INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Licenciatura en Ciencia de Datos

Analítica y Visualización de Datos

Profesor: Alejandro López Gómez

Proyecto Final

Análisis Estratégico de Datos de E-Commerce mediante Procesamiento de Señales y Segmentación de Clientes

Alumno

Antonio Eugenio Daniel

Domínguez Espinoza Juan Pablo 2024630346

Grupo: 5AM1

Fecha: 7 de enero de 2026

Contenido

Tabla de figuras 2
Introducción 3
Objetivos 3
Objetivo general 3
Objetivos específicos 4
Metodología 4
Fuente de datos y herramientas 4
Preparación y Calidad de Datos (ETL) 4
Evaluación de Métodos de Suavizado y Análisis Espectral 5
Estrategia de Segmentación de Clientes (Modelo RFM) 5
Validación y Reducción de Dimensionalidad 5
Resultados 6
Detección de Ciclos y Filtrado de Ruido 6
Segmentación de Clientes (Clustering) 7
Dashboard interactivo de Inteligencia de Negocios 8
Conclusiones 8

Tabla de figuras

Figura 1. Comparativa de técnicas de reducción de ruido: Medias Móviles vs. Filtro Espectral FFT. 6

Figura 2. Proyección de los segmentos de clientes basada en métrica RFM 7

Figura 3. Interfaz principal del Dashboard de Analítica Estratégica 8

Introducción

En la era actual de la economía digital, el comercio electrónico (*E-Commerce*) se ha consolidado como un generador masivo de información transaccional. Sin embargo, el volumen de datos por sí mismo no garantiza el éxito empresarial; el verdadero valor reside en la capacidad de procesar, limpiar y transformar estos datos crudos en conocimiento accionable (Inteligencia de Negocios). Las organizaciones que basan sus estrategias en la intuición corren el riesgo de perder competitividad frente a aquellas que adoptan una cultura basada en datos (*Data-Driven*), capaces de predecir ciclos de demanda y personalizar la experiencia del cliente.

El presente proyecto se centra en el análisis integral de un conjunto de datos transaccional de un minorista con sede en el Reino Unido, abarcando operaciones realizadas entre 2010 y 2011. La problemática central abordada es la necesidad de identificar patrones ocultos de comportamiento de compra y segmentar eficazmente la base de clientes para optimizar la toma de decisiones estratégicas.

Para lograr esto, se ha diseñado e implementado un flujo de trabajo de Ciencia de Datos (*Pipeline*) que va más allá del análisis descriptivo tradicional. La metodología integra técnicas avanzadas de **Procesamiento de Señales**, utilizando la Transformada Rápida de Fourier (FFT) para detectar matemáticamente la estacionalidad y ciclos de venta semanales, superando las limitaciones de los métodos de suavizado estadístico convencionales como las medias móviles. Asimismo, se emplean técnicas de **Reducción de Dimensionalidad** mediante Análisis de Componentes Principales (PCA) para la reconstrucción de señales y eliminación de ruido.

Finalmente, para la caracterización de usuarios, se aplica el modelo de segmentación **RFM** (Recencia, Frecuencia y Monto) potenciado por algoritmos de aprendizaje no supervisado (**K-Means Clustering**). Todos los hallazgos y modelos se consolidan en una herramienta de visualización interactiva (Dashboard web), permitiendo la exploración dinámica de los resultados y facilitando la detección de oportunidades comerciales en mercados internacionales.

Objetivos

Objetivo general

Analizar el comportamiento de ventas de una tienda de comercio electrónico (*E-Commerce*) utilizando técnicas de ciencia de datos, con el fin de descubrir patrones de consumo ocultos y agrupar a los clientes en segmentos clave para mejorar la toma de decisiones.

Objetivos específicos

1. **Limpieza y Preparación de Datos:** Procesar la base de datos original para corregir errores, eliminar transacciones canceladas y manejar datos faltantes, asegurando que la información sea confiable antes de analizarla.
2. **Evaluación de Métodos y Detección de Ciclos:** Comparar distintas técnicas de suavizado de datos (como medias móviles y PCA) para seleccionar la más efectiva y, posteriormente, aplicar la **Transformada de Fourier** para identificar con precisión los ciclos de venta repetitivos.
3. **Segmentación de Clientes:** Clasificar a los usuarios en grupos diferenciados (como clientes VIP o frecuentes) aplicando el método **RFM** (Recencia, Frecuencia y Monto) junto con el algoritmo de agrupamiento **K-Means**.
4. **Visualización de Resultados:** Desarrollar un **Dashboard interactivo** que permita explorar gráficamente los hallazgos del análisis, los segmentos de clientes y el rendimiento de las ventas por región.

Metodología

Para dar cumplimiento a los objetivos planteados, se diseñó un flujo de trabajo analítico dividido en cinco etapas secuenciales, desde la adquisición de los datos hasta la visualización de resultados.

Fuente de datos y herramientas

Se utilizó un conjunto de datos transnacional (*dataset*) perteneciente a una empresa minorista del Reino Unido, obtenido del repositorio público **Kaggle**:

https://www.kaggle.com/datasets/carrie1/ecommerce-data/data

El registro abarca transacciones realizadas entre el **01/12/2010 y el 09/12/2011**. El procesamiento y análisis se llevó a cabo utilizando el lenguaje de programación **Python**, empleando librerías especializadas para cálculo numérico y ciencia de datos, incluyendo **Pandas** (manipulación de datos), **NumPy** (álgebra lineal) y **Scikit-Learn** (modelado estadístico).

Preparación y Calidad de Datos (ETL)

Para garantizar la integridad del análisis, se aplicó un estricto proceso de limpieza y transformación que incluyó:

- **Depuración de duplicados:** Se identificaron y eliminaron 5,268 registros duplicados que podían sesgar los resultados.

- **Filtrado de transacciones inválidas:** Se excluyeron facturas con prefijo 'C' (cancelaciones), así como registros con precios unitarios o cantidades negativas/cero.
- **Tratamiento de valores faltantes:** Se eliminaron las filas sin *CustomerID*, ya que la identificación del cliente es un requisito indispensable para la fase de segmentación.
- **Ingeniería de características:** Se generó la variable *TotalAmount* calculando el producto de la cantidad por el precio unitario.

Evaluación de Métodos de Suavizado y Análisis Espectral

Con el objetivo de detectar estacionalidad en las ventas diarias, se procedió en dos fases:

1. **Fase Exploratoria de Denoising:** Se compararon distintas técnicas para reducir el ruido de la señal de ventas, incluyendo **Medias Móviles** (7 y 30 días) y reconstrucción mediante **PCA**. Se observó que los métodos tradicionales introducían un retardo temporal (*lag*) no deseado.
2. **Fase de Análisis Espectral:** Basado en lo anterior, se seleccionó la **Transformada Rápida de Fourier (FFT)**. Esta técnica permitió descomponer la serie de tiempo en el dominio de la frecuencia, identificando con precisión matemática los ciclos de periodicidad (patrones semanales) sin perder la alineación temporal de los datos.

Estrategia de Segmentación de Clientes (Modelo RFM)

Se implementó una segmentación conductual basada en el modelo **RFM** (Recencia, Frecuencia, Monto):

1. **Cálculo de Métricas:** Se determinaron los días desde la última compra, la cantidad de transacciones y el valor monetario total por cliente.
2. **Preprocesamiento del Modelo:** Dado el sesgo en la distribución de los datos, se aplicó una **transformación logarítmica** para normalizar las variables, seguida de una estandarización **Z-Score** para igualar las escalas.
3. **Clustering:** Se aplicó el algoritmo no supervisado **K-Means**, agrupando a los clientes en clústeres homogéneos según sus hábitos de consumo.

Validación y Reducción de Dimensionalidad

Para corroborar la calidad de la segmentación, se ejecutaron dos procesos de validación:

- **Análisis de Correlación:** Se aplicaron los coeficientes de Pearson y Spearman para entender la relación lineal y monótona entre las variables RFM.

- **Visualización con PCA:** Se utilizó el Análisis de Componentes Principales (PCA) para reducir las 3 dimensiones del modelo RFM a 2 componentes principales. Esto permitió proyectar los clústeres en un plano cartesiano (2D), facilitando la validación visual de la separación y cohesión entre los segmentos de clientes.

Resultados

A continuación, se presentan los hallazgos principales derivados del procesamiento de datos y la implementación de los modelos analíticos.

Detección de Ciclos y Filtrado de Ruido

Mediante la comparativa de métodos de suavizado, se demostró que las medias móviles (de 7 y 30 días) introducían un retraso temporal significativo que dificultaba la identificación precisa de puntos de inflexión en las ventas.

En contraste, la aplicación de la Transformada Rápida de Fourier (FFT) permitió aislar las frecuencias dominantes de la serie temporal. El análisis espectral reveló un pico de magnitud significativo en la frecuencia aproximada de 0.14, lo que corresponde matemáticamente a un periodo de 7 días (1/0.14 = 7.14).

Este hallazgo confirma estadísticamente la existencia de un ciclo de ventas semanal robusto, validando la necesidad de estrategias de inventario y marketing diferenciadas por día de la semana.

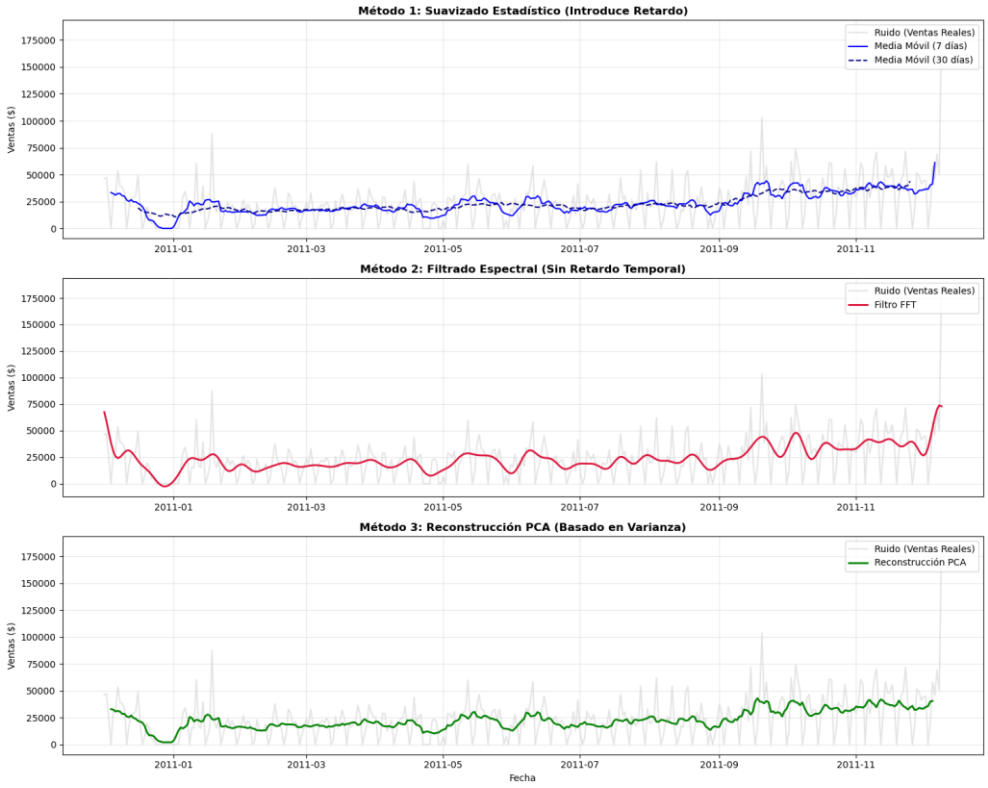

Figura 1. Comparativa de técnicas de reducción de ruido: Medias Móviles vs. Filtro Espectral FFT.

Segmentación de Clientes (Clustering)

El modelo K-Means, alimentado con las variables RFM normalizadas, identificó tres grupos (clústeres) de clientes con comportamientos claramente diferenciados. La reducción de dimensionalidad mediante PCA permitió visualizar estos grupos en un plano bidimensional, confirmando una separación efectiva con una superposición mínima.

Los segmentos identificados se caracterizan de la siguiente manera:

- **Clientes VIP (Oro):** Usuarios con alta frecuencia de compra, alto valor monetario y recencia muy baja (compraron hace poco). Representan el segmento más valioso para fidelizar.
- **Clientes Recurrentes (Plata):** Grupo con frecuencia y gasto promedio. Son usuarios estables que requieren estrategias de *cross-selling* para incrementar su valor.
- **Clientes Ocasionales/En Riesgo (Bronce):** Usuarios con baja frecuencia y alta recencia (hace mucho que no compran). Este segmento es candidato para campañas de reactivación.

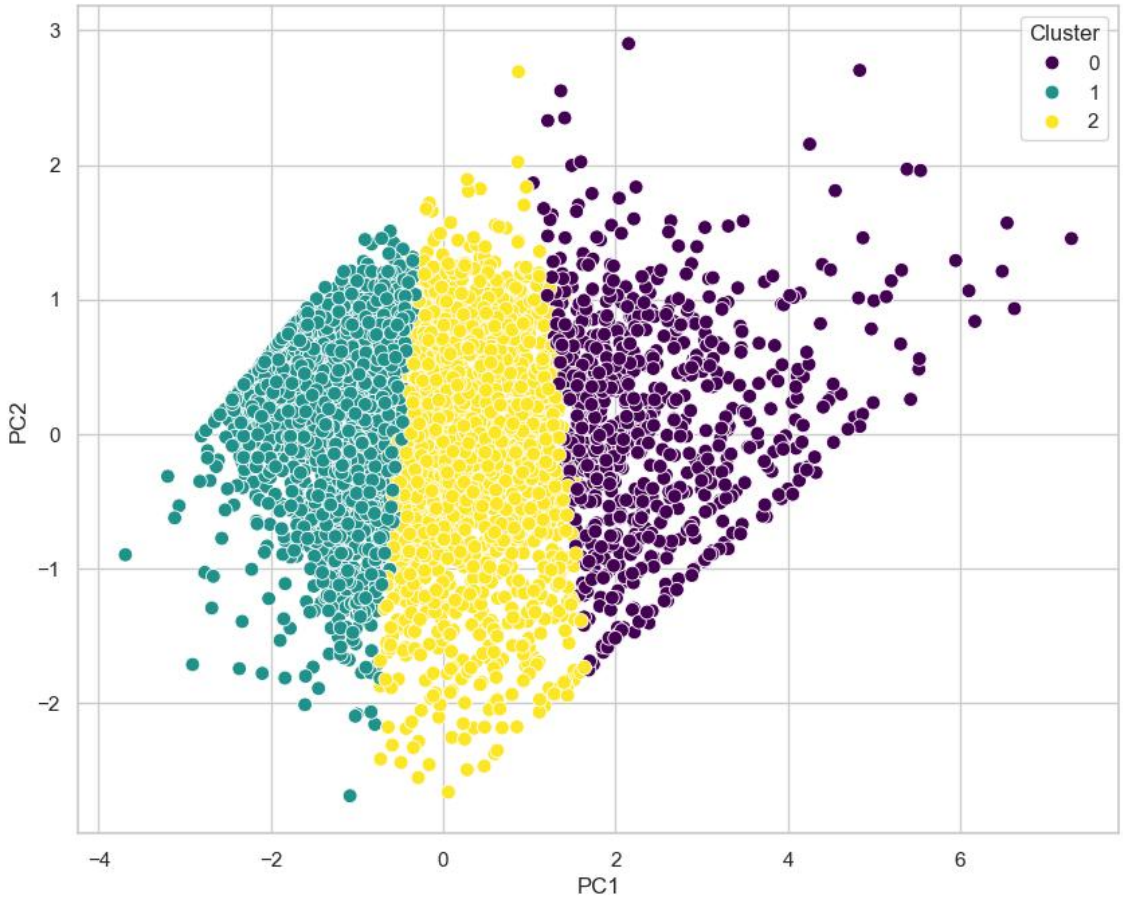

Figura 2. Proyección de los segmentos de clientes basada en métrica RFM

Dashboard interactivo de Inteligencia de Negocios

Como producto final, se desarrolló e implementó una aplicación web utilizando el framework **Streamlit**, la cual integra los modelos matemáticos procesados. Esta herramienta permite a los usuarios de negocio:

1. **Monitorear la salud de los datos:** Visualización de la serie temporal original frente a la señal filtrada por FFT.
2. **Explorar la segmentación:** Filtrado dinámico de clientes por clúster para analizar métricas específicas.
3. **Análisis Geográfico:** Un mapa de calor global que identifica los mercados internacionales con mayor penetración fuera del Reino Unido.

Figura 3. Interfaz principal del Dashboard de Analítica Estratégica

Conclusiones

El desarrollo de este proyecto ha permitido validar cómo la integración de técnicas avanzadas de ciencia de datos transforma registros transaccionales crudos en una herramienta estratégica de alto valor para la toma de decisiones empresariales. A través de la implementación de un flujo de trabajo completo, desde la limpieza de datos hasta el despliegue de una aplicación web, se alcanzaron los siguientes hitos:

En primer lugar, **la superioridad del análisis espectral sobre los métodos estadísticos tradicionales.** La fase experimental demostró que, si bien las medias móviles son útiles para visualizar tendencias generales, introducen un retardo temporal (*lag*) que puede oscurecer la reactividad del análisis. La aplicación de la **Transformada de Fourier (FFT)** resultó ser una solución más robusta desde la perspectiva de la ingeniería de señales, permitiendo identificar

matemáticamente un ciclo de venta semanal de 7 días con precisión, eliminando el ruido aleatorio sin comprometer la alineación temporal de los datos.

En segundo lugar, **la segmentación efectiva como motor de estrategia.** La combinación del modelo RFM con el algoritmo de aprendizaje no supervisado **K-Means** logró desglosar una base de datos heterogénea en tres clústeres accionables. La validación visual mediante **PCA (en 2D y 3D)** confirmó que estos grupos no son arbitrarios, sino que representan comportamientos de consumo estructuralmente distintos (VIP, Recurrentes y Ocasionales), lo cual habilita a la empresa para dejar de usar estrategias genéricas y pasar a campañas de marketing personalizadas.

Finalmente, **la democratización del dato mediante visualización interactiva.** La construcción del Dashboard en **Streamlit** cierra la brecha técnica entre el modelo matemático y el usuario final. Al permitir la exploración dinámica de los clústeres y la geografía de las ventas, la herramienta deja de ser un ejercicio académico para convertirse en un prototipo funcional de Inteligencia de Negocios (*Business Intelligence*), demostrando que el valor de los datos no reside solo en su análisis, sino en su capacidad de ser comunicados eficientemente.