

# Revisión del Proyecto

## ¿Qué salió bien?

- **Integración exitosa**  
La información estaba originalmente separada en múltiples fuentes y con formatos distintos. Integrarla permitió reconstruir el comportamiento diario de cada vaca. Esto fue clave porque sin un dataset unificado no era posible modelar patrones productivos o reproductivos consistentes.
- **Desarrollo de un pipeline reproducible**  
Un pipeline reproducible evita ambigüedades, errores manuales y cambios involuntarios en los datos. Garantiza que cualquier miembro del equipo pueda repetir los pasos y obtener los mismos resultados, lo cual es importante para el mantenimiento.
- **Selección correcta de técnicas**  
Los modelos basados en árboles manejan bien datos ruidosos, no lineales y con valores faltantes, que representan exactamente las condiciones del dataset del CAETEC. Esto permitió obtener resultados robustos y estables.
- **Evaluación rigurosa y comparativa entre modelos**  
Realizar la evaluación bajo las mismas condiciones y métricas permitió seleccionar el modelo adecuado sin sesgos. Aseguró que la decisión de elegir XGBoost fuera sólida y respaldada por evidencia.
- **Colaboración continua con stakeholders**  
La opinión de Lupita fue necesaria para interpretar el comportamiento fisiológico y reproductivo de las vacas, especialmente al reconstruir estados o validar resultados del modelo.
- **Construcción de un dashboard funcional**  
El dashboard permitirá a Lupita ver resultados prácticos, garantizando que el proyecto tuviera aplicación real y no solo teórica.

## ¿Qué salió mal?

- **Dependencia del dataset aumentado**  
El modelo aprendió en parte reglas derivadas del propio proceso de augmentación, lo que limita su capacidad de generalizar a datos reales y nuevos.
- **Dataset inicial incompleto y mal estructurado**  
Los datos venían fragmentados por vaca, lo que dificultó la integración y generó muchos pasos de transformación que podrían haberse evitado con un formato

uniforme.

- **Riesgo de leakage en la validación**

Las particiones aleatorias incluyeron registros del mismo animal tanto en entrenamiento como en validación, lo que infló artificialmente las métricas del modelo.

- **Altos niveles de nulos y ruido operativo**

Esto generó la necesidad de limpieza extensa, retrasos y más pasos de preprocesamiento. Muchos valores no describían estados fisiológicos útiles.

- **Iteraciones adicionales para reconstruir el estado reproductivo**

El estado reproductivo no venía explícitamente en los archivos. Hubo que inferirlo, lo cual tomó tiempo y generó complejidad adicional.

- **Falta de nuevos datos reales para validar generalización**

Esto impidió una validación robusta del modelo fuera del dataset existente.

## ¿Qué se hizo bien?

- **Aplicación de CRISP-DM**

Permitió que el proyecto avanzara de manera ordenada, con entregables formales por fase y claridad sobre decisiones tomadas.

- **Documentación extensa del pipeline**

Esto garantiza trazabilidad y auditoría, ya que cualquier transformación está registrada y puede replicarse o revisarse si un error aparece.

- **Validación por expertos del dominio**

Los datos del hato no pueden interpretarse sin conocimiento veterinario. Su intervención aseguró interpretaciones correctas.

- **Uso de técnicas robustas como XGBoost**

Permitió controlar mejor el overfitting y manejar la complejidad del dataset.

## ¿Qué se necesita mejorar?

- **Mejor calidad y estructura de datos originales**

Tener datos en múltiples archivos, con columnas ambiguas o inconsistentes, aumenta el retrabajo y reduce la calidad del modelo.

- **Menor dependencia del aumento artificial de datos**

Los modelos aprenden patrones artificiales que no siempre se reflejan en la operación real.

- **Mejor planeación inicial sobre requisitos de datos**  
Muchos datos necesarios (eventos, inseminaciones, partos) no estaban contemplados al inicio, y debieron pedirse después.

## Pitfalls

- **Múltiples archivos por vaca**  
Esto complicó la integración; cada archivo requería parseo manual.
- **Columnas sin explicación clara**  
Muchas columnas del robot eran duplicadas o inconsistentes, lo que generó confusión y errores en la limpieza.
- **Patrón aprendido influenciado por datos aumentados**  
El modelo aprendió reglas derivadas, no necesariamente reales.
- **Estados reproductivos no explícitos**  
Hubo que inferirlos combinando eventos de las fichas técnicas, lo cual agrega incertidumbre.

## Enfoques Engañosos

- **Tratar el dataset crudo como un recurso suficientemente estructurado**  
Se descubrió que era necesario limpiarlo profundamente antes de modelar.
- **Intentar modelar directamente con registros individuales de ordeño**  
La variabilidad era demasiado alta; fue necesario hacer agregación diaria.
- **Intentar modelar sin reconstruir el estado reproductivo**  
El modelo rendía mal porque faltaba información fisiológica esencial.
- **Asumir que los archivos del robot tenían una estructura estable**  
Los nombres de columnas cambiaban entre archivos.
- **Intentar resolver el problema sin pedir más datos**  
Fue necesario solicitar más información al CAETEC para:
  - reconstruir inseminaciones
  - calcular DEL
  - identificar partos
  - asociar días en gestación

## Pistas para Seleccionar Técnicas Adecuadas en Proyectos Similares

- Si hay alto riesgo de sobreajuste: preferir modelos con regularización interna (XGBoost).

- Si las clases son desbalanceadas: evitar modelos lineales simples o MLP sin ajuste de pesos.
- Si hay dependencias temporales o fisiológicas: reconstruir estados antes de modelar.