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

Con base a los resultados obtenidos en la iteración anterior, consideramos que debíamos realizar cambios a los datos para lograr obtener mejores resultados del modelo. En este documento, se hablará a detalle de la preparación de los nuevos datos para el desarrollo de una nueva versión del modelo.

1.1 Trasfondo

Durante la fase de modelado (modeling II, III y IV), a pesar de los cambios en el tamaño del dataset, las arquitecturas utilizadas y los ajustes realizados a los parámetros de los modelos, no logramos obtener resultados satisfactorios. Por lo que fue necesario buscar otro enfoque para tratar de obtener un desempeño decente. Como solución, se cambió el etiquetado del dataset, pasando de ser clasificación a segmentación, utilizando Bounding - Boxes para marcar la ubicación exacta de cada vaca en la imagen por medio de coordenadas y de esta forma contar la cantidad de vacas en la imagen.

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.

No requerimos crear atributos derivados, realizar transformaciones complejas o integrar múltiples fuentes de datos como sugiere CRISP-DM. Esto se debe a las decisiones tomadas para usar las imagenes en sus condiciones reales.

1.3 Criterios

Para definir cómo prepararemos los datos debemos de tener presente nuestros objetivos 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.

2.0 Dataset y Limpieza

Para preparar los datos hemos cambiado la forma del etiquetado de las imágenes, cambiando de clasificación a segmentación, utilizando las 8,000 imágenes originales del dataset creado en la primera iteración de preparación de los datos .

2.1 Segmentación

Se reclasificaron las imágenes cambiando del conteo de número de vacas por imagen, a marcar con recuadros, en dónde se encuentran las vacas en cada imagen. En la preparación original de los datos, colaboramos con otro equipo para realizar el etiquetado, ambos equipos etiquetando para clasificación y segmentación, teniendo como resultado dos Datasets disponibles para ambos equipos. En las iteraciones pasadas, estábamos utilizando el primer dataset (clasificación), pero los resultados de los modelos no fueron eficientes, por lo que cambiamos al segundo dataset (segmentación).

El proceso de segmentación se llevó a cabo utilizando Roboflow, una herramienta especializada para el etiquetado y preparación de datasets en visión por computadora. El objetivo principal fue segmentar las imágenes, por medio de Bounding - Boxes, encerrando en un recuadro a cada vaca presente en la imagen de la fila de espera para ordeña en un momento específico del día.

Para asegurar la consistencia, 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)

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
    - 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 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.