Deployment

Adrián Matute, Pablo Martínez, Osvaldo Del Valle, Andres Callealta, Jorge Martínez
Tecnológico de Monterrey Campus Querétaro, México

Plan de entrega

Opciones para la implantación del modelo híbrido (un solo modelo)

- Instalación por parte del equipo
    - Se agendará una visita al CAETEC donde el equipo de trabajo realizará la instalación del sistema en el entorno de producción. Posteriormente se dará una capacitación al personal y usuarios finales para el uso del sistema.

- Instalación por parte del cliente
    - Se entregará un repositorio de github con los modelos y el sistema de gestión. Estos incluirán:
        - Manual de instalación: para la que el sistema pueda ser instalado por parte del cliente.
        - Manual de usuario: Describe el funcionamiento y operación del sistema para que los usuarios finales puedan hacer uso de este.
        - Manual técnico: Descripción detallada de todos los componentes del sistema para su mantenimiento y/o trabajo futuro.

De las dos opciones consideradas para la entrega del modelo, aplicaremos la segunda (Instalación por el cliente). Esto vendrá con su respectivo manual y scripts para poder montar el sistema automáticamente.

Plan de monitoreo y de mantenimiento

El objetivo es asegurar que el sistema de detección de vacas y monitoreo de aglomeraciones continúe siendo preciso y relevante para optimizar la producción de leche en CAETEC.
Los encargados de operar el sistema considerarán los siguientes puntos y en caso de que no se cumplan se podrán tomar una serie de decisiones:

Estrategia de monitoreo

1. Definir métricas clave para monitoreo continuo:
    a. Precisión de modelo en condiciones de día y noche.
    b. Frecuencia de detecciones incorrectas (falsos positivos/negativos).
    c. Tiempos promedios de respuesta para identificar aglomeraciones.
    d. Impacto en la optimización de las filas de ordeño.
2. Establecer un sistema de alertas:
    a. Alertas automáticas cuando la precisión del modelo caiga por debajo del umbral de 85%.
    b. Notificaciones si las aglomeraciones persisten durante períodos grandes de tiempo.
3. Implementar pruebas periódicas:
    a. Comparación de datos reales de conteo contra las predicciones del modelo.

Mantenimiento del sistema

1. Actualización del modelo:
    a. Añadir imágenes de las diferentes estaciones del año para garantizar adaptabilidad
2. Revisión de hardware y software:
    a. Verificar que las cámaras se encuentren en buenas condiciones.
3. Documentación:
    a. Mantener un registro de todas las actualizaciones, problemas y soluciones implementadas.
    b. Generar reportes periódicos con recomendaciones de acuerdo a los datos recolectados.

Criterios para dejar de usar el sistema

1. La precisión del modelo cae de forma constante por debajo del umbral definido y no es recuperable con reentrenamientos.
2. El sistema no logra identificar de manera correcta aglomeraciones de vacas en diferentes momentos del día.
3. Cambios en los objetivos del negocio o en la estructura del proceso de ordeño.

Acciones ante fallas

1. Si el sistema no puede seguir usándose:
    a. Revisar los datos y realizar un reentrenamiento.
    b. Proponer nuevas arquitecturas de modelos.
    c. En caso de tener fallas críticas, iniciar un nuevo proyecto enfocado en cubrir los aspectos con deficiencias.
2. Si los datos recolectados ya no reflejan condiciones actuales:
    a. Realizar una nueva recolección de datos para ajustar el modelo.

Dado que la meta final del proyecto, es ayudar al CAETEC a mejorar la producción de leche, es importante hacer una revisión constante del sistema para saber si los datos generados ayudan a evitar aglomeraciones. Además, es importante determinar si los beneficios económicos provenientes del sistema justifican que se le dé un mantenimiento de manera continua.

Reporte de producto final

Resultados obtenidos:
El proyecto de detección y conteo de vacas en la fila de ordeño, utilizando la arquitectura de YOLOv8, logró resultados sobresalientes en términos de precisión y desempeño. Las métricas principales para cada modelo desarrollado y aprobado son las siguientes:

- ***Modelo diurno:*** Precisión del 96.39% en condiciones de luz natural.
- ***Modelo nocturno:*** Precisión del 92.9% en ambientes con poca iluminación.

- ***Modelo híbrido (seleccionado como solución final):***

Métrica	Valor
Recall	95%
mAP50	96%
mAP@50-95	84%

Estas métricas reflejan un modelo sólido y adaptable para monitorear el flujo de vacas en la fila de ordeño bajo diferentes condiciones de iluminación, cumpliendo con los objetivos establecidos.

Descripción del Proceso y Costos
El desarrollo del proyecto siguió la metodología CRISP-DM, asegurando un enfoque estructurado:

1. Preparación y limpieza de datos:
    - Se identificaron problemas críticos en el dataset, como imágenes demasiado oscuras o sobreexpuestas, que fueron eliminadas para evitar impactos negativos en el entrenamiento del modelo.
    - Se implementaron técnicas de mejora para aumentar la calidad de los datos disponibles.

2. Selección y evaluación de modelos:
    - Se investigaron y probaron arquitecturas como YOLOv8 y Faster R-CNN, lo que requirió un tiempo significativo para implementar y evaluar las alternativas.
    - Finalmente, se seleccionó YOLOv8 híbrido por su robustez, precisión y capacidad de generalización en condiciones mixtas.

3. Evaluación y ajuste del modelo:
    - Se probaron diversos modelos separados para día y noche, antes de optar por el modelo híbrido, ahorrando tiempo y recursos computacionales.

4. Planes de despliegue y mantenimiento:
    - Se desarrolló un plan detallado de implementación, que incluye la entrega del repositorio en GitHub con el código completo y un plan de monitoreo y mantenimiento para garantizar el uso continuo y efectivo del sistema.

La descripción del proceso incluyendo los costos:
Para llevar a cabo este proyecto, el equipo ha seguido la metodología CRISP-PM. Hemos empezado tratando los datos, preparándonos para los modelos y comprendiendo los, viendo que es lo que podría causar problemas, borrando los outliers (imágenes muy oscuras o muy claras). Luego hemos investigado cuáles son los modelos que podrían ser los más adecuados para nuestro problema, lo cual nos tomó bastante tiempo, causando gran pérdida de tiempo y energía, ya que pasamos bastante tiempo tratando de implementar dos modelos por separado, gastando unidades de cálculo. En este proyecto, los costos se cuentan en tiempo, ya que no hemos pagado para acceder a recursos tipo GPUs. Al haber tratado de entrenar el modelo híbrido y los dos modelos con YOLO v8, y Fast R-CNN, hemos procedido a la evaluación de cada uno de ellos, para llegar a la conclusión que el modelo híbrido de YOLO v8 nano era el más robusto. Para terminar con el proyecto según la metodología, hemos finalmente diseñado un plan de despliegue, que va de la entrega del repositorio GitHub del proyecto, con todo el código, hasta el plan de monitoreo y mantenimiento.

Desviaciones del Plan Original
A lo largo del proyecto, surgieron varias desviaciones importantes:

- **Identificación individual de vacas**: Inicialmente, se planteó la posibilidad de identificar cada vaca individualmente y rastrear su tiempo de espera. Sin embargo, esta idea se descartó por la falta de un dataset etiquetado a ese nivel de detalle y por los recursos necesarios para entrenar un modelo más avanzado.
- **Separación de modelos por condiciones de iluminación**: Aunque inicialmente se planearon modelos separados para día y noche, finalmente se optó por un modelo híbrido más eficiente y generalizable.
- **Diseño del dashboard**: Gracias al contacto frecuente con el cliente, el diseño del panel fue ajustado continuamente con base en comentarios, logrando una interfaz más funcional y alineada con las necesidades de CAETEC.

Planes de Implementación

El modelo final será implementado con los siguientes elementos clave:

- Entrega de un repositorio completo en GitHub con el código documentado para facilitar su uso y mantenimiento.
- Implementación de un dashboard interactivo que permita visualizar métricas clave, como conteo de vacas y posibles aglomeraciones en tiempo real.
- Integración de un plan de monitoreo y mantenimiento para garantizar que el modelo se ajuste a los cambios en el entorno operativo, como nuevas condiciones de iluminación o configuraciones en la fila de ordeño.

Recomendaciones para Trabajo Futuro
Para mejorar y extender el alcance del proyecto, se sugieren las siguientes acciones:

- Implementar técnicas avanzadas de data augmentation, como ajustes en el contraste de imágenes, para mejorar la capacidad del modelo en condiciones de iluminación variables.
- Experimentar con arquitecturas más complejas, como versiones avanzadas de YOLOv8 (small o medium), para explorar mejoras adicionales en precisión y capacidad de generalización.
- Obtener acceso a recursos computacionales más robustos, como GPUs de mayor potencia, para entrenar modelos más complejos y realizar ajustes finos con mayor eficiencia.

Review del proyecto

Documentación de la experiencia

En un proyecto como este donde se tiene el enfoque en la detección de vacas y prevención de aglomeraciones para optimizar la producción lechera, hemos tenido diferentes momentos a lo largo del desarrollo de la solución los cuales han sido trampas potenciales que han afectado a cada fase del proyecto.

Desde el inicio, uno de los principales riesgos que tuvimos durante el entendimiento del negocio fue definir de forma incompleta el problema y los objetivos de negocio que buscamos impactar. Si no se comprende con claridad cómo las aglomeraciones impactan la producción de leche, el proyecto podría centrarse en aspectos secundarios.

En la fase de entendimiento, recolección y preparación de los datos también tuvimos ciertos problemas. Imágenes con mala iluminación llegaron a causar una limitación del desempeño del modelo, por eso se tomó la decisión de eliminar esas imágenes con valores extremos de iluminación ya sean altos o bajos. Por otro lado, también tuvimos cierta problemática en el etiquetado de las imágenes, ya que en una primera iteración de esta tarea, los miembros del equipo sin darnos cuenta etiquetamos las mismas imágenes, por lo que tuvimos que hacerlo de nuevo ya de manera correcta, cada quien con imágenes diferentes.

Durante la etapa de modelado, el seleccionar un modelo inadecuado para los recursos tecnológicos disponibles o para la complejidad de la solución a la que queremos llegar, también puede llevar a resultados desfavorables. Es por ello que probamos con distintos modelos y configuraciones, para evitar la confianza en un solo modelo para no comprometer el desempeño del sistema.

Enfoque engañoso

Entrenar dos modelos para un mismo problema fue un error estratégico que incrementó la complejidad del proyecto sin proporcionar beneficios sustanciales. En el futuro, es preferible abordar problemas similares con un único modelo robusto y bien optimizado, utilizando técnicas como el aumento de datos, transferencia de aprendizaje y ajustes hiper paramétricos para abordar las variaciones en los datos (como las condiciones diurnas y nocturnas). Esto reduce la duplicación de esfuerzos y facilita tanto la implementación como el mantenimiento del sistema.

Lecciones Aprendidas y Recomendaciones para Proyectos Similares

En el desarrollo de un proyecto de minería de datos como este, es fundamental documentar las mejores prácticas y los aprendizajes obtenidos para facilitar la selección de técnicas adecuadas en condiciones similares en el futuro. Estas recomendaciones, basadas en la experiencia del proyecto, proporcionan una guía clara para enfrentar problemas análogos de manera eficiente y efectiva:

- No tratar de separar lo datos para hacer varios modelos, pero más bien tratar de entender mejor los datos, preparandolos mejor, usando técnicas de data augmentation y modification (por ejemplo cambiar el contraste de las imágenes que sean más oscuras que un cierto límite) para aumentar la varianza y que el modelo pueda más fácilmente entender patrones.
- Pasar más tiempo entrenando otros modelos con otras arquitectura, encontrar el mejor modelo para un cierto problema es intentar otras arquitecturas, y solamente luego mejorarlo.
- En caso de que sea un proyecto importante, en donde la precisión es importante, obtener recursos más grandes, GPUs con mayor potencia, para poder tener tiempo de entrenar modelos más complejos y poder hacer más ajustes.

Participación de cada miembro del equipo:

Adrián Matute:

Durante el proyecto, fui elegido como líder del equipo, lo que me permitió tener una visión general de todo el proceso y coordinar las tareas entre los miembros. Tuve una participación constante y activa desde las primeras fases. En el entendimiento del negocio y de los datos, trabajé junto al equipo para definir los objetivos del proyecto, tanto a nivel de negocio como en términos de minería de datos, asegurándonos de que nuestras metas estuvieran alineadas con las necesidades de CAETEC. También colaboré en la elaboración y seguimiento del plan del proyecto, y participé en la primera visita al CAETEC para comprender mejor el contexto y las necesidades de la organización.

En la fase de preparación de datos, mi participación fue clave en varias tareas. A mí y a mis compañeros nos tocó etiquetar aproximadamente 800 fotos, pero debido a un error en la asignación de imágenes, terminé etiquetando el doble de lo planeado. Además, me encargué de limpiar el dataset eliminando imágenes que tenían problemas de luminosidad, ya fueran demasiado claras o demasiado oscuras, lo que podría haber afectado el rendimiento del modelo. También ayudé a separar las imágenes en dos conjuntos: uno para el día y otro para la noche, basado en los nombres de los archivos que indican la hora de captura. Para hacer esto, definimos un umbral de luminosidad que nos permitió clasificar las imágenes de manera más precisa, lo cual resultó ser fundamental para mejorar la calidad del dataset y facilitar el proceso de entrenamiento.

Esta fase fue crucial, ya que aseguramos que los datos con los que trabajamos fueran lo más limpios y organizados posible, lo que nos permitió avanzar con los modelos de manera más efectiva.

En la fase de modelado, me encargué de monitorear y entrenar el modelo de YOLOv8, específicamente el que se utilizaría para las imágenes con buena iluminación, es decir, las tomadas durante el día en el rancho. Además de este, entrené por separado otro modelo, el Faster R-CNN, para evaluar su rendimiento. Sin embargo, este último no logró los resultados esperados en términos de efectividad para abordar la problemática y cumplir con los objetivos que teníamos. Fue un proceso interesante, pero los resultados no fueron los más adecuados para el contexto del proyecto.

A lo largo de estas semanas de trabajo, estuve presente en casi todas las clases y formé parte activa de las reuniones con el socio formador, lo que me permitió mantenerme alineado con los avances del proyecto y con las expectativas del CAETEC.

En cuanto a la evaluación del modelo, decidí realizar una segunda iteración con Faster R-CNN, con el fin de mejorar los resultados. Sin embargo, a pesar de los ajustes realizados, los resultados seguían sin ser los más efectivos en comparación con otros enfoques. Fue una lección importante sobre la necesidad de probar diferentes modelos y ajustar sus parámetros para encontrar la mejor solución.

Finalmente, en la fase de deployment, me encargué de desarrollar el plan de monitoreo y mantenimiento del modelo. Este plan es clave para garantizar que el modelo se mantenga funcional y preciso en el tiempo. También apoyé en la redacción del reporte final, en el que se detallaron los objetivos alcanzados en la fase de "business understanding" y se definieron los próximos pasos para el proyecto, incluyendo posibles mejoras tanto en los modelos como en la metodología de trabajo.

Jorge Martínez López:

A lo largo del proyecto, mi rol en el equipo fue de soporte, ayudar al equipo a evitar contratiempos y tener el proyecto en tiempo y forma, esto conlleva a ver participado poder ayudar al equipo en lo que se pueda y ser el integrador del equipo.

Podré decir que tuve un impacto significativo en el proyecto debido a que aporte en el desarrollar de la mayoría de las fases:

Business Understanding, me encargué de señalizar la metodología que se occupará en el proyecto, además me dediqué al refinamiento del documento, elementos de corrección, formato, agregar información escencial.

Data Understanding , tuve la oportunidad de escribir la parte de exploración de los datos, hacer los análisis pertinentes para saber las características de los datos como definir los outliers.

Data Preparation, en esta fase describí la técnica ocupada para hacer la limpieza de datos, y encontrar los outlier, con ayudada de matute, generamos un script que separa los outliers y genere un folder con los datos limpios.

Modelado, escribí la pre liminar de la documentación de la construcción del modelos, no obstante, debido a los correcciones se generaron otros documentos que ocuparon la base de lo había escrito y de ahí refinaron los parrafos.

Evaluation, tuve poca intervención o fue nula, debido a que estuve construyendo el readme de github.

Deployment, mi participación fue generar el reporte ejecutivo que se le entregará a Arturo, a su vez diseñé la presentación.

Fui responsable de desempeñar como intermediario con el socio formador, presentando los avances del proyecto y asegurando una comunicación efectiva. Dentro de mis responsabilidades, diseñé la carta de validación de objetivos de negocio, así como la carta de avisos de privacidad, para lo cual realicé una investigación exhaustiva sobre los regímenes fiscales aplicables al manejo de información de terceros. Además, gestioné y mantuve actualizada la bitácora de avisos y privacidad, garantizando la integridad de los registros.

Durante el desarrollo, enfrentamos problemas que retrasaron la implementación de los modelos, específicamente debido a un mal etiquetado de las vacas en el dataset, lo que detuvo los avances por varios días. En ese periodo, intenté entrenar un modelo utilizando la arquitectura Faster R-CNN con un dataset similar disponible en Kaggle, pero los resultados no fueron favorables debido a la falta de de datos en el dataset. Posteriormente, mi compañero Matute retomó este modelo, logrando avances significativos al trabajar con un dataset completo y depurado.

Enseñanzas y aprendizajes clave:

Uno de los mayores aprendizajes fue la importancia de la gestión efectiva del equipo, en la que cada integrante desempeñó un rol especializado. En mi caso, me encargué de garantizar la integridad del proyecto, evitando la exposición de información sensible en plataformas como GitHub o en las carpetas compartidas de Drive. También validé y promoví las buenas prácticas en el manejo de datos. Además, este proyecto me permitió dominar una metodología que considero fundamental para el desarrollo de aplicaciones tecnológicas en la industria. Este aprendizaje me posiciona con las habilidades necesarias para actuar como project manager en futuros desarrollos tecnológicos.

Pablo Martínez:
Durante el proyecto me desempeñé en varios aspectos como realizar el plan de mitigación de riesgos para considerar eventos que pudieran afectar la realización exitosa del proyecto. Igualmente tuvo un papel en la propuesta de la arquitectura para implementar la solución y la modelación de los diagramas de despliegue.

Al momento de preparar y etiquetar los datos colaboré con el etiquetado de mi porción del dataset. Posteriormente junté el dataset de bounding boxes y verifiqué la calidad y el orden de este, sometiendo revisiones en errores de formato que fueron presentados. Trabajé igualmente en el reporte donde se documenta este proceso (data preparation).

Para nuestro primer modelado realicé los pasos de separación en conjuntos de entrenamiento y validación así como el entrenamiento del modelo híbrido. Dejando este proceso documentado y de una forma reproducible para compartirlo posteriormente y usarlo en entrenamientos futuros. Con el objetivo de presentar resultados progresivamente al socio use herramientas para la generación de videos con las inferencias generadas por el modelo.

Al momento del despliegue me encargué de programar el script que genera las inferencias del modelo y se comunica con el resto de sistemas para el almacenamiento de las inferencias. Para esto tuve comunicación con el socio para entender la forma en la que funcionan los sistemas actuales, fotos tomadas cada 5 minutos, para brindar una solución que trabaje de la mano con sus scripts actuales y tome en cuenta el ambiente de producción.

Trabajé en modularizar el servidor encargado de guardar y consultar los datos. Esto con la intención de que pudiera ser escalable en caso de que se quiera extender el proyecto. Esta modularización también surgió con la intención de darle al socio la opción de trabajar con una base de datos local ligera (sqlite) o usar un proveedor para acceder a los datos de forma global. Asimismo, desarrollé un script de instalación para montar todos los sistemas en el hardware objetivo.

Osvaldo Del Valle:
Durante el desarrollo de este proyecto, tuve participación en diferentes aspectos del proyecto, primeramente colabore en la creación del plan de trabajo, después era encargado de actualizar este documento dándole seguimiento a las actividades planeadas, esto fue importante ya que el equipo pudo tener un orden con prioridades para saber que se debería de hacer a continuación así como conocer el estatus del proyecto en ese momento. Además ayudé a la creación del dataset las dos veces que fue requerido, lo cual fue clave ya que era la base de nuestro modelo.

Hablando de la parte de reportes y metodología, tuve participación en algunos documentos como el reporte de resultados e instalación para el socio formador, en el reporte 1 de modelado y en la escritura de la bitácora de privacidad y seguridad.

Ahora bien, en el área donde considero que más me desempeñe fue en el desarrollo orientado a cómo se iban a mostrar los datos a los interesados, para ello se creó una vista a manera de dashboard, donde se puedo observar información sobre el comportamiento de las vacas a través de un dia y esto mismo pero a través de varios días. Esto fue realmente importante ya que nos ayudó a tener una forma atractiva de mostrar resultados y funcionamiento al cliente, al mismo tiempo que a nosotros nos ayudaba para obtener hallazgos en base a lo que se observa analizando la información. Además de ayudar en la creación de los elementos necesarios para el despliegue de nuestro proyecto en producción, lo que implicó la creación de scripts que instalen el proyecto y unan desde la parte de él tomar una foto, realizar predicciones, almacenar en una base de datos hasta mostrar información en el dashboard. Adicionalmente fui encargado de la creación y gestión de la base de datos remota que se utilizará para guardar datos de las predicciones así como el registro del acceso a la base de datos, lo

cual fue indispensable para la parte de la bitácora de privacidad.

Considero que en el desarrollo de este proyecto he adquirido conocimientos que me ayuda a gestionar correctamente un proyecto de inteligencia artificial, porque entiendo cuales son las fases de CRISP-DM y lo más importante en que ayuda cada una de estas fases, en esto algo que pude darme cuenta es cómo cada fase por más corta que sea comparada a la siguiente es importante y de verdad ayuda a un progreso más eficiente, por ejemplo la construcción del dashboard de información fue más sencillo en cuanto a definir qué cosas eran importantes de considerar, esto gracias a que sabíamos claramente la forma en que el CAETEC trabaja así como los datos que tenemos y en base a nuestros objetivos propuestos. Finalmente considero que puedo poner en práctica ciertas partes de la metodología en diferentes áreas, y que esto es de valor.

Andres Callealta:

Además de haber puesto en práctica la metodología CRISP-PM a lo largo del proyecto, tuve la oportunidad de escribir partes de todos los proyectos, lo que me permitió entender más en profundidad la metodología, comprendiendo y dándome la cuenta de la importancia que es de seguir un “plano” detallado, sobre todo en proyectos grandes como este, en los cuales hay muchas diferentes fases diferentes, donde el tiempo cuenta mucho, un pequeño retraso en una fase impide a veces poder pasar a la siguiente, tanto al nivel personal que grupal. Me di cuenta de la importancia de hacer las cosas con un orden predefinido, y no lanzarme directamente los ojos cerrados a las partes del proyecto que me gustan más. El proceso aplicado es sobre todo útil en proyectos no personales, donde el cliente interviene especificando sus necesidades o sugerencias. A nivel personal, la parte que más me gusto y donde más aprendí, fue entrenar el modelo nocturno, comprendiendo los resultados y métricas para poder volver a iterar, aplicando técnicas de regularización, y todo esto en un contexto de big data, donde uno no puede permitirse múltiples iteraciones ya que el set de datos es muy grande y lanzar una época gasta mucho tiempo y recursos. Para terminar, al principio del proyecto me di cuenta de la amplitud de él, sobre todo preparando, etiquetando las imágenes, y preparando los diferentes sets de datos, training ,validación y test, lo cual tomó mucho más tiempo que lo que pensaba.