ML para optimizar pronósticos de ventas de refrescos con segmentación y variables climáticas

Marcos Aquino, Natalia Olvera, Ericka Rodríguez, Carol Rendón & Rolando Ruiz

Emails: {A00835576, A01285367, A01571463, A01425341, A00835733}@tec.mx

Tecnológico de Monterrey – TC3007C.102 Inteligencia Artificial para la Ciencia de Datos II, Monterrey, N.L.

Noviembre de 2025

Resumen

El presente proyecto tiene como objetivo mejorar la predicción de ventas de refrescos familiares en canales tradicionales de venta (abarrotes, kioskos y tienditas) mediante la aplicación de modelos de aprendizaje máquina que incorporan variables climáticas, particularmente la temperatura. Para ello, se cuenta con datos históricos de ventas, información meteorológica asociada a cada canal de venta y las predicciones originales de la empresa, las cuales se busca optimizar mediante la inclusión de variables ambientales. La evaluación del desempeño de los modelos se realizará utilizando el error medio absoluto (MAE) como métrica de comparación. El proyecto es factible gracias a la disponibilidad de datos suficientes para las etapas de entrenamiento, validación y prueba, así como al uso de herramientas de análisis predictivo como Python, Pandas y Scikit-Learn. El equipo de desarrollo está conformado por estudiantes de ingeniería en ciencia de datos, bajo la supervisión de profesores y con retroalimentación directa del socio formador. La iniciativa es relevante, ya que contribuye directamente a mejorar la toma de decisiones en la gestión de inventarios, reduciendo tanto el exceso como el desabasto, y en la planeación de la distribución de refrescos en los canales tradicionales. La metodología propuesta consta de cinco etapas: (1) análisis exploratorio de los datos, (2) agrupamiento de regiones con condiciones climáticas similares, (3) desarrollo de un modelo simple como línea base, (4) construcción de modelos de aprendizaje máquina como regresión lineal y XGBoost, y (5) comparación y evaluación de los resultados obtenidos.

Palabras clave: key, words, palabras, clave

1 Introducción

En la industria de bebidas, la precisión en los pronósticos de venta es un factor crítico para la eficiencia operativa y la rentabilidad. Los márgenes suelen ser estrechos y la demanda presenta una alta variabilidad, influenciada por múltiples factores externos. Los errores en la previsión, ya sea por desabastecimiento o *sobrestock*, generan consecuencias directas en la producción, la distribución y la satisfacción del cliente, afectando la competitividad de toda la cadena de valor.

La cadena de suministro en este sector puede describirse como un flujo continuo desde la embotelladora, que planifica la producción y gestiona inventarios; pasando por el distribuidor, encargado de la logística y asignación de volúmenes; hasta llegar al punto de venta, donde se enfrenta la demanda real del consumidor final de los clientes. En este sistema interconectado, las decisiones tomadas en etapas tempranas dependen fuertemente de la capacidad para anticipar la demanda futura con la mayor precisión posible.

Diversos factores son conocidos por afectar el consumo de bebidas: las condiciones climáticas (temperatura, precipitaciones, humedad), el precio y la disponibilidad de los productos, las estrategias de marketing y promociones, y los hábitos de consumo asociados a la salud y al estilo de vida. No obstante, los modelos de pronóstico actualmente implementados en muchas empresas del sector se basan principalmente en series históricas de ventas y variables comerciales, de-

jando de lado el impacto del clima, a pesar de su evidente influencia en la demanda, especialmente en productos de alta estacionalidad.

El problema que aborda este estudio consiste en mejorar la precisión de los pronósticos de venta B2B en la industria de bebidas mediante la incorporación de variables climáticas. Específicamente, se busca predecir la demanda semanal por categoría de producto a nivel de punto de venta, integrando datos históricos y condiciones meteorológicas locales. De esta forma, se pretende evaluar el valor agregado que ofrecen las variables climáticas en la predicción, comparando el desempeño de modelos tradicionales frente a modelos ampliados con información ambiental.

2 Planteamiento del problema

En la industria de bebidas, los modelos de pronóstico utilizados para planificar producción, distribución y abastecimiento enfrentan una fuerte limitación: aunque la demanda está estrechamente relacionada con las condiciones climáticas, los sistemas actualmente implementados por la empresa se basan casi exclusivamente en información histórica de ventas y variables comerciales. Como consecuencia, dichos modelos presentan errores sistemáticos en semanas donde ocurren variaciones meteorológicas significativas, tales como olas de calor, descensos abruptos de temperatura o periodos de lluvia prolongada.

El reto específico de este proyecto surge a partir de la información proporcionada por la embotelladora, la cual incluye: ventas reales por cliente y por *SKU* durante 2024, predicciones semanales generadas por un modelo interno, de naturaleza desconocida, y un conjunto detallado de variables climáticas observadas y pronosticadas (con horizontes de hasta seis días de anticipación), asociadas a distintas estaciones meteorológicas. Cada cliente está vinculado a una estación, lo que permite relacionar de manera directa la demanda con las condiciones ambientales locales.

El problema central consiste en determinar en qué medida la integración de variables climáticas puede mejorar la precisión del modelo original proporcionado por la empresa. Sin embargo, la mejora cuantitativa no es el único objetivo: también es necesario analizar los patrones detrás de los aciertos y fallas del modelo base, identificar qué variables meteorológicas explican con mayor fuerza las fluctuaciones de demanda y caracterizar los casos en los que el modelo interno se equivoca en mayor o menor medida. En otras palabras, se busca desarrollar un enfoque que combine precisión, simplicidad y un alto grado de explicabilidad.

3 Justificación

La inclusión de información climática en modelos de pronóstico de ventas está ampliamente sustentada en la literatura, especialmente en sectores de consumo altamente sensibles a la temperatura, la humedad o las precipitaciones. En productos como sodas, aguas y bebidas isotónicas, el clima influye directamente en la intensidad y frecuencia del consumo. No obstante, el sistema actual de la empresa, basado en series históricas comerciales, no captura adecuadamente estas variaciones, lo que limita su capacidad para anticipar picos o caídas abruptas de demanda.

Además, el impacto del clima no es homogéneo en todas las regiones del país. Las estaciones meteorológicas presentan comportamientos diferenciados, por lo que la segmentación mediante técnicas de agrupamiento permite identificar tipologías de clima similares y construir modelos más representativos y robustos. Esto no solo mejora la calidad del pronóstico, sino que también reduce la complejidad del sistema al evitar la necesidad de desarrollar un modelo completamente independiente para cada cliente.

La relevancia de este proyecto es tanto operativa como estratégica. En un nivel operativo, contar con pronósticos más precisos contribuye a optimizar inventarios, reducir costos logísticos y minimizar pérdidas por desabasto o exceso de producto. Desde una perspectiva estratégica, comprender la relación entre clima y demanda permite a la empresa anticipar tendencias, planificar con mayor certidumbre y fortalecer su ventaja competitiva frente a competidores que no integran variables ambientales en sus modelos.

Finalmente, incluso en el escenario en que la

incorporación de clima no genere mejoras sustanciales en la precisión global, el análisis detallado de los errores del modelo interno (cuándo falla, por qué y bajo qué condiciones) aporta un valor significativo para la toma de decisiones y para el diseño de futuros sistemas de predicción más transparentes, interpretables y confiables.

Este estudio será realizado con base en los productos CRFR y CRFNR en los subcanales de mayor concurrencia: Hogar con tienda, abarrotes y estanquillos.

4 Marco Teórico

Con el fin de sustentar la incorporación de variables climáticas en modelos de pronóstico de ventas B2B dentro de la industria de bebidas, es necesario revisar los principales avances en torno al *forecasting* de demanda y el uso de información externa en series temporales. La literatura existente abarca diversas líneas de investigación que aportan perspectivas complementarias: los modelos de pronóstico utilizados en retail y bienes de consumo masivo (CPG), los estudios que han analizado la elasticidad entre clima y consumo de bebidas, los métodos empleados para integrar datos externos en modelos de predicción, y las estrategias de *feature engineering* que permiten representar de forma adecuada las condiciones meteorológicas. Esta revisión busca identificar las contribuciones más relevantes en cada una de estas áreas, así como las brechas existentes que justifican la presente investigación.

El pronóstico de la demanda es un componente esencial en la planeación operativa de las empresas manufactureras y de retail, ya que permite optimizar inventarios y niveles de servicio al cliente. Tradicionalmente, los modelos estadísticos como ARIMA o los de suavizamiento exponencial fueron la base del *forecasting*; sin embargo, en años recientes se ha incrementado el uso de algoritmos de aprendizaje automático para capturar patrones no lineales y dependencias complejas en los datos. En este contexto, resulta relevante considerar el paradigma del aprendizaje supervisado, el cual se caracteriza por utilizar pares entrada–salida etiquetados para entrenar un modelo capaz de realizar predicciones precisas. A diferencia del aprendizaje tradicional, donde el usuario define explícitamente las reglas, el aprendizaje supervisado permite que estas se generen automáticamente a partir de los datos, aumentando la objetividad y reduciendo el sesgo en los procesos de toma de decisiones (Verma, Nagar, y Mahapatra, 2021). Debido a su capacidad para obtener resultados precisos en tareas de regresión, este enfoque se ha consolidado como uno de los más utilizados en problemas de pronóstico de series temporales.

En este contexto, Mitra, Jain, Kishore, y Kumar (2022) realizaron un estudio comparativo de cinco técnicas de regresión basadas en *machine learning* (*Random Forest*, *XGBoost*, *Gradient Boosting*, *AdaBoost* y *ANN*) frente a un modelo híbrido propuesto (RF–XGBoost–LR) para el pronóstico de ventas semanales en una cadena de Estados Unidos. Los autores consideraron variables como la temperatura de la región y el tamaño de las tiendas, observando que el modelo híbrido superó a los demás en métricas de desempeño. Este trabajo demuestra la capacidad de los enfoques combinados para capturar mejor la variabilidad de la demanda en entornos complejos.

De manera complementaria, Islek y Oguducu (2015) propusieron una metodología para el pronóstico de demanda en almacenes de distribución que combina técnicas de minería de datos con aprendizaje automático. Su enfoque agrupa los almacenes con comportamientos de venta similares mediante un algoritmo de clusterización y, posteriormente, aplica un modelo híbrido que combina promedios móviles con redes bayesianas. Los resultados muestran una mejora sustancial en la precisión de las estimaciones, especialmente en contextos con múltiples centros de distribución y amplios catálogos de productos. Este tipo de metodologías resalta la importancia de incorporar enfoques jerárquicos y combinaciones de modelos para capturar diferencias sustanciales en las regiones disponibles.

Diversos estudios han mostrado que las condiciones climáticas, en particular las temperaturas extremas y las precipitaciones, influyen significativamente en los patrones de consumo de bebidas. En este sentido, Ho, Ko, Liu, y Wu (2024)

analizaron la relación entre las expectativas de condiciones meteorológicas extremas y el comportamiento de compra de productos de consumo masivo (FMCG). Los autores encontraron que la anticipación de eventos meteorológicos extremos incrementa de manera significativa el consumo, impulsado por mecanismos psicológicos como la percepción de escasez futura. Pese a centrarse principalmente en eventos extremos, los hallazgos ofrecen evidencia de que las variables climáticas no sólo afectan el comportamiento del consumidor desde una perspectiva física (temperatura o lluvia), sino también desde una psicológica, vinculada a la percepción del riesgo y la disponibilidad futura.

Más allá de estos efectos situacionales, existe evidencia adicional de que parámetros climáticos cotidianos, como la radiación solar o la temperatura, influyen en patrones más finos del comportamiento del consumidor. En particular, Tian, Zhang, y Zhang (2018) demostraron que menores niveles de luz solar incrementan la propensión de los consumidores a buscar variedad, mientras que las temperaturas altas tienden a aumentar dicha conducta en comparación con temperaturas frías. Estos hallazgos resultan relevantes para la industria de bebidas, pues sugieren que no solo los extremos climáticos, sino también variaciones moderadas del clima diario, pueden modificar las decisiones de compra y la composición del portafolio demandado.

La incorporación de variables exógenas en modelos de series temporales se ha consolidado como una estrategia eficaz para mejorar la precisión del pronóstico en diversos sectores. En el ámbito energético, por ejemplo, Vu, Muttaqi, y Agalgaonkar (2015) demostraron que la selección adecuada de variables climáticas, como temperatura, humedad y días lluviosos, puede mejorar de forma notable el desempeño de los modelos de predicción de demanda eléctrica. Los autores propusieron un enfoque de regresión múltiple que combina el análisis de multicolinealidad y (*backward elimination*) para seleccionar las variables más relevantes, reduciendo así el ruido y la redundancia entre predictores.

El modelo, aplicado a datos mensuales de Australia, mostró errores de predicción dentro de márgenes aceptables, destacando la importancia de la selección rigurosa de variables exógenas. Aunque el estudio se centra en la electricidad, la metodología puede trasladarse al contexto del pronóstico de ventas en bebidas, donde las variables meteorológicas presentan correlaciones y multicolinealidades similares. En este sentido, la integración controlada de datos externos permite capturar patrones estacionales y variaciones locales que los modelos univariados no logran reflejar adecuadamente.

El tratamiento adecuado de las variables climáticas representa un aspecto importante para la mejora de modelos de pronóstico. No basta con incorporar directamente la temperatura o la precipitación, es necesario representar de forma más expresiva sus efectos sobre el comportamiento de consumo. Yilmaz (2024) abordaron este problema en el contexto del pronóstico de ventas de bebidas, proponiendo un conjunto de variables derivadas que capturan tanto la variabilidad diaria como los efectos estacionales del clima.

Entre las transformaciones más relevantes se incluyen: una variable binaria para registrar si hubo lluvia en el día; categorías discretas de temperatura para capturar no linealidades; una medida de desviación respecto a la temperatura media mensual para identificar condiciones inusuales; y un modificador de fin de semana que ajusta las variables meteorológicas según el comportamiento diferencial de consumo en fines de semana. Adicionalmente, se incorporaron promedios móviles con el fin de capturar efectos retardados y tendencias recientes. Este tipo de ingeniería de características no sólo permite reducir los errores de predicción, sino también entender cómo las condiciones meteorológicas afectan de forma diferenciada el consumo en distintos periodos y contextos.

5 Objetivos

En esta sección, se describirán tanto el Objetivo General como los Objetivos Específicos del Proyecto.

5.1 Objetivo General

Fortalecer la exactitud de los pronósticos de ventas de refrescos en tiendas de abarrotes y "tienditas" utilizando técnicas de aprendizaje automático supervisado, incorporando factores meteorológicos y análisis por regiones climáticas.

5.2 Objetivos Específicos

- Identificar las variables climáticas que tienen mayor afectación a la venta de refrescos en los distintos canales comerciales.
- Proponer un modelo de predicción de venta de refrescos que sea explicable y reutilizable, que mejore una predicción previa que no toma en cuenta variables climáticas.
- Establecer recomendaciones específicas de absatecimiento de producto por tienda, basadas en su tipo de clima.

6 Hipótesis

Integrar variables climáticas en el modelo de pronóstico mejora la precisión de las ventas de refrescos, ya que el clima influye de forma distinta según el producto y la región.

7 Metodología

La metodología seguida para incorporar las variables climáticas en el pronóstico de ventas se estructuró en varias etapas clave. Primero se segmentan las estaciones meteorológicas mediante un análisis de clústeres para identificar perfiles climáticos similares. Posteriormente, se detalla el proceso de ingeniería de datos aplicado para la integración de las variables semanales. Finalmente, se presentan y comparan los resultados de los modelos de ajuste implementados: un modelo base, una regresión lineal y un modelo XGBoost.

7.1 Clusterización por caracterización climática

Con el objetivo de agrupar las treinta y tres estaciones meteorológicas según sus patrones climáticos semanales, se desarrolló un proceso de clusterización basada en características climáticas. Este procedimiento se llevó a cabo en dos etapas principales: la preparación de los datos (la cual se hizo en el Entregable 1) y la aplicación de técnicas de agrupamiento jerárquico.

7.2 Preparación y procesamiento de datos

Para integrar la información temporal, se vinculó el dataset meteorológico con el calendario, incorporando atributos de día de la semana, semana y mes. De este procedimiento se obtuvieron dos archivos intermedios en formato `pickle`:

- **weather_complete.pkl:** conjunto completo de registros meteorológicos con variables de calendario.
- **weather_groupby.pkl:** agregación semanal del clima por estación meteorológica, filtrando únicamente los registros reales.

Esta agregación se realizó mediante un promedio semanal de las 12 variables climáticas para cada estación (`highest_temp`, `lowest_temp`, `avg_daily_all_hours`, ...), generando una caracterización temporal de cada estación. Los datos se transformaron para que cada estación fuera representada como una única fila, creando un vector que describe su perfil climático completo.

7.3 Clusterización jerárquica

Con los datos procesados en `weather_groupby.pkl`, se construyó una matriz en formato estación $\times$ (semana $\times$ variable), la cual representa la evolución semanal de las principales variables climáticas por estación.

Se utilizó *StandardScaler* para estandarizar las diferentes escalas, asegurando que contribuyeran de manera equitativa al análisis siguiente, posteriormente se calculó la matriz de distancias dinámicas (*Dynamic Time Warping*, DTW) entre estaciones, con el fin de capturar similitudes en la forma temporal de las series.

El agrupamiento final se realizó mediante un modelo de **Clustering Jerárquico Aglomerativo** (`Agglomerative Clustering`) con métrica

euclidiana y enlace de tipo Ward, seleccionando seis clústeres como número óptimo de grupos según el análisis de inercia y el coeficiente de silueta.

Adicionalmente, se exploraron configuraciones alternativas utilizando el *método del Codo* (*Elbow Method*) y el índice de *Silhouette* con el algoritmo `K-Means`, para comparar la estabilidad de los resultados. Los valores de silueta oscilaron entre 0.15 y 0.32 para distintos números de clústeres, mientras que el modelo jerárquico alcanzó una silueta de 0.074 (calculada sobre la matriz de distancias DTW), reflejando una separación moderada entre grupos, coherente con la naturaleza continua de los datos climáticos.

Los resultados del agrupamiento se almacenaron en el archivo `stat_weather.pkl` (renombrado posteriormente como `cluster_agglomerative.pkl`), el cual asocia cada estación meteorológica con el clúster correspondiente.

Finalmente, se visualizaron las series semanales de temperatura promedio de las estaciones pertenecientes a cada clúster, evidenciando diferencias en la estacionalidad y la magnitud de las temperaturas promedio entre grupos.

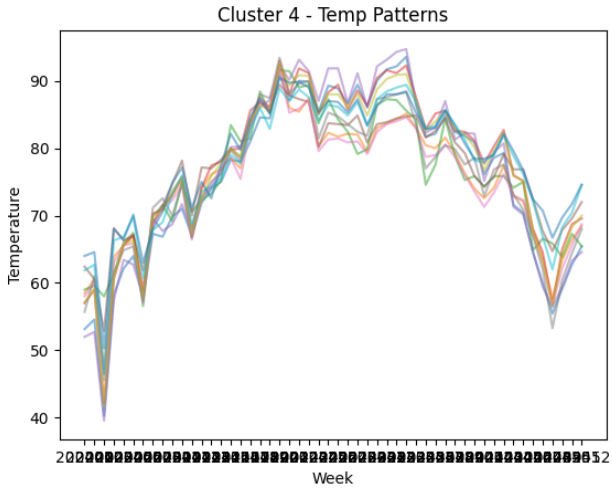

Figura 1: Serie temporal promedio de temperatura para el Clúster 4.

Se observan los patrones semanales característicos de este grupo de estaciones, con variaciones estacionales marcadas y una tendencia térmica distintiva respecto a otros clústeres.

De esta forma, el proceso permitió obtener una **caracterización climática agrupada** de las estaciones meteorológicas, que servirá como insumo para análisis posteriores, tales como la incorporación de variables climáticas en los modelos de predicción de ventas, obteniendo así la participación de 23 estaciones de clima.

7.4 Ingeniería de datos

Para el desarrollo de los modelos con inclusión de variables climáticas, primero se aplicaron las transformaciones pertinentes al conjunto de datos para obtener un registro semanal del clima, tanto real como pronosticado, de tal forma que corresponda con la granularidad de las ventas. Para los datos reales, se tomaron los datos climáticos, como temperatura promedio diaria y precipitación, para cada día de cada semana. De esta manera, se pasa a un formato horizontal, donde cada registro es referente a cada semana disponible y las columnas son factores climáticos para los días lunes a sábado. En cambio, para los datos pronosticados, para apegarse a un ejercicio real de uso del modelo, se tomaron los datos cuyo pronóstico fue realizado el día domingo de la semana anterior. Por tanto, al igual que los datos reales, se tiene un registro para cada semana y estación de clima, en este caso pronósticos para días lunes a sábado.

No obstante, los datos disponibles se encuentran de forma diaria para cada factor, lo cual no es relevante cuando se consideran datos de venta semanales. Por esta razón, se opta por generar variables que representen información que caracterice cada semana. En este resumen, se consideran estadísticas descriptivas de media, desviación estándar, máximo y mínimo; igualmente, se capturaron conteos de días que identifican la naturaleza de la semana, como días con temperaturas registradas por encima de 30°C, días de lluvia; y dinámicas encontradas como días con incrementos y decrementos de la temperatura, máximo salto de temperatura, volatilidad, etc.

7.5 Modelos implementados

Una vez realizado el procesamiento de los datos, se llevaron a cabo modelos de distinta complejidad para realizar predicciones de ventas semanales por cliente, con la disposición de los datos a lo largo del año.

Modelo base Los datos de predicciones generados por la empresa fueron complementados con una capa de ajuste con el objetivo de corregir el sesgo sistemático en función de una variable climática óptima identificada. El objetivo de este modelo fue evaluar si un factor multiplicativo por rango climático, segmentado por el producto, cluster y demanda, podría reducir el Error Absoluto Medio (MAE).

Metodología del Modelo Base La metodología empleada fue un **Out-of-Time (OOT) Segmentado** por rangos, buscando la estabilidad del ajuste.

- **Segmentación Jerárquica:** La unidad de análisis se definió por la combinación de producto, clúster climático y tamaño del cliente (Pequeno, Mediano, Grande, Muy Grande). El factor se calculó de manera independiente para cada segmento.
- **Validación OOT (80/20):** Los datos se dividieron por la dimensión temporal (`week`) en un 80 % para el entrenamiento del factor y un 20 % para la prueba.
- **Discretización Climática:** La variable climática óptima (Temp_{opt}) se discretizó en rangos fijos (*bins*) para calcular el factor de ajuste específico para cada condición.
- **Cálculo del Factor de Corrección:** El factor de ajuste (F) se calculó de la manera:

$$F = 1 + \text{Mediana}\left(\frac{\text{Venta Real} - \text{Predicción Base}}{\text{Predicción Base}}\right)$$

- **Aplicación Granular:** El factor resultante (F) se aplicó como un multiplicador directo a la predicción base del 20 % OOT.

Regresión lineal La regresión lineal multivariada se utilizó como primer modelo paramétrico para estimar la demanda semanal. Este modelo asume una relación lineal entre la variable objetivo (ventas semanales por cliente) y el conjunto de predictores utilizados, los cuales incluyen: (1) variables climáticas agregadas semanalmente, (2) retardos de ventas (*lags*) que capturan dependencia temporal de uno a tres periodos previos, y (3) la predicción base. Para evitar sesgos temporales, el entrenamiento se ejecutó por clúster de clientes utilizando validación temporal tipo *TimeSeriesSplit*, lo que garantiza que el modelo siempre predice valores futuros sin acceder a información del futuro durante el entrenamiento. Además, dentro de cada partición se aplicó un escalado estándar (*StandardScaler*) para asegurar estabilidad numérica en los coeficientes. Aunque la regresión lineal es un modelo simple, resulta útil como línea base analítica porque permite interpretar de manera directa la contribución marginal de cada predictor y comparar su desempeño contra métodos más complejos.

XGBoost El modelo XGBoost (*Extreme Gradient Boosting*) se empleó como alternativa no lineal y más expresiva para capturar relaciones complejas entre clima, ventas pasadas y patrones de interacción cliente–semana. XGBoost es un algoritmo basado en ensambles de árboles de decisión que optimiza gradualmente una función de pérdida. Su capacidad para modelar interacciones no lineales y efectos jerárquicos lo convierte en una herramienta adecuada para series temporales con múltiples clientes y alta variabilidad entre clústeres. En este trabajo, el modelo se entrenó por clúster, incorporando las mismas características que la regresión lineal: clima, lags y predicción base. También se utilizó *TimeSeriesSplit* para respetar la estructura temporal. XGBoost incluye regularización explícita y parámetros de control de complejidad (como profundidad máxima, tasa de aprendizaje y proporción de muestreo), lo cual reduce el riesgo de sobreajuste y mejora su capacidad de generalización. Este modelo permite capturar relaciones no lineales entre clima y demanda, efectos acu-

mulativos, y variaciones de comportamiento específicas por cliente o periodo del año, lo que lo hace potencialmente más robusto que la regresión lineal en escenarios dinámicos.

8 Recursos Utilizados

Para el desarrollo del análisis y los modelos predictivos se emplearon diversos recursos técnicos, computacionales y de conocimiento especializado. En cuanto a **librerías**, el procesamiento de datos y la ingeniería de características se realizaron con `pandas`, `numpy` y `pickle`, mientras que la estandarización y los modelos base se implementaron con `scikit-learn`. Para los métodos de clusterización, se utilizaron `scipy`, `tslearn` (para DTW) y `sklearn.cluster`. La visualización de resultados se apoyó en `matplotlib` y `seaborn`. El modelo XGBoost se implementó mediante la librería `xgboost`.

Respecto a los **datos disponibles**, se emplearon registros meteorológicos y datos proporcionados por la empresa, junto con datos de ventas históricas por cliente. Además, se generaron bases intermedias en formato `pickle` para optimizar la reproducibilidad del flujo de trabajo.

En cuanto a **expertos consultados**, el proceso de validación metodológica y del pipeline de modelado contó con apoyo y retroalimentación del equipo docente de la materia, en particular sobre técnicas de clusterización, manejo de series de tiempo y evaluación de modelos, también contamos con el apoyo de expertos de la empresa para adaptarnos a las necesidades y objetivos que buscamos solucionar.

Finalmente, se utilizaron **recursos computacionales** provistos por Google Colab y entornos locales, permitiendo entrenar modelos, realizar análisis exploratorios y generar visualizaciones de forma eficiente.

9 Resultados

Debido a que el modelo base fue entrenado con metodología OOT mientras que los modelos de aprendizaje máquina usaron *TimeSeriesSplit* sus resultados se presentan en tablas diferentes, siendo su mejora directamente comparable sobre diferentes bases de error. A pesar de haber encontrado una mayor cantidad de clústers, solamente se muestran los resultados de aquellos con la mayor cantidad de estaciones climáticas y con un desempeño más representativo.

Antes de proceder con el análisis, se recuerda que la columna 'Error Original' en los cuadros de resultados representa el Error Absoluto Medio (MAE) de la predicción base actual de la empresa, la cual no incorpora variables climáticas.

En el Cuadro 1 se compara el error original contra el error del modelo base.

Cuadro 1: MAE en Validación OOT Segmentada (modelo base)

Error Original	**Error Modelo**
18.3	19.7

En el Cuadro 2 se presenta un resumen general del error absoluto promedio obtenido para los otros modelos, así como la mejora relativa comparada contra el modelo original.

Cuadro 2: MAE en Validación TimeSeriesSplit (Lineal/XGBoost)

Modelo	**Error original**	**Error Modelo**
Regresión Lineal	7.62	8.05
XGBoost	7.62	8.51

Los resultados muestran que, en promedio, ninguno de los tres modelos propuestos supera al modelo original. Tanto el modelo base, como la regresión lineal y el XGBoost presentan un error absoluto superior con un incremento promedio aproximado de 1.4, 0.43 y 0.89 puntos en el error absoluto.

A pesar de que estos resultados globales sugieren que los modelos no generan una mejora promedio, esta conclusión debe interpretarse con precaución. Primero, los valores agregados esconden variabilidad significativa entre clústeres, clientes y temporadas, es decir, existen clientes o semanas donde los modelos sí superan al modelo

original. También, el modelo original ya incorpora información operativa histórica que no necesariamente está contenida explícitamente en las variables climáticas. Por último, la dificultad del problema aumenta debido al tamaño reducido de cada serie temporal, lo cual limita la capacidad de aprendizaje de modelos multivariados.

En consecuencia, aunque el desempeño promedio no supera a la línea original, estos resultados motivan un análisis más granular por clúster, cliente y condiciones climáticas, para identificar periodos o segmentos donde los modelos sí aportan valor y comprender los factores que afectan diferencialmente el error.

Los Cuadros 3 y 4 presentan los resultados agregados para los dos productos analizados (CRFR y CRFNR). Se reporta el error absoluto del modelo original, el error generado por cada modelo y la mejora relativa respecto al modelo original.

Cuadro 3: Resultados comparativos por producto del modelo base.

Producto	**Error Original**	**Error Modelo**	**Mejora**
CRFNR	7.24	7.48	-3.30
CRFR	7.48	7.55	-0.93

Cuadro 4: Resultados comparativos por producto modelo Lineal vs XGBoost.

Producto	Modelo	Error original	Error modelo	Mejora
CRFNR	Lineal	7.54	8.00	-0.47
CRFNR	XGBoost	7.53	8.46	-0.93
CRFR	Lineal	7.69	8.09	-0.40
CRFR	XGBoost	7.70	8.56	-0.86

Confirmando la tendencia inicial del modelo base, ni la regresión lineal ni el XGBoost superan al modelo original. En todos los casos, el error absoluto del modelo entrenado es mayor que el del modelo original, con disminuciones relativas cercanas al 0.4–0.9 unidades.

Los resultados desagregados por clúster permiten evaluar si el desempeño de los modelos varía según el tipo de cliente agrupado por patrones de consumo y condiciones asociadas. Los Cuadros 5 y 6 presentan el error absoluto promedio de la predicción original y de cada modelo, junto con la métrica de mejora relativa frente al modelo original.

Cuadro 5: Desempeño del modelo base por clúster.

Cluster	**Error Original**	**Error Modelo**
0	16.80	18.09
4	21.96	23.86
5	15.32	16.63

Cuadro 6: Desempeño de los otros modelos por clúster

Cluster	**Modelo**	**Error original**	**Error modelo**
0	Lineal	6.98	7.51
0	XGBoost	6.98	7.90
4	Lineal	9.18	9.47
4	XGBoost	9.18	10.04
5	Lineal	6.47	6.85
5	XGBoost	6.47	7.35

Los resultados muestran que ninguno de los modelos supera de forma sistemática al modelo original en los distintos clústeres. La regresión lineal tiende a presentar un menor deterioro en comparación con XGBoost, especialmente en el clúster 5, donde la estructura de la señal parece ser más estable y cercana a una tendencia lineal. Por otro lado, en clústeres como el 4, caracterizados por mayor variabilidad en el consumo, los tres modelos presentan un desempeño inferior respecto a la linea original.

La Figura 2 presenta la evolución del error absoluto medio por semana para ambos modelos —Regresión Lineal y XGBoost— considerando únicamente el error del modelo (no el baseline)

Ambos modelos presentan un incremento abrupto del error alrededor de la semana 25, lo cual sugiere la presencia de un comportamiento atípico en la demanda (posiblemente asociado a efectos climáticos extremos, cambios estacionales o eventos comerciales específicos). Después de este pico, los errores regresan a niveles similares a los observados semanas antes, indicando que dicha anomalía no implica un cambio estructural permanente.

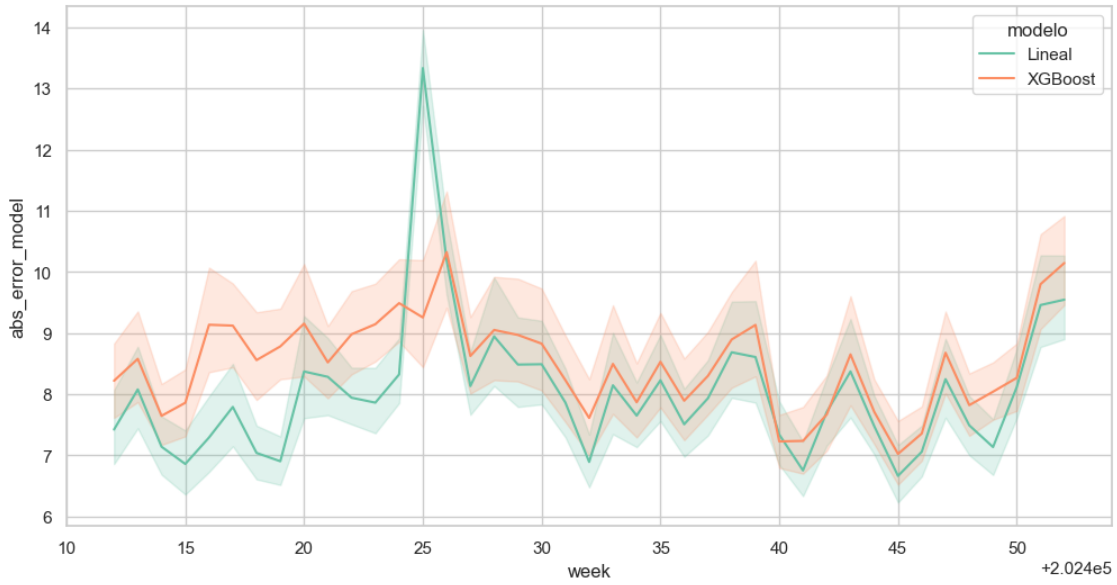

Figura 2: Error de modelo por semana

Hacia el final del periodo (semanas 48–52), ambos modelos muestran un incremento moderado del error, consistente con la variabilidad típica de la demanda en periodos de cierre de año. A pesar de ello, la regresión lineal sigue mostrando un mejor ajuste relativo frente a XGBoost.

La Figura 3 muestra el mapa de calor de la mejora del modelo respecto al original (definida como la reducción del error absoluto medio). Los valores positivos representan semanas en las que el modelo supera a la predicción base, mientras que los valores negativos indican un desempeño inferior.

En general, el mapa de calor permite identificar con claridad los periodos y clústeres donde el modelo presenta mayor dificultad, así como aquellos donde la aproximación es más estable. Esta visualización destaca la importancia de explorar variables adicionales, reconsiderar la estructura de lags o incluso segmentar el entrenamiento de forma distinta para obtener mejoras significativas en periodos críticos.

10 Conclusiones

El objetivo principal del proyecto fue fortalecer la exactitud de los pronosticos de ventas de refrescos en canales de venta tradicionales mediante la incorporación de variables climáticas. de que la integración de variables climáticas mejoraría la precisión de las ventas de refrescos al influir el clima de forma distinta según el producto y la región, no pudo ser validada de forma sistemática con los modelos implementados.

En la evaluación general, y utilizando el Error Absoluto Medio (MAE) como métrica, se encontró que ninguno de los enfoques desarrollados logró superar consistentemente el desempeño de la predicción original de la empresa. Los resultados desagregados por producto y por clúster climático confirmaron esta tendencia de deterioro en la precisión a nivel global, sugiriendo que el modelo original ya incorpora información operativa e histórica clave que no fue complementada del todo por las variables climáticas utilizadas.

No obstante, esta conclusión general debe interpretarse con el matiz de que sí se identificaron clientes específicos donde los modelos lograron superar la precisión de la predicción original, demostrando que la inclusión de variables climáticas sí tiene un impacto positivo localizado.

El análisis detallado de los errores y la visualización del desempeño semanal permitieron identificar periodos y segmentos críticos donde el desempeño es inferior, así como aquellos donde la predicción mejoró. Este conocimiento, sobre

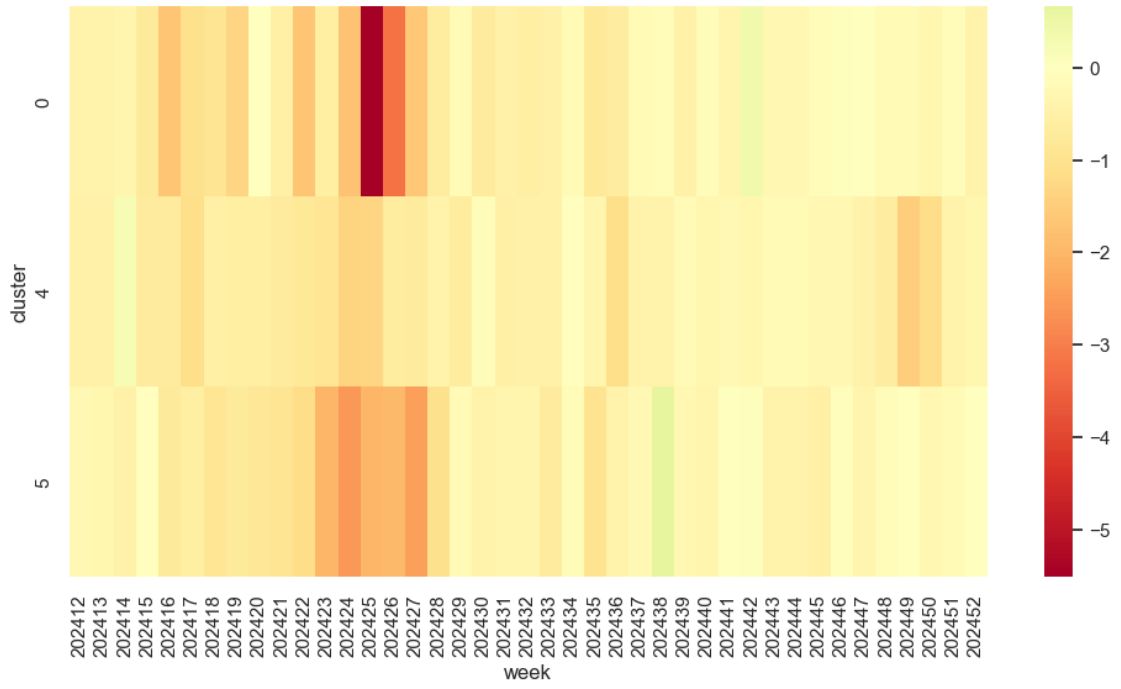

Figura 3: Comparación de los modelos vs. predicciones originales

cuándo y dónde el modelo interno se equivoca o acierta, aporta un valor significativo para la toma de decisiones, la gestión de inventarios y, crucialmente, para el diseño de futuros sistemas de predicción más transparentes, interpretables y robustos. El proyecto motiva un análisis más granular de los datos para implementar los modelos solo en aquellos segmentos donde se ha comprobado su mejora.

Referencias

Ho, C., Ko, K., Liu, S., y Wu, C. (2024). *Stormy sales: the influence of weather expectations on fmcg consumption.* Descargado de `https://www.emerald.com/jpbm/article-abstract/33/7/801/1218352/Stormy-sales-the-influence-of-weather-expectations?redirectedFrom=fulltext`

Islek, I., y Oguducu, S. G. (2015). *A retail demand forecasting model based on data mining techniques.* Descargado de `https://ieeexplore.ieee.org/abstract/document/7281443`

Mitra, A., Jain, A., Kishore, A., y Kumar, P. (2022). *A comparative study of demand forecasting models for a multi-channel retail company: A novel hybrid machine learning approach.* Descargado de `https://doi.org/10.1007/s43069-022-00166-4`

Tian, J., Zhang, Y., y Zhang, C. (2018). Predicting consumer variety-seeking through weather data analytics. *Electronic Commerce Research and Applications*, 194-207.

Verma, R., Nagar, V., y Mahapatra, S. (2021). Introduction to supervised learning. *Data Analytics in Bioinformatics: A Machine Learning Perspective*, 1-34.

Vu, D. H., Muttaqi, K. M., y Agalgaonkar, A. P. (2015). *A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables.* Descargado de `https://doi.org/10.1016/j.apenergy.2014.12.011`

Yilmaz, O. (2024). *An investigation of weather impact on beverage sales forecasting.* Descargado de `http://arno.uvt.nl/show.cgi?fid=169249`