Hackathon de Inteligencia de Datos: Predicción de Demanda en la Industria

Marissa E. Luna[1], Ximena A. Cantón[1] and Nubia G. Barajas[1]

[1] *Instituto Tecnológico y de Estudios Superiores de Monterrey, Escuela de Ingenieria y Ciencias, Guadalajara, Jalisco*

Abstract—Se presenta un análisis de ventas históricas de Interlub, incluyendo segmentación por cliente, artículo y unidad de venta. Se entrenaron modelos de predicción con XGBoost para kilogramos, litros y piezas, mejorando su desempeño tras eliminar outliers. Finalmente, se proponen variables externas e internas adicionales para futuras mejoras en la precisión del modelo.

Keywords—análisis exploratorio de datos, predicción de ventas, XGBoost, outliers, series temporales, unidades de venta, métricas de desempeño

I. Introducción

La predicción de la demanda es clave para optimizar la producción, logística y compras en la industria. En este trabajo se analiza el historial de ventas de Interlub, empresa especializada en lubricantes industriales, con el objetivo de anticipar la cantidad futura de ventas por artículo.

El análisis incluyó un estudio exploratorio para detectar patrones y anomalías, seguido del desarrollo de modelos predictivos con *XGBoost*, empleando codificación de variables, eliminación de outliers mediante IQR y transformaciones logarítmicas para mejorar la estabilidad del modelo.

II. Metodología

El análisis exploratorio de datos (EDA) se llevó a cabo con el objetivo de comprender el comportamiento histórico de las órdenes de venta registradas por la empresa, identificando patrones generales, irregularidades y posibles errores en los datos. Las etapas principales de la metodología fueron las siguientes:

Limpieza y estructuración del dataset

- El archivo original contenía todos los campos unidos en una sola columna. Se realizó un preprocesamiento para dividirlo correctamente en: *Orden de Venta*, *Creación Orden de Venta*, *Código Cliente*, *Artículo*, *Cantidad* y *Unidad de venta*.
- Se detectaron y eliminaron registros duplicados y aquellos con cantidad igual a cero.
- Adicionalmente, se corrigió un error sistemático en los datos: cuando la columna *Cantidad* presentaba más de tres ceros consecutivos al final (e.g., 270000), estos ceros eran eliminados para reflejar la cantidad real (270).

Conversión de tipos de datos

- Se transformaron los campos *Cantidad* a tipo numérico y *Creación Orden de Venta* a tipo fecha (`datetime64`).
- Para facilitar el análisis mensual, se generó una nueva columna *AñoMes* en formato de texto (`YYYY-MM`).

División por unidades de venta

- Se identificó que los productos se vendían en tres unidades diferentes: **KG**, **L** y **PZA**.
- Para evitar mezclas que pudieran sesgar los análisis, el dataset se segmentó por unidad, y cada uno se analizó de manera independiente.

Análisis exploratorio estructurado

Se abordaron las siguientes dimensiones:

- Análisis de clientes
- Análisis de artículos
- Análisis de cantidad de pedidos
- Análisis temporal mensual
- Detección de valores atípicos (*outliers*) mediante *boxplots* y estadística robusta (IQR)

III. Análisis Exploratorio de Datos

Antes de iniciar con el análisis exploratorio, se realizó un proceso de revisión y limpieza de los datos. El conjunto original contenía 31,156 registros distribuidos en 6 columnas clave: número de orden, fecha de creación, cliente, artículo, unidad de venta y cantidad.

No se encontraron valores nulos en ninguna de las columnas, lo que indica que el dataset estaba completo. Sin embargo, se identificaron **4,957 registros duplicados**, los

cuales fueron eliminados, reduciendo el total a **26,199 entradas únicas**. También se detectaron **20 registros con cantidad igual a cero**, que fueron descartados por no representar ventas reales.

Estos datos se anexaron a dataframe de errores, y se encontró que todos los artículos y clientes de estas ventas de 0, ya habían realizado o sido parte de pedidos válidos, descratando la posibilidad de existencia de productos erróneos o clientes inactivos.

Posteriormente, se transformó la columna `Creación Orden de Venta` al tipo de dato `datetime` para permitir un análisis temporal adecuado. Finalmente, los datos fueron segmentados en tres subconjuntos según la unidad de venta: **kilogramos (KG)**, **litros (L)** y **piezas (PZA)**. Esto permitió aplicar estadísticas descriptivas y visualizaciones más precisas para cada tipo de producto.

A continuación, se presentan las estadísticas básicas de la cantidad vendida en cada unidad:

- **Kilogramos (KG)** – `12,826 observaciones`
  - Media: 596.4 Desviación estándar: 1,515
  - Mínimo: 1 Mediana: 135 Máximo: 37,800
- **Litros (L)** – `7,702 observaciones`
  - Media: 366.1 Desviación estándar: 746.1
  - Mínimo: 1 Mediana: 142 Máximo: 17,280
- **Piezas (PZA)** – `5,651 observaciones`
  - Media: 25.9 Desviación estándar: 77.2
  - Mínimo: 1 Mediana: 9 Máximo: 4,500

Como se observa, los productos vendidos en kilogramos y litros presentan mayores volúmenes de venta y una mayor dispersión. En cambio, los productos vendidos por pieza tienden a comercializarse en cantidades bajas, con una fuerte concentración cerca del mínimo y una mediana significativamente inferior al promedio, lo cual indica una distribución altamente sesgada a la derecha.

Este preprocesamiento fue esencial para garantizar la calidad del análisis posterior y obtener conclusiones confiables sobre el comportamiento de ventas, clientes y productos.

a. Análisis de clientes

Se identificaron **907 clientes únicos** en el dataset.

La mayoría de los clientes realizan pocos pedidos: la mediana es de **4 pedidos**, pero el promedio es **29**, lo que indica la presencia de clientes con comportamiento altamente frecuente.

El cliente con mayor número de pedidos registró **1,913 órdenes**, mientras que varios clientes tienen solo una.

Al observar las cantidades vendidas por cliente, se identificaron diferencias relevantes dependiendo de la unidad de medida. Por ejemplo, en productos vendidos por kilogramo, los histogramas revelaron una alta concentración de ventas en rangos bajos de cantidad, con algunos valores extremos que influyen en la media.

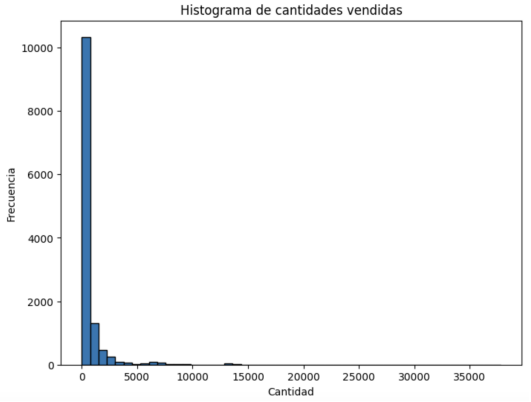

Fig. 1: Histograma de cantidades vendidas en kilogramos. Se observa una fuerte concentración de ventas en valores bajos, pero con una larga cola hacia la derecha que indica la presencia de outliers.

Al aplicar escala logarítmica a los histogramas, se visualizó con mayor claridad la tendencia general de los datos.

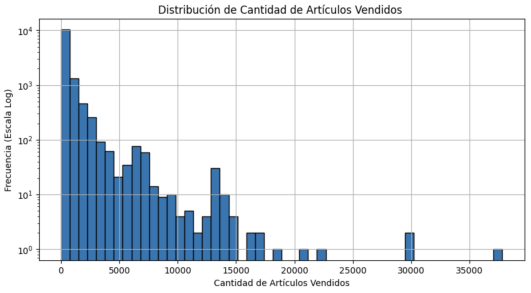

Fig. 2: Histograma logarítmico de cantidades vendidas en kilogramos. La transformación permite apreciar la concentración real de los datos en rangos bajos de venta.

Un análisis similar con productos vendidos en litros mostró un patrón parecido, lo cual refuerza la tendencia de que la mayoría de los clientes adquiere cantidades pequeñas por transacción.

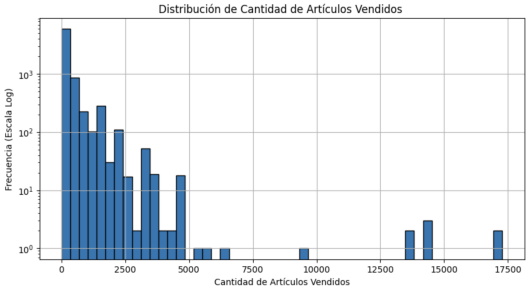

Fig. 3: Histograma de cantidades vendidas en litros. La distribución presenta un comportamiento similar al de los kilogramos: alta concentración en valores bajos y presencia de valores extremos.

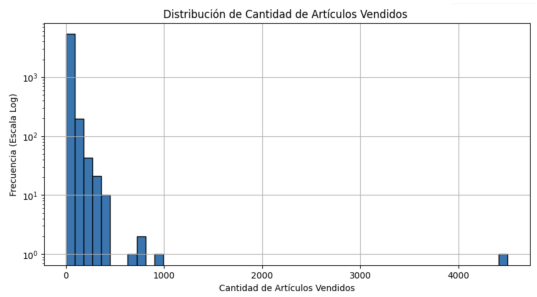

Fig. 4: Histograma de cantidades vendidas en piezas. Se observa una alta frecuencia en valores pequeños, con algunos casos atípicos en el extremo derecho de la distribución.

La distribución de las cantidades vendidas en piezas refuerza la tendencia observada en otros tipos de unidades de medida. La mayoría de las compras se realizan en pequeñas cantidades, lo que sugiere que los clientes tienden a adquirir solo lo necesario en cada transacción. Sin embargo, la presencia de valores extremos indica que existen ciertas compras atípicas con volúmenes significativamente mayores. Esto podría deberse a compras al mayoreo o a clientes con necesidades específicas que requieren una cantidad superior a la media.

Distribución de Pedidos por Cliente

La siguiente gráfica muestra la distribución del número de pedidos realizados por cliente:

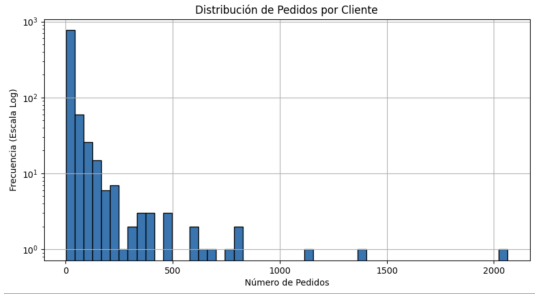

Fig. 5: Histograma de pedidos por cliente.

Observaciones clave:

- Distribución sesgada a la derecha: La gran mayoría de los clientes realizó un número reducido de pedidos. A medida que aumenta el número de pedidos, la cantidad de clientes disminuye rápidamente.
- Clientes ocasionales vs. clientes frecuentes: Hay una base amplia de clientes que realizaron pocos pedidos, lo que sugiere un comportamiento de compra esporádico o de prueba. En contraste, existe un grupo muy pequeño de clientes que realizaron entre 500 y más de 2000 pedidos, lo que indica una alta frecuencia de compra y un valor estratégico para el negocio.
- Outliers valiosos: Se observan clientes atípicos con más de 1000 pedidos. Aunque representan menos del 1% del total, su contribución a los ingresos puede ser desproporcionadamente alta.

Es recomendable segmentar la base de clientes según su frecuencia de pedidos para diseñar estrategias diferenciadas:

- Fidelización y retención para los clientes frecuentes.
- Campañas de reactivación o promoción para los clientes ocasionales.

Top 10 Clientes con Más Pedidos

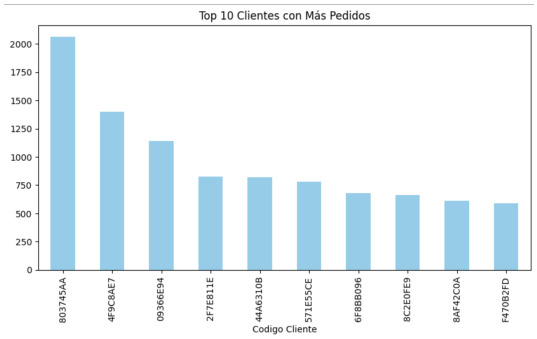

Fig. 6: Los 10 clientes con mayor cantidad de pedidos registrados. Cada barra representa un cliente identificado por su código, ordenado de mayor a menor número de pedidos.

Observaciones clave:

- Alta concentración de pedidos en pocos clientes: El cliente con código `803745AA` destaca claramente con más de 2000 pedidos, seguido por otros clientes con volúmenes que van desde aproximadamente 600 hasta 1400 pedidos.
- Disminución progresiva: Existe una disminución escalonada en la cantidad de pedidos conforme se desciende en el ranking, lo que refuerza la importancia de los primeros lugares.
- Clientes clave para el negocio: Estos 10 clientes probablemente representan una parte significativa del volumen total de pedidos. Su comportamiento de compra sostenido sugiere relaciones comerciales consolidadas o alto nivel de fidelización.

Estos clientes son estratégicos para el negocio. Se recomienda darles seguimiento cercano, ya sea mediante programas de lealtad, atención personalizada o condiciones comerciales preferenciales. También conviene analizar si estos clientes pertenecen a un mismo segmento, región o tipo de industria, para identificar oportunidades de crecimiento en perfiles similares.

b. Análisis de artículos

En total, se identificaron **889 artículos únicos** dentro del conjunto de datos. Para facilitar el análisis, los artículos se agruparon por unidad de venta: kilogramos (kg), litros (L) y piezas (PZA). Esto permitió comparar no solo la frecuencia de venta de cada artículo, sino también la cantidad vendida en función de su unidad de medida.

Estadísticas descriptivas por unidad

A continuación se muestran los principales estadísticos de la cantidad vendida por artículo según su unidad:

- **Kilogramos (KG):** media de 17,075, mediana de 1,080 y máximo de 897,102. Existe un fuerte sesgo a la derecha.
- **Litros (L):** media de 8,623, mediana de 1,670 y máximo de 207,724. La distribución sigue un patrón similar a KG.
- **Piezas (PZA):** media de 808, mediana de 18 y máximo de 33,548. Alta concentración en valores bajos.

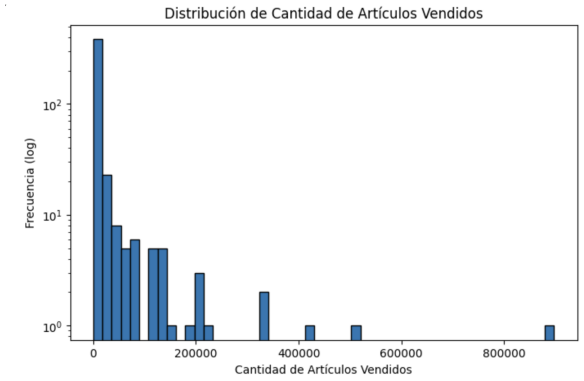

Fig. 7: Distribución de la cantidad vendida por artículo (KG). Se observa un fuerte sesgo a la derecha: la mayoría de los artículos se vende en cantidades pequeñas, pero algunos superan con creces las 100,000 unidades.

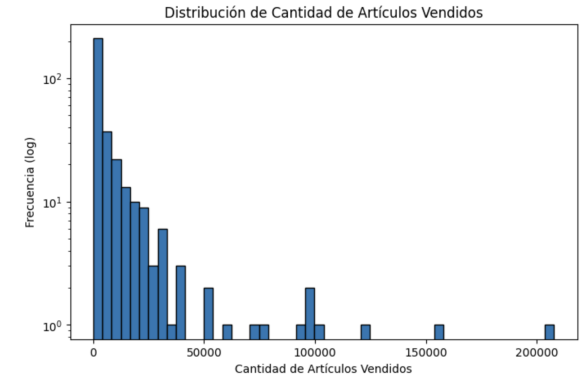

Fig. 8: Distribución de la cantidad vendida por artículo (L). La concentración en valores bajos también es evidente, aunque con menor dispersión que en KG.

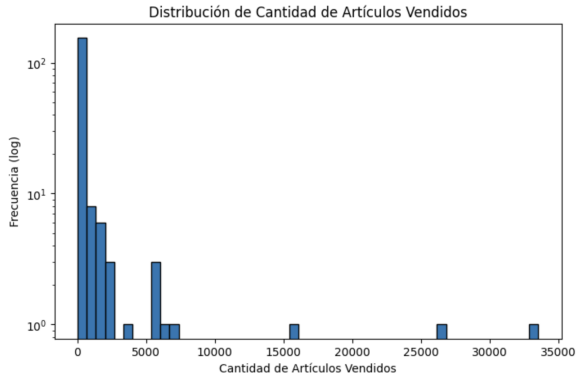

Fig. 9: Distribución de la cantidad vendida por artículo (PZA). Se muestra una altísima concentración en ventas pequeñas, con una larga cola de artículos que superan las 10,000 unidades.

Artículos más vendidos por cantidad

A continuación se presentan los **10 artículos más vendidos** según su cantidad total, desglosados por unidad de medida.

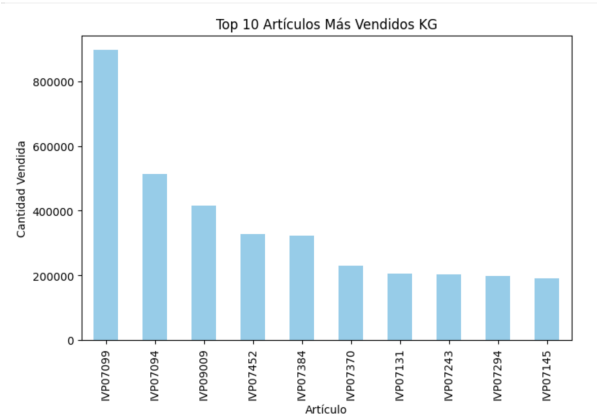

Fig. 10: Top 10 artículos más vendidos en KG. El artículo `IVP07099` domina claramente con más de 800,000 kg vendidos.

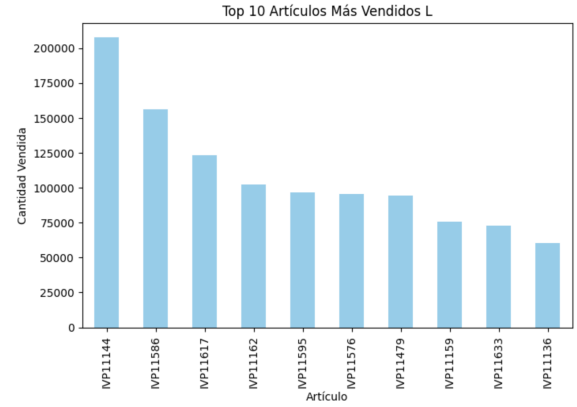

Fig. 11: Top 10 artículos más vendidos en L. El artículo `IVP1144` encabeza la lista con más de 200,000 litros vendidos.

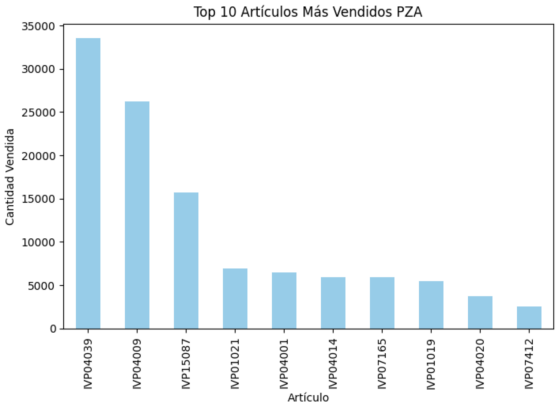

Fig. 12: Top 10 artículos más vendidos en PZA. El artículo `IVP04039` supera las 30,000 piezas vendidas, muy por encima de los demás.

Artículos más vendidos por número de pedidos

Además del volumen total, se identificaron los artículos que aparecen con mayor frecuencia en pedidos. Esto permite distinguir productos de alto consumo recurrente, incluso si se venden en volúmenes pequeños.

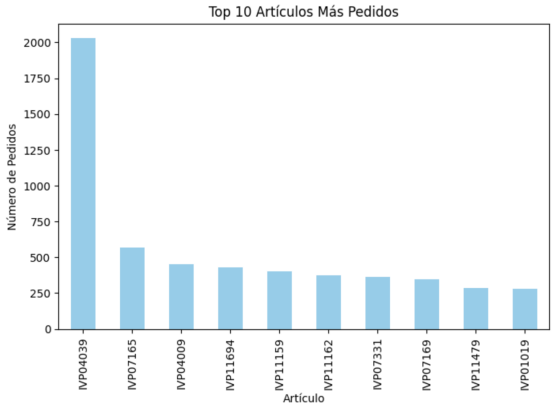

Fig. 13: Top 10 artículos con más pedidos. `IVP04039` lidera con más de 2,000 pedidos registrados, lo que lo posiciona como un artículo de alta frecuencia de compra.

Conclusiones del análisis de artículos

- Las distribuciones de cantidad vendida por artículo están fuertemente sesgadas a la derecha, especialmente en piezas, lo que indica que pocos productos concentran el mayor volumen.
- Algunos artículos son líderes tanto en volumen como en frecuencia de pedidos, pero también se identifican productos que, aunque no se vendan en grandes cantidades, tienen alta rotación.
- Estas diferencias son importantes para la gestión de inventario, promociones y predicción de demanda futura.

c. Análisis temporal

Se realizó un análisis de la evolución de las ventas a lo largo del tiempo, utilizando la variable `Fecha` y agregando la cantidad total vendida por día. Este enfoque permite detectar patrones de comportamiento, posibles estacionalidades o eventos atípicos que afecten la demanda.

Tendencia por unidad de medida

Las ventas se desagregaron por unidad de medida: kilogramos, litros y piezas. Esto permitió observar diferencias en los patrones de comportamiento temporal de cada tipo de producto.

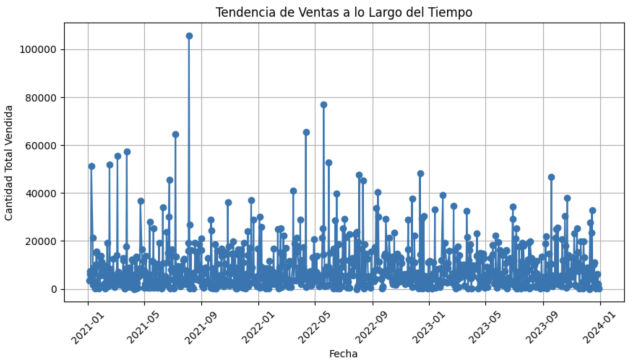

Fig. 14: Tendencia de ventas para productos vendidos en kilogramos. Se observan múltiples picos de venta, especialmente en los primeros trimestres de 2021 y 2022, lo que podría sugerir patrones de estacionalidad o campañas específicas.

La serie de kilogramos presenta una fuerte variabilidad diaria, con valores que oscilan entre 0 y más de 100,000 unidades vendidas en un solo día. Esta dispersión sugiere que las ventas no se distribuyen de manera uniforme en el tiempo.

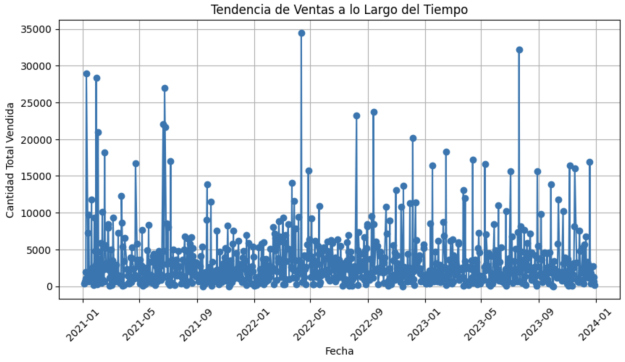

Fig. 15: Tendencia de ventas para productos vendidos en litros. Se aprecia una distribución más estable con algunos picos que podrían estar relacionados con campañas o clientes específicos.

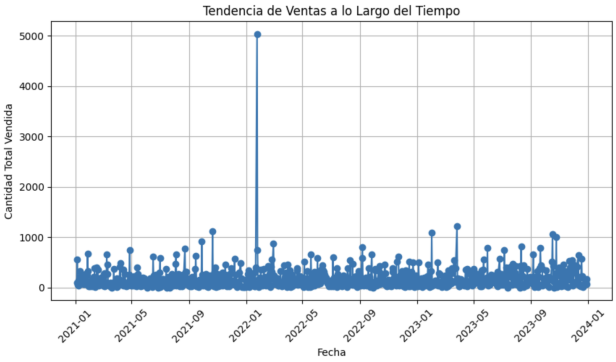

Fig. 16: Tendencia de ventas para productos vendidos en piezas. La actividad es generalmente baja, con algunos incrementos puntuales importantes.

Conclusiones del análisis temporal

- Las ventas en kilogramos muestran una alta variabilidad, con ciertos momentos de mayor intensidad que podrían estar relacionados con estacionalidad, promociones o necesidades específicas del cliente.
- El comportamiento varía significativamente según la unidad de medida. Las ventas en litros son más estables, mientras que en piezas los volúmenes son bajos y más esporádicos.
- Este tipo de análisis es fundamental como base para modelos de predicción de demanda o para la planificación logística y comercial.

IV. Modelos Predictivos

a. XGBoost

Corrección y ajuste del modelo

Se desarrolló un modelo predictivo utilizando el algoritmo **XGBoost**, con el objetivo de estimar la cantidad futura de ventas por artículo. Para mejorar su desempeño, se aplicó

una corrección basada en la eliminación de **valores atípicos** utilizando el rango intercuartílico (IQR), con lo cual se logró una reducción significativa en la dispersión de los datos.

Cada modelo se entrenó por separado para las unidades de venta: kilogramos (KG), litros (L) y piezas (PZA). Se codificaron las variables categóricas mediante `LabelEncoder`, y se utilizó una partición del 70% para entrenamiento y 30% para prueba.

Evaluación del modelo

La siguiente tabla muestra las métricas obtenidas en los conjuntos de entrenamiento y prueba:

TABLE 1: MÉTRICAS DE DESEMPEÑO DEL MODELO XGBOOST POR UNIDAD DE VENTA

Unidad	Conjunto	MAE	RMSE	R^2
KG	Entrenamiento	64.28	129.86	0.797
	Prueba	106.85	194.34	0.528
L	Entrenamiento	38.76	72.56	0.850
	Prueba	89.24	143.09	0.405
PZA	Entrenamiento	2.18	4.25	0.911
	Prueba	5.87	9.83	0.506

En todos los casos se observa un desempeño superior en el conjunto de entrenamiento respecto al de prueba, lo que indica cierto grado de **sobreajuste**. A pesar de ello, los resultados son razonables considerando la variabilidad inherente a las series de ventas y la simplicidad del modelo.

Predicción de ventas futuras

Con los modelos entrenados, se realizaron predicciones para los próximos 30 días para los artículos más vendidos en cada unidad. Los resultados se muestran en las siguientes gráficas:

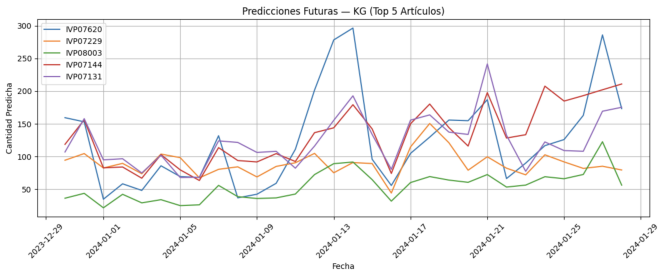

Fig. 17: Predicción de ventas futuras para los artículos más vendidos en KG

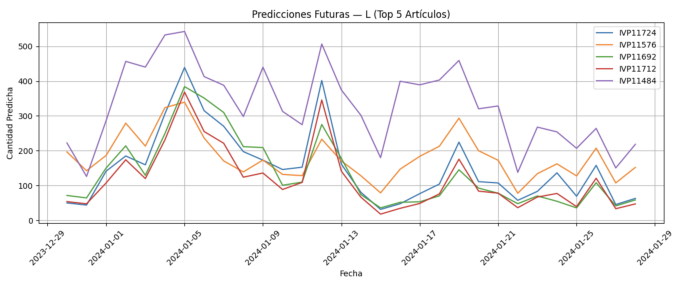

Fig. 18: Predicción de ventas futuras para los artículos más vendidos en L

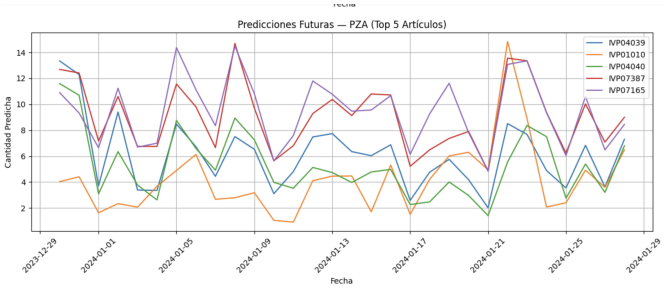

Fig. 19: Predicción de ventas futuras para los artículos más vendidos en PZA

Propuesta de variables adicionales para mejorar el modelo

Para incrementar la precisión del modelo, se propone incorporar variables adicionales que capturen mejor el contexto en el que ocurren las ventas. Entre ellas se incluyen:

- **Tendencias históricas:** Promedios móviles, ventas del mismo periodo del año anterior o indicadores de estacionalidad.
- **Factores externos:**
  - Presencia de *promociones* u ofertas especiales
  - *Días festivos* o fechas clave en el sector
  - *Disponibilidad de proveedores* o cambios en materia prima
- **Condiciones económicas:** Inflación, tipo de cambio o costos logísticos que afectan el comportamiento de compra.
- **Datos internos adicionales:**
  - Nivel de *inventario disponible* en cada fecha
  - *Plazos de crédito* o condiciones comerciales por cliente

Estas variables permitirían enriquecer el conjunto de datos, reducir el error del modelo y aumentar su capacidad para anticipar escenarios futuros de forma más robusta.

Validación cruzada (K-Fold)

Con el fin de evaluar la estabilidad del modelo y reducir el riesgo de una división afortunada o desafortunada de los datos, se aplicó validación cruzada con 5 *folds*. Este proceso no tiene como objetivo entrenar un nuevo modelo, sino comprobar si la configuración actual de hiperparámetros es confiable y generaliza bien.

Al realizar la validación cruzada se obtuvieron los siguientes resultados:

TABLE 2: RESULTADOS DE VALIDACIÓN CRUZADA (5 FOLDS) POR UNIDAD DE VENTA

Unidad	RMSE Promedio	RMSE por Fold
KG	1.00	[0.98, 0.97, 1.01, 1.00, 1.01]
L	1.02	[0.99, 1.00, 1.05, 1.04, 1.01]
PZA	0.73	[0.72, 0.72, 0.72, 0.76, 0.71]

La validación cruzada permitió verificar que:

- No existía un sobreajuste grave.
- Los resultados fueron consistentes entre diferentes combinaciones de datos de entrenamiento y prueba.
- No fue necesario ajustar los hiperparámetros actuales, ya que se obtuvo un rendimiento estable y satisfactorio.

Este enfoque proporcionó evidencia objetiva para mantener la configuración del modelo sin cambios adicionales, lo que incrementa la confianza en su capacidad de generalización.

Conclusión: Los hiperparámetros fueron evaluados usando validación cruzada con 5 folds. Se observó un RMSE promedio estable, sin varianza significativa entre los folds, por lo que se decidió mantener dicha configuración para el modelo final.

V. Conclusiones

El presente análisis permitió desarrollar modelos de predicción por unidad de venta (KG, L y PZA) con resultados satisfactorios y estables. A través del análisis exploratorio de datos (EDA) se identificaron errores sistemáticos en las cantidades registradas, lo cual subraya la importancia del preprocesamiento y limpieza como paso previo al modelado.

Tras aplicar la eliminación de valores atípicos, codificación de variables categóricas y una transformación logarítmica de la variable objetivo, se entrenaron modelos con el algoritmo *XGBoost*, alcanzando buenos niveles de precisión. La validación cruzada (K-Fold) mostró un bajo nivel de varianza entre los distintos folds, confirmando que los modelos generalizan de manera adecuada sin incurrir en sobreajuste.

El modelo entrenado por unidad permitió realizar predicciones de demanda para los siguientes 30 días, destacando artículos con alto volumen de ventas y permitiendo identificar patrones de consumo relevantes. Estas predicciones tienen el potencial de apoyar decisiones estratégicas relacionadas con inventario, producción y atención al cliente.

En resumen, el modelo desarrollado representa un primer paso sólido hacia una solución escalable de predicción de demanda en Interlub. Futuras mejoras pueden incluir la incorporación de variables externas (estacionales, macroeconómicas o de campañas de marketing), así como el análisis más profundo del comportamiento por cliente o por familia de producto.

A. Anexo

Documentación de librerías utilizadas en la elaboración del código: [1] [2] [3] [4]

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Duchesnay, *Scikit-learn: Machine Learning in Python*, 2011. [Online]. Available: https://scikit-learn.org/

[2] T. Chen and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, 2016. [Online]. Available: https://xgboost.readthedocs.io/

[3] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, *Array programming with NumPy*, 2020. [Online]. Available: https://numpy.org/

[4] J. D. Hunter, *Matplotlib: A 2D Graphics Environment*, 2007. [Online]. Available: https://matplotlib.org/