GRADO EN ECONOMÍA

CURSO ACADÉMICO 2024-2025

TRABAJO DE FIN DE GRADO

Business Analytics: Definición, evolución y aplicación de técnicas de análisis en una empresa de venta de hardware:

Un enfoque práctico mediante árboles de decisión, Random Forest y clustering para optimización comercial

Business Analytics: Definition, Evolution, and Application of Predictive Techniques in a Hardware Retail Company

A Practical Approach Using Decision Trees, Random Forest, and Clustering for Business Optimization

AUTOR: LUCÍA URIA YABAR

DIRECTORA: ELIANA ROCÍO ROCHA BLANCO

CONVOCATORIA DE DEFENSA: JUNIO,2025

DECLARACIÓN RESPONSABLE

La persona que ha elaborado el TFG que se presenta es la única responsable de su contenido. La Universidad de Cantabria, así como quien ha ejercido su dirección, no son responsables del contenido último de este Trabajo.

En tal sentido, Don/Doña LUCÍA URIA YABAr *se hace responsable:*

1. *De la AUTORÍA Y ORIGINALIDAD del trabajo que se presenta.*
2. *De que los DATOS y PUBLICACIONES en los que se basa la información contenida en el trabajo, o que han tenido una influencia relevante en el mismo, han sido citados en el texto y en la lista de referencias bibliográficas.*

Asimismo, declara que el Trabajo Fin de Grado tiene una extensión de máximo 10.000 palabras, excluidas tablas, cuadros, gráficos, bibliografía y anexos.

Fdo.:

RESUMEN

Este trabajo aplicó técnicas de Business Analytics a datos simulados de una empresa B2B de hardware, analizando patrones de compra, segmentando clientes y prediciendo el abandono. Se observó una mayor frecuencia de compra en productos más económicos. Los modelos predictivos lograron precisiones moderadas (53 –59%), siendo útiles a pesar del desbalance de clases. La segmentación de clientes reveló perfiles diferenciados, como el Clúster 1, que son las empresas grandes y propensas a productos premium pero con mayor abandono. El precio fue el factor más influyente en el abandono, aportando *insights* valiosos para adaptar estrategias por sector y país.

ABSTRACT

This study uses Business Analytics on simulated B2B hardware data to analyze purchase behavior, customer segments, and churn. Lower-priced products had higher purchase frequency. Predictive models achieved moderate accuracy (53–59%) despite class imbalance. Clustering identified key profiles, such as Cluster 1—large firms ideal for premium products but with higher churn risk. Price emerged as the main churn factor, informing strategic decisions by sector and region.

1. INTRODUCCIÓN

En la actual era del dato, la competitividad de las empresas depende en gran medida de su capacidad para interpretar y aprovechar la información disponible. El volumen de datos crece de forma exponencial, siguiendo patrones como los descritos por la Ley de Moore, lo que exige el uso de herramientas analíticas robustas que permitan transformar esos datos en valor estratégico (Asllani, 2014). No obstante, a pesar del creciente interés en estas tecnologías, persiste una brecha significativa entre la demanda de profesionales especializados en *Business Intelligence* (BI) y *Business Analytics* (BA) y la oferta existente, lo cual limita su adopción generalizada en muchas organizaciones (Salazar y Kunc, 2025).

El concepto de BI fue introducido por Luhn en 1958 y más tarde popularizado por Dresner. Desde entonces, ha evolucionado hacia el BA, incorporando técnicas estadísticas y modelos predictivos que permiten tomar decisiones más proactivas (Davenport, 2007). Actualmente, estas herramientas se aplican ampliamente en sectores como la salud, la manufactura o el comercio minorista. En este contexto, el *prescriptive analytics*, considerado la fase más avanzada del análisis, se utiliza para optimizar desde tratamientos médicos personalizados hasta rutas logísticas en la cadena de suministro (Moesmann y Pedersen, 2024). La reciente incorporación de la inteligencia artificial generativa (GenAI) ha revolucionado aún más este campo, automatizando la construcción de modelos y generando recomendaciones precisas de manera autónoma (Salazar et al., 2025).

En el caso de España, aunque muchas pequeñas y medianas empresas (PYMES) han comenzado a adoptar soluciones de BI, aún enfrentan importantes desafíos a la hora de adaptar los modelos de madurez analítica a sus recursos limitados, tanto económicos como tecnológicos (González-Varona et al., 2024).

Este trabajo aplica de forma práctica los conceptos de BA en una empresa ficticia dedicada a la venta de hardware informático. Utilizando herramientas analíticas desarrolladas en Python (a través de la plataforma de Google Colab), se emplean técnicas como árboles de decisión, *Random Forest* y *clustering*. Los objetivos del análisis incluyen técnicas supervisadas y no supervisadas, la predicción de compras de productos de gama alta, la segmentación de clientes y el agrupamiento de productos con patrones de comportamiento similares.

Los resultados obtenidos muestran cómo el uso del BA, incluso en entornos simulados, puede contribuir a optimizar la toma de decisiones estratégicas. Esto subraya el potencial de estas herramientas para mejorar el rendimiento empresarial, facilitando una gestión más eficiente y basada en datos objetivos.

2. CONCEPTOS FUNDAMENTADELS DE BUSINESS ANALYTICS

2.1. DEFINICIÓN Y EVOLUCIÓN DE BUSINESS ANALYTICS

El concepto de Business Analytics (BA) ha ido evolucionando a lo largo de las últimas décadas, complicando la tarea de dar una definición precisa y universal. Power et al. (2018) destacan que no existe un consenso claro sobre su significado, ya que diferentes disciplinas y sectores han adoptado el concepto con matices distintos. Según ellos, BA es un término emergente con raíces en estadística, informática y gestión empresarial, lo que ha llevado a una proliferación de definiciones que varían en alcance, detalle y aplicabilidad.

En el artículo *Defining business analytics: an empirical approach* (Power et al., 2018), se identifica un problema recurrente en la literatura: la ambigüedad en torno a la relación de BA con términos afines como *Business Intelligence (BI)* y *Data Analytics (DA)*. A través del análisis de tendencias y definiciones recopiladas de diversas fuentes, los autores concluyen que BA es una disciplina en evolución que combina métodos analíticos avanzados con aplicaciones empresariales orientadas a la toma de decisiones. Es decir, se trata del conjunto de métodos, tecnologías y procesos utilizados para analizar datos empresariales y transformarlos con el objetivo de mejorar la toma de decisiones.

Para explicar la distinción entre Business Analytics (BA) y Business Intelligence (BI), es necesario remontarse a 1958, cuando Hans Peter Luhn utilizó por primera vez el término *Business Intelligence* en la publicación de su artículo *A Business Intelligence System* (Luhn,1958). En este, definía el concepto como “la capacidad de aprehender las interrelaciones de los hechos presentados de tal manera que guíen la acción hacia un objetivo predeterminado”.

Más tarde, en 1989, Howard Dresner describió BI como un término general que abarca conceptos y métodos destinados a mejorar la toma de decisiones comerciales mediante el uso de sistemas de soporte basados en hechos (Klimberg et al, 2010).

Con la llegada del nuevo milenio, el avance de las tecnologías y el incremento en la capacidad de almacenamiento y procesamiento de datos, surgió el concepto de *BA*, el cual utiliza *BI* como base para la realización de cálculos estadísticos e inferencias. BI se enfoca en el

análisis descriptivo y en la generación de informes a partir de datos históricos (analítica reactiva), su propósito principal es ofrecer un resumen claro y estructurado de la información histórica y actual de la empresa, con el fin de comprender qué ha ocurrido y qué está ocurriendo en el negocio. BA emplea modelos predictivos y prescriptivos con el propósito de anticipar eventos futuros y facilitar la toma de decisiones óptima (analítica proactiva). (Tableau ,s.f; Olsys Careers ,2025; BARC,2024)

Business Intelligence	Business Analytics
Analítica Reactiva	Analítica Proactiva
Responde a las siguientes preguntas:	
• ¿qué sucedió? • ¿cuándo? • ¿quién? • ¿cuántos?	• ¿por qué sucedió? • ¿ocurrirá otra vez? • ¿qué ocurriría si...? • ¿qué otras cosas dicen los datos que nunca se nos ocurrió preguntar?

Tabla 1. Diferencia entre BI y BA. Fuente: Cano, I.R. (2018).

En la tabla 1 se ejemplifica la diferencia entre BI y BA a través de una serie de preguntas, diferenciándolos en función de su enfoque y tipo de análisis. Como se mencionó anteriormente, *BI* se asocia con una analítica reactiva, centrada en la descripción de eventos pasados a través de preguntas como qué sucedió, cuándo, quién estuvo involucrado y cuántos eventos ocurrieron. En cambio, *BA* se orienta hacia una *Analítica Proactiva*, buscando no solo explicar las causas de los eventos, sino también predecir su recurrencia y explorar escenarios hipotéticos a partir de los datos. Mientras que *BI* permite comprender el pasado con base en información histórica, *BA* va más allá al utilizar modelos predictivos y prescriptivos para anticipar tendencias y mejorar la toma de decisiones estratégicas.

Los tipos de análisis dentro de *BA* pueden clasificarse en tres categorías, de acuerdo con sus objetivos o usos: descriptivo, predictivo y prescriptivo (Liberatore & Luo, 2011).

El *análisis descriptivo* se basa en el uso de herramientas para examinar datos y comprender el rendimiento empresarial, aplicando técnicas para categorizar, caracterizar y clasificar la información de manera útil. Sus resultados suelen representarse en gráficos e informes que permiten identificar patrones y tendencias. Por esta razón, las preguntas que aborda la analítica descriptiva están relacionadas con el rendimiento actual y pasado, como qué, cuánto y cuál.

El *análisis predictivo* se enfoca en anticipar el rendimiento futuro mediante el examen de datos históricos, la identificación de relaciones entre ellos y su posterior extrapolación. Sus principales resultados incluyen la detección de patrones ocultos en grandes volúmenes de datos, lo que facilita la predicción de comportamientos y tendencias. Las preguntas típicas que responde están vinculadas al análisis de sensibilidad de la situación actual, como “¿qué pasaría si…?”.

El *análisis prescriptivo* emplea herramientas como la optimización y la simulación para determinar las mejores alternativas con el fin de alcanzar los objetivos empresariales. Además, incorpora técnicas matemáticas y estadísticas utilizadas en la analítica predictiva, combinándolas con métodos de optimización para gestionar la incertidumbre en los datos. Las preguntas abordadas en este tipo de análisis están relacionadas con el impacto de los cambios operacionales en el rendimiento futuro.

En la tabla 2, basada en Bordawekar, R., Blainey, B. y Apte, C. (2014) muestra ejemplos de aplicaciones de BA clasificadas según el tipo de análisis mencionados anteriormente. Asimismo, cada categoría se asocia con herramientas específicas y objetivos funcionales, desde la fijación de precios personalizados hasta la reducción del desabastecimiento de inventario. Además, se detalla el tipo de problema abordado, como el análisis de datos estructurados, la predicción de ingresos o el modelado de riesgos.

Aplicación analítica	**Herramienta analítica**	**Clasificación**	**Objetivo funcional**	**Tipos de problema**
Gestión de la cadena de suministro, es decir, programación de productos, enrutamiento	Prescriptiva	Prescripción	Lograr eficiencia en la cadena de suministro mejorando el servicio y reduciendo costos	Optimización
Predicción de ingresos	Predictiva	Predicción	Reducción de desabastecimientos de inventario y mejora en la gestión del flujo de efectivo	Aprendizaje/Estadística descriptiva e inferencial
Análisis de ventas, informes financieros y presupuestación	Descriptiva	Reporte	Análisis de datos estructurados / no estructurados	Análisis financiero trimestral y presupuestación
Modelado de riesgos en seguros y crédito	Prescriptiva	Simulación	Reducción de pérdidas por fraude y fijación de precios personalizados	Modelado y Simulación / Estadística Inferencial

Tabla 2. Ejemplos de aplicaciones de analítica de negocios, objetivos funcionales y tipos de problemas analíticos. Fuente: Bordawekar, R., Blainey, B. y Apte, C. (2014).

3. HERRAMIENTAS Y TECNOLOCÍAS EN BUSINESS ANALYTICS

3.1 BIG DATA Y SU RELACIÓN CON BUSINESS ANALYTICS

A finales de la década de los 80, la tecnología de almacenamiento de datos o *data warehouse* en inglés, consistía en almacenar grandes volúmenes de datos fuera de las bases de datos de producción para mantener la agilidad y rendimiento. Para ello, múltiples copias de los datos se ubicaban en varios servidores de bases de datos conocidos como "*data marts*", que podían ser independientes o estar integrados en un almacén de datos corporativo.

A pesar de que estos data warehouse hayan creado valor para las empresas (Davenport et. al, 2014) el traslado constante de datos por la red resultaba tedioso y los resultados se obtenían transcurrido un largo periodo de tiempo. Además, el volumen de datos que se puede almacenar es limitado. Actualmente, a esto hay que añadir que la información se genera continuamente, dando lugar a lo que hoy en día conocemos como *Big Data*.

El *Big Data* ha logrado llamar la atención de muchos diversos sectores: el gobierno, las finanzas, ingenierías, la salud, y mayormente, en el ámbito empresarial. Los datos generados en este contexto se caracterizan por su incapacidad para ser categorizados por bases de datos relacionales tradicionales, por su volumen, y sobre todo, por ser generados, capturados y procesados de manera muy rápida. Por ello, uno de los desafíos más relevantes que enfrentan las organizaciones actualmente, es el de desarrollar métodos apropiados que permitan manejar y analizar grandes volúmenes de datos, con el fin de optimizar la toma de decisiones. Por tanto, al aplicar las técnicas de BA en Big Data, se obtienen numerosas herramientas avanzadas que permiten visualizar y transformar los datos en información útil. De tal manera que, la analítica de Big Data o *DA* consiste en un conjunto de técnicas y recursos diseñados para gestionar grandes cantidades de datos no estructurados que los sistemas tradicionales de bases de datos no pueden manejar. En consecuencia, las soluciones de *DA* permiten a las organizaciones a detectar cambios e innovar en tiempo real.

3.2 TÉCNICAS DE ANÁLISIS DE DATOS EN BA

Los procesos ETL (*Extract, Transform, Load*) son fundamentales para integrar datos de múltiples sistemas, siendo especialmente relevantes en la creación de almacenes de datos, donde representan una de las fases más costosas debido a la complejidad de unificar información heterogénea (Duque Méndez et al., 2016). A pesar de la disponibilidad de herramientas ETL, muchas no permiten la personalización necesaria para contextos complejos.

Paralelamente, el preprocesamiento de datos se considera una etapa crítica en todo proyecto de ciencia de datos,ya que mejora la calidad del conjunto inicial y garantiza resultados más precisos (García et al., 2016). Incluye tareas como limpieza, integración y normalización, así como la reducción de datos para simplificar sin perder información clave.

Por último, la visualización de datos permite representar la información de forma gráfica, facilitando su interpretación y apoyando la toma de decisiones de manera intuitiva (Therón Sánchez, R., 2021).

3.2.1 Técnicas de Aprendizaje Supervisado: Algoritmos de clasificación más utilizados

El aprendizaje automático (*Machine Learning*) tiene como objetivo dotar a los sistemas informáticos de la capacidad para resolver problemas a partir de datos o experiencias previas. Una de las metodologías más comunes es el aprendizaje supervisado (*Supervised Machine Learning, SML*), que requiere conjuntos de datos etiquetados por expertos, lo que permite que el modelo aprenda patrones entre las variables de entrada y una salida conocida para luego hacer predicciones con nuevos datos (Shetty et al., 2022).

Dentro del aprendizaje supervisado se distinguen dos enfoques principales: la clasificación, cuando la variable objetivo es categórica, y la regresión, cuando se trata de valores numéricos continuos (Ajah & Nweke, 2019). La regresión es especialmente útil para analizar relaciones entre variables y prever tendencias futuras.

Una técnica destacada de clasificación son los árboles de decisión, modelos no paramétricos que dividen los datos en subconjuntos mediante reglas basadas en los valores de las variables, lo que permite llegar a una predicción tanto en problemas de clasificación como de regresión (Arana, 2021). Una extensión más robusta de este modelo es el algoritmo *Random Forest*, que construye múltiples árboles con subconjuntos aleatorios del mismo conjunto de datos y realiza una votación entre ellos para generar la predicción final, mejorando la precisión y reduciendo el sobreajuste (Breiman, 2001).

Otro algoritmo de aprendizaje supervisado ampliamente utilizado en clasificación y regresión es el modelo Máquina de Soporte Vectorial (Support Vector Machine - SVM), cuyo objetivo es encontrar el mejor límite para separar dos clases, utilizando funciones kernel —como el Gaussiano— que permiten transformar los datos a espacios de mayor dimensión para una mejor separación. Este modelo se basa en los llamados vectores de soporte para definir sus límites de decisión (Betancourt, G. A., 2005).

Finalmente, las redes neuronales son algoritmos inspirados en el funcionamiento del cerebro humano, compuestos por nodos interconectados que ajustan sus pesos durante el entrenamiento para asociar entradas y salidas. Este aprendizaje puede realizarse por lotes o de forma incremental y destaca por su capacidad para abordar tareas complejas como la clasificación de imágenes, reconocimiento de voz o predicción de series temporales (Hagan, 2014).

3.2.2 Algunos ejemplos de aplicación empresarial: segmentación de clientes, predicción de abandono y detección de fraude

Diversos estudios han demostrado la eficacia de los modelos de árboles de decisión y *Random Forest* en diferentes sectores. Camacho, Ramallo y Ruiz (2018) aplicaron árboles de decisión para identificar los factores que influyen en el precio de la vivienda en Madrid, encontrando que la superficie es el más determinante, seguido por el número de baños y habitaciones.
En el ámbito financiero, el estudio *Comparison of Random Forest*, *Logistic Regression* and *Multilayer Perceptron Methods on Classification* en la clasificación del cierre de cuentas de clientes bancario (Khoirunissa et. al, 2021) mostró que Random Forest obtuvo la mejor precisión (91 %) frente al perceptrón multicapa y la regresión logística. Zhao (2023) también utilizó *Random Forest* y árboles de decisión para predecir el abandono de clientes bancarios, confirmando nuevamente la alta eficacia del primero.

En telecomunicaciones, Thorat y V.R.S. (2022) aplicaron *Random Forest* para predecir la pérdida de clientes, empleando ingeniería de características y optimización de hiperparámetros, logrando alta precisión y utilidad práctica en entornos reales.

Finalmente, en el comercio electrónico, Zhai (2025) desarrolló un sistema de detección automática de fraude basado en *Random Forest*, que alcanzó una precisión del 98,76 % y un

recall del 97,53 %, superando otros modelos y demostrando su idoneidad ante problemas complejos y con datos desbalanceados.

3.2.3 Técnicas de Aprendizaje No Supervisado: Definición y diferencias con el aprendizaje supervisado.

El aprendizaje no supervisado permite identificar patrones en datos no etiquetados, sin intervención humana (IBM, 2021). Una técnica clave es el *clustering*, que agrupa datos similares y es útil para segmentar clientes o detectar comportamientos comunes. K-Means es uno de los algoritmos más populares en este campo (Shobayo & Ogunleye, 2023).

John, Shobayo y Ogunleye (2023) analizaron datos de un minorista online usando RFM[1] (Recency Frequency y Monetary value por sus siglas en inglés) y varios algoritmos de clustering. El modelo GMM fue el más eficaz para diferenciar perfiles de clientes. En el sector bancario, Barkhordar (2021) propuso un enfoque híbrido con redes LSTM (Long Short-Time Memory)[2] y *dynamic time warping* seguido de K-Means, logrando una segmentación precisa basada en patrones de transacción.

También destaca el *clustering jerárquico*, que permite explorar estructuras complejas mediante enfoques aglomerativos o divisivos, siendo el primero el más utilizado (IBM, 2021).

Las reglas de asociación, como las del algoritmo Apriori, permiten encontrar patrones frecuentes en datos transaccionales, aplicándose en análisis de cesta de mercado o recomendaciones (IBM, 2021). Por último, el Análisis de Componentes Principales (PCA) reduce la dimensionalidad de los datos manteniendo la información relevante, lo que mejora la eficiencia analítica (IBM, 2021).

3.3 HERRAMIENTAS

Los lenguajes de programación son esenciales para aplicar técnicas de *Data Analytics* y *Business Intelligence* (BI), ya que permiten manipular, analizar y visualizar grandes volúmenes de datos con eficiencia. Entre los más utilizados se encuentran Python y R, debido a su versatilidad y a la amplia disponibilidad de bibliotecas especializadas que facilitan el análisis de datos.

Python destaca por su sencillez y legibilidad, lo que lo hace especialmente popular en entornos científicos, educativos y de software libre. Su sintaxis de alto nivel y sus estructuras integradas, como listas y diccionarios, permiten resolver tareas complejas de forma clara y concisa. Además, es un lenguaje fácil de aprender, eficiente y compatible con múltiples plataformas. Incluso sin una base sólida en programación, los analistas pueden usar Python

[1] El modelo RFM (Recency, Frequency, Monetary) es una metodología de segmentación de clientes que evalúa tres dimensiones clave del comportamiento de compra: la recencia (tiempo desde la última compra), la frecuencia (número de compras realizadas) y el valor monetario (cantidad total gastada). A cada cliente se le asigna una puntuación en cada dimensión, lo que permite identificar a los más valiosos para la empresa. (Murphy, C. ,2024 What is recency, frequency, monetary value (RFM) in marketing?, Investopedia)

[2] Las redes LSTM (Long Short-Term Memory) son un tipo de red neuronal recurrente capaz de conservar información a largo plazo mediante una estructura de memoria interna, lo que las hace eficaces para el procesamiento de datos secuenciales (MathWorks)

para escribir scripts que realicen tareas de limpieza, manipulación y visualización de datos. Otro punto a favor es su amplia comunidad de desarrolladores, que ofrece gran variedad de bibliotecas, herramientas y apoyo a través de foros en línea (Holguín Londoño, Moreno Gallón y Holguín Londoño, 2024).

En el ámbito empresarial, Python resulta especialmente atractivo, ya que permite realizar procesos como la limpieza de datos, la selección y extracción de características relevantes, la asignación de etiquetas para facilitar la interpretación, el cálculo de métricas estadísticas, y la representación visual mediante tablas y gráficos, como histogramas o de barras (https://aws.amazon.com/es/what-is/python/).

R, por su parte, también es un lenguaje de alto nivel, orientado específicamente al análisis estadístico. Aunque su sintaxis requiere mayor precisión —ya que distingue entre mayúsculas, símbolos y números—, es igualmente potente para la gestión y análisis de datos. Las diferencias entre Python y R suelen centrarse en aspectos como la visualización o la forma de gestionar los datos, pero en general son mínimas.

En el análisis de datos, se emplean diversos formatos de archivos, cada uno con características y usos específicos. Los archivos CSV (Comma Separated Values) son ampliamente utilizados para almacenar datos tabulares en texto plano, y pueden manipularse fácilmente en Python con el módulo csv o con librerías como Pandas. Los archivos Excel, que incluyen fórmulas y múltiples formatos, se pueden trabajar en Python con herramientas como openpyxl. Por su parte, los archivos XML (Extensible Markup Language), que almacenan datos en forma semiestructurada mediante etiquetas, pueden ser manipulados con el módulo xml.etree.ElementTree o también con Pandas (Holguín Londoño et al., 2024).

En este contexto, también es importante distinguir entre el software de código abierto, cuyo código fuente puede ser utilizado, modificado y distribuido libremente, y el software bajo licencia, que restringe su uso a quienes han adquirido una autorización.

Entre las herramientas más utilizadas en Business Intelligence se encuentran Microsoft Power BI, SAP Business y Oracle Analytics Cloud. Estas plataformas interpretan los lenguajes de programación para transformar los datos en visualizaciones útiles que facilitan la toma de decisiones. Cada una ofrece características distintas según el tamaño y las necesidades de la organización.

En la Figura 1 se presenta un modelo conceptual del funcionamiento de Power BI. En ella se muestra cómo esta herramienta permite conectar múltiples fuentes de datos, que luego se publican en el servicio en la nube de Power BI. A través de *Power BI Gateways*, se accede de forma segura a los datos locales de la organización. Finalmente, los usuarios pueden consultar informes y *dashboards* desde distintos dispositivos mediante el servicio en la nube.

-  *Microsoft Power BI:*

Es un software proporcionado por Microsoft que realiza transforma datos en objetos visuales para facilitar la toma de decisones. Power BI es un conjunto de aplicaciones que permite analizar datos y compartir información, que es actualizada en tiempo real y está disponible

en todos los dispositivos. Ofrece tres tipos de licencias: Fabric (gratuita), Power BI Pro y Power BI Premium por usuario (PPU).

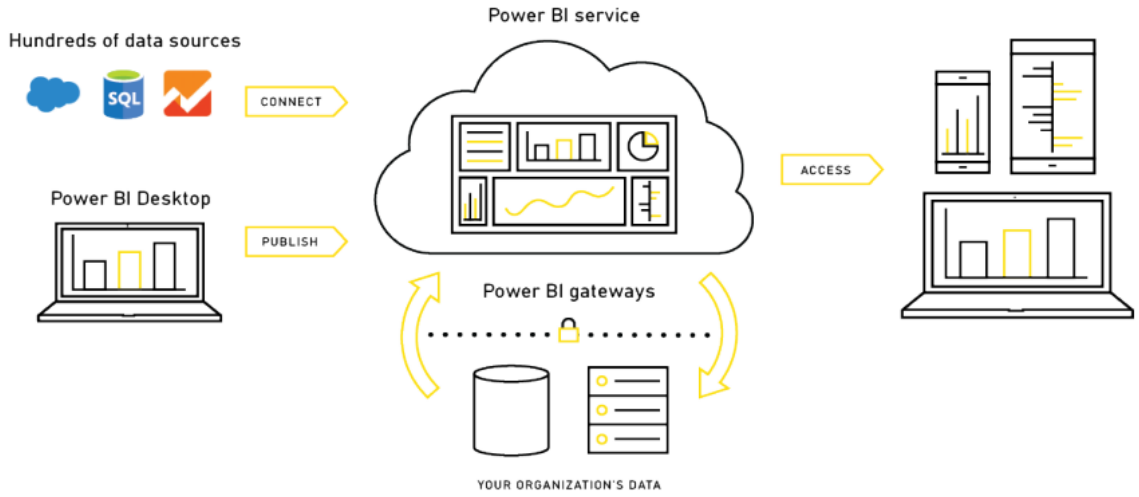


Figura 1. Modelo conceptual Power BI. Fuente. Mamani, 2018

- *SAP Business* :

SAP (*Systems, Applications, and Products in Data Processing*) es un software empresarial desarrollada por SAP SE diseñado como un ERP (Enterprise Resource Planning)[3] modular que permite a las empresas adaptar sus funciones según sus necesidades.Su usd requiere licencia propia.

- *Oracle Analytics Cloud:*

Es una plataforma bajo licencia de análisis de datos en la nube desarrollada por Oracle. Ofrece funciones de BI, análisis predictivo y visualización, integrando inteligencia artificial y aprendizaje automático. Permite trabajar con grandes volúmenes de datos en tiempo real y genera automáticamente explicaciones e insights.

En resumen, las herramientas de Business Analytics analizadas —Microsoft Power BI, SAP Business y Oracle Analytics Cloud— ofrecen enfoques distintos según el tipo de empresa. Power BI es ideal para pequeñas y medianas empresas por su facilidad de uso, bajo coste e integración con otros productos de Microsoft. SAP Business está orientado a grandes corporaciones que requieren una integración profunda con sistemas ERP, aunque implica

[3] Enterprise Resource Planning : La planificación de recursos empresariales (ERP) es un sistema de software que ayuda a las organizaciones a optimizar sus procesos de negocio centrales —incluyendo finanzas, RR. HH., fabricación, cadena de suministro, ventas y procurement— con una visión unificada de la actividad y una única fuente de verdad (SAP, s.f.)

mayor complejidad y coste. Oracle Analytics Cloud, por su parte, es adecuada para medianas y grandes empresas con infraestructura Oracle, destacando por sus capacidades avanzadas de análisis, inteligencia artificial y aprendizaje automático. Cada plataforma se adapta a diferentes necesidades según el nivel de especialización y recursos disponibles.

4. Estudio de Caso práctico de Business Analytics

4.1 APLICACIÓN PRÁCTICA DE BA

4.1.1 Objetivo

Este trabajo tiene como objetivo principal mostrar cómo se pueden aplicar técnicas de Business Analytics (BA) a un conjunto de datos simulados de una empresa ficticia dedicada a la venta de hardware informático. Para ello, se han planteado varios objetivos específicos que se desarrollan utilizando herramientas de análisis y aprendizaje automático con Python en la plataforma de Google Colab. Entre ellos destacan: analizar los clientes que son más rentables, ver qué productos se venden más según la zona, identificar a los agentes de ventas más eficaces, estudiar si el precio influye en la frecuencia de compra, predecir si un cliente comprará productos de gama alta y segmentar clientes . Con todo esto, se busca mostrar cómo BA puede servir para mejorar las decisiones, entender mejor a los clientes y optimizar los procesos de negocio.

4.1.2. Descripción del dataset

En cuanto, a los datos, se obtuvieron del área de exploración de datos de Maven Analytics,[4] una plataforma de aprendizaje y comunidad en línea centrada en el desarrollo de habilidades en analítica de datos. Para este caso práctico, se emplean datos del pipeline de ventas B2B de una empresa ficticia que vende hardware, incluyendo información sobre agentes de ventas, productos y oportunidades de ventas *(Free data sets & dataset samples, 2024)*. El conjunto de datos abarca negociaciones realizadas entre el 20 de octubre de 2016 y el 31 de diciembre de 2017. En conjunto, hay cinco ficheros en formato CSV: clientes, productos, pipeline de ventas, equipo de ventas y diccionario. En la Figura xx se muestra el número de registros y atributos de cada fichero.

[4] https://mavenanalytics.io/data-playground?order=date_added%2Cdesc&page=3&pageSize=5

```
Clientes:
  Filas: 85
  Columnas: 7
------------------------
Productos:
  Filas: 7
  Columnas: 3
------------------------
Oportunidades:
  Filas: 8800
  Columnas: 8
------------------------
Equipo de ventas:
  Filas: 35
  Columnas: 3
```

Figura 2. Resumen tablas. Fuente: Elaboración propia

Por tanto, el conjunto de datos se estructura en cuatro tablas principales: *clientes, productos, oportunidades y equipo de ventas*. La tabla de *clientes* contiene 85 filas y 7 columnas; la de *productos* incluye 7 filas y 3 columnas; la de *oportunidades* está compuesta 8.800 filas y 8 columnas; y, por último, la de *equipo de ventas* cuenta con 35 filas y 3 columnas. En este caso, *oportunidades* es la tabla central, que actúa como un registro detallado de los intentos de venta realizados por la empresa. Esta tabla se encuentra relacionada con las otras tres mediante claves externas, lo que permite realizar análisis con información adicional proveniente de los clientes, productos y agentes comerciales.

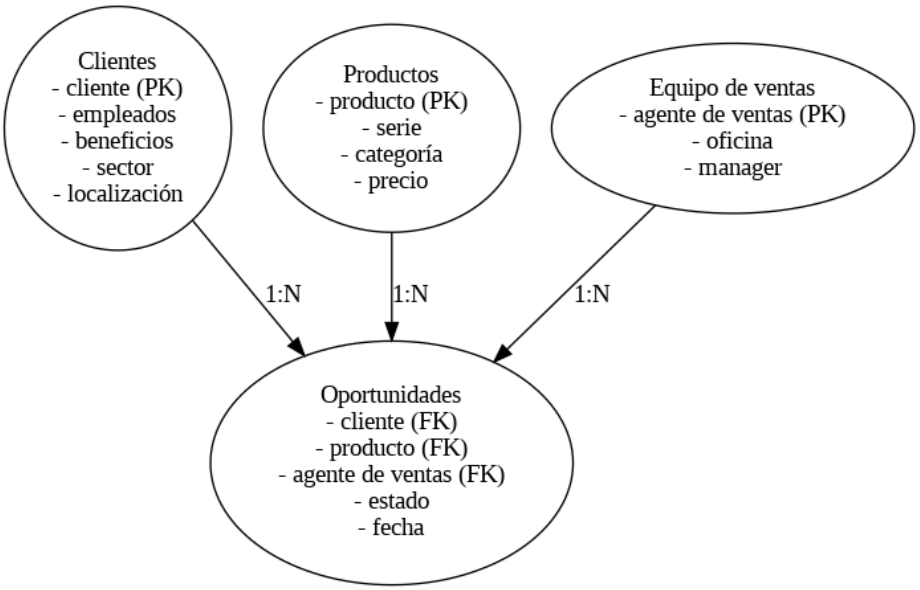


Figura 3. Diagrama ER. Fuente: Elaboración propia.

En la Figura 3 se muestra el diagrama E/R (Entidad Relación) elaborado a través de la librería de Python graphviz. Se reflejan las claves primarias y foráneas mencionadas anteriormente. Además, se establecen las relaciones de uno a muchos, indicando que un registro en las tablas de clientes, productos y equipo de ventas se puede asociar a varios registros de la tabla oportunidades. Los campos cliente, producto y agente de ventas en la tabla *"Oportunidades"* son claves foráneas que establecen relaciones con sus respectivas tablas

padre (Clientes, Productos y Agentes). Estas relaciones garantizan la integridad referencial del modelo de datos, asegurando que solo se puedan asignar a las oportunidades valores existentes en las tablas vinculadas.

4.1.3. Descripción de los atributos o variables

En la Figura 4, se muestra la descripción de la tabla "*clientes". H*ay 85 filas (clientes) y siete columnas: nombre del cliente; el sector al que pertenece; año en el que se estableció; beneficios; número de empleados; localización; y de quién es subsidiario. Las variables categóricas, como la localización o el sector pertenecen al tipo de dato object, que es el más general. La variable beneficios, se categoriza como float64 ya que recoge números de punto flotante (reales). Las variables empleados y año de establecimiento, al ser números enteros, se guardan en int64. A excepción del atributo subsidiaria tan solo tiene 15 valores no nulos, los demás atributos no cuentan con atributos nulos.

```
RangeIndex: 85 entries, 0 to 84
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   cliente              85 non-null     object
 1   sector               85 non-null     object
 2   ano_establecimiento  85 non-null     int64
 3   beneficios           85 non-null     float64
 4   empleados            85 non-null     int64
 5   localizacion         85 non-null     object
 6   subsidiaria_de       15 non-null     object
dtypes: float64(1), int64(2), object(4)
```

Figura 4. Información tabla clientes. Fuente: Elaboración propia

En la figura 5, se muestran las características de la tabla "*productos"* incluyendo los productos de la empresa vendedora, la categoría a la que pertenecen y su precio. Tanto el producto como la serie son de tipo object, mientras que el precio de venta es de tipo int. Asimismo, no se encuentran valores nulos.

```
productos.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7 entries, 0 to 6
Data columns (total 3 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   producto      7 non-null      object
 1   serie         7 non-null      object
 2   precio_venta  7 non-null      int64
dtypes: int64(1), object(2)
```

Figura 5. Información tabla productos. Fuente: Elaboración propia

En la Tabla 3 se detalla el nombre de los productos, así como la serie y sus respectivos precios.

Producto	Serie	Precio de venta
GTX Basic	GTX	550
GTX Pro	GTX	4821
MG Special	MG	55
MG Advanced	MG	3393
GTX Plus Pro	GTX	5482
GTX Plus Basic	GTX	1096
GTK 500	GTK	26768

Tabla 3. Productos empresa

En total, la empresa ofrece siete productos con distintos rangos de precios, que pueden pertenecer a cuatro series distintas. Además, dentro de cada serie encontramos el producto "*lowcost*" y uno de mayor categoría, a excepción de la serie GTK, que cuenta con un único producto de alta gama. Aquí, la clave primaria es producto, la cual se vincula a su vez con la tabla *oportunidades*.

En la tabla de "*oportunidades"* (Tabla 3) hay un total de 8800 filas y ocho columnas, de las cuales id de oportunidad es un identificador alfanumérico str (cadena de texto) que incluye ocho caracteres y combina letras y números (p.ej. 1C1I7A6R), de tal manera que sea el identificador único o *primary key.* En esta tabla solo valor de cierre no pertenece al tipo object, ya que es un número de punto flotante. En cuanto a valores nulos, clientes cuenta con 1425 valores nulos, fecha de la creación de la negociación con 500 y fecha de cierre junto con valor de cierre con 2089.

```
oportunidades.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8800 entries, 0 to 8799
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id_oportunidad  8800 non-null   object
 1   agente_ventas   8800 non-null   object
 2   producto        8800 non-null   object
 3   cliente         7375 non-null   object
 4   estado_deal     8800 non-null   object
 5   fecha_creacion  8300 non-null   object
 6   fecha_cierre    6711 non-null   object
 7   valor_cierre    6711 non-null   float64
dtypes: float64(1), object(7)
```

Figura 6. Información tabla oportunidades. Fuente: Elaboración propia

Como puede apreciarse en la Figura 7, en suma, se han vendido 4238 productos y se han perdido 2473 oportunidades. Además, hay 1589 oportunidades en *engaging, e*s decir, clientes con los que todavía se está negociando, y 500 clientes potenciales a los que aún se está intentando captar, también conocido como en fase *prospecting*. Al ser ésta la tabla núcleo que registra cada intento de venta, contiene claves foráneas que enlazan con las otras tres tablas *(cliente, producto, agente de ventas)*.

```
estado_deal
Won            4238
Lost           2473
Engaging       1589
Prospecting     500
Name: count, dtype: int64
```

Figura 7. Información variable estado_deal. Fuente: Elaboración propia

Finalmente, la tabla (figura 8), "*equipo de ventas"* se compone de 35 entradas, es decir, 35 agentes de ventas. El tipo de dato de cada atributo es el mismo ya que tanto agente de ventas como el mánager son variables nominales, mientras que la oficina en la que operan es policotómica (central, este u oeste). La oficina central cuenta con 11 trabajadores, mientras que en las otras dos hay 12.

```
equipo.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   agente_ventas  35 non-null     object
 1   manager        35 non-null     object
 2   oficina        35 non-null     object
dtypes: object(3)
```

Figura 8. Información tabla equipo de ventas. Fuente: Elaboración propia

En concreto, se puede observar en la figura 10, que hay 6 mánagers repartidos en tres oficinas (dos por oficina) En la oficina central se encuentran Dustin Brinkman y Melvin Marxen, en la Cara Losch junto con Rocco Neubert y finalmente en la oeste Celia Rouche y Summer Sewald (ver figura 10). En cuanto a los agentes de ventas, en la oficina central hay 11 agentes y tanto en las oficinas del este como en la oeste hay 12 (ver figura 11).

```
Número total de empleados por oficina:
   oficina  numero_empleados
0  Central                11
1     East                12
2     West                12
```

Figura 9. Número empleados por oficina. Fuente: Elaboración propia

```
Managers por oficina:
   oficina                              manager
0  Central  [Dustin Brinkmann, Melvin Marxen]
1     East        [Cara Losch, Rocco Neubert]
2     West      [Celia Rouche, Summer Sewald]
```

Figura 10. Mánagers por oficina. Fuente: Elaboración propia

En este caso, la clave principal es agente de ventas.

4.2 JUSTIFICACIÓN TÉCNICAS APLICADAS

Para realizar los objetivos propuestos en este análisis, se han seleccionado tres técnicas principales: clasificación: árboles de decisión (Decision Tree) y Random Forest y clustering. Cada técnica está orientada a resolver distintos tipos de problemas dentro del ámbito de Business Analytics.

En primer lugar, se ha aplicado clustering para segmentar tanto a los clientes como a los productos. Esta técnica permite agrupar entidades con características similares, lo que resulta especialmente útil para identificar perfiles de clientes o líneas de productos con comportamientos o atributos comunes. Esta segmentación facilita una mejor personalización de las acciones comerciales y una comprensión más profunda del mercado objetivo.

Por otro lado, para predecir si un cliente comprará un producto de gama alta, se emplearon Decision Tree y Random Forest.Asimismo, también se aplicó *Random Forest* para predecir el posible abandono de clientes. Estas técnicas permiten modelar de forma eficiente la relación entre múltiples variables explicativas (como ingresos, sector o historial de compras) y una variable objetivo categórica. El árbol de decisión aporta interpretabilidad y transparencia, mientras que *Random Forest* ofrece una mayor robustez y precisión, al combinar múltiples árboles para mejorar la generalización del modelo.

4.3. FASES DE UN PROYECTO DE BA

4.3.1 Pre- procesamiento de datos (ETL)

Los datos, como se ha mencionado previamente, provienen de Maven Analytics - Data Playground-, un repositorio abierto.La base de datos pertenece a la categoría de CRM, lo que simula el funcionamiento de un sistema de relación con clientes. Por tanto, aunque la fuente original real no es un sistema productivo, los datos están estructurados como si provinieran de un CRM, un tipo de sistema interno común en entornos empresariales. Para la extracción de los datos se implementó la librería *pandas* de Python. En cuanto a la frecuencia de extracción, al tratarse de un dataset estático alojado en un repositorio educativo, la frecuencia de actualización es puntual: se descarga una única vez y no se actualiza en tiempo real ni periódicamente.

4.3.2 Revisión y limpieza

A la hora de realizar la revisión y la limpieza de los datos, se encontraron los siguientes valores nulos (Ver figura 11):

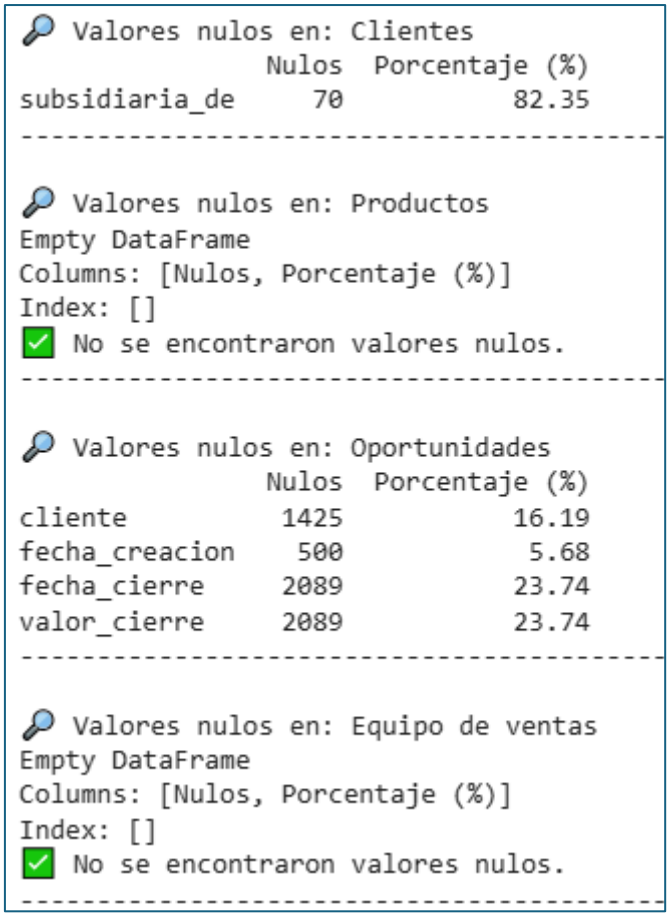


Figura 11.Valores nulos. Fuente: Elaboración propia

En resumen, en clientes se encontraron 70 valores nulos en la variable subsidiaria de, que representa el 82,35%, siendo éste el porcentaje más alto de nulos en toda la muestra. En oportunidades se encontraron a su vez valores nulos, un 16,19% en clientes (1425 valores); 5,68% en la fecha de creación de la negociación (500 valores); 23,74% en fecha de apertura de negociación (2089); y, por último, el mismo porcentaje para la fecha de cierre, 23,74%. Finalmente, no se encontraron duplicados en los datos.

Siguiendo el procedimiento habitual, se procedió a la normalización de las variables categóricas. La normalización consiste en quitar las mayúsculas, espacios y signos de puntuación. En esta fase, se asignan valores a las variables categóricas y se normaliza el texto. A continuación, en las tablas 4,5,6,7,8 y 9 se muestran las codificaciones de las variables categóricas:

1	Central
2	Este
3	Oeste

Tabla 4. Codificación oficina (equipo).Fuente:Elaboración propia

1	Ganado
2	Perdido
3	Engaging
4	Prospecting

Tabla 5. Codificación estado de la oportunidad (oportunidades).Fuente:Elaboración propia

0	Empleo
1	Entretenimiento
2	Finanzas
3	Marketing
4	Médico
5	Comercio minorista
6	Servicios
7	Software
8	Tecnología
9	Telecomunicaciones

Tabla 6. Codificación sector (clientes).Fuente: Elaboración propia

0	Bélgica
1	Brasil
2	China
3	Alemania
4	Italia
5	Japón
6	Jordania
7	Kenia
8	Corea
9	Noruega
10	Panamá
11	Filipinas
12	Polonia
13	Rumanía
14	Estados Unidos

Tabla 7. Codificación de localización (clientes). Fuente: Elaboración propia

0	Acme Corporation
1	Bubba Gump
2	Golddex
3	Inity
4	Massive Dynamic
5	Sonron
6	Warephase

Tabla 8. Codificación subsidiara de (clientes).Fuente: Elaboración propia

0	GTK
1	GTX
2	MG

Tabla 9. Codificación serie (productos). Fuente: Elaboración propia

4.3.3 Identificación de patrones y valores atípicos

Para ello, se estimaron el primer cuartil y el tercer cuartil, para así obtener el IQR (*Interquartile Range* o Rango Intercuartílico), que representa el rango 50% de los datos (Figura 11).

```
Outliers detectados en: Clientes
  - beneficios: 6 valores atípicos
  - empleados: 8 valores atípicos
------------------------------------------------
Outliers detectados en: Productos
  - precio_venta: 1 valores atípicos
------------------------------------------------
Outliers detectados en: Oportunidades
  - valor_cierre: 15 valores atípicos
------------------------------------------------
Outliers detectados en: Equipo de ventas
  No se encontraron outliers numéricos.
```

Figura 12. Número outliers por tabla. Fuente. Elaboración propiaEn *clientes,* se encontraron 6 valores atípicos en beneficios y 8 valores atípicos en empleados (Figura 12).

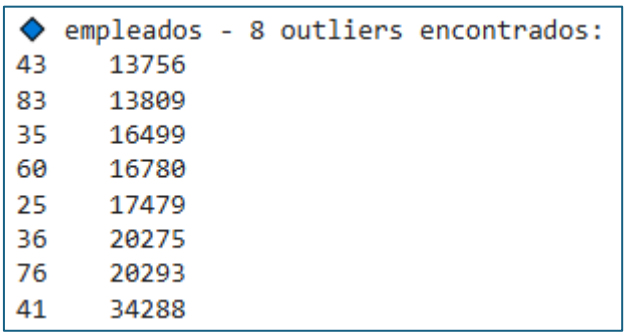


Figura 13. Outliers tabla empleados. Fuente: Elaboración propia

Estos datos sugieren que algunos clientes de esta empresa cuentan con un número de empleados considerablemente mayor en comparación con el resto. Es importante destacar que estos empleados no pertenecen a la empresa de hardware, sino que son trabajadores de los propios clientes.

En las figuras 14 y 15 se muestra la distribución de beneficios de los clientes. Dichas distribuciones se muestran en un diagrama *boxplot* junto con un gráfico de barras con la función de densidad. En el *boxplot (figura 14)*, los límites de la caja representan el primer y tercer cuartil, siendo la línea horizontal que divide la caja la mediana. Los bigotes representan el valor mínimo y el valor máximo, y los círculos los outliers. Se puede observar como la mayoría de los clientes tiene beneficios entre aproximadamente 0 y 45000 unidades monetarias y que la mediana esta ligeramente por debajo del centro, lo que sugiere que la mayor parte de los clientes cuentan con beneficios moderados. En este gráfico, los *outliers* están muy por encima de la banda verde. En la figura 15 se confirma la existencia de *outliers*, ya que la cola derecha es larga. Por tanto, se pone de manifiesto la existencia de clientes con altos recursos a los que se les podría priorizar en estrategias comerciales y de fidelización.

En "*productos"*, tan sólo se encontró un valor atípico en el precio de venta, que corresponde con el precio del artículo GTK 500. En "*oportunidades"* se encontraron 15 valores atípicos en valor de cierre. Estos valores se corresponden con curiosamente tres trabajadoras (Elease Gluck, Markita Hansen y Rosalina Dieter) de la misma oficina (oeste) y el mismo producto (GTK 500). Por otro lado, al ser el producto cuyo precio es el más elevado, no sorprende que se clasifique como valor atípico. En *equipo de ventas* no se encontraron *outliers*.

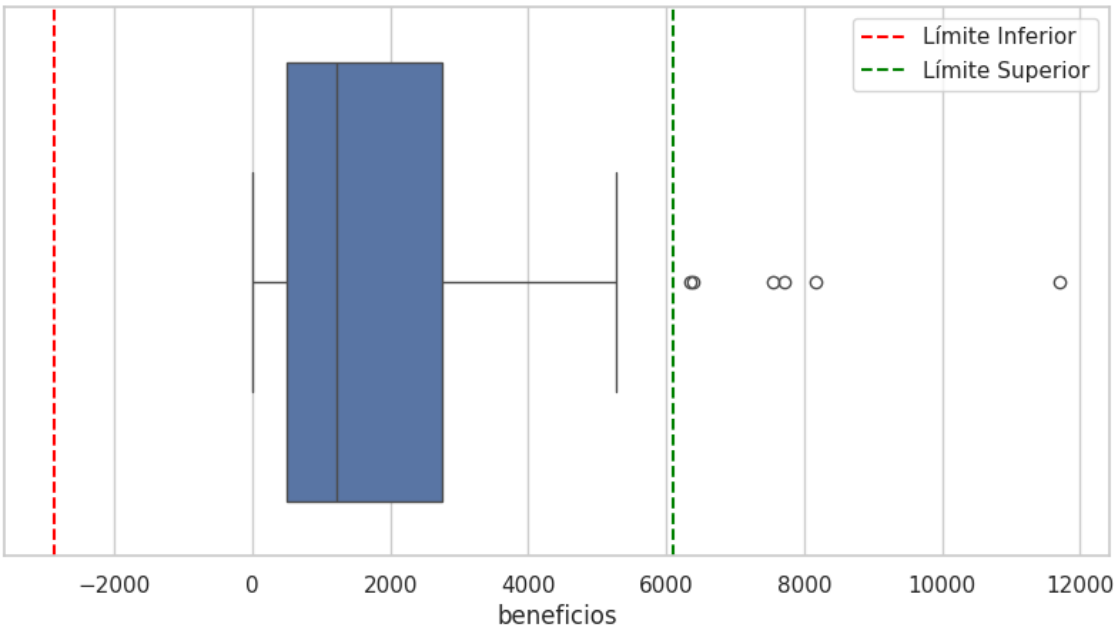


Figura 14. Beneficios clientes. Fuente. Elaboración propia

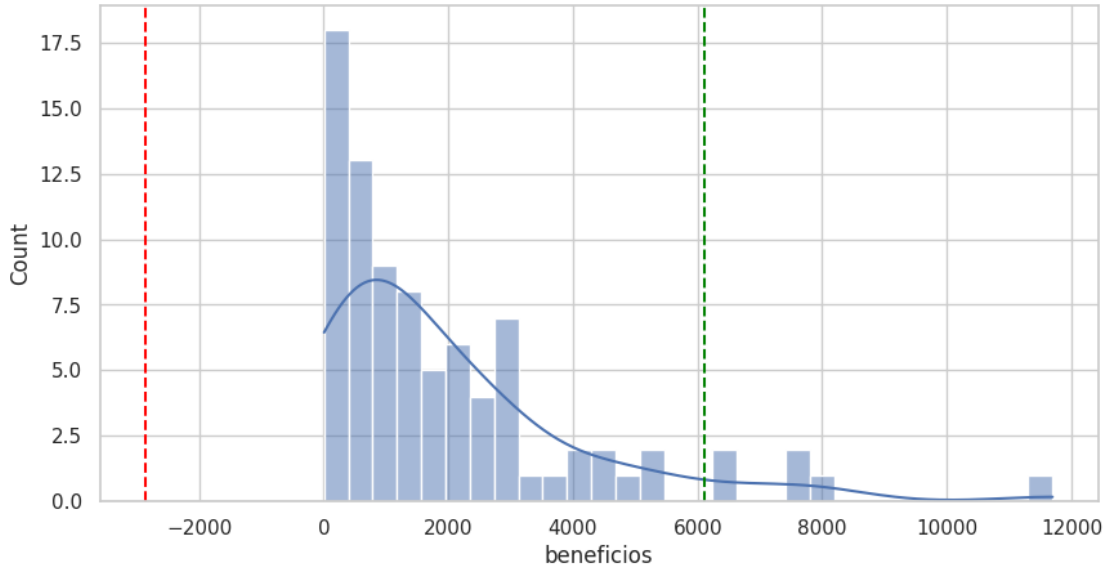


Figura 15. Distribución beneficios clientes. Fuente. Elaboración propia

4.3.4 Análisis exploratorio

Al ser una empresa cuyos productos son hardware de ordenadores, los clientes pueden dividirse en diferentes sectores. En la figura 16 se muestra los estadísticos descriptivos de la tabla *"clientes"* El rango de la fecha de establecimiento de los clientes va desde el año 1979 y 2017, estando la mayoría de las empresas fundadas entre 1989 y 2002.En los beneficios, los estadísticos descriptivos se ven afectados por la existencia de los *outliers*, por lo que la

media no es representativa del grupo. En los empleados también se observa una gran desviación (de 5700 empleados aproximadamente), donde el rango va desde los 9 hasta los 34,288 empleados. En cuanto a la localización, el 75% de los clientes están en la localización codificada como 14, que se corresponde con Estados Unidos.

Estadísticos Descriptivos de Clientes:

	ano_establecimiento	beneficios	empleados	sector_cod	localizacion_cod
count	85.000000	85.000000	85.000000	85.000000	85.000000
mean	1996.105882	1994.632941	4660.823529	4.800000	12.764706
std	8.865427	2169.491436	5715.601198	2.548576	3.246416
min	1979.000000	4.540000	9.000000	0.000000	0.000000
25%	1989.000000	497.110000	1179.000000	3.000000	14.000000
50%	1996.000000	1223.720000	2769.000000	5.000000	14.000000
75%	2002.000000	2741.370000	5595.000000	7.000000	14.000000
max	2017.000000	11698.030000	34288.000000	9.000000	14.000000

Figura 16. Estadísticos descriptivos tabla clientes. Fuente: Elaboración propia

En los estadísticos descriptivos de “oportunidades” (Figura 17), el 75% de las situaciones de negociación pertenecen a la categoría codificada como 2, que representa oportunidad perdida. En cuanto al valor del cierre, la media se sitúa en 1490,92 u.m.

Estadísticos Descriptivos de Oportunidades:

	fecha_creacion	fecha_cierre	valor_cierre	estado_deal_cod
count	8300	6711	6711.000000	8800.000000
mean	2017-06-14 08:35:06.216867584	2017-08-01 03:32:25.641484032	1490.915512	1.812614
min	2016-10-20 00:00:00	2017-03-01 00:00:00	0.000000	1.000000
25%	2017-04-04 00:00:00	2017-05-18 00:00:00	0.000000	1.000000
50%	2017-06-24 00:00:00	2017-08-02 00:00:00	472.000000	2.000000
75%	2017-08-27 00:00:00	2017-10-18 00:00:00	3225.000000	2.000000
max	2017-12-27 00:00:00	2017-12-31 00:00:00	30288.000000	4.000000
std	NaN	NaN	2320.670773	0.924346

Figura 17. Estadísticos oportunidades. Fuente: Elaboración propia

Si se analiza “productos” (Figura 18), el producto más vendido en la oficina central ha sido MG Special, habiéndose vendido 825 unidades. En la oficina oeste el producto más vendido ha sido GTX Basic, con 755 unidades. Por otro lado, la oficina este parece estar quedándose un poco atrasada, ya que su producto más vendido ha sido GTX Pro, que no llega a las 500 ventas.

```
Producto más vendido por oficina:
    oficina    producto  ventas
6   central  mg special     825
15     west   gtx basic     755
11     east     gtx pro     473
```

Figura 18. Producto más vendido por oficina. Fuente. Elaboración propia

En cuanto al producto más vendido por agente en cada oficina (Figura 19), el agente de ventas más productivo fue Darcel Schlecht, con 747 ventas de GTXPro. En la oficina oeste, Vicki Flames vendió 451 MG Special, y en la este Cassey Cress logró vender 346 MG Advanced.

```
Producto más vendido por agente en cada oficina:
    oficina    agente_ventas  ventas producto más vendido
2   central  darcel schlecht     747              gtx pro
28     west   vicki laflamme     451           mg special
11     east     cassey cress     346         mg advanced
```

Figura 19. Producto más vendido por agente en cada oficina. Fuente. Elaboración propia

4.3.5 Visualización exploratoria de los datos

Como se puede observar en la figura 20, los beneficios de los clientes de los sectores de software y tecnología son muy variados al tener una caja muy grande, siendo también ambos los que mayor beneficio tienen. Por otro lado, el sector de servicios es el que menor beneficio obtiene, situándose su mediana en el nivel más bajo comparado con el resto.

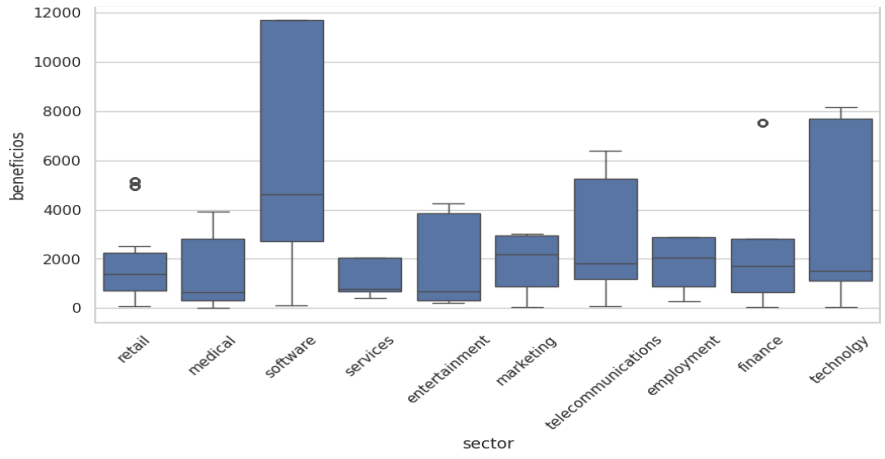


Figura 20. Distribución beneficios por sector. Fuente: Elaboración propia

En la figura 21 se muestra la frecuencia de compra de los productos diferenciándolos por colores. El producto más vendido fue el GTX Basic seguido por MG Special. No sorprenden

estos resultados, ya que son los dos productos más baratos y que la mayoría de los clientes cuentan con ingresos moderados. El tercer producto con mayor frecuencia de compra es GTX Pro seguido de cerca por el MG Advanced, que a su vez está prácticamente a la par con el GTX Plus Pro.

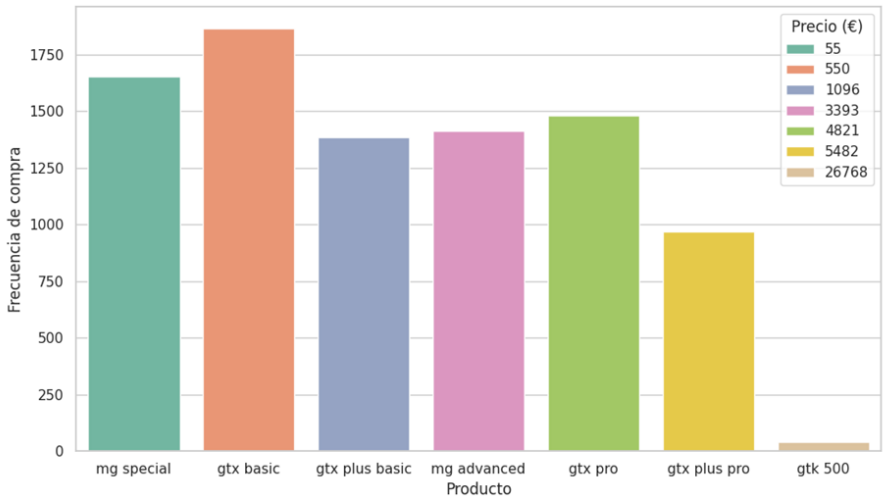


Figura 21. Frecuencia de compra por producto (coloreada por precio). Fuente: Elaboración propia

Al comparar la frecuencia de compra de los productos agrupándolos según su rango de precio, se comprueba que los productos más vendidos son los de rango de precio bajo seguido por los de rango medio. Ambos grupos tienen poca diferencia en el precio, en especial los productos de rango medio. Por el contrario, los productos de rango elevado se venden poco, además de diferencia entre sus precios que se puede observar dado el tamaño de la caja (Ver figura 22).

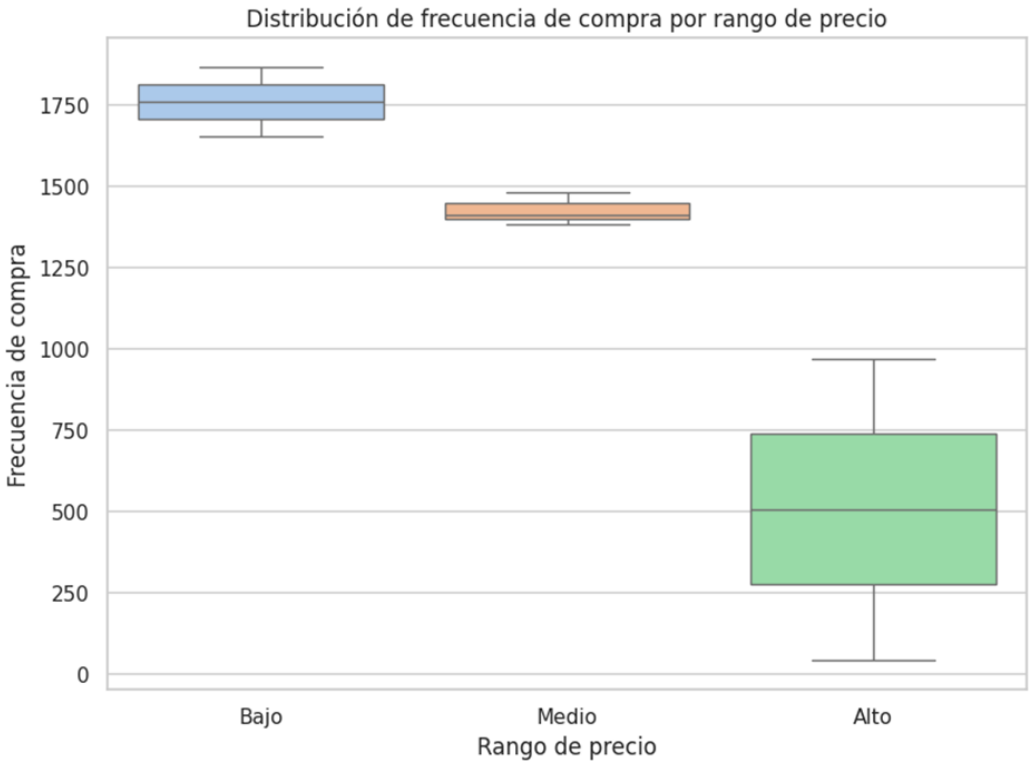


Figura 22. Distribución de frecuencia de compra por rango de precio. Elaboración propia

4.3.6. Aplicación de las técnicas de BA

Objetivo 1: Predecir si un cliente comprará un producto de gama alta

Dada la baja frecuencia de compra de los productos de gama alta, se aplicaron tanto *Random Forest* como *RandomForestClassifier* para tratar de predecir si un cliente comprará un producto de gama alta, establecido como aquellos productos cuyo precio sea mayor que 5000 u.m (unidades monetarias). Sin embargo, dado que hay más productos baratos que caros, había un desbalance que en un primer momento entrenó al modelo a predecir siempre la clase 0 (productos que no son de alta gama). Por ello, se añadieron las variables sector del cliente y oficina, así como el uso de SMOTE (*Synthetic Minority Over-sampling Technique*). SMOTE es un algoritmo de *oversampling* que produce instancias sintéticas para balancear la distribución de clases en el conjunto de datos, utilizando el enfoque del k-vecino más cercano para generar nuevas muestras (Moreno et al., 2009).

Los resultados muestran que el modelo tiene un rendimiento moderado, logrando identificar correctamente a una proporción relativamente baja de los compradores reales (*recall*). El F1-score en ambas clases indica que el modelo mantiene equilibrio relativo entre identificar correctamente a los compradores y evitar clasificaciones erróneas.

```
Clasificación con Árbol de Decisión (J48)
[[ 824 1169]
 [ 662 1263]]
              precision    recall  f1-score   support

           0       0.55      0.41      0.47      1993
           1       0.52      0.66      0.58      1925

    accuracy                           0.53      3918
   macro avg       0.54      0.53      0.53      3918
weighted avg       0.54      0.53      0.53      3918
```

Figura 23. Clasificación con Árbol de Decisión. Fuente: Elaboración propia

En total, el modelo predijo 824 verdaderos negativos (no compra no alta gama) y 1169 falsos positivos. Por otro lado, predijo correctamente 1263 verdaderos positivos y 662 falsos positivos... Por tanto, el modelo acierta más al predecir quién comprará que al identificar quién no lo hará. Sin embargo, tiene una tasa de error elevada al predecir clase 0, clasificando incorrectamente muchos no compradores como compradores (Ver figuras 23 y 24).

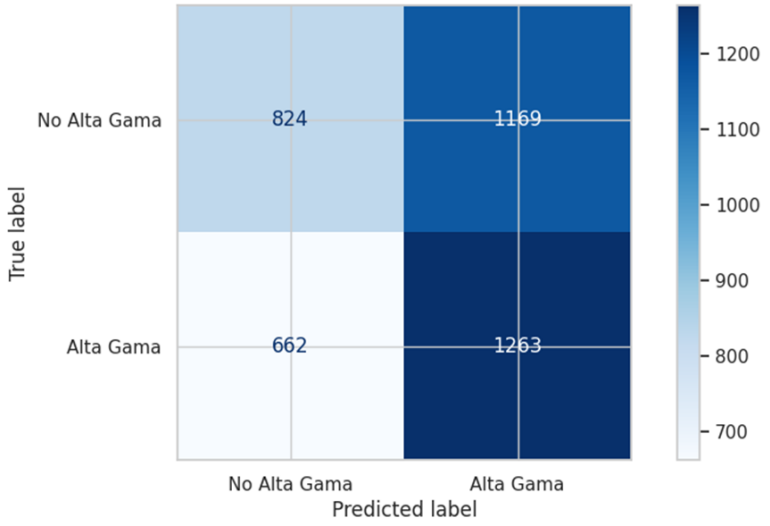


Figura 24. Matriz de confusión Árbol de decisión. Fuente: Elaboración propia

El análisis de importancia de atributos en el modelo de árbol de decisión muestra que las variables más determinantes para predecir la compra de un producto de gama alta son el año de establecimiento del cliente (53,9%), seguido por el número de empleados (17,2%), la localización geográfica (11,7%) y la oficina comercial asignada (10,1%). Estos factores, relacionados con la estructura y el contexto de la empresa cliente, son los que más influyen en las decisiones del modelo.

📊 Importancia de atributos para el Árbol de Decisión:

	atributo	importancia
0	ano_establecimiento	0.539196
2	empleados	0.171668
12	localizacion_brazil	0.117401
26	oficina_east	0.101101
6	sector_medical	0.041784
3	sector_entertainment	0.016546
1	beneficios	0.006019
27	oficina_west	0.005494
11	sector_telecommunications	0.000792
7	sector_retail	0.000000

Figura 25. Matriz de confusión Árbol de decisión. Fuente: Elaboración propia

En contraste, variables como el sector económico, los beneficios o la ubicación de otras oficinas tienen una influencia mucho menor. Esto sugiere que, el tamaño, la antigüedad y la región del cliente son más relevantes que su industria concreta a la hora de identificar patrones de compra de productos de gama alta.

La figura 26 muestra el árbol de decisión utilizado para predecir si un cliente comprará un producto de gama alta, limitado a una profundidad de tres niveles para facilitar su análisis. Cada nodo representa una decisión basada en una variable, mostrando el número de muestras, la distribución entre clases (value = [No, Sí]) y el índice Gini, que mide la impureza del nodo (0 = totalmente puro, 0.5 = mezcla equitativa).

En la raíz, el árbol divide los datos según si el cliente está localizado en Brasil (localizacion_brazil <= 0.5), con un índice Gini de 0.5 y clases muy equilibradas (5436 "No", 4604 "Sí"). A medida que el árbol se ramifica, se consideran variables como oficina_east, empleados y ano_establecimiento.

Por ejemplo, en la rama izquierda, si un cliente pertenece a la oficina Este y tiene menos de 18.877 empleados, se llega a nodos más homogéneos. Uno de ellos agrupa 6194 clientes con un reparto casi equilibrado (3200 “No” frente a 2994 “Sí”). En cambio, en la rama derecha, encontramos nodos más puros, como uno con 52 muestras en el que 44 son “No” y solo 8 son “Sí” (Gini = 0.26).

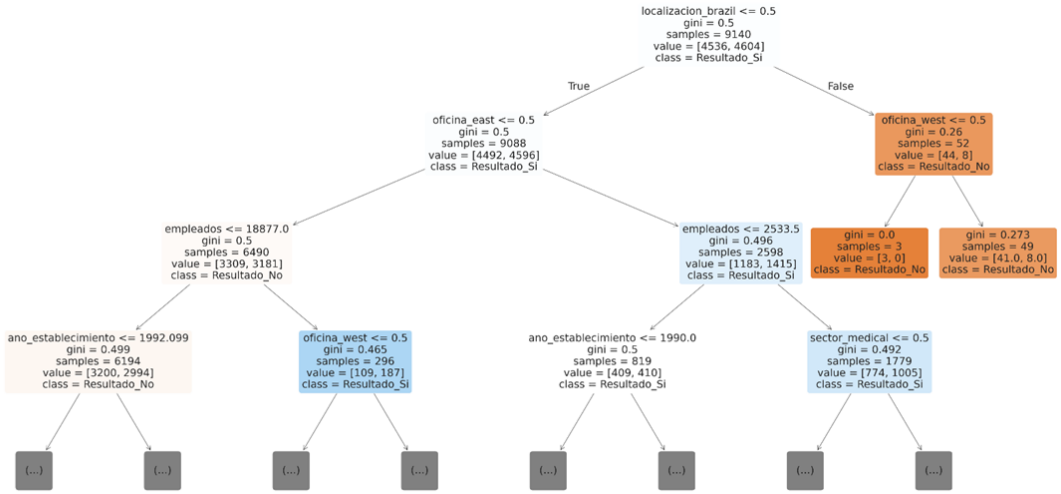


Figura 26. Visualización árbol de decisión. Fuente: Elaboración propia

Este análisis aporta transparencia sobre cómo el modelo toma decisiones.

Al comparar los modelos de Árbol de Decisión y Random Forest, se observa que ambos presentan un rendimiento moderado, aunque con diferencias en la forma de clasificar las clases. El modelo de *Random Forest* ofreció un desempeño más balanceado. Los indicadores de calidad son mejores que los obtenidos con el árbol de decisión,mostrando una precisión del 63 % para la clase 0 y 56 % para la clase 1, con un f1-score medio de 0.58 (Ver figura 27).

```
Clasificación con Random Forest
[[ 938 1055]
 [ 562 1363]]
              precision    recall  f1-score   support

           0       0.63      0.47      0.54      1993
           1       0.56      0.71      0.63      1925

    accuracy                           0.59      3918
   macro avg       0.59      0.59      0.58      3918
weighted avg       0.60      0.59      0.58      3918
```

Figura 27. Clasificación con Random Forest. Fuente: Elaboración propia

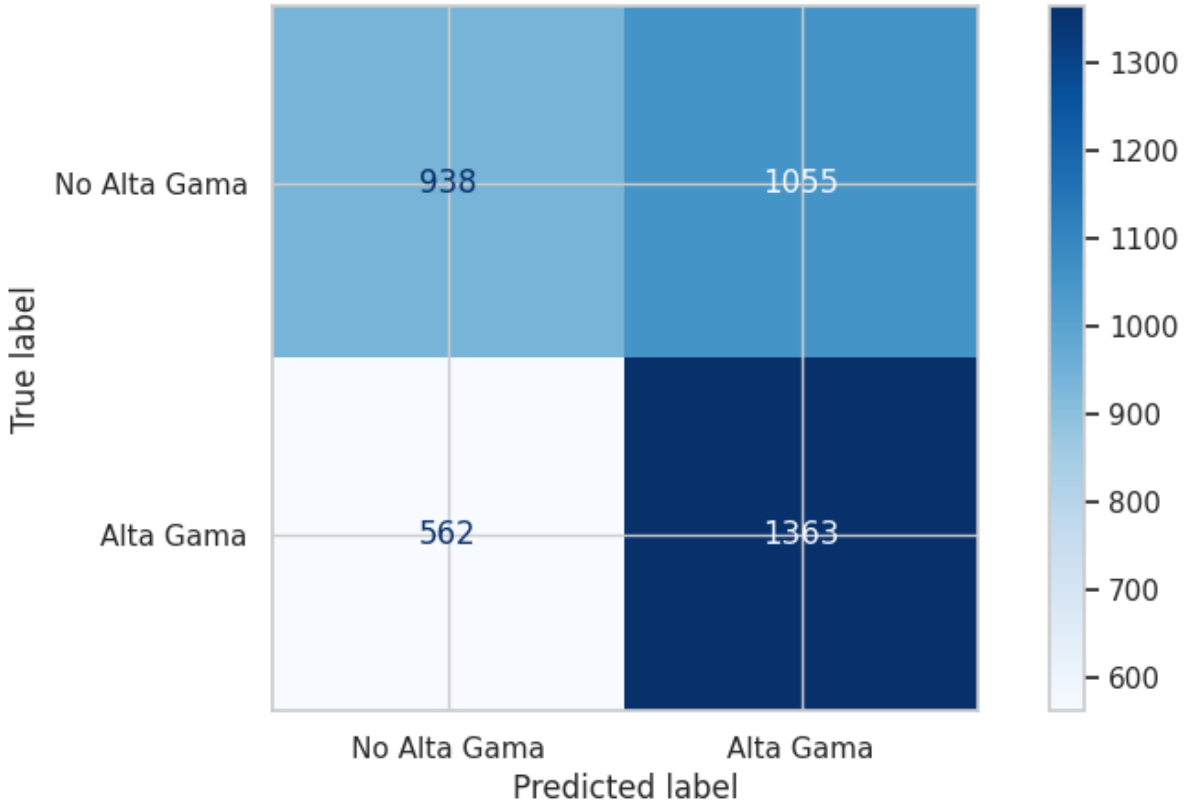


Figura 28. Matriz de confusión Random Forest. Fuente: Elaboración propia

La matriz de confusión del modelo Random Forest muestra que el modelo predijo correctamente 938 casos de clase 0 y 1363 de clase 1. Sin embargo, se produjeron 1055 falsos positivos y 562 falsos negativos, lo que indica un rendimiento moderado con bastantes errores de clasificación en ambas clases (Figura 28). A pesar de ello, *Random Forest* ha predicho mejor que el árbol de decisión (mayor número de verdaderos negativos y menor de falsos positivos).

En resumen, el modelo de *Random Forest* es más robusto en escenarios con ruido o datos más complejos, gracias a su naturaleza de ensamble de árboles múltiples

Los atributos más relevantes para *Random Forest* difieren en su mayoría de los del árbol de decisión, salvo año establecimiento y empleados, que coinciden en ambos modelos aunque en orden inverso (Figura 29).

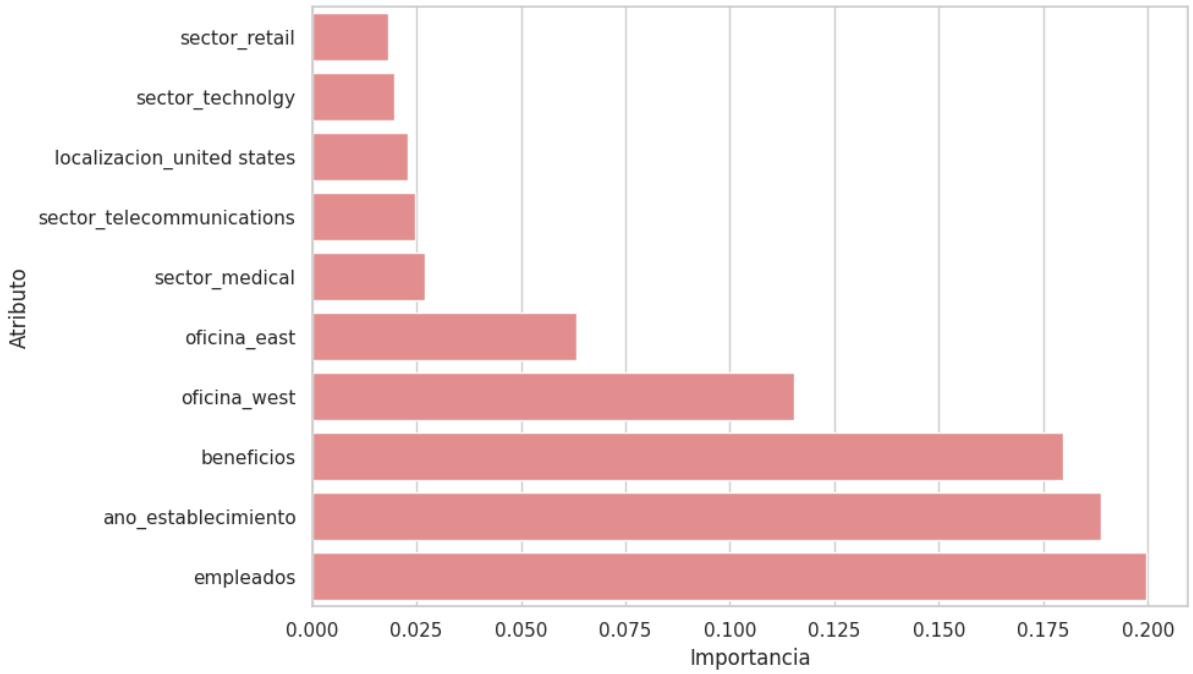


Figura 29. Importancia de los atributos *Random Forest*. Fuente: Elaboración propia

Objetivo 2: Segmentación de clientes

El análisis exploratorio evidenció una marcada diversidad en la base de clientes, reflejada en variables como sector, número de empleados, beneficios y localización. Esta heterogeneidad, observable en los amplios rangos y desviaciones estándar de los datos, sugiere que un enfoque único no sería adecuado para todos los casos. Aunque en una primera etapa se analizaron relaciones básicas —por ejemplo, los beneficios según el sector—, no se llegó a identificar grupos de clientes con combinaciones similares de características. En este contexto, el análisis de clustering resultó clave.

El método del codo ayuda a elegir el número óptimo de clústeres al identificar el punto donde añadir más grupos ya no mejora significativamente la cohesión interna. En este caso indica que tres clústeres es el número óptimo (Figura 30).

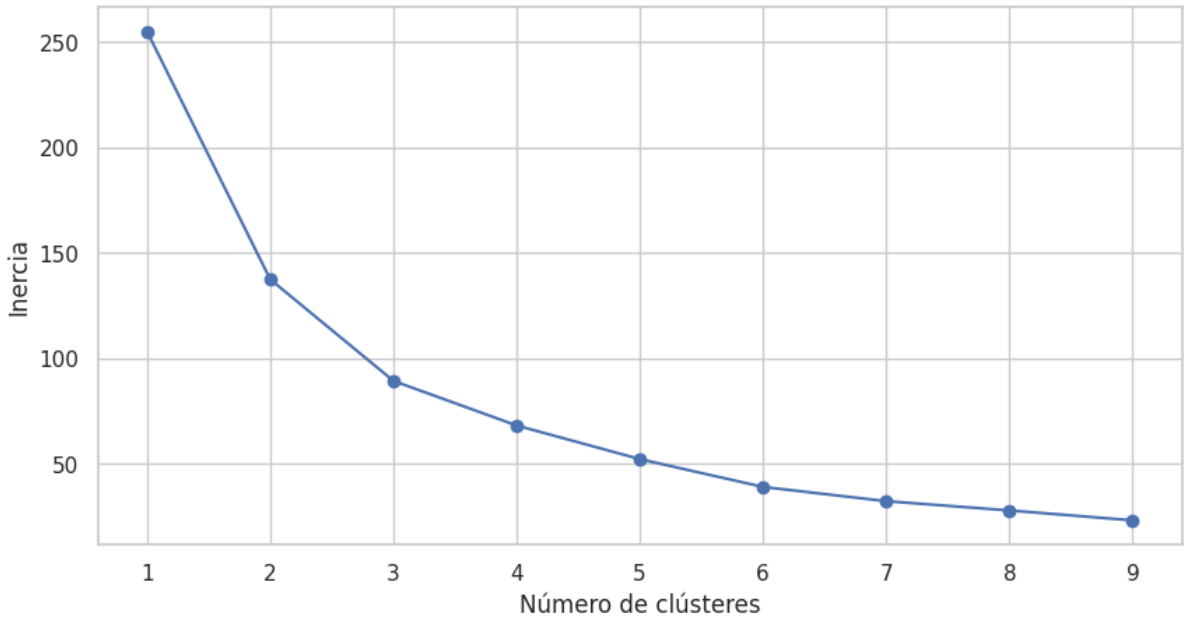


Figura 30. Método del codo para seleccionar K grupos. Fuente: Elaboración propia

El gráfico de dispersión (Figura 31) muestra la relación entre el número de empleados y los beneficios de cada cliente, con cada punto coloreado según el clúster al que pertenece. En primer lugar, el Clúster 0, representado por puntos verdes, es el grupo más numeroso. Se caracteriza por clientes con un número de empleados y beneficios generalmente bajos o medios, aunque con una dispersión considerable. La media de empleados en este grupo es de aproximadamente 2,543, mientras que la media de beneficios es de 1,153.42 u.m.

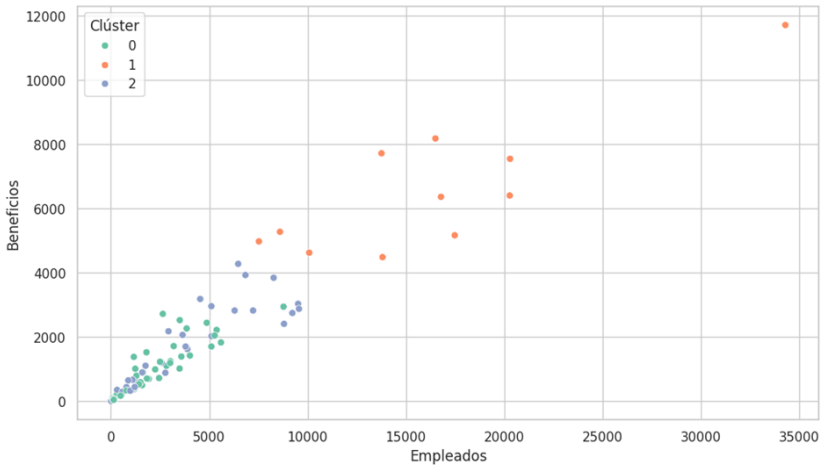


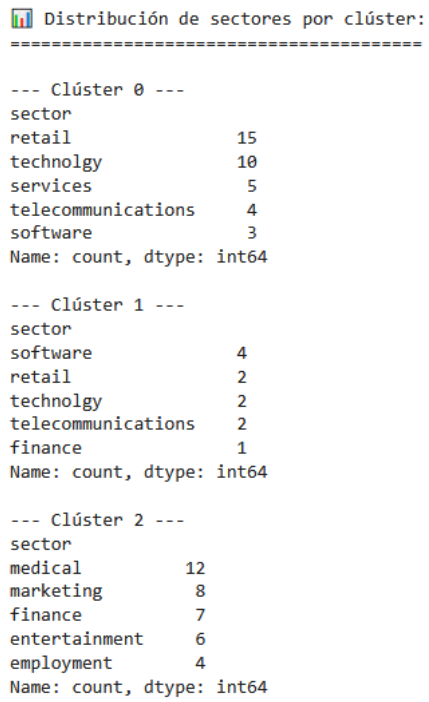
Figura 31. Clústeres de clientes según empleados y beneficios. Fuente: Elaboración propia

Por otro lado, el Clúster 1, identificado con puntos naranjas y ubicado en la parte superior derecha del gráfico, agrupa a los clientes más grandes y rentables. Este grupo es más reducido, pero destaca por tener una media de empleados mucho mayor, alrededor de 16,307, y una media de beneficios que alcanza los 6,577.65 u.m.

Finalmente, el Clúster 2, representado por puntos morados, es similar al Clúster 0 en cuanto a número de empleados y beneficios, aunque muestra una ligera tendencia a contar con más

empleados. En este grupo, la media de empleados es de aproximadamente 3,316 y la de beneficios es de 1,473.33 u.m.

En conjunto, el análisis de clústeres parece haber separado con éxito a los clientes más grandes y rentables (clúster 1) del resto, compuesto principalmente por empresas de tamaño medio o pequeño (clústeres 0 y 2). Esto indica que segmentar a los clientes según empleados y beneficios puede ser una estrategia útil para identificar a los de mayor valor (Figura 31).



```
📊 Distribución de sectores por clúster:
========================================

--- Clúster 0 ---
sector
retail                15
technolgy             10
services               5
telecommunications     4
software               3
Name: count, dtype: int64

--- Clúster 1 ---
sector
software              4
retail                2
technolgy             2
telecommunications    2
finance               1
Name: count, dtype: int64

--- Clúster 2 ---
sector
medical          12
marketing         8
finance           7
entertainment     6
employment        4
Name: count, dtype: int64
```

Figura 32. Distribución de sectores por clúster. Fuente: Elaboración propia

Se observa que los sectores representados varían entre los tres clústeres. El clúster 0 presenta una composición bastante diversa, con mayor presencia en los sectores de *retail*, tecnología, servicios, telecomunicaciones y software. En contraste, el clúster 1, previamente identificado como el de mayor tamaño en términos de empleados y beneficios, concentra a sus clientes principalmente en los sectores de software, *retail*, tecnología y telecomunicaciones, aunque también incluye algunos del sector financiero. Esto sugiere que los sectores tecnológicos tienen un peso relevante entre los clientes más grandes.

Por otro lado, el clúster 2 agrupa principalmente a empresas de los sectores médico, marketing, financiero, entretenimiento y empleo, mostrando una composición distinta a los otros dos.

En conjunto, el análisis sectorial refuerza que los clústeres no solo se distinguen por el tamaño de los clientes, sino también por la industria a la que pertenecen. Esta segmentación puede resultar útil para diseñar estrategias comerciales y de marketing más focalizadas.

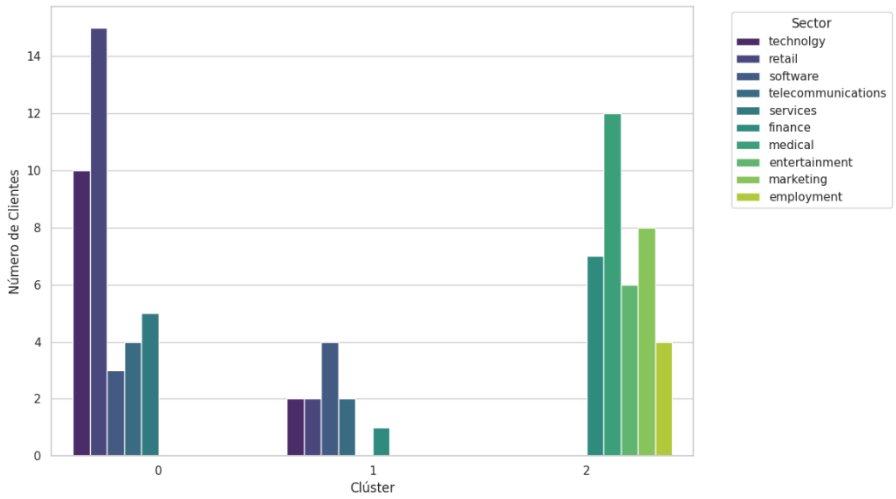


Figura 33. Representación de la distribución de sectores por clúster clientes. Fuente: Elaboración propia

Al observar la distribución de localizaciones por clúster (Figura 32), representada en la segunda gráfica generada (Figura 33), se aprecia una fuerte concentración de clientes en Estados Unidos en los tres grupos. El clúster 0 muestra una mayoría clara de clientes en este país, aunque también se detecta una presencia muy reducida en lugares como Italia, Noruega, Brasil, Alemania, Panamá, Bélgica, Rumanía y China.

```
📊 Distribución de localizaciones por clúster:
========================================

--- Clúster 0 ---
localizacion
united states    29
italy             1
norway            1
brazil            1
germany           1
panama            1
belgium           1
romania           1
china             1
Name: count, dtype: int64

--- Clúster 1 ---
localizacion
united states    9
japan            1
korea            1
Name: count, dtype: int64

--- Clúster 2 ---
localizacion
united states    33
kenya             1
philipines        1
jordan            1
poland            1
Name: count, dtype: int64
```

Figura 34. Distribución de localización por clúster. Fuente. Elaboración propia.

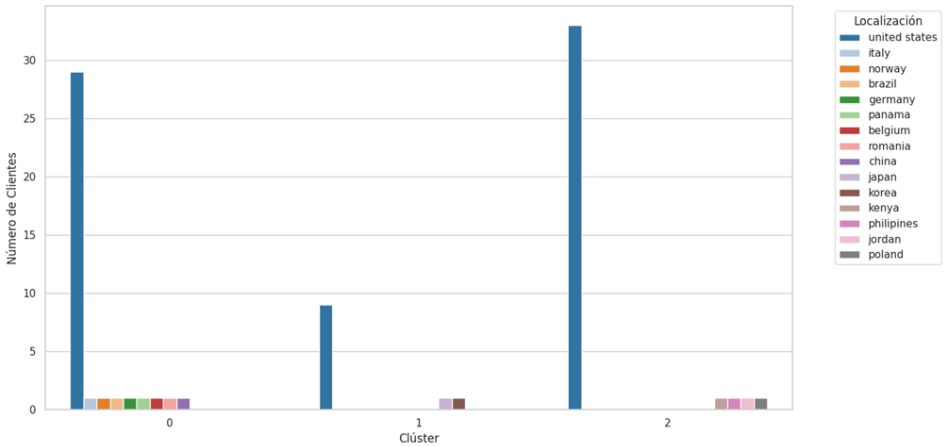


Figura 35. Representación de la distribución de localización por clúster. Fuente. Elaboración propia

El clúster 1 sigue un patrón similar, con mayoría en Estados Unidos y una presencia muy limitada en países como Japón y Corea. Por su parte, el clúster 2 también presenta una alta concentración en Estados Unidos, aunque incluye algunos clientes en Kenia, Filipinas, Jordania y Polonia.

En conjunto, la localización geográfica no parece ser un factor determinante en la diferenciación entre clústeres, al menos cuando se analiza a nivel de país.

El análisis del año de establecimiento (Figuras 36 y 37) muestra que el clúster 1 agrupa a las empresas más antiguas, con una media cercana a 1988, lo que sugiere mayor consolidación. En cambio, los clústeres 0 y 2 tienen medias más recientes (alrededor de 1995–1998), indicando que agrupan a organizaciones fundadas principalmente en los años 90 y 2000. Estos datos refuerzan la diferenciación del clúster 1 como el más veterano, lo que puede ser útil para definir estrategias específicas según la antigüedad del cliente.

📊 Estadísticas del año de establecimiento por clúster:

Cluster	count	mean	std	min	25%	50%	75%	max
0	37.0	1998.837838	8.301651	1982.0	1993.0	2000.0	2005.0	2013.0
1	11.0	1988.727273	6.856981	1979.0	1982.0	1990.0	1994.5	1997.0
2	37.0	1995.567568	8.764006	1983.0	1989.0	1994.0	2000.0	2017.0

Figura 36. Estadísticas año establecimiento por clúster. Fuente: Elaboración propia

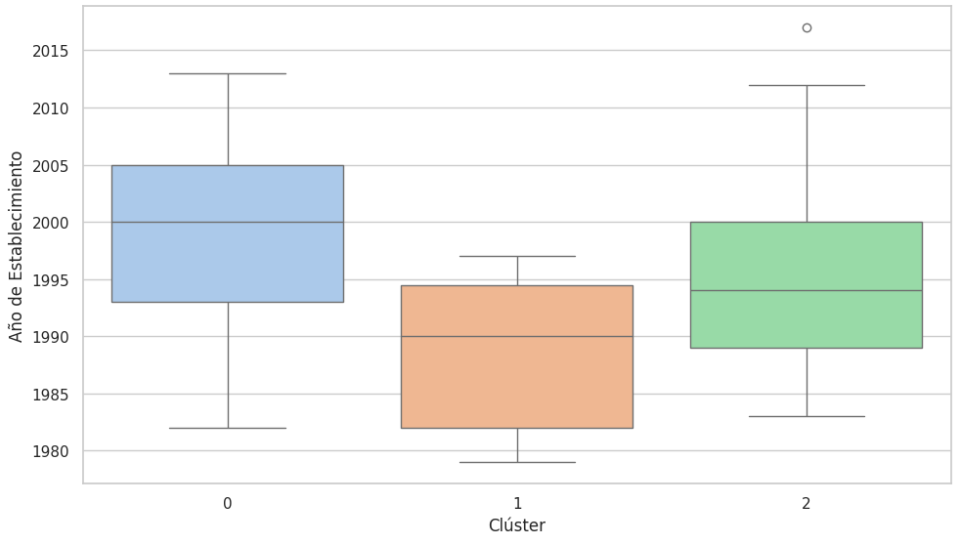


Figura 37. Distribución del año de establecimiento por clúster. Fuente: Elaboración propia

El análisis del número de oportunidades por cliente muestra diferencias claras entre los clústeres. El clúster 1 destaca con una media de 120,18 oportunidades, indicando una alta actividad comercial con clientes grandes y consolidados. En cambio, los clústeres 0 y 2 presentan medias más bajas y similares (alrededor de 80), lo que sugiere un volumen de oportunidades más moderado (Ver figuras 38 y 39). Estos resultados refuerzan la singularidad del clúster 1 como el grupo con mayor intensidad en la relación comercial.

📊 Estadísticas del número de oportunidades por cliente por clúster:

Cluster	count	mean	std	min	25%	50%	75%	max
0	37.0	80.405405	19.302359	51.0	62.0	79.0	94.0	121.0
1	11.0	120.181818	49.215482	70.0	83.0	105.0	153.5	200.0
2	37.0	83.189189	23.780054	52.0	64.0	83.0	94.0	170.0

Figura 38. Estadísticas de oportunidades por cliente por clúster. Fuente: Elaboración propia

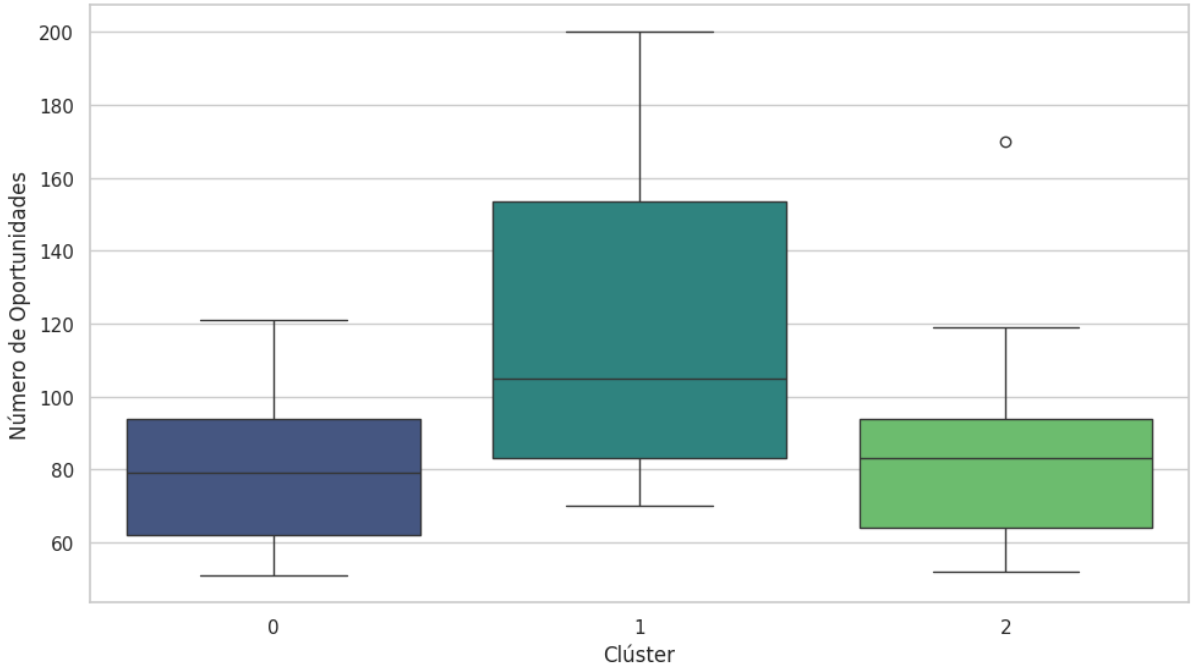


Figura 39. Distribución del número de oportunidades por cliente por clúster. Fuente: Elaboración propia

Objetivo 3: Identificar productos con comportamiento similar, así como encontrar los productos más y menos vendidos

Durante la fase de análisis exploratorio, se identificó una amplia variabilidad en el precio como en la frecuencia de compra de los productos, lo que indica diferencias en comportamiento. Por ello, aplicar técnicas de clustering basadas en estas variables permite agrupar los productos en segmentos con comportamientos similares.

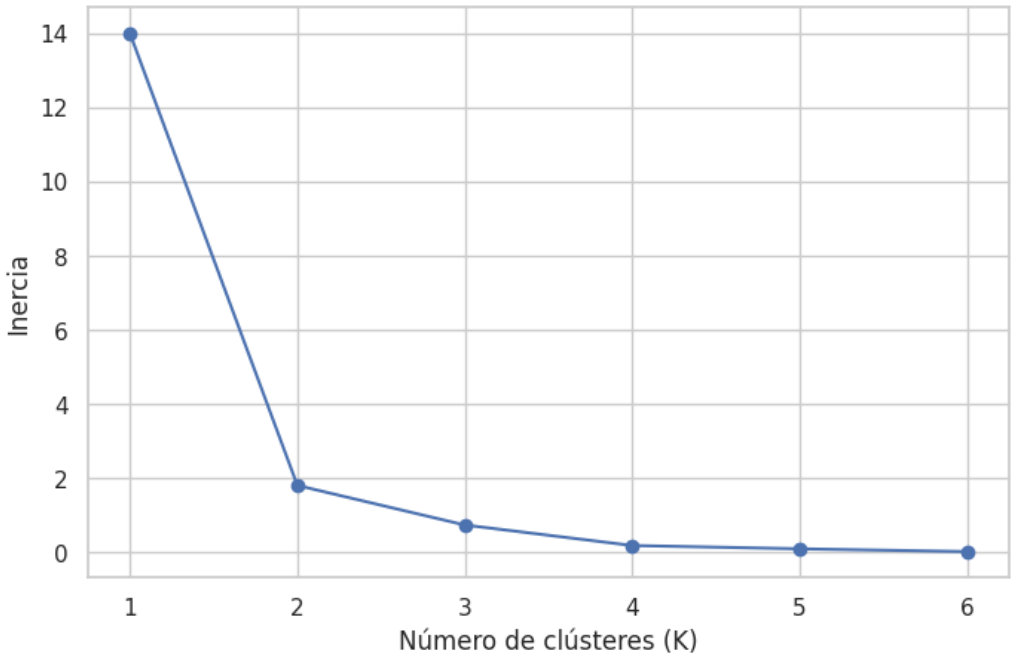


Figura 40. Método del codo para clústeres de productos. Fuente: Elaboración propia

El gráfico de la inercia muestra que el punto de inflexión se encuentra en K=3, lo que sugiere que tres clústeres es una elección adecuada para agrupar los productos según su precio y frecuencia de compra.

	producto	precio_venta	Frecuencia	Cluster
0	gtx basic	550	1866	2
1	gtx pro	4821	1480	2
2	mg special	55	1651	2
3	mg advanced	3393	1412	2
4	gtx plus pro	5482	968	0
5	gtx plus basic	1096	1383	2
6	gtk 500	26768	40	1

Figura 41. Productos en cada clúster según su precio y frecuencia de compra. Fuente: Elaboración propia

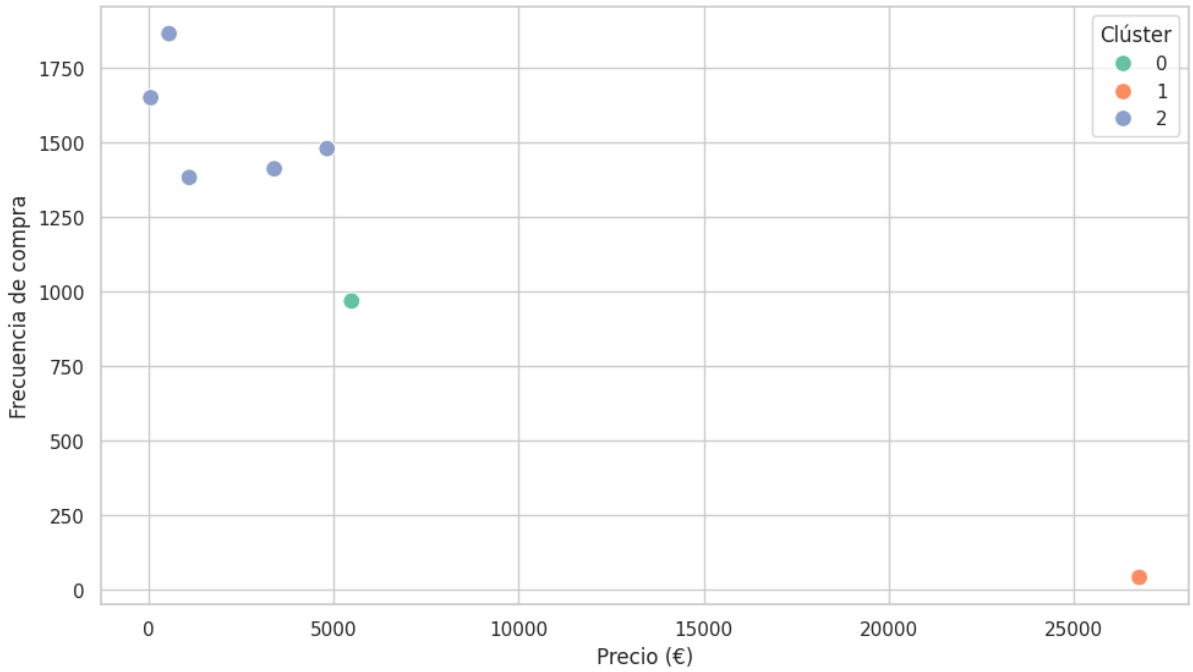


Figura 42. Clústeres de productos según precio y frecuencia de compra. Fuente: Elaboración propia

Al analizar la tabla y el gráfico de dispersión de los clústeres de productos, se observan tres grupos bien definidos. El clúster 0 incluye solo al producto *gtx plus pro*, que tiene un precio elevado (5482 €) y una frecuencia de compra moderadamente baja (968). El clúster 1 también contiene un único producto, *gtk 500*, que destaca por tener el precio más alto de todos (26.768 €) y la frecuencia de compra más baja (40) (Figura 41).

Por otro lado, el clúster 2 agrupa a la mayoría de los productos: *gtx basic*, *gtx pro*, *mg special*, *mg advanced* y *gtx plus basic*. Estos productos presentan precios más bajos en comparación con los clústeres y frecuencias de compra significativamente más altas, lo que los define como parte de la gama principal o de mayor volumen.

En conjunto, los clústeres reflejan una segmentación clara entre productos de volumen, alta gama y ultra alta gama (Figura 42).

4.5 Resultados

El análisis se basó en cuatro tablas principales (clientes, productos, oportunidades y equipo de ventas) con un total de 85 clientes, 8800 oportunidades y 35 agentes. Se identificaron valores nulos y atípicos.

En el análisis descriptivo, el estado de las oportunidades fue mayoritariamente "Won" (4238), y el periodo analizado abarcó 437 días. Los sectores más rentables fueron software y tecnología. Los productos más vendidos variaron por oficina. Se observó que los productos de menor precio se vendían con mayor frecuencia.

Se construyeron dos modelos para predecir productos de alta gama:

- Árbol de Decisión: precisión del 53% (F1: 0.47 y 0.58), principales predictores: año de establecimiento, empleados y localización.
- *Random Forest*: precisión del 59% (F1: 0.54 y 0.63), predictores más relevantes: empleados, año de fundación y beneficios.

En segmentación, se identificaron tres clústeres de clientes; el Clúster 1 agrupó a los más grandes y rentables. También se agruparon productos según precio y frecuencia de compra.

Para predecir el abandono de oportunidades, se desarrolló un modelo Random Forest (precisión del 57%, F1 = 0.57 en ambas clases). Las variables más importantes fueron el precio de venta y la frecuencia de compra. Se detectaron tasas de abandono más altas en sectores como finanzas y telecomunicaciones, y en países como Corea o Noruega, destacando el Clúster 1 como el de mayor riesgo.

5.DISCUSIÓN

Los resultados obtenidos a lo largo del análisis y la modelización son, en general, coherentes con lo que cabría esperar en un entorno de ventas B2B con una estructura organizativa amplia y una oferta de productos diversificada.

En primer lugar, el reparto de oportunidades de venta entre los diferentes estados ("Won", "Lost", "Engaging", "Prospecting") muestra una distribución lógica, con una proporción significativa de oportunidades ganadas (48%) frente a las perdidas (28%) y en fases intermedias. Esta distribución es habitual en procesos de ventas con múltiples etapas y ciclos relativamente largos, como los observados durante los 437 días de operación.

También es razonable que los productos de menor precio se asocien con una mayor frecuencia de compra. Esto sugiere una segmentación natural del mercado, en la que productos más accesibles son adoptados por un mayor número de clientes o se compran de forma más recurrente.

Los modelos predictivos de clasificación lograron resultados moderados. La precisión de los modelos (53% para el Árbol de Decisión y 59% para Random Forest) no es especialmente elevada, pero es consistente con la complejidad del problema y la naturaleza de los datos dado el desequilibrio de clases. La relevancia de variables como *empleados*, *beneficios* y *año de establecimiento* es coherente con el hecho de que empresas más consolidadas y de mayor tamaño suelen ser más propensas a adquirir productos de mayor valor.

La segmentación de clientes permitió identificar perfiles diferenciados en función del tamaño y el sector, lo cual es fundamental para personalizar estrategias comerciales. El Clúster 1, compuesto por empresas más grandes y con mayores beneficios, podría ser un objetivo prioritario para productos de alta gamas.

6.CONCLUSIONES, LIMITACIONES Y FUTURAS LÍNEAS DE TRABAJO.

Este trabajo ha permitido explorar el potencial del Business Analytics aplicado a un entorno de ventas B2B mediante el análisis, segmentación y modelización de datos empresariales

reales. A través de técnicas de análisis descriptivo, modelado predictivo y segmentación, se han obtenido hallazgos relevantes para la toma de decisiones estratégicas.

Uno de los principales aportes del trabajo ha sido demostrar cómo, a partir de datos estructurados y mediante BA, es posible extraer valor tangible para la empresa en forma de recomendaciones prácticas: desde el rediseño de estrategias comerciales hasta la priorización de segmentos clave.

Sin embargo, se identifican algunas limitaciones que conviene destacar. La disponibilidad limitada de variables (sin datos textuales, temporales o de interacción directa con clientes) ha reducido el alcance de los modelos predictivos.

De cara a futuras líneas de trabajo, se propone:

- Ampliar la base de datos con nuevas fuentes (por ejemplo, datos temporales más precisos, información de CRM, actividad digital de clientes o respuestas a campañas).
- Explorar el análisis de sentimientos o minería de texto sobre comunicaciones con clientes (correos, encuestas, tickets) para enriquecer el conocimiento sobre la experiencia del cliente.
- Desarrollar sistemas de apoyo a la toma de decisiones en tiempo real para los equipos de ventas, integrando los modelos desarrollados en aplicaciones empresariales.
- Detectar el posible abandono de clientes.

En resumen, este trabajo pone de relieve cómo la analítica de datos puede ser una herramienta poderosa para guiar decisiones empresariales, y abre la puerta a seguir integrando soluciones basadas en inteligencia artificial para una gestión más inteligente, personalizada y eficiente de los procesos comerciales

8. BIBLIOGRAFÍA

1. **Ajah, I.A. and Nweke, H.F.** (2019) 'Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications', *Big Data and Cognitive Computing*, 3(2). Disponible en: https://doi.org/10.3390/bdcc3020032.
2. **Arana, C.** (2021) *Modelos de aprendizaje automático mediante árboles de decisión*. Buenos Aires: Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA). Disponible en: https://hdl.handle.net/10419/238403.
3. **Asllani, A.** (2014) *Business analytics with management science models and methods*. Upper Saddle River, NJ: Pearson FT Press.
4. **BARC** (2024) *Data, BI and Analytics Trend Monitor 2025: Data Security and Quality Lead Global Priorities*. BARC News, 27 November. Available at: https://barc.com/news/barc-releases-data-bi-and-analytics-trend-monitor-2025-data-security-and-quality-lead-global-priorities/?utm_source=chatgpt.com
5. **Barkhordar, E., Shirali-Shahreza, M.H. y Sadeghi, H.R.** (2021) "Clustering of Bank Customers using LSTM-based encoder-decoder and Dynamic Time Warping", *arXiv* [cs.LG]. Disponible en: http://arxiv.org/abs/2110.11769.
6. **Bordawekar, R., Blainey, B. y Apte, C.** (2014) "Analyzing analytics", *SIGMOD Record*, 42(4), pp. 17–28. Disponible en: https://doi.org/10.1145/2590989.2590993.

7. **Breiman, L.** (2001) "Random Forests", *Machine Learning*, 45(1), pp. 5–32. Disponible en: https://doi.org/10.1023/a:1010933404324.
8. **Camacho, M., Ramallo, S. y Ruiz Marín, M.** (2021) «Árboles de decisión en economía: una aplicación a la determinación del precio de la vivienda», en *Nuevos métodos de predicción económica con datos masivos*. Madrid: FUNCAS, pp. 61-92.
9. **Cano, I.R.** (2018) *Historia y evolución de la analítica de negocio*. Viewnext.com. Disponible en: https://www.viewnext.com/historia-y-evolucion-de-la-analitica-de-negocio/.
10. **Challenger-Pérez, I., Díaz-Ricardo, Y. y Becerra-García, R.A.** (2014) "El lenguaje de programación Python", *Ciencias Holguín*, 20(2), pp. 1-13. Disponible en: https://www.redalyc.org/articulo.oa?id=181531232001.
11. **Davies, T.M.** (2016) *The Book of R: A First Course in Programming and Statistics*. 1ª ed. San Francisco, CA: No Starch Press.
12. **Duque Méndez, N.D. et al.** (2016) "Modelo para el proceso de extracción, transformación y carga en bodegas de datos. Una aplicación con datos ambientales", *Ciencia e Ingeniería Neogranadina*, 26(2), pp. 95–109. Disponible en: https://doi.org/10.18359/rcin.1799.
13. **García, S. et al.** (2016) "Big Data Preprocessing: Methods and Prospects", *Big Data Analytics*, 1(1). Disponible en: https://doi.org/10.1186/s41044-016-0014-0.
14. **Gonzalez-Varona, J.M. et al.** (2024) "Applicability of Business Intelligence Maturity Models to SMEs", *arXiv* [econ.GN]. Disponible en: http://arxiv.org/abs/2406.01616.
15. **Hagan, M.T., Demuth, H.B. y Beale, M.H.** (2014) *Neural Network Design*. 2ª ed. Martin Hagan.
16. **Hermawan, D.R. et al.** (2021) "Comparative Study of J48 Decision Tree Classification Algorithm, Random Tree, and Random Forest on In-Vehicle Coupon Recommendation Data", en *2021 International Conference on Artificial Intelligence and Big Data Analytics*. IEEE, pp. 1–6.
17. **Holguín Londoño, G.A. et al.** (2024) *Introducción a la ciencia y análisis de datos con Python: una visión empresarial*. Pereira: Universidad Tecnológica de Pereira. Disponible en: https://repositorio.utp.edu.co/server/api/core/bitstreams/4097dc4c-50ef-477d-a360-c3f4934b260a/content.
18. **John, J.M., Shobayo, O. & Ogunleye, B.** (2023) 'An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market', *Analytics*, 2(4), pp. 809–823. Disponible en: https://doi.org/10.3390/analytics2040042.
19. Khoirunissa, H.A., Widyaningrum, A.R. y Maharani, A.P.A. (2021) "Comparison of Random Forest, Logistic Regression, and MultilayerPerceptron methods on classification of bank customer account closure", Indonesian Journal of Applied Statistics, 4(1), p. 14. Disponible en: https://doi.org/10.13057/ijas.v4i1.41461
20. **Klimberg, R.K. & Miori, V.** (2010) *Back in Business*. OR/MS Today. Disponible en: https://www.mydral.com/downloads/docs/klimberg - back in business - 2010.pdf.
21. **Kunc, M.** (2019) "The Role of Business Analytics in Supporting Strategy Processes: Opportunities and Limitations", *Journal of the Operational Research Society*, 70(6), p. 974. Disponible en: https://doi.org/10.1057/s41274-018-0155-8.
22. **Liberatore, M. & Luo, W.** (2010) 'The Analytics Movement: Implications for Operations Research', *Interfaces*, 40(4), pp. 313–324. Disponible en: https://doi.org/10.1287/inte.1100.0502.
23. **Luhn, H.P.** (1958) 'A Business Intelligence System', *IBM Journal of Research and Development*, 2(4), pp. 314–319. Disponible en: https://doi.org/10.1147/rd.24.0314.

24. **Moesmann, M. y Pedersen, T.B.** (2024) "Data-Driven Prescriptive Analytics Applications: A Comprehensive Survey", *arXiv* [cs.DB]. Disponible en: http://arxiv.org/abs/2412.00034.
25. **Power, D.J. et al.** (2018) "Defining Business Analytics: An Empirical Approach", *Journal of Business Analytics*, 1(1), pp. 40–53. Disponible en: https://doi.org/10.1080/2573234X.2018.1507605.
26. **Salazar, A. y Kunc, M.** (2025) "The Contribution of GenAI to Business Analytics", *Journal of Business Analytics*, 8(2), pp. 79–92. Disponible en: https://doi.org/10.1080/2573234X.2024.2435835.
27. **Shetty, S.H. et al.** (2022) "Supervised Machine Learning: Algorithms and Applications", en *Fundamentals and Methods of Machine and Deep Learning*. Wiley, pp. 1–16. Disponible en: https://doi.org/10.1002/9781119821908.ch1.
28. **Tableau** (s. f.) *Business intelligence: a complete overview*. Tableau. Disponible en:https://www.tableau.com/es-es/learn/articles/business-intelligence/bi-business-analytics
29. **The Olsys Team** (2025) *Business Intelligence 2025: Emerging Trends and Innovations*. Olsys Careers, 20 February. Available at: https://careers.olsysltd.com/business-intelligence-2025-emerging-trends-and-innovations/?utm_source=chatgpt.com
30. **Therón Sánchez, R.** (2021) "Visualización de datos: Caminos de ida y vuelta entre arte y ciencia en la producción y consumo de imágenes", *Fonseca Journal of Communication*, (23), pp. 39–60. Disponible en: https://doi.org/10.14201/fjc2021233960.
31. **Thorat, A.S. y V.R.S.** (2022) 'A Random Forest Churn Prediction Model: An Investigation of Machine Learning Techniques for Churn Prediction and Factor Identification in the Telecommunications Industry', *Mathematical Statistician and Engineering Applications*, 71(4), pp. 12662–12666. Disponible en: https://doi.org/10.17762/msea.v71i4.2434.
32. **Wilder, C.R. & Ozgur, C.O.** (2015) "Business Analytics Curriculum for Undergraduate Majors", *INFORMS Transactions on Education*, 15(2), pp. 180–187.
33. **Zhai, J.** (2025) 'Automatic Identification Method for Fraudulent Users on E-Commerce Websites Based on Random Forest Algorithm', *Journal of Computational Methods in Sciences and Engineering*, 25(2), pp. 1147–1154. Disponible en: https://doi.org/10.1177/14727978241295575.
34. **Zhao, S.** (2023) "Customer Churn Prediction Based on the Decision Tree and Random Forest Model", *BCP Business & Management*, 44, pp. 339–344. Disponible en: https://doi.org/10.54691/bcpbm.v44i.4840.
35. **"What is Unsupervised Learning?"** (2025) *IBM.com*, 12 mayo. Disponible en: https://www.ibm.com/think/topics/unsupervised-learning (Consultado: 16 junio 2025).