"¿Cómo influye la estratificación en la reducción de la varianza del error estimado cuando la variable objetivo presenta una distribución altamente sesgada (ej. exceso de ceros o clases minoritarias)?"

"¿Qué diferencias metodológicas se reportan al aplicar Stratified K-Fold frente a un K-Fold convencional en estudios de predicción de supervivencia o series temporales cortas?"

"Analiza la importancia de mantener la proporción de la distribución original en cada partición (fold) para evitar artefactos en la medición del rendimiento en datasets desbalanceados."

"¿Bajo qué condiciones específicas de distribución de datos la literatura científica desaconseja el uso de validación cruzada sin estratificación?"

La estratificación reduce varianza y sesgos del error estimado cuando la distribución es muy sesgada o dependiente, pero puede ser contraproducente si rompe la estructura real de los datos.

1. Estratificación y varianza del error con distribuciones sesgadas

En K-Fold aleatorio, los folds pueden tener proporciones de clases o rangos de y muy diferentes del conjunto completo, generando **dataset/target shift** y alta varianza del error estimado, especialmente con clases minoritarias o exceso de ceros (López et al., 2014; Sáez & Romero-Béjar, 2022; Yücelbaş & Yücelbaş, 2023).

- En clasificación desbalanceada, la estratificación por clase reduce la probabilidad de folds sin minoría y hace que el entrenamiento y test sean más representativos, disminuyendo sesgo y desviación estándar del error (Szeghalmy & Fazekas, 2023; López et al., 2014; Widodo et al., 2022; Yücelbaş & Yücelbaş, 2023; Pan et al., 2020).
- En regresión con distribución muy asimétrica, la estratificación del regresando (dividir y en estratos) produce folds con distribuciones de salida más similares y mejora **bias y desviación** de la estimación de RMSE/MAE respecto a CV no estratificado (Sáez & Romero-Béjar, 2022).

Importancia de mantener la proporción original

Mantener la proporción de la distribución original en cada fold evita artefactos como:

- Medidas AUC/F1 artificialmente optimistas o pesimistas cuando algún fold casi no tiene positivos (Szeghalmy & Fazekas, 2023; López et al., 2014; Parker et al., 2007; Widodo et al., 2022).
- Variaciones grandes entre folds, que obligan a repetir muchas veces el CV para estabilizar la estimación (Szeghalmy & Fazekas, 2023; Sáez & Romero-Béjar, 2022; Yücelbaş & Yücelbaş, 2023; Spezia & Mendoza, 2025).

2. Stratified K-Fold vs K-Fold convencional en supervivencia y series cortas

En predicción de supervivencia con eventos raros y n pequeño, se recomienda **K-Fold (a menudo estratificado por evento/censura)** para obtener estimaciones más estables que train-test o bootstrap, especialmente en alta dimensión (Dubray-Vautrin et al., 2025). Estratificar por estado del evento (y a veces por centros/hospital) mantiene la tasa de eventos en cada fold y mejora la estabilidad de C-index y Brier Score frente a particiones aleatorias (Parker et al., 2007; Bahrami et al., 2024; Dubray-Vautrin et al., 2025).

En series temporales cortas, la estratificación clásica por clase se considera inadecuada: la literatura de datos temporales/espaciales enfatiza que **lo prioritario es respetar la estructura temporal o espacial**, usando "block" o "temporal" CV; el K-Fold aleatorio (estratificado o no) subestima groseramente el error de predicción futura (Roberts et al., 2017; Beigaitė et al., 2022).

3. ¿Cuándo se desaconseja CV sin estratificación?

Se desaconseja K-Fold aleatorio sin estratos cuando:

- Hay **fuerte desbalance de clases** o muy pocos ejemplos de la clase minoritaria: pueden aparecer folds sin esa clase, rompiendo el entrenamiento y sesgando AUC/F1/recall (Szeghalmy & Fazekas, 2023; López et al., 2014; Widodo et al., 2022; Yücelbaş & Yücelbaş, 2023; Pan et al., 2020).
- La variable objetivo continua es muy sesgada o multimodal y se busca buena estimación de error medio: la estratificación del regresando reduce el "target shift" entre train/test (Sáez & Romero-Béjar, 2022).
- Con **muestras pequeñas y baja señal** (p.ej. microarrays, pronóstico clínico), incluso pequeñas desviaciones en proporciones por fold producen **sesgos importantes** en AUC y error; se recomiendan esquemas estratificados y balanceados en lugar de CV no estratificado o LOOCV estándar (Parker et al., 2007; Dubray-Vautrin et al., 2025).
- En datasets clínicos multi-centro, estratificar por fuente o bloques (hospital, región) o usar leave-source-out es preferible a CV aleatorio si el objetivo es generalizar a nuevos centros (Bahrami et al., 2024; Leinonen et al., 2024; Roberts et al., 2017).

En cambio, con grandes muestras y distribuciones relativamente equilibradas, la diferencia entre CV estratificado y no estratificado tiende a ser pequeña, sobre todo si se repite el K-Fold muchas veces (Szeghalmy & Fazekas, 2023; Sáez & Romero-Béjar, 2022; Yücelbaş & Yücelbaş, 2023).

Conclusión

En datasets desbalanceados, con excesos de ceros, eventos raros o baja señal, la validación sin estratificación introduce más varianza y sesgo en el error estimado. Mantener la proporción (por clases o por rangos de la variable objetivo) en cada fold es clave para evitar artefactos, siempre que no se rompan dependencias temporales/espaciales, donde se recomiendan estrategias de "block" o por fuente en lugar de estratificación clásica.

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

References

Szeghalmy, S., & Fazekas, A. (2023). A Comparative Study of the Use of Stratified Cross-Validation and Distribution-Balanced Stratified Cross-Validation in Imbalanced Learning. *Sensors (Basel, Switzerland)*, 23. https://doi.org/10.3390/s23042333

López, V., Fernández, A., & Herrera, F. (2014). On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed. *Inf. Sci.*, 257, 1-13. https://doi.org/10.1016/j.ins.2013.09.038

Sáez, J., & Romero-Béjar, J. (2022). Impact of Regressand Stratification in Dataset Shift Caused by Cross-Validation. *Mathematics*. https://doi.org/10.3390/math10142538

Parker, B., Günter, S., & Bedő, J. (2007). Stratification bias in low signal microarray studies. *BMC Bioinformatics*, 8, 326 - 326. https://doi.org/10.1186/1471-2105-8-326

Widodo, S., Brawijaya, H., & Samudi, S. (2022). Stratified K-fold cross validation optimization on machine learning for prediction. *Sinkron*. https://doi.org/10.33395/sinkron.v7i4.11792

Bahrami, P., Tanbakuchi, D., Afzalaghaee, M., Ghayour-Mobarhan, M., & Esmaily, H. (2024). Development of risk models for early detection and prediction of chronic kidney disease in clinical settings. *Scientific Reports*, 14. https://doi.org/10.1038/s41598-024-83973-5

Leinonen, T., Wong, D., Vasankari, A., Wahab, A., Nadarajah, R., Kaisti, M., & Airola, A. (2024). Empirical investigation of multi-source cross-validation in clinical ECG classification. *Computers in biology and medicine*, 183, 109271. https://doi.org/10.1016/j.compbiomed.2024.109271

Yücelbaş, C., & Yücelbaş, Ş. (2023). Enhanced Cross-Validation Methods Leveraging Clustering Techniques. *Traitement du Signal*. https://doi.org/10.18280/ts.400626

Roberts, D., Bahn, V., Ciuti, S., Boyce, M., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J., Schröder, B., Thuiller, W., Warton, D., Wintle, B., Hartig, F., & Dormann, C. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40, 913-929. https://doi.org/10.1111/ecog.02881

Dubray-Vautrin, A., Gravrand, V., Marret, G., Lamy, C., Klijanienko, J., Vacher, S., Ahmanache, L., Kamal, M., Choussy, O., Servant, N., Dupain, C., Tourneau, L., & Mullaert, J. (2025). Internal validation strategy for high dimensional prognosis model: A simulation study and application to transcriptomic in head and neck tumors. *Computational and Structural Biotechnology Journal*, 27, 3792 - 3802. https://doi.org/10.1016/j.csbj.2025.08.035

Spezia, A., & Mendoza, M. (2025). Comparing Cluster-Based Cross-Validation Strategies for Machine Learning Model Evaluation. *ArXiv*, abs/2507.22299. https://doi.org/10.48550/arxiv.2507.22299

Pan, T., Zhao, J., Wu, W., & Yang, J. (2020). Learning imbalanced datasets based on SMOTE and Gaussian distribution. *Inf. Sci.*, 512, 1214-1233. https://doi.org/10.1016/j.ins.2019.10.048

Beigaitė, R., Mechenich, M., & Žliobaitė, I. (2022). Spatial Cross-Validation for Globally Distributed Data. **, 127-140. https://doi.org/10.1007/978-3-031-18840-4_10