

UNIVERSIDAD
YACHAY
TECH



Sistema de Recomendación FNN (TOP-3)

Maestría en Inteligencia Artificial

Módulo: Deep Learning

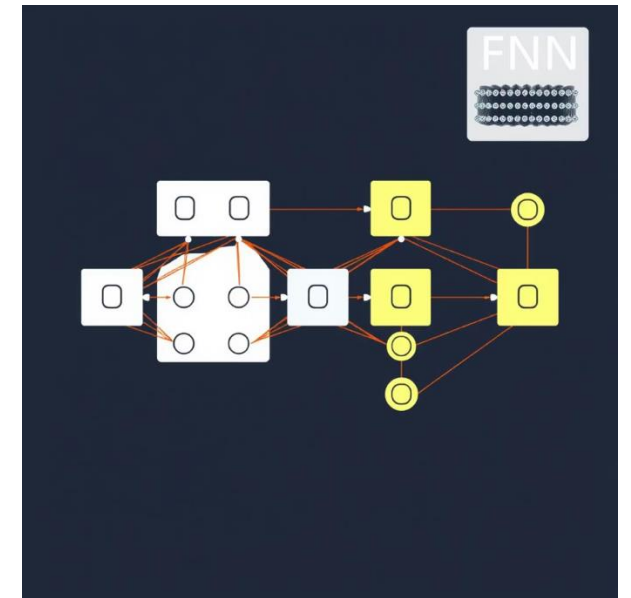
Profesor: Dr. Oscar Chang

Estudiantes: Juan Brito, Alexander González

Período Académico: 2025 – 2026

Sistema de Recomendación FNN (TOP-3)

Producción | Corte 2025-11-09



Análisis Exploratorio de Datos (EDA)

Comprensión del Comportamiento de Compra por Familia

El objetivo del EDA fue analizar el comportamiento histórico de compra por familia y subcategoría para identificar patrones que permitan predecir la probabilidad de compra futura.

Dataset Analizado

- +690.000 registros históricos.
- Variables: Recencia, Frecuencia, Volumen, SOW, Estacionalidad.
- Identificadores clave: Código Familia, Código Subcategoría.
- Variable objetivo: Target binario (compra/no compra).

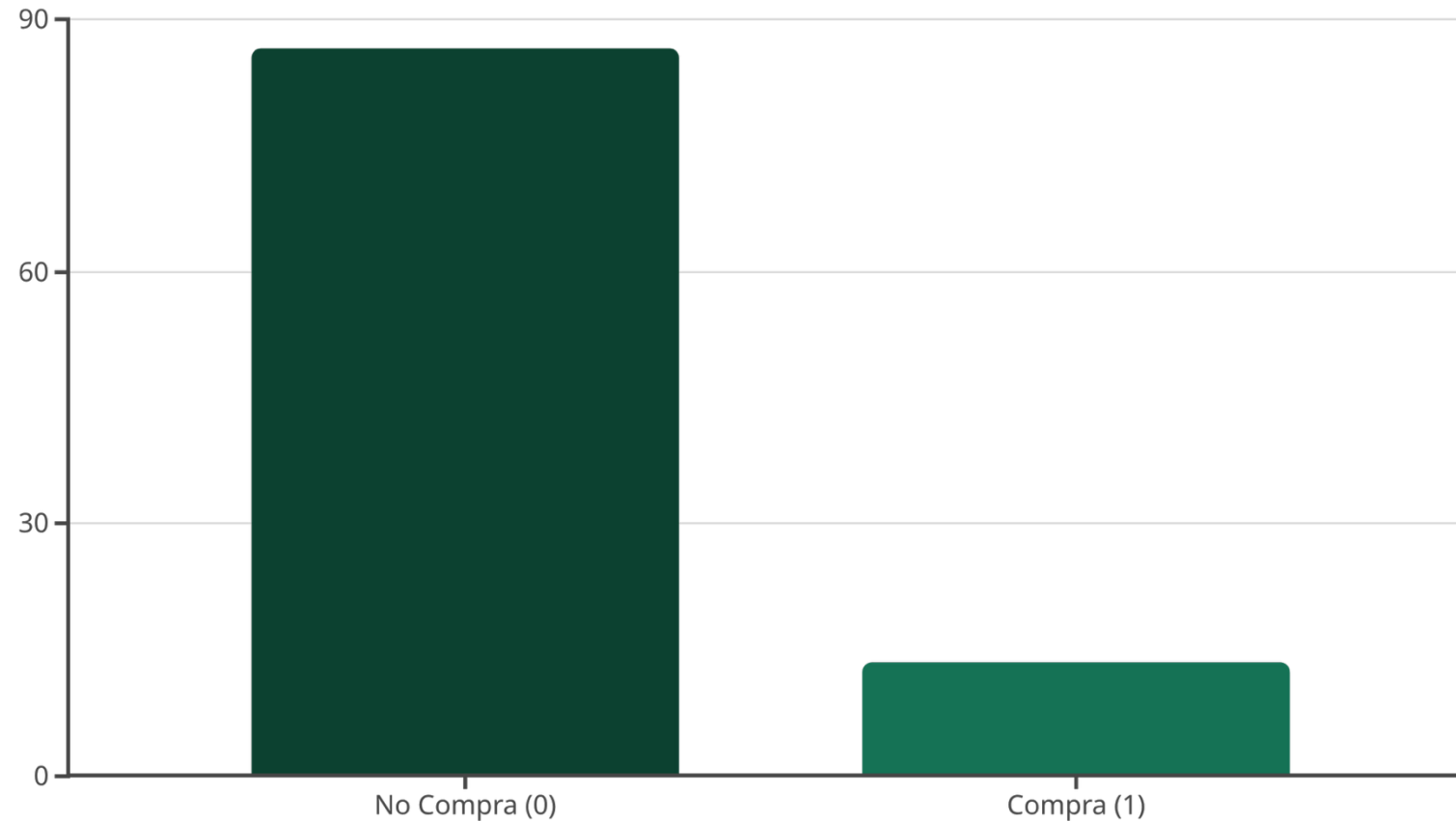
Acciones Clave del EDA

- Validación de calidad de datos (NaN, outliers).
- Análisis de distribución del target (desbalance).
- Evaluación de variables numéricas y su relación con la compra.
- Identificación de variables estratégicas para el modelo.
- Preparación de dataset para embeddings, escalado, imputación y split por familia.

El análisis confirmó un problema de clasificación binaria desbalanceada, justificando el uso de BCEWithLogits + pos_weight, split por CODIGO_FAMILIA y embeddings para variables categóricas, sentando una base sólida para el modelo FNN.

Análisis Exploratorio de Datos – Distribución de la Variable Objetivo

El análisis exploratorio inicial se centra en la variable objetivo, que indica la ocurrencia de una compra futura. La distribución de esta variable clave revela un desbalance significativo entre las clases, un patrón común en contextos de negocio donde las transacciones de compra son eventos menos frecuentes que la ausencia de compra.



Este desbalance subraya la necesidad de un enfoque cuidadoso en el modelado, justificando la implementación de estrategias específicas para asegurar un rendimiento óptimo del sistema de recomendación.

→ Métricas de Evaluación Robustas

Empleo de AUC, Recall y F1-score para una evaluación más precisa, evitando métricas sesgadas por el desbalance.

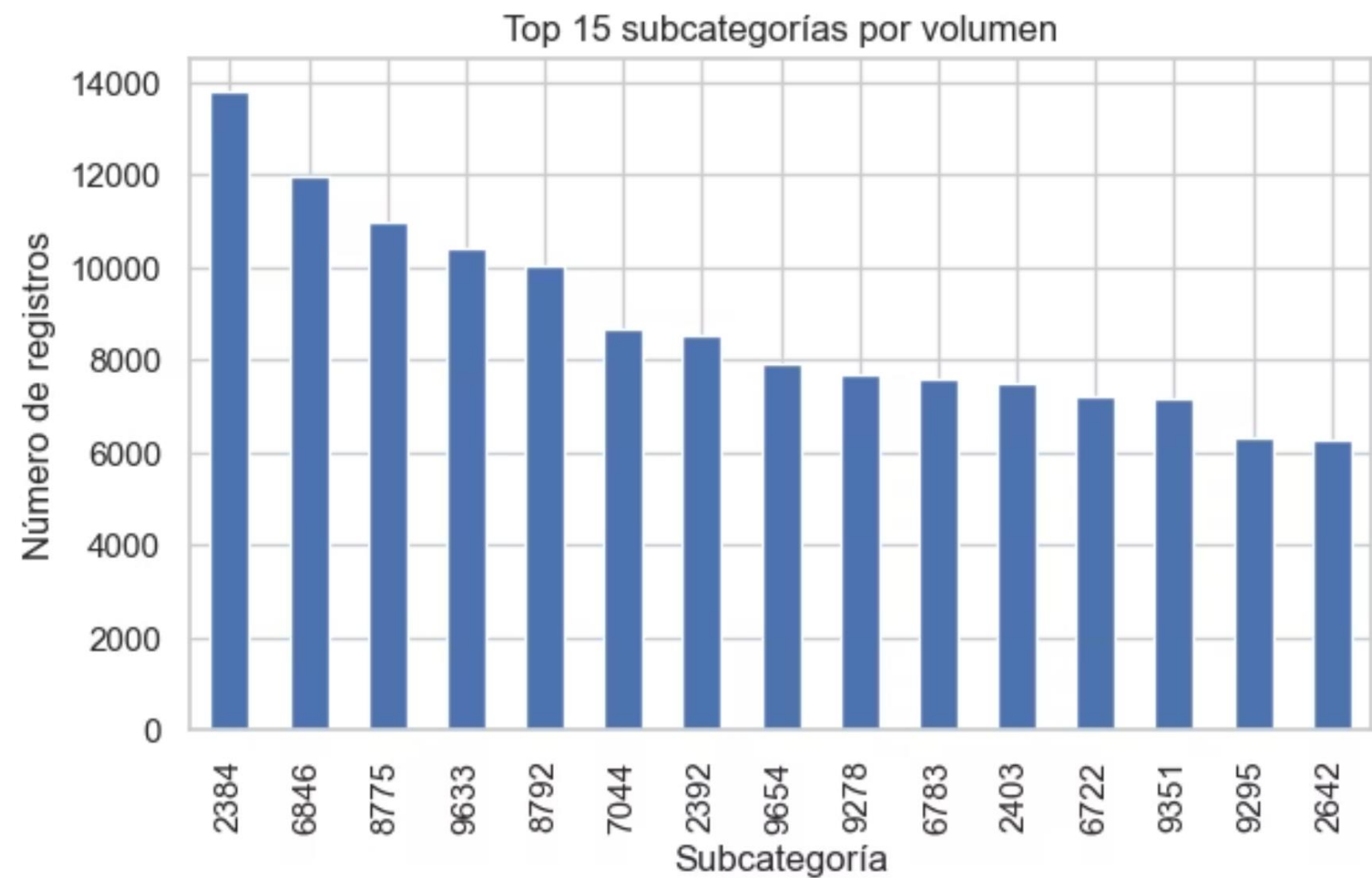
→ Compensación del Desbalance

Uso de ponderación de clases (`pos_weight`) en la función de pérdida del modelo para dar mayor relevancia a la clase minoritaria de 'compra'.

Distribución de Subcategorías por Volumen de Registros

El análisis por subcategoría permite identificar los segmentos con mayor volumen de información histórica, revelando las áreas donde se concentra la actividad de compra.

Se observa una alta concentración de registros en un conjunto reducido de subcategorías, lo que sugiere patrones de compra recurrentes y posibles oportunidades de segmentación para futuras estrategias comerciales.



- Priorizar Subcategorías Estratégicas
Enfocar esfuerzos en las subcategorías con mayor volumen para optimizar el rendimiento del modelo de recomendación.
- Comprender la Heterogeneidad
Reconocer las diferencias en el comportamiento de compra entre subcategorías para adaptar las estrategias.
- Justificar Uso de Embeddings
Este patrón refuerza la necesidad de utilizar embeddings categóricos en el modelo FNN para capturar estas relaciones complejas.

Contexto del Proyecto

Objetivo

Recomendar 3 productos (subcategorías) por familia para maximizar aciertos de compra de las próximas 2 semanas.

Caso de Uso

Sistema de recomendación para retail/farmacia.

Datos Utilizados

Histórico: Historico_08122025.csv
(compras Nov 2023 a Nov 2025, ~500,000 registros)

Test Real: data_test.xlsx
(compras reales 2025-12-01 a 2025-12-09, solo evaluación)

Flujo End-to-End del Sistema



Datos Históricos

Historico_08122025.csv



Feature Engineering

4 features: recencia, frecuencia, sow, estacionalidad



Modelo FNN

4 inputs → 64 → 32 → 1 output



Predicción TOP-3

ordenar y seleccionar top 3 por familia



Evaluación

comparar contra compras reales

4 Features Clave del Modelo

Recencia (recencia_hl)

Intuición: Mide cuán reciente fue la última compra (1 = muy reciente, 0 = antigua)

Frecuencia (freq_score)

Intuición: >1 compra más seguido; =1 igual; <1 menos seguido

Share of Wallet (sow_24m)

Intuición: Proporción de gasto en esa subcategoría vs total

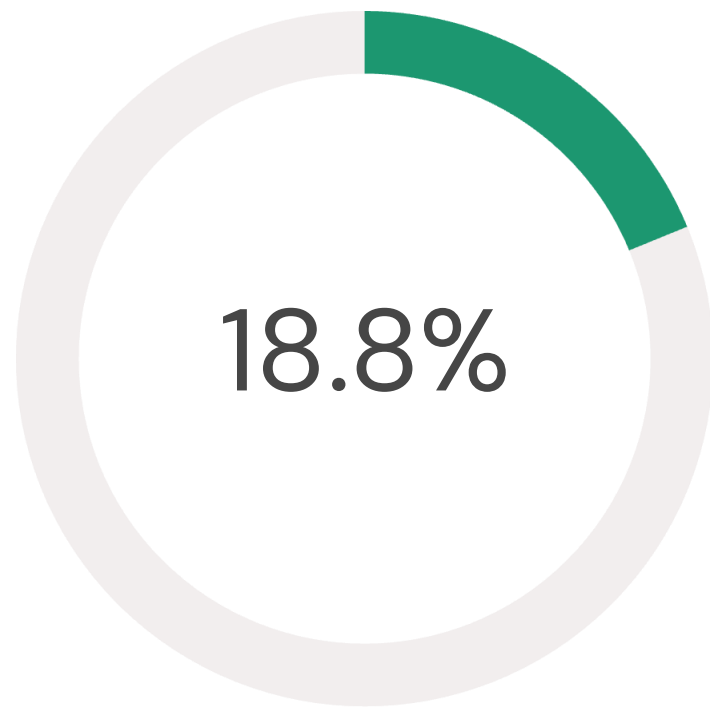
Estacionalidad (season_ratio)

Intuición: Patrón estacional de compra en ese mes

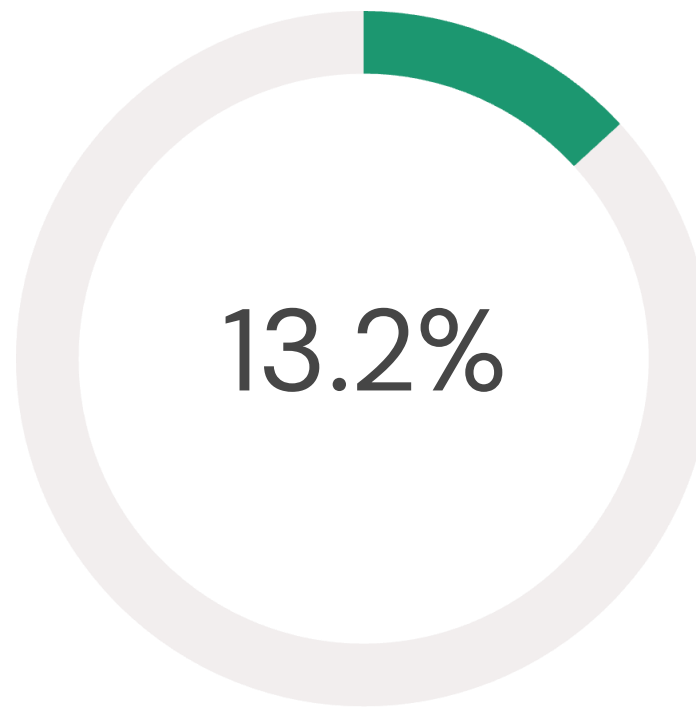
Baseline: Modelo Lineal

$$\text{score} = 0.4 \times \text{recencia} + 0.3 \times \text{frecuencia} + 0.1 \times \text{sow} + 0.2 \times \text{estacionalidad}$$

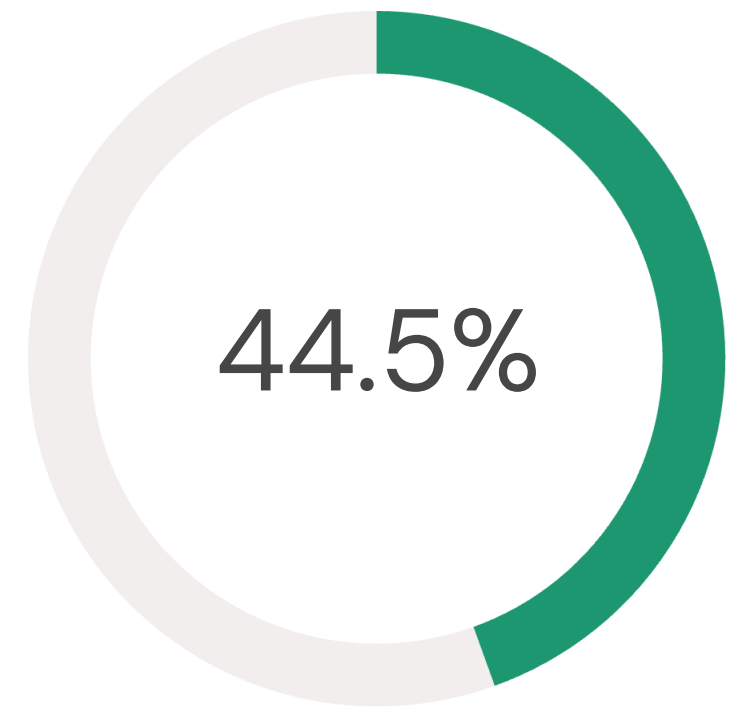
Resultados (Corte 2025-11-09):



Precision@3



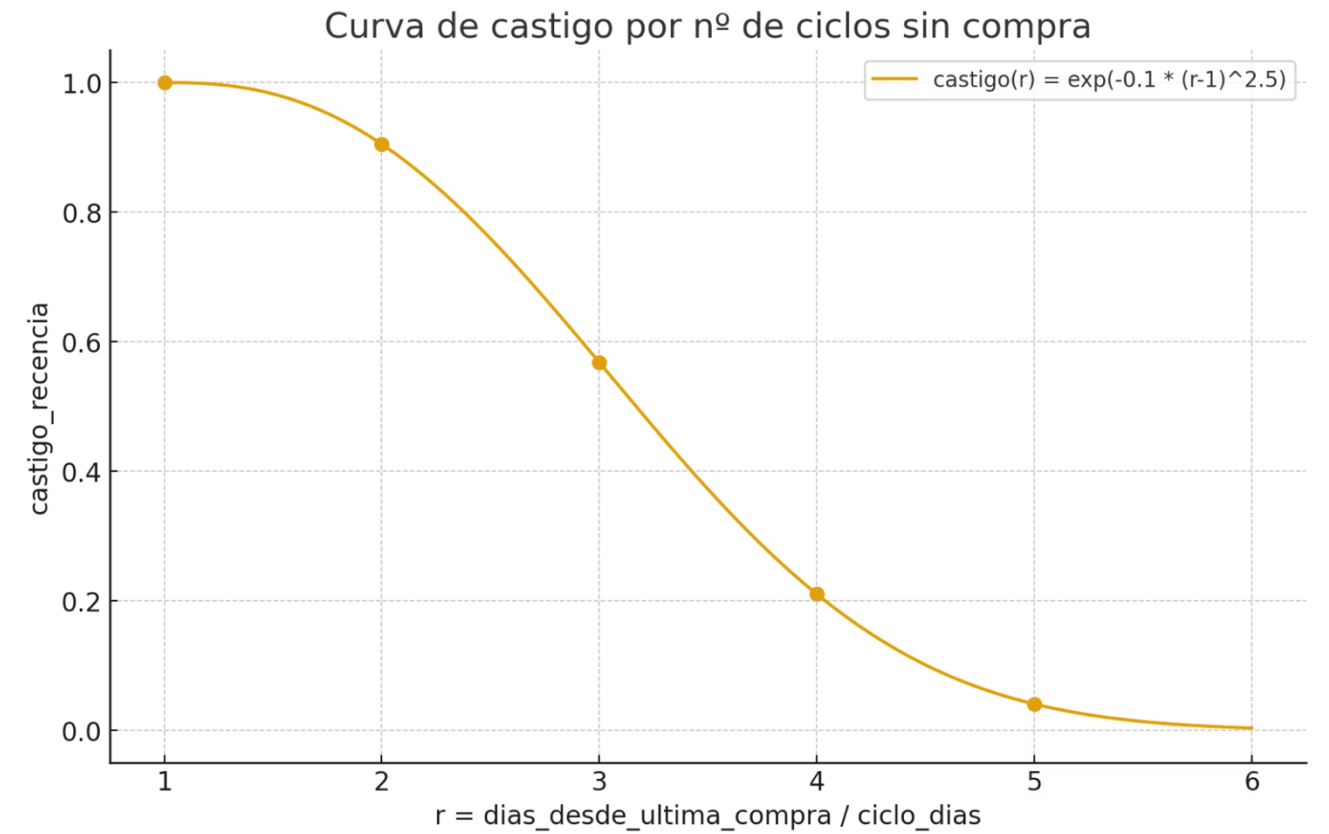
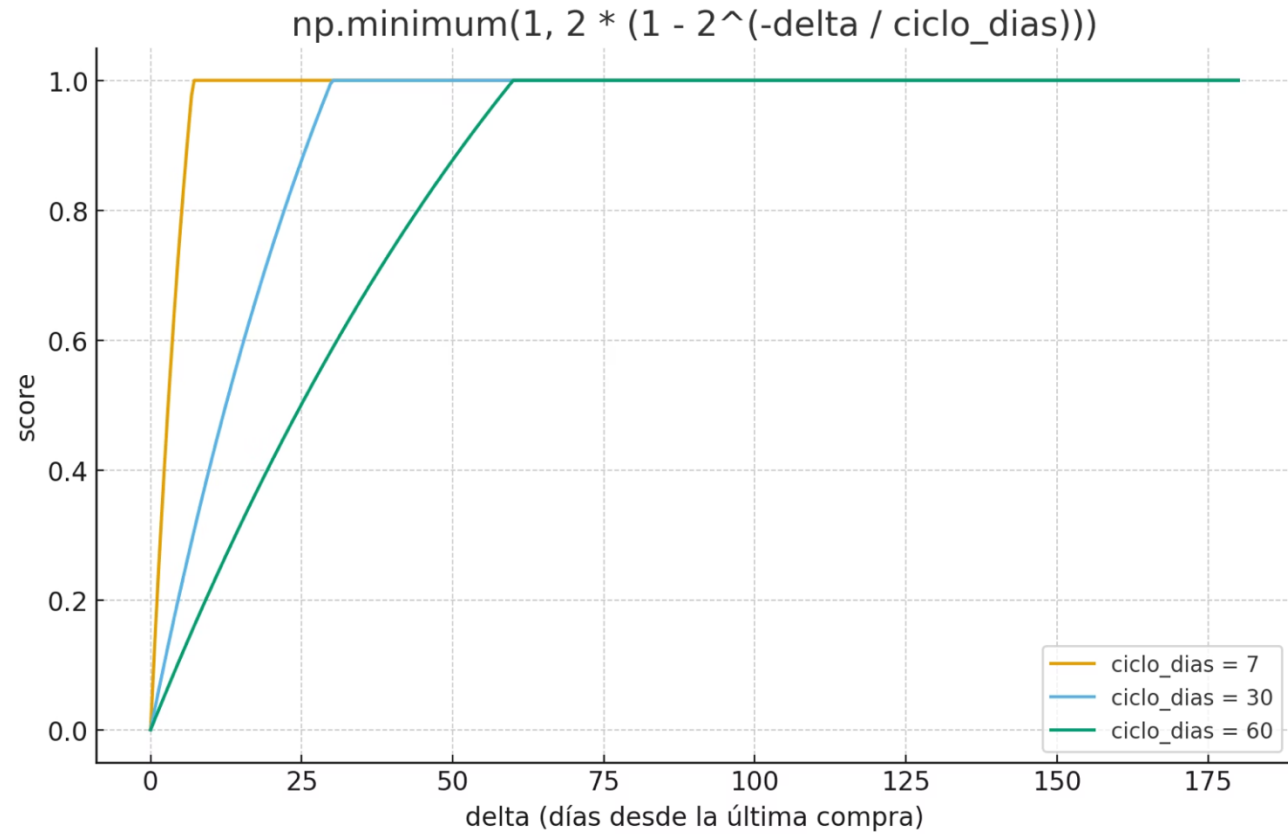
Recall@3



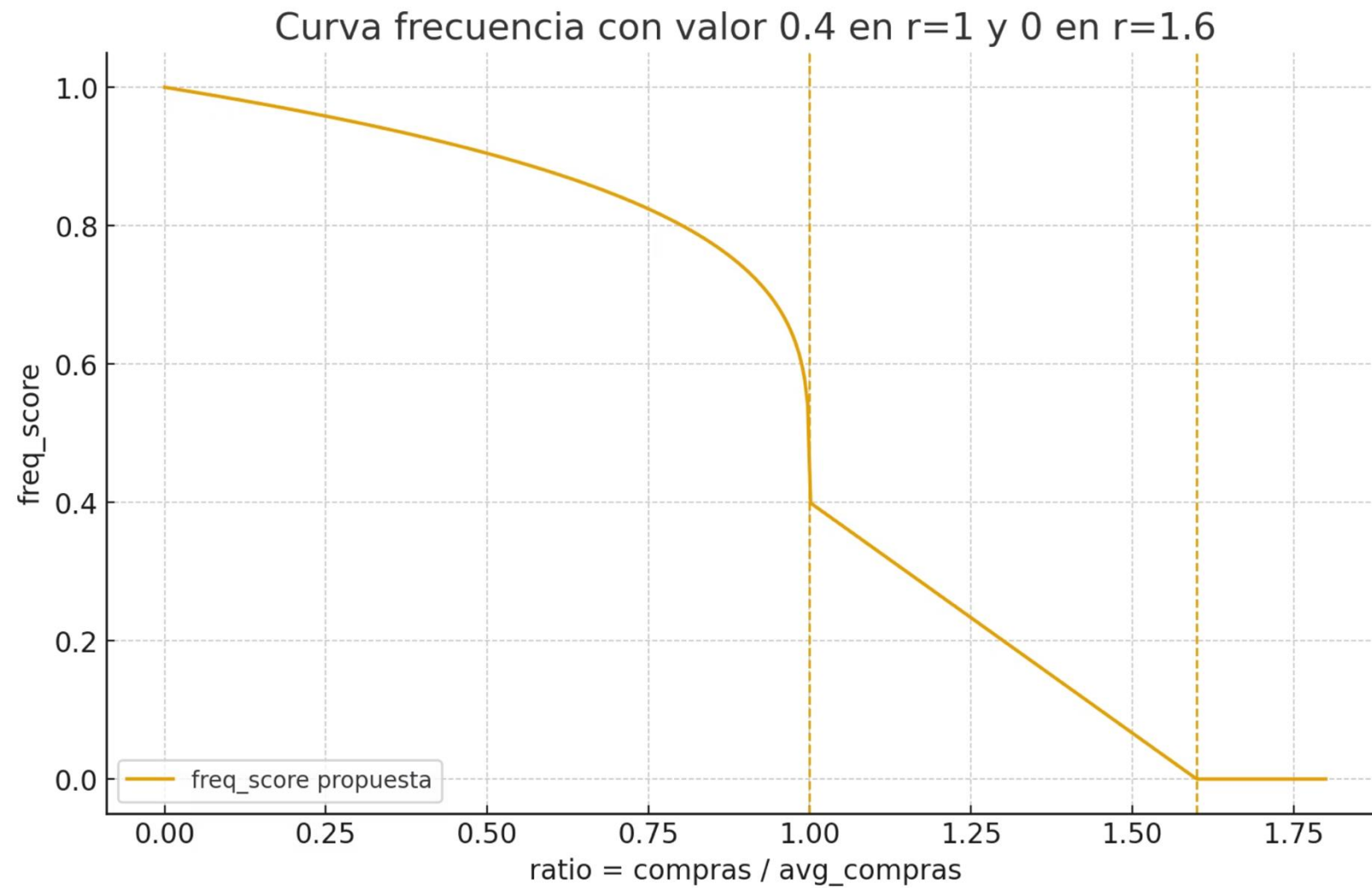
Hit Rate@3

Propósito: Referencia para comparar mejora del modelo FNN.

Recencia



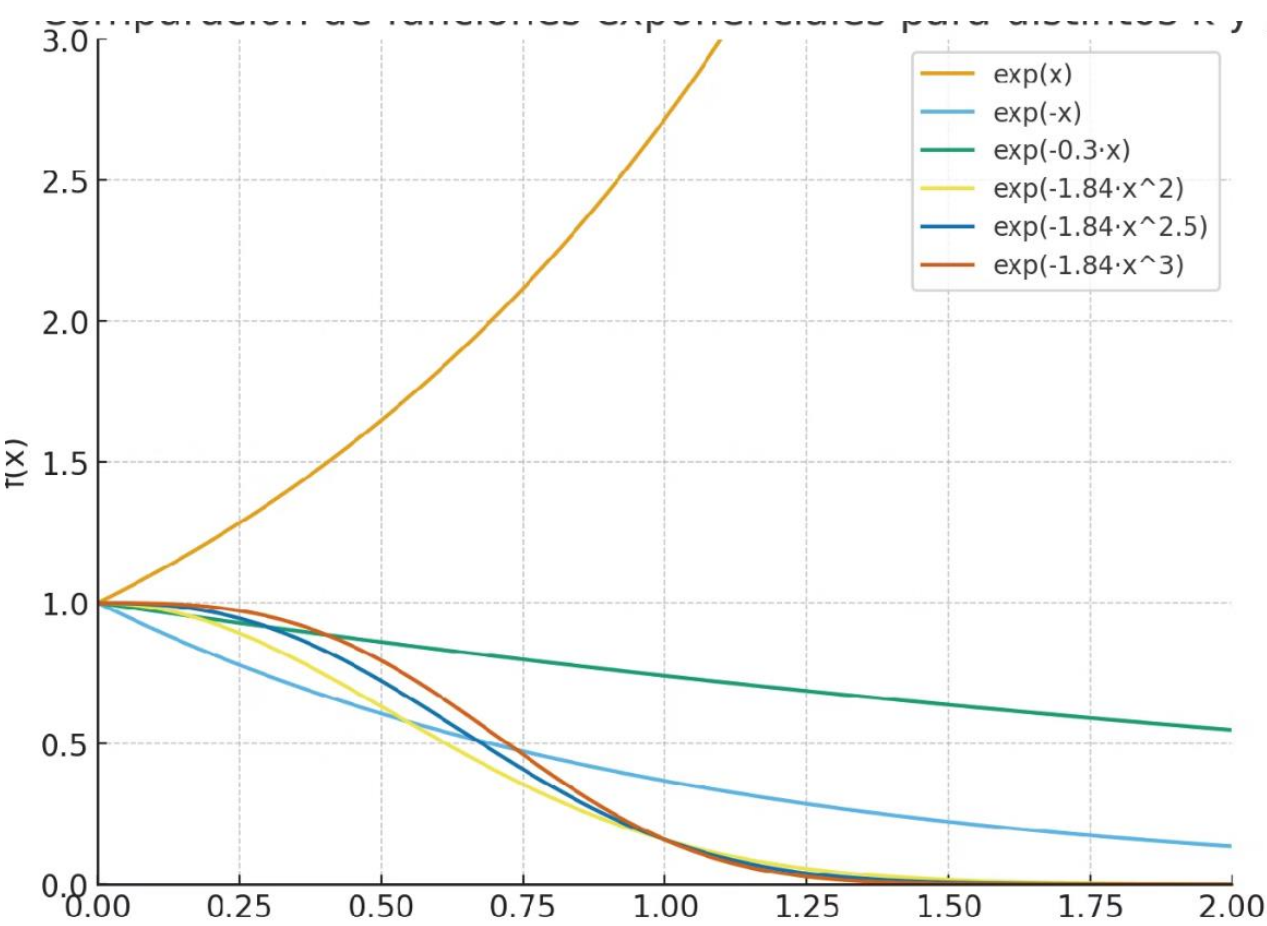
Frecuencia



SOW

Se cuenta el número de transacciones que se realizo por cada subcategoria y se ve la proporción que representan estas con respecto al total de transacciones de la familia, luego se rescala el valor de modo que sea 1 para la subcategoria que mas gasta la familia y el resto su correspondiente proporcional.

Estacionalidad



Construcción del Dataset y Arquitectura FNN

Construcción del Dataset

01

Período de Features
(Histórico menos últimos 21 días)

- Se utilizó el histórico COMPLETO MENOS los últimos 21 días (Nov 10-30).
- En este período se calcularon las 4 features: recencia, frecuencia, SOW, estacionalidad.
- Esto asegura que las features sean "limpias" sin contaminar con el período de evaluación.

02

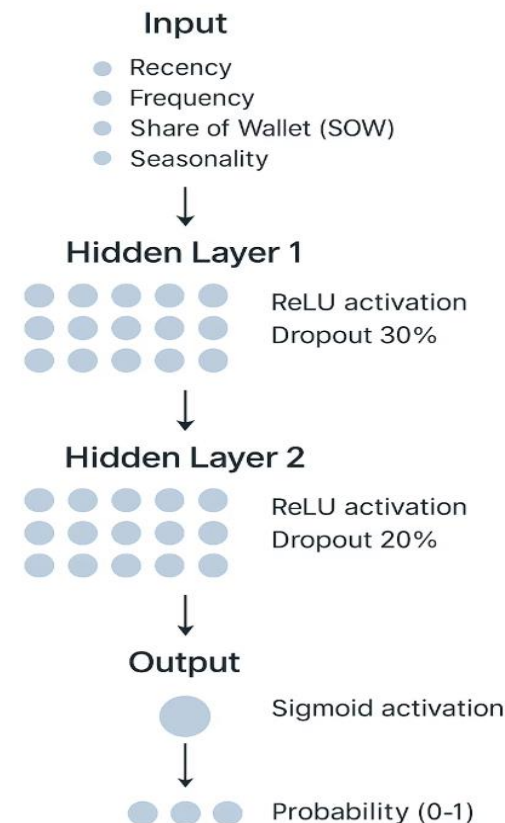
Período de Target (Nov 10-30)

- Los últimos 21 días se reservaron EXCLUSIVAMENTE para calcular el TARGET.
- TARGET = 1: Si la familia compró esa subcategoría en Nov 10-30.
- TARGET = 0: Si la familia NO compró esa subcategoría en Nov 10-30.

03

Dataset Final

- Cada fila = (familia, subcategoría, 4 features, target).
- Total: 60,205 registros.
- Ejemplo:
Familia 100045509 | Subcat 9353 |
recencia=0.933 | freq=1.33 | SOW=0.20 |
season=1.0 | TARGET=1
Familia 100045509 | Subcat 9278 |
recencia=0.450 | freq=0.80 | SOW=0.10 |
season=0.5 | TARGET=0



Entrenamiento del Modelo FNN

Dataset

Total: 60,205 registros (corte de features al 2025-11-09)

Split 80/20: Train 48,164 | Test 12,041

Configuración

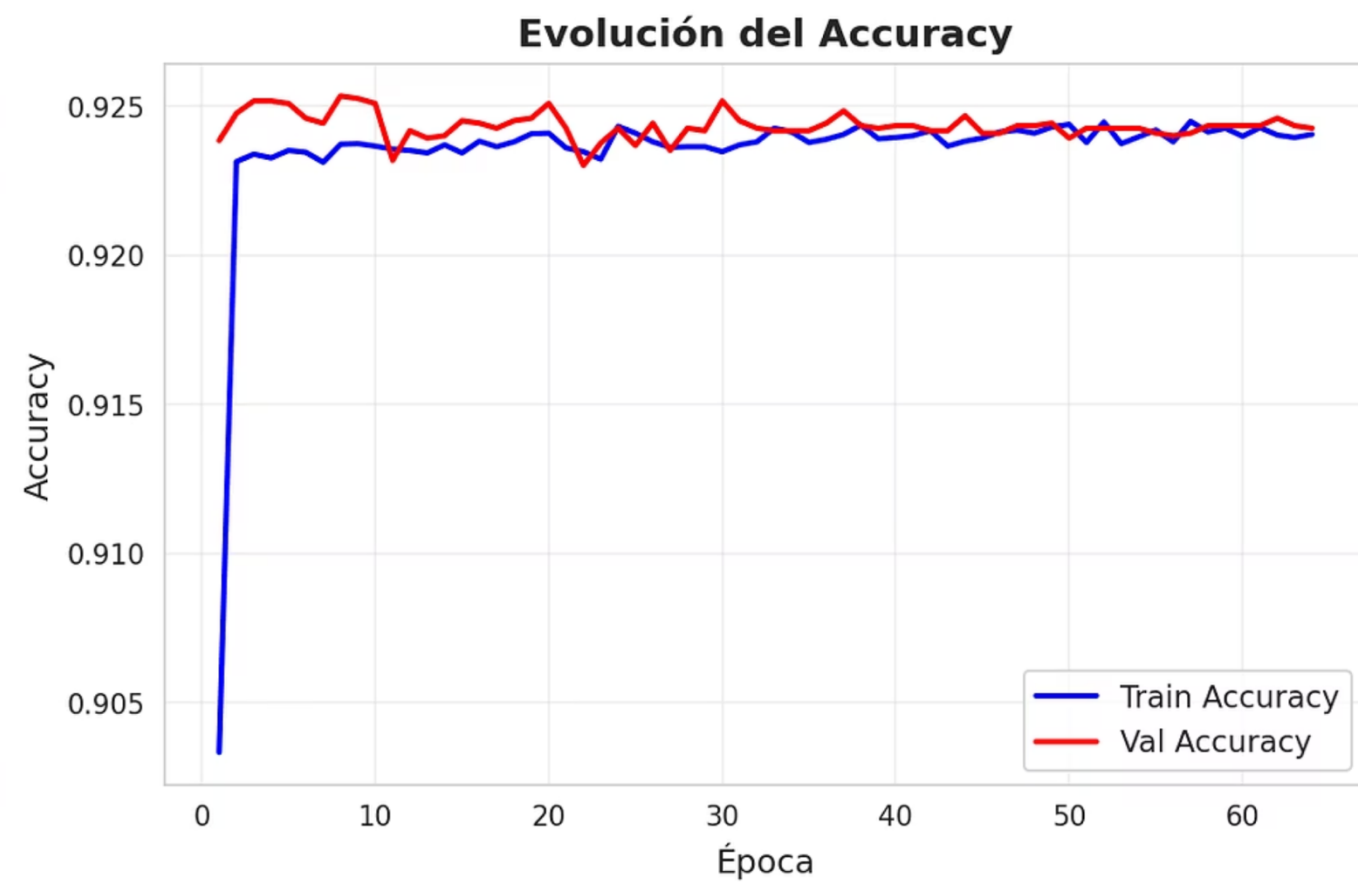
Normalización: Estandarización (mean=0, std=1) usando scaler

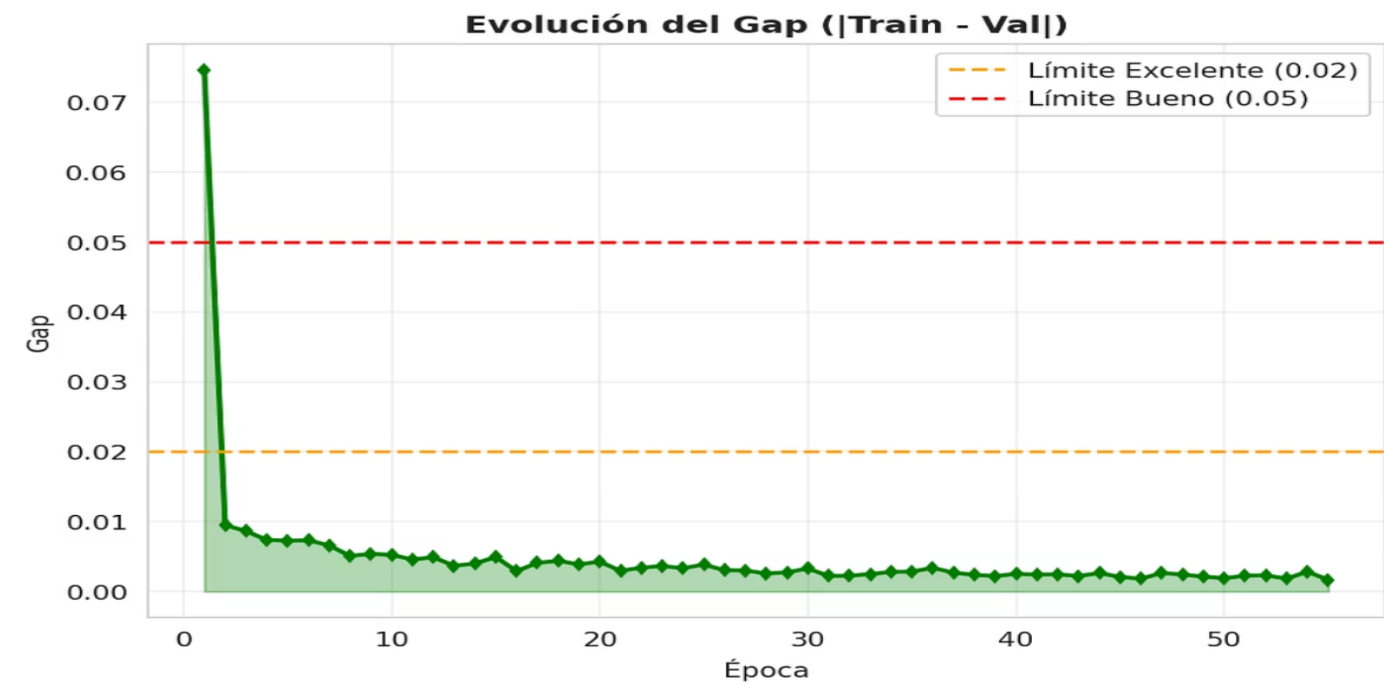
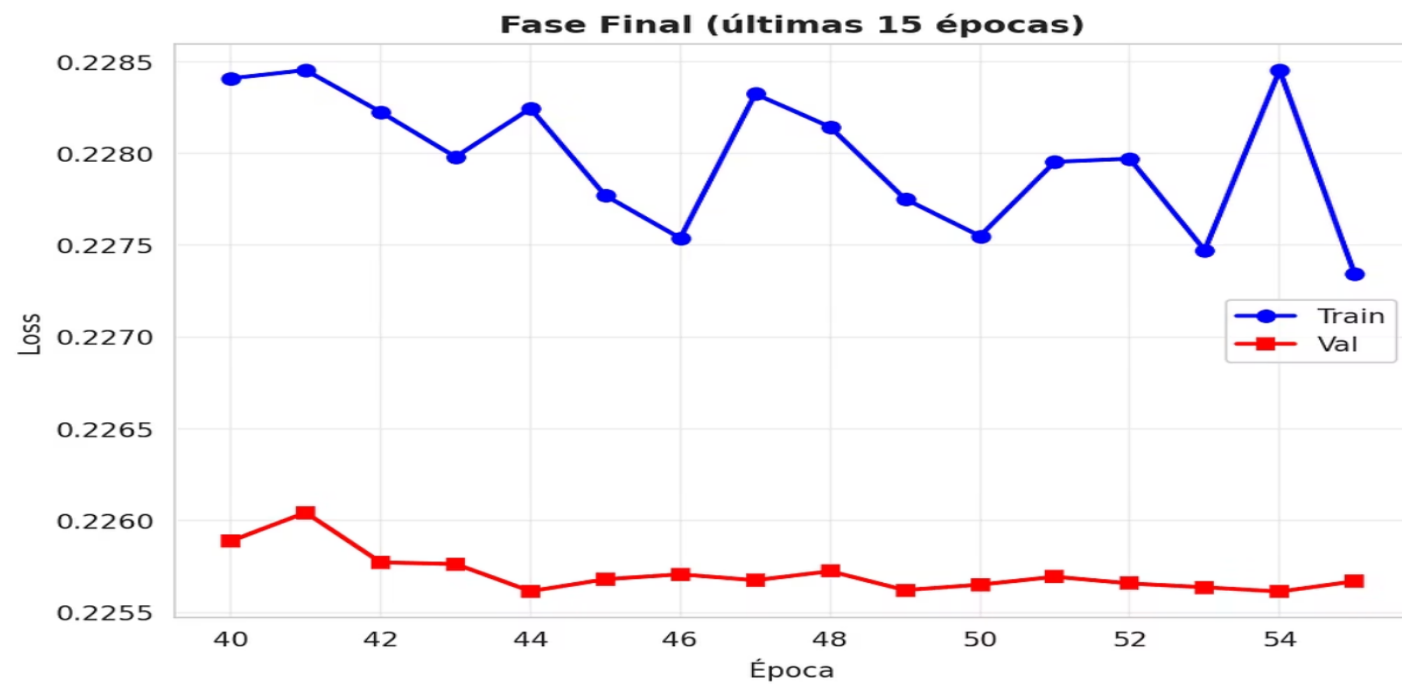
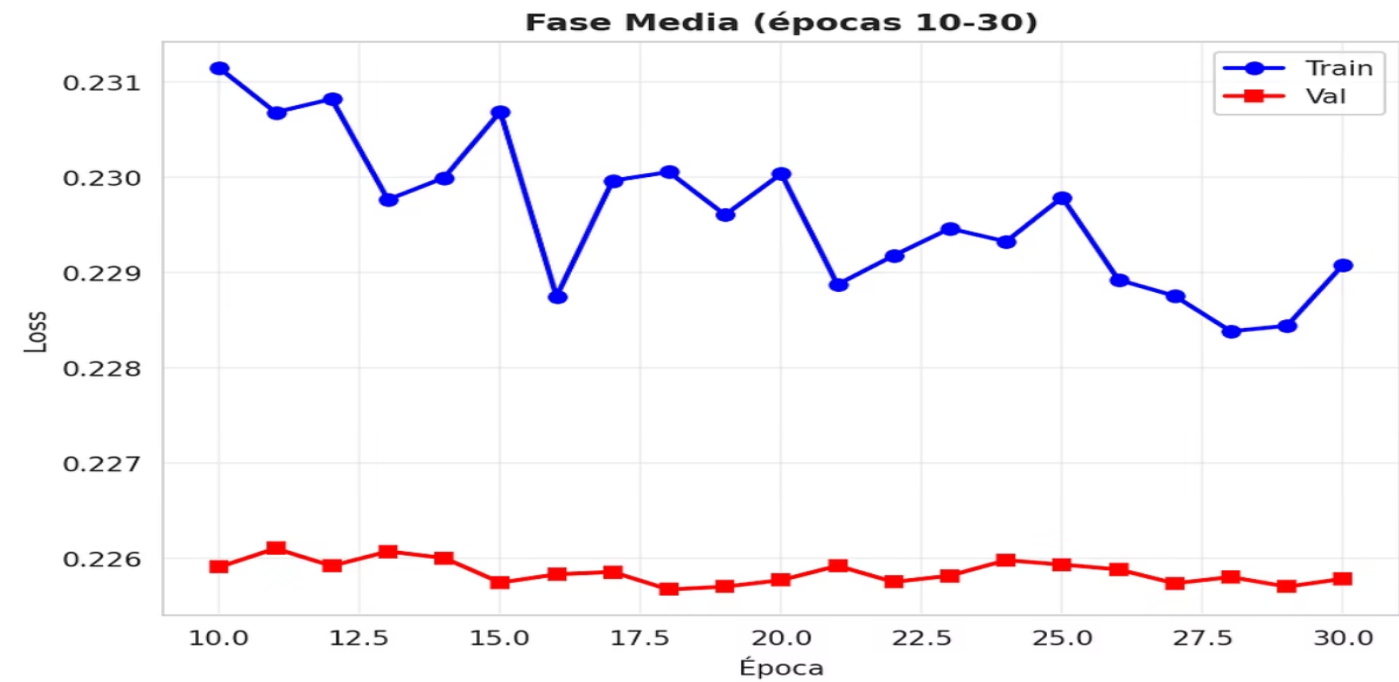
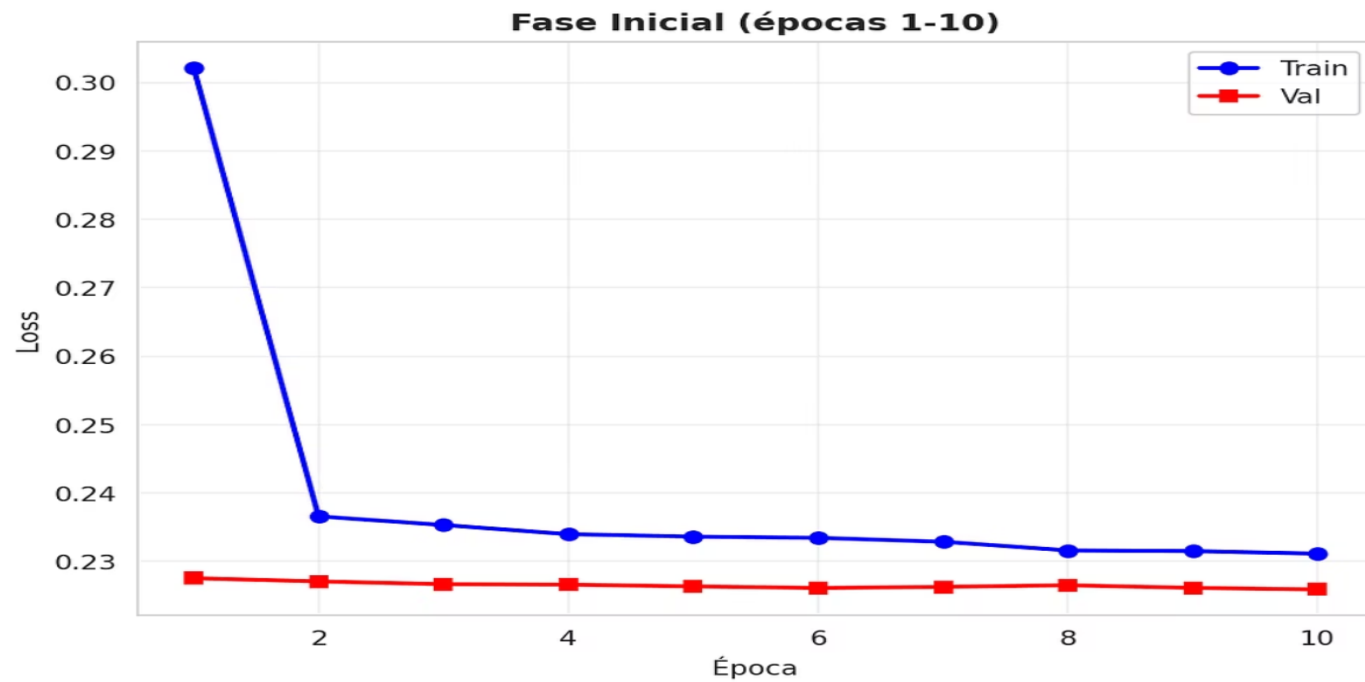
Epochs: Hasta 100 con early stopping

Resultado: Terminó en 55 epochs

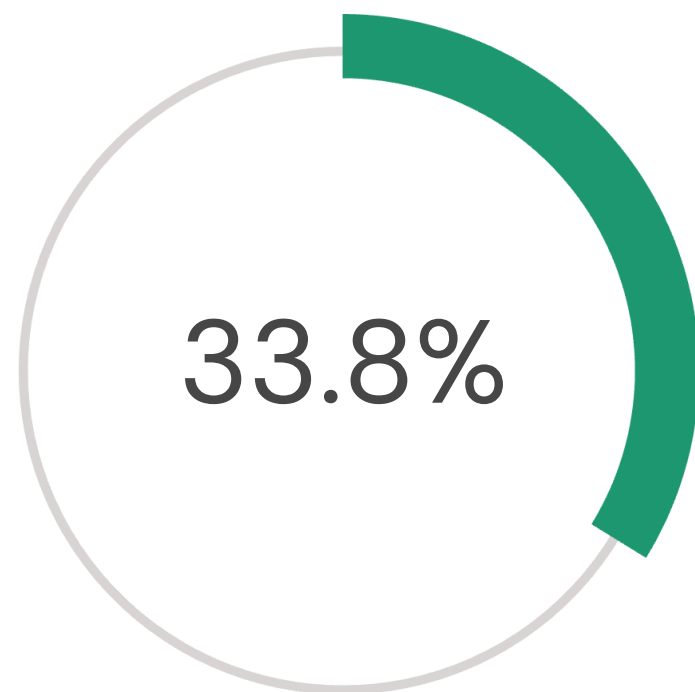
Loss Final

Train Loss	0.2273
Val Loss	0.2257
Gap	0.0016 (excelente convergencia, sin overfitting)

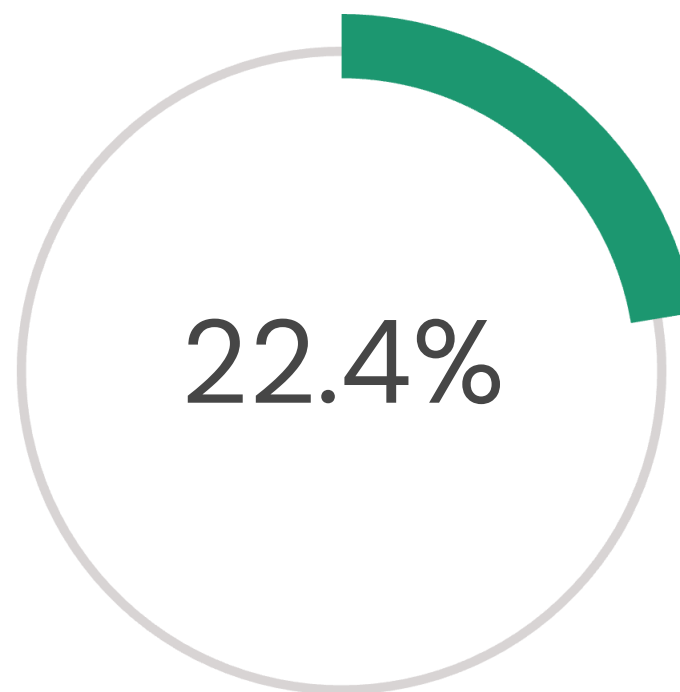




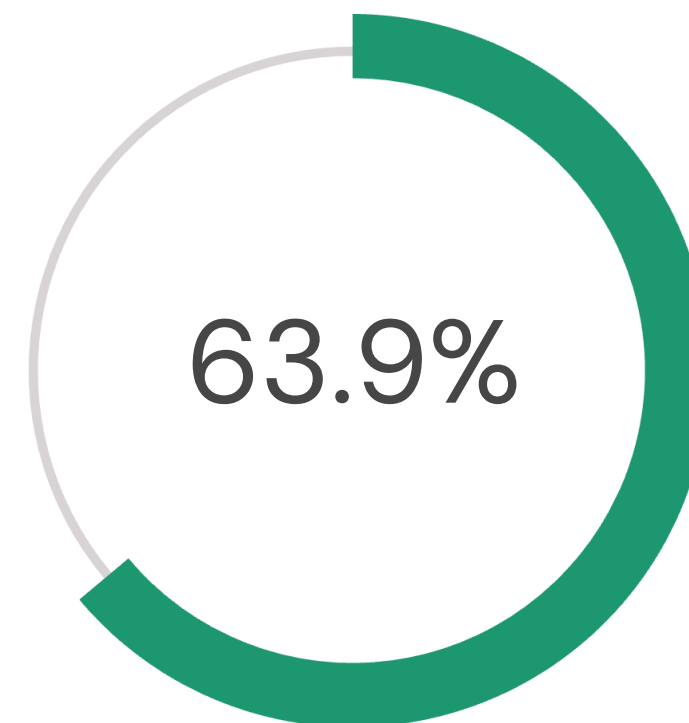
Resumen Ejecutivo



Precision@3



Recall@3



Hit Rate@3

Mejora vs Baseline (Linear): +79.7% Precision | +70.0% Recall | +43.5% Hit Rate

Evaluación: Proceso TOP-3



Predecir Probabilidad

Para todas las subcategorías de cada familia.con las características calculadas con el dataset completo.



Seleccionar TOP-3

Las 3 con mayor probabilidad,por familia por subcategoría.



Comparar

Contra compras reales (2025-12-01 a 2025-12-09).

Métricas

Precision@3

Proporción de recomendaciones TOP-3 que fueron compradas.

Recall@3

Proporción de compras reales que estaban en TOP-3.

Hit Rate@3

Porcentaje de familias con al menos 1 acierto en TOP-3.

Resultados: Linear vs FNN (Corte 2025-11-09)

Métrica	Linear	FNN	Mejora
Precision@3	18.8%	33.8%	+79.7%
Recall@3	13.2%	22.4%	+70.0%
Hit Rate@3	44.5%	63.9%	+43.5%

El modelo FNN supera significativamente al baseline lineal en todas las métricas, demostrando la capacidad de la red neuronal para capturar relaciones no lineales en los datos de compra.

Conclusiones y Próximos Pasos

Conclusiones:

✓ Modelo FNN supera baseline lineal en todas las métricas

✓ Precision@3 de 33.8% indica recomendaciones de alta calidad

✓ Hit Rate@3 de 63.9% demuestra cobertura efectiva

✓ Convergencia excelente sin overfitting

Próximos Pasos:



Monitorear performance mensualmente



Reentrenar con datos nuevos cada periodo de tiempo por ejemplo cada 2 meses o trimestre



Explorar features adicionales si es necesario



Considerar ensemble con otros modelos en futuro