# Clasificación de Especies de Iris mediante k-Vecinos más Cercanos

**Análisis de Aprendizaje Automático - Dataset IRIS**

## Resumen

Este estudio implementa el algoritmo k-Vecinos más Cercanos (k-NN) para clasificar especies de iris basándose en características morfológicas. El modelo alcanzó una precisión del 100% identificando patrones distintivos entre *Iris setosa*, *Iris versicolor*, e *Iris virginica*. Las medidas de pétalos demostraron ser los predictores más discriminatorios para la clasificación exitosa.

## Análisis Exploratorio del Notebook

### Revisión del Dataset y Procesamiento

El dataset IRIS comprende 150 muestras balanceadas (50 por especie) con cuatro características morfológicas: longitud y ancho del sépalo, y longitud y ancho del pétalo. El análisis exploratorio confirmó la ausencia de valores faltantes y distribuciones bien diferenciadas entre especies. El preprocesamiento incluyó la eliminación de la columna ID y la división aleatoria en conjuntos de entrenamiento (120 muestras) y prueba (30 muestras) con random\_state=42.

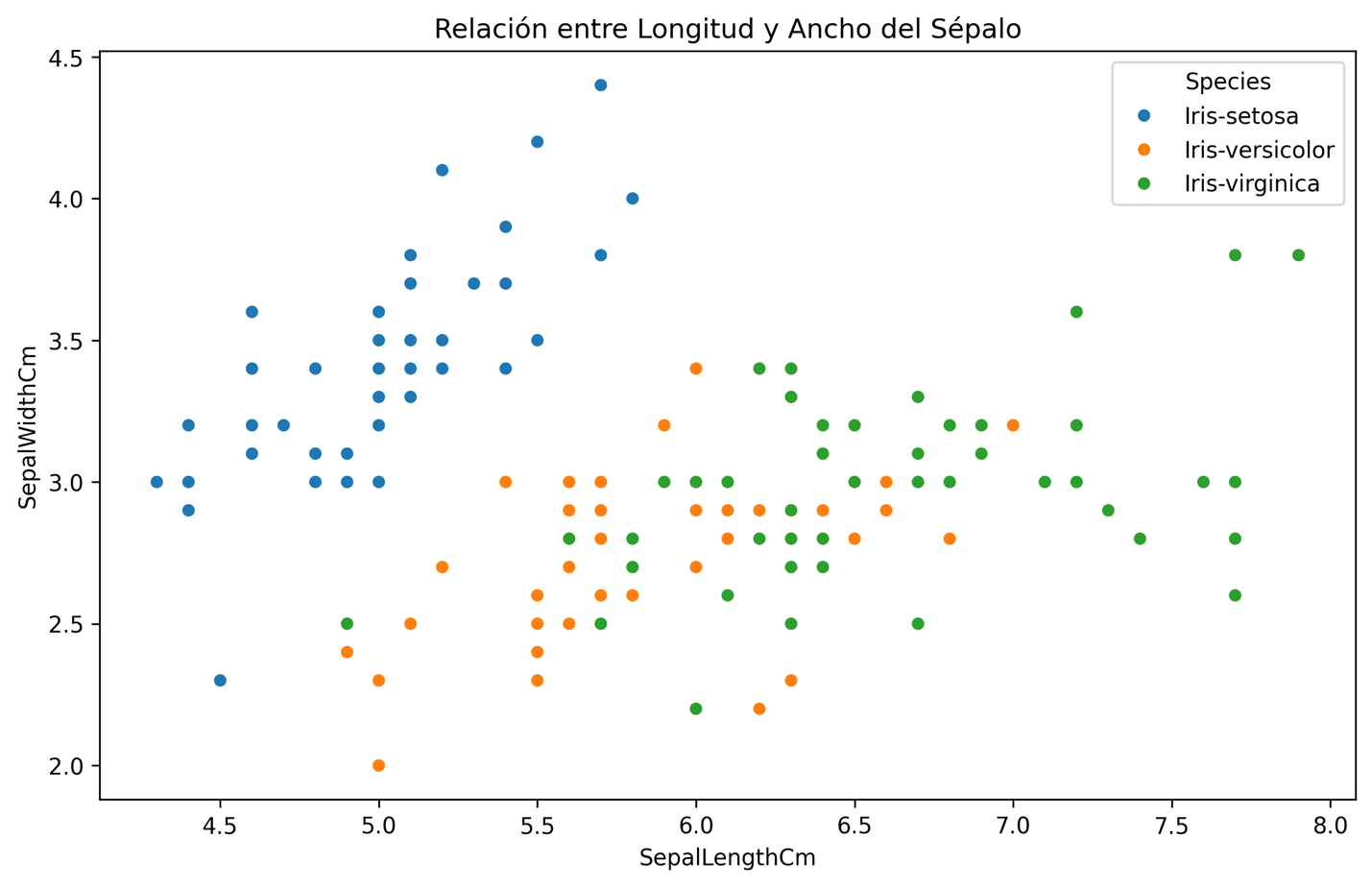
### Implementación del Modelo k-NN

Se implementó el algoritmo k-NN con k=5 vecinos, entrenando sobre 120 muestras y evaluando en 30 muestras de prueba. El modelo clasificó correctamente las 30 muestras del conjunto de prueba, alcanzando una precisión del 100%. La matriz de confusión mostró clasificación perfecta: 10 *Iris setosa*, 9 *Iris versicolor*, y 11 *Iris virginica*, todas correctamente identificadas.

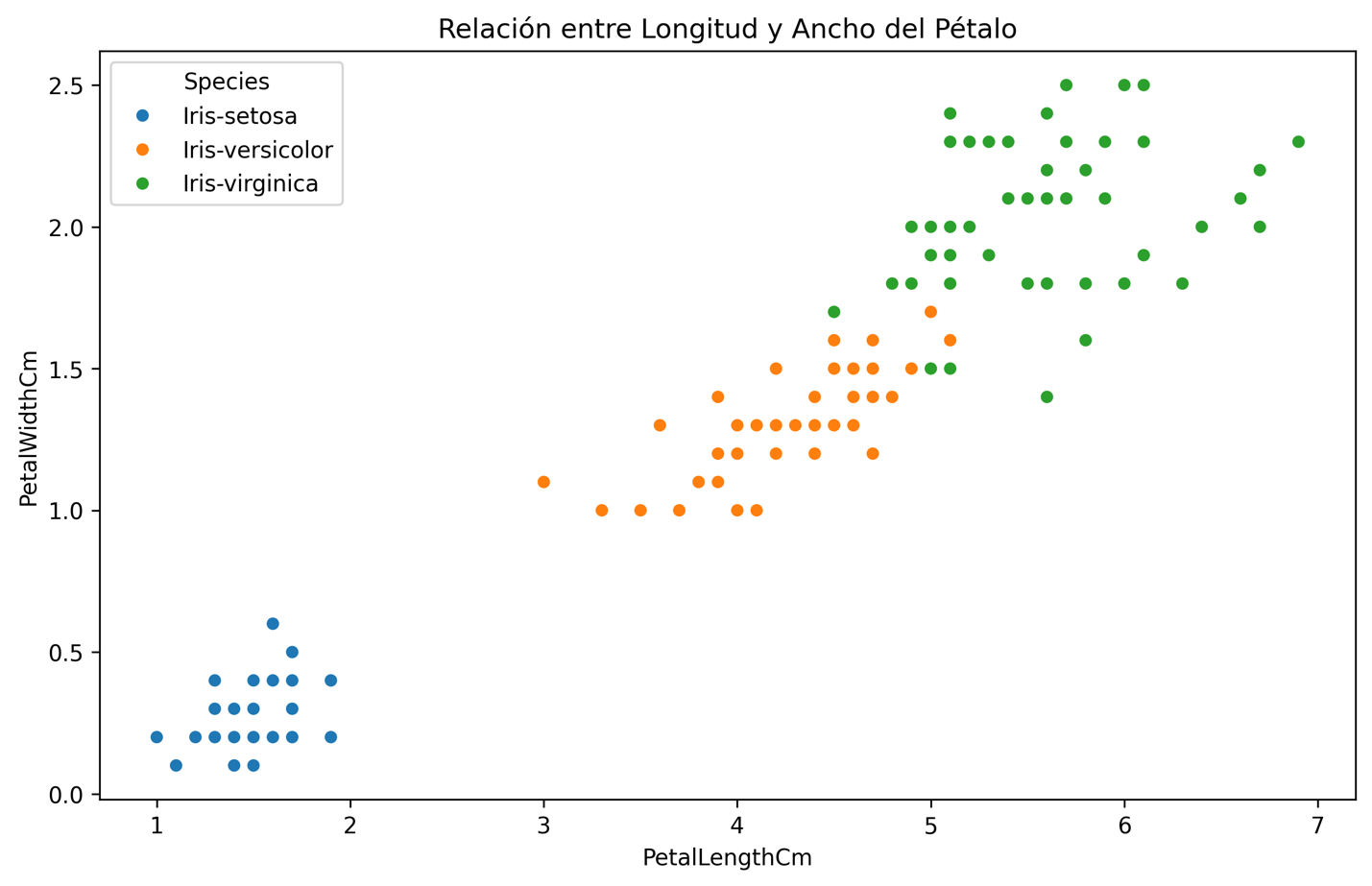
## Hallazgos y Observaciones Clave

### ¿Qué patrones se observan en los gráficos?

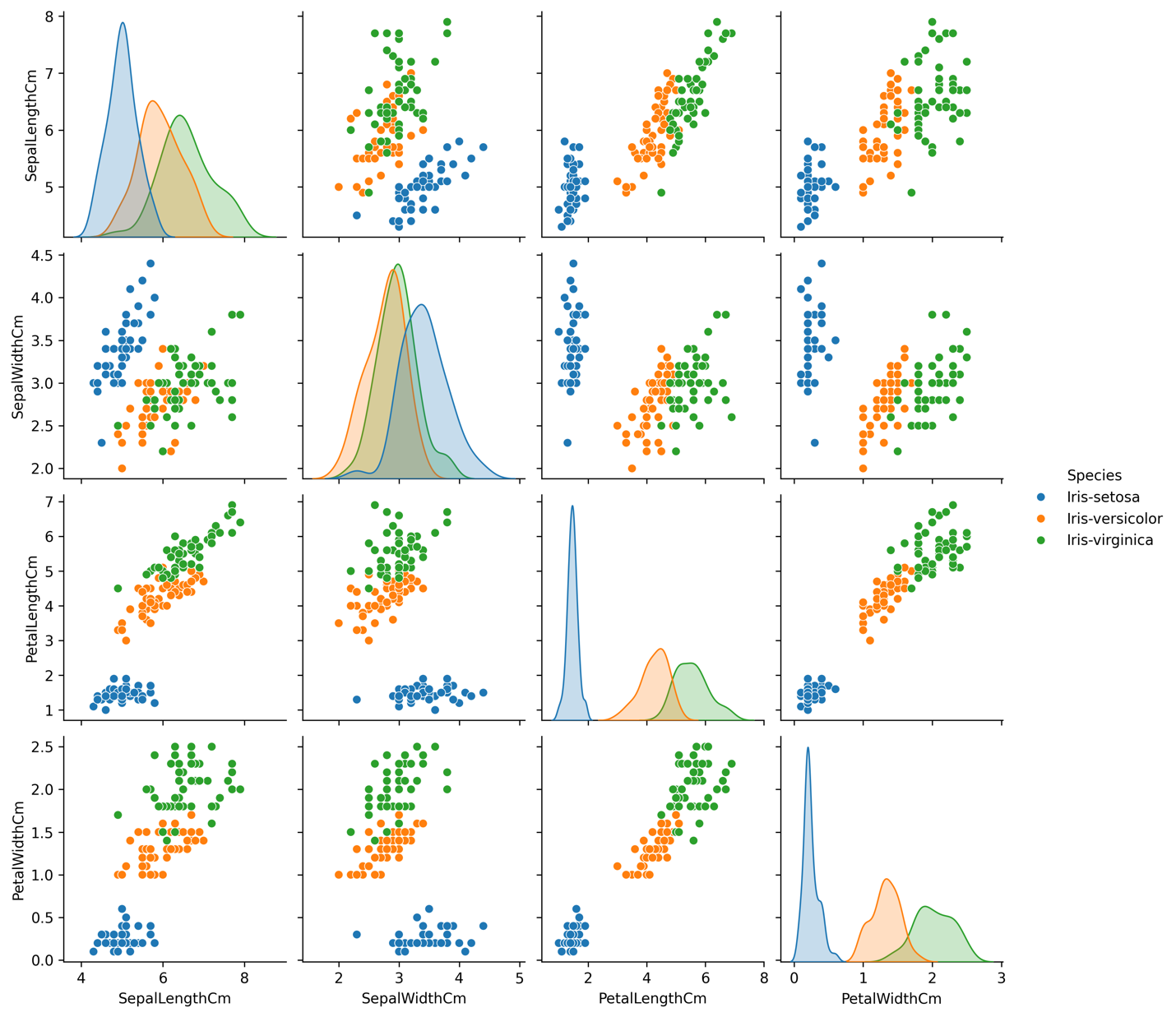
**Gráficos de Sépalo (Longitud vs Ancho):** Los gráficos de características de sépalo revelan solapamiento considerable entre *Iris versicolor* e *Iris virginica*, mientras que *Iris setosa* forma un cluster relativamente bien definido. *Iris setosa* se caracteriza por sépalos más cortos (4-6 cm) pero proporcionalmente más anchos (2.5-4.5 cm). *Iris virginica* tiende hacia sépalos más largos (6-8 cm) pero más estrechos (2-4 cm), con *Iris versicolor* ocupando una posición intermedia con superposición significativa.



**Gráficos de Pétalo (Longitud vs Ancho):** Las características de pétalos muestran separación casi perfecta entre las tres especies. *Iris setosa* forma un cluster extremadamente compacto con pétalos pequeños (1-2 cm longitud, 0.1-0.6 cm ancho). *Iris versicolor* ocupa una región intermedia bien delimitada (3-5 cm longitud, 1.0-1.8 cm ancho), mientras que *Iris virginica* se distingue por pétalos grandes (4.5-7 cm longitud, 1.5-2.5 cm ancho) con mínimo solapamiento entre especies.



**Matriz de Dispersión (Pairplot):** El pairplot confirma estos patrones y revela correlaciones positivas fuertes entre longitud y ancho de pétalos. Los histogramas en la diagonal muestran distribuciones claramente separadas para variables de pétalos, validando la superioridad discriminatoria de estas características sobre las de sépalos.



### ¿Qué columnas son más útiles para diferenciar las especies?

**Jerarquía de Poder Discriminatorio:**

1. **Longitud del Pétalo (PetalLengthCm)** - Predictor más potente con separación casi perfecta entre especies. Rangos distintivos: *Setosa* (1.0-1.9 cm), *Versicolor* (3.0-5.1 cm), *Virginica* (4.5-6.9 cm), con overlaps mínimos únicamente entre versicolor y virginica.
2. **Ancho del Pétalo (PetalWidthCm)** - Excelente complemento discriminatorio con gaps visibles entre grupos. Rangos: *Setosa* (0.1-0.6 cm), *Versicolor* (1.0-1.8 cm), *Virginica* (1.5-2.5 cm).
3. **Longitud del Sépalo (SepalLengthCm)** - Utilidad moderada, permite distinguir *Setosa* de otras especies y *Virginica* de las restantes, pero con considerable solapamiento entre *Versicolor* y *Virginica*.
4. **Ancho del Sépalo (SepalWidthCm)** - Menor utilidad discriminatoria debido a alta variabilidad intraespecie y patrón inverso (*Setosa* con sépalos más anchos).

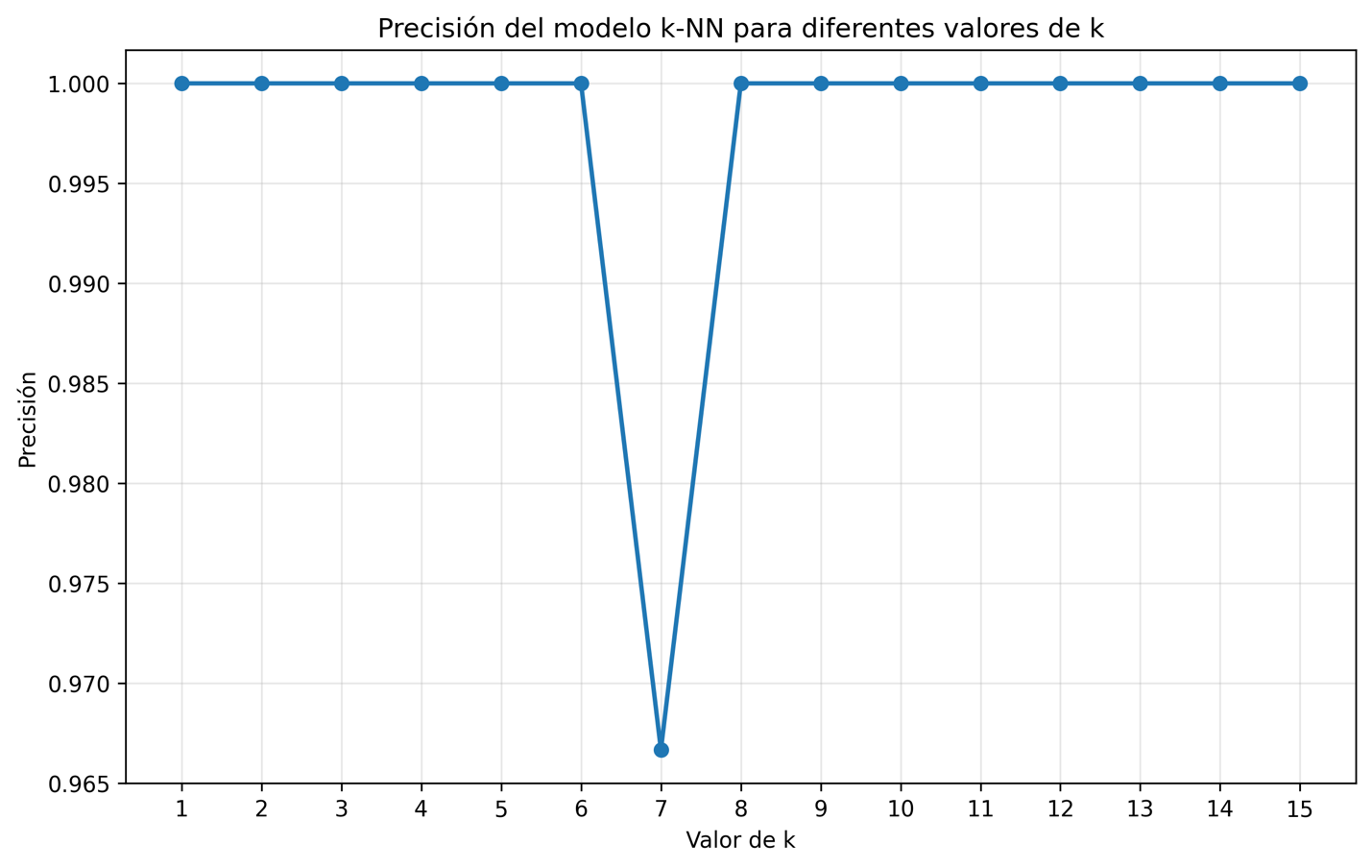
## Reflexión y Análisis Técnico

### ¿Qué utilidad tienen los gráficos para distinguir las especies?

Los gráficos de dispersión y el pairplot proporcionan valor fundamental para la comprensión de separabilidad entre especies. Permiten identificar clusters naturales y detectar regiones de solapamiento potencial, particularmente entre *Iris versicolor* e *Iris virginica* en variables de sépalo. Los gráficos de pétalos demuestran separación casi lineal, validando la idoneidad del algoritmo k-NN para este problema de clasificación.

La visualización facilita la selección de características óptimas y proporciona interpretabilidad al modelo, permitiendo entender por qué el algoritmo clasifica correctamente. La matriz de dispersión revela correlaciones entre variables y confirma que un modelo basado únicamente en características de pétalos podría ser suficiente para esta tarea de clasificación.

### ¿Cuál fue la precisión del modelo y cómo interpretarla?



**Precisión Obtenida: 100% (30/30 predicciones correctas)**

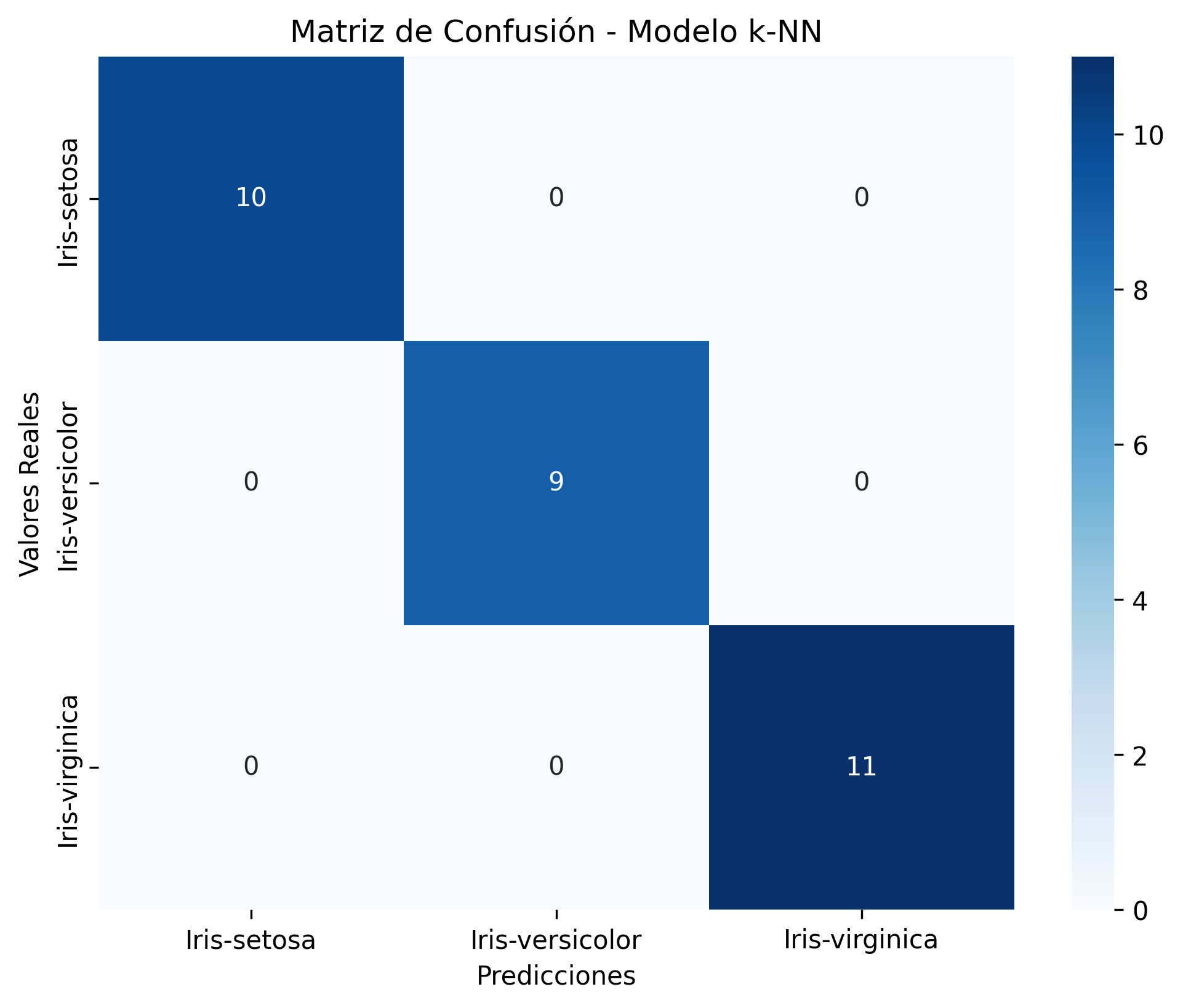
Desde una perspectiva estadística, este resultado debe interpretarse considerando el intervalo de confianza amplio inherente al tamaño limitado del conjunto de prueba. Con 30 muestras, el intervalo de confianza al 95% se extiende aproximadamente desde 88% hasta 100%.

Desde la perspectiva del aprendizaje automático, este resultado es esperado dado que IRIS es un dataset naturalmente muy separable. La experimentación con diferentes valores de k (k=1 a k=15) confirmó la robustez del resultado, con solo k=7 mostrando una ligera disminución a 96.7%, indicando que el resultado no es artificialmente perfecto.

En aplicaciones reales, esperaríamos precisiones en el rango 90-98% con muestras más grandes y mayor variabilidad ambiental. El resultado del 100% es creíble para este experimento específico pero debe interpretarse dentro del contexto de las limitaciones del dataset.

### ¿Qué información brinda la matriz de confusión sobre aciertos y errores?

**Análisis de la Matriz de Confusión:**



La matriz de confusión revela clasificación perfecta para todas las especies con ausencia completa de falsos positivos y falsos negativos. Cada especie muestra sensibilidad y especificidad del 100%: *Setosa* (10/10), *Versicolor* (9/9), y *Virginica* (11/11).

**Fortalezas Evidenciadas:** - El modelo no muestra sesgo hacia ninguna clase específica - Rendimiento consistente independiente del tamaño de muestra por clase - Discriminación efectiva incluso entre *Versicolor* y *Virginica*, que históricamente presentan mayor confusión

**Limitaciones del Análisis:** - La ausencia de errores impide identificar patrones específicos de confusión para mejoras futuras - El tamaño limitado (30 muestras) no captura completamente la variabilidad poblacional potencial - Puede generar expectativas no realistas para la aplicación a nuevos datos con mayor variabilidad

## Conclusiones

Este análisis demostró la efectividad del algoritmo k-NN para la clasificación de especies de iris, alcanzando precisión perfecta mediante la identificación de patrones morfológicos distintivos. Las características de pétalos emergieron como predictores superiores, proporcionando separación casi perfecta entre especies y validando su relevancia biológica en la diferenciación taxonómica.

Los resultados confirman la aplicabilidad de métodos de aprendizaje automático para problemas de clasificación botánica y proporcionan insights sobre la importancia relativa de diferentes características morfológicas en la identificación de especies.

## Referencias

*Ronggon, A.A., & Rahman, M.S. (2025). Performance Analysis and Noise Impact of a Novel Quantum KNN Algorithm for Machine Learning.*

*Park, G., Lee, Y.-G., Yoon, Y.-S., Ahn, J.-Y., Lee, J.-W., & Jang, Y.-P. (2022). Machine Learning-Based Species Classification Methods Using DART-TOF-MS Data for Five Coniferous Wood Species. Forests, 13(10), 1688.* [*https://doi.org/10.3390/f13101688*](https://doi.org/10.3390/f13101688)

*Nguyen, B., & Zhang, Y. (2025). A comparative study of diabetes prediction based on lifestyle factors using machine learning. doi:10.48550/ARXIV.2503.04137*

*Jumaah, M. A., Ali, Y. H., & Rashid, T. A. (2025). Artificial Liver Classifier: A new alternative to conventional machine learning models. doi:10.48550/ARXIV.2501.08074*