Model Overview (Primera iteración)

Esta fase consiste en implementar pipelines reproducibles que integren preprocesado (imputación, codificación, escalado), entrenamiento y búsqueda de hiper parámetros con control de versiones y seeds. Para que de esta forma, los candidatos a modelo final se evalúan de la misma forma.

3.1. Variable Objetivo y Predictores

Variable dependiente (Y):

- `DEL (kg)`: Cantidad total de leche obtenida por día de ordeña, indicador clave de productividad y eficiencia del sistema.

Variables predictoras (X): La selección se basó en el análisis de correlación previo:

- `Duración_seg`: Tiempo de ordeña convertido a segundos
- `Número de ordeño`: Frecuencia o turno de ordeña
- `DI/DD/TI/TT - media de flujos`: Flujo promedio por cuarto de ubre (Delantera Izquierda, Delantera Derecha, Trasera Izquierda, Trasera Trasera)
- `DI/DD/TI/TT - conductividad`: Indicadores de salud de la glándula mamaria

3.2. Preprocesamiento Integrado

Cada pipeline incluye las siguientes etapas estandarizadas:

Imputación: SimpleImputer con estrategia mediana para manejar valores faltantes de forma robusta.

Escalado: StandardScaler para normalizar variables numéricas con diferentes rangos (excepto en Random Forest, que no lo requiere).

Transformación temporal: Conversión de formato "mm:ss" a segundos totales para `Duración`.

Limpieza de datos: Eliminación de registros con valores nulos en X o Y mediante máscara booleana.

3.2. Estrategia de Validación

Se implementó **validación cruzada K-Fold** con las siguientes especificaciones:

- **k = 5 folds**: Balance entre varianza y costo computacional
- **shuffle = True**: Aleatorización de muestras antes de dividir
- **random_state = 42**: Garantiza reproducibilidad de resultados
- **Métricas evaluadas:**

- **R^2 (coeficiente de determinación)**: Mide la proporción de varianza explicada por el modelo
- **RMSE (Root Mean Squared Error)**: Cuantifica el error promedio en kg de producción

En cada iteración, el modelo se entrena con 4 folds (80% de datos) y se evalúa con el fold restante (20%), reportando promedios y desviaciones estándar.

3.3. Modelos Candidatos

Se evaluaron cuatro algoritmos de regresión con sus respectivos pipelines:

3.3.1 Linear Regression

Pipeline base sin regularización, usado como benchmark de referencia.

3.3.2 Ridge Regression

Regresión lineal con regularización L2 para controlar sobreajuste y manejar multicolinealidad entre predictores.

3.3.3 Lasso Regression

Regresión lineal con regularización L1, capaz de realizar selección automática de características (max_iter=10000).

3.3.4 Random Forest Regressor

Ensamble de árboles de decisión, capaz de capturar relaciones no lineales e interacciones entre variables (n_jobs=-1 para paralelización).