Tecnológico de Monterrey

Inteligencia artificial avanzada para la ciencia de datos 2

Gpo 501

Docentes

Dr. Benjamín Valdés Aguirre

Ma. Eduardo Daniel Juárez Pineda

Dr. Ismael Solis Moreno

Dr. José Antonio Cantoral Ceballos

Dr. Carlos Alberto Dorantes Dosamantes

Integrantes

Carlos Rodrigo Salguero Alcántara	A00833341
Diego Perdomo Salcedo	A01709150
Dafne Fernández Hernández	A01369230
José Emiliano Riosmena Castañón	A01704245
Luis Arturo Rendón Iñarritu	A01703572

Querétaro, Querétaro

1.0 Preparación de los datos

Considerando los resultados de la primera iteración del modelo consideramos que era necesario hacer unos cambios a los datos para mejorar los resultados de los modelos. Este documento detalla cómo preparamos los datos para una tercera iteración.

1.1 Trasfondo

Durante la primera iteración de modelado, se identificaron limitaciones significativas en el aprendizaje de ambos modelos (diurno y nocturno). El tamaño reducido de los datasets especializados (aproximadamente 4,000 imágenes cada uno) era insuficiente para un entrenamiento efectivo. Como solución, se implementó una estrategia de ampliación del dataset: se reclasificaron las imágenes y se expandió la colección hasta alcanzar 8,000 imágenes para cada condición de iluminación, efectivamente duplicando el tamaño del dataset original.

1.2 Adaptaciones de CRISP-DM

La adaptación de la fase de preparación de datos de CRISP-DM en nuestro proyecto fue significativamente más simple que lo sugerido en la metodología estándar debido a la naturaleza específica de nuestros datos y objetivos.

Nuestro dataset consistía únicamente en imágenes y sus clasificaciones (número de vacas y momento del día), lo que eliminó la necesidad de muchas actividades tradicionales de preparación de datos. No requerimos crear atributos derivados, realizar transformaciones complejas o integrar múltiples fuentes de datos como sugiere CRISP-DM. La única transformación necesaria fue la categorización entre día y noche basada en la hora del archivo.

La organización de datos se limitó a una estructura de carpetas simple pero efectiva, categorizando por número de vacas y condición día/noche. Esta estructura fue suficiente para facilitar el entrenamiento del modelo sin necesidad de formateos o transformaciones adicionales sugeridas por CRISP-DM.

1.3 Criterios

Para definir cómo prepararemos los datos debemos de tener presente nuestros objetivo de negocio y de minería de datos y los criterios de éxito de cada uno.

Objetivo de Negocio

- Identificar el número de vacas en cada fila en un periodo de tiempo determinado.
  - Determinar con alta precisión la cantidad de vacas en una imagen. Arturo o Ivo determinarán si la precisión es satisfactoria.

Objetivo de Minería de Datos

- Determinar la cantidad de vacas en cada imagen en cualquier condición.
  - Un modelo para condiciones diurnas con un 80% de precisión.
  - Un modelo para condiciones nocturnas con un 50% de precisión.

Clasificaremos las imagenes de dos formas, por la cantidad de vacas en la imagen y en condiciones diurnas y nocturnas. Esta decisión se fundamentó en la hipótesis de que dos modelos especializados (uno para condiciones diurnas y otro para nocturnas) podrían superar el rendimiento de un único modelo generalista.

2.0 Dataset y Limpieza

Para la clasificación de las imágenes usamos el mismo proceso que se usó en la segunda iteración de preparación de los datos, agregando también las nuevas imágenes. La diferencia más grande es que esta clasificación se realizó solamente con el equipo Vaqueros de Datos sin ayuda externa.

2.1 Clasificación

Se reclasificaron las imágenes y se expandió la colección hasta alcanzar más de 9,000 imágenes para cada condición de iluminación, efectivamente duplicando el tamaño del dataset original. Dado que la iluminación es un desafío, ya que afecta la visibilidad y dificulta contar con precisión el número de vacas, priorizamos aumentar la cantidad de imágenes en el dataset nocturno, aunque también incrementamos el número de fotos en el conjunto diurno.

El proceso de clasificación se llevó a cabo utilizando Roboflow, una herramienta especializada para el etiquetado y preparación de datasets en visión por computadora. El objetivo principal fue categorizar las imágenes según el número de vacas presentes en la fila de ordeño en un momento específico.

Para asegurar la consistencia en la clasificación, se establecieron criterios específicos que definen cuándo se debe considerar la presencia de una vaca. Se determinó que una vaca debe ser contabilizada cuando al menos el 20% de su cuerpo es visible en la imagen. Por otro lado, se establecieron criterios claros para elementos que no deben ser considerados como vacas:

Criterios de Exclusión:

- Partes aisladas del animal (cola, pata, oreja)
- Personal de trabajo
- Desechos animales
- Aves en el entorno
- Sombras (de vacas o personas)
- Elementos del entorno (botas, escaleras, puertas)

Una vez completada la clasificación manual siguiendo estos criterios, se implementó una estructura de organización basada en carpetas numeradas. Cada carpeta se identificó con un número que corresponde a la cantidad de vacas presentes en las imágenes contenidas. Por ejemplo:

- Carpeta "0": Imágenes sin vacas presentes
- Carpeta "1": Imágenes con una vaca
- Carpeta "3": Imágenes con tres vacas
- Y así sucesivamente

Esta estructura organizada facilita el acceso y la gestión eficiente de las imágenes durante las fases posteriores del proyecto, especialmente durante el entrenamiento y validación del modelo.

2.2 Almacenamiento

Para gestionar eficientemente este trabajo distribuido, se implementó un sistema de almacenamiento centralizado utilizando Google Drive como plataforma principal.

La estrategia de almacenamiento se desarrolló de la siguiente manera:

1. **Trabajo Distribuido:**
    - Cada miembro realizó la clasificación de un subconjunto específico de imágenes
    - Se mantuvieron los criterios de clasificación estandarizados mencionados anteriormente
    - El trabajo se realizó de manera asíncrona para optimizar tiempos
2. **Consolidación de Datos:**
    - Se estableció un Google Drive compartido como punto central de almacenamiento
    - Cada colaborador incorporó sus imágenes clasificadas en carpetas designadas
    - Se mantuvo una estructura organizada para facilitar la integración

Este enfoque colaborativo permitió la creación eficiente de un dataset unificado y coherente, listo para su uso en la fase de modelado.

2.3 Dia y noche

Para este punto los datos están divididos adecuadamente por cantidad de vacas.
Generamos un programa que organiza el conjunto de imágenes basándose en el momento del día (día/noche) determinado por la hora en el nombre del archivo. Procesa imágenes de los conjuntos de entrenamiento, validación y prueba, y las redistribuye en una nueva estructura de directorios manteniendo su división original pero agregando una subcategoría de tiempo (día/noche). El programa considera como "día" las imágenes tomadas entre las 6:00 y 17:59 horas, y como "noche" las demás. El proceso se realiza de manera paralela utilizando múltiples hilos para mejorar la eficiencia, y mantiene la estructura de clasificación original dentro de cada categoría de tiempo.

2.4 Limpieza

A pesar de los resultados obtenidos, evaluamos que debemos continuar trabajando con la misma limpieza que antes. La eliminación de outliers y la filtración de imágenes con condiciones extremas de iluminación pueden ser una dificultad. Sin embargo, tras un análisis, se tomó la decisión estratégica de mantener estas imágenes en el dataset.

Esta decisión se fundamenta en varios aspectos clave:

1. **Representatividad de Condiciones Reales:**

- El sistema final operará en un entorno real donde estas condiciones "no ideales" son frecuentes
- La variabilidad en iluminación y calidad de imagen son inherentes al contexto operativo

2. **Objetivo de Generalización:**
    - El propósito fundamental es desarrollar un modelo capaz de determinar la cantidad de vacas bajo cualquier condición de iluminación
    - La eliminación selectiva de datos podría crear una brecha entre el rendimiento en entrenamiento y el rendimiento en producción
3. **Robustez del Modelo:**
    - La exposición a condiciones desafiantes durante el entrenamiento debería fortalecer la capacidad de adaptación del modelo
    - La inclusión de casos extremos puede mejorar la robustez general del sistema

Por lo tanto, se optó por mantener la integridad del dataset original, incluyendo aquellas imágenes que podrían considerarse outliers o subóptimas. Esta decisión, aunque podría afectar las métricas de rendimiento durante el entrenamiento, está alineada con el objetivo final de desarrollar un sistema confiable y práctico para su implementación en condiciones reales de operación.