¿Cuál es el fundamento teórico detrás del compromiso entre sesgo y varianza al seleccionar el valor de K en validación cruzada para conjuntos de datos con $N < 2000$?""En base a la literatura actual, ¿por qué un valor de $K=10$ reduce el sesgo de la estimación pero tiende a aumentar su varianza en muestras pequeñas en comparación con valores de K más bajos?""¿Qué evidencia empírica existe en artículos científicos que compare el error de generalización en modelos entrenados con muestras pequeñas utilizando $K=5$ frente a $K=10$?""¿Cómo afecta el sobreajuste (overfitting) a la estimación del rendimiento cuando el tamaño de la muestra es limitado y se emplean configuraciones de K elevadas?"

Sesgo–varianza en elegir K en validación cruzada con muestras pequeñas

Fundamento teórico del compromiso sesgo–varianza

En k-fold CV se estima el error promedio entrenando en ((K-1)/K) de los datos y validando en (1/K).

- **Sesgo del estimador de error**: al entrenar con menos datos que el modelo final (entrenado en todo el conjunto), el error de CV tiende a **sobreestimar** el error real (sesgo al alza); este sesgo disminuye al aumentar (K), pues el tamaño del conjunto de entrenamiento se acerca más a (N) (Fushiki, 2011; Górriz et al., 2024).
- **Varianza del estimador**: al aumentar (K), el fold de test se hace más pequeño, lo que hace cada estimación de error más ruidosa y sensible a qué observaciones caen en test; además los conjuntos de entrenamiento se solapan fuertemente, por lo que los errores entre folds están correlacionados, limitando la reducción de varianza por promediado (Rodríguez et al., 2010; Nti et al., 2021).

Con (N<2000), estos efectos se intensifican: cada fold es pequeño y la variabilidad por partición domina (Vabalas et al., 2019; Varoquaux, 2017).

Por qué (K=10) ↓ sesgo pero ↑ varianza (vs. K más bajo)

Estudios clásicos y recientes muestran:

- Incrementar (K) (hasta LOOCV) **reduce sesgo** del error de CV porque los modelos se entrenan con más datos (Fushiki, 2011; Nti et al., 2021).
- Pero la **varianza del error** puede crecer o no disminuir de forma monotónica, sobre todo cuando los conjuntos de test son pequeños (Rodríguez et al., 2010; Nti et al., 2021; Moss et al., 2018).

Simulaciones teóricas y empíricas en clasificación y regresión indican que 5 y 10 folds suelen dar errores de generalización similares, con 10-fold más cercano al error “verdadero” pero con error estándar algo mayor en muestras pequeñas (Rodríguez et al., 2010; Nti et al., 2021; Jiang & Wang, 2017). En regresión lineal/LASSO con datos normales, un estudio reciente encuentra que tanto sesgo como varianza del estimador de error decrecen con (K), pero el óptimo de (K) depende de (n) y del modelo, y puede ser >10 para (n\ge 100) (Vasilopoulos & Matthews, 2024), lo que cuestiona la regla simple “10 = más varianza”. En la práctica aplicada (neurociencia, ML), sin embargo, se observa empíricamente que con (N) pequeños la dispersión entre repeticiones de 10-fold suele ser mayor que con 5-fold, precisamente por el tamaño de los folds de test (Vabalas et al., 2019; Rodríguez et al., 2010; Varoquaux, 2017; Moss et al., 2018).

Comparaciones empíricas K=5 vs K=10

Estudios que comparan múltiples (K) en muchos datasets y algoritmos encuentran:

Estudio	Modelos/datos	Hallazgos clave sobre K=5 vs K=10	Citas
Rodríguez et al. 2010	Naive Bayes, kNN; datos artificiales	Analizan descomposición de varianza: relación entre K, tamaño muestral y varianza; muestran que el efecto de K es complejo y depende del clasificador y n (Rodríguez et al., 2010).	(Rodríguez et al., 2010)
Jiang & Wang 2017	5 algoritmos, 20 datasets	Derivan relación cuantitativa entre varianza de CV, n, K y repeticiones; muestran cómo K y repeticiones regulan varianza y proponen estrategias para minimizarla (Jiang & Wang, 2017).	(Jiang & Wang, 2017)
Nti et al. 2021	GBM, LR, DT, KNN; múltiples datasets	Empíricamente: K=7 suele equilibrar mejor precisión y coste; 5 y 10 dan rendimientos comparables; describen explícitamente el patrón “K bajo: +sesgo, –varianza; K alto: –sesgo, +varianza” (Nti et al., 2021).	(Nti et al., 2021)
Marcot & Hanea 2020	Redes Bayesianas discretas	Error de clasificación decrece con K y se estabiliza cerca de K=10; para n grande, K=5 suele bastar (Marcot & Hanea, 2020).	(Marcot & Hanea, 2020)

En conjuntos pequeños, varios trabajos en neuroimagen y ML muestran que k-fold CV (K=5–10) tiende a producir **estimaciones demasiado optimistas y muy variables** en comparación con nested CV o train/test fijo (Vabalas et al., 2019; Varoquaux, 2017; Li, 2023), lo que refleja la alta varianza y sobreajuste de la métrica de CV más que del modelo en sí.

Efecto del sobreajuste con K altos y N limitado

Con (N) pequeño y (K) grande:

- Cada modelo se entrena en ((K-1)/K N) (p.ej., 90% con K=10), lo que facilita ajustar ruido idiosincrático.
- La métrica de CV se usa a menudo para **seleccionar hiperparámetros o modelos**; con alta varianza del estimador, es fácil “sobreajustar” la propia métrica de CV (seleccionar el modelo que maximizó un valor inflado por azar) (Li, 2023; Teodorescu & Brasoveanu, 2025).
- Simulaciones con datos puramente ruido muestran que k-fold CV puede producir accuracies por encima del azar, especialmente con N pequeño y búsqueda intensa de hiperparámetros (Vabalas et al., 2019; Li, 2023).

Estudios en selección de modelos demuestran que, en muestras finitas, se puede sobreajustar cualquier criterio de selección (incluida la media de CV), degradando el rendimiento out-of-sample una vez que se “tunea demasiado” contra un estimador ruidoso (Teodorescu & Brasoveanu, 2025). Recomendaciones habituales para N<~2000 incluyen:

- Usar **nested CV** para separar evaluación y tuning (Vabalas et al., 2019; Li, 2023; Górriz et al., 2024; Teodorescu & Brasoveanu, 2025).
- Considerar **K moderados (5–10) y, si es posible, repetir CV varias veces (J-K-fold)** para reducir la varianza del estimador en lugar de aumentar K sin límite (Jiang & Wang, 2017; Moss et al., 2018).

Resumen práctico para N<2000

- Teóricamente, aumentar K **reduce el sesgo** del error de CV porque el entrenamiento se aproxima más al uso de todo el conjunto (Fushiki, 2011; Nti et al., 2021; Górriz et al., 2024).
- El efecto de K sobre la **varianza** es más complejo; en aplicaciones con N pequeño, K altos suelen implicar **errores de test más ruidosos y mayor variabilidad entre particiones**, lo que se percibe empíricamente como "más varianza" (Vabalas et al., 2019; Rodríguez et al., 2010; Varoquaux, 2017; Nti et al., 2021; Moss et al., 2018).
- Comparaciones empíricas muestran que **K=5 y K=10 suelen producir errores de generalización muy similares**, con diferencias menores que las debidas al algoritmo o al propio muestreo; K≈7–10 suele ser un compromiso razonable, pero no universal (Jiang & Wang, 2017; Marcot & Hanea, 2020; Nti et al., 2021).
- Con muestras pequeñas y tuning intensivo, el **sobreajuste a la métrica de CV** se vuelve crítico; nested CV y/o repeticiones de CV son más importantes que elegir entre 5 y 10 folds (Vabalas et al., 2019; Li, 2023; Moss et al., 2018; Teodorescu & Brasoveanu, 2025).

These papers were sourced and synthesized using Consensus, an AI-powered search engine for research. Try it at https://consensus.app

References

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE*, 14. https://doi.org/10.1371/journal.pone.0224365

Jiang, G., & Wang, W. (2017). Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recognit.*, 69, 94-106. https://doi.org/10.1016/j.patcog.2017.03.025

Marcot, B., & Hanea, A. (2020). What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?. *Computational Statistics*, 36, 2009 - 2031. https://doi.org/10.1007/s00180-020-00999-9

Vasilopoulos, A., & Matthews, G. (2024). Cross-validation Optimal Fold-Number for Model Selection. *American Journal of Undergraduate Research*. https://doi.org/10.33697/ajur.2024.123

Rodríguez, J., Martínez, A., & Lozano, J. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 569-575. https://doi.org/10.1109/tpami.2009.187

Li, J. (2023). Asymptotics of K-Fold Cross Validation. *J. Artif. Intell. Res.*, 78, 491-526. https://doi.org/10.1613/jair.1.13974

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21, 137-146. https://doi.org/10.1007/s11222-009-9153-8

Varoquaux, G. (2017). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, 180, 68-77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Nti, I., Yarko-Boateng, O., & Aning, J. (2021). Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science*. https://doi.org/10.5815/ijitcs.2021.06.05

Moss, H., Leslie, D., & Rayson, P. (2018). Using J-K-fold Cross Validation To Reduce Variance When Tuning NLP Models. **, 2978-2989.

Górriz, J., Segovia, F., Ramírez, J., Ortíz, A., & Suckling, J. (2024). Is K-fold cross validation the best model selection method for Machine Learning?. *ArXiv*, abs/2401.16407. https://doi.org/10.48550/arxiv.2401.16407

Teodorescu, V., & Brasoveanu, L. (2025). Assessing the Validity of k-Fold Cross-Validation for Model Selection: Evidence from Bankruptcy Prediction Using Random Forest and XGBoost. *Comput.*, 13, 127. https://doi.org/10.3390/computation13050127