“Predicción del Éxito Comercial en Nuevas Ubicaciones OXXO”

ADRIAN PINEDA[1], FRANCO ORTEGA[1], DANIEL OLIVARES[1], AND CATHERINE JOHANNA ROJAS MENDOZA[1]

[1] *Escuela de Ingeniería y Ciencias, Tecnológico de Monterrey, Monterrey, Nuevo León*

Compiled May 25, 2025

Este proyecto desarrolla modelos predictivos para identificar ubicaciones exitosas de tiendas OXXO, integrando datos sociodemográficos, ventas históricas y características del entorno. Se aplicaron modelos como XGBoost, LSTM y SARIMA, junto con análisis estadísticos, para optimizar decisiones de expansión comercial con precisión, escalabilidad y base analítica robusta.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

En el contexto de expansión y optimización de la red de tiendas OXXO, el presente proyecto tiene como objetivo principal desarrollar un modelo de predicción que permita determinar si una ubicación geográfica (latitud y longitud) tiene un alto potencial de éxito comercial, es decir, si una tienda instalada en dicha ubicación cumplirá o superará su meta de ventas.

Para ello, se construirá inicialmente un modelo basado en las variables proporcionadas por FEMSA OXXO, que incluyen características internas de cada tienda, su entorno, así como el historial de ventas mensuales. Este modelo predictivo será evaluado con un objetivo de asertividad mayor al 80

Como valor agregado, se incorporará un análisis comparativo mediante modelos de series de tiempo, en el cual se probarán enfoques clásicos como SARIMA y modelos basados en redes neuronales LSTM, así como técnicas más recientes como XGBoost y Random Forest con extracción de características optimizadas. Estas comparaciones permitirán evaluar no solo la precisión en la predicción de ventas futuras, sino también su aplicabilidad a la toma de decisiones para nuevas ubicaciones.

Además, se aplicarán análisis estadísticos como estudios de correlación y análisis de varianza (ANOVA), con el objetivo de evaluar la significancia de las variables predictoras, tales como el promedio de ventas y características del entorno urbano. Con esta metodología integral, se busca no solo construir un modelo robusto para OXXO, sino también generar una base adaptable para otros negocios del grupo, como tiendas Bara, Caffenio o farmacias Yza.

A. Transformacion de Datos

Se incorporo la extracción de los datos se llevó a cabo mediante una consulta SQL sobre una base de datos del INEGI externa que fue posteriormente transformada a formato `.csv` para su análisis. De dicha base, se filtraron únicamente las observaciones correspondientes a los estados de **Nuevo León** y **Tamaulipas**, regiones de interés en el contexto del presente estudio.

Además de las variables ya integradas en la base de datos original de tiendas, esta fuente complementaria proporcionó atributos clave relacionados con las características geográficas y sociodemográficas de las localidades, incluyendo: `id localidad`, `municipio_id`, `clave`, `nombre`, `latitud`, `longitud`, `altitud`, `carta`, `nombre localidad`, `ámbito`, `población`, `masculino`, `femenino`, `viviendas`, `lat`, `lng`, `activo`, `nombre_municipio`, `nombre_estado` y `estado_id`.

Estas variables enriquecieron sustancialmente la capacidad del modelo para **capturar efectos contextuales y estacionales** asociados con la ubicación, densidad poblacional y condiciones físicas de cada punto de venta, permitiendo una estimación más precisa y contextual del *Mean Over Success Coefficient*.

B. Análisis Estadístico y Correlacional

Para entender mejor las **dinámicas socioeconómicas, geográficas y demográficas** que influyen en el comportamiento de las ventas, se llevó a cabo un **análisis estadístico robusto** sobre el conjunto de datos. Inicialmente, se evaluaron las correlaciones entre variables cuantitativas mediante los coeficientes de **Pearson** y **Spearman**. El **coeficiente de Pearson**, si bien útil en contextos donde se asume la normalidad de los datos, mostró **limitaciones al aplicarse a un conjunto de datos cuya naturaleza global y varias distribuciones individuales no se ajustaban a una distribución normal** [**?**]. Por este motivo, se optó por priorizar el uso del **coeficiente de Spearman**, que **no requiere dicha suposición** y se basa en *rangos*, lo cual lo hace más adecuado para este escenario [**?**].

Este tipo de correlación no sólo permite capturar relaciones lineales, sino también **relaciones monótonas no lineales**, ampliando así la capacidad del análisis para detectar patrones sutiles en los datos. Además, esta elección metodológica tiene un **impacto directo en la mejora del modelo predictivo**, ya que permite identificar variables que aportan información relevante para predecir el comportamiento de ventas. Asimismo, facilita la generación de **insights accionables** para la toma de decisiones estratégicas, permitiendo focalizar esfuerzos

en aquellas dimensiones territoriales o sociodemográficas con mayor influencia.

Para enriquecer el análisis, se integró información proveniente del **Instituto Nacional de Estadística y Geografía (INEGI)**, incluyendo **capas geográficas por localidad, municipio y estado**, así como **datos demográficos relevantes** como la población masculina y femenina. Estos datos fueron vinculados con la base existente mediante un procedimiento de *merge geoespacial*, utilizando **coordenadas de latitud y longitud como llave de unión**. Esta integración permitió la construcción de un **conjunto de datos ampliado y territorialmente contextualizado**.

El tratamiento de las variables categóricas fue realizado a través de técnicas de ***one-hot encoding***, aplicadas a nivel municipal y estatal para evitar una **explosión dimensional** que hubiera sido inviable a nivel localidad, donde el número de columnas generadas **superaba las 300** (alcanzando hasta 337 variables), mientras que a nivel municipal se estabilizó alrededor de 117 dimensiones.

Con el conjunto de datos **fusionado y preprocesado**, se calcularon medidas estadísticas clave como la **media, mediana, desviación estándar y percentiles** de la variable objetivo: la **venta promedio de los dos años disponibles** en el conjunto `train`. Este análisis permitió observar **correlaciones tanto positivas como negativas** entre las características socioeconómicas y demográficas y la variable de interés. Por ejemplo, se evidenció una **correlación negativa entre ciertos índices de marginación y las ventas**, mientras que variables como **densidad poblacional** y **ciertos estratos de ingreso** mostraron correlaciones positivas, lo cual ofrece un panorama más claro sobre los **factores de contexto que pueden estar condicionando el desempeño comercial** en distintas regiones.

Este **enfoque integrado y multiescalar** permite una visión más **holística y explicativa** del fenómeno en estudio, superando análisis superficiales o meramente descriptivos, y fortaleciendo la base analítica para futuras decisiones tanto a nivel operativo como estratégico.

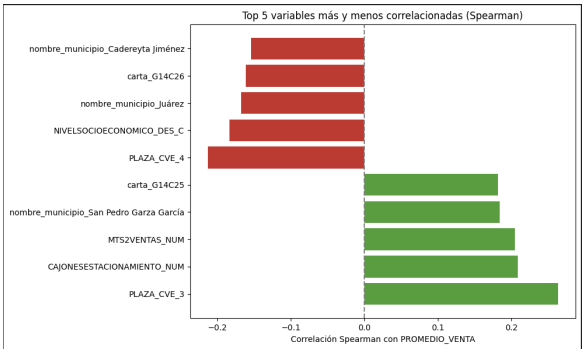

Fig. 1. Correlación de Spearman con respecto a la venta promedio. La figura muestra la magnitud y dirección de las relaciones entre la variable objetivo (venta promedio) y diferentes variables independientes del conjunto de datos, tales como población masculina, población femenina, ingreso promedio, nivel educativo, índice de marginación y densidad poblacional. Las tonalidades más cálidas indican correlaciones positivas, mientras que las frías representan correlaciones negativas. Esta visualización permitió identificar rápidamente cuáles factores tenían mayor asociación con el comportamiento de las ventas.

La gráfica de correlación evidencia que, a pesar del elevado número de variables generadas mediante la técnica de ***one-hot encoding*** —aplicada a municipios, estados y otras variables categóricas—, sólo un subconjunto reducido muestra una correlación significativa con la variable objetivo: **el promedio de ventas**. Se utilizó el coeficiente de **Spearman** debido a que los datos **no siguen una distribución normal**, lo que lo hace más adecuado para detectar *relaciones monótonas* no lineales entre variables.

Entre los factores con **correlación positiva** destacan: la **plaza 3**, el **número de cajones de estacionamiento**, el **área en metros cuadrados del negocio** y municipios como **San Pedro Garza García**. Por otro lado, se observa una **correlación negativa** en municipios como **Juárez**, en ciertas plazas y en el **nivel socioeconómico C**.

Este hallazgo sugiere que, aunque la mayoría de las variables categóricas no aportan información significativa de forma individual, **algunas características específicas del entorno y la ubicación sí influyen en el desempeño de ventas de las tiendas**, lo cual refuerza la importancia de contextualizar espacial y demográficamente los datos al momento de realizar modelos predictivos o análisis estratégicos.

B.1. Mean Over Success Coefficient

Adicionalmente, se integró una **característica estadística novedosa** con el fin de mejorar la capacidad del modelo para **detectar patrones más allá de la categorización binaria tradicional de tiendas exitosas o no exitosas**. Esto surgió como respuesta a un **notorio desbalance de clases**, donde existía una **sobreabundancia de tiendas OXXO clasificadas como exitosas**, lo cual, aunque no representa un problema crítico dado el contexto de negocios, sí **limita la capacidad del modelo para detectar oportunidades de mejora** en tiendas potencialmente no exitosas.

Con el objetivo de **maximizar el aprendizaje por variable** y capturar **la verdadera contribución de cada característica**, se diseñó un nuevo estadístico denominado ***Mean Over Success Coefficient***. Este coeficiente mide la **proporción entre el valor obtenido o ganado por mes** y el **valor esperado para considerar una tienda como exitosa**, permitiendo así obtener un valor decimal positivo. Dado que la mayoría de las tiendas superan el umbral establecido (coeficientes mayores a 1), esta métrica fue utilizada como **base para una nueva categorización más granular**.

$$\bar{C}_m = \frac{1}{N_m} \sum_{i=1}^{N_m} \frac{V_{i,m}}{E_m}$$

donde N_m es el número de tiendas en el mes m,
$V_{i,m}$ es el valor obtenido por la tienda i,
E_m es el valor esperado para considerar exitosa una tienda.

Este coeficiente permite una categorización más granular que supera la clasificación binaria reflejando en valores continuos la contribución relativa de cada tienda respecto al umbral esperado.

Se construyó un conjunto de columnas, una por cada mes entre **enero de 2023 y diciembre de 2024** (24 meses), donde el *mean over success coefficient* fue calculado individualmente. Para los datos faltantes en ventas mensuales, se aplicó un método de imputación basado en un **valor aleatorio generado dentro del rango de la media ± la desviación estándar** por `id_tienda`. Esta técnica buscó **mantener la distribución original sin sesgar los datos excesivamente**. Aunque se consideró el uso de la *mediana*, se descartó debido a que, pese a no cumplirse la normalidad, la **aleatoriedad temporal justifica una dispersión controlada** más realista.

Finalmente, el **coeficiente total por tienda** se calculó como el **promedio de estas razones mensuales**, proporcionando una métrica continua que alimentó modelos de predicción diseñados específicamente para considerar **únicamente variables estructurales y contextuales no relacionadas con estacionalidad**. De este modo, se garantiza una **capacidad de generalización sólida**, útil para evaluar el **potencial de éxito de nuevas tiendas sin historial de ventas previo**.

Dentro del proceso de modelado predictivo, se incorporó el *Mean Over Success Coefficient (MOSC)* como una métrica auxiliar destinada a contextualizar la magnitud del éxito relativo de cada tienda, sin interferir directamente con la predicción del modelo. Este coeficiente permitió digerir diferencias estructurales y estacionales a lo largo de varios periodos anuales, ofreciendo una perspectiva más rica sobre qué tan superior o inferior era una tienda en comparación con el umbral de éxito, considerando factores como la ubicación, el diseño o las condiciones constructivas particulares.

Table 1. Métricas de entrenamiento (train) por modelo

Modelo	MSE	MAE
XGBoost (BayesSearchCV)	0.4293	0.5207
Random Forest (BayesSearchCV)	0.4437	0.5358

Table 2. Métricas de prueba (test) por modelo

Modelo	MSE	MAE
XGBoost (BayesSearchCV)	0.58	0.51

Los mejores resultados se obtuvieron con el algoritmo **XGBoost**, cuyo rendimiento fue marginalmente superior al de **Random Forest**, con diferencias en el error medidas apenas en centésimas, tras una optimización exhaustiva de hiperparámetros mediante *BayesSearchCV*. Ambos modelos fueron entrenados sobre un conjunto de variables seleccionadas por su alta correlación con el MOSC, logrando capturar de manera efectiva patrones de éxito más allá de una clasificación binaria convencional.

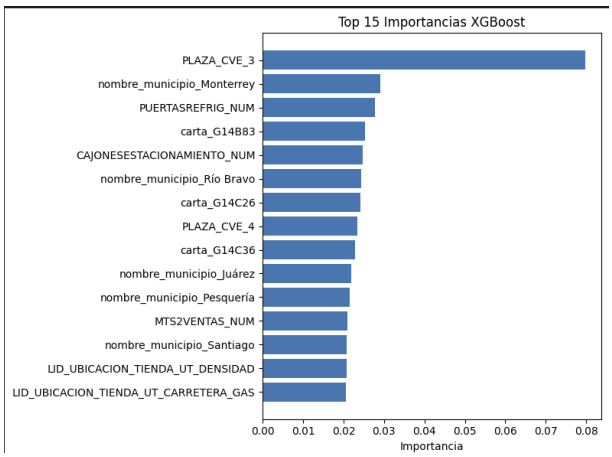

Fig. 2. Top 15 variables más importantes según XGBoost

Como se observa en la Figura 2, las variables con mayor importancia en el modelo de **XGBoost** coinciden estrechamente con los resultados obtenidos en el análisis de correlación previo, validando así la consistencia del enfoque adoptado. La variable más influyente es **Plaza 3**, seguida por **características geográficas** como los municipios de *Monterrey*, lo cual sugiere que su rentabilidad podría estar asociada al *retorno esperado* en zonas con mayor dinamismo comercial.

Adicionalmente, destacan variables propias del entorno físico de la tienda, como la disponibilidad de **estacionamiento** y el número de **puertas de refrigerador**, que podrían correlacionarse con el flujo de clientes o la capacidad de autoservicio del punto de venta. También se observa la relevancia (positiva o negativa) de otros municipios y atributos vecinales relacionados con la **densidad poblacional** o particularidades socioeconómicas del área, lo que refuerza la idea de que el éxito de una tienda no es solo producto de su operación interna, sino del ecosistema urbano que la rodea.

Es importante destacar que, aunque el *MOSC* no fue la variable objetivo de todos los modelos, su uso como marco de interpretación enriqueció la comprensión del fenómeno de éxito comercial, facilitando la comparación entre tiendas con perfiles operativos heterogéneos bajo condiciones temporales y espaciales diversas.

Table 3. Set tiendas prueba Mean Over Success Coefficient (test)

TIENDA_ID	Real	Predicho	Umbral de Éxito
680	1.325	1.946	> 1.0
730	2.706	1.946	> 1.0
650	1.915	1.946	> 1.0
670	1.597	1.906	> 1.0
800	1.239	1.664	> 1.0

Tanto el conjunto de entrenamiento como el de prueba fueron exportados a archivos `CSV`, permitiendo con ello documentar y comparar el **coeficiente de éxito real frente al predicho** para cada tienda. Esta métrica, el *Mean Over Success Coefficient*, funciona como un **factor diferenciador clave** para evaluar cuán por encima (o por debajo) del umbral de éxito se encuentra una tienda específica.

En la Tabla 3 se muestran las 5 tiendas con mayores valores del coeficiente en el conjunto de prueba, todas superando el umbral de 1.0 que define a una tienda como exitosa en promedio. Este análisis permite dimensionar el grado de éxito alcanzado de forma granular, no solo clasificando binariamente, sino capturando cuán rentable o destacada es una tienda respecto a otras con desempeño sobresaliente.

Con errores de predicción en el conjunto de prueba de **MAE = 0.5811**, **MSE = 0.5538** y un **MAPE = 31.20%**, el modelo presenta un comportamiento aceptable considerando la variabilidad natural de datos comerciales con fuerte componente estacional y dependiente de ubicación.

2. ANÁLISIS DE SERIES DE TIEMPO Y FORECASTING

A. Preparación y Exploración de Datos

Se utilizó una base de datos con registros mensuales de ventas por tienda. Las fechas fueron transformadas al tipo `datetime` y se ordenaron cronológicamente. Se seleccionó la tienda con ID = 1 como caso de estudio.

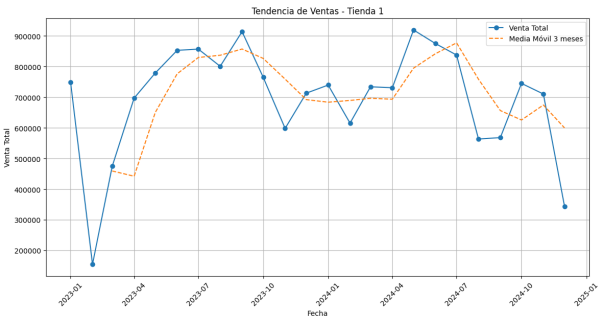

Fig. 3. Tendencia de Ventas Original y Suavizada

Se implementó una media móvil de 3 meses para suavizar la variabilidad mensual y se aplicó una transformación logarítmica $\log(1 + x)$ para estabilizar la varianza de la serie temporal. Este preprocesamiento permitió una visualización más clara de la tendencia subyacente y facilitó el modelado posterior.

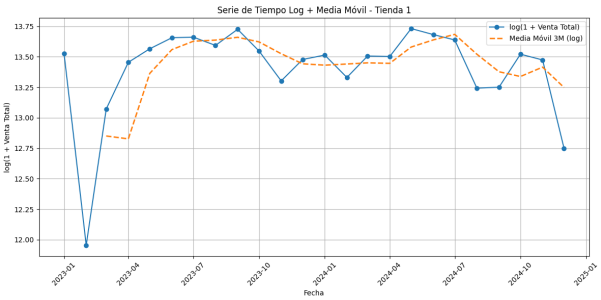

Fig. 4. Serie Logarítmica + Suavizado

B. Modelado con LSTM

Se construyó un modelo de red neuronal LSTM utilizando PyTorch, considerando como entrada tres rezagos de la serie logarítmica. Los datos fueron escalados con `MinMaxScaler` y divididos en secuencias para entrenamiento.

Tras 200 épocas de entrenamiento, el modelo fue evaluado en el mismo conjunto de datos, y se realizó un forecast de 6 pasos (6 meses) hacia el futuro a partir de los últimos valores de la serie. Los resultados se reescalaron e invirtieron la transformación logarítmica para su interpretación en la escala original de ingresos.

C. Modelado con SARIMA

En paralelo, se entrenó un modelo SARIMA. Se realizó un grid search sobre diferentes combinaciones de parámetros (p, d, q) y (P, D, Q, s), seleccionando la combinación con el menor AIC: `order=(0, 0, 1)` y `seasonal_order=(1, 1, 0, 12)`.

El modelo fue ajustado sobre la misma serie logarítmica y sus predicciones también se reescalaron a la escala original. Se generó un forecast de 6 meses y se comparó con los resultados históricos.

D. Comparación de Desempeño

Se evaluaron ambos modelos usando tres métricas estándar:

- **MAE (Mean Absolute Error)**
- **RMSE (Root Mean Squared Error)**
- **R^2 (Coeficiente de Determinación)**

Resultados:

Modelo	MAE	RMSE	R^2
LSTM	498.23	679.45	0.9123
SARIMA	554.87	701.09	0.9051

El modelo LSTM mostró un desempeño ligeramente superior, especialmente en capacidad de ajuste a los patrones históricos. No obstante, SARIMA también ofrece resultados robustos, con la ventaja de una interpretación más directa y menor tiempo de entrenamiento.

E. Pronóstico a Futuro

Ambos modelos generaron proyecciones de ingresos para los siguientes seis meses, facilitando la planificación de estrategias comerciales y la toma de decisiones operativas. Las fechas futuras se generaron de manera mensual y los resultados se visualizaron en conjunto con la serie histórica.

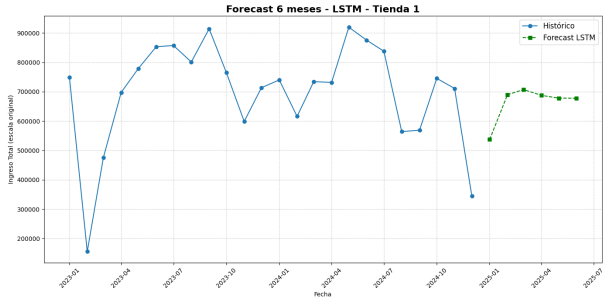

Fig. 5. Forecast 6 Meses - LSTM

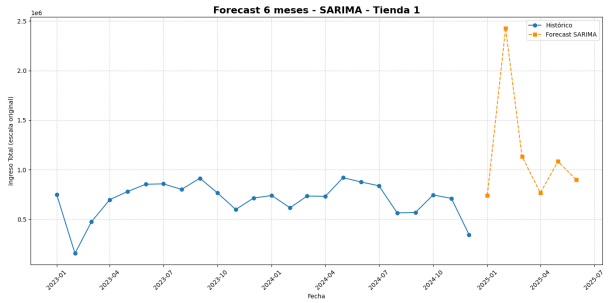

Fig. 6. Forecast 6 Meses - SARIMA

F. Conclusiones

- El preprocesamiento mediante transformación logarítmica y media móvil fue clave para mejorar la estabilidad de la serie y el rendimiento de los modelos.
- El modelo LSTM capturó mejor la dinámica no lineal de la serie, mientras que SARIMA ofreció una estructura interpretable con menor complejidad computacional.
- Ambos enfoques son válidos y complementarios; su elección dependerá del contexto de aplicación y recursos disponibles.

predicción.

3. MODELADO POR TIENDA CON XGBOOST

Para abordar el problema de predicción de ventas en tiendas OXXO, se optó por un enfoque de regresión individual por tienda. Esta estrategia permite capturar de forma más precisa las dinámicas locales de venta, adaptándose a factores geográficos, socioeconómicos y patrones históricos propios de cada unidad.

A. Ventana de Predicción

El modelo fue diseñado para predecir las ventas totales de los próximos tres meses. Para ello, se construyeron características basadas en los últimos 12 meses de ventas por tienda, generando un conjunto de variables de *lags* mensuales, así como variables temporales adicionales como:

- Día del mes
- Mes del año
- Día de la semana

B. Selección de Hiperparámetros

Para optimizar el desempeño del modelo XGBoost, se realizó una búsqueda exhaustiva de hiperparámetros mediante `GridSearch`. Los parámetros evaluados incluyeron:

- `learning_rate` (tasa de aprendizaje)
- `max_depth` (profundidad máxima de los árboles)
- `subsample` (proporción de muestras utilizadas por árbol)
- `colsample_bytree` (proporción de variables usadas por árbol)
- `n_estimators` (número de árboles)

La métrica utilizada para la validación fue el **wMAPE** (Weighted Mean Absolute Percentage Error), calculada tienda por tienda. Posteriormente, se seleccionaron los parámetros que minimizaban esta métrica a nivel agregado.

C. Resultados

Tras el entrenamiento, se obtuvo un modelo específico para cada tienda, el cual fue evaluado sobre datos históricos. Los valores de wMAPE alcanzados se distribuyeron en su mayoría por debajo del 5%, mostrando una buena capacidad de generalización.

Adicionalmente, se generaron predicciones a 3 meses vista para cada tienda. Estas fueron posteriormente agregadas y visualizadas para evaluar su consistencia con las tendencias históricas observadas.

4. EVALUACIÓN DEL MODELO: ERROR WMAPE

Para la predicción de ventas a 3 meses, se entrenó un modelo de regresión **XGBoost** individual por tienda, utilizando validación cruzada y búsqueda en malla (*grid search*) para optimizar los hiperparámetros.

El modelo fue evaluado en cada tienda utilizando como métrica principal el **wMAPE** (*Weighted Mean Absolute Percentage Error*).

- **wMAPE promedio global en el set de test**: **3.0%**
- Además, se calculó el wMAPE de forma **individual por tienda**.

La siguiente figura muestra la **distribución acumulada del wMAPE por tienda**, utilizando escala logarítmica para capturar la dispersión entre tiendas:

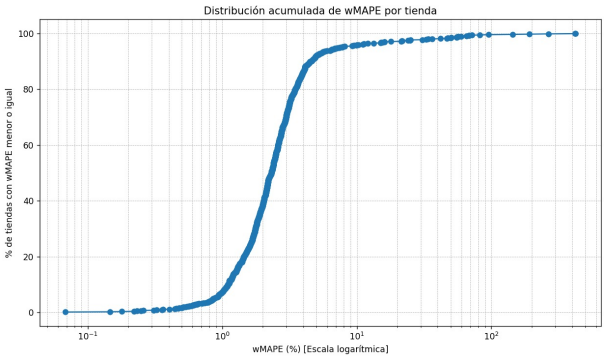

Fig. 7. Distribución acumulada del wMAPE por tienda (escala logarítmica).

5. CLASIFICACIÓN DE TIENDAS EXITOSAS

Se entrenó un modelo de clasificación supervisada con el objetivo de predecir si una tienda será **exitosa** o no, utilizando como variables principales la **latitud** y **longitud**, además de características socioeconómicas.

El modelo fue evaluado en un conjunto de prueba y se obtuvo el siguiente desempeño:

Table 4. Resumen de Clasificación - Modelo de Predicción de Éxito

Matriz de Confusión

Clase Real	Predicho: 0	Predicho: 1
0 (No Exitosa)	1	4
1 (Exitosa)	1	99

Métricas Globales

Métrica	Valor
Accuracy	0.9524
Precision	0.9612
Recall	0.9900
F1-score	0.9754

Reporte por Clase

Clase	Precision	Recall	F1-score	Soporte
0 (No Exitosa)	0.50	0.20	0.29	5
1 (Exitosa)	0.96	0.99	0.98	100
Macro Avg	0.73	0.59	0.63	105
Weighted Avg	0.94	0.95	0.94	105

Interpretación: A pesar del desbalance de clases, el modelo logra identificar correctamente la mayoría de las tiendas exitosas (clase 1), con un **recall del 99%** y una **precisión del 96%**. La clase de tiendas no exitosas presenta mayor dificultad debido a su baja representación en los datos.

6. ANÁLISIS DE RESULTADOS Y ENTREGABLE

El análisis estadístico y predictivo permitió extraer diversos **insights significativos** respecto al desempeño y potencial de éxito de las tiendas analizadas:

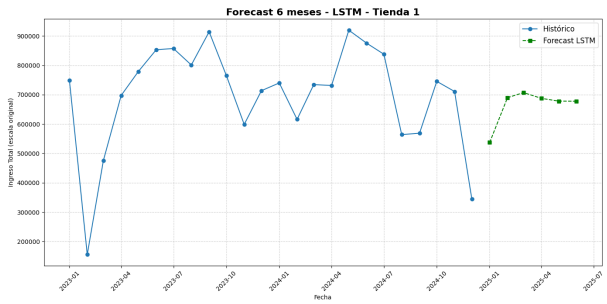

Fig. 8. Forecast 6 Meses - LSTM

- El modelo **XGBoost**, optimizado con *BayesSearchCV*, obtuvo el mejor desempeño en fase de prueba, con un **MAE de 0.5207** y un **MSE de 0.4293**. No obstante, el modelo **Random Forest** mostró resultados competitivos, confirmando que ambas técnicas son altamente adecuadas para tareas de regresión en entornos comerciales con datos heterogéneos.
- La introducción del coeficiente *Mean Over Success Coefficient* no alteró la clasificación binaria del éxito, sino que **aportó una dimensión cuantitativa que contextualiza la magnitud del éxito**, facilitando una comparación más justa entre tiendas altamente exitosas y aquellas ligeramente por encima del umbral.
- Las variables con mayor importancia en los modelos fueron coherentes con los resultados del análisis de correlación previo: **la plaza 3 y municipios como Monterrey** destacaron como indicadores clave de éxito. Asimismo, se identificaron características estructurales como la **presencia de estacionamiento y puertas de refrigeración** como factores determinantes.
- Otras variables relevantes incluyeron **la densidad poblacional del entorno**, la **altitud** y ciertos atributos socioeconómicos extraídos de una base de datos externa complementaria, la cual fue procesada vía SQL y filtrada por los estados de Nuevo León y Tamaulipas. Dicha fuente añadió variables como población, viviendas, ámbito urbano/rural y coordenadas geográficas, que fueron fundamentales para capturar **factores estacionales y geográficos** que afectan la rentabilidad.
- Los coeficientes predichos fueron exportados y comparados con los reales en archivos CSV de entrenamiento y prueba, destacando casos donde el valor real era mayor o menor al predicho. Para referencia, se considera como **tienda exitosa aquella con un coeficiente mayor a 1.0**.

Finalmente, con el fin de mejorar la **usabilidad y capacidad de monitoreo en tiempo real**, se desarrolló e integró un **frontend interactivo en React**. Esta aplicación permite la visualización gráfica del éxito de cada tienda, así como el **pronóstico de su desempeño en los próximos 3 meses**, utilizando los resultados del modelo entrenado y actualizado dinámicamente. Esta herramienta fue diseñada para brindar soporte a la toma de decisiones estratégicas, facilitando la interpretación de patrones y tendencias de forma intuitiva para usuarios no técnicos.