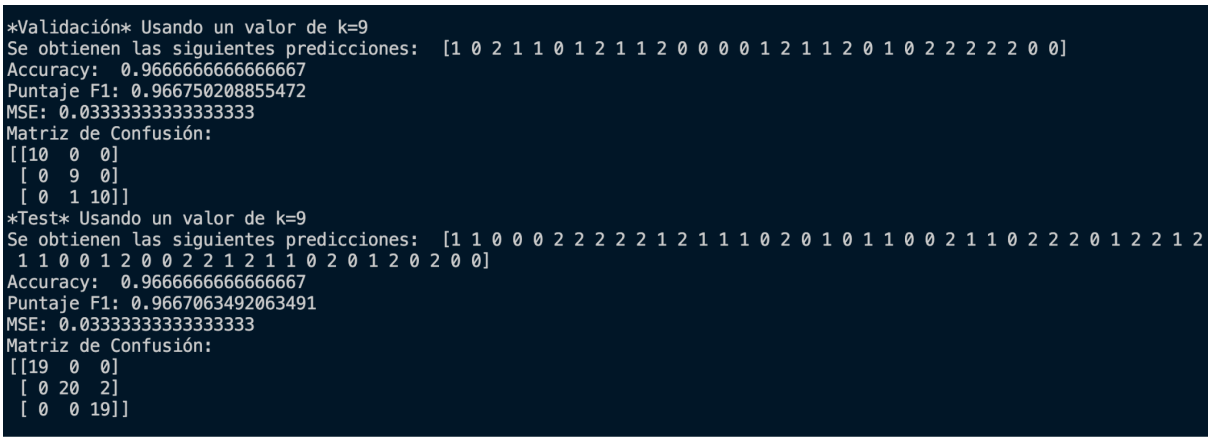

Figura 10. Datos de validación y prueba con k = 9

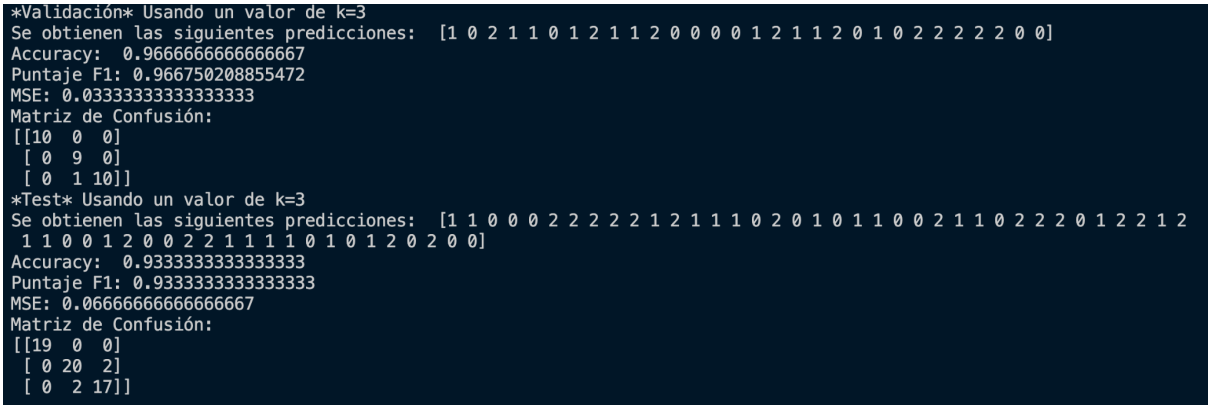

Figura 11. Datos de validación y prueba con k = 3

Se puede observar como en los valores de 6 y 9 la varianza es muy baja, mientras que para el valor de k = 21, la varianza tiene un cambio significativo. Se puede concluir entonces que el cambio de varianza está determinado por el valor de k, y que valores muy altos o muy bajos la pueden modificar drásticamente.

4. **Diagnóstico y explicación el nivel de ajuste del modelo**

Se puede observar que en general el modelo implementado presenta buenas métricas siempre y cuando los valores de las k, estén entre 6, 9 y 15. Dejando así de lado, los valores que modifican la varianza y el sesgo de los datos (como k=3 o k=21). Cabe resaltar que el uso de k está dentro de los valores impares ya que en caso de utilizar valores pares, hay una

posibilidad de empate de algunas predicciones. Al estar ajustando manualmente los parámetros a utilizar para el modelo, no hay necesidad de hacer un ajuste extra. Así el parámetro más optimizable es de k=6, ya que valores más grandes nos arrojan resultados con más errores y con una varianza mayor. Para justificar el modelo optimizado, se hizo un cross validation que nos permite estimar la capacidad de generalización del modelo y evaluar su rendimiento de una manera más confiable, obteniendo los siguientes resultados.

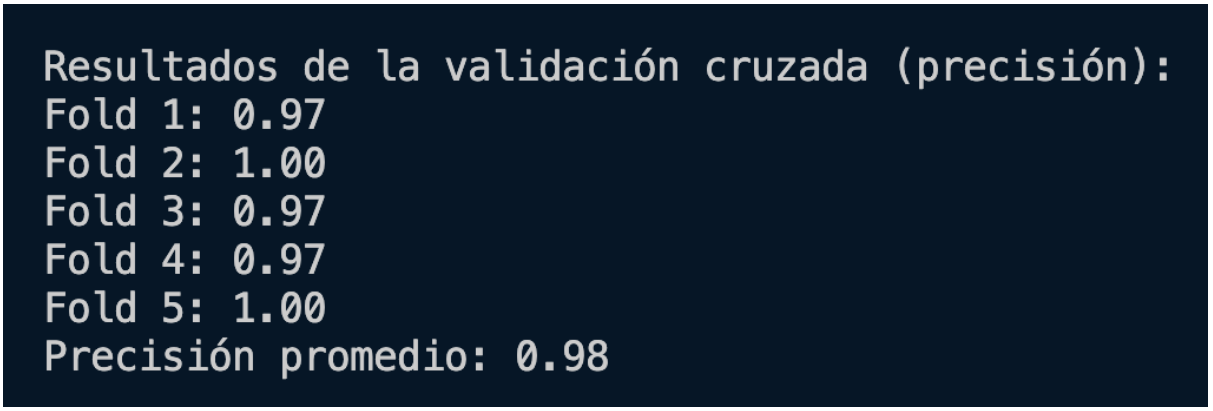

Figura 12. Resultados de cross-validation

Con una precisión promedio del 98% se puede concluir que el modelo tiene un excelente rendimiento. Así, el modelo es capaz de realizar predicciones en datos no vistos con una alta precisión.

En conclusión, es importante considerar el cambio de parámetros para los modelos, ya que esto nos permite observar las diferencias que nos ayuden a validar si el modelo ha mejorado o ha empeorado. Así mismo, es importante la división de los datos que nos permita hacer una evaluación del modelo. Si bien, la división presentada en este documento hubiera sido diferente, es probable que el modelo resultante tuviera métricas y resultados diferentes aunque cercanos.

Anexo

A continuación se muestran algunas de las gráficas obtenidas del modelo que también ayudaron a tener una mejor visión sobre el desempeño del mismo.

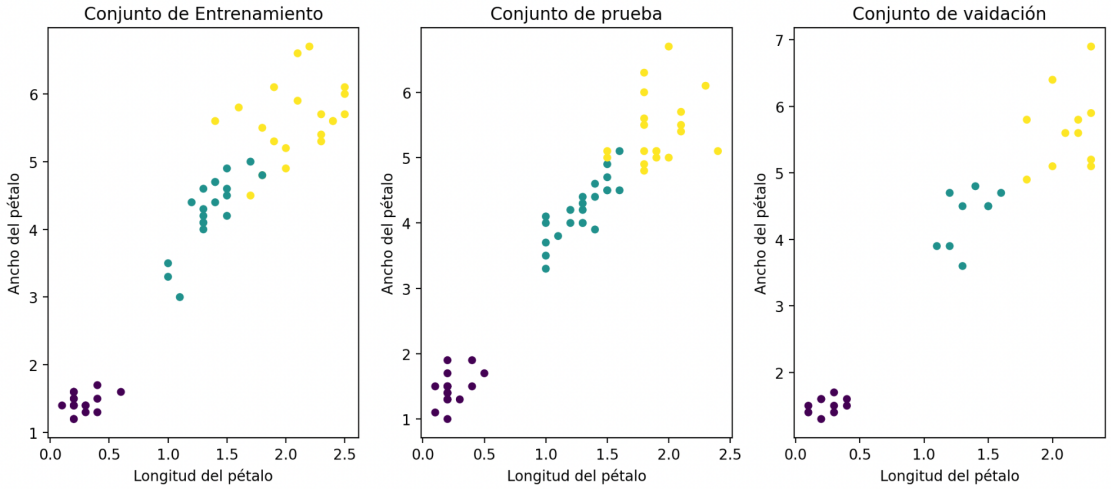

Figura 13. Distribución de los datos en los conjuntos de prueba, entrenamiento y validación

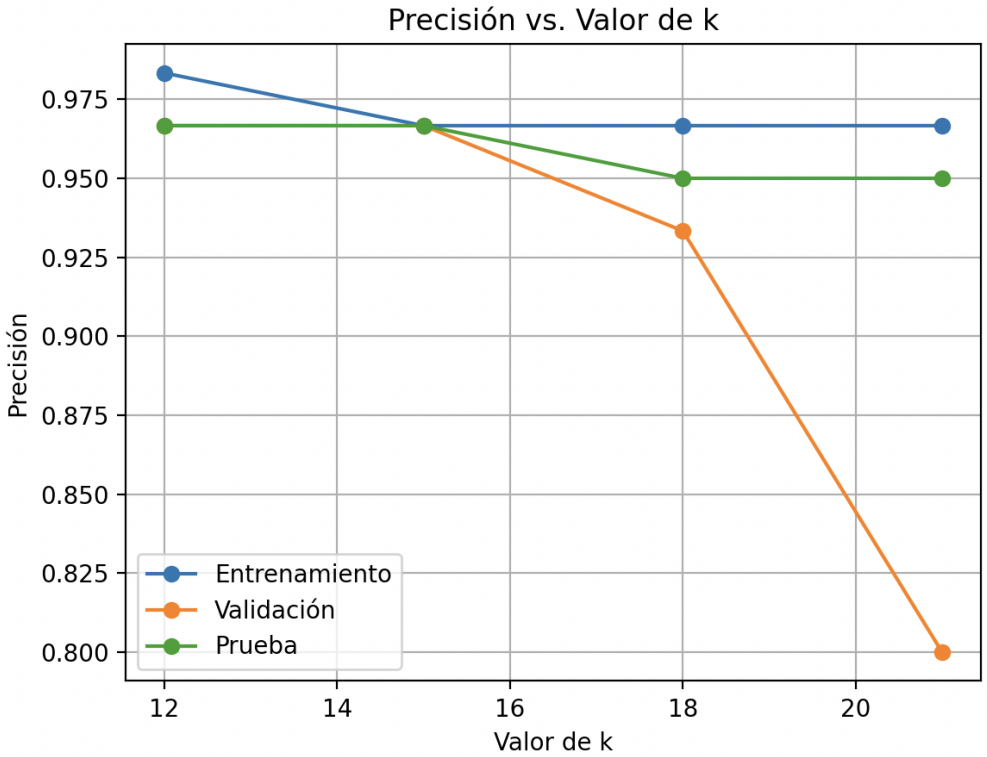

Figura 14. Precisión vs el valor de k

En la figura 14 se observa cómo cuando el valor de k aumenta, la precisión disminuye.

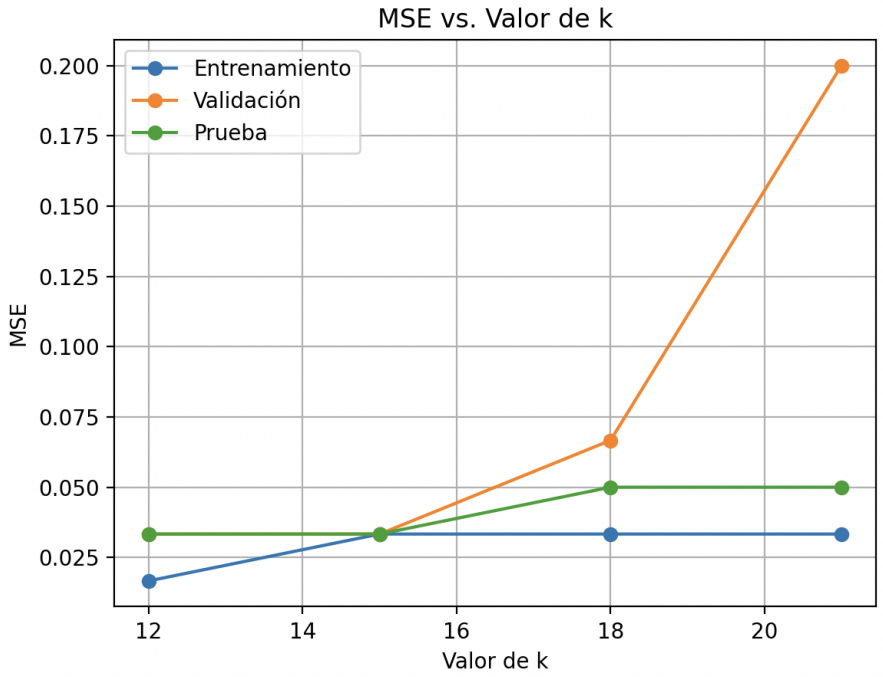

Figura 15. MSE vs Valor de k

Así mismo, en la figura 15 se observa cómo cuando aumenta el valor de k, también aumenta el valor del MSE.

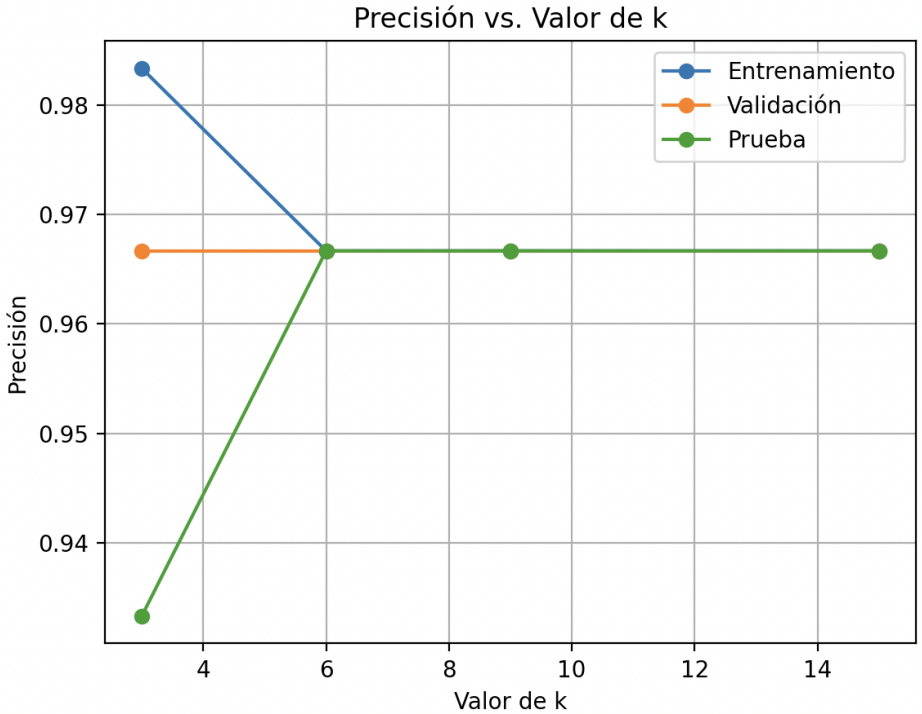

Figura 16. Precisión vs Valor de k. Valores entre 3 y 15

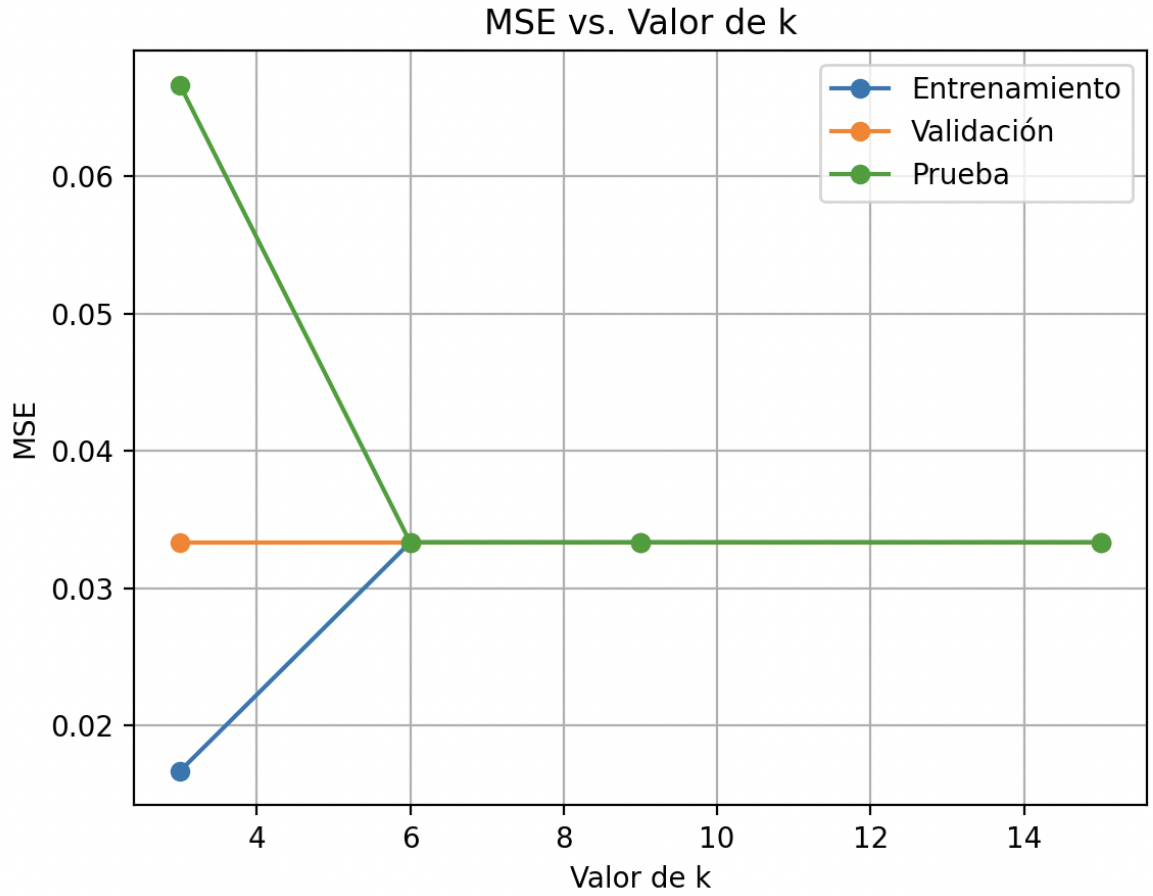

Figura 17. MSE vs valor de K. Valores entre 3 y 15.

En está última gráfica se puede observar que el valor del MSE disminuye cuando el valor de k es 6.