TÉCNICAS DE ESCALADO

Nombre: Dina Susan Calcina Aquino

Factores a Tener en Cuenta para Elegir una Técnica de Escalado en Datos Heterogéneos

Uno de los factores es la naturaleza de las variables. Los datos pueden contener una mezcla de tipos, como variables numéricas y categóricas. Las técnicas de escalado tradicionales, como la normalización Min-Max o la estandarización Z-score, son aplicables principalmente a variables numéricas. Para las categóricas, es esencial una codificación previa (por ejemplo, One-Hot Encoding) antes de cualquier transformación conjunta [1].

La distribución de los datos es otro factor. La estandarización Z-score, que transforma los datos para tener una media de cero y una desviación estándar de uno, asume una distribución aproximadamente normal y es sensible a los valores atípicos. En contraste, el escalado robusto (Robust Scaler), que utiliza la mediana y el rango intercuartílico, es más resistente a la presencia de outliers, lo que lo hace idóneo para distribuciones sesgadas o con valores extremos [2].

El objetivo del análisis también influye en la elección. Algunos algoritmos de aprendizaje automático, especialmente aquellos basados en distancias (como K-Means, SVM o PCA), son altamente sensibles a la escala de las características. Un escalado inapropiado puede llevar a que las variables con rangos más amplios dominen el proceso de aprendizaje, sesgando los resultados. Por lo tanto, el escalado busca homogenizar la influencia de cada característica en el modelo [3].

Finalmente, la prevención de la fuga de datos (data leakage) es una consideración metodológica vital. Los parámetros de escalado (media, desviación estándar, mínimos y máximos) deben calcularse exclusivamente sobre el conjunto de datos de entrenamiento y luego aplicarse para transformar tanto el conjunto de entrenamiento como el de prueba. Esto asegura que el modelo no tenga acceso a información del conjunto de prueba durante su fase de preparación, manteniendo la integridad de la evaluación [4].

Referencias Bibliográficas

Referencias

[1] Azevedo, K., Quaranta, L., Calefato, F., Castelluccio, M., & Catolino, G. (2024). A multivocal literature review on the benefits and limitations of automated machine learning tools. *arXiv preprint*. https://doi.org/10.48550/arXiv.2024.15342

[2] Ghazwani, M., & Hani, U. (2024). Data driven analysis of tablet design via machine learning for evaluation of impact of formulations properties on the disintegration time. *Ain Shams Engineering Journal*, *15*(8), 102847. https://doi.org/10.1016/j.asej.2024.102847

[3] Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*, 350-361. https://doi.org/10.1016/j.neucom.2016.12.013

[4] Flores-Martín, D., Lemus-Prieto, F., & Rico-Gallego, J. A. (2024). PULSE: A modular framework for predictive energy efficiency in heterogeneous data centers. *SoftwareX*, *27*, 101789. https://doi.org/10.1016/j.softx.2024.101789