Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Estado de México

Inteligencia artificial avanzada para la ciencia de datos I

# Análisis y Reporte sobre el desempeño del modelo.

Nadia Paola Ferro Gallegos - A01752013

Profesor:

Jorge Adolfo Ramírez Uresti



**11 de Septiembre, 2023**

## Selección del conjunto de datos "Iris"

Para la realización de esta tarea se utilizó el algoritmo KNN el cual se refiere a k vecinos mas cercanos. KNN o k-NN es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual. La clasificación la realiza haciendo suposiciones de que se pueden encontrar puntos similares cerca uno del otro. Debido a la forma en la que clasifica la información el KNN se utiliza principalmente como un algoritmo de clasificación a pesar de que también puede ser utilizado como algoritmo de regresión.

Para los problemas de clasificación, se asigna una etiqueta de clase en base a la etiqueta que se representa con mayor frecuencia alrededor de un punto de datos determinado. Para los problemas de regresión se utiliza un proceso similar, sin embargo, se toma un promedio de los k vecinos mas cercanos para hacer una predicción. También es importante aclarar que para que se pueda hacer la clasificación, se debe definir la distancia, en este caso vamos a utilizar la distancia euclidiana.

Para esta tarea se utilizaron diferentes conjuntos de datos para probar el algoritmo programado, pero para la realización de este reporte se va a utilizar el conjunto de datos "Iris". La elección del conjunto de datos "Iris" se basa en varias consideraciones para poder demostrar la generalización de un modelo de aprendizaje automático.

Para empezar el conjunto de datos "Iris" es un conjunto de clasificación multiclase que consta de tres diferentes clases de plantas iris (setosa, versicolor y virginca) y cuatro características (longitud de sépalo, ancho de sépalo, longitud de pétalo y ancho de pétalo). Esta característica multiclase permite evaluar como el KNN generaliza en una tarea de clasificación con múltiples clases.

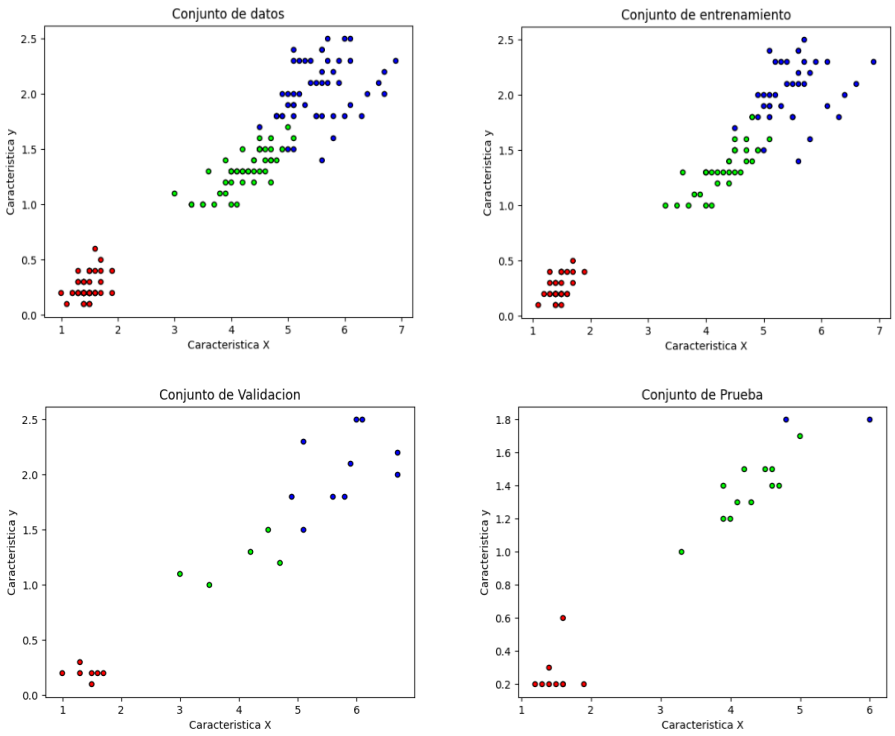
Por otro lado, el conjunto de datos "Iris" tiene un tamaño moderado y las características del conjunto de datos iris son numéricas y no requieren un procesamiento complejo. Estas características lo hacen manejable y facilita la implementación y evaluación del modelo KNN.

## 1. Separación y evaluación del modelo

Para la separación y evaluación del modelo primero se dividió el conjunto de datos en conjuntos de entrenamiento, validación y prueba. Esto permite entrenar al modelo, ajustar hiperparametros y evaluar su rendimiento en datos no vistos. El conjunto de entrenamiento se utilizó para entrenar el modelo.

Para empezar, vamos a presentar como se hizo la distribución de los datos en tres diferentes conjuntos. El conjunto de entrenamiento se utilizo para entrenar el modelo KNN, se selecciono el 70% de los datos para que el modelo aprendiera de estos datos. Después el conjunto de validación se utilizo para ajustar los hiperparametros y evaluar el rendimiento del modelo en datos no vistos durante el entrenamiento, para este conjunto se destino el 15% de los datos. A pesar de que el modelo no se entrena con estos datos se utilizan para tomar decisiones sobre la configuración optima del modelo, el cual mostraremos mas adelante. Por ultimo el conjunto de datos de prueba se utilizo para evaluar el rendimiento optimo del modelo, donde se simula como el modelo se desempeñaría en una situación real.

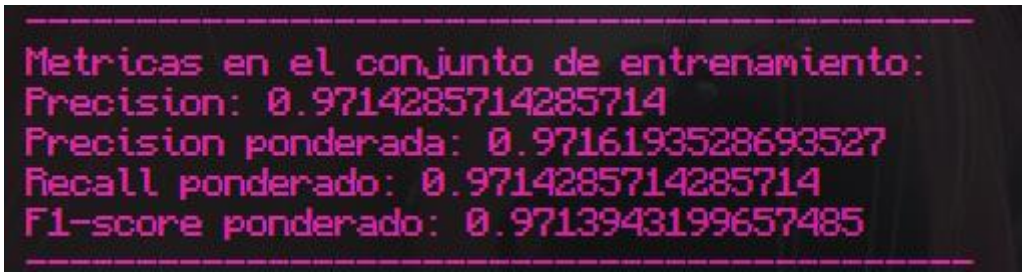
A continuación, se puede observar el conjunto de datos completo y la distribución de los datos en los tres conjuntos de entrenamiento, validación y pruebas:



## 2. Diagnóstico y explicación del sesgo

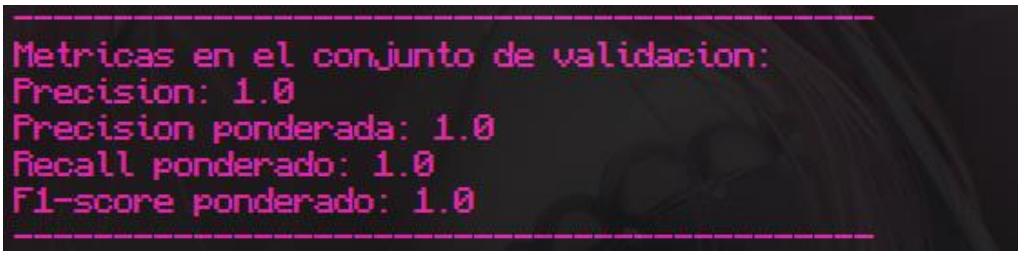
Al momento de hablar de los sesgos en loa modelos de aprendizaje nos referimos a las diferencias entre las predicciones del modelo y los valores reales. Para evaluar el desempeño del modelo, se utilizaron métricas en los conjuntos de entrenamiento, validación y prueba. Los cuales nos van a ayudar a determinar si el modelo sufre de bajo sesgo, un sesgo medio o un sesgo alto. El bajo sesgo ocurre cuando el modelo no esta capturando la complejidad de los datos, por lo que se podría decir que es demasiado simple. El sesgo medio indica un equilibrio, por lo que se considera aceptable. Por último, el sesgo alto se produce cuando el modelo es muy complejo, por lo que a pasar de que se adapta a los datos de entrenamiento a la perfección puede tener problemas al momento de generalizar.

Para empezar, se realizaron predicciones en el conjunto de entrenamiento y se calcularon las siguientes métricas:



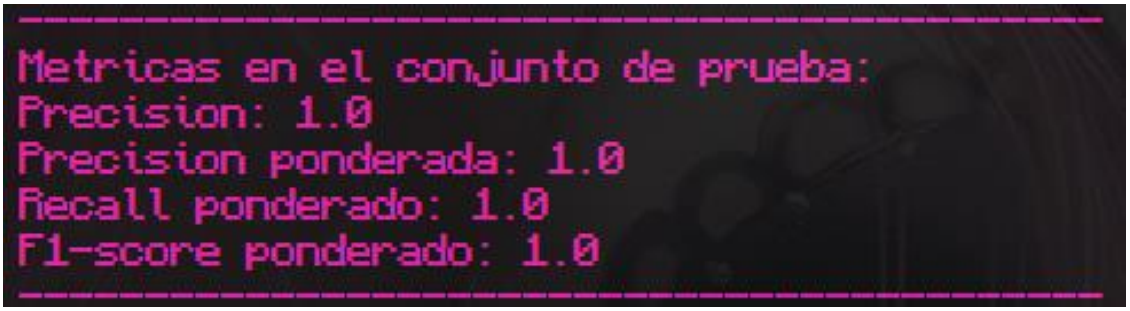

El conjunto de entrenamiento logro una precisión cercana al 97.14% lo que indica que es capaz de predecir de la forma correcta los datos de entrenamiento. El Recall ponderado también es alto por lo que es eficaz en la recuperación de muestras. Por ultimo el valor de F1-Score también es elevado, por lo que nos podría indicar un sesgo bajo en el conjunto de entrenamiento.

Luego se realizaron predicciones en el conjunto de validación y se calcularon las siguientes métricas:



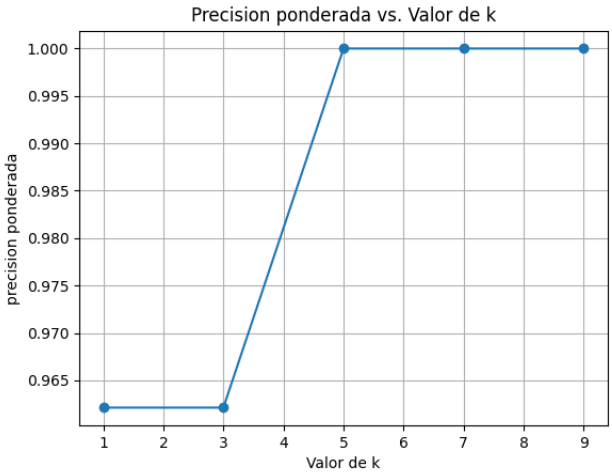
En el conjunto de validación podemos observar que se alcanzó una precisión del 100%, además el recall ponderado y el F1-Score también son perfectos. Por lo que se podría tratar de un posible sobreajuste.

Luego se realizaron predicciones en el conjunto de prueba y se calcularon las siguientes métricas:



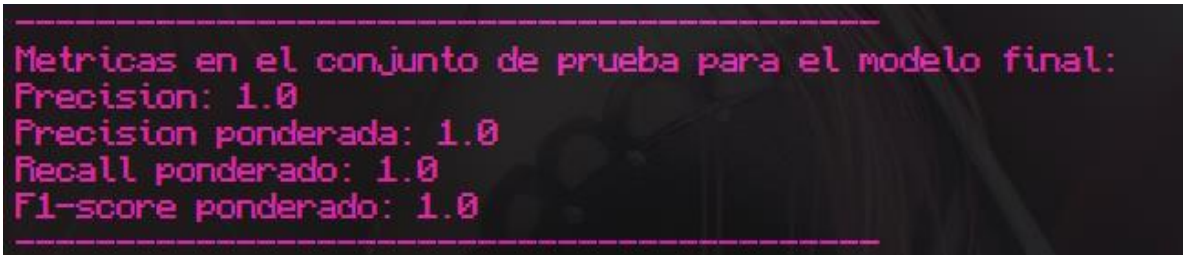

En el conjunto de prueba podemos observar que se alcanzó una precisión del 100%, además el recall ponderado y el F1-Score también son perfectos. Por lo que se podría tratar de un posible sobreajuste en los datos de prueba.

Después, se realizó una búsqueda de hiperparametros variando el valor de k (1, 3, 5, 7, 9) y calculando las métricas de evaluación en el conjunto de validación para cada valor de k logramos obtener la siguiente grafica comparativa de la precisión ponderada versus el valor de k:



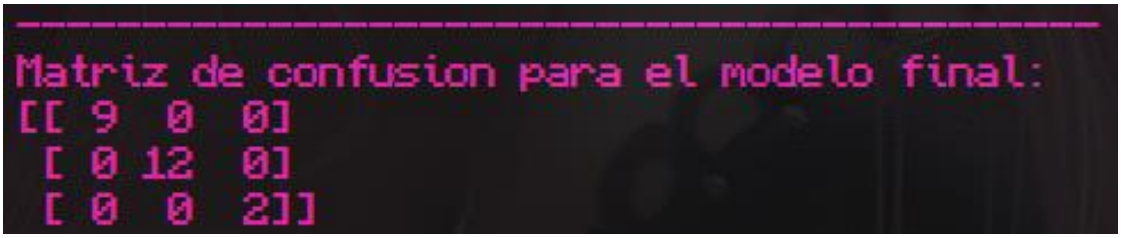
Como se puede observar el mejor valor de k en el conjunto de validación fue cuando k=5, por lo que finalmente se entrenó el modelo KNN con todos los datos de entrenamiento utilizando el mejor valor de k encontrado.

Se realizaron las mismas predicciones que en el conjunto de validación y se calcularon las siguientes métricas para el conjunto de prueba para el modelo final:




En el conjunto de prueba para el modelo final también podemos observar que se alcanzó una precisión del 100%, el recall ponderado y el F1-Score también son perfectos. Por lo que también se podría tratar de un posible sobreajuste en estos datos.

Por último, se despliega la siguiente matriz de confusión para el modelo final, donde se muestra la distribución de las predicciones del modelo en cada clase:




Estos resultados en general muestran el rendimiento del KNN en el conjunto de entrenamiento, validación y en el conjunto de prueba. La precisión ponderada mide la precisión de un modelo en todas las clases, en este caso podemos observar que las métricas muestran una precisión perfecta en la mayoría de los conjuntos, por lo que el modelo predice un gran porcentaje de las muestras. La métrica de Recall ponderado mide la capacidad del modelo para recuperar muestras positivas en todas las clases, de acuerdo con los resultados en las métricas se puede observar que el modelo es capaz de identificar todas las muestras

relevantes. El F1-Score Ponderado, combina la precisión y el Recall, por lo que de acuerdo con los resultados en las métricas se puede observar un gran equilibrio entre la precisión y la capacidad de recuperación en las métricas.

Por último, la matriz de confusión revela que en el modelo final todas las muestras se clasificaron correctamente en sus respectivas clases y también nos muestra que el valor de k=5 que se seleccionó funciono bien con el conjunto de prueba.

Basándonos en los resultados generales se podría decir que este modelo KNN muestra un bajo sesgo en todos los conjuntos, ya que es capaz de ajustarse bien a todos los datos que le enviamos, las altas métricas pueden confirmar esta declaración.

## 3. Diagnóstico y explicación de la varianza

El grado de varianza se refiere a la capacidad del modelo para manejar la variabilidad de datos. Una alta varianza indica que el modelo es sensible a tener pequeñas fluctuaciones por lo que indica que le puede costar generalizar en datos no vistos. Por otro lado, una varianza baja indica que el modelo es estable y tiende a generalizar mejor.

En nuestro caso, en el conjunto de entrenamiento se logró una precisión cercana al 97.14%, después en el conjunto de validación y prueba el modelo obtuvo una precisión del 100%. De acuerdo con estos datos se podría interpretar como una baja varianza o como un sobreajuste, ya que se esta logrando un rendimiento perfecto. En este caso la aplicación de técnicas de regularización podría ayudar a controlar la varianza del modelo y mejorar su capacidad de generalización de nuevos datos. Por lo que, en términos de varianza podríamos clasificar el modelo actual como medio debido a la posibilidad de sobreajuste, pero hay una posibilidad de mejora mediante la optimización de hiperparametros y estrategias de control de varianza.

## 4. Diagnóstico y explicación del nivel de ajuste

El nivel de ajuste se refiere a la capacidad de un modelo para capturar y representar correctamente la relación entre las características de entrada y la variable objetivo. Actualmente, existen tres niveles de ajuste Underfitting, Fitting y Overfitting. Cuando un modelo esta en Underfitting implica que no ha aprendido adecuadamente las características y la variable objetivo, por lo que resulta en un modelo demasiado simple y no se pueden

hacer predicciones precisas. Un modelo que esta en un ajuste adecuado o en Fitting, es capaz de ajustar adecuadamente las características de entrada y la variable objetivo, por lo que el modelo generaliza bien los datos y tiene un buen equilibrio entre sesgo y varianza. Cuando un modelo esta en Overfitting significa que se ha adaptado en exceso a los datos de entrenamiento, como resultado el modelo muestra un rendimiento excepcionalmente alto en el conjunto de entrenamiento, pero presenta un rendimiento deficiente en el resto de los conjuntos.

Como se menciono anteriormente el modelo muestra un rendimiento alto en el conjunto de entrenamiento y en los conjuntos de validación y prueba muestra un rendimiento perfecto. Lo que podría indicar que hay un sobreajuste en los conjuntos de prueba y validación. Por lo que se podría calificar al modelo actual como Overfitting, debido a su alta precisión. De la misma manera, como se mencionó anteriormente se podrían utilizar técnicas de regularización o ajuste de parámetros para lograr un mayor equilibrio en el nivel de ajuste.

## Uso de técnicas de regularización

Para mejorar el rendimiento del modelo KNN que tenemos y arreglar el problema de Overfitting que encontramos anteriormente vamos a utilizar tres técnicas de regularización. La primera técnica es el ajuste de hiperparametros, donde vamos a utilizar una cuadricula para probar diferentes valores de k y encontrar el que optimice el rendimiento en el conjunto de validación. Después utilizamos la validación cruzada para evaluar el modelo de manera mas profunda y precisa.