



## MAT306 - PROYECTOS ESTADÍSTICOS

*2do. Semestre 2025 / Programa Diurno / Campus Casa Central Valparaíso*

---

# Informe de Formulación

---

***Autores:***

CRISOSTO LEÓN, Carlos Matías  
MELÍAN DÍAZ, Diego Ignacio  
MESÍAS CARRIZO, Camilo Javier  
PIZARRO OLIVARES, Mauricio Alejandro  
RIVERA RAASCH, Nolberto Patricio

***Profesor:***

Francisco Antonio  
ALFARO MEDINA

TABLA DE CONTENIDOS

<b>1. Contenido científico/tecnológico</b>	<b>3</b>
1.1. Problema y oportunidad . . . . .	3
1.2. Resultados previos . . . . .	3
1.3. Hipótesis . . . . .	4
1.4. Estado del arte . . . . .	4
1.5. Criterios de éxito . . . . .	5
<b>2. Estrategia de desarrollo e impacto</b>	<b>6</b>
2.1. Producto final . . . . .	6
2.2. Usuarios/beneficiarios . . . . .	6
2.3. Ventajas . . . . .	6
2.4. Impacto esperado . . . . .	6
<b>3. Capacidades y organización</b>	<b>7</b>
3.1. Equipo y roles . . . . .	7
3.2. Coordinación . . . . .	7
<b>4. Objetivos y métricas</b>	<b>8</b>
4.1. Objetivo General . . . . .	8
4.2. Objetivos Específicos . . . . .	8
4.3. Métricas de Evaluación . . . . .	8
4.3.1. Métrica Principal: RMSSE (Root Mean Squared Scaled Error) . . . . .	9
4.3.2. Métricas Secundarias . . . . .	9
4.4. Criterios de Éxito y Alineación . . . . .	10
<b>5. Metodología</b>	<b>11</b>
5.1. Estrategia de Validación . . . . .	11
5.2. Métricas de Evaluación . . . . .	11
5.2.1. RMSSE (Root Mean Squared Scaled Error) . . . . .	12
5.2.2. Métricas Auxiliares . . . . .	12
5.3. Interpretabilidad del Modelo . . . . .	12
5.4. Gestión de Sesgos y Data Leakage . . . . .	13
5.4.1. Prevención de Leakage Temporal . . . . .	13

5.4.2. Mitigación del Sesgo de Intermitencia . . . . .	13
<b>6. Ética, permisos y datos</b>	<b>14</b>
6.1. Datos . . . . .	14
6.2. Uso responsable . . . . .	14
<b>7. Resultados e hitos</b>	<b>15</b>
7.1. Resultados . . . . .	15
7.2. Hitos . . . . .	15

## 1. Contenido científico/tecnológico

### 1.1. Problema y oportunidad

En el presente proyecto se trabajó con el dataset *M5 Forecasting*, el cual consiste en un arreglo de series temporales que indican el número de ventas diarias de un producto determinado en una tienda determinada de la cadena de supermercados Walmart, durante 1913 días seguidos.

Una exploración preliminar del dataset revela que existen

1. 3 estados (de los EE.UU.): 'CA', 'TX', 'WI'.
2. 10 tiendas por estado, lo cual da  $3 \times 10 = 30$  tiendas en total, identificadas por un `store_id`.
3. 3.049 productos distintos disponibles, identificados por un `store_id`.

Las series temporales del dataset se crean sólo cuando ese producto realmente se vende en esa tienda, lo cual explica por que el total de series que se tiene es menor a  $3 \times 10 \times 3,049$ . En otras palabras, el conjunto de índices que identifica a cada serie es un subconjunto del producto cartesiano  $\{\text{store\_id}\} \times \{\text{item\_id}\}$ .

El problema que aqueja a la cadena Walmart es la incertidumbre respecto a las ventas que de cada producto tendrá cada día, por cada tienda. El pronóstico M5 permite ayudar a resolver dicha incertidumbre, dado que el dataset incluido hace las veces de insumo para un análisis predictivo de las ventas de la cadena (análisis predictivo cuya realización se plantea en el presente proyecto).

### 1.2. Resultados previos

El Análisis Exploratorio de Datos (EDA) y las pruebas de concepto iniciales han revelado dinámicas estructurales determinantes para la estrategia de modelado:

1. **Diagnóstico de Intermittencia:** Se cuantificó que aproximadamente el 68 % de las observaciones diarias en el conjunto de entrenamiento ( $d_1$  a  $d_{1913}$ ) corresponden a ventas cero. Este hallazgo descarta el uso de funciones de pérdida cuadráticas simples (MSE) como único optimizador, sugiriendo la necesidad de funciones de pérdida asimétricas o enfoques probabilísticos.
2. **Sensibilidad al Calendario:** Se identificó una fuerte correlación entre los picos de ventas y la variable categórica de eventos.
3. **Heterogeneidad Geográfica:** El análisis por estado reveló que la activación de los beneficios SNAP genera impulsos de demanda asincrónicos entre California (CA), Texas (TX) y Wisconsin (WI), validando la necesidad de segmentar o incluir el estado como variable categórica fundamental.

4. **Estabilidad de la Tendencia:** A través de medias móviles de 28 y 56 días, se observó que, si bien la volatilidad diaria es alta, la tendencia subyacente es relativamente estable, lo que refuerza la hipótesis de que los modelos basados en ventanas rodantes (*rolling windows*) tendrán un alto poder predictivo.

### 1.3. Hipótesis

Para la resolución del problema, se plantean las siguientes hipótesis de trabajo que guiarán la ingeniería de características y la selección del modelo:

- **H1 - Superioridad de la No-Linealidad:** Se postula que un modelo de ensamble basado en árboles de decisión (LightGBM o XGBoost) superará en desempeño (menor RMSSE) a los modelos lineales regularizados (Ridge/Lasso) y al *baseline* estacional (Naive). Esto se debe a la capacidad de los árboles para capturar interacciones no lineales entre el precio de venta y los eventos de calendario sin necesidad de especificar la forma funcional de la relación.
- **H2 - Memoria de Corto Plazo:** Se hipotetiza que las variables de rezago (*Lags*) correspondientes al ciclo semanal (lag 7) y al ciclo mensual (lag 28), junto con sus estadísticos rodantes (medias y desviaciones estándar), constituirán los predictores con mayor ganancia de información, superando a las variables estáticas del producto.
- **H3 - Efecto Precio:** La varianza en el precio relativo (precio actual vs. promedio histórico) actuará como una señal fuerte de elasticidad-precio. Se espera que una reducción del precio correlacione positivamente con el volumen de ventas, y que el modelo sea capaz de aprender esta elasticidad por categoría.

### 1.4. Estado del arte

El dominio del pronóstico de series temporales en retail ha evolucionado desde métodos estadísticos puros hacia enfoques híbridos y de aprendizaje profundo. Actualmente, el panorama se divide en:

1. **Métodos Estadísticos Clásicos (Benchmark):** Modelos como ARIMA, SARIMA y Suavizamiento Exponencial (ETS) siguen siendo robustos para series individuales con fuerte estacionalidad. Sin embargo, su incapacidad para compartir información entre series y su alto costo computacional al escalar a 30,000 SKUs limitan su uso en este contexto.
2. **Modelos de Aprendizaje Profundo (Deep Learning):** Arquitecturas como *DeepAR* (Amazon), *N-BEATS* y *Temporal Fusion Transformers* (TFT) representan la vanguardia académica. Son capaces de generar pronósticos probabilísticos y capturar dependencias de largo plazo. No obstante, su complejidad de entrenamiento y requerimientos de hardware (GPU) suelen ser prohibitivos para implementaciones ágiles con recursos limitados.
3. **Gradient Boosted Decision Trees (GBDT):** Representado por librerías como LightGBM, XGBoost y CatBoost. Este enfoque se considera el **estándar de facto en competencias**

**de ciencia de datos** para datos tabulares y series de tiempo estructuradas. Su ventaja radica en la eficiencia de memoria (mediante el uso de histogramas), velocidad de entrenamiento, manejo nativo de valores faltantes y variables categóricas, y alta interpretabilidad mediante métricas de importancia de características (Gain/Split). Este proyecto se alinea con este enfoque, buscando optimizar la implementación de LightGBM para el problema M5.

## 1.5. Criterios de éxito

La evaluación del éxito del proyecto se regirá por métricas cuantitativas estrictas y criterios de calidad de software:

- **Métrica Principal (RMSSE - Root Mean Squared Scaled Error):** Siguiendo los estándares de la competencia M5, el criterio principal de éxito será obtener un RMSSE en el conjunto de validación (últimos 28 días) **inferior a 0.80**.

$$RMSSE = \sqrt{\frac{\frac{1}{h} \sum_{t=n+1}^{n+h} (Y_t - \hat{Y}_t)^2}{\frac{1}{n-1} \sum_{t=2}^n (Y_t - Y_{t-1})^2}}$$

Un valor  $< 1,0$  indica que el modelo supera al método *Naive* estacional, validando que el aprendizaje automático está aportando valor real sobre la inercia histórica.

- **Métrica Secundaria (RMSE):** Se monitoreará la Raíz del Error Cuadrático Medio para evaluar la magnitud absoluta del error en unidades de producto, buscando minimizarlo consistentemente a través de un esquema de validación cruzada temporal (*Walk-Forward Validation*).
- **Eficiencia Computacional:** El *pipeline* de preprocesamiento y entrenamiento debe ser ejecutable dentro de los límites de memoria de un entorno estándar (ej. 12GB - 16GB RAM) sin recurrir a sub-muestreos que comprometan la representatividad de los datos. Esto implica el uso exitoso de técnicas de reducción de memoria (downcasting).
- **Reproducibilidad:** El éxito también se define por la entrega de un código modular y documentado que permita replicar los resultados desde la ingesta de datos crudos hasta la generación del archivo de envío (*submission*), garantizando que no exista fuga de datos (*data leakage*) del futuro hacia el pasado.

## 2. Estrategia de desarrollo e impacto

### 2.1. Producto final

El producto final que se entregará corresponderá al conjunto de entregables de la Parte Técnica del proyecto, es decir:

1. un repositorio, en la forma de un archivo .ipynb, que documentará los códigos generados en las distintas etapas del proyecto; y
2. una presentación beamer que destacará los resultados más importantes del proyecto.

### 2.2. Usuarios/beneficiarios

La presentación será expuesta frente al profesor Alfaro y los estudiantes de la versión 2025-2 del ramo de Proyectos Estadísticos de la carrera de Ingeniería Civil Matemática, MAT306. Esto será parte de un ciclo de presentaciones donde cada grupo del ramo expondrá su avance, de manera que los principales beneficiarios del presente proyecto serán los demás estudiantes de MAT306, en tanto que nuestra presentación será un insumo para aquel ciclo, la intención formativa del cual es exponer distintas formas de abordar el mismo desafío M5.

La industria y la academia podrían verse beneficiadas por nuestro proyecto en la medida que el profesor Alfaro (quien dirigirá dicho ciclo) invite como audiencia a otros profesores y/o a agentes del mundo empresarial, para evaluar y ser comunicados de nuestro avance del proyecto, que es una aplicación de herramientas que, en sí mismas, son de interés tanto académico como industrial.

### 2.3. Ventajas

Nuestro proyecto, frente a otras modalidades de análisis predictivo de series de tiempo, ofrece las ventajas de reproducibilidad y claridad. El énfasis en la creación de un archivo .ipynb, debidamente comentado y probadamente reproducible, hace más fácil la revisión de las distintas etapas del proyecto, aumentando la claridad e interpretabilidad del mismo.

### 2.4. Impacto esperado

El impacto esperado es que el resultado de nuestro proyecto (en la forma de los entregables de la Parte Técnica del proyecto) sirva como insumo a la actividad de carácter formativo a realizar en el ramo MAT306. Idealmente, en la medida que a dicha actividad acudan invitados de la industria y la academia, es posible y deseable que la exposición de nuestro avance resulte en la posibilidad de una aplicación de las técnicas empleadas durante el mismo a algún nuevo proyecto, provisto por alguno de los expertos invitados.

### 3. Capacidades y organización

#### 3.1. Equipo y roles

Los roles asignados en el equipo fueron los siguientes:

- Líder: Mauricio Pizarro
- Analistas:
  - Carlos Crisosto
  - Camilo Mesías
  - Nolberto Rivera
- Redactor de la presentación: Diego Melián

#### 3.2. Coordinación

La primera etapa, de definición del problema, será concertada, primordialmente, por el Líder del proyecto, para determinar explícitamente la problemática a abordar y los criterios de evaluación de los distintos hitos.

La segunda (y más extensa) etapa, correspondiente al desarrollo del proyecto, será realizada por los analistas. Esta consistirá en la creación de un único .ipynb que documentará todos los códigos realizados durante el proyecto. El código será trabajado comunitariamente en Google Colab, ocupando también para la comunicación a la hora de programar un canal de Discord expresamente generado para dicho fin.

Finalmente, la última etapa, correspondiente a la preparación del material de difusión, será realizada por un integrante del equipo a quien será designada esa tarea en específico.

Vale la pena aclarar que también durante las últimas dos etapas el líder tendrá la labor de armonizar y guiar al resto de integrantes del proyecto.



## 4. Objetivos y métricas

La definición de objetivos y métricas para el problema de pronóstico M5 no es trivial debido a la naturaleza jerárquica y esparsa de los datos. A continuación, se detalla la estructura teleológica del proyecto y el marco matemático de evaluación.

### 4.1. Objetivo General

Diseñar, implementar y validar un sistema de pronóstico de demanda a nivel granular (SKU por tienda) utilizando algoritmos de ensamblaje de árboles de decisión (*Gradient Boosting*), optimizando la función de pérdida para distribuciones de datos con alta intermitencia (ceros) y validando su robustez mediante un esquema de evaluación temporal rodante (*Walk-Forward*) para minimizar el error escalado en un horizonte de 28 días.

### 4.2. Objetivos Específicos

Para alcanzar el objetivo general, se descomponen las metas operativas en los siguientes puntos:

1. **OE1 - Caracterización de la Intermitencia:** Cuantificar y diagnosticar el nivel de esparsidad (porcentaje de ceros) y la estacionalidad de las 30,490 series temporales para justificar la selección de la función objetivo (distribución Tweedie vs. Poisson/Gaussiana).
2. **OE2 - Ingeniería de Características Temporales:** Construir un espacio de características (*feature space*) robusto que capture la dinámica de corto y largo plazo mediante la generación de rezagos (*lags*) y ventanas móviles (*rolling statistics*), mitigando el riesgo de fuga de datos (*data leakage*).
3. **OE3 - Modelado Eficiente en Memoria:** Entrenar y ajustar hiperparámetros de un modelo LightGBM capaz de procesar el dataset completo (aprox. 50 millones de filas tras la transformación) dentro de las restricciones de memoria (RAM < 16 GB), utilizando técnicas de optimización de tipos de datos (*downcasting*).
4. **OE4 - Validación Robusta:** Implementar una estrategia de validación cruzada temporal (*Walk-Forward Validation*) con al menos 3 particiones de 28 días, para estimar el error de generalización en periodos no vistos y evaluar la estabilidad del modelo ante cambios de tendencia.
5. **OE5 - Interpretabilidad y Decisión:** Identificar los impulsores clave de la demanda (precio, eventos, calendario) mediante métricas de importancia de características (Gain/Split) para proporcionar *insights* accionables sobre la sensibilidad de las ventas.

### 4.3. Métricas de Evaluación

Dada la heterogeneidad de las series (productos que venden miles de unidades vs. productos que venden casi cero), las métricas tradicionales como el RMSE o el MAPE presentan sesgos signifi-

cativos. Este proyecto adopta el marco de evaluación oficial de la competencia M5, priorizando métricas independientes de la escala.

#### 4.3.1. Métrica Principal: RMSSE (Root Mean Squared Scaled Error)

El RMSSE es la métrica rectora de este proyecto. A diferencia del RMSE estándar, el RMSSE penaliza el error en relación con la variabilidad histórica de la propia serie. Esto evita que los productos de alto volumen dominen la métrica de error global.

La fórmula del RMSSE para una serie individual  $i$  se define como:

$$RMSSE_i = \sqrt{\frac{\frac{1}{h} \sum_{t=n+1}^{n+h} (Y_t - \hat{Y}_t)^2}{\frac{1}{n-1} \sum_{t=2}^n (Y_t - Y_{t-1})^2}} \quad (1)$$

Donde:

- $Y_t$  es la venta real en el tiempo  $t$ .
- $\hat{Y}_t$  es el pronóstico en el tiempo  $t$ .
- $n$  es la longitud del historial de entrenamiento.
- $h$  es el horizonte de pronóstico (28 días).
- El denominador representa el **Error Cuadrático Medio del método Naive** en el conjunto de entrenamiento (asumiendo pronóstico  $Y_t = Y_{t-1}$ ).

#### Interpretación del RMSSE:

- $RMSSE = 1$ : El modelo tiene el mismo desempeño que el método *Naive* (repetir el valor de ayer).
- $RMSSE < 1$ : El modelo aporta valor predictivo superior al *Naive*.
- $RMSSE > 1$ : El modelo es peor que simplemente repetir el último valor observado.

#### 4.3.2. Métricas Secundarias

- **RMSE (Root Mean Squared Error)**: Se utilizará como función de pérdida durante el entrenamiento (o su aproximación mediante Tweedie) y para la interpretación operativa directa (unidades de error).

$$RMSE = \sqrt{\frac{1}{h} \sum_{t=1}^h (Y_t - \hat{Y}_t)^2}$$

- **WRMSSE (Weighted RMSSE):** Aunque el cálculo completo del WRMSSE (que pondera por volumen de venta en dólares) es costoso computacionalmente para iteraciones rápidas, se utilizará como referencia final. El objetivo es asegurar que el modelo no solo prediga bien los productos de baja rotación, sino que mantenga la precisión en los productos que generan mayores ingresos.

4.4. Criterios de Éxito y Alineación

El éxito del proyecto se dictamina en función de umbrales cuantitativos y la viabilidad técnica de la solución. La siguiente tabla resume la alineación entre objetivos y criterios de aceptación:

Objetivo	Descripción del Entregable	Criterio de Éxito (KPI)
Precisión del Modelo	Modelo LightGBM optimizado	<b>RMSSE</b> < 0,80 en el conjunto de validación (superando el benchmark Naive por un 20 %).
Estabilidad	Validación Walk-Forward (3 folds)	La desviación estándar del RMSE entre folds debe ser inferior al 10 % de la media del error.
Eficiencia (OE3)	Pipeline de procesamiento en memoria	Ejecución completa del <i>training</i> sin errores <b>MemoryError</b> en un entorno de 12GB RAM.
Valor de Negocio	Análisis de Importancia de Variables	Identificación correcta de patrones conocidos (ej. impacto SNAP, caída en Navidad) en el top 10 de <i>Feature Importance</i> .

Tabla 4.1: Matriz de alineación de objetivos y criterios de éxito.

## 5. Metodología

El diseño metodológico de este proyecto se fundamenta en el reconocimiento de la naturaleza estocástica y jerárquica de las series temporales de retail. A diferencia de los problemas de regresión estándar, donde se asume independencia entre observaciones (i.i.d.), el pronóstico de la demanda en el contexto M5 requiere un tratamiento que respete la estructura secuencial de los datos, la interdependencia entre productos y la alta prevalencia de valores cero (intermitencia).

Para abordar estos desafíos, se ha optado por un enfoque de aprendizaje supervisado utilizando algoritmos de *Gradient Boosting Decision Trees* (GBDT), específicamente LightGBM, optimizado mediante una función de pérdida de la familia Tweedie. A continuación, se detallan los esquemas de validación, las métricas de desempeño seleccionadas, los mecanismos de interpretabilidad y los controles de riesgo.

### 5.1. Estrategia de Validación

La validación de modelos predictivos en series temporales exige prevenir el uso de información futura. Por consiguiente, se descartaron técnicas como la validación cruzada aleatoria ( $K$ -Fold Cross Validation) en favor de una estrategia de **Validación Walk-Forward** (o de origen rodante).

El esquema implementado respeta estrictamente la causalidad temporal bajo las siguientes directrices:

- **Partición Cronológica:** El conjunto de datos, que abarca 1,913 días, se divide en un conjunto de entrenamiento ( $D_{train}$ ) y un conjunto de validación ( $D_{val}$ ).
- **Horizonte de Pronóstico ( $h$ ):** Se define un horizonte fijo de  $h = 28$  días, alineado con los requerimientos operativos de la competencia M5.
- **Ventana de Entrenamiento:** Se utilizan los datos históricos comprendidos entre el día  $d_1$  y el día  $d_{1885}$  para el ajuste de los parámetros del modelo.
- **Ventana de Validación (Out-of-Sample):** La evaluación del desempeño se realiza sobre los días  $d_{1886}$  a  $d_{1913}$ . Este periodo actúa como un sustituto del futuro desconocido, permitiendo estimar la capacidad de generalización del modelo ante patrones estacionales y tendencias recientes no vistas durante el entrenamiento.

Este diseño permite simular un entorno de producción real, donde el modelo se re-entrena periódicamente con nueva información disponible para predecir el mes subsiguiente.

### 5.2. Métricas de Evaluación

Dada la heterogeneidad en la escala de ventas (productos de alta rotación frente a productos de la “cola larga”), el uso exclusivo de métricas absolutas como el RMSE puede inducir sesgos hacia los artículos de mayor volumen. Por ello, se ha adoptado el marco de evaluación oficial de la competencia M5.

### 5.2.1. RMSSE (Root Mean Squared Scaled Error)

La métrica principal seleccionada es el RMSSE. Esta métrica es invariante a la escala y penaliza el error de pronóstico en relación con la volatilidad histórica de la serie. Se calcula dividiendo el Error Cuadrático Medio (MSE) del modelo por el MSE de un pronóstico de referencia ingenuo (*Naive Forecast*) dentro de la muestra.

Para una serie temporal individual  $i$ , el RMSSE se define matemáticamente como:

$$RMSSE_i = \sqrt{\frac{\frac{1}{h} \sum_{t=n+1}^{n+h} (Y_t - \hat{Y}_t)^2}{\frac{1}{n-1} \sum_{t=2}^n (Y_t - Y_{t-1})^2}} \quad (2)$$

Donde:

- $Y_t$  es la demanda real en el tiempo  $t$ .
- $\hat{Y}_t$  es la predicción del modelo.
- $n$  es la longitud del historial de entrenamiento.
- El denominador corresponde a la varianza de las diferencias de primer orden (error del modelo *Naive* estacional).

Un  $RMSSE < 1,0$  indica que el modelo propuesto supera en precisión al método simple de “repetir el último valor observado”, justificando así la complejidad computacional añadida del Machine Learning.

### 5.2.2. Métricas Auxiliares

Para complementar la evaluación y facilitar la interpretación operativa, se monitorean:

- **RMSE (Root Mean Squared Error):** Proporciona una medida del error en las unidades originales del producto, útil para dimensionar el stock de seguridad.
- **MAE (Mean Absolute Error):** Ofrece una perspectiva del error promedio, siendo menos sensible a valores atípicos (*outliers*) que el RMSE.

## 5.3. Interpretabilidad del Modelo

A pesar de que los modelos de ensamble suelen considerarse “cajas negras”, en este proyecto se han implementado técnicas para desagregar la contribución de las variables y garantizar la transparencia de las predicciones:

1. **Importancia de Características (Feature Importance by Gain):** Se analiza la ganancia total de información (reducción de entropía) aportada por cada variable en los nodos de los árboles de decisión. Esto permite identificar los *drivers* fundamentales de la demanda.
  - Se espera que las variables de *Lags* y *Rolling Means* dominen la importancia, seguidas por las variables de calendario (día de la semana) y precios.
2. **Análisis de Residuos:** Se examina la distribución de los errores ( $Y_t - \hat{Y}_t$ ) para detectar sesgos sistemáticos (e.g., subestimación constante en días de promoción) y verificar la homocedasticidad.

## 5.4. Gestión de Sesgos y Data Leakage

La integridad del modelo depende críticamente de la prevención de fugas de información (*data leakage*) y la mitigación de sesgos estadísticos inherentes a los datos de retail.

### 5.4.1. Prevención de Leakage Temporal

El riesgo más severo en series temporales es utilizar información del futuro para predecir el pasado. Se han implementado los siguientes controles:

- **Restricción de Lags:** Se prohíbe el uso de rezagos menores al horizonte de pronóstico ( $lag < 28$ ). Esto garantiza que, al momento de inferir el día  $t + 1$ , no se requiera el dato real del día  $t$ , el cual no estaría disponible en un escenario productivo real. Las características dinámicas se construyen exclusivamente sobre  $t - 28$  y anteriores.
- **Codificación Cautelosa:** Las codificaciones de variables categóricas (*Target Encoding*) se realizan basándose estrictamente en los datos de entrenamiento, evitando que la media de ventas del conjunto de validación contamine los predictores.

### 5.4.2. Mitigación del Sesgo de Intermitencia

Dada la alta proporción de ceros en el dataset (aproximadamente 68%), un modelo entrenado con una función de pérdida cuadrática (MSE) tendería a converger hacia cero para minimizar el error global, subestimando los picos de demanda real.

Para corregir este sesgo, se utiliza la función de pérdida **Tweedie**. Esta distribución es un caso especial de los Modelos Lineales Generalizados (GLM) que modela una masa puntual en cero (Poisson) y una distribución continua para valores positivos (Gamma). Al ajustar el parámetro de varianza  $p$  (donde  $1 < p < 2$ ), el modelo aprende a discriminar entre la probabilidad de venta (clasificación implícita) y la magnitud de la venta (regresión), adaptándose naturalmente a la dispersión de los datos del M5.

## 6. Ética, permisos y datos

### 6.1. Datos

El dataset ocupado corresponderán a los provistos por el desafío M5 de Kuggle. Este dataset es completamente públicos, y no contiene datos sensibles de los consumidores.

Aunque no se citen expresamente, ocuparemos herramientas en las lectures del ramo MAT306, de autoría del profesor Alfaro.

### 6.2. Uso responsable

El único aspecto que creemos importante mencionar respecto a la aplicación de este tipo de herramientas fuera del entorno académico, es que la categorización de los productos como aptos o no para SNAP puede ser un tema problemático, por la trascendencia en la opinión pública de dicho programa.

## **7. Resultados e hitos**

### **7.1. Resultados**

### **7.2. Hitos**

Los hitos (o actividades) del 1 al 9 detallados en la Carta Gantt tienen como entregable el archivo .ipynb en los repositorios de cada uno de los integrantes del equipo.