Data Understanding

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

Tecnológico de Monterrey Campus Querétaro, México

A. Recolección de datos inicial

En este proyecto, los datos principales consisten en **imágenes aéreas ortogonales** tomadas en un rancho ganadero, cuyo propósito es monitorear el comportamiento de las vacas y su tiempo de espera en la fila de ordeño.

- **Ubicación de los datos**: Las imágenes están almacenadas en Google Drive en formato JPG, organizadas por nombre de archivo que incluye la fecha y la hora de captura.
- **Métodos de adquisición**: Las imágenes fueron capturadas a través de cámaras instaladas en el techo del rancho.
- **Problemas para adquirir los datos**: Las imágenes no están etiquetadas ni organizadas por la cantidad de vacas en cada foto. Además, la calidad de las imágenes nocturnas puede dificultar el análisis visual. Para resolver esto, se identificó la cantidad de vacas que se encuentran en cada foto dentro de los datos y se clasificaron y agruparon las fotos por la cantidad de vacas ubicadas.

B. Descripción de los datos

- **Formato de los datos**: Imágenes en formato JPG, con nombres de archivo que contienen la fecha y hora exacta de captura.
- **Cantidad de datos**: El dataset consta de aproximadamente 8115 fotos, cada una representando distintos momentos del día, capturadas en intervalos de cinco minutos. Lo que equivale a 56,4 días de captura.
- **Características descubiertas**: Existen imágenes con demasiado ruido que dificulta la visualización de las vacas. Puede que sean imágenes con una oscuridad o luminosidad bastante alta.
- **Cualidad intrínseca de los datos:** Cada imagen mide 1920 por 1080 píxeles, pesando unos 0.23 MB, y tienen una profundidad de color (RGB) de 24 bits.

C. Exploración de los datos

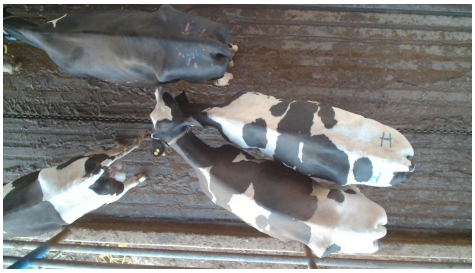
Las condiciones del establo produjeron fotos casi completamente obscuras.

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.

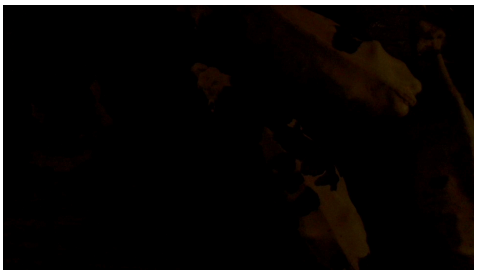
- **Técnicas de consulta**: Ocuparemos la plataforma Roboflow con el objetivo de explotar las imágenes y empezar las transformaciones pertinentes para la construcción de la dataset.
    - **Separación de las imágenes**: Son 8115 imágenes, separamos estas en batches de 811 imágenes, el split nos genera 10 batches, con un batch con 812 imágenes.
    - **Pre-análisis de batch**: Analizaremos con roboflow cada una de los batches. En el pre análisis ocuparemos la herramienta de bounding-box AI, que subraya el etiquetado que queremos, que en este caso vaca. 4
    - **Correcciones batch**: Analizaremos las anotaciones que se hizo en el pre-análisis, para ello, visualizamos batch por batch e imagen por imagen. Para mayor correcciones ajustaremos los parámetros de iluminación y contraste cuando se requiera, a su vez, modificar las anotaciones que hizo la AI, y guardar la imagen con su respectivo batch.
    - **Construcción dataset**: Agrupamos los batches en un solo dataset para procesarlo en código para hacer una separación del conjunto de prueba, validación y entrenamiento.
    - **Descubrimientos**: Al analizar las imágenes en la noche no se alcanza a diferenciar las vacas, además de tener imágenes que el pre-análisis con AI subraya un conjunto de vaca como una supervaca.
    - **Impacto en el proyecto:** Buscamos construir un modelo que haga un conteo de vacas en cualquier instante del día con precisión mayor al 75% para resaltar los cuellos de botella, no obstante, planteamos hacer un solo modelo, pero las imágenes de noche introducirán ruido al modelo por la variedad de iluminación, por ende, construiremos dos modelos, uno que esté funcionando en el día y uno en la noche.

Visualización de Datos



Esta primera imagen muestra una situación de iluminación clara, donde se puede identificar fácilmente las vacas. Este tipo de imagen será fundamental para el entrenamiento del modelo.

:



Las dos imágenes nocturnas representan escenas con poca luz. Se demuestra claramente el problema de falta de visibilidad, lo que dificulta la correcta identificación de las vacas. Las vacas se vuelven difíciles de distinguir, lo que confirma la necesidad de preprocesamiento o un modelo especializado para trabajar con este tipo de imágenes.

D. Calidad de los datos

- **Datos completos**: Durante el proceso de captura de imágenes, se observaron ciertas omisiones debido a fallas eléctricas que ocurrieron durante el día y la noche. Estas interrupciones en el suministro eléctrico afectaron la operación de la cámara, lo que provocó que se omitieran ciertos lapsos de tiempo sin tomar ninguna foto. Como resultado, el dataset no está completamente cubierto en todos los periodos del día que inicialmente se esperaban.
- **Datos correctos:** Las imágenes dentro del dataset a utilizar son las correctas, no existen datos o imágenes corruptos o imágenes duplicadas que puedan interferir con el entrenamiento del modelo. Las fotos representan cierta continuidad temporal del día y noche dentro de la fila de espera.
- **Errores comunes:** Existen imágenes con cierta luminosidad o aparecimiento de sombras las cuales pueden generar cierto error en las características esperadas de una imagen dentro del dataset.
- **Falta de datos:** Imágenes del dataset las cuales fueron tomadas de noche, generan cierta desinformación y falta de datos gracias a la falta de luz, con lo que genera ciertas complicaciones en el entrenamiento del modelo con este conjunto de datos. Además de periodos de corte de luz en el rancho, por lo que no existen fotos de ciertos momentos en el tiempo.

Posibles soluciones:

- Aplicar técnicas de mejora de imagen como el ajuste de brillo, contraste o el uso de filtros para reducir las sombras y estandarizar la luminosidad en las imágenes.
- Uso de técnicas de aumentación de datos que simulan variaciones de luz, contraste y sombras, lo que permite aprender a manejar mejor las variaciones.
- Implementar un proceso de preselección de las imágenes para filtrar aquellas con sombras o iluminación extremas que podrían afectar el rendimiento del modelo. Estas imágenes se pueden tratar por separado.
- Entrenar modelo específico para imágenes nocturnas, usando técnicas de preprocesamiento avanzadas para mejorar la calidad de imagen, y otro modelo para las imágenes diurnas.