BUSINESS UNDERSTANDING

Adrián Matute, Pablo Martínez, Osvaldo Del Valle, Andres Callealta, Jorge Martínez

Tecnológico de Monterrey Campus Querétaro, México

A. Metodología CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) o Proceso estándar interprofesional para la extracción de datos, es una metodología de trabajo que describe las fases del proyecto y sus tareas necesarias de cada fase y una explicación de las relaciones entre las tareas.

Business Understanding; definir los objetivos del proyecto e identificar los requerimiento desde una perspectiva de negocio, y diseñar un plan para lograr los objetivos,

- Determinar los objetivos de negocio
  - Contexto
  - Objetivos de negocio
  - Criterios de éxito
- Evaluar la situación
  - Recursos disponibles
  - Requerimientos y limitaciones
  - Riesgos y contingencias
  - Terminología
  - Costos y beneficios
- Determinar los objetivos minería de datos
  - Objetivos de minería de datos
  - Criterios de éxito de minería de datos
- Planificar el proyecto
  - Plan del proyecto
  - Evaluación de herramientas disponibles y técnicas

Data Understanding; familiarizarse con los datos e identificar problemas en los datos con el objetivo de formular una hipótesis preliminar.

- Recolección de datos
  - Reporte recolección de datos
- Descripción de datos
  - Reporte descripción de los datos
- Exploración de los datos
  - Reporte exploración de los datos
- Verificación de la calidad de los datos
  - Reporte calidad de los datos

Data Preparation; definir los datos que se utilizarán para el análisis y los criterios que se deben considerar (relevancia, calidad y restricciones).

- Dataset
  - Descripción de la dataset
- Selección de los datos
  - Justificación para inclusión exclusión
- Limpiar los datos
  - Reporte de las modificaciones de los datos
- Construcción de los datos
  - Generar registros
  - Correlaciones de atributos
- Integración de los datos
  - Conjuntar la información
- Formato de los datos
  - Reconstruir el formato de los datos

Modeling; Seleccionar y aplicar varias técnicas de modelado para lograr los objetivos de minería de datos.

- Seleccionar las técnicas de modelado
  - Técnicas de modelado
  - Modelización de la hipótesis
- Diseñar un modelo de prueba
  - Modelo de prueba
- Construcción de modelo
-
  - Configuración de parámetros
  - Modelos
  - Descripción de modelo
- Evaluación del modelo
  - Revisión de la configuración de los parámetros
  - Evaluación del modelo

Evaluation; lograr los objetivos de negocio y determinar si hay alguna razón de negocio por la cual el modelo es deficiente, a su vez compara los resultados.

- Evaluar resultados
  - Evaluación de la minería de datos
  - Probar el modelo
- Revisión del proceso
  - Revisión del proceso
- Determinar el siguientes pasos
  - Lista de posibles acciones
  - Decisiones

Deployment; Presentar los resultados obtenidos y cómo esto se puede implementar en entornos de producción, y su escalabilidad.

- Plan de implementación
  - Documentar un plan para su implementación
- Seguimiento y mantenimiento del plan
  - Documentar un plan de supervisión y mantenimiento para evitar problemas durante la fase operativa.
- Elaborar informe final
  - Presentación de la minería de datos.
- Proyecto de revisión
  - Mejoras a futuro.
  - Retrospectiva del proyecto.

B. Situación del negocio

CAETEC es un centro de investigación agroindustrial que busca mejorar la eficiencia y sostenibilidad de la producción agrícola y

ganadera mediante tecnologías avanzadas. En este contexto, el negocio está centrado en la optimización de la producción de leche, el bienestar de las vacas, y la reducción de costos operativos a través del uso de inteligencia artificial y sensores para monitorear a las vacas lecheras.

El rancho cuenta con robots de ordeño que monitorean la producción de leche de manera automatizada. Además, se utilizan podómetros y sensores para registrar el movimiento y descanso de las vacas.

La doctora veterinaria Guadalupe López del CAETEC tiene en claro que la productividad de las vacas lecheras dependerá de los estímulos de su entorno. Se prioriza la comodidad de las vacas mediante el monitoreo de su tiempo de descanso, acceso al agua, uso de camas, y su interacción con el entorno.

Problemas actuales:

- **Desperdicio de energía:** Las vacas pasan tiempo de pie en filas para ser ordeñadas, lo que resulta en un uso ineficiente de su energía. Este comportamiento puede reducir la cantidad de leche que producen.
- **Uso de camas:** Las vacas deben descansar en camas de arena para maximizar su comodidad y, con eso, su productividad. Sin embargo, no se tiene un control preciso sobre cuántas vacas están usando las camas, cuánto tiempo descansan, o si están utilizando correctamente las instalaciones.
- **Optimización de ordeño:** El proceso de ordeño está automatizado con robots, pero no existe un mecanismo robusto para identificar cuántas vacas están listas para ser ordeñadas o cuánto tiempo esperan en fila, lo cual afecta la eficiencia general de la producción.

Puntos de interés:

- El CAETEC está interesado en replicar en sus establos un sistema de cámaras instaladas en el techo con el fin de monitorear la actividad de la vaca, consiste en tomar imágenes periódicas de las vacas en el sector de las camas y de la fila para el ordeño.
- Implementación de sistemas inteligentes para analizar las imágenes y contar el número de vacas echadas o de pie, en el sector de descanso, y el número de vacas que se encuentran en espera de ser ordeñadas, optimizando así estos espacios y mejorando su producción.
- Los resultados de este proyecto tienen implicaciones directas en la comercialización de la leche, lo que beneficiaría económicamente al CAETEC.

C. *Objetivos de negocio y objetivos de minería de datos*

Objetivos de negocio:

Identificar patrones significativos de aglomeraciones en la fila de espera para el ordeño.

- El socio formador determinará si las aglomeraciones afectan la eficiencia del proceso de ordeño o el bienestar animal.

Identificar los cuellos de botella en la fila de ordeño.

- El socio formador determinará si los hallazgos proporcionan información útil para optimizar la programación e identificar los cuellos de botella en la fila.

Objetivos de minería de datos:.

Hacer un conteo de las vacas en la fila de espera para identificar patrones en su comportamiento.

Analizar las aglomeraciones en cuestión del tiempo para determinar cuellos de botella.

D. Recursos disponibles para el proyecto

Expertos.

- Arturo González de Cosío: Director de CAETEC. Responsable de la supervisión general del proyecto, asegurando que esté alineado con los objetivos estratégicos de CAETEC.
- Ivo Neftali Ayala García: Ingeniero electrónico. Se encargará de la integración y mantenimiento de los sistemas de cámaras instalados en el rancho, además de un segundo supervisor general del proyecto.
- Guadalupe López Rendón: Especialista pecuaria. Aporta su experiencia en el manejo del ganado bovino
- Dr. Benjamin Valdes Aguirre: Profesor de técnicas y arquitecturas de deep learning.
- Dr. Ismael Solis Moreno: Profesor de Big data y cómputo en la nube.
- Dr. José Antonio Cantoral: Profesor de herramientas para el procesamiento del lenguaje natural.
- Dr. Carlos Alberto Dorantes: Profesor de estadística avanzada para la ciencia de datos.
- Ma. Eduardo Daniel Juárez Pineda: Profesor de Metodologías de proyectos de ciencia de datos.

Datos:

Imágenes aéreas: Las imágenes obtenidas de las cámaras instaladas en el techo del rancho, enfocadas en la sala de espera de ordeño. Este conjunto de datos será clave para entrenar los modelos.

E. *Recursos de cómputo:*

Ejecución de modelo para detección

- Raspberry pi 3 contamos con una placa de desarrollo con las que cuenta el CAETEC para correr los modelos y simular el sistema implementado en el corral.

Cómputo para entrenamiento de modelos CNN

- Pocas unidades (~20) de cálculo en Google Collab Pro, lo que no permitirá entrenar unos modelos con los GPUs T4, L4 o A100, todos de NVIDIA
- Dos laptops gamers con GPU
    - NVIDIA GTX960M
    - Geforce RTX

Cómputo para programar

- Cinco laptops para desarrollar los modelos CNN.

F. Software

Etiquetado de imágenes

- Roboflow; software que ayuda al etiquetado de imágenes para cualquier modelo de CNN.
- Labelme, herramientas de etiquetado.
- LabelGPT, herramienta con AI para etiquetado de imágenes.

Entrenamiento de modelos

- Kaggle, herramienta en web para entrenar modelos CNN, en caso de no tener GPU.
- Colab, herramienta en web para poder entrenar y correr modelos de AI.

Desarrollo de modelos CNN

- Tensorflow, librería de python para construir modelos CNN.
- Pytorch, biblioteca de python para la construcción de modelos.
- Ultralystics, librería con la que cuenta con arquitectura YOLO.

Almacenamiento de datos

- Google Drive, almacenamiento de los avances del proyecto y de los datos.

G. *Requerimientos para el proyecto*

Calendario de entrega

Entendimiento del negocio: 5 de octubre

Preparación de la base de datos: 8 de octubre

Entendimiento de los datos: 14 de octubre

Modelado: 28 de octubre

Evaluación:11 de noviembre

Despliegue:21 de noviembre

Entrega del proyecto: 26 de noviembre.

Comprensibilidad y calidad de los datos

Las imágenes utilizadas en el proyecto están en formato JPG y contienen información clave en su nombre de archivo, como la fecha y hora en que fueron capturadas (ejemplo: 2024-04-14-23-00-03.jpg). Esta nomenclatura facilita el orden cronológico y permite asociar cada imagen con eventos específicos del comportamiento de las vacas, como el tiempo que pasan en la fila de espera para el ordeño.

Gracias al formato de los nombres de archivo, que incluye la fecha y minuto exacto, es fácil organizar las imágenes y correlacionarse con el comportamiento diario de las vacas en diferentes momentos del día. Sin embargo, es necesario considerar que las vacas no están identificadas individualmente.

Aunque las imágenes están en formato JPG, la resolución de las imágenes deberá evaluarse para asegurar que sea suficiente para identificar claramente el comportamiento de las vacas. Cualquier falta de claridad o resolución podría afectar la precisión del modelo de IA.

Una parte de las imágenes fue tomada en condiciones nocturnas. Esto presenta un desafío adicional debido a la baja visibilidad. El análisis efectivo de estas imágenes requerirá preprocesamiento adicional para ajustar el brillo o contraste, o para utilizar técnicas avanzadas de mejora en imágenes que permitan una mejor detección de las vacas en ambientes oscuros.

Seguridad

Aspectos legales

H. Supuestos del proyecto

Supuestos sobre los datos

a. Disponibilidad de los datos: Se asume que la cámara instalada en el rancho estará operativa durante todo el periodo del proyecto, proporcionando un flujo continuo de imágenes en intervalos de 5 minutos.

Supuestos del negocio

b. Colaboración del personal del rancho: Se asume que los operadores y el personal encargado de la gestión del rancho estarán disponibles para proporcionar acceso a los datos, colaborar en el análisis de los resultados y aplicar las recomendaciones derivadas del proyecto.
c. Infraestructura tecnológica: Se supone que el rancho tiene acceso a la infraestructura de cómputo necesaria para almacenar y procesar grandes volúmenes de datos sin afectar la operatividad diaria.

I. Restricciones del proyecto

Disponibilidad de los recursos: El dataset no está etiquetado, ni con el número de vacas presentes en la imagen, y todavía menos con la identificación de cada vaca por imagen.

Aspectos técnicos: Tomando fotos cada cinco minutos sin identificar las vacas, tendremos que aproximar el tiempo de

espera.

Calidad de los datos: Buena parte de la imágenes son durante la noche, cosa que aumenta la dificultad en detectar las vacas.

J. Plan de proyecto

El plan de proyecto se estructura siguiendo el modelo CRISP-DM, con fases claramente definidas. Cada fase incluye ciertas tecnologías y herramientas adecuadas, así como los detalles específicos de las actividades a realizar.

1. Comprensión del negocio

Objetivo: Identificar y entender los objetivos de negocio de CAETEC, centrados en la optimización del proceso de ordeño y el uso de camas para mejorar la producción de leche.

Actividades:

- Visita al CAETEC y reuniones con expertos para establecer los criterios de éxito y definir el impacto del proyecto.
- Identificación de los problemas relacionados con la fila de espera y el descanso de las vacas.

Herramientas y tecnologías:

- Herramientas de colaboración en la nube (Google Drive) para documentar los objetivos y requisitos del negocio.

Fecha de entrega estimada: 5 de octubre.

2. Comprensión de los datos

Objetivo: Entender la estructura y calidad de las imágenes recopiladas, así como las condiciones técnicas bajo las cuales se obtuvieron.

Actividades:

- Revisar la estructura de los nombres de los archivos que contienen fecha y hora.
- Evaluar la calidad de las imágenes (resolución y condiciones de luz).
- Identificar datos faltantes o corruptos.

Herramientas y tecnologías:

- Kaggle.
- Librería de papers sobre temas relacionados.

Fecha de entrega estimada: 14 de octubre

3. Preparación de los datos

Objetivo: Preprocesar las imágenes, incluyendo el ajuste de brillo y contraste de imágenes nocturnas, y etiquetar los datos para el entrenamiento del modelo.

Actividades:

- Preprocesar las imágenes usando técnicas de mejora de imágenes (ajuste a brillo y contraste).
- Etiquetar manualmente las imágenes para identificar vacas.
- Organizar las imágenes en conjuntos de entrenamiento, validación y prueba.

Herramientas y tecnologías:

- Roboflow para etiquetar imágenes.
- Python para el preprocesamiento de las imágenes.

Fecha de entrega estimada: 8 de octubre.

4. Modelado

Objetivo: Desarrollar y entrenar modelos de redes neuronales para detectar el comportamiento de las vacas.

Actividades:

- Cargar el dataset en una carpeta de recursos compartidos.
- Correr el entrenamiento del modelo con las arquitecturas propuestas.

Herramientas y tecnologías:

- Google Colab para entrenar modelos
- Equipos de cómputo locales con GPUs
- Carpetas compartidas en la nube

Fecha de entrega estimada: 28 de octubre

5. Evaluación

Objetivo: Evaluar la precisión del modelo y asegurar que cumpla con los objetivos del negocio.

Actividades:

- Valoración de la eficacia del modelo
- Benchmarks en el hardware objetivo
- Análisis con

Herramientas y tecnologías:

- Raspberry pi 3
- Ffmpeg
- PyTorch Mobile

Fecha de entrega estimada: 4 de noviembre

6. Despliegue

Objetivo: Implementar el modelo en el entorno del rancho para su uso en tiempo real, integrando las cámaras y generando reportes automáticos.

Actividades:

- Montar la solución en el entorno objetivo.
- Recabar datos del negocio real.

Herramientas y tecnologías:

- Raspberry pi 3
- Camara web
- Servidores de Google Drive para almacenar información

Fecha de entrega estimada: 21 de noviembre

K. Glosario de términos

CAETEC: Centro de investigación agroindustrial que se enfoca en la eficiencia y sostenibilidad de la producción agrícola y ganadera mediante el uso de tecnologías avanzadas del Tecnológico de Monterrey

Inteligencia Artificial (IA): Tecnología que simula la inteligencia humana en máquinas, permitiendo la automatización de procesos, análisis de datos y toma de decisiones.

Podómetros: Dispositivos que miden el número de pasos o la actividad física, en este caso, utilizados para monitorear el movimiento de las vacas.

Robots de ordeño: Máquinas automatizadas que ordeñan a las vacas sin intervención humana, optimizando el proceso y monitoreando la producción de leche.

Sensores: Dispositivos que capturan datos sobre el entorno o los animales, como el movimiento o el descanso de las vacas.

CNN (Red Neuronal Convolucional): Tipo de red neuronal usada comúnmente para el análisis y procesamiento de imágenes, clave para el reconocimiento de patrones en fotografías.

YOLO (You Only Look Once): Arquitectura de red neuronal para la detección rápida de objetos en imágenes, utilizada en la visión por computadora.

Etiquetado de imágenes: Proceso de asignar etiquetas o anotaciones a elementos específicos dentro de una imagen, necesario para entrenar modelos de aprendizaje automático.

Big Data: (Conjunto de tecnologías que permiten manejar, almacenar y analizar grandes volúmenes de datos para obtener información útil.) Se refiere a datos que tienen un volumen, velocidad o complejidad tan elevados que resulta complicado o imposible manejarlos usando métodos convencionales.

Cómputo en la nube: Provisión de recursos de computación como almacenamiento y procesamiento a través de internet, facilitando el uso de grandes infraestructuras sin necesidad de adquirir hardware costoso.

GPU (Unidad de Procesamiento Gráfico): Hardware especializado en la manipulación de gráficos y cálculos paralelos, usado en el entrenamiento de modelos de inteligencia artificial.

Roboflow: Software que facilita el etiquetado y procesamiento de imágenes para entrenar modelos de inteligencia artificial, específicamente redes neuronales convolucionales (CNN).

Colab: Plataforma web proporcionada por Google para desarrollar y entrenar modelos de machine learning, permitiendo acceso a recursos computacionales como GPUs.

Kaggle: Plataforma web que ofrece acceso a datasets y herramientas de machine learning, usada para entrenar y probar modelos cuando no se dispone de hardware adecuado.

Labelme: Herramienta de código abierto para etiquetar imágenes manualmente, permitiendo la anotación de objetos para su uso en modelos de inteligencia artificial.

Infraestructura tecnológica: Conjunto de tecnologías, tanto hardware como software, necesarias para soportar los procesos de almacenamiento, procesamiento y análisis de datos.

Supuestos del proyecto: Condiciones que se toman como ciertas para poder llevar a cabo el desarrollo del proyecto, como la disponibilidad de datos o recursos.