Tecnológico de Monterrey

Campus Querétaro, Querétaro

Reporte final - Vacas Saturno Saturnitas 🐮🪐

Inteligencia artificial avanzada para la ciencia de datos II | TC3007

Alumnos:

Kevin Alejandro Ramírez Luna	A01711063
Diego Antonio García Padilla	A01710777
José Eduardo Viveros Escamilla	A01710605
Fidel Alexander Bonilla Montalvo	A01798199
Guadalupe Paulina López Cuevas	A01701095
Ángel Mauricio Ramírez Herrera	A01710158
Cristian Chávez Guía	A01710680

19 oct 2025

Introducción

Este proyecto desarrolló un sistema analítico para mejorar la toma de decisiones en el manejo del hato lechero de Campo AGRO Experimenta del Tecnológico de Monterrey, campus Qro (CAETEC). El objetivo principal se centró especialmente en la determinación del momento óptimo de secado de las vacas.

> *"El secado en las vacas se refiere al periodo de reposo que se establece entre dos lactaciones, durante el cual se deja de ordeñar a la vaca para permitir la regeneración del tejido mamario y la recuperación de sus reservas corporales."*

Esta decisión es crítica porque influye directamente en los costos de operación: una vaca en secado o con baja producción que sigue usando el robot de ordeña genera uso ineficiente de la máquina, consume energía, incrementa la carga operativa y ralentiza a todo el sistema. Por ello, se necesitaba un mecanismo objetivo que clasificara el estado productivo de cada vaca con base en datos reales del robot DeLaval VMS V300 y de los registros del hato, las fichas técnicas.

Para resolver este problema, se construyó un pipeline completo de minería de datos siguiendo la metodología CRISP-DM(Cross Industry Standard Process for Data Mining), desde la comprensión del negocio, integración de datos hasta el modelado, evaluación y despliegue de nuestro modelo.

Se eligió esta metodología porque proporciona trazabilidad, control de versiones y una estructura clara para comunicación con stakeholders, lo cual era esencial dado que CAETEC requería documentación replicable para otros hatos.

Como resultados del proyecto se incluye un modelo predictivo (XGBoost), un dashboard interactivo, scripts de preparación de datos y documentación técnica completa para que el sistema pueda operar localmente desde el primer día y crecer en fases posteriores.

Contexto y Objetivos del Proyecto

El CAETEC opera bajo estándares de producción de Alpura, lo que exige un control riguroso de eficiencia productiva en la leche para el consumo humano. Un reto recurrente es identificar cuándo una vaca está entrando al periodo de pre-secado, ya que durante esta fase su rendimiento cae y su uso del robot debe reducirse. Hasta ahora, esta decisión dependía del criterio de las veterinarias, lo cual, aunque valioso, introduce variabilidad y posibles retrasos en la implementación de medidas.

El proyecto se diseñó para dar soporte a tres objetivos estratégicos:

- reducir días improductivos
- estandarizar decisiones veterinarias
- incrementar el aprovechamiento del robot de ordeña.

Para lograrlo, los objetivos de minería de datos se definieron en torno a automatizar la clasificación del estado productivo y reconstruir variables fisiológicas que no estaban disponibles directamente en los registros.

La selección de estos objetivos no fue arbitraria: respondieron a la necesidad de disponer de un sistema que integrará productividad, fisiología y reproducción en un marco consistentemente analizable, algo que no era posible con la estructura original de los datos.

Resultados del Modelado y Evaluación

El proyecto exploró diversos enfoques a lo largo de tres iteraciones, comenzando con modelos de regresión y series de tiempo, los cuales después de un proceso detallado y específico de evaluación se descartaron porque la producción láctea diaria presenta una variabilidad que no se puede predecir de forma estable únicamente con secuencias temporales. Esto justificó el cambio hacia un enfoque de clasificación, que se alineaba mejor con cómo las veterinarias toman decisiones: no predicen litros diarios, sino **estados fisiológicos**.

Con esta reformulación, se entrenaron diversos modelos, yendo desde los más simples de Machine Learning (Regresión Logística, XGBoost, Random Forest) a arquitecturas más complejas de Deep Learning (TabNet e MLP). La elección final del modelo usado en el dashboard se sustentó en la obtención de las mejores métricas que den mayor valor predictivo a nuestra problemática. Cosa que recayó en XGBoost al ofrecer el mejor equilibrio entre desempeño y robustez: teniendo un accuracy de aproximadamente 0.97% y un macro-F1 cercano a 0.95%. Pero lo más importante no fueron las métricas, sino su **capacidad de regularizar**, manejar datos ruidosos y evitar overfitting, lo cual era crucial dada la mezcla de datos reales y los aumentados en el proceso de entrenamiento.

Aunque Random Forest logró métricas más altas, se descartó como modelo primario porque sus valores superiores eran resultado de un mayor sobreajuste (overfitting); en cambio, se conserva como modelo de referencia por su interpretación interna.

Modelos más simples (como Regresión Logística) o más complejos (como TabNet o MLP) se mantuvieron como baselines, ya que ninguno superó la estabilidad del modelo basado en métodos de ensamble con árboles.

- **XGBoost**
  - Accuracy ≈ 0.97
  - Macro F1 ≈ 0.95
  - Alta estabilidad y regularización
  - Menor riesgo de overfitting que Random Forest
  - Mejor balance entre desempeño y robustez
- **Random Forest**
  - Métricas incluso más altas en el dataset,
  - Pero con **riesgo de overfitting**, por lo que se usa como modelo de referencia.
- **Modelos alternativos** (MLP, TabNet, LogReg)
  - Correctos como baseline, pero menos estables.

Resultados respecto a los Objetivos de Negocio

El sistema final cumple con los objetivos del CAETEC al automatizar la clasificación del estado productivo, desacoplando la valoración clínica de la carga operativa diaria. El modelo permite detectar caídas anormales en la producción, anticipar estados de pre-secado y priorizar vacas que requieren intervención.

Si bien el dataset utilizado depende parcialmente de datos aumentados, esta decisión se justificó porque el CAETEC no contaba con suficiente información real y continua del hato. El Data Augmentation se diseñó cuidadosamente para respetar reglas fisiológicas validadas con las veterinarias, y en la fase de evaluación se demostró que, con los datos actuales, este enfoque era el único viable para obtener modelos confiables que respalden las necesidades del CAETEC y de cualquier hato con el robot de DELAVAL.

La justificación y uso del data augmentation fue indispensable para el modelo. Ya que sin este último habría sido inservible el modelo por falta de representatividad en ciertas clases.

En síntesis, el sistema automatiza un proceso que antes requería horas de revisión humana y lo transforma en una herramienta objetiva, consistente y utilizable día a día.

Valor agregado de estos resultados

- Por primera vez, el CAETEC tiene un **sistema cuantitativo** para identificar el momento de secado.
- El modelo traduce métricas fisiológicas y productivas en categorías clínicas útiles para la toma de decisiones veterinarias.

Resultados respecto a los Objetivos de Negocio

El sistema permite:

- Automatizar la clasificación del estado productivo, reduciendo variabilidad entre veterinarias.
- Detectar **caídas anormales** de producción con suficiente anticipación.
- Priorizar revisiones clínicas y redistribuir recursos (alimentación, tiempo de ordeño).
- Reducir días improductivos antes del secado.
- Tomar decisiones informadas con manejo con información objetiva.

Descripción del Proceso y Costos

El proyecto siguió CRISP-DM porque esta metodología permite capturar decisiones, supuestos y transformaciones de manera ordenada. Esto era importante porque los datos del robot y de las fichas del hato no estaban diseñados para análisis predictivo; por lo tanto, se necesitaba un proceso estructurado para documentar cada paso.

Durante la fase de comprensión de datos se identificaron problemas clave: múltiples archivos CSV por vaca, columnas duplicadas, valores faltantes y registros con semántica ambigua (por ejemplo, mediciones por cuarto de la ubre difíciles de interpretar). Detectar esto tempranamente evitó errores posteriores.

La fase de preparación de datos requirió más esfuerzo que lo inicialmente planeado porque fue necesario reconstruir estados reproductivos a partir de eventos que no venían codificados explícitamente, extraer fechas y horas en formatos inconsistentes y consolidar registros diarios. Esta reconstrucción se justificó porque sin ella el modelo no dispondría del contexto fisiológico indispensable para identificar el periodo de pre-secado.

En la fase de modelado, se seleccionaron técnicas diversas para evitar sesgos y se llevaron a cabo pruebas exhaustivas para elegir el mejor modelo. La evaluación formal consolidó estos hallazgos, justificando el avance al deployment.

Los costos del proyecto se concentraron en horas de trabajo de análisis, ingeniería de datos y desarrollo de dashboards. No se requirió inversión en infraestructura, lo que representa una ventaja estratégica para el CAETEC.

Costos y Esfuerzo del Proyecto

- **Infraestructura**
  - No requirió inversión en hardware adicional.
  - Se utilizó infraestructura existente: robot, computadoras institucionales, archivos históricos.
- **Costos principales**
  - Horas de trabajo en:
    - limpieza de datos,
    - reconstrucción del estado reproductivo,
    - modelado,
    - iteración y validación,
    - desarrollo del dashboard.
- **Beneficios esperados**
  - Reducción de días improductivos por vaca.
  - Mejor uso del robot DeLaval.
  - Ahorro en energía y alimentación.
  - Menos tiempo invertido por veterinarias en análisis manual.
  - Toma de decisiones más estandarizada.

Desviaciones del Plan Original

El plan inicial contemplaba un proceso lineal de regresión, modelado y deployment, pero esta ruta no era adecuada dada la naturaleza del problema. Reformular el problema como clasificación multiclase fue necesario porque los resultados de regresión eran inestables y no aportaban información útil.

Otra desviación importante fue la calidad de los datos: la integración del dataset real y la reconstrucción del estado reproductivo tomó más tiempo de lo estimado, pero fue esencial para obtener un modelo funcional.

Siguiendo la metodología CRISP-DM, el recorrido que tomamos como equipo fue el siguiente. Donde la ruta de trabajo, como se mencionó anteriormente, se realizó a lo largo de 3 iteraciones para asegurarnos que nuestros objetivos se cumplieran.

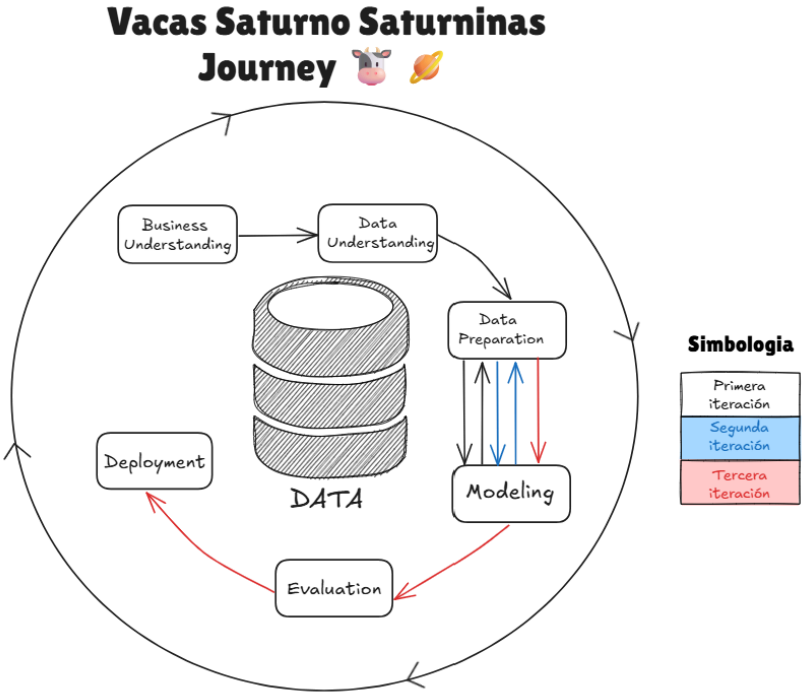

Figura 1. Recorrido de VSS a través de CRISP-DM

En la parte de Data Preparation se tenían diversos datasets con las siguientes caracteristicas:

- múltiples archivos por vaca
- columnas inconsistentes
- nombres distintos entre periodos
- nombres de columnas sin un nombre claro
- estados reproductivos no explícitos

Durante las dos primeras iteraciones del proyecto fue necesario reconstruir variables, calcular atributos derivados y solicitar más datos al CAETEC.

Finalmente, el plan de deployment inicial contemplaba integrar el sistema directamente en los sistemas del CAETEC. Esto se pospuso porque se determinó que el enfoque más seguro era entregar un sistema local, replicable y de uso inmediato, antes de proceder a una integración de mayor complejidad. Sin embargo, por la falta de acceso al sistema cerrado de DeLaval se propone como implementación futura la solicitud del API para extracción de datos y cargador automático al sistema y pipeline de limpieza. Por el momento solo se cuenta con una carga manual de los datos.

Plan de Deployment y Mantenimiento

El sistema entregado consta de un modelo XGBoost entrenado, un modelo Random Forest de respaldo, scripts de preparación y un dashboard. Todo esto se entrega en un repositorio que puede ejecutarse de manera local. Este enfoque se eligió porque minimiza riesgos, evita dependencias con los sistemas actuales del CAETEC y permite que el sistema funcione desde el primer día.

El plan de monitoreo incluye validaciones periódicas, comparación con decisiones clínicas reales y protocolos para reentrenar el modelo cuando lleguen datos nuevos. Este monitoreo es esencial porque el modelo fue entrenado con datos aumentados y porque el comportamiento productivo del hato puede cambiar a lo largo del tiempo.

Recomendaciones para Trabajo Futuro

Las recomendaciones se enfocan en mejorar la capacidad predictiva y operativa del sistema. La más importante es recolectar más datos reales para reducir la dependencia del augmentation y permitir validaciones robustas. También se recomienda revisar periódicamente las reglas fisiológicas y explorar nuevos modelos para otros problemas clínicos (mastitis, eficiencia, persistencia).

Otra recomendación es diseñar un plan de integración progresiva con los sistemas internos del CAETEC, permitiendo automatizar la lectura de archivos y la generación de alertas.

Finalmente, se sugiere realizar un análisis económico formal que cuantifique el ahorro exacto generado por el modelo, lo que ayudará a justificar futuras inversiones en infraestructura.