Feature engineering (Segunda iteración)

Dataset de Registros de Ordeño

El objetivo de nuestra feature engineering:

- Reducir la asimetría (skewness) de variables de producción altamente positivas.
- Estabilizar la varianza de mediciones fisiológicas.
- Facilitar el uso posterior de modelos que asumen distribuciones más cercanas a la normalidad.

Se definen tres grupos de variables:

1. Variables de producción (production_vars)

   Incluyen la producción total por ordeño y las producciones por cuarto de ubre/

   Para cada una de estas columnas, se genera una nueva variable logarítmica usando np.log1p (log(1+x)), lo cual permite manejar valores cercanos a cero sin problemas numéricos.

2. Variables de sangre (blood_vars)

   Representan indicadores de presencia de sangre en la leche por cuarto de ubre.

   Estas variables también se transforman mediante logaritmo natural (con log1p), generando columnas nuevas con nombres del tipo log_DI_sangre. El propósito es capturar mejor la dinámica relativa de estos valores, que pueden ser muy dispersos.

3. Variables de flujo (flow_vars)

   Incluyen tanto flujos instantáneos como medias de flujo por cuarto.

   La transformación logarítmica sobre estos campos ayuda a suavizar picos extremos y resaltar proporciones relativas entre cuartos, más que sus valores absolutos crudos.

En conjunto, estas nuevas variables logarítmicas constituyen la primera capa de features derivados sobre el dataset de ordeño individual, preparando el terreno para análisis estadísticos más robustos y para modelos posteriores que se benefician de distribuciones menos sesgadas.