**UNIVERSIDAD TECNOLÓGICA DE PANAMÁ**

**VICERRECTORÍA DE INVESTIGACIÓN, POSTGRADO Y EXTENSIÓN DE LA FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES**

MAESTRÍA EN ANALÍTICA DE DATOS

**CURSO: MODELOS PREDICTIVOS**

PROYECTO FINAL

ANÁLISIS DE COMPORTAMIENTO DE COMPRA Y PREDICCIÓN DE MEDIOS DE PAGO EN UN ENTORNO DIGITAL

ESTUDIANTE:

VICTORIA VIOLETA RODRÍGUEZ GALLARDO 4-772-749

PROFESOR:

JUAN M. CASTILLO, PhD

GRUPO:

AÑO

2025

**Contenido**

[Introducción 3](#_Toc194409100)

[Justificación 4](#_Toc194409101)

[Antecedentes 5](#_Toc194409102)

[Definición del problema 6](#_Toc194409103)

[Avance de análisis predictivo 7](#_Toc194409104)

[1. Limpieza de datos y preparación del dataset 7](#_Toc194409105)

[2. Análisis exploratorio de datos (EDA) 7](#_Toc194409106)

[3. Modelado con Random Forest 7](#_Toc194409107)

[Gráficas de visualización 8](#_Toc194409108)

[1. Exploración general del comportamiento de compra 8](#_Toc194409109)

[2. Segmentación por características del cliente 9](#_Toc194409110)

[3. Visualización contable 11](#_Toc194409111)

# **Introducción**

Comprender el comportamiento de los usuarios dentro de una plataforma de comercio electrónico resulta esencial para optimizar procesos, mejorar la experiencia del cliente y aumentar la eficiencia de las estrategias comerciales. En este estudio se desarrolló un análisis profundo con base en un conjunto de datos que recoge información variada de los consumidores, incluyendo datos demográficos como edad, género y nivel de ingresos, así como aspectos conductuales como el dispositivo utilizado para comprar, el tipo de producto adquirido y la intención detrás de la compra. El foco principal se centra en predecir el método de pago que el usuario elige al realizar una compra. Para esto, se emplearon técnicas de análisis exploratorio de datos, visualización gráfica y modelos predictivos construidos en lenguaje R. A su vez, se realizó un análisis contable y de series de tiempo para complementar la visión de negocio, proporcionando indicadores que permiten observar patrones de ingresos, preferencias y comportamientos relevantes. Este trabajo se enmarca en un contexto práctico donde el uso de la ciencia de datos permite anticipar decisiones y diseñar respuestas personalizadas desde las plataformas de ecommerce.

# **Justificación**

Las decisiones que toma un consumidor durante su experiencia de compra en línea están influenciadas por múltiples factores. Entender estos elementos, y especialmente el método de pago elegido, puede marcar una diferencia significativa en el éxito de las estrategias comerciales. Esta información permite no solo personalizar campañas de marketing, sino también ajustar la experiencia del usuario para hacerla más fluida y satisfactoria. Desde el punto de vista operativo y financiero, conocer con antelación el método de pago puede facilitar la planificación de flujos de efectivo, anticipar comisiones por uso de plataformas externas (como PayPal) y mejorar la negociación con proveedores de servicios de pago. Por otra parte, al incluir en el análisis variables como el uso de descuentos, preferencias de envío e intención de compra, se amplía la perspectiva más allá de la simple predicción, abarcando aspectos contables y estratégicos. Esta justificación parte de la premisa de que una empresa con conocimiento profundo de sus clientes y de su comportamiento puede tomar mejores decisiones, incrementar la fidelidad de sus usuarios y optimizar cada etapa del proceso comercial.

# **Antecedentes**

En los últimos años, el análisis de datos aplicado al comercio electrónico se ha enfocado principalmente en predecir si un usuario comprará o no, cuánto gastará, qué producto elegirá o si abandonará su carrito. Sin embargo, pocas investigaciones han abordado de manera específica la predicción del método de pago, a pesar de que este aspecto tiene consecuencias directas en la conversión de ventas, la rentabilidad y la estructura de costos operativos. Investigaciones relacionadas han utilizado técnicas como regresión logística para clasificación binaria, redes neuronales profundas para detectar patrones complejos y árboles de decisión para interpretaciones simples y visuales. En este trabajo, se seleccionó el modelo Random Forest por su equilibrio entre precisión, interpretabilidad y robustez frente a datos ruidosos o con múltiples variables categóricas. Este modelo, basado en ensamblado de árboles, permite obtener una visión global del comportamiento de los usuarios y de los factores que más influyen en su decisión de pago. Además, se complementa con un enfoque exploratorio que permite detectar relaciones no evidentes en los datos antes de aplicar el modelado predictivo.

# **Definición del problema**

El objetivo central del análisis es construir un modelo predictivo capaz de anticipar qué método de pago utilizará un cliente en una transacción digital. Este tipo de predicción se apoya en un conjunto de variables observadas en el comportamiento histórico de usuarios, que permiten inferir sus preferencias futuras. Las variables utilizadas abarcan desde datos personales (edad, género, estado civil, nivel de educación e ingresos) hasta aspectos más dinámicos como el tipo de dispositivo usado, la categoría del producto, la intención detrás de la compra (necesidad, deseo o impulso), si se utilizó o no un descuento, y las preferencias de envío seleccionadas. A partir del cruce de esta información se busca identificar patrones consistentes que permitan establecer correlaciones claras entre el perfil del usuario y el medio de pago utilizado. El objetivo final es que esta predicción pueda integrarse en sistemas de recomendación o de optimización interna para anticiparse al comportamiento del consumidor y mejorar procesos logísticos y financieros.

# **Análisis predictivo**

1. Determinación de la base de datos:

La base de datos utilizada en este análisis corresponde a un conjunto de datos sintético de comportamiento de consumidores en comercio electrónico. Este dataset recopila información detallada sobre miles de transacciones realizadas por usuarios con características demográficas y conductuales diversas. Cada observación representa una compra individual e incluye variables como edad del cliente, género, nivel de ingresos, nivel educativo, estado civil, categoría del producto adquirido, monto de la compra, dispositivo utilizado para realizar la transacción, intención de compra, tipo de envío seleccionado, método de pago utilizado, entre otros.

Este conjunto de datos es altamente adecuado para realizar un análisis predictivo, ya que contiene tanto variables numéricas como categóricas que pueden correlacionarse con el método de pago. Además, su estructura permite aplicar técnicas de minería de datos y machine learning sin requerir una transformación excesiva. La información fue tratada y manipulada usando el lenguaje de programación R, el cual permite integrar tareas de limpieza, visualización y modelado en un solo entorno de trabajo. La elección de esta base de datos responde a la necesidad de contar con una representación realista del entorno de compra digital, permitiendo simular con precisión escenarios de decisión del consumidor y facilitando el entrenamiento de modelos predictivos con fines académicos y estratégicos.

## 

## **Limpieza de datos y preparación del dataset**

1. Pre-procesamiento y limpieza

El preprocesamiento de los datos fue una etapa clave para asegurar la calidad del análisis y la efectividad del modelo predictivo. Se inició con la inspección del dataset para identificar posibles valores faltantes, inconsistencias o formatos incorrectos. Una de las primeras tareas fue convertir la variable Purchase\_Amount, que originalmente se encontraba en formato de texto con símbolo monetario, a formato numérico. Para ello, se eliminaron los signos de dólar y comas que interferían con el tratamiento estadístico, convirtiendo los valores a tipo numeric.

Posteriormente, se transformaron las fechas de compra (Time\_of\_Purchase) al formato de fecha estándar usando funciones del paquete lubridate. Esta transformación permitió realizar análisis temporales más adelante, como la evolución de compras por mes. Asimismo, se eliminaron duplicados que podrían haber generado sesgos en el modelo y se reordenaron los datos para facilitar su manipulación.

También se aseguraron los tipos de datos apropiados para cada columna. Las variables categóricas como Gender, Income\_Level, Education\_Level, Payment\_Method, entre otras, fueron convertidas a factores (factor) para que el modelo de Random Forest pudiera procesarlas correctamente. Además, se convirtieron las columnas booleanas como Discount\_Used y Customer\_Loyalty\_Program\_Member a tipo lógico (TRUE/FALSE).

Como parte del preprocesamiento se descartaron columnas irrelevantes para la predicción, como Customer\_ID y Location, ya que no aportaban información útil al modelo y podían generar ruido. También se realizó una imputación simple de valores faltantes donde fue necesario, aunque en general el dataset presentaba una calidad aceptable.

Esta fase de preparación de datos sentó las bases para un análisis confiable, eliminando errores, homogenizando formatos y asegurando que todas las variables estuvieran en condiciones de ser utilizadas tanto para el análisis descriptivo como para el modelado predictivo.

1. Análisis descriptivo

Con los datos ya limpios y transformados, se realizó un análisis descriptivo con el objetivo de entender la distribución y comportamiento general de las variables clave. Se calcularon estadísticas como la media, mediana y desviación estándar del monto de compra (Purchase\_Amount), identificando un valor medio alrededor de $286.94, una mediana de $270.26 y una desviación estándar de $159.39. Esto indica una distribución moderadamente dispersa con presencia de valores extremos, los cuales podrían representar compras de alto valor en ciertas categorías.

Se utilizó la librería ggplot2 para representar visualmente la información. Un histograma permitió observar la concentración de compras en rangos bajos y medios, mientras que una gráfica de barras evidenció que los métodos de pago más utilizados fueron tarjeta de crédito, débito y PayPal. A través de gráficos segmentados, se analizaron también variables como género, nivel de ingresos e intención de compra, mostrando diferencias importantes en el comportamiento de pago según el perfil del usuario.

Adicionalmente, se exploraron las relaciones entre variables como el uso de descuentos y el método de pago, el tipo de envío elegido, y la categoría del producto comprado. Estos análisis permitieron identificar patrones que luego serían útiles para alimentar el modelo predictivo. Por ejemplo, se observó que los usuarios que utilizan PayPal tienden a comprar en categorías como tecnología y entretenimiento, mientras que los pagos con débito fueron más comunes en compras de bajo monto.

Este análisis descriptivo no solo sirvió como una aproximación inicial al comportamiento de los datos, sino que también permitió seleccionar las variables con mayor poder explicativo para la fase de modelado. Además, ofreció un panorama visual y estadístico de cómo interactúan los factores demográficos y conductuales en el proceso de compra online.

d. Selección de variables

La selección de variables es una etapa fundamental dentro del proceso de modelado predictivo, ya que permite identificar qué características del dataset aportan valor al momento de predecir la variable objetivo: el método de pago (Payment\_Method). A partir del análisis descriptivo y la exploración gráfica, se identificaron las variables que tienen mayor potencial explicativo sobre la decisión del usuario.

Entre las variables seleccionadas se encuentran datos demográficos como Age, Gender, Income\_Level, Education\_Level y Marital\_Status, ya que estos atributos suelen estar directamente relacionados con hábitos de consumo y capacidad financiera. A nivel transaccional, se incorporaron variables como Purchase\_Amount, Purchase\_Category, Device\_Used\_for\_Shopping, Shipping\_Preference, Time\_Spent\_on\_Site, Pages\_Viewed, Product\_Viewed, Purchase\_Intention, Discount\_Used y Customer\_Loyalty\_Program\_Member, por su vínculo directo con el proceso de compra.

Se descartaron variables como Customer\_ID y Location, ya que no aportaban información predictiva útil y podían introducir ruido o generar sobreajuste en el modelo. La variable Time\_of\_Purchase fue excluida en la versión base del modelo, aunque se mantuvo para análisis de series de tiempo.

La relevancia de las variables seleccionadas fue posteriormente validada mediante el análisis de importancia de variables proporcionado por el modelo Random Forest, el cual confirmó que atributos como Income\_Level, Device\_Used\_for\_Shopping, Purchase\_Category y Discount\_Used son altamente influyentes en la predicción del método de pago. Esta selección cuidadosa de variables permitió mejorar la eficiencia del modelo y garantizar que sus predicciones se basen en patrones consistentes y explicables dentro del comportamiento del consumidor.

1. Selección de modelo

Para resolver el problema planteado de predecir el método de pago que utilizará un cliente en una transacción de comercio electrónico, se seleccionó el modelo de clasificación Random Forest. Este modelo pertenece a la familia de algoritmos de aprendizaje supervisado y se caracteriza por su robustez, capacidad de manejar variables categóricas y numéricas, y su rendimiento confiable en problemas de clasificación multiclase como el presente.

Random Forest opera construyendo múltiples árboles de decisión sobre subconjuntos aleatorios del conjunto de entrenamiento, y combinando sus predicciones para obtener una respuesta más precisa y estable. Su mecanismo de votación o promedio permite mitigar el sobreajuste que puede ocurrir en modelos de árbol individuales. Además, una de sus ventajas más relevantes es la capacidad de medir la importancia relativa de cada variable predictora en la construcción del modelo, lo cual se aprovechó en este análisis para validar la selección de atributos.

El modelo fue entrenado utilizando el 80% del dataset limpio y preprocesado, reservando el 20% restante para evaluar su desempeño. La variable objetivo fue Payment\_Method, mientras que las variables predictoras incluyeron datos demográficos, comportamentales y transaccionales seleccionados en el apartado anterior. Se utilizaron 200 árboles (ntree = 200) como configuración inicial, y se habilitó la medición de importancia de variables (importance = TRUE).

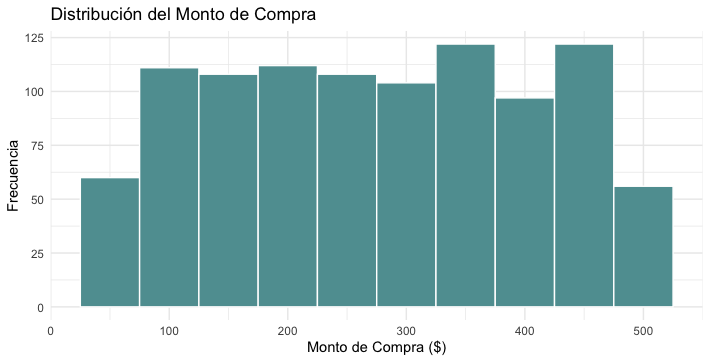
La evaluación del modelo se realizó a través de una matriz de confusión, que permitió analizar el nivel de acierto por clase. Los resultados indicaron una alta precisión en la predicción de métodos de pago frecuentes como tarjeta de crédito y PayPal, mientras que los métodos menos comunes presentaron una tasa de acierto ligeramente inferior, lo cual es esperable en problemas de clasificación desbalanceada.

Además, se graficó la importancia de las variables para visualizar cuáles fueron los atributos más influyentes en la predicción. Este análisis confirmó que variables como Income\_Level, Device\_Used\_for\_Shopping, Purchase\_Category y Discount\_Used fueron determinantes para predecir el comportamiento de pago del usuario. Estas visualizaciones fueron realizadas en R utilizando la librería randomForest junto con funciones complementarias de ggplot2 y caret, lo cual permitió integrar el modelado y la interpretación de resultados de forma visual e intuitiva.

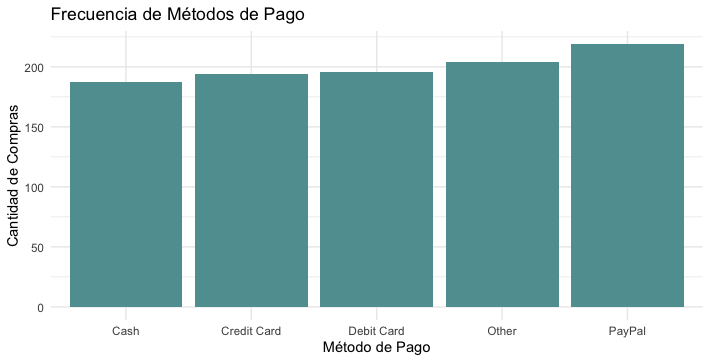
# **Gráficas de visualización**

## **Exploración general del comportamiento de compra**

Se generó un histograma para analizar la distribución del monto de compra entre los usuarios, revelando una concentración significativa en rangos medios, con algunos valores atípicos hacia extremos superiores.

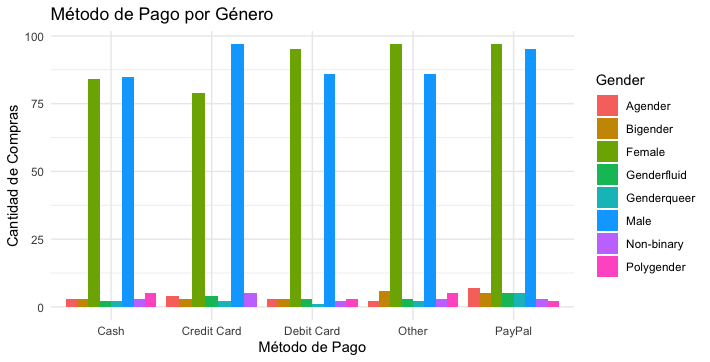


Se elaboró una gráfica de barras que muestra la frecuencia de uso de cada método de pago, permitiendo observar cuál es el medio preferido por los consumidores.

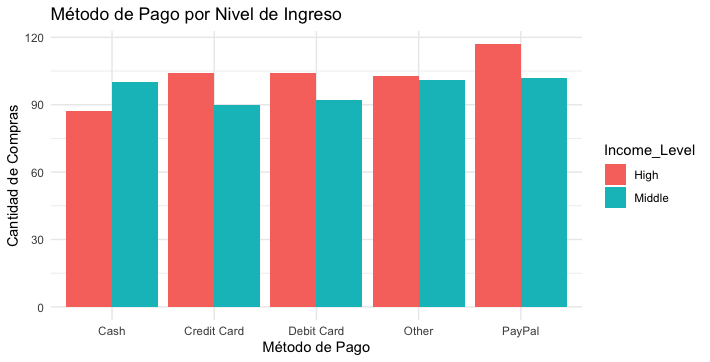


## **Segmentación por características del cliente**

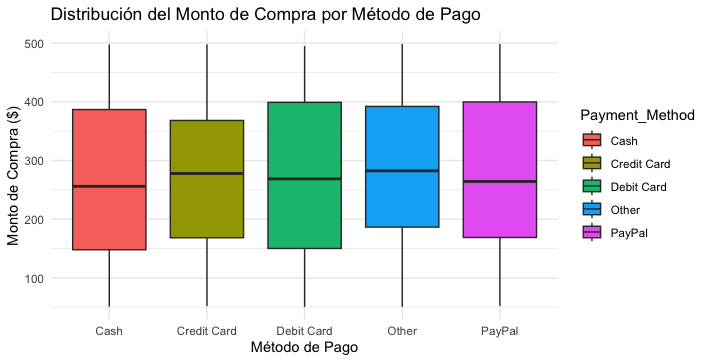
Se realizaron múltiples visualizaciones segmentadas para entender mejor el comportamiento de los usuarios según características específicas. Se comparó el método de pago elegido por hombres y mujeres, mostrando diferencias leves pero relevantes.



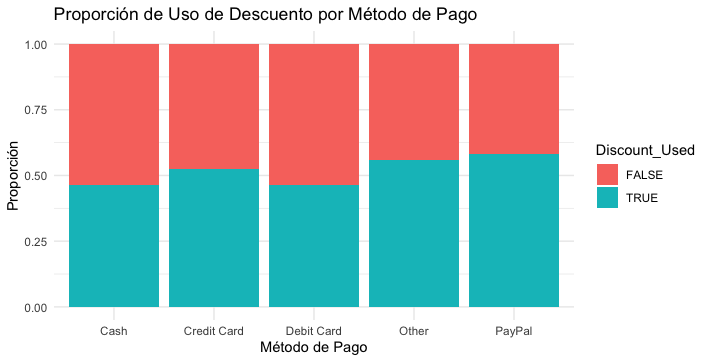
Se graficó la relación entre nivel de ingresos y el tipo de pago, evidenciando que usuarios con mayores ingresos tienden a usar con más frecuencia tarjetas de crédito.



El boxplot permitió observar la dispersión del monto de compra según el método de pago, destacando diferencias sustanciales entre medios.



Finalmente, se graficó el uso de descuentos por método de pago, revelando que los usuarios que utilizan PayPal y débito aprovechan más las promociones.



## **Visualización contable**

Desde una perspectiva financiera, se construyeron gráficos que muestran los ingresos totales por categoría de producto, identificando cuáles son los rubros más rentables.