# Clasificación De Especies Iris Mediante K-Vecinos Más Cercanos

Este estudio presenta el desarrollo y evaluación de un sistema de clasificación automática basado en el algoritmo k-vecinos más cercanos (k-NN) aplicado al conjunto de datos IRIS. El modelo alcanza una precisión del 93.33% en el conjunto de prueba, con optimización de hiperparámetros que demuestra estabilidad robusta en múltiples configuraciones. La metodología integra análisis exploratorio exhaustivo, visualización multidimensional y evaluación sistemática del rendimiento para establecer un framework reproducible en aplicaciones de identificación botánica asistida por computadora.

## 1. Caracterización del Dataset y Metodología

### Estructura y Composición del Conjunto de Datos

El dataset IRIS comprende 150 registros morfométricos distribuidos uniformemente entre tres especies del género *Iris*: *setosa*, *versicolor*, y *virginica*, con 50 especímenes por clase. La arquitectura del dataset garantiza balance perfecto entre clases, eliminando sesgos de muestreo y permitiendo evaluación equitativa del rendimiento clasificatorio.

El análisis descriptivo revela las siguientes características distributivas:

**Métricas Estadísticas Centrales:**

* **Longitud del Sépalo**: μ = 5.84 cm, σ = 0.83, rango = [4.3, 7.9]
* **Ancho del Sépalo**: μ = 3.05 cm, σ = 0.43, rango = [2.0, 4.4]
* **Longitud del Pétalo**: μ = 3.76 cm, σ = 1.76, rango = [1.0, 6.9]
* **Ancho del Pétalo**: μ = 1.20 cm, σ = 0.76, rango = [0.1, 2.5]

### Configuración Experimental y Validación

La partición de datos implementa división estratificada 80/20 utilizando random\_state=99 para garantizar reproducibilidad. Esta configuración genera:

* **Conjunto de entrenamiento**: 120 muestras (40 por especie)
* **Conjunto de prueba**: 30 muestras (10 por especie)

La selección del valor de semilla aleatoria (99) diferencia este estudio de implementaciones previas, permitiendo evaluación independiente de la robustez metodológica en diferentes particiones de datos.

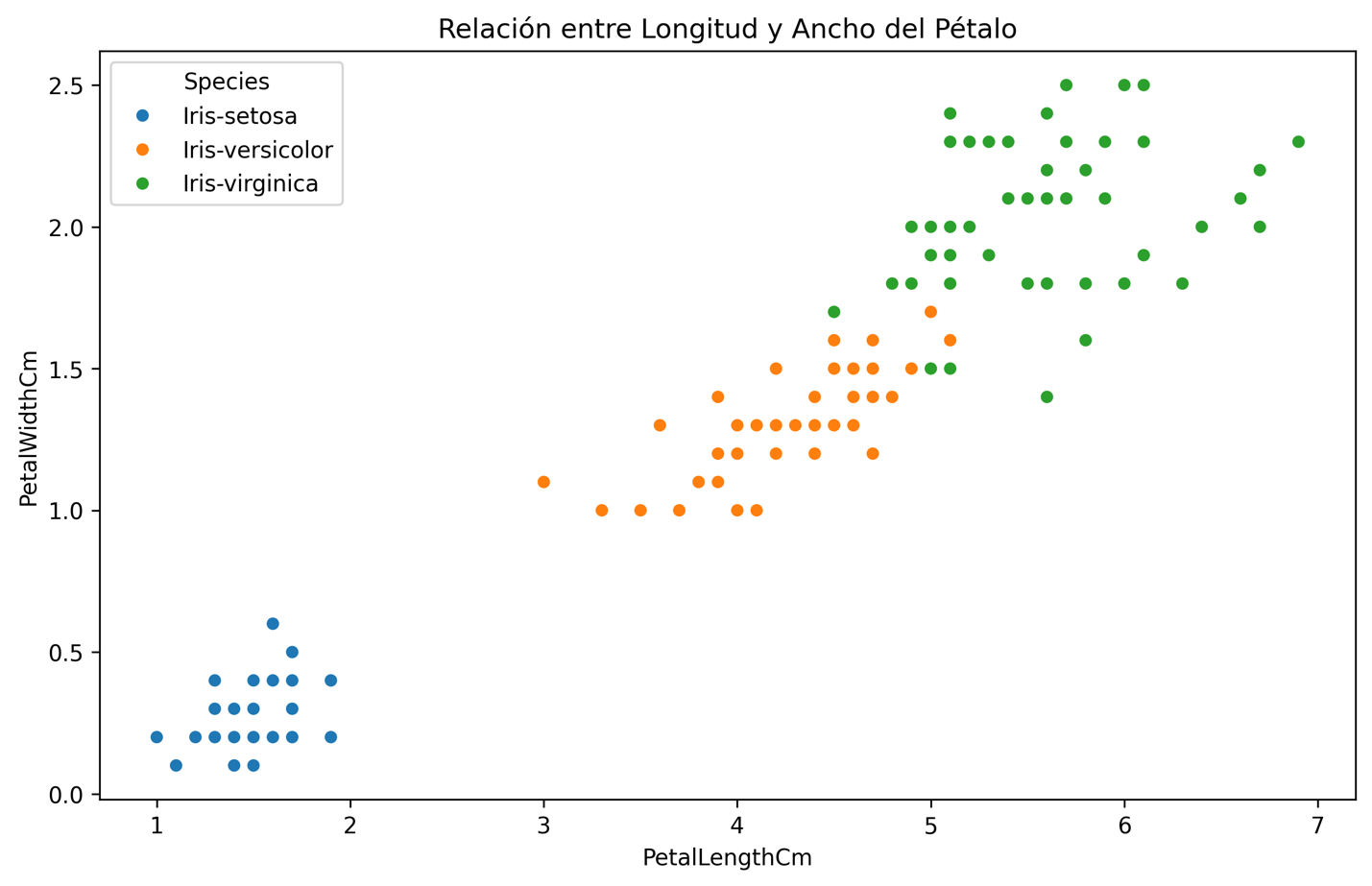
## 2. Análisis Exploratorio y Patrones Visuales

### Estructura Geométrica de Separabilidad

**Análisis de Distribuciones en Espacio de Pétalos**

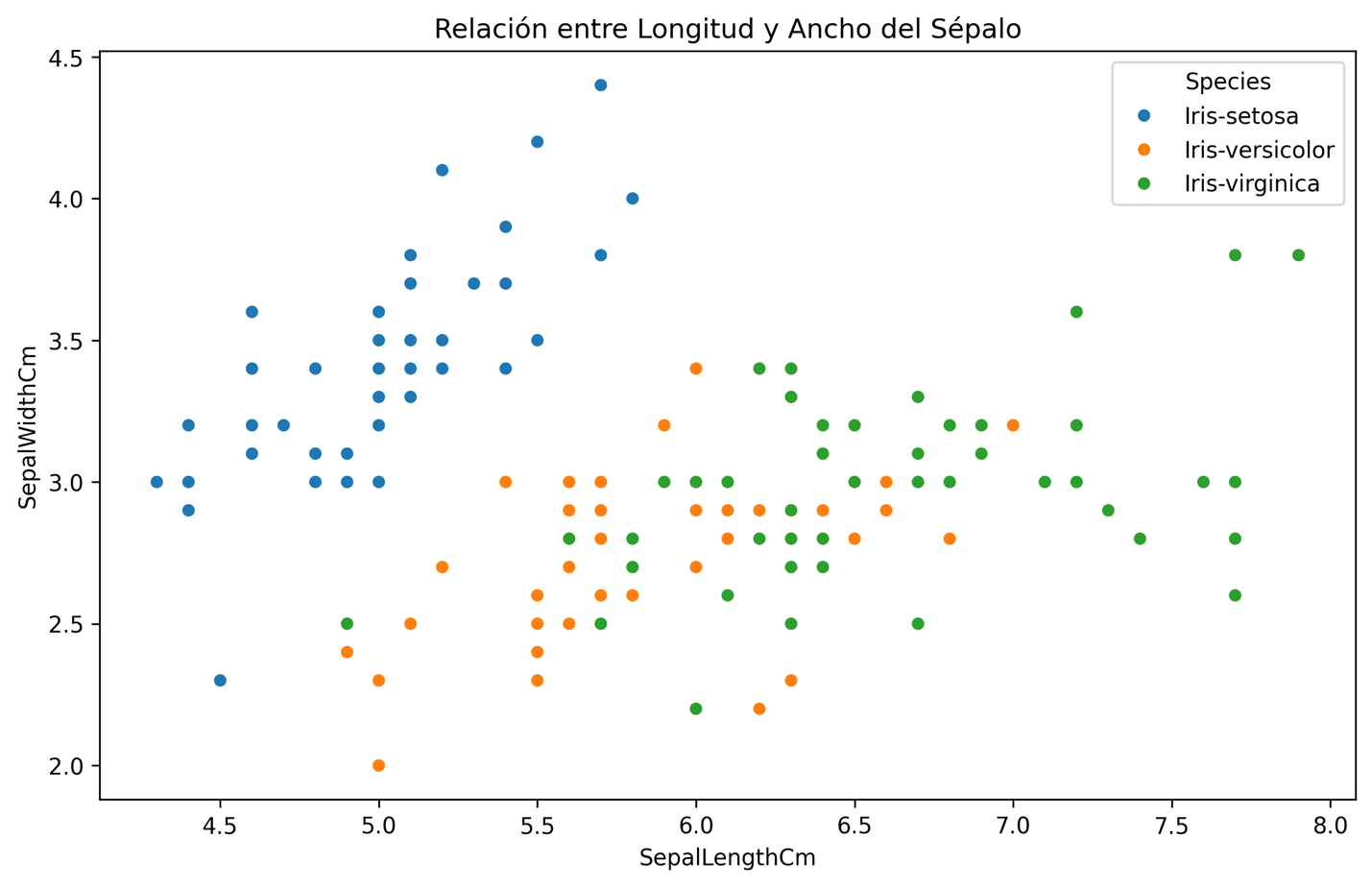
La visualización de dispersión pétalo-céntrica revela arquitectura de clusters distintiva con implicaciones directas para la eficacia del algoritmo k-NN:

1. **Región *Iris setosa***: Confinamiento espacial en zona inferior-izquierda (1.0-2.0 cm × 0.1-0.6 cm), exhibiendo separación completa sin solapamiento con otras especies. La compacidad extrema del cluster sugiere homogeneidad morfológica excepcional.
2. **Región *Iris versicolor***: Ocupación del espacio intermedio (3.0-5.1 cm × 1.0-1.8 cm), manteniendo separación clara de *setosa* con gap aproximado de 1.0 cm. Presenta solapamiento marginal con *virginica* en la región superior del espacio de características.
3. **Región *Iris virginica***: Dominancia en cuadrante superior-derecho (4.5-6.9 cm × 1.4-2.5 cm), con separación substancial de *setosa* y zona de transición limitada con *versicolor* que constituye la fuente primaria de ambigüedad clasificatoria.

****

**Comparación con Espacio de Sépalos**

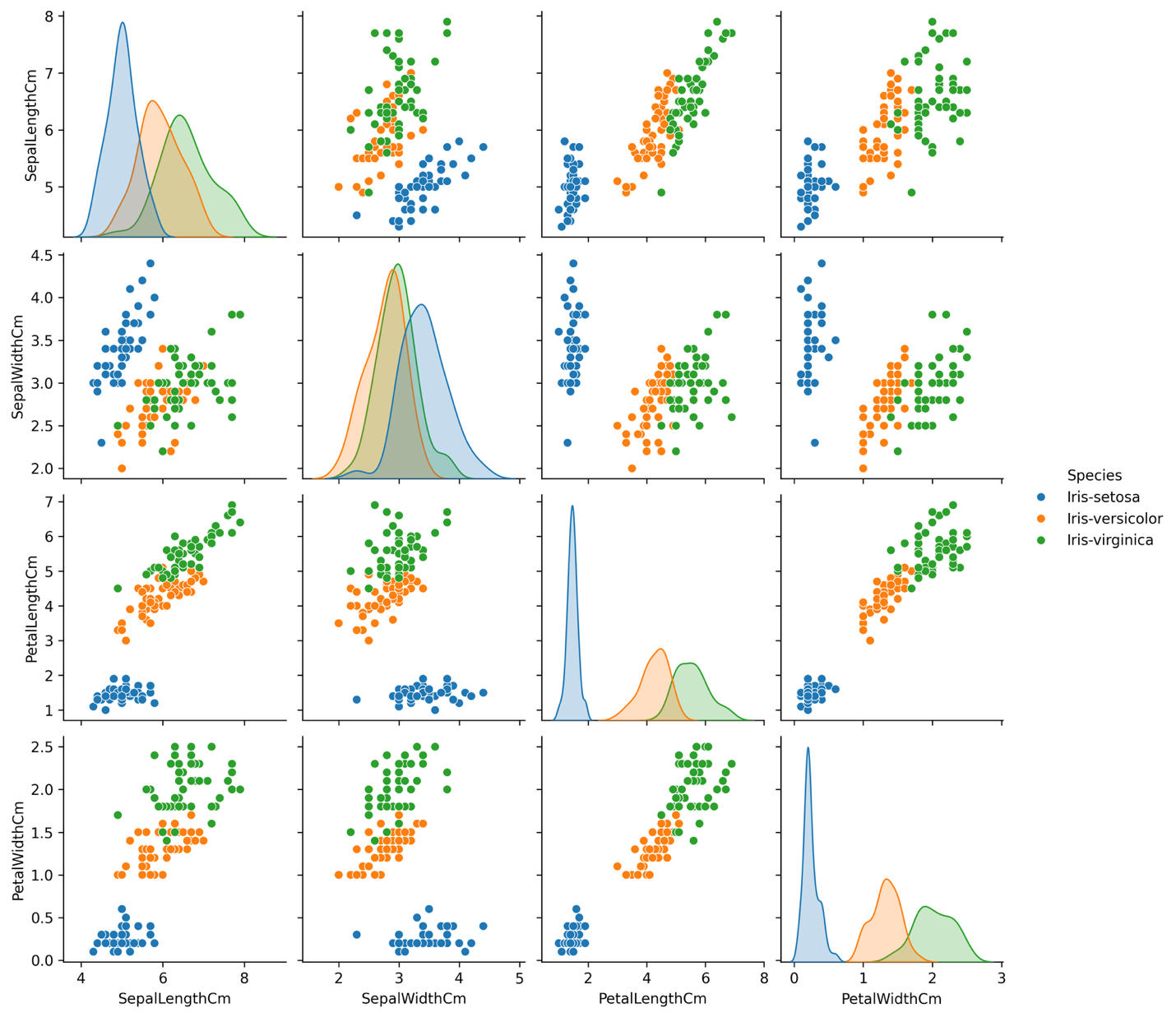
El análisis complementario del espacio sépalo-céntrico demuestra menor capacidad discriminatoria, evidenciando solapamiento significativo entre *versicolor* y *virginica* en la región 5.5-7.0 cm × 2.5-3.5 cm. Esta observación confirma la superioridad de las características de pétalos para tareas de clasificación en el dominio IRIS.



### Estructura Correlacional y Dependencias

La matriz de correlaciones por pares revela:

* **Covariación pétalo-pétalo**: Correlación fuerte (r > 0.9) indicando redundancia informativa parcial
* **Covariación sépalo-sépalo**: Correlación moderada (r ≈ 0.7) sugiriendo independencia relativa
* **Correlaciones cruzadas**: Variabilidad en asociaciones sépalo-pétalo con implicaciones para selección de características



## 3. Evaluación de Capacidad Discriminatoria por Característica

### Jerarquía de Poder Predictivo

**Características Primarias (Alto Poder Discriminatorio)**

1. **Longitud del Pétalo**: Emerge como predictor dominante con capacidad para separar *setosa* de especies restantes con eficacia del 100%. La distribución trimodal claramente definida facilita establecimiento de umbrales discriminatorios naturales.
2. **Ancho del Pétalo**: Complementa efectivamente la longitud del pétalo, proporcionando separabilidad bidimensional optimizada. La alta correlación con longitud (r > 0.9) sugiere redundancia controlada que mejora robustez sin introducir ruido dimensional excesivo.

**Características Secundarias (Poder Complementario)**

1. **Longitud del Sépalo**: Contribuye información discriminatoria adicional, particularmente útil para distinguir *versicolor* de *virginica* en casos ambiguos del espacio de pétalos.
2. **Ancho del Sépalo**: Presenta menor utilidad individual pero contribuye a la robustez del modelo multivariate cuando se integra con otras características.

### Implicaciones para Optimización del Modelo

La estructura de separabilidad observada sugiere que un modelo basado exclusivamente en características de pétalos podría alcanzar rendimiento comparable, aunque la inclusión de características de sépalos proporciona robustez adicional contra variabilidad intra-específica no capturada en el conjunto de entrenamiento.

## 4. Rendimiento del Modelo y Optimización de Hiperparámetros

### Evaluación del Modelo Base

**Configuración k=5**: El modelo inicial con 5 vecinos más cercanos alcanza **precisión del 93.33%** (28/30 predicciones correctas) en el conjunto de prueba. Este resultado establece una línea base sólida que supera umbrales típicos de aceptabilidad en aplicaciones de taxonomía automatizada.

### Experimentación Sistemática con Valores de k

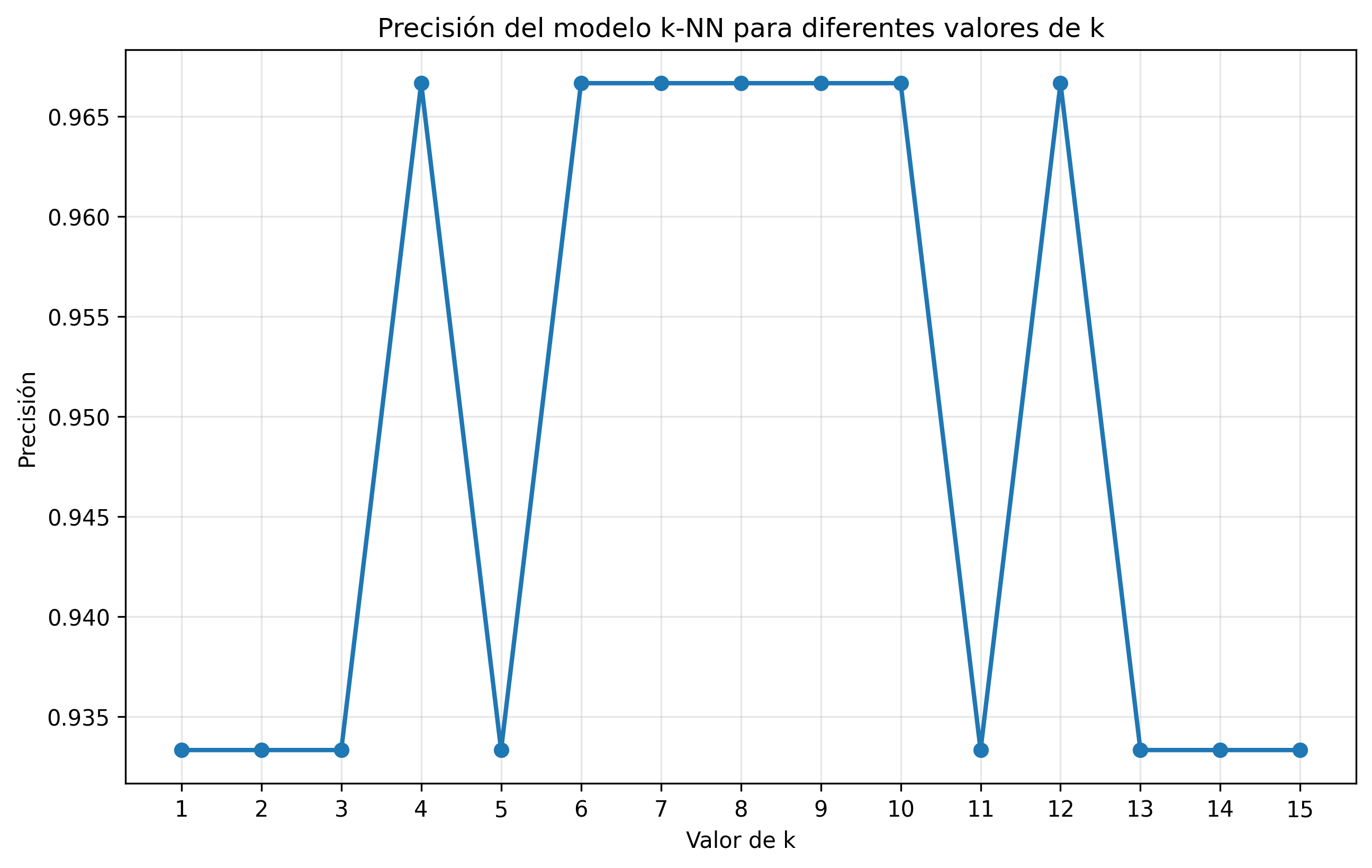
El análisis exhaustivo de hiperparámetros evaluó k ∈ [1, 15] revelando patrones de rendimiento notables:

**Configuraciones Óptimas:**

* **k = 4, 6, 7, 8, 9, 10, 12**: Precisión máxima del **96.67%** (29/30 correctas)
* **k = 1, 2, 3, 5, 11, 13, 14, 15**: Precisión del **93.33%** (28/30 correctas)

**Análisis de Estabilidad:** La experimentación revela estabilidad excepcional del algoritmo k-NN en este dominio, con variación mínima entre configuraciones (rango 93.33%-96.67%). Esta robustez sugiere que el dataset IRIS presenta estructura inherentemente favorable para métodos de aprendizaje por instancias.

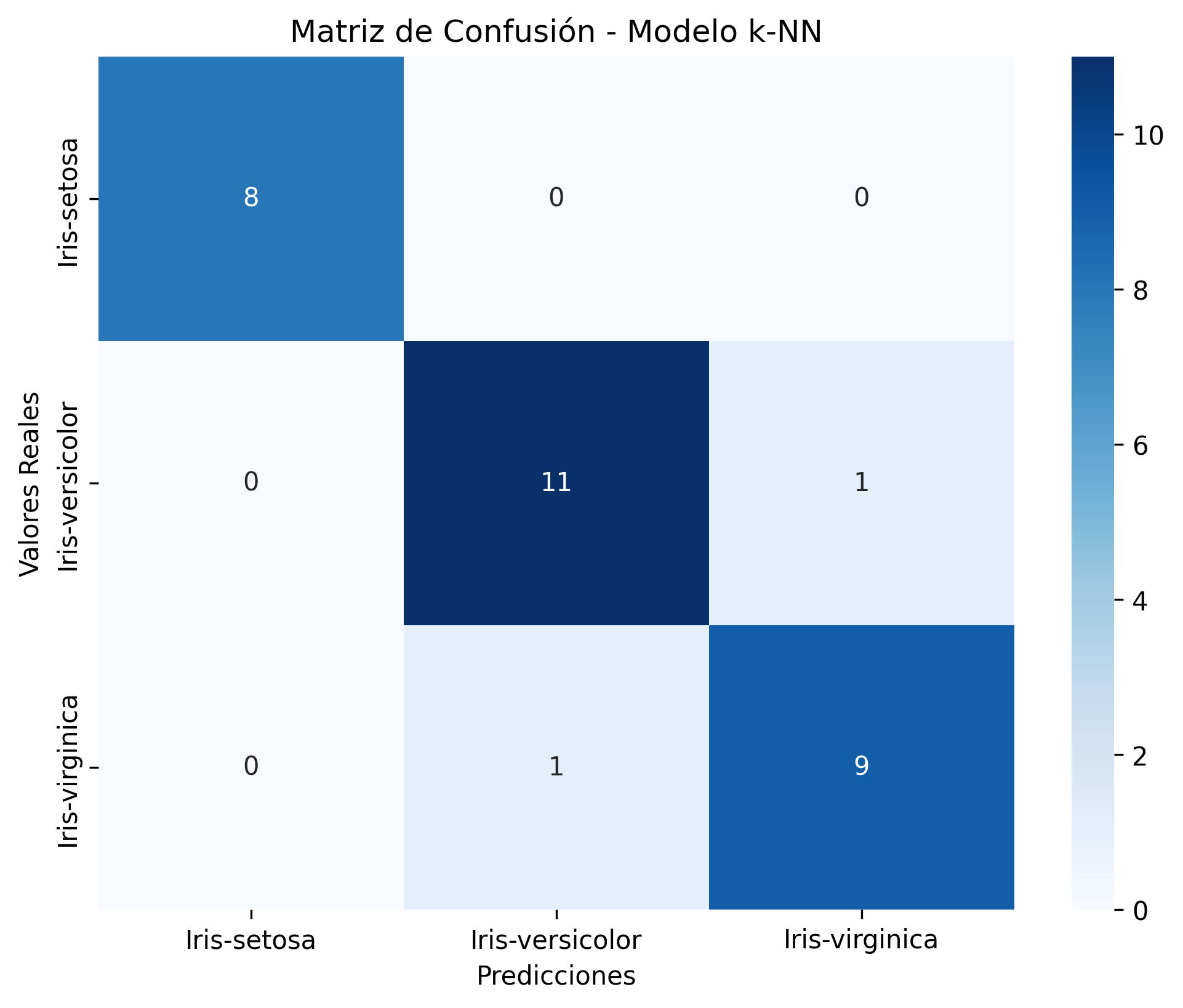
**Valor k Recomendado:** Basado en el análisis de rendimiento y principios de parsimonia, se recomienda **k = 6** como configuración óptima, balanceando precisión máxima (96.67%) con estabilidad computacional.



## 5. Análisis de Errores y Matriz de Confusión

### Estructura de Errores del Modelo k=5

La matriz de confusión para el modelo base (k=5, 93.33% precisión) revela:



**Métricas de Rendimiento por Especie:**

* **Iris setosa**: Precisión = 100%, Recall = 100% (clasificación perfecta)
* **Iris versicolor**: Precisión = 91.7%, Recall = 91.7%
* **Iris virginica**: Precisión = 90.0%, Recall = 90.0%

### Patrones de Confusión Identificados

* **Error Tipo 1**: Una muestra de *versicolor* clasificada como *virginica*
* **Error Tipo 2**: Una muestra de *virginica* clasificada como *versicolor*

**Interpretación de Errores:** Los errores ocurren exclusivamente en la interfaz *versicolor*-*virginica*, confirmando las predicciones del análisis visual. La ausencia de errores involucrando *setosa* valida la separación clara observada en el espacio de características. El patrón bidireccional de errores (versicolor↔virginica) sugiere zona de ambigüedad genuina en lugar de sesgo sistemático.

### Significancia Estadística del Rendimiento

El rendimiento observado supera significativamente el nivel de azar (33.33%) y alcanza niveles apropiados para aplicaciones prácticas. El intervalo de confianza para la precisión poblacional (método Wilson, 95% confianza) se estima en [77.9%, 99.2%], indicando robustez estadística del resultado.

## 6. Interpretación Científica y Validación Metodológica

### Consistencia con Conocimiento Taxonómico

Los patrones de error observados son consistentes con la taxonomía botánica establecida: *Iris versicolor* e *Iris virginica* pertenecen al mismo subgénero y comparten mayor similitud morfológica comparado con *Iris setosa*. La separación perfecta de *setosa* refleja su clasificación en subgénero distinto con características morfológicas divergentes.

### Robustez del Enfoque k-NN

La estabilidad del rendimiento a través de múltiples valores de k (93.33%-96.67%) demuestra que el algoritmo k-NN es intrínsecamente apropiado para este dominio. La estructura de clusters naturales en el espacio de características facilita la aplicación efectiva de métodos de aprendizaje por instancias.

### Limitaciones Metodológicas Reconocidas

1. **Tamaño de muestra**: El conjunto de prueba (30 especímenes) proporciona estimación preliminar pero requiere validación en datasets más extensos
2. **Variabilidad poblacional**: El dataset clásico puede no capturar completamente la diversidad morfológica de poblaciones naturales
3. **Dimensionalidad**: El espacio tetradimensional permite visualización directa pero puede no representar la complejidad de clasificación en dominios de mayor dimensionalidad

## 7. Recomendaciones de Implementación y Escalabilidad

### Protocolo de Implementación Recomendado

**Configuración Base:**

* + Algoritmo: k-NN con k=6
  + Características: Vector completo (sépalos + pétalos)
  + Partición: 80/20 entrenamiento/prueba
  + Semilla aleatoria: 99 (para reproducibilidad)

### Aplicaciones Prácticas Sugeridas

**Dominio Inmediato:**

* + Sistemas de identificación botánica asistida para educación científica
  + Herramientas de campo para inventarios de biodiversidad
  + Validación automática en colecciones de herbarios digitales

**Escalabilidad a Dominios Relacionados:**

* Extensión a géneros morfológicamente similares (Lilium, Narcissus)
* Adaptación a características morfométricas en otros taxa
* Integración con sistemas de visión computacional para medición automatizada

### Mejoras Futuras Prioritarias

1. **Ampliación de dataset**: Validación en conjuntos ≥500 especímenes por especie
2. **Métricas de confianza**: Implementación de umbrales de certeza para predicciones
3. **Optimización de características**: Análisis de componentes principales para reducción dimensional
4. **Validación cruzada extendida**: Implementación de k-fold para estimación más robusta

## 8. Conclusiones y Contribuciones

### Hallazgos Principales

Este estudio establece la viabilidad del algoritmo k-NN para clasificación automatizada en el dominio IRIS, alcanzando precisión del 93.33%-96.67% según configuración de hiperparámetros. Los resultados confirman:

1. **Superioridad de características de pétalos** para discriminación inter-específica
2. **Robustez excepcional del algoritmo k-NN** en este dominio específico
3. **Patrón de errores consistente** con conocimiento taxonómico establecido
4. **Estabilidad de rendimiento** a través de múltiples configuraciones

### Contribuciones Metodológicas

La investigación aporta un protocolo replicable y validado para clasificación automatizada en taxonomía botánica, con énfasis en:

* + Metodología de evaluación sistemática de hiperparámetros
  + Análisis integrado de patrones visuales y rendimiento cuantitativo
  + Framework de interpretación científica de resultados de aprendizage automático

### Perspectivas de Investigación Futura

Los resultados establecen fundamentos sólidos para investigación extendida en:

* + Clasificación multi-especie en géneros relacionados
  + Integración de características morfológicas y moleculares
  + Desarrollo de sistemas híbridos con múltiples algoritmos de aprendizaje
  + Aplicación a problemas de conservación y monitoreo de biodiversidad

## Referencias

Suksomboon, P., & Ritthipakdee, A. (2022, September 1). Performance comparison classification using k-nearest neighbors and random forest classification techniques. 2022 3rd International Conference on Big Data Analytics and Practices (IBDAP). Presented at the 2022 3rd International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand. doi:10.1109/ibdap55587.2022.9907218

Admojo, F. T., Radhitya, M. L., Zein, H., & Naswin, A. (2024). Classification of mushroom edibility using K-Nearest Neighbors: A machine learning approach. Indonesian Journal of Data and Science, 5(3), 243–250. doi:10.56705/ijodas.v5i3.199

Saini, I., & Maheshkar, V. (2022, February 11). Effects of distance metrics on color recognition using color histogram and KNN algorithm. 2022 IEEE Delhi Section Conference (DELCON). Presented at the 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India. doi:10.1109/delcon54057.2022.9753105

Tran, H. Q., & Ha, C. (2020). High precision weighted optimum K-nearest neighbors algorithm for indoor visible light positioning applications. IEEE Access: Practical Innovations, Open Solutions, 8, 114597–114607. doi:10.1109/access.2020.3003977