UNIVERSIDAD TECNOLÓGICA DE PANAMÁ FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES MAESTRIA EN ANALITICA DE DATOS

Machine Learning y Analítica Prescriptiva "Predicción y Alerta de Variación de Precios en la Canasta Básica (Ciudad de Panamá)"

> Presentado por: Cibyl Ángeles Aquino Rodríguez Jesús González Castillo

> > Profesor Juan Montenegro

Proyecto Final

Introducción

Este proyecto aborda la predicción y la alerta temprana de variaciones de precios en la Canasta Básica de Alimentos de la Ciudad de Panamá, un problema con impacto directo en el poder adquisitivo de los hogares y en la toma de decisiones de política pública. Se trabajó con la base de datos publicada por ACODECO en el portal de datos abiertos, integrando los archivos mensuales de 2021 y 2022 que reportan precios por producto, presentación, establecimiento y fecha. A partir de esa materia prima se construyó un flujo reproducible que automatiza la descarga e integración, estandariza nombres y formatos, transforma las tablas de la estructura original a un formato largo adecuado para series de tiempo y aplica criterios de calidad para entregar un conjunto coherente y consistente.

Datos y Fuente

Los datos provienen de la Autoridad de Protección al Consumidor y Defensa de la Competencia (ACODECO), a través del portal de datos abiertos de Panamá. Se recopilaron los archivos mensuales de precios de la Canasta Básica de Alimentos (CBA) para los años 2021 y 2022 (con disponibilidad parcial en 2020). Los archivos estaban en formato CSV/XLS, uno por mes, con información a nivel de producto, presentación, supermercado y fecha.

Descarga e integración de datos

- Se diseñó un script automatizado en Python para descargar y organizar los archivos en la estructura de carpetas del proyecto.
- Se unificaron 14 archivos mensuales, resultando en un dataset consolidado.

Análisis Exploratorio (EDA)

El análisis exploratorio comenzó por ordenar el conjunto de datos: se unificaron nombres de variables a un esquema sencillo con producto, presentación, supermercado, mes y precio; la fecha se llevó a un marcador mensual homogéneo y el precio se normalizó a formato numérico. Con esa base se revisó la calidad, comprobando tipos, presencia de nulos y duplicados en la combinación producto, presentación, supermercado y mes, además de listar precios imposibles o claramente fuera de rango. Para controlar valores extremos se aplicó un criterio robusto basado en el rango intercuartílico calculado dentro de cada grupo producto, presentación y supermercado, con un plan de respaldo cuando los grupos eran pequeños que recurrió a límites por producto y, en última instancia, a límites globales. El resultado es un conjunto coherente y trazable, acompañado por tablas de control y un archivo limpio listo para modelar.

Las gráficas ayudan a entender el comportamiento del mercado sin perderse en el detalle. La tendencia mensual del precio muestra una trayectoria suavemente ascendente a lo largo del periodo, mientras que la versión con mediana y banda intercuartílica permite distinguir el nivel típico de cada mes y cuánta dispersión hubo alrededor. Al comparar supermercados mediante boxplots se observan diferencias claras tanto en nivel como en variabilidad: hay tiendas sistemáticamente más caras o baratas y otras con

distribuciones más extendidas y presencia de extremos. El diagrama de dispersión que resume la mediana por producto revela una heterogeneidad marcada entre artículos, con muchos concentrados en rangos bajos y un subconjunto con niveles notablemente más altos. El coeficiente de variación complementa esta lectura al señalar qué productos son relativamente más inestables y por tanto conviene monitorear o tratar con técnicas robustas. También se estimó el porcentaje de atípicos por supermercado como señal de consistencia de captura y se construyó un mapa de calor de medianas producto por supermercado que pone en evidencia patrones de nivel sostenidos en ciertas tiendas y productos.

Todo lo anterior orienta decisiones prácticas para el modelado. La evidencia de una dinámica temporal sugiere validar por tiempo y considerar variables de calendario, las diferencias persistentes entre tiendas recomiendan incluir el efecto supermercado o normalizar por punto de venta, y la heterogeneidad entre productos invita a usar resúmenes robustos. Con el conjunto limpio, las tablas de validación y las figuras, el proyecto avanza con una base sólida y con una narrativa de datos clara que respalda las siguientes etapas.

Procesamiento de Datos (Limpieza)

En esta etapa se realizó la preparación y limpieza de los datos, asegurando que el conjunto fuera consistente, completo y apto para el análisis y modelado. Se aplicaron procesos de normalización, filtrado y estructuración de la información, con el objetivo de garantizar la calidad de los registros y facilitar la posterior generación de indicadores y características para los modelos predictivos.

Normalización y limpieza inicial

- Se eliminaron columnas vacías y se estandarizaron los nombres de variables (snake_case, sin caracteres especiales).
- Se aplicó un proceso wide → long (melt), reestructurando la tabla para que cada fila representara una observación la estructura original de los archivos (donde cada columna correspondía a un supermercado/establecimiento) a un formato largo. En este nuevo esquema, cada fila representa una observación única con las variables (producto, medida, establecimiento, precio, anio_mes). Esta transformación es fundamental ya que:
 - Simplifica el manejo de series temporales y el cálculo de indicadores por productoestablecimiento.
 - Nos permite aplicar funciones de agrupación, imputación y creación de lags de forma eficiente.
 - Estandariza el dataset para que cada registro corresponda a un evento, el precio de un producto específico en un establecimiento y mes determinado.
- Se normalizaron strings de productos, medidas y establecimientos: minúsculas, sin espacios extra ni símbolos no deseados.
- El campo precio se convirtió a numérico (float), corrigiendo formatos con comas y descartando valores no válidos.

Filtrado, consistencia e imputación de valores faltantes

Se eliminaron registros con precios nulos o menores/iguales a 0.

- Se identificaron valores atípicos.
 - o Rango intercuartílico (IQR): se consideraron atípicos los precios fuera del intervalo Q1 $1.5 \times IQR$, Q3 + $1.5 \times IQR$.
 - Desviación estándar: de manera complementaria, se marcaron como sospechosos los precios que superaban ±3σ respecto a la media del producto en ese establecimiento.
- Se generó la malla completa de combinaciones (producto, establecimiento, mes) para el rango de fechas disponible (2021–2022). Esto permitió garantizar que cada serie tuviera registros en todos los meses, incluso si originalmente había huecos en los archivos de ACODECO.
 - La malla se construyó con un producto cartesiano entre la lista de productos, establecimientos y el conjunto de periodos (anio_mes).
 - Al unirla con los datos originales, se obtuvieron observaciones faltantes que luego se completaron mediante las siguientes imputaciones.
 - Interpolación lineal: Se utiliza dentro de cada grupo de establecimiento × producto. Rellena los valores faltantes en los periodos intermedios basándose en los precios anteriores y posteriores, creando una transición suave entre los datos conocidos.
 - Propagación hacia adelante y hacia atrás: (ffill() y bfill()). Se aplican en los extremos de la serie, donde la interpolación lineal no puede estimar valores.
 - Este enfoque asegura que los modelos puedan calcular lags (1, 3, 6 meses), medias móviles y volatilidades sin romper la continuidad temporal.

Diccionario de Datos

Tras la etapa de procesamiento y depuración de datos, se obtuvo un conjunto limpio y estructurado con las siguientes variables:

COLUMNA	TIPO	DESCRIPCION		
establecimiento	TEXTO	Nombre del supermercado donde se obtuvo a observación		
producto	TEXTO	Producto observado		
medida	TEXTO	Si el producto tiene algún tipo de medida (kg, l, lb)		
source_file	TEXTO	Indica de donde viene la información (si es imputado o el archivo de origen)		
year	NUMBER	Año de la observación		
month	NUMBER	Mes de la observación		
precio	FLOAT	Precio de producto		
periodo	NUMBER	Identificador del periodo en formato YYYYMM, obtenido de year y month.		

Ingeniería de Características

Para capturar la dinámica temporal y la variabilidad de los precios de la canasta básica, se implementaron las siguientes transformaciones:

 Trabajo por serie temporal: Cada combinación (producto, establecimiento) se trató como una serie independiente, lo que permitió calcular métricas de comportamiento histórico de manera específica por producto y supermercado.

- Lags de precios: Se generaron variables lag_1, lag_3 y lag_6, que corresponden al precio del producto 1, 3 y 6 meses atrás, respectivamente. Estas variables reflejan la dependencia temporal de los precios y permiten capturar patrones recientes y estacionales.
- Medias móviles y volatilidad: Se calcularon medias móviles (MA_3, MA_6) y volatilidades (VOL_3, VOL_6) sobre ventanas de 3 y 6 meses. Estas métricas describen la tendencia y la variabilidad de los precios, facilitando la predicción de movimientos futuros.
- Objetivo de predicción y alerta: Se definió como objetivo el precio del próximo mes (y_next) y la variación porcentual correspondiente (pct_change_next). Además, se generó una variable binaria de alerta (alert_next) para identificar incrementos del 5% o más en el precio mensual.
- Comparación con mediana mensual: Se calculó la mediana de precios por producto considerando todos los establecimientos, y se creó la variable ratio_precio_mediana para indicar si un precio estaba por encima o por debajo de la tendencia central del mes.
- Variables temporales: Se agregaron year y month como variables adicionales para capturar efectos de estacionalidad.
- Limpieza final: Se eliminaron las filas sin objetivo (y_next nulo), correspondientes al último mes de cada serie, garantizando que todos los registros tengan información completa para entrenamiento y validación.

Modelado

Para la predicción de precios y la alerta de incrementos significativos, se construyeron modelos de regresión y clasificación, utilizando las variables temporales y estadísticas generadas en la ingeniería de características.

División temporal de los datos

Para asegurar la integridad temporal y evitar fugas de información futura, los datos se dividieron en conjuntos de entrenamiento, validación y prueba siguiendo el orden cronológico de los meses.

- El conjunto Train contiene datos hasta octubre de 2021, representando aproximadamente el 60% de la serie temporal.
- El conjunto Valid incluye los meses desde noviembre de 2021 hasta enero de 2022, utilizado para ajustar hiperparámetros y evaluar desempeño durante el desarrollo del modelo.
- El conjunto Test abarca los datos a partir de febrero de 2022, destinado a la evaluación final del modelo sobre información no vista.

Esta división estricta permite simular un escenario de predicción, donde el modelo solo dispone de información histórica para anticipar precios futuros y generar alertas.

Modelo de Referencia (Baselines)

Antes de entrenar modelos complejos, se evaluaron baselines simples basados en valores pasados del precio, como referencia inicial de desempeño:

1. **Baseline LV (lag 1)**: Este modelo naive predice que el precio del próximo mes será igual al precio del mes actual.

2. **Baselines estacionales (lag 3 y 6)**: Estos modelos consideran los precios de 3 y 6 meses atrás, capturando posibles patrones estacionales.

Los resultados de evaluación se presentan a continuación:

Conjunto	Baseline	Lag	RMSE	MAE
TRAIN	LV	1	0.2172	0.0548
TRAIN	Estacional	3	0.3294	0.1107
TRAIN	Estacional	6	0.3896	0.1379
VALID	LV	1	0.2522	0.0639
VALID	Estacional	3	0.3661	0.1191
VALID	Estacional	6	0.3909	0.1379
TEST	LV	1	0.3562	0.0943
TEST	Estacional	3	0.4383	0.1481
TEST	Estacional	6	0.4741	0.1647

Observaciones:

- El baseline lag 1 es el más competitivo, especialmente en errores absolutos (MAE), mostrando que el precio mensual tiene fuerte persistencia.
- Los lags más largos (3 y 6 meses) reflejan tendencias estacionales, pero incrementan el error al no capturar cambios recientes.
- Estos resultados sirven como referencia mínima: cualquier modelo de machine learning deberá superar estas métricas para considerarse útil.

Preparación de Features y Pipelines

En la construcción de modelos de predicción, se definieron features numéricas y categóricas:

- Numéricas: precio actual, lags 1, 3 y 6, medias móviles (MA_3, MA_6), volatilidades (VOL_3, VOL 6), ratios respecto a la mediana mensual, mediana por producto, año y mes.
- Categóricas: producto, establecimiento.

Se implementó un pipeline de preprocesamiento usando scikit-learn:

- Para features numéricas: se imputaron valores faltantes con la mediana y se escalaron con StandardScaler.
- Para features categóricas: se imputaron valores faltantes con la categoría más frecuente y se codificaron con OneHotEncoder ignorando categorías desconocidas.

Se generaron conjuntos finales para regresión (y_next) y clasificación (alert_next) de la siguiente manera:

Conjunto	Regresión (y_next)	Clasificación (alert_next)
Train	X_train, y_train	X_train, y_train_cls
Valid	X_valid, y_valid	X_valid, y_valid_cls
Test	X_test, y_test	X_test, y_test_cls

- La variable de clasificación alert_next indica un aumento del precio ≥ 5% respecto al mes actual.
- Este pipeline asegura que todos los modelos posteriores reciban datos consistentes, manejando correctamente valores faltantes y escalamiento de features.

Modelos de Regresión

Se entrenaron y evaluaron modelos de regresión para predecir el precio del próximo mes (y_next), usando los pipelines de features definidos previamente.

Modelos considerados:

Modelo	Baseline
Ridge	Regresión lineal con regularización L2 (alpha=1.0)
UCPPog	HistGradientBoostingRegressor con 300 iteraciones, un modelo basado en boosting
HGBReg	para capturar relaciones no lineales

Métricas de desempeño:

Modelo	Conjunto	RMSE	MAE	R ²
Ridge	VALID	0.2393	0.0834	0.988
Ridge	TEST	0.3345	0.1126	0.978
HGBReg	VALID	0.2308	0.0782	0.989
HGBReg	TEST	0.3331	0.1116	0.978

Comentarios:

- El modelo HGBReg presenta mejor desempeño que Ridge en validación, reflejando que los datos contienen patrones no lineales y efectos estacionales que Ridge no puede capturar.
- En el conjunto de prueba, ambos modelos presentan una ligera caída en desempeño, especialmente Ridge, lo que sugiere cierta sensibilidad a datos fuera de la ventana de entrenamiento.
- HGBReg logra un R² de 0.978 en test, indicando que el modelo explica cerca del 97.8% de la variabilidad del precio, con errores promedio razonables.

Para predicción de precios, los modelos basados en boosting como HGBReg son más adecuados que regresiones lineales simples, especialmente cuando se manejan series con tendencias y estacionalidades complejas.

Modelos de Clasificación (Alertas de Precios)

Se entrenaron modelos para predecir la **alerta de aumento de precio (alert_next)**, utilizando las mismas features y pipelines que para regresión.

Modelos considerados:

Modelo	Baseline
LogReg	Regresión logística con balanceo de clases
HGBCls	HistGradientBoostingClassifier con 300 iteraciones, capaz de capturar relaciones
11020.5	complejas no lineales

Métricas de desempeño:

Modelo	Conjunto	Accuracy	F1	ROC-AUC
LogReg	VALID	0.764	0.380	0.852

LogReg	TEST	0.740	0.374	0.806	
HGBCls	VALID	0.907	0.282	0.855	
HGBCls	TEST	0.887	0.239	0.821	

Interpretación:

- El modelo HGBCls supera a LogReg en Accuracy y ROC-AUC, mostrando que los efectos no lineales y las combinaciones de features temporales y categóricas son relevantes para predecir aumentos significativos de precio.
- El F1-score relativamente bajo de ambos modelos refleja el desbalance de clases; aunque la Accuracy es alta, la capacidad de detectar correctamente los aumentos (positivos) es limitada.
- El ROC-AUC de aproximadamente 0.855 en validación para HGBCls indica buena discriminación entre casos con y sin alerta.
- La regresión logística mantiene un desempeño aceptable, pero se ve superada por los modelos basados en boosting, que capturan mejor la complejidad del problema.

Para la predicción de alertas de precios, los modelos de boosting son más adecuados, en escenarios de series temporales con múltiples productos y establecimientos y presencia de desbalance en las clases.

Optimización de Hiperparametros

Para mejorar el desempeño de los modelos de regresión, se realizó una búsqueda de hiperparámetros usando GridSearchCV con TimeSeriesSplit. Esto respeta la secuencia temporal de los datos, evitando que información futura "filtre" en entrenamiento.

Ridge Regression

Se implementó un pipeline que combina preprocesamiento numérico y categórico con Ridge Regression. El objetivo fue encontrar el mejor valor de alpha que minimice el error cuadrático medio (RMSE) en validación temporal.

• Se probaron distintos valores de alpha = [0.1, 1.0, 10.0, 100.0].

Validación temporal: 3 folds

Métrica: RMSE

• Mejor resultado en CV: alpha = 10.0

Desempeño final:

Conjunto	RMSE	MAE
VALID	0.2399	0.0836
TEST	0.3342	0.1125

El modelo Ridge ofrece un baseline robusto y estable, reduciendo la sensibilidad a valores atípicos y proporcionando predicciones razonablemente precisas en periodos futuros. Aunque menos flexible que modelos complejos, es útil para comparar mejoras posteriores.

HistGradientBoostingRegressor (HGB)

Se entrenó un HistGradientBoostingRegressor, que permite capturar relaciones no lineales y patrones complejos en series temporales de precios.

• Grid de búsqueda:

o max_depth = [3, 5, None]

o min_samples_leaf = [20, 50]

o learning_rate = [0.01, 0.1]

• Validación temporal: 3 folds

Métrica: RMSE

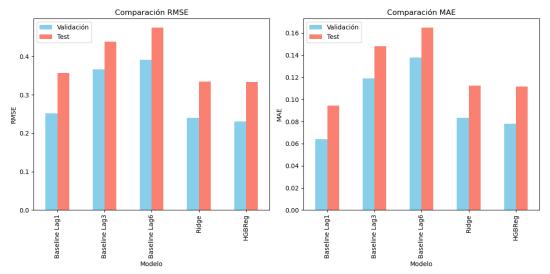
• Mejores parámetros encontrados: learning_rate = 0.1, max_depth = 5, min_samples_leaf = 20

Desempeño final:

Conjunto	RMSE	MAE
VALID	0.2339	0.0792
TEST	0.3345	0.1123

HGB logra un mejor ajuste, capturando efectos no lineales y estacionales. Se observa menor error en validación y una buena generalización en test, aunque la predicción fuera de muestra sigue siendo más desafiante (RMSE test > RMSE valid). La combinación de features temporales y categóricas, junto con un modelo no lineal como HGB, mejora significativamente la predicción de precios futuros frente a modelos lineales simples. La validación temporal garantiza una evaluación realista de desempeño en producción.

Baseline vs Modelado

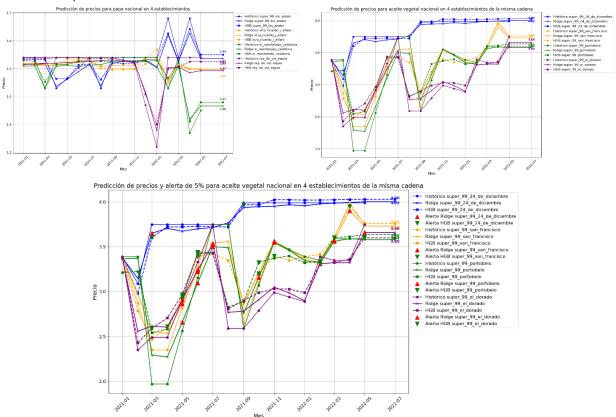


La comparación de métricas evidencia que los modelos basados en *Gradient Boosting* (HGBReg) y Ridge superan de forma consistente a los baselines construidos con rezagos simples (lag 1, 3 y 6). Mientras que los baselines muestran errores crecientes a medida que el rezago se amplía, HGBReg logra mantener los menores valores de RMSE y MAE tanto en validación como en prueba, confirmando su capacidad para capturar relaciones no lineales y patrones de estacionalidad. Ridge también presenta un buen desempeño, aunque con ligeras limitaciones frente a variaciones más complejas. Estos resultados validan la ventaja de modelos más sofisticados para anticipar precios en la canasta básica frente a enfoques ingenuos basados únicamente en históricos directos.

Reproducibilidad (Modelos y Features)

Para asegurar la reproducibilidad del sistema, se guardaron los modelos entrenados, los *pipelines* de preprocesamiento y los conjuntos de datos procesados. Esto garantiza consistencia en las transformaciones, facilita la reutilización en nuevas predicciones y permite replicar el flujo completo sin rehacer el procesamiento previo.

Predicciones y Alertas Utilizando los Modelos Entrenados



Los resultados obtenidos muestran que los modelos desarrollados logran anticipar con solidez la evolución de los precios en productos representativos de la canasta básica. En el caso de la papa nacional y el aceite vegetal, tanto Ridge como HistGradientBoosting siguen de manera cercana la tendencia observada en los establecimientos, capturando los cambios de nivel y las fluctuaciones en distintos periodos. Además, para el aceite vegetal se evidencia la utilidad del esquema de alerta temprana: los modelos de clasificación identifican los meses en los que el incremento supera el umbral del 5%, aportando señales accionables que complementan la predicción de precios. Este doble enfoque —pronóstico de valores futuros y detección de variaciones significativas— refuerza la aplicabilidad práctica del sistema para el monitoreo continuo de la canasta básica en Panamá.

Evaluación y Resultados

En esta etapa se cargaron los modelos entrenados (Ridge y HistGradientBoosting para regresión, e HistGradientBoostingClassifier para clasificación), junto con sus preprocesadores y listas de variables,

además de recuperar los conjuntos de datos de entrenamiento, validación y prueba con sus respectivos objetivos para ambas tareas.

Evaluación de Métricas

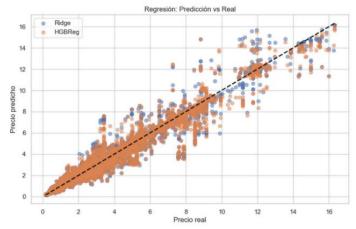
Se evaluó el desempeño de los modelos utilizando métricas estandarizadas para cada tarea. En regresión se analizaron RMSE, MAE y R², mientras que para clasificación se consideraron Accuracy, F1-macro y ROC-AUC, permitiendo comparar Ridge y HGBReg, y evaluar HGBCls para alertas de precio.

Tarea	Modelo	RMSE ↓	MAE ↓	R² ↑	Accuracy ↑	F1- macro 个	ROC- AUC ↑	Interpretación
Regresión	Ridge	0.334	0.113	0.978	-	-	-	Desempeño robusto como baseline lineal.
	HGBReg	0.335	0.112	0.978	-	-	-	Capta patrones no lineales; desempeño ligeramente superior.
Clasificación	HGBCls	-	-	-	0.887	0.589	0.821	Buen balance entre sensibilidad y especificidad; útil para alertas.

Los resultados muestran que los modelos basados en Gradient Boosting son más flexibles y capaces de capturar relaciones complejas en los datos, tanto en regresión como en clasificación. En regresión, aunque Ridge ofrece un baseline estable, HGBReg logra un ajuste más preciso frente a patrones no lineales y estacionales. Para la detección de alertas de precio, HGBCls proporciona una alta discriminación, demostrando su utilidad en escenarios desbalanceados y dinámicos.

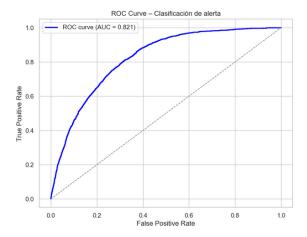
Análisis visual de la regresión

La gráfica de dispersión entre los precios reales y los predichos permite comparar el desempeño de los dos modelos evaluados. En ambos casos, los puntos se concentran cerca de la diagonal de referencia, lo que indica una buena capacidad de ajuste. Sin embargo, se observa que HGBReg (naranja) presenta una mayor alineación con la línea ideal, sobre todo en los rangos intermedios y altos de precios, reduciendo la dispersión y los errores sistemáticos. Por el contrario, el modelo Ridge (azul) tiende a subestimar o sobreestimar con

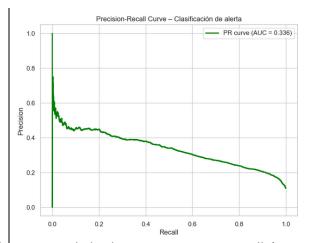


mayor frecuencia en valores extremos, lo que refleja la limitación de los enfoques lineales para capturar patrones más complejos. Este grafica respalda las métricas cuantitativas previamente reportadas, confirmando que los modelos basados en *Gradient Boosting* logran un mejor ajuste y mayor capacidad de generalización en el problema planteado.

Análisis visual de la clasificación



La curva ROC muestra un desempeño sólido del modelo de clasificación basado en *Histogram Gradient Boosting*, con un área bajo la curva (AUC = 0.821). Esto indica una alta capacidad de discriminación entre las clases, incluso en escenarios con diferentes umbrales de decisión, manteniendo una buena relación entre verdaderos positivos y falsos positivos.



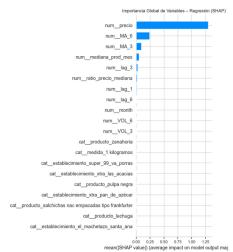
Por otro lado, la curva Precision–Recall (AUC = 0.336) revela una mayor complejidad en el balance entre precisión y cobertura. Aunque el modelo logra identificar correctamente una gran proporción de los casos positivos (recall), la precisión disminuye conforme se amplía la cobertura, lo cual es común en problemas con desbalance de clases.

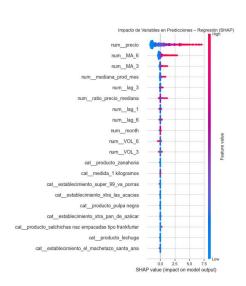
En conjunto, ambas curvas confirman que el modelo es adecuado para la detección de alertas, pero también resaltan la importancia de ajustar el umbral de decisión según el objetivo específico: maximizar la detección de casos relevantes (recall) o minimizar las falsas alarmas (precisión).

Análisis SHAP

SHAP es una técnica que utiliza los valores de Shapley, un concepto de la teoría de juegos, para atribuir a cada característica del modelo su contribución a la predicción final. En otras palabras, SHAP responde a la pregunta: ¿qué tan importante es cada variable para el resultado que predice el modelo?

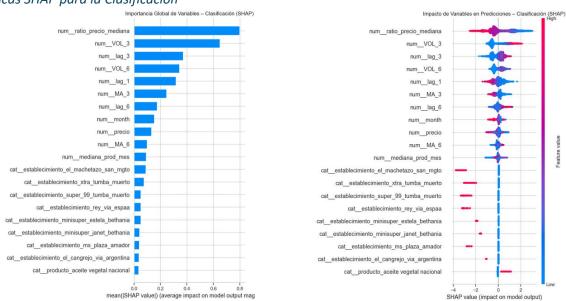
Graficas SHAP para la regresión





Los resultados del análisis SHAP evidencian que las variables numéricas relacionadas con el comportamiento histórico del precio son las más influyentes en el modelo de regresión. En particular, el precio previo (num_precio) y las medias móviles de 6 y 3 meses (num_MA_6, num_MA_3) explican la mayor parte de la variabilidad en las predicciones, seguidas por indicadores de tendencia como la mediana mensual del producto y los rezagos de 3 meses. Por otro lado, las variables categóricas asociadas a productos y establecimientos presentan un impacto mucho menor en comparación con las series temporales de precios, lo que confirma que la dinámica histórica del precio es el principal determinante de las predicciones.





Los resutados de SHAP para clasificacion obtenidos permiten comprender la importancia y el efecto de las variables en el modelo de clasificación. En la figura de importancia global se observa que las variables más influyentes son principalmente indicadores relacionados con precios y volúmenes, destacando num_ratio_precio_mediana, num_vol_3 y num_log_3. Asimismo, el gráfico de dispersión muestra cómo los valores altos o bajos de estas variables impactan en la probabilidad de clasificación, lo que aporta interpretabilidad al modelo y refuerza la confianza en sus resultados.

Análisis de error por segmento

Analisis de Error por Productos:

Categoría	Producto	RMSE Normalizado	R²	Observaciones
Mejor predicción	Lechuga	0.154	0.961	Error relativo más bajo, precios estables
	Carne molida de primera	0.157	0.970	Modelo captura muy bien la variabilidad
	Codito	0.165	0.973	Alta precisión relativa y buen ajuste
Peor predicción	Queso blanco prensado bajo en sal	0.203	0.955	Mayor error relativo, precios más volátiles
	Papa nacional	0.204	0.953	Dificultad para capturar fluctuaciones
	Lentejas	0.196	0.942	Aunque R ² alto, error relativo más elevado

El análisis de error por producto muestra que el modelo logra un ajuste muy alto en todos los casos (R² entre 0.94 y 0.97), con errores absolutos relativamente bajos. Los productos más estables, como lechuga, carne molida de primera y codito, presentan los menores errores relativos (RMSE normalizado entre 0.15 y 0.17), lo que indica que el modelo predice sus precios con gran precisión. En contraste, productos más

volátiles como papa nacional, queso blanco prensado bajo en sal y lentejas registran los mayores errores relativos (RMSE normalizado alrededor de 0.20). Esto sugiere que, aunque el modelo explica bien la variabilidad, la capacidad predictiva disminuye en productos sujetos a fluctuaciones de mercado o estacionalidad más marcadas.

Análisis de Error por Establecimiento:

Categoría	Establecimiento	RMSE Normalizado	R²	Observaciones
Mejor predicción	super_99_camino_real	0.102	0.983	Más bajo error relativo, excelente ajuste
	super_99_san_francisco	0.124	0.972	Alta precisión, precios promedio más altos (~4.4)
	super_99_ricardo_j_alfaro	0.144	0.977	Muy buen ajuste, error relativo bajo
Peor predicción	xtra_las_acacias	0.274	0.960	Mayor error relativo, precios más bajos amplifican el error
	super_99_chanis	0.257	0.947	Predicción débil relativa, precios bajos (~1.46) lo explican
	xtra24_diciembre	0.170	0.957	Error relativo medio, aunque R ² aceptable

El análisis de error por establecimiento confirma que el modelo también se desempeña de forma sólida en todos los supermercados evaluados, con R² entre 0.94 y 0.98. Se observa que los locales con precios promedio más altos, como super_99_camino_real, super_99_san_francisco y super_99_ricardo_j_alfaro, obtienen los mejores resultados (RMSE normalizado entre 0.10 y 0.14), lo que indica predicciones muy consistentes. Por el contrario, establecimientos con precios promedio más bajos, como xtra_las_acacias y super_99_chanis, muestran errores relativos más elevados (0.25–0.27), aunque manteniendo un buen nivel de ajuste global. Esto revela que el modelo es más confiable en segmentos de precios altos, mientras que en los de precios bajos pequeñas desviaciones impactan más en la métrica relativa.

Conclusiones

El proyecto permitió construir un flujo reproducible de integración, limpieza y modelado de precios de la canasta básica, garantizando consistencia en las series temporales y trazabilidad de los datos. El análisis exploratorio confirmó una tendencia alcista moderada, heterogeneidad entre productos y diferencias persistentes entre establecimientos, lo que justifica la incorporación de variables temporales y categóricas en los modelos. Los resultados evidencian que los enfoques lineales (Ridge) ofrecen un baseline sólido, pero los modelos basados en *Gradient Boosting* capturan de manera más efectiva las relaciones no lineales y efectos de interacción, alcanzando en regresión un R² cercano al 0.98 y en clasificación un ROC-AUC superior a 0.82. El bajo F1-score observado en clasificación refleja el desbalance natural de los eventos de alza, lo que plantea la necesidad de estrategias adicionales de manejo de clases.

Más allá de las métricas, el valor práctico está en la ruta completa: un pipeline que parte de datos abiertos, garantiza calidad, entrega indicadores y visualizaciones interpretables y produce un dataset listo para modelado que puede utilizarse para monitoreo y para respaldar decisiones de protección al consumidor. Como líneas de mejora se recomienda ampliar el horizonte temporal, incorporar variables exógenas macroeconómicas y logísticas, afinar la imputación en series con alta estacionalidad y desarrollar tableros operativos con validación fuera de tiempo y control de deriva de datos. Con estas extensiones, el sistema puede evolucionar hacia una solución de alerta temprana robusta y mantenible, útil tanto para análisis estratégico como para acciones tácticas en el seguimiento de la canasta básica.

Limitaciones

- Cobertura temporal reducida: aunque los datos de ACODECO están disponibles desde 2020, el proyecto se enfocó en 2021–2022, priorizando la calidad de los registros y excluyendo el periodo de pandemia, que pudo distorsionar los precios. Esta decisión limita la detección de patrones de largo plazo y estacionalidades completas.
- Variables exógenas no incluidas: los modelos se basaron exclusivamente en precios históricos. No se incorporaron factores externos como inflación, costos logísticos, producción agrícola o tipo de cambio, que pueden influir en las variaciones de precios.
- Desbalance de clases en alertas: los incrementos de precio ≥5% representan una minoría en comparación con la estabilidad o reducciones, lo que afectó la sensibilidad de los modelos de clasificación (F1 bajo pese a ROC-AUC alto).
- Limitación en granularidad: aunque se incluyeron productos y supermercados, no se consideraron variaciones por distrito o región fuera de Ciudad de Panamá, lo que restringe la generalización del modelo a nivel nacional.
- Predicciones sensibles a series volátiles: en productos con alta variabilidad (ej. papa, queso blanco bajo en sal), el modelo mostró mayor error relativo, indicando la necesidad de estrategias específicas para series con alta estacionalidad o choques de oferta.

Recomendaciones

- Ampliación de horizonte temporal: incorporar datos históricos adicionales y extender el rango hacia años posteriores para robustecer la detección de patrones estacionales y estructurales.
- Incorporación de variables exógenas: integrar factores macroeconómicos (inflación, tipo de cambio) y logísticos (costos de transporte, producción agrícola) que puedan influir en las dinámicas de precios.
- Manejo del desbalance de clases: aplicar técnicas como *SMOTE*, *undersampling*, *class weights* o *threshold tuning* para mejorar la sensibilidad en la predicción de alertas.
- Validación fuera de tiempo: implementar esquemas de backtesting más extensos y pruebas de robustez ante distintos escenarios de predicción.
- Despliegue y monitoreo: desarrollar tableros interactivos de alerta temprana y establecer mecanismos de control de deriva de datos que aseguren estabilidad del modelo en producción.