Consideraciones clave para el escalado de datos heterogéneos en Machine Learning

Al elegir técnicas de escalado para conjuntos de datos con variables heterogéneas, se deben considerar múltiples factores críticos que impactan significativamente el rendimiento del modelo de machine learning.

Distribución de los datos: El escalado robusto utiliza la mediana y el rango intercuartílico (IQR) en lugar de la media y desviación estándar, haciéndolo robusto a valores atípicos y distribuciones sesgadas Feature Engineering: Scaling, Normalization, and Standardization - GeeksforGeeks. Para distribuciones normales, la estandarización Z-score es preferible, mientras que para distribuciones asimétricas o con outliers, el escalado robusto resulta más adecuado.

Naturaleza de las variables: Es esencial para conjuntos de datos con características de rangos, unidades o magnitudes variables What is Feature Scaling and Why is it Important? - Analytics Vidhya. Variables categóricas requieren codificación previa, mientras que variables numéricas continuas y discretas necesitan estrategias diferenciadas de escalado.

Algoritmo de machine learning seleccionado: Aunque los modelos basados en árboles son (casi) no afectados por el escalado Importance of Feature Scaling — scikit-learn 1.7.1 documentation, algoritmos como SVM, redes neuronales y k-NN son sensibles a la escala de las características.

Presencia de valores atípicos: Los desafíos asociados con conjuntos de datos como datos faltantes, datos inconsistentes y datos mixtos son obstáculos frecuentes Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. El escalado Min-Max es vulnerable a outliers, mientras que el escalado robusto mantiene estabilidad.

Interpretabilidad del modelo: La elección entre normalización (0-1) y estandarización (media=0, std=1) afecta la interpretación de coeficientes y

importancia de características. Esta transformación mejora el rendimiento de modelos de clasificación, pero existen varias técnicas de escalado disponibles, y esta elección generalmente no se realiza cuidadosamente [The choice of scaling technique matters for classification performance - ScienceDirect].

Referencias bibliográficas:

- Singh, D., & Singh, B. (2021). Effect of data scaling methods on machine learning algorithms and model performance. Technologies, 9(3), 52.
- Pinheiro, J. M. H., et al. (2025). The impact of feature scaling in machine learning: Effects on regression and classification tasks. arXiv preprint arXiv:2506.08274.
- Moreira-Matias, L., et al. (2023). The choice of scaling technique matters for classification performance. Applied Soft Computing, 133, 109924.
- Yang, Z., et al. (2022). Heterogeneous feature fusion based machine learning on shallow-wide and heterogeneous-sparse industrial datasets. Conference proceedings.
- Kumar, S., et al. (2022). Impact of feature scaling on machine learning models for the diagnosis of diabetes. IEEE Conference Publication.