UNIVERSIDAD DEL VALLE DE GUATEMALA

CC3105 – Machine Learning Engineering

Ing. Diego Jossué Contreras Méndez

Laboratorio 1 - CRISP-DM

Integrantes:

Arturo Argueta	21527
Edwin de León	22809
Diego Leiva	21752
José Pablo Orellana	21970

Guatemala, 24 de julio de 2025

Índice

Buisiness Understanding

Objetivos comerciales

General

Optimizar la gestión de ventas y clientes para una tienda basada en Reino Unido mediante el análisis histórico de transacciones con el fin de mejorar las decisiones de marketing, inventario y fidelización de clientes.

Específicos

- Identificar los productos más vendidos y menos vendidos para optimizar el inventario
- Analizar el comportamiento de los clientes por segmentos
- Detectar patrones de compra para campañas de marketing
- Estimar valor de vida del cliente

Inventario de recursos

Humanos

- Analista de datos
- Científico de datos
- Ingeniero de datos
- Ingeniero en sistemas

Recursos Técnicos

- Ambiente de Python:
    - Pandas
    - Scikitlearning
    - Seaborn
    - Excel (database)
- Ambiente en R
- Ordenador:
    - RAM >=16 GBS
    - Almacenamiento: 20 GB
    - Acceso a internet para repositorios
- Repositorios
    - Github
    - Gitlab
- Supuestos:
    - Se cuenta con permiso para uso de datos
    - Los datos han sido limpiados
    - Los datos no han sido alterados

Objetivos técnicos

- Limpiar y preparar conjunto de datos
- Realizar un anális exploratorio inicial
- Implementación de RFM para segmentación de los clientes
- Visualización de datos históricos para tendencias de venta
- Aplicar clustering para agrupación de clientes

Data Understanding

Datos iniciales

- Invoice: Número de factura.
- StockCode: Código del producto.
- Description: Descripción del producto.
- Quantity: Cantidad comprada.
- InvoiceDate: Fecha y hora de la transacción.
- Price: Precio unitario.
- Customer ID: Identificador del cliente.
- Country: País del cliente.

Datos complementarios

- Canal de venta (ej. web, móvil, email).
- Información demográfica del cliente (edad, género, ingresos).
- Costo del producto (para calcular ganancia).
- Descuentos/promociones aplicadas.
- Forma de pago.
- Frecuencia de visitas al sitio web

Descripción de datos

1. Invoice (Número de Factura)
    - Tipo de dato: Texto (string)
    - Representa el identificador único de cada transacción.
    - Las facturas que inician con la letra "C" indican cancelaciones o devoluciones.
    - Se identificaron 25,900 facturas únicas.

2. StockCode (Código de Producto)
    a. Tipo de dato: Texto
    b. Representa el código asignado a cada producto.
    c. Se encontraron 4,038 códigos únicos, lo que sugiere una amplia variedad de productos en el catálogo.

3. Description (Descripción del Producto)
    a. Tipo de dato: Texto

b. Describe en lenguaje natural el nombre o tipo de producto.
c. Hay 1,454 valores nulos, posiblemente por registros incompletos o errores de carga.
d. Se identificaron 4,008 descripciones únicas, lo que concuerda con la diversidad de códigos de producto.

4. Quantity (Cantidad Comprada)
a. Tipo de dato: Entero
b. Indica cuántas unidades del producto se vendieron o devolvieron en la transacción.
c. Valores negativos representan devoluciones.
d. Presenta 4360 valores distintos, lo que refleja múltiples combinaciones de cantidades por producto.

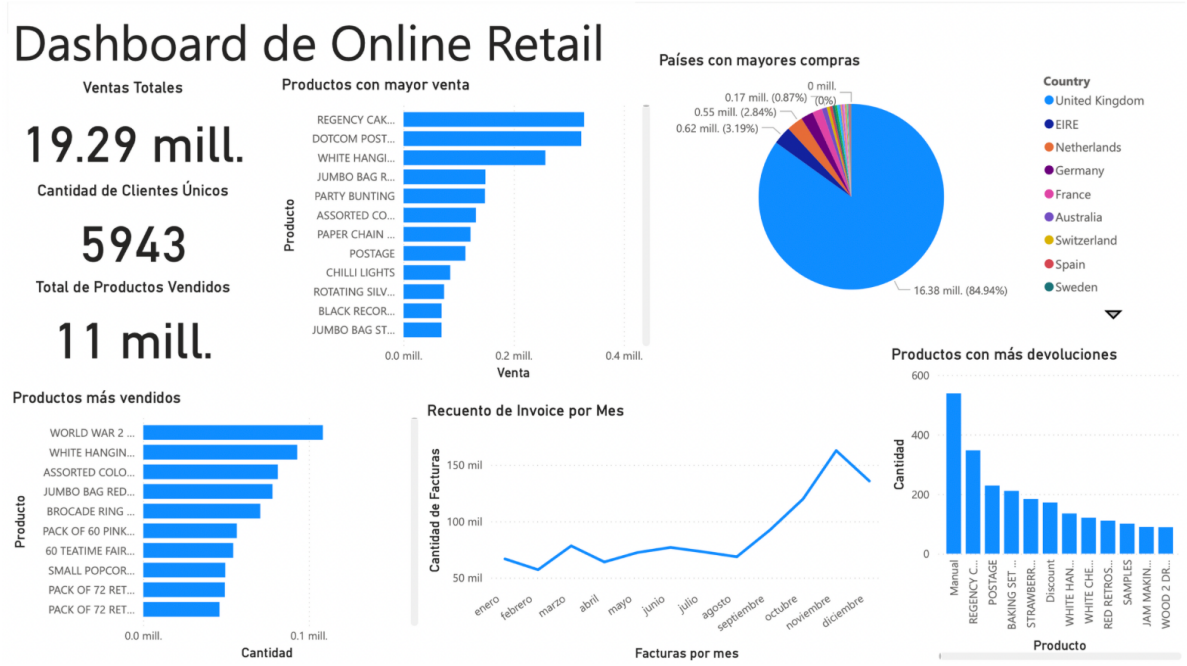
5. InvoiceDate (Fecha y Hora de la Factura)
a. Tipo de dato: Fecha y hora (datetime)
b. Marca el momento exacto en que se realizó la compra o devolución.
c. Cubre el periodo de diciembre 2010 a diciembre 2011.
d. Existen 20,621 fechas distintas, lo cual es esperable si se incluyen horas, minutos y segundos.

6. Price (Precio Unitario)
a. Tipo de dato: Decimal (float)
b. Representa el precio de una sola unidad del producto (antes de aplicar cantidad).
c. Hay 2,893 precios únicos, lo cual indica variedad en la estrategia de precios o promociones.
d. Es fundamental para calcular ingresos y márgenes.

7. Customer ID (Identificador del Cliente)
a. Tipo de dato: Decimal (aunque representa un identificador entero)
b. Se identificaron 5,874 clientes únicos.
c. Hay 133,600 registros sin ID, posiblemente ventas sin cliente registrado (compras anónimas o datos faltantes).

8. Country (País del Cliente)
a. Tipo de dato: Texto
b. Indica el país desde donde se realizó la compra.
c. Se encontraron 38 países diferentes, destacando United Kingdom como el país con mayor volumen de registros.

Dashboard



Calidad de los Datos

1. Completitud

- **Customer ID**:
  - Tiene **133,600 valores nulos** (24.65% del total), lo que representa una pérdida considerable de información sobre los clientes.
- **Description**:
  - Contiene **1,454 valores nulos**. Aunque es menor en proporción, impacta en el análisis de productos.
- Todas las demás columnas están **completas (sin valores nulos)**.

Conclusión: El dataset **no está completamente completo**, siendo Customer ID la columna con mayor impacto.

2. Consistencia

- **Invoice**:
  - Formato mixto: algunas facturas comienzan con “C” (devoluciones), lo cual es esperable, pero se deben manejar por separado.
- **Quantity**:
  - Existen **valores negativos** (devoluciones), lo cual es válido, pero debe tratarse adecuadamente en los análisis financieros.
- **Price**:
  - Hay registros con precio = 0 o negativo. Esto puede reflejar:
    - Promociones
    - Errores de captura
    - Productos de muestra

Conclusión: Datos mayormente consistentes, pero hay casos que requieren validación o tratamiento especial.

3. Exactitud

- No se puede verificar directamente sin fuentes externas, pero hay **sospechas de errores**:
  - Descripciones faltantes.
  - Precios en cero o negativos.
  - Cantidades demasiado altas o atípicas podrían ser errores de entrada.

Conclusión: Hay **riesgo moderado de errores de captura**, sobre todo en precios y descripciones.

4. Validez

- **Fechas:** InvoiceDate tiene formato correcto y rango válido (2010–2011).
- **Country:** 38 valores válidos, sin errores de codificación aparente.
- **Quantity y Price:** tipo de dato correcto, aunque algunos valores no son válidos lógicamente (negativos o extremos).

Conclusión: La mayoría de los datos son válidos según el dominio, pero deben controlarse outliers.

5. Unicidad

- No hay un identificador único por fila (no existe un TransactionID).
- Combinaciones como Invoice + StockCode podrían repetirse (una factura puede incluir varios productos).
- Hay **duplicados exactos de filas**: esto podría indicar errores de registro.

Conclusión: Falta un identificador único por transacción; se recomienda revisar duplicados.

Data Prep

Datos Excluidos

Durante el proceso de limpieza, se eliminaron diferentes registros por las razones indicadas.

- Registros sin Customer ID, ya que estas representan compras anonimas que no permiten un buen analisis de segmentacion.
- Filas con Quantity<=0 o Price <=0, ya que esto se considera como un posible error de entrada o alguna promoción de la que no se tiene conocimiento.
- Registros con Description nulo, debido a que no es posible identificar a que producto hace referencia la compra.
- Duplicados exactos, se identificaron y eliminaron registros duplicados que podrían deberse a errores de carga.

Durante la limpieza de datos, se eliminaron 261,822 registros (~24.5% del total). Si bien esta cifra podría parecer alta, la eliminación se basó principalmente en que los registros sin Customer ID no permiten analizar el comportamiento individual ni segmentar clientes, por lo que no aportan valor al análisis. Las descripciones vacías impiden identificar productos. Y finalmente los registros con Quantity o Price menores o iguales a cero representan devoluciones, promociones o errores, y su imputación con valores promedio distorsionaría los ingresos reales.

Por lo tanto, se consideró que la eliminación era preferible a la imputación, preservando la integridad y validez del análisis. La base final conserva más de 805 mil registros de alta calidad, suficientes para extraer conclusiones representativas.

Nuevas Variables Creadas

Para enriquecer el análisis se crearon nuevas variables:

- TotalSale, deriva de la multiplicación de Quantity y Price, lo cual representa el ingreso total de la venta de ese producto.
- Devolución, surge del valor booleano que indica si la factura es una cancelación o no.

Pipeline de Transformación

La transformación se documentó en un Jupyter Notebook, siguiendo este flujo:

- Carga del archivo unificado combined_online_retail.csv
- Conversión de tipos de datos (InvoiceDate a datetime, Customer ID a entero, etc.)
- Limpieza:
    - Eliminación de nulos en Customer ID y Description
    - Filtrado de Quantity y Price negativos o cero
    - Remoción de duplicados exactos
- Creación de variables nuevas: TotalVenta, EsDevolucion, Mes
- Exportación de dataset limpio para análisis en Power BI

Modeling

Técnicas posibles para el problema

Dado que el objetivo es predecir el valor de ventas (TotalSale) en base a variables como cantidad (Quantity), precio (Price), país (Country), y otros atributos, se trata de un problema de regresión supervisada. Las técnicas posibles incluyen:

- Regresión Lineal
- Árboles de Decisión
- Random Forest Regressor
- XGBoost Regressor

En este caso, se optó por Random Forest Regressor, una técnica robusta y eficaz para relaciones no lineales y con variables categóricas codificadas.

Definición del error (métrica de evaluación)

Se utilizaron varias métricas de regresión para medir el desempeño del modelo:

- MSE (Mean Squared Error): 74.19
- RMSE (Root Mean Squared Error): 8.61
- MAE (Mean Absolute Error): 3.26
- MAPE (Mean Absolute Percentage Error): 42.06%
- MedAE (Mediana del Error Absoluto): 1.84
- R^2 Score: 0.984
- Explained Variance Score: 0.984

La métrica principal considerada fue R^2, que indica qué tan bien el modelo explica la variación en los datos. Un valor cercano a 1 confirma un buen desempeño.

Modelo aplicado

Se utilizó un modelo de **Random Forest Regressor** con los siguientes parámetros:

- n_estimators = 30 (número de árboles)
- max_depth = 10
- random_state = 42

Este modelo permitió capturar relaciones complejas y no lineales entre las variables. Se evitó el sobreajuste al limitar la profundidad y el número de árboles.

Evaluation

Los resultados obtenidos tras entrenar el modelo Random Forest Regressor sobre el dataset de ventas retail son los siguientes

```
MSE: 74.33
R²: 0.9836
MAE: 3.26
RMSE: 8.62
MedAE: 1.84
MAPE: 42.06%
Explained Variance: 0.9836
```

Estos valores reflejan un muy buen desempeño del modelo:

- Un R^2 de 0.9836 indica que el modelo explica el 98.36% de la variabilidad del valor de venta total.

- La gráfica de dispersión muestra que la mayoría de los puntos están muy cerca de la línea ideal, lo que sugiere que el modelo tiene alta capacidad de generalización.

- Las métricas de error como MAE (3.26) y MedAE (1.84) indican que el modelo comete errores pequeños en términos absolutos en la mayoría de los casos.

- Sin embargo, el MAPE de 42.06% sugiere que en proporción a ciertos valores reales bajos, los errores relativos pueden ser altos. Esto es típico cuando existen ventas muy pequeñas en el dataset.

¿Se puede publicar la solución?

Sí, el modelo puede ser publicado y utilizado en un entorno de producción controlado, con las siguientes consideraciones:

- Es robusto y generaliza bien para datos similares a los del conjunto original.
  Puede integrarse en un sistema de predicción de ingresos o ventas para apoyar decisiones comerciales o estrategias de inventario.
  La solución puede ser incorporada en pipelines de MLOps con métricas de monitoreo para validar su desempeño en tiempo real.
  Limitaciones del Modelo

Aunque el desempeño es alto, el modelo presenta algunas limitaciones importantes:

1. Dependencia del histórico: Está entrenado con datos pasados, por lo tanto no responde bien a cambios bruscos de mercado o comportamiento del consumidor no registrado en el dataset.
2. Datos de entrada limitados: No se consideraron variables temporales avanzadas (como estacionalidad, días festivos, campañas), lo cual podría mejorar las predicciones.

Deployment

Monitores Sugeridos

- **Monitoreo de segmentos de clientes:** rastrear la proporción de clientes en cada segmento. Un cambio abrupto podría indicar cambios en el comportamiento de los clientes o problemas de datos.
- **Monitoreo de rendimiento del modelo:** para modelos supervisados, se medirían métricas como error medio absoluto o accuracy y se registrarían a lo largo del tiempo para detectar degradación.
- **Monitoreo de negocio:** métricas de negocio como ingresos diarios, tasa de retorno, tasa de conversión y valores promedio de factura deben compararse con los objetivos comerciales; caídas inexplicables podrían indicar problemas de datos o en el modelo.
- **Monitoreo de infraestructura:** supervisar tiempos de respuesta y uso de recursos del servicio de inferencia para garantizar que el sistema sea escalable y esté disponible.

Sugerencia para utilizar modelo en producción

- **Ingesta de datos:** un proceso automatizado carga periódicamente las nuevas transacciones desde el sistema de facturación hacia un almacén de datos.
- **Preprocesamiento:** el pipeline de preparación limpia registros duplicados y valores faltantes, calcula variables de recencia, frecuencia y valor monetario, y normaliza las variables numéricas. Este pipeline se guarda como componente reutilizable.

- **Predicción y segmentación:** el modelo de segmentación genera la etiqueta de segmento para cada cliente. Si se utilizan modelos de predicción de valor de vida, el sistema calcula el valor esperado para cada cliente.