



Tecnológico de Monterrey

BimboNET

Avance 7. Resumen Ejecutivo



Análisis de Ataques BimboNet

Equipo 14

Integrantes del equipo:

Giancarlo Franco Carrillo A01638108

Luis Fernando Rivera Albarrán A01209133

Brandon Vladimir Montoya López A01209115

Resumen Ejecutivo

El presente trabajo detalla el diseño, desarrollo y evaluación de un sistema de Visión por Computadora para la segmentación de instancias del producto "Mantecadas" en entornos de retail. El objetivo de negocio principal es la verificación de la presencia o ausencia del producto en anaquel, una tarea crítica para la gestión de la cadena de suministro y la estrategia de ventas. Ante el desafío de la recolección masiva de datos etiquetados, se implementó una novedosa metodología iterativa, comenzando con un pipeline de generación de datos semi-automático (SAM+OCR) y culminando en un ciclo de etiquetado asistido por IA sobre un dataset de más de 3,400 imágenes. Los hallazgos demostraron de forma concluyente que la calidad, especificidad y escala de los datos curados manualmente eran el factor determinante para el éxito, por encima de la simple escala de la arquitectura del modelo. El modelo definitivo, un YOLOv8m-seg entrenado sobre el dataset final verificado por humanos, alcanzó un rendimiento de nivel de producción de 79.0% mAP50-95, validando la estrategia data-céntrica. Se concluye con una discusión sobre la viabilidad del modelo para su implementación, una hoja de ruta para su mejora continua, y un análisis estratégico para su despliegue en una infraestructura de nube.

1. El Desafío Estratégico: La Incertidumbre en el "Último Metro"

Para una empresa de consumo masivo como Grupo Bimbo, el "último metro" de la cadena de suministro —el anaquel en el punto de venta (PDV)— es el momento de la verdad. La falta de visibilidad a esta escala, multiplicada por miles de tiendas, genera una incertidumbre costosa: ¿Está el producto en el anaquel? ¿Se ha agotado? ¿Está exhibido correctamente? Tradicionalmente, responder a estas preguntas ha dependido de auditorías manuales, un proceso lento, caro y con un margen de error humano considerable. La capacidad de automatizar esta verificación no solo optimiza la gestión de inventarios, sino que también proporciona datos valiosos para la estrategia de ventas y marketing.

El objetivo de este proyecto fue resolver este desafío mediante el desarrollo de un sistema de IA capaz de verificar automáticamente la presencia o ausencia del producto "Mantecadas" en las imágenes enviadas por los usuarios de la plataforma T-Conecta, sentando las bases para una auditoría de anaquel escalable y en tiempo real. Se optó por un enfoque de segmentación de instancias para proporcionar una prueba visual irrefutable de cada detección, generando así un alto grado de confianza y abriendo la puerta a análisis más sofisticados como el conteo de unidades y la diferenciación de variantes del producto.

2. Hallazgos y Metodología: Un Viaje Centrado en los Datos

El proyecto se concibió como un viaje iterativo, donde cada paso informaba al siguiente, siguiendo una filosofía de IA centrada en los datos.

- **2.1. El Desafío del "Lienzo en Blanco" y el Experimento con Datos Sintéticos** El proyecto se vio inicialmente obstaculizado por la falta de datos reales (antes del 15 de mayo), lo que obligó a una fase de experimentación con datos sintéticos generados en BlenderProc. El intento de usar clasificadores convencionales (ResNet18, etc.) en este set fracasó, evidenciando una brecha insalvable entre el entorno simulado y la complejidad del mundo real. Esta etapa aportó una lección crítica: la necesidad de datos reales y modelos más sofisticados.



Ilustración 1 Simulación en Blenderproc

Se entrenaron varios modelos Yolo11 con el dataset sintético y los resultados obtenidos por la matriz de confusión muestran como muchas de las predicciones se están interpretando como el fondo de la escena, dando resultados muy poco fiables.

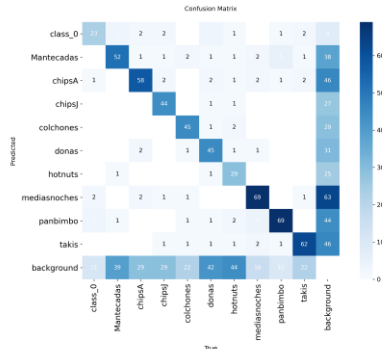
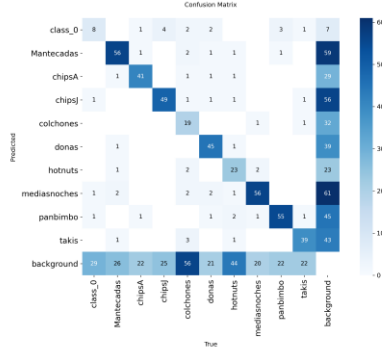
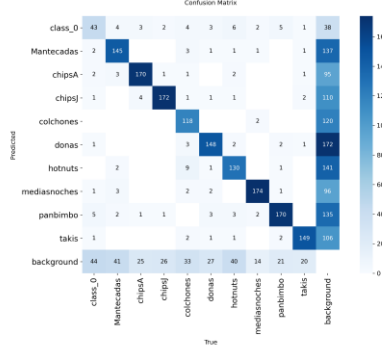
Modelo y parámetros

Yolo11n
Detección
150 imágenes
~250-350 instancias x clase
100 épocas

Matriz de confusión

Confusion Matrix

| | class_0 | Mantecadas | chipsA | chipsJ | colchones | donas | hotnuts | mediasnoches | panbimbo | takis | background |
|--------------|---------|------------|--------|--------|-----------|-------|---------|--------------|----------|-------|------------|
| class_0 | 75 | 2 | | | | | | | | | |
| Mantecadas | 90 | 2 | 1 | | | 2 | | | | 1 | 50 |
| chipsA | 1 | 50 | | | | | 1 | | | | 55 |
| chipsJ | 1 | 1 | 45 | | 1 | 1 | | | | | 44 |
| colchones | | | | 42 | 1 | 1 | 1 | | | 1 | 38 |
| donas | | 1 | 1 | 2 | 50 | | | 1 | | | 49 |
| hotnuts | | 2 | 1 | | 1 | | 37 | | | 2 | 40 |
| mediasnoches | | | 1 | | 1 | | 1 | | 41 | | 41 |
| panbimbo | | 2 | | | | | | 1 | 40 | | 38 |
| takis | | | | | | | | | | 34 | 59 |
| background | 55 | 55 | 58 | 58 | 51 | 49 | | 51 | 51 | 41 | |

| | | |
|---|---|--|
| <p>Yolo11n-seg Segmentación 150 imágenes ~250-350 instancias x clase 100 épocas</p> |  | |
| <p>Yolo11n-seg -> Goods-segmentation weights Dataset: https://universe.roboflow.com/qun-li/goods-segmentation Segmentación 200 imágenes ~300-350 instancias x clase 100 épocas</p> |  | |
| <p>Yolo11n-seg -> Goods-segmentation weights Dataset: https://universe.roboflow.com/qun-li/goods-segmentation Segmentación 500 imágenes ~650-850 instancias x clase 150 épocas</p> |  | |

A pesar de entrenar con pesos pre entrenados y agregar más escenas, los resultados no fueron los mejores, los modelos seguían interpretando y confundiendo el fondo como productos.

Se intento simular abolladuras a los modelos mediante el uso de deformadores, pero BlenderProc no tiene una implementación directa aun desarrollada para ello, aun así, intentamos agregar una variación en la posición de vértices aleatorias en los modelos, pero debido a que los modelos generados por instant mesh tienen una malla muy compleja, al hacerlo de esta manera los resultados no simulan correctamente una abolladura natural y solo agregan más ruido, por lo que descartamos esta implementación.



Ilustración 1.2 Abolladuras a modelos con distintas intensidades

- 2.2. Generación del Dataset Base y el Análisis Exploratorio (EDA)** Tras recibir las 3,478 imágenes reales, el EDA reveló una alta variabilidad en calidad, con un 39% de imágenes potencialmente borrosas y un 21% oscuras. Este hallazgo confirmó que se necesitaba una solución robusta al "ruido" visual. Se procedió a implementar el pipeline híbrido de **OCR+SAM** para generar un dataset inicial de 1,506 máscaras candidatas.



*Ilustración 1 Ejemplo de imagen
removida en el análisis Exploratorio*

- **2.3. Modelos Baseline y el Diagnóstico del Cuello de Botella** Se entrenaron dos modelos yolov8 (n y m) sobre este primer dataset semi-curado. Ambos modelos se estancaron en un rendimiento bajo (~15% mAP50-95). Una revisión manual de las predicciones reveló la causa: el dataset inicial, aunque masivo, contenía un alto grado de ruido (máscaras imprecisas, falsos positivos como "Panqués"). Esta conclusión fue el pivote del proyecto: el problema no era la capacidad del modelo, sino la calidad de las etiquetas.



*Ilustración 2 Mascara removida en el
proceso semi-asistido.*

- **2.4. Creación del "Dataset Dorado" y el "Modelo Guía" Multi-Clase.** Se desarrolló una herramienta de curación manual para revisar y clasificar un subconjunto de los datos, creando un "Dataset Dorado" de 478 instancias de alta calidad, divididas en 4 clases. Al entrenar un modelo yolov8m-seg sobre este set, el rendimiento se disparó, alcanzando un **mAP50-95 general de 78.8%**. Esto probó de forma concluyente que la especificidad y la limpieza de los datos eran la clave para el éxito.

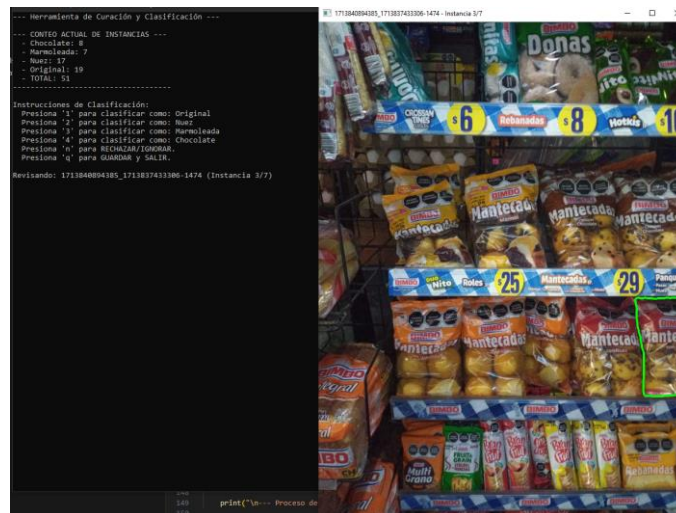


Ilustración 3. Herramienta desarrollada en CMD para verificar las máscaras a usar.

- 2.5. El Ciclo de "Bootstrapping" y el Modelo Definitivo**
 Finalmente, se aplicó una técnica de etiquetado asistido por IA. El "Modelo Guía" (con 78.8% mAP) se utilizó para generar pseudo-etiquetas en todo el conjunto de 3,478 imágenes. Estas ~2,400 etiquetas candidatas fueron luego revisadas y corregidas manualmente por el equipo, un proceso que resultó en el dataset final, masivo y de alta fidelidad, sobre el cual se entrenó el modelo definitivo.

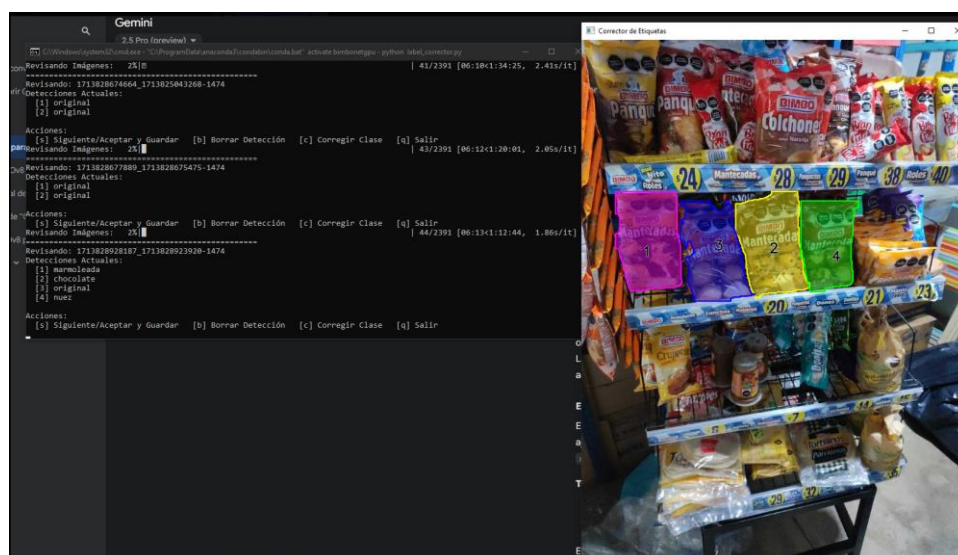


Ilustración 4. Proceso de curacion final.

3. Resultados y Evolución Cuantitativa del Rendimiento

La siguiente tabla resume la evolución del rendimiento a través de las fases clave del proyecto, demostrando el impacto de cada decisión estratégica.

| Modelo / Iteración | Dataset Utilizado | Calidad del Dataset | mAP50-95 (Mask) | Conclusión Clave |
|---------------------|-----------------------|--------------------------------|-----------------|--|
| 1 v1/v2 (Baseline) | ~1,900 imgs, 1 clase | Crudo, generado por SAM+OCR | ~15% | Rendimiento base bajo. La calidad de los datos es el cuello de botella. |
| 2 v3 ("Guía") | 327 imgs, 4 clases | 478 instancias, curado manual | 78.80% | La calidad y especificidad de los datos es el factor más importante. |
| 3 v4 ("Definitivo") | ~1,345 imgs, 4 clases | >2,300 instancias, verificadas | 79.00% | Robustez y Escalabilidad Confirmada. El alto rendimiento se mantiene y estabiliza en un dataset 4 veces más grande. |

Ilustración 5. Tabla comparativa de modelos

El modelo definitivo alcanzó un rendimiento general de 79.0% en la métrica mAP50-95 para las máscaras, con un desglose por clase que demuestra su alta capacidad de especialización:

- **original:** 90.7%
- **nuez:** 78.5%
- **chocolate:** 75.6%
- **marmoleada:** 71.4%

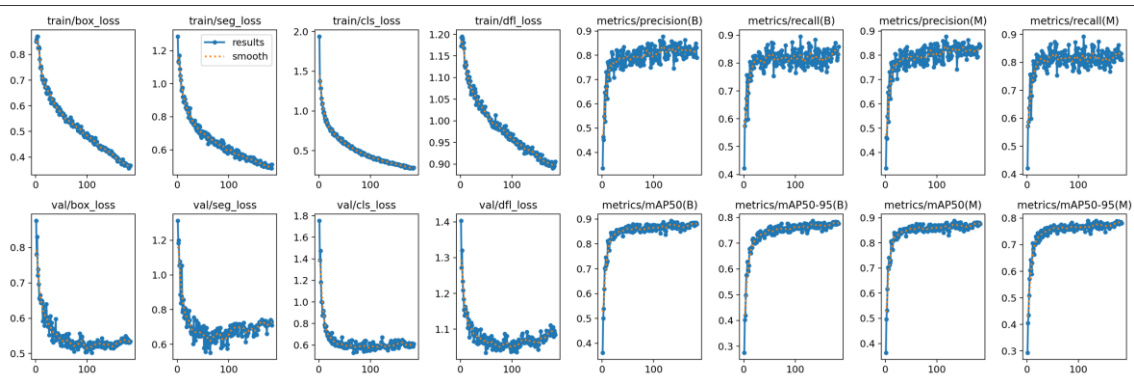


Ilustración 6. Evolucion de las metricas en los 150 epochs.

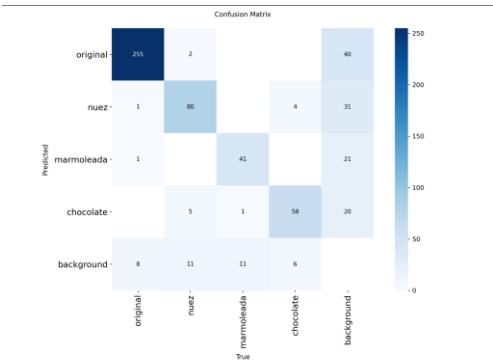


Ilustración 7. Matriz de confusion por clase

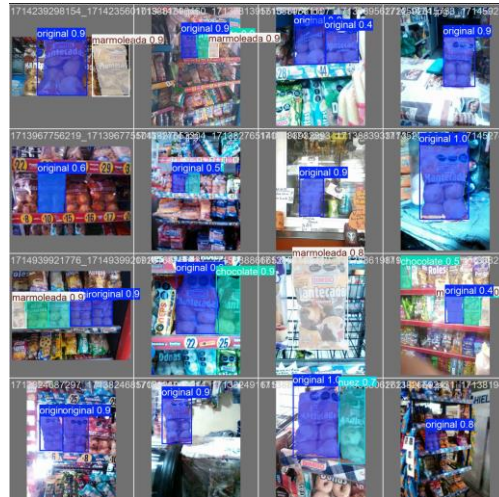


Ilustración 8. Imágenes usadas para validación del modelo.

4. Discusión y Recomendaciones Estratégicas

4.1. Rendimiento del Modelo para Producción

Con un mAP50-95 general del 79% y un rendimiento superior al 90% para la clase principal, el modelo actual es un candidato fuerte para un despliegue en un programa piloto. Ha superado el umbral del 75% que se considera robusto para producción. Su capacidad para identificar correctamente 9 de cada 10 mantecadas "originales" y más de 7 de cada 10 de las variedades más raras lo convierte en una herramienta fiable para el objetivo de negocio de determinar la presencia/ausencia de producto.

A pesar del resultado, siempre hay margen de mejora:

1. **Enriquecer con "Negativos Difíciles":** Analizar los errores del modelo actual (ej. falsos positivos) y añadir explícitamente imágenes de esos productos de confusión al dataset para un re-entrenamiento futuro. Estos negativos difíciles incluyen:
 1. Mantecadas muy encimadas, una junto con otra.
 2. Mantecadas que están una sobre otra y solo se ve la parte superior de la mantecada del fondo.
 3. Fotos oscuras, donde el SAM + OCR tienen dificultad para leer, la modelo está cegado en fotos con buena luz.
2. **Creación de un "Golden Test Set":** Para obtener una métrica de Recall académicamente pura y libre de cualquier sesgo del pipeline de generación de datos, se recomienda la creación de un conjunto de prueba pequeño y perfectamente etiquetado desde cero.

4.3. Recomendaciones Clave para Implementación

1. **API de Inferencia:** Encapsular el modelo best.pt en una API RESTful (usando FastAPI) que sirva como el cerebro central del sistema.
2. **Infraestructura en la Nube:** Desplegar la API en **Google Cloud Platform (GCP)**, utilizando servicios como Vertex AI Endpoints, dada su facilidad de uso y su potente ecosistema de IA.
3. **Aplicación Cliente y Piloto:** Desarrollar la aplicación móvil para los usuarios de "Chambitas" y colaborar con el departamento de negocio para diseñar un programa piloto en un grupo selecto de tiendas para medir el impacto real.
4. **Ciclo de Mejora Continua:** Implementar un sistema de retroalimentación donde las predicciones corregidas por los usuarios se recolecten para futuros re-entrenamientos.

4.4. Tareas Accionables para los Stakeholders

- **Definir Reglas de Negocio:** Traducir las predicciones del modelo en acciones concretas. (Ej: "Si el modelo no detecta mantecadas 'originales' en una tienda de alta rotación por dos días seguidos, generar una alerta de stock al distribuidor").
- **Explotar la Data Granular:** Diseñar estrategias de marketing o ventas basadas en la data de qué variantes son más o menos prevalentes en ciertas zonas geográficas o tipos de tienda.
- **Planificar el Futuro:** Utilizar el pipeline de IA desarrollado en este proyecto como un activo reutilizable para identificar y rastrear el siguiente producto de alto valor para la compañía.

4. Análisis Costo-Beneficio

El caso de negocio para este proyecto es excepcionalmente sólido.

La inversión principal ha sido en capital humano. Estimamos un total de 240 horas-hombre de trabajo especializado en las fases de entendimiento del negocio y estrategia de datos, sumadas a unas 45 horas de trabajo intensivo en la curación y etiquetado del "Dataset Dorado". El costo de cómputo para el entrenamiento fue nulo al utilizar hardware local de alto rendimiento. Para la fase de operación, se proyecta un costo mensual de entre \$100 y \$200 USD para el despliegue de la API en un entorno de nube piloto.

Estos costos se ven minimizados por los beneficios potenciales. En el frente operativo, la automatización reducirá el tiempo de validación por imagen de minutos a segundos, generando un ahorro de más de 16 horas-hombre por cada 1,000 imágenes procesadas. Estratégicamente, el beneficio es aún mayor.

Al proporcionar una visión casi en tiempo real del estado del anaquel, el sistema permite una reacción rápida ante los agotados en stock (stockouts). Una reducción conservadora de esta métrica podría traducirse en un aumento estimado del 1% al 3% en las ventas de los productos monitoreados. A esto se suman los beneficios intangibles, pero cruciales, de mejorar la retención de los usuarios de T-Conecta y la creación de un activo de datos sin precedentes para la inteligencia de mercado.

5. Riesgos y Desafíos Estratégicos

Todo proyecto de IA conlleva riesgos inherentes. Nuestra metodología se ha enfocado en mitigar los más importantes:

- **Riesgos de Datos y Confianza:** El principal riesgo identificado fue el "ground truth incompleto" heredado de nuestro pipeline de generación automática, lo que podría inflar métricas como el Recall. La mitigación directa fue el proceso de curación manual intensivo. Para una validación final y auditable, se recomienda la creación de un "Golden Test Set" perfectamente etiquetado. Adicionalmente, el riesgo de que el modelo se confunda con productos similares se aborda con la naturaleza visual de la segmentación: cada detección es una prueba en sí misma, construyendo confianza con los stakeholders.
- **Riesgos Operacionales y de Cumplimiento:** Existe el riesgo de que usuarios intenten "engañar" al sistema. Esto se mitiga monitoreando los puntajes de confianza de las predicciones y marcando las anomalías para revisión. Asimismo, para garantizar el cumplimiento de la privacidad, se debe implementar un paso de anonimización en la API que desenfoque automáticamente cualquier rostro que pudiera aparecer en las imágenes antes de su procesamiento.