

Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos

Entendimiento de negocio

Los concentrados v2

Daniel Queijeiro Albo - A01710441

Diego Alfaro Pinto - A01709971

Diego Isaac Fuentes Juvera - A01705506

Jesus Ramirez Delgado - A01274723

Mauricio Anguiano Juarez - A01703337

Luis Adrián Uribe Cruz - A01783129

Índice

Adaptaciones CRISP-DM.....	3
Objetivos de negocio.....	3
Contexto.....	3
Objetivos de negocio.....	4
Criterios de éxito de objetivos de negocio.....	4
Situación actual.....	4
Inventario de recursos.....	4
Requisitos, supuestos y restricciones.....	5
Riesgos y contingencias.....	6
Terminología.....	6
Minería de datos.....	7
Objetivo de minería de datos.....	7
Criterios de éxito de minería de datos.....	7
Plan de proyecto.....	8
Plan.....	8
Evaluación inicial de herramientas y técnicas.....	8

Adaptaciones CRISP-DM

En la fase de **Business Understanding** se mantuvieron todos los pasos estipulados por la metodología CRISP-DM, siguiendo su estructura original y su traducción al español:

- Business objectives Objetivos de negocio.
- Assess the situation Situación actual.
- Data mining goals Minería de datos.
- Project plan Plan de proyecto.

No se realizaron adaptaciones ni modificaciones los componentes, exceptuando la actividad “Cost and Benefits”, ya que el proyecto no tiene un propósito monetario, sino un fin académico y beneficio del Tecnológico de Monterrey.

1. Objetivos de negocio

1.1 Contexto

En este proyecto trabajaremos con el Campo Agroexperimental del Tecnológico de Monterrey (CAETEC), el cual cuenta con uno de los 3 establos robotizados del TEC de un total de 5 en México que nos permitirá llevar el proyecto a cabo; lo que lo posiciona como referente en ganadería de precisión y tecnología aplicada al sector lechero. En las instalaciones cuentan con ganado bovino Holstein destinado a la producción lechera con sección genética multi-generacional, el cual es ordeñado mediante un sistema automatizado VMS (Voluntary Milking System) de la empresa DeLaval. Este sistema opera con 3 robots de ordeño (BMC 1, 2 y 3) donde las vacas acuden voluntariamente mediante un sistema de incentivos con alimento peletizado.

La operación genera una producción diaria de 5,000-6,000 litros de leche con un promedio de 35 litros por vaca al día aproximadamente, entregados diariamente a Alpura, quien evalúa la calidad del producto.

A pesar de la automatización avanzada, el CAETEC enfrenta varios desafíos críticos que limitan su eficiencia operativa. La mastitis subclínica puede pasar desapercibida hasta manifestarse clínicamente, haciendo que las vacas afectadas no puedan producir leche durante 15 días, además de generar pérdidas por leche descartada y lavados del robot (45-60 minutos sin producción por ciclo). Adicional a esto existe una variabilidad estacional ya que en el verano las vacas suelen estar más incómodas y son más propensas a ser molestadas por moscas, haciendo que den más patadas y se lastimen lo que resulta en un aumento de casos de mastitis de hasta 2.5%. Alpura puede detener la recolección si detecta colonias bacterianas elevadas o presencia de sangre, afectando directamente el ingreso al establo. A partir de estos datos y desafíos, el objetivo del proyecto es desarrollar un modelo de inteligencia artificial

que permita atender estas necesidades específicas del sector mediante la generación de conocimiento útil para la toma de decisiones, la mejora en la producción de leche y la prevención de enfermedades de temporada en el ganado.

Problemas a los que se enfrentan

1. Fallas o paros en el robot de ordeña, estos causan lavados largos, calibraciones incorrectas y mantenimiento, a su vez generando un tiempo muerto de 45-60 min durante el ciclo de lavado del robot, que a su vez deriva en largas filas de espera para las vacas haciendo que se estresen y producen menos leche.
2. Mastitis clínica y subclínica que puede pasar desapercibida y causar pérdidas de leche y tiempos muertos.
3. Estrés térmico y presencia de moscas en verano que incrementa las patadas y la incidencia de mastitis en las vacas.
4. El descarte de la leche dependiendo si hay porcentaje de antibiótico o sangre dentro de la leche lo que se traduce en pérdida de producto y dinero.
5. Dependencia de DeLaval para la obtención, medición e interpretación de métricas.

1.2 Objetivos de negocio

Además de contar con sistemas DeLaval y el software DelPro para el monitoreo del ganado y control de la producción, el CAETEC busca potenciar el aprovechamiento de los datos generados por estos sistemas.

El objetivo de negocio del proyecto es:

Generar indicadores para la mejora genética y/o manejo del hato.

1.3 Criterios de éxito de objetivos de negocio

- Se identifican características conductuales o fisiológicas estadísticamente asociadas con:
 - Mayor producción de leche.
 - Menor incidencia de problemas de salud.
 - Mejor adaptación al sistema robotizado.
- Se califican las vacas según sus características genéticas. Debido a la subjetividad de los objetivos, su cumplimiento debe ser aprobado por lo menos por uno de los siguientes expertos:
 - Dra. Guadalupe Lopez Rendón
 - Dr. Ivo Neftali Ayala García

- Ing. Sergio Sebastián Caballero Chávez

2. Situación actual

2.1 Inventario de recursos

Contamos con los siguientes recursos:

- El sistema DelPro, donde se registran todos los datos de ordeño provenientes
- Dataset de imágenes: 57,870 imágenes verticales y horizontales de vacas al salir de la máquina de ordeña (para clasificación de condición corporal y detección de anomalías físicas).
- Dataset CSV: Información detallada sobre cada vaca y sus ordeñas (producción, consumo, patadas, MDI, sangre, incompletos, estado reproductivo, etc...).
- Infraestructura:
 - Robots DeLaval, collares de rumia (parcialmente disponibles), dashboard actual.
 - Infraestructura de CAETEC.
- Recursos humanos:
 - Personal de campo, veterinarios, operarios y equipo de análisis de datos.

2.2 Requisitos, supuestos y restricciones

- Requisitos
 - Diseñar un modelo de predicción de desempeño individual que utilice datos históricos de producción, comportamiento y salud.
 - Asegurar la trazabilidad y limpieza de los datos antes del modelado, documentando todo el flujo de procesamiento.
 - Optimizar el uso de recursos computacionales, priorizando la eficiencia sobre la complejidad algorítmica.
- Suposiciones
 - Los datos obtenidos por el sistema DeLaval reflejan adecuadamente la productividad y el comportamiento real de las vacas.
 - El comportamiento observado (patadas, rechazos, tiempos de ordeño, etc.) tiene relación directa con el bienestar y la eficiencia productiva.
 - Las vacas mantienen rutinas relativamente estables en cuanto a horario, alimentación y frecuencia de ordeño, lo que permite modelar patrones consistentes.
 - Los factores externos como el clima o el estado reproductivo influyen, pero no anulan las tendencias generales de comportamiento y producción.
 - Los datos históricos disponibles son suficientes para representar distintos perfiles de vacas dentro del hato (baja, media y alta productividad).
- Restricciones

- Algunos registros pueden estar incompletos o contener valores anómalos por fallos de sensores o cambios en la configuración de los robots.
- Las condiciones del entorno físico (temperatura, iluminación, disposición del corral) pueden haber variado entre períodos de registro, afectando la homogeneidad de los datos.
- No todos los factores que influyen en la productividad están registrados (por ejemplo, alimentación detallada, jerarquía social o estrés por traslado).

2.3 Riesgos y contingencias

En el siguiente enlace se puede encontrar la matriz de riesgos y planes de contingencia:
enlace.

 PVG concentración IA

2.4 Terminología

Término	Descripción
Antibiótico en leche	Leche de vacas bajo tratamiento (no se destina a consumo).
Balance energético negativo	Estado metabólico en el que la vaca produce más energía en forma de leche de la que consume en alimento, lo que causa pérdida de condición corporal.
BCS (Body Condition Score)	Evaluación visual del estado corporal de la vaca en escala de 1 (muy flaca) a 5 (muy gorda).
Collar de rumia	Dispositivo que mide actividad, tiempo de rumia, tiempo echada/parada y comportamiento reproductivo.
Condición corporal estable	Cambios graduales en la escala BCS, sin caídas bruscas, deseables para salud reproductiva.
Conductividad eléctrica	Medida de la leche en cada cuarto; valores altos se asocian con mastitis o inflamación.
CMT (California Mastitis Test)	Prueba química de campo para detectar mastitis.
DeLaval	Empresa sueca especializada en tecnología para la producción lechera. El sistema DeLaval VMS™ (Voluntary Milking System) automatiza el ordeño y registra datos sobre producción, salud y comportamiento de las vacas.
DEL (Días en leche)	Número de días transcurridos desde el último parto.

DEL 305	Estándar de producción de leche a 305 días en leche; usado para comparar el rendimiento de vacas
Días abiertos	Tiempo transcurrido entre el parto y una inseminación fecundante.
Días preñada	Tiempo transcurrido desde la fecundación hasta el parto esperado.
Hato	Conjunto de cabezas de ganado.
Incompletos	Ordeños en los que al menos un cuarto de la ubre no fue ordeñado correctamente.
MDI (Mastitis Detection Index)	Índice calculado por el robot; valores ≥ 2 sugieren mastitis clínica o subclínica.
Patadas	Eventos registrados cuando la vaca patea y una pezonera se desprende durante el ordeño.
Pezonera	Componente de ordeño que se coloca en cada pezón al momento de la ordeña; Puede desprendérse por patadas.
Periodo seco	Intervalo previo al parto en que la vaca no es ordeñada, permitiendo recuperación de la glándula mamaria
Periodo de transición	Primeros 20 días postparto; etapa crítica por riesgo de balance energético negativo y enfermedades
Rechazo	Vaca que su ordeño fue negado, por no reconocimiento de la vaca o no se pudo concretar la ordeña.
Sangre en leche	Detectada automáticamente por el robot; a partir de 800 ppm la leche es descartada.
SCC / OCC (Somatic Cell Count)	Conteo celular somático (no medido directamente por el robot). Se utiliza como indicador de salud de la ubre y calidad de la leche.
Hacinamiento	Situación en la que se mantienen demasiados animales en un espacio reducido, lo que supera los límites necesarios para su salud, producción y bienestar.
Heredabilidad	Proporción de la variación observada en un rasgo (como la producción de leche) que se debe a factores genéticos y puede ser transmitida a la descendencia.

3. Minería de datos

3.1 Objetivo de minería de datos

Identificar patrones y relaciones que permitan construir indicadores para apoyar la toma de decisiones sobre mejora genética y/o manejo del hato.

3.2 Criterios de éxito de minería de datos

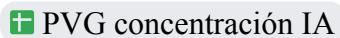
- Se identifican características conductuales y/o fisiológicas con asociaciones estadísticamente significativas ($p < 0.05$) con:
 - Mayor producción de leche.
 - Menor incidencia de problemas de salud.
 - Mejor adaptación al sistema robotizado.

Estas asociaciones permiten generar indicadores cuantitativos por cada dimensión (producción, salud y comportamiento).

- Integración de los indicadores: Los indicadores obtenidos se combinan en un esquema unificado que permite calificar a las vacas según su potencial o mérito genético, facilitando decisiones de **retención, supervisión o descarte**, de acuerdo con un índice cuantitativo validado por al menos uno de los expertos del proyecto, previamente mencionados.

4. Plan de proyecto

4.1 Plan



4.2 Evaluación inicial de herramientas y técnicas

Para el desarrollo del modelo predictivo y el análisis de datos del hato, se emplearán herramientas y librerías de ciencia de datos y aprendizaje automático.

- **Pandas:** Será la herramienta principal para la manipulación, limpieza y análisis exploratorio de los datos. Permitirá integrar información proveniente de los registros de producción, comportamiento y salud de las vacas, así como realizar transformaciones y filtrados.
- **NumPy:** Se utilizará para operaciones matemáticas y de álgebra lineal a gran escala, así como para la optimización de procesos numéricos durante el preprocesamiento de datos.

- **Scikit-learn:** Se empleará como la herramienta principal inicial para el desarrollo de los modelos de comportamiento y sanidad. Se prioriza esta librería debido a que incluye implementaciones directas de los algoritmos base seleccionados, como Random Forest e Isolation Forest, permitiendo un prototipado y validación rápidos.
- **TensorFlow y PyTorch:** Estas librerías se utilizarán en caso de ser necesario implementar modelos más complejos o ajustar las arquitecturas a necesidades específicas que requieran mayor personalización de la que ofrece Scikit-learn.