

# Model Overview (Segunda 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 evalúen de la misma forma.

## 3.1. Variable objetivo y predictores

El modelo clasifica una vaca en relación a su secado, una variable categórica que representa si una vaca se debe secar o ya no, es decir, clasifica a partir de los días restantes desde la observación actual hasta el momento óptimo de secado de la vaca. El rango oscila entre 0 días (secado inmediato) y más de 300 días (inicio del ciclo), transformando la pregunta "¿cuándo secar?" en una tarea de regresión sobre series temporales.

### Variables predictoras (x)

El modelo procesa más de 50 características organizadas en categorías funcionales:

- **Features temporales:** `Días en leche (DIM)`, `Dur_min`, `` semana/mes de lactancia, componentes estacionales.
- **Ventanas móviles de producción:** Medias, desviaciones estándar y tendencias en ventanas de 3, 7, 14, 21 y 30 días.
- **Características por cuarto de ubre:** Flujos individuales y coeficientes de variación.
- **Atributos derivados para la detección del pico de producción:** Ratio producción actual vs. pico histórico, días desde el pico, tas.
- **Indicadores operativos:** Duración del ordeño, eficiencia (kg/min), detección de sangre.
- **Fases de lactancia:** Clasificación binaria en temprana (0-100d), media (100-200d) y tardía (200+d).

## 3.2. Arquitectura del modelo lstm

### 3.2.1 Estructura de la Red

El modelo implementa una arquitectura LSTM de tres capas con reducción progresiva.

Python

```
Input: (batch, 30 timesteps, 50+ features)
LSTM Layer 1: 128 units (dropout 0.3, recurrent_dropout 0.3)
Dropout: 0.2
```

```
LSTM Layer 2: 64 units
Dropout: 0.2

LSTM Layer 3: 32 units (return_sequences=False)
Dropout: 0.2

Dense Layer 1: 32 units (ReLU)
Dropout: 0.2
Dense Layer 2: 16 units (ReLU)

Output: 1 unit (Linear) → Predicción en días
```

### 3.2.2 Hiper Parámetros Clave

- Optimizador Adam (lr=0.001)
- Estandarización para convergencia estable
- LossHuber Robusto a outliers (MSE para errores pequeños, MAE para grandes)
- Tamaño del Batch de 32
- Balance memoria GPU y estabilidad de gradientes
- 100 Epocas de entrenamiento (con early stop)
- Típicamente converge en 40-60 epochs
- Regularización agresiva por dataset limitado
- Dropout Dense0.2
- Prevención de overfitting en capas finales

### 3.3. Preprocesamiento

**Generación de secuencias:** Ventana deslizante de 30 días consecutivos como input, prediciendo el día 31 como target. Cada ciclo de 200 días genera ~170 secuencias independientes.

**Normalización:** MinMaxScaler [0,1] ajustado SOLO en training set. Apropiado para funciones de activación LSTM (tanh, sigmoid).

**Manejo de nulos:** Forward-fill como estrategia primaria (último valor conocido), luego se reemplazó por 0. Ciclos con >20% nulos se excluyen.

### 3.3. Estrategia de entrenamiento

Los datos se dividieron siguiendo el siguiente split:

- **Training Set (80%)**

- **Test Set (20%)**
- **Validation (20% del train):** Subdivisión interna para early stopping y reduce LR

Aplicamos un shuffle a los datos debido a que las secuencias de diferentes vacas son independientes. No estamos prediciendo el futuro de la misma serie, sino generalizando a nuevas vacas.

**No se usa K-Fold:** Por costo computacional (15-30 min por entrenamiento  $\times$  K folds) y porque para redes profundas el estándar es validation set fijo con early stopping.




### Callbacks de Control

#### Early Stopping:

- Monitorea `val_loss` con patience = 15 epochs
- Restaura pesos del mejor epoch (no el último)
- Previene overfitting cuando train loss baja pero val loss sube

### 3.4. Sistema de reglas

Además del LSTM, se implementa un sistema de reglas expertas que clasifica vacas en bandas de riesgo:

-  **ROJO** kg/min < 0.9 AND pendiente  $\leq$  0 AND DIM  $\geq$  240 Secado inmediato.
-  **ÁMBAR** kg/min < 1.0 AND pendiente  $\leq$  0 AND DIM  $\geq$  200 Monitoreo intensivo.
-  **VERDE** Resto de casos Operación normal.

#### Ventaja de la complementariedad:

- **LSTM:** Predicciones numéricas precisas (15-30 días de anticipación) para planificación estratégica.
- **Reglas:** Alertas binarias interpretables para decisiones operativas diarias.
- **Robustez:** Reglas actúan como red de seguridad si LSTM falla.

### 3.5. Métricas de evaluación

Además de las métricas clásicas como la  $R^2$ , F-Score y el MSE se implementaron otras métricas. Tales como:

Matriz de confusión binaria  $\rightarrow$  Alerta (ROJO/ÁMBAR) y No Alerta (VERDE)

- **Precisión:** De las alertas emitidas, % correctas (minimiza tiempo desperdiciado).
- **Recall:** De las vacas que necesitan secado, % detectadas (minimiza pérdidas por secado tardío).
- **Trade-off:** Priorizar Recall > Precisión (mejor prevenir que lamentar en ganadería).