



Pontificia Universidad
JAVERIANA
Cali

**MODELO DE APRENDIZAJE AUTOMÁTICO PARA SEGMENTACIÓN DE CLIENTES ACTUALES Y
PROSPECCIÓN DE NUEVOS CLIENTES POTENCIALES EN UNA EMPRESA DE SERVICIOS DE
TECNOLOGÍA.**

Karen Sofía Mendoza Valero
Código 9026708

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
Cristian Alejandro Torres Valencia

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 16 DE 2026

TABLA DE CONTENIDO

1	CONTEXTUALIZACIÓN DEL PROYECTO	2
1.1	DEFINICIÓN DEL PROBLEMA	2
1.1.1	PLANTEAMIENTO DEL PROBLEMA	2
1.1.2	FORMULACIÓN DEL PROBLEMA.....	3
1.1.3	SISTEMATIZACIÓN	3
2	OBJETIVOS DEL PROYECTO	4
2.1	OBJETIVO GENERAL	4
2.2	OBJETIVOS ESPECÍFICOS	4
3	MARCO TEÓRICO Y ANTECEDENTES.....	5
3.1	CONTEXTO ORGANIZACIONAL	5
3.2	MARCO TEÓRICO	6
3.3	ANTECEDENTES	15
4	CARACTERIZAR LOS DATOS DE LOS CLIENTES ACTUALES UTILIZANDO LA INFORMACIÓN DEL CRM SALESFORCE, CON EL FIN DE IDENTIFICAR VARIABLES RELEVANTES QUE PERMITAN CONSTRUIR PERFILES ESTRATÉGICOS.....	18
4.1	ESTRUCTURACIÓN DE LA BASE DE DATOS	18
4.2	CARACTERIZACIÓN DE LA BASE DE DATOS	23
4.3	ANÁLISIS EXPLORATORIO DE DATOS (EDA) - CARACTERIZACIÓN POR ÁREAS DE PORTAFOLIO Y SUBPRODUCTOS.....	27
5	MODELOS DE APRENDIZAJE AUTOMÁTICO PARA IDENTIFICAR Y PRIORIZAR PROSPECTOS COMERCIALES	37
5.1	MODELO NO SUPERVISADO	39
5.1.1	PROCESAMIENTOS.....	39
5.1.2	RESULTADOS DE LOS CLUSTERS POR BU	56
5.1.3	PRINCIPALES HALLAZGOS DE LOS PERFILES	60

5.2	MODELO SUPERVISADO	61
5.2.1	ESTRATEGIA DE CROSS SELLING	62
5.2.1.1	Surgimiento de la Necesidad de Negocio.....	62
5.2.1.2	Formulación del problema	62
5.2.1.3	Desarrollo iterativo del modelo: tres fases metodológicas.....	64
5.2.1.4	Resultados	67
5.2.2	MODELO SUPERVISADO PARA PROSPECTOS EXTERNOS	77
5.2.2.1	Técnicas de aprendizaje supervisado utilizadas	84
6	EVALUAR EL DESEMPEÑO DE LOS MODELOS CONSTRUIDOS MEDIANTE MÉTRICAS CUANTITATIVAS Y CRITERIOS DE NEGOCIO, CON EL FIN DE APOYAR DECISIONES COMERCIALES BASADAS EN DATOS.....	103
6.1	EVALUACIÓN DEL MODELO NO SUPERVISADO (CLUSTERING).....	103
6.2	EVALUACIÓN DEL MODELO SUPERVISADO	108
6.2.1	Evaluación del Modelo de Cross-Selling (Clientes Actuales).....	108
6.2.2	Evaluación del modelo híbrido de Prospectos Externos	113
7	CONCLUSIONES Y TRABAJOS FUTUROS	118
7.1	CONCLUSIONES	118
7.1.1	ESTRUCTURACIÓN DE DATOS Y PERFILES (OBJETIVO 1).....	118
7.1.2	MODELOS DE SEGMENTACIÓN Y PROSPECCIÓN (OBJETIVO 2).....	120
7.1.2.1	Conclusiones del segundo objetivo — modelo no supervisado	120
7.1.2.2	Modelo de cross-selling e impacto en el crecimiento comercial	121
7.1.2.3	Modelo supervisado para prospectos externos.....	124
7.1.3	EVALUACIÓN INTEGRAL DE LOS MODELOS (OBJETIVO 3)	125
7.2	TRABAJOS FUTUROS	126
7.2.1	Expansión del modelo a escala latinoamericana	126
7.2.2	Enriquecimiento del modelo con variables de contexto tecnológico	127
7.2.3	Tablero de inteligencia comercial integrado con Salesforce	127
7.2.4	Actualización periódica del modelo	128
8	REFERENCIAS BIBLIOGRÁFICAS	129

9	ANEXOS	132
	ANEXO A. REPOSITORIO DEL PROYECTO	132
	ANEXO B. PROMPT PARA CONSTRUCCIÓN DE LA BASE DE PROSPECTOS EXTERNOS	133

LISTA DE FIGURAS

Figura 1. Análisis gráfico: Método del Codo y Silhouette.....	41
Figura 2. Proyección PCA — Segmentación global con $k=7$ (solapamiento entre clústeres).....	52
Figura 3. Heatmap de Afinidad Multi-BU — Top 20 Clientes Estratégicos.....	70
Figura 4. Correlación entre Scores de Afinidad por BU	71
Figura 5. Variables más importantes — Random Forest (Soluciones SAP).....	73
Figura 6. Variables más importantes — Random Forest (Cyber Security).....	73
Figura 7. Variables más importantes — Random Forest (Digital Business)	74
Figura 8. Variables más importantes — Random Forest (Digital Workplace).....	74
Figura 9. Variables más importantes — Random Forest (Private Cloud).....	75
Figura 10. Variables más importantes — Random Forest (Cloud Pública).....	75
Figura 11. Distribución del Score Total de Similitud — Prospectos externos.....	91
Figura 12. Mapa de Calor de Afinidad Híbrida por BU — Top 20 Prospectos Externos.....	98
Figura 13. Correlación entre Scores Finales de Afinidad por BU (Modelo Híbrido).....	100
Figura 14. Dispersión de Scores Finales de Afinidad Híbrida — SAP vs Private Cloud.....	101
Figura 15. PCA del clustering y curva Silhouette vs k — Cloud Pública	104
Figura 16. PCA del clustering y curva Silhouette vs k — Cyber Security.....	105
Figura 17. PCA del clustering y curva Silhouette vs k — Digital Business / Aplicaciones	105
Figura 18. PCA del clustering y curva Silhouette vs k — Digital Workplace	106
Figura 19. PCA del clustering y curva Silhouette vs k — Private Cloud / Infra Tradicional.....	107
Figura 20. PCA del clustering y curva Silhouette vs k — Soluciones SAP.....	107
Figura 21. Distribución del AUC bajo CV repetida estratificada — Cloud Pública (Cross-selling).....	111
Figura 22. Distribución del AUC bajo CV repetida estratificada — Cyber Security (Cross-selling)	112
Figura 23. Distribución del AUC bajo CV repetida estratificada — Digital Workplace (Cross-selling).....	113
Figura 24. Comparación de AUC-ROC por modelo y Unidad de Negocio (Prospectos externos).....	115
Figura 25. Mejor modelo por Unidad de Negocio — AUC-ROC (Prospectos externos)	115

LISTA DE TABLAS

Tabla 1. Estructura general del dataset	24
Tabla 2. Columnas por BU	27
Tabla 3. Portafolio de productos por BU	28
Tabla 4. Distribución General por BU	29
Tabla 5. Subproductos por BU	30
Tabla 6. Caracterización inicial por BU	33
Tabla 7. Caracterización con tres variables por BU.	33
Tabla 8. Hiperparámetros	37
Tabla 9. Bases de datos utilizadas en el proyecto	38
Tabla 10. Configuración k=2.....	41
Tabla 11. Configuración k=4.....	42
Tabla 12. Configuración con k = 8	44
Tabla 13. Comparación técnica y estratégica de las configuraciones	45
Tabla 14. Distribución de clientes y principales BU por cluster (k = 4)	46
Tabla 15. Distribución de clientes y BU's predominantes (k = 8)	47
Tabla 16. Resultados consolidados del clustering por Unidad de Negocio — métricas escaladas a [0–1].....	55
Tabla 17. Top 20 clientes con mayor potencial de cross-selling según score de propensión agregado	67
Tabla 18. Resultados del proceso de depuración de la base de prospectos externos	79
Tabla 19. Distribución de clases y tamaños de muestra por Unidad de Negocio	82
Tabla 20. Comparación de AUC-ROC por modelo y Unidad de Negocio (Prospectos Externos)	88
Tabla 21. Rangos de scores por componente y Unidad de Negocio (572 prospectos externos)	92
Tabla 22. Top 10 prospectos — Private Cloud / Infra Tradicional (Score_Final híbrido)	92
Tabla 23. Top 10 prospectos — Digital Workplace (Score_Final híbrido)	93
Tabla 24. Top 10 prospectos — Cyber Security (Score_Final híbrido)	93
Tabla 25. Top 10 prospectos — Digital Business / Aplicaciones (Score_Final híbrido)	94
Tabla 26. Top 10 prospectos — Cloud Pública (Score_Final híbrido)	94
Tabla 27. Top 10 prospectos — Soluciones SAP (Score_Final híbrido)	95
Tabla 28. Ranking global de prospectos externos — Top 20 por Score_Total_Final	96
Tabla 29. Resultados comparativos de AUC-ROC por modelo y BU (Cross-Selling)	109
Tabla 30. Validación cruzada repetida de los modelos Random Forest con mayor AUC en cross-selling	110

LISTA DE ESQUEMAS

Esquema 1. Metodología del Objetivo 1: caracterización de los clientes actuales	18
Esquema 2. Metodología del Objetivo 2: modelado no supervisado y supervisado	37
Esquema 3. Metodología del Objetivo 3: evaluación del desempeño de los modelos	103

LISTA DE ANEXOS

ANEXO A. REPOSITORIO DEL PROYECTO	128
ANEXO B. PROMPT PARA CONSTRUCCIÓN DE LA BASE DE PROSPECTOS EXTERNOS	129

INTRODUCCIÓN

La filial en Colombia de una multinacional del sector tecnológico enfrentaba al inicio de este proyecto un problema recurrente en la gestión estratégica de sus datos comerciales: la ausencia de una segmentación formal de clientes y la falta de herramientas analíticas para priorizar oportunidades. Esta situación dificultaba la toma de decisiones informadas, limitaba la personalización de las estrategias comerciales y reducía la efectividad del pipeline de ventas, en un contexto donde el sistema CRM Salesforce concentraba un volumen significativo de información histórica sin un aprovechamiento sistemático.

Frente a esta realidad, el presente trabajo aplicó técnicas de ciencia de datos para transformar la información del CRM en conocimiento accionable. Se desarrolló un sistema analítico de segmentación y prospección de clientes apoyado en algoritmos de aprendizaje automático: en una primera fase se realizó la caracterización de los 167 clientes activos del período 2018–2025 mediante técnicas de clustering (K-Means) por Unidad de Negocio, y en una segunda fase se entrenaron modelos supervisados (Random Forest, XGBoost y Regresión Logística) junto con un componente de similitud coseno, integrados en un score híbrido que se aplicó sobre 572 prospectos externos para priorizar oportunidades comerciales con mayor afinidad al perfil de los clientes existentes.

El proyecto, de carácter aplicado, priorizó el diseño e implementación de soluciones alineadas a las necesidades reales del negocio. Como resultado, se obtuvieron rankings priorizados de prospectos por BU y un ranking agregado global, así como una herramienta de cross-selling sobre la base instalada que el equipo comercial puede utilizar directamente. El sistema desarrollado quedó documentado de forma reproducible y sentó las bases para una cultura organizacional orientada al uso estratégico de la información, con extensiones futuras hacia campañas de marketing dirigido, análisis de mercado y diseño de portafolios de servicios personalizados.

1 CONTEXTUALIZACIÓN DEL PROYECTO

1.1 DEFINICIÓN DEL PROBLEMA

1.1.1 PLANTEAMIENTO DEL PROBLEMA

La filial Colombia de una empresa especializada en servicios de tecnología enfrenta actualmente importantes desafíos en la gestión de sus clientes y en la identificación de nuevas oportunidades comerciales. Aunque la organización dispone de un CRM robusto como Salesforce, que almacena información clave sobre cuentas, oportunidades y actividades comerciales, estos datos no han sido utilizados de forma analítica ni estratégica para comprender patrones de comportamiento ni para establecer perfiles diferenciados de clientes.

En la operación actual, los equipos comerciales tratan a todos los clientes de forma homogénea, sin diferenciar entre industrias, líneas de servicio, niveles de madurez digital ni potencial de crecimiento. Este enfoque uniforme limita la posibilidad de personalizar ofertas, diseñar estrategias de fidelización efectivas, identificar oportunidades de cross-selling o upselling y priorizar adecuadamente a los clientes con mayor valor estratégico. Como resultado, se invierte tiempo y esfuerzo en cuentas que no representan oportunidades reales o relevantes para el negocio.

Un segundo aspecto crítico es la forma en que se generan nuevos prospectos comerciales. Actualmente, estas oportunidades provienen principalmente de áreas como marketing o referidos, sin que exista un mecanismo analítico que permita evaluar si dichos prospectos se alinean con los perfiles de clientes exitosos que ya atiende la empresa. Además, no se ha implementado un enfoque proactivo para buscar nuevas empresas en el mercado que compartan características con estos perfiles ideales. Esto genera un pipeline comercial con gran cantidad de prospectos poco calificados, lo que desgasta los recursos comerciales y técnicos en oportunidades con baja o nula posibilidad de cierre, disminuye la tasa de conversión y desvía el foco estratégico.

Desde la perspectiva de la ciencia de datos, esta problemática representa un desafío técnico que requiere un enfoque analítico riguroso basado en técnicas de modelado. Es fundamental estructurar y transformar los datos históricos del CRM para generar conocimiento útil que facilite la comprensión profunda del comportamiento de los clientes actuales y permita evaluar con objetividad nuevas oportunidades comerciales. En este contexto, el proyecto propone integrar dos enfoques complementarios de aprendizaje automático: uno no supervisado, orientado a la identificación de segmentos estratégicos dentro de la base de clientes actual; y otro supervisado, enfocado en la priorización de prospectos externos que presenten afinidad con los perfiles definidos. Esta articulación metodológica define el carácter del proyecto como una propuesta clara de ciencia de datos, alineada con los desafíos técnicos y estratégicos identificados.

En síntesis, el problema se manifiesta en tres frentes interrelacionados:

- La falta de segmentación efectiva de los clientes actuales, lo cual limita la capacidad de la organización para diseñar estrategias comerciales diferenciadas y alineadas con el valor estratégico de cada cliente.
- La ausencia de mecanismos predictivos objetivos que permitan identificar y priorizar nuevos prospectos con base en su afinidad con los perfiles más exitosos.
- La necesidad de estructurar un enfoque integral desde la ciencia de datos, que abarque desde el tratamiento y análisis de datos históricos hasta la evaluación sistemática de modelos, como soporte a decisiones comerciales más precisas y eficientes.

1.1.2 FORMULACIÓN DEL PROBLEMA

¿Cómo diseñar un modelo de ciencia de datos, basado en técnicas de aprendizaje automático, que permita segmentar los clientes actuales de una empresa de servicios tecnológicos a partir de datos del CRM Salesforce, e identificar y priorizar nuevos prospectos comerciales en fuentes externas, con base en su similitud con los perfiles exitosos previamente definidos?

1.1.3 SISTEMATIZACIÓN

1. ¿Cómo agrupar los clientes actuales mediante técnicas de segmentación que permitan caracterizar perfiles estratégicos a partir de sus datos históricos en Salesforce?
2. ¿Qué variables presentes en los datos del CRM resultan más útiles para diferenciar y caracterizar los perfiles estratégicos de clientes en el contexto de servicios tecnológicos?
3. ¿Cómo utilizar los perfiles de clientes exitosos para identificar y priorizar empresas externas con potencial de conversión, empleando técnicas predictivas o métricas de similitud?
4. ¿Qué métricas cuantitativas permiten validar la calidad de los segmentos y la efectividad del modelo de prospección predictiva, desde el punto de vista técnico y comercial?

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Diseñar un modelo de ciencia de datos, basado en técnicas de aprendizaje automático, que permita segmentar a los clientes actuales de una empresa de servicios tecnológicos utilizando datos del CRM Salesforce, con el fin de identificar y priorizar nuevos prospectos comerciales en fuentes externas que presenten similitud con los perfiles exitosos definidos.

2.2 OBJETIVOS ESPECÍFICOS

1. Caracterizar los datos de los clientes actuales utilizando la información del CRM Salesforce, con el fin de identificar variables relevantes que permitan construir perfiles estratégicos.
2. Desarrollar dos modelos de aprendizaje automático: uno no supervisado para segmentar a los clientes actuales y otro supervisado o basado en similitud para identificar y priorizar prospectos comerciales en fuentes externas.
3. Evaluar el desempeño de los modelos construidos mediante métricas cuantitativas y criterios de negocio, con el fin de apoyar decisiones comerciales basadas en datos.

3 MARCO TEÓRICO Y ANTECEDENTES

3.1 CONTEXTO ORGANIZACIONAL

La compañía es una multinacional tecnológica con sede en Brasil, presente en el mercado latinoamericano desde hace más de 20 años y operando en 10 países de la región. Su portafolio de servicios incluye soluciones en ciberseguridad, infraestructura, cloud, plataformas, Digital Workplace (mesas de ayuda), SAP, monitoreo, gestión de crisis, entre otras [1]. En Colombia, la operación inició formalmente con la adquisición de la empresa Synapsis en el período 2014–2015, consolidándose como un actor relevante en el sector de servicios tecnológicos [2].

El enfoque comercial de la compañía se estructura en torno a una red de ejecutivos comerciales responsables de las cuentas y un equipo horizontal de consultores preventa encargados del diseño de propuestas técnicas y económicas. Adicionalmente, el equipo de marketing participa en la generación de leads y la organización se ha fortalecido recientemente con la inclusión de unidades de negocio (BU) que integran perfiles como los Business Developer Managers (BDM), quienes trabajan junto con comerciales y preventas.

Salesforce es el CRM de la organización que centraliza la información de clientes y oportunidades. En él se documentan datos de las empresas, sus sectores, el ciclo de ventas de las oportunidades (desde calificación hasta cierre), y la trazabilidad de los leads generados por marketing. No obstante, la plataforma no ha sido explotada analíticamente para identificar patrones de comportamiento, evaluar la madurez de los clientes o establecer perfiles diferenciados por industria, servicio o volumen.

Actualmente, la compañía atiende principalmente a grandes y medianas empresas del sector público, privado, educativo y mixto. Entre sus clientes activos más destacados se encuentran organizaciones como Ecopetrol, Vanti, la Federación Nacional de Cafeteros, varias aseguradoras, EPS y universidades. Las soluciones ofrecidas, por su complejidad y costo, no son fácilmente adoptadas por pymes o emprendimientos. A pesar de contar con cerca de 60 clientes activos al año, se estima que el número acumulado en los últimos cinco años ha oscilado entre 140 y 170 clientes.

Una de las principales problemáticas comerciales es la ausencia de una segmentación efectiva de clientes, lo que impide aplicar estrategias diferenciadas y priorizar adecuadamente las ofertas. Como resultado, se invierte tiempo y esfuerzo en oportunidades con baja probabilidad de conversión, debido a la falta de criterios objetivos para calificar y filtrar prospectos. Esto impacta negativamente en la efectividad del pipeline comercial.

Adicionalmente, la prospección de nuevos clientes se lleva a cabo de forma reactiva, a través de referidos, campañas de marketing o iniciativas individuales de los ejecutivos, sin el respaldo de un modelo predictivo que permita identificar y priorizar empresas con alto potencial de conversión.

3.2 MARCO TEÓRICO

El presente proyecto se fundamenta en conceptos centrales de la ciencia de datos y el análisis comercial estratégico. A continuación, se presentan los pilares teóricos que sustentan la propuesta, con sus respectivas definiciones:

- **Aprendizaje automático**

El aprendizaje automático permite construir modelos a partir de datos para encontrar patrones útiles que mejoren la toma de decisiones en entornos reales [3]. Esta capacidad de aprendizaje basada en datos permite aplicar técnicas como clasificación, agrupamiento o recomendación a problemas complejos como la gestión de clientes. Asimismo, los algoritmos de aprendizaje automático pueden adaptarse a datos históricos para desarrollar sistemas predictivos aplicables a sectores como el comercial, financiero o tecnológico [4].

- **Segmentación de clientes**

La segmentación consiste en dividir un mercado en grupos distintos de compradores que pueden requerir productos o estrategias de marketing diferenciadas [6]. Esta práctica permite agrupar una base de clientes en subconjuntos homogéneos según características comunes, lo que facilita la personalización de estrategias de marketing, fidelización y ventas. Una segmentación efectiva permite entender mejor las necesidades de los clientes y diseñar ofertas más ajustadas a cada grupo [6].

En el contexto de la ciencia de datos, este proceso se realiza mediante algoritmos de aprendizaje no supervisado, especialmente K-Means, que asigna clientes a clústeres con base en similitud de atributos [3].

- **Prospección predictiva**

La prospección predictiva busca identificar posibles clientes potenciales a partir de patrones extraídos de clientes existentes. En el entorno B2B, esto implica comparar nuevos registros con los perfiles históricos de clientes exitosos. Esta tarea puede abordarse mediante modelos supervisados como Random Forest o mediante técnicas de similitud. Estas estrategias permiten priorizar prospectos con alta probabilidad de conversión, mejorando el rendimiento comercial [5].

- **Enfoques del aprendizaje automático en ciencia de datos**

El aprendizaje automático es una rama de la inteligencia artificial que permite a los sistemas aprender patrones a partir de datos y hacer predicciones o tomar decisiones sin programación explícita [7]. Existen dos enfoques principales:

- **Aprendizaje supervisado:** Se entrena un modelo con datos etiquetados para aprender la relación entre variables independientes y una variable objetivo. Técnicas comunes incluyen regresión logística, árboles de decisión, Random Forest y redes neuronales [4]. En este proyecto, se utiliza para predecir qué prospectos externos son más similares a los clientes exitosos, facilitando su priorización.
- **Aprendizaje no supervisado:** Se analiza información sin etiquetas conocidas con el fin de descubrir estructuras ocultas en los datos, como la agrupación de clientes. Entre las técnicas utilizadas se encuentran:
 - K-Means, que agrupa datos minimizando la varianza interna de cada grupo [3].
 - DBSCAN, que detecta agrupaciones con base en densidades locales.
 - Gaussian Mixture Models (GMM), que asume distribuciones gaussianas y asigna pertenencia probabilística.

Estas técnicas permiten identificar patrones de comportamiento y construir perfiles estratégicos de clientes.

Métricas de evaluación

Para validar los modelos se utilizan distintas métricas, tanto para segmentación como para predicción:

- **Modelos de clustering:** Se emplea el Silhouette Score, que mide la cohesión de un objeto respecto a su grupo comparado con otros. Valores cercanos a 1 indican una buena separación [3].
- **Modelos supervisados:** Se utilizan:
 - Precisión: proporción de verdaderos positivos sobre los predichos como positivos.
 - Recall: proporción de verdaderos positivos sobre los casos positivos reales.
 - F1-Score: media armónica entre precisión y recall, útil cuando se busca equilibrio entre ambos.
 - Métricas de similitud como distancia euclidiana o coseno, aplicables en modelos basados en perfiles.

Estas métricas permiten evaluar tanto el desempeño técnico como el valor práctico de los modelos en decisiones comerciales. Además, un modelo útil suele presentar rendimientos entre el 80% y el 90% en datos de prueba.

Fundamentos Matemáticos de los Modelos Utilizados

Esta sección presenta el fundamento matemático de cada uno de los algoritmos de aprendizaje

automático empleados en el proyecto, con el propósito de sustentar formalmente las decisiones metodológicas adoptadas y proveer al lector el marco teórico necesario para comprender e interpretar los resultados. Para cada modelo se describe su función objetivo, sus principales componentes matemáticos y las métricas de evaluación asociadas, con sus respectivas referencias bibliográficas.

1. K-Means Clustering

K-Means es un algoritmo de agrupamiento no supervisado propuesto por MacQueen [11] y formalizado computacionalmente por Lloyd [12]. Su objetivo es particionar un conjunto de n observaciones en K grupos (clústeres), de modo que cada observación pertenezca al grupo cuyo centroide es el más cercano. El algoritmo minimiza la inercia intra-clúster, definida como la suma de las distancias cuadradas de cada punto a su centroide asignado.

Trabajos comparativos como el de Arunachalam y Kumar [20] han evaluado el desempeño de aproximaciones alternativas (clustering jerárquico, K-Medoids, fuzzy clustering y mapas auto-organizativos), reportando ventajas particulares de SOM y fuzzy clustering en datasets ordinales de gran cardinalidad. Sin embargo, en aplicaciones B2B con muestras pequeñas como la presente —167 clientes—, K-Means mantiene su preferencia por la interpretabilidad directa de los centroides, su eficiencia computacional y la disponibilidad de métricas de calidad bien establecidas (Silhouette, Calinski-Harabasz, Davies-Bouldin), criterios que han sido validados en literatura reciente sobre segmentación de mercados industriales [21].

Función objetivo (inercia):

$$J(C_1, \dots, C_K) = \sum_{i=1}^K \sum_{x \in C_i} \|x - \mu_i\|^2$$

donde:

- K es el número de clústeres,
- C_i es el conjunto de puntos asignados al clúster i ,
- x es un vector de características de una observación,
- μ_i es el centroide (media vectorial) del clúster i ,
- $\|\cdot\|$ denota la norma euclidiana.

El centroide de cada clúster se actualiza iterativamente como la media aritmética de todos los puntos asignados a ese grupo:

$$\mu_i = (1/|C_i|) \times \sum_{x \in C_i} x$$

El algoritmo alterna entre dos pasos hasta converger: (1) asignación de cada observación al clúster con el centroide más cercano, y (2) recalcu de los centroides. Si bien la convergencia al mínimo global no está garantizada, en la práctica se obtienen soluciones satisfactorias mediante múltiples inicializaciones aleatorias con semillas distintas. En este proyecto se emplearon entre 10 y 30 inicializaciones (n_{init}) por configuración.

Preprocesamiento requerido: dado que K-Means utiliza distancias euclidianas, el algoritmo es sensible a la escala de las variables. Por esta razón se aplicó transformación logarítmica ($\log_{10}$) y estandarización con RobustScaler, que escala las variables utilizando la mediana y el rango intercuartílico (IQR), reduciendo la influencia de valores atípicos:

$$x_{no}^{rm} = (x - mediana) / IQR$$

Métricas de validación interna del clustering

índice de Silueta (Silhouette Score) [16]: mide simultáneamente la cohesión interna de un clúster y su separación respecto a los demás. Para cada punto i se define:

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}$$

donde $a(i)$ es la distancia media del punto i a todos los demás puntos de su mismo clúster (cohesión), y $b(i)$ es la distancia media mínima del punto i a los puntos del clúster más próximo (separación). El índice global S es el promedio de $s(i)$ sobre todos los puntos. Valores cercanos a +1 indican clústeres bien separados y compactos; valores cercanos a 0 indican solapamiento; valores negativos indican asignación incorrecta.

índice de Calinski-Harabasz (CH) [17]: evalúa la razón entre la dispersión inter-clúster y la dispersión intra-clúster. Se define como:

$$CH(K) = [SS_B / (K-1)] / [SS_W / (n-K)]$$

donde SS_B es la suma de cuadrados entre grupos (traza de la matriz de dispersión entre clústeres), SS_W es la suma de cuadrados dentro de los grupos, n es el número de observaciones y K el número de clústeres. Valores más altos de CH indican mejor segmentación.

índice de Davies-Bouldin (DB) [18]: mide la similitud promedio entre cada clúster y el que más se le parece. Se define como:

$$DB = (1/K) \times \sum_{i=1}^K \max_{j \neq i} \{ (s_i + s_j) / d_{ij} \}$$

donde s_i es la dispersión promedio del clúster i , s_j la dispersión del clúster j , y d_{ij} la distancia entre los centroides de los clústeres i y j . Valores más bajos de DB indican mejor separación entre clústeres.

Estas tres métricas fueron empleadas en el Procesamiento 4 (segmentación avanzada por BU) para seleccionar el número óptimo de clústeres en cada unidad de negocio, tal como se presenta en la Tabla 15 (Resultados consolidados del clustering por unidad de negocio).

2. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad que proyecta los datos en un espacio de menor dimensión preservando la mayor parte de la varianza original [15]. En este proyecto, PCA se utilizó exclusivamente para la visualización bidimensional de los clústeres obtenidos, facilitando la interpretación cualitativa de los resultados del K-Means.

Sea X una matriz de datos de dimensión $n \times p$ (n observaciones, p variables). PCA calcula la descomposición espectral de la matriz de covarianza:

$$\Sigma = (1/(n-1)) X^T X = V \Lambda V^T$$

donde V es la matriz ortogonal de vectores propios (componentes principales) y Λ es la matriz diagonal de valores propios (eigenvalues), ordenados de mayor a menor: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. La proyección de los datos sobre las primeras k componentes principales es:

$$Z = X V_k$$

donde V_k contiene las k columnas de V correspondientes a los k mayores valores propios. La varianza explicada acumulada por las k primeras componentes se calcula como:

$$VE(k) = \sum_{i=1}^k \lambda_i / \sum_{i=1}^p \lambda_i$$

En este proyecto se utilizaron $k = 2$ componentes principales para la visualización, lo que permite representar los clústeres en un plano cartesiano e identificar visualmente el grado de solapamiento o separación entre grupos. Esta visualización fue fundamental en el Procesamiento 3 para identificar casos de sobre-segmentación con siete clústeres, donde el PCA evidenció un alto solapamiento entre grupos.

3. Regresión Logística

La regresión logística es un modelo lineal generalizado diseñado para estimar la probabilidad de pertenencia a una clase binaria [4]. A diferencia de la regresión lineal, transforma la combinación lineal de las variables predictoras mediante la función sigmoide (logística), garantizando que la salida del modelo se encuentre en el rango $[0, 1]$:

$$P(Y=1 | X) = \sigma(\beta^T X) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)})$$

donde β_0 es el término de intercepción, β_i son los coeficientes asociados a cada variable predictora x_i , y $\sigma(\cdot)$ es la función sigmoide. La región de decisión se obtiene fijando un umbral (generalmente 0.5): si $P(Y=1 | X) \geq 0.5$ se predice la clase positiva.

Estimación de parámetros por máxima verosimilitud

Los coeficientes β se estiman maximizando la log-verosimilitud del modelo sobre los datos de entrenamiento:

$$\ell(\beta) = \sum_i [y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)]$$

donde $y_i \in \{0, 1\}$ es la etiqueta real y $\hat{y}_i = P(Y=1 | X_i)$ es la probabilidad predicha. Este problema de optimización no tiene solución analítica cerrada, por lo que se resuelve numéricamente mediante descenso de gradiente (o variantes como L-BFGS).

Regularización L2 (Ridge)

Para evitar el sobreajuste y mejorar la generalización, especialmente con datasets de alta dimensionalidad como el utilizado en este proyecto (producto del One-Hot Encoding), se incorpora una penalización L2 sobre los coeficientes en la función objetivo:

$$Obj(\beta) = -\ell(\beta) + \lambda \times \sum_i \beta_i^2 = -\ell(\beta) + \lambda ||\beta||^2$$

donde $\lambda > 0$ es el parámetro de regularización que controla el compromiso entre ajuste a los datos y penalización de la complejidad del modelo. La regularización L2 contrae los coeficientes hacia

cero sin anularlos, preservando todas las variables en el modelo. En scikit-learn, este parámetro se controla mediante $C = 1/\lambda$, donde valores pequeños de C implican mayor regularización.

Interpretación: los coeficientes β_i del modelo representan el log-odds (log del cociente de probabilidades) por unidad de cambio en x_i , lo que proporciona una interpretación directa de la influencia de cada variable firmográfica en la probabilidad de consumo de una BU. Variables con coeficientes positivos grandes aumentan la probabilidad de adopción; variables con coeficientes negativos la reducen.

4. Random Forest

Random Forest es un método de conjunto (ensemble) propuesto por Breiman [13] que combina múltiples árboles de decisión independientes para producir predicciones más robustas y con menor varianza que un único árbol. El método se fundamenta en dos principios: Bootstrap Aggregating (Bagging) y aleatorización de características.

Construcción del árbol base: cada árbol de decisión divide el espacio de características recursivamente seleccionando la variable y el umbral que maximizan la ganancia de información. Para problemas de clasificación se utiliza la impureza de Gini como criterio de división:

$$Gini(t) = 1 - \sum_i p_i(t)^2$$

donde $p_i(t)$ es la proporción de observaciones de la clase i en el nodo t . La división se selecciona minimizando la impureza ponderada de los nodos hijo:

$$\Delta Gini = Gini(t) - [(n_L/n) \times Gini(t_L) + (n_R/n) \times Gini(t_R)]$$

donde n_L y n_R son los números de observaciones en los nodos izquierdo y derecho, respectivamente.

Principio de Bagging:

Para cada árbol T^b , se genera una muestra bootstrap D^b de tamaño n con reemplazamiento a partir del conjunto de entrenamiento D . Adicionalmente, en cada nodo se selecciona aleatoriamente un subconjunto de m variables (donde $m \ll p$) para buscar la mejor división, lo que decorrelaciona los árboles y reduce la varianza del conjunto:

$$\hat{f}_{RF}(x) = (1/B) \times \sum_{b=1}^B T^b(x; D^b)$$

donde B es el número de árboles y $T^b(x; D^b)$ es la predicción del árbol b . Para clasificación, la predicción final se obtiene por votación mayoritaria:

$$\hat{y}(x) = \operatorname{argmax}_c \sum_b \mathbb{I}(T^b(x) = c)$$

Ponderación de clases (class_weight = 'balanced_subsample'):

En presencia de desbalance de clases, Random Forest puede sesgarse hacia la clase mayoritaria. El parámetro `balanced_subsample` ajusta los pesos de clase dentro de cada bootstrap, de modo que la clase minoritaria (clientes que consumen la BU) recibe mayor peso en la función de pérdida:

$$wc = n / (K \times nc)$$

donde n es el número total de muestras, K el número de clases (2 en este caso) y nc el número de

muestras de la clase c. Esto hace que los errores en la clase minoritaria sean penalizados proporcionalmente más, mejorando el recall sobre clientes que consumen la BU.

Esta estrategia de ponderación interna se eligió en lugar de técnicas de remuestreo como SMOTE o ADASYN, dado que la literatura especializada en clasificación con clases desbalanceadas reporta que el oversampling sintético puede introducir ruido y artefactos en datasets de tamaño moderado, mientras que la ponderación de clases en algoritmos basados en árboles preserva la estructura original de la muestra y produce resultados más estables [22].

Importancia de variables:

Una ventaja adicional de Random Forest es su capacidad para estimar la importancia de cada variable predictora mediante el promedio de la reducción de impureza de Gini que produce cada variable a lo largo de todos los árboles del bosque. Esta información puede ser utilizada para identificar qué atributos firmográficos son más discriminativos para predecir la adopción de cada BU.

5. XGBoost (Extreme Gradient Boosting)

XGBoost, propuesto por Chen y Guestrin [14], es un algoritmo de boosting por gradiente que construye un modelo aditivo secuencial, donde cada nuevo árbol corrige los errores residuales del modelo actual. Formalmente, el modelo es una suma de K funciones (árboles de decisión):

$$\hat{y}(x) = \sum_{k=1}^K f^k(x), \text{ con } f^k \in \mathcal{F}$$

donde $\mathcal{F}$ es el espacio de árboles de regresión con estructura fija, y cada f^k es el q-ésimo árbol.

Función objetivo con regularización:

La función objetivo de XGBoost incorpora explícitamente un término de regularización que controla la complejidad de los árboles, a diferencia del Gradient Boosting clásico:

$$Obj(\theta) = \sum_i \ell(y_i, \hat{y}_i) + \sum_k \Omega(f^k)$$

donde $\ell(\cdot)$ es la función de pérdida (log-loss para clasificación binaria) y $\Omega(f)$ es el término de regularización del árbol:

$$\Omega(f) = \gamma T + (1/2) \lambda ||w||^2$$

donde T es el número de hojas del árbol, w es el vector de puntuaciones (pesos) de las hojas, γ penaliza la creación de nuevas hojas (podando el árbol) y λ es la penalización L2 sobre los pesos de las hojas.

Aprendizaje por descenso de gradiente de segundo orden:

XGBoost utiliza una aproximación de Taylor de segundo orden de la función de pérdida para determinar la estructura óptima de cada nuevo árbol, lo que lo hace más eficiente que los métodos de primer orden:

$$Obj^{Tn}(q) \approx \sum_i \ell''(y_i) [g_i w^\psi + (1/2)(h_i + \lambda) w^{\psi^2}] + \gamma T$$

donde I^j es el conjunto de instancias en la hoja j , $g_i = \partial \ell / \partial \hat{y}^{n-1}$ y $h_i = \partial^2 \ell / \partial (\hat{y}^{n-1})^2$ son los gradientes de primer y segundo orden de la pérdida respecto a la predicción del paso anterior.

Hiperparámetros clave:

- **learning_rate ($\eta = 0.05$):** escala la contribución de cada árbol: $\hat{y}^n = \hat{y}^{n-1} + \eta \times f^n(x)$. Valores bajos requieren más árboles pero reducen el sobreajuste.
- **max_depth = 4:** limita la profundidad máxima de cada árbol, controlando la complejidad del modelo.
- **subsample = 0.9:** fracción de observaciones utilizadas en cada iteración de boosting, introduciendo estocasticidad.
- **colsample_bytree = 0.9:** fracción de variables seleccionadas para cada árbol, similar a la aleatorización de Random Forest.

El uso de regularización L2 nativa, el aprendizaje de segundo orden y la paralelización eficiente de la construcción de árboles hacen de XGBoost uno de los algoritmos de mayor desempeño en competiciones de ciencia de datos con datos tabulares.

6. Similitud por Coseno

La similitud por coseno es una métrica de similitud entre dos vectores en un espacio n -dimensional, definida como el coseno del ángulo que forman. A diferencia de la distancia euclidiana, la similitud coseno evalúa la orientación relativa de los vectores, independientemente de su magnitud [19]. Esta propiedad es especialmente útil cuando se comparan perfiles de empresas con distintos niveles de completitud de información.

Para dos vectores A y B en $\mathbb{R}^n$, la similitud coseno se define como:

$$\cos(A, B) = (A \cdot B) / (||A|| \times ||B||) = \sum_i A_i B_i / (\sqrt{\sum_i A_i^2} \times \sqrt{\sum_i B_i^2})$$

donde $A \cdot B$ es el producto punto, $||A||$ y $||B||$ son las normas euclidianas de los vectores.

Propiedades:

- **Rango:** $\cos(A, B) \in [-1, +1]$, donde $+1$ indica dirección idéntica, 0 indica ortogonalidad y -1 indica dirección opuesta.
- **Invariancia a la escala:** $\cos(\lambda A, B) = \cos(A, B)$ para cualquier escalar $\lambda > 0$, lo que permite comparar vectores de diferente magnitud.

Aplicación en este proyecto:

Para cada Unidad de Negocio, se calcula el centroide del cluster estratégico (aquel con mayor ACV promedio) en el espacio de características escaladas:

$$\mu^* = (1/|C^*|) \times \sum_{x \in C^*} \phi(x)$$

donde $\phi(x)$ es el vector de características de un cliente tras la transformación y escalamiento, y C^* es el clúster estratégico. El score de afinidad de un prospecto externo z respecto a esa BU se define como:

$$Score_Afinidad_BU(z) = \cos(\varphi(z), \mu^*) = (\varphi(z) \cdot \mu^*) / (|| \varphi(z) || \times || \mu^* ||)$$

Este score tiene una interpretación geométrica clara: mide qué tan alineado está el vector del prospecto con el perfil promedio de los clientes más valiosos de cada BU en el espacio firmográfico. Los prospectos con mayor similitud coseno son aquellos cuya estructura organizacional más se asemeja a la del segmento estratégico, independientemente de su tamaño absoluto.

7. Métrica AUC-ROC para Modelos Supervisados

El Área Bajo la Curva ROC (AUC-ROC) es la métrica principal de evaluación para los modelos supervisados de este proyecto. La curva ROC (Receiver Operating Characteristic) traza la Tasa de Verdaderos Positivos (TPR o Recall) en función de la Tasa de Falsos Positivos (FPR) a medida que se varía el umbral de clasificación $\tau \in [0, 1]$:

$$TPR(\tau) = VP(\tau) / (VP(\tau) + FN(\tau)) = Recall(\tau)$$

$$FPR(\tau) = FP(\tau) / (FP(\tau) + VN(\tau)) = 1 - Especificidad(\tau)$$

donde VP = verdaderos positivos, FN = falsos negativos, FP = falsos positivos, VN = verdaderos negativos.

El AUC es el área bajo esta curva, calculada mediante la regla del trapecio. Su interpretación estadística es la probabilidad de que el modelo asigne un score más alto a una instancia positiva elegida al azar que a una instancia negativa elegida al azar:

$$AUC = P(score(x+) > score(x-))$$

Esta propiedad hace al AUC invariante al desbalance de clases, convirtiéndola en la métrica más adecuada para los modelos de este proyecto, donde las clases presentan alta asimetría (especialmente SAP: 10.8% positivos). Valores de $AUC \geq 0.70$ son considerados aceptables; ≥ 0.80 buenos; ≥ 0.90 excelentes para aplicaciones comerciales.

Complementariamente, se utilizan las siguientes métricas para caracterizar el desempeño en un umbral fijo ($\tau = 0.5$):

$$Precisión = VP / (VP + FP)$$

$$Recall = VP / (VP + FN)$$

$$F1 = 2 \times (Precisión \times Recall) / (Precisión + Recall)$$

En el contexto de prospección comercial, el Recall es particularmente relevante, ya que mide la proporción de prospectos genuinamente afines que el modelo logra identificar. Un recall alto minimiza el riesgo de omitir oportunidades de negocio valiosas.

8. Score Híbrido de Priorización de Prospectos

El score híbrido integra los resultados del modelo de similitud coseno y del mejor modelo

supervisado en un único indicador de priorización comercial por prospecto y BU. Su formulación es un promedio ponderado con pesos iguales:

$$Score_Final_BU(z) = \alpha \times Score_Afinidad_BU(z) + (1-\alpha) \times Score_Supervisado_BU(z)$$

con $\alpha = 0.5$. La elección de pesos iguales refleja un supuesto de equidistancia entre la perspectiva geométrica (similitud de perfil organizacional) y la perspectiva estadística (probabilidad de adopción aprendida de los patrones históricos). Esta ponderación puede ser ajustada en versiones futuras del modelo a medida que se acumule evidencia empírica sobre la efectividad relativa de cada componente.

Esta arquitectura sigue el principio de los sistemas de recomendación híbridos formalizado por Burke [23], quien demostró que la combinación de aproximaciones content-based (en este caso, similitud coseno sobre variables firmográficas) con aproximaciones supervisadas (clasificación basada en patrones históricos) produce sistemas más robustos que cualquiera de los componentes por separado, particularmente en escenarios cold-start donde los nuevos elementos —en nuestro caso, prospectos externos sin historial transaccional— carecen de evidencia comportamental directa.

El ranking global de prospectos se construye sumando los scores híbridos a través de todas las BUs:

$$Score_Total(z) = \sum_{BU} Score_Final_BU(z)$$

y ordenando los prospectos de mayor a menor $Score_Total$. La BU de mayor afinidad por prospecto se determina como:

$$BU^* = \operatorname{argmax}_{BU} Score_Final_BU(z)$$

Esta información orienta al equipo comercial sobre la línea del portafolio más adecuada para iniciar la conversación con cada prospecto priorizado.

3.3 ANTECEDENTES

Se presentan a continuación tres trabajos previos que abordan problemáticas similares o que, aunque en contextos distintos, emplean técnicas que pueden aplicarse al proyecto actual:

Título: Factores que inciden en el “Churn” de una empresa de telecomunicaciones: Un análisis basado en Machine Learning

Institución: Universidad de Manizales

Año: 2023

Enlace: <https://ridum.umanizales.edu.co/bitstreams/64524b76-39a7-47cf-8411-e4f4e51b63df/download>

Este estudio aborda la problemática de abandono de clientes (churn) en una empresa de telecomunicaciones de Caldas. A través del análisis de datos históricos, se entrenaron varios modelos de aprendizaje automático, entre ellos Random Forest, regresión logística y XGBoost,

para predecir la probabilidad de que un cliente abandone. El modelo permite anticipar estos casos con suficiente antelación y plantea estrategias específicas para la retención de clientes [8].

Aporte: Este trabajo valida la utilidad del aprendizaje supervisado para analizar patrones de comportamiento del cliente y anticipar decisiones de compra o abandono.

Diferencias: Se enfoca en una problemática de retención, y no de prospección, las técnicas utilizadas son útiles para la identificación de prospectos potenciales en el proyecto actual.

Metodología y hallazgos: el estudio aplicó técnicas de Machine Learning en Python —incluyendo regresión logística, árboles de decisión y Random Forest— sobre datos de clientes de una empresa de telecomunicaciones en los municipios de Manizales, Villamaría y Chinchiná (Caldas). El análisis identificó los factores de mayor influencia en el abandono del servicio y permitió proponer estrategias específicas de retención diferenciadas por perfil de cliente. El trabajo valida la aplicabilidad del aprendizaje supervisado para anticipar comportamientos de abandono a partir de datos históricos del CRM, enfoque directamente transferible al presente proyecto en el análisis de patrones de consumo de servicios tecnológicos.

Título: Desarrollo de un modelo analítico para la segmentación de clientes en una cooperativa financiera

Institución: Universidad Nacional de Colombia

Año: 2020

Enlace: https://repositorio.unal.edu.co/bitstream/handle/unal/79648/98644734_2020.pdf

Este estudio tuvo como objetivo desarrollar un modelo de segmentación para una cooperativa financiera colombiana. Utilizando técnicas de minería de datos y métodos de agrupamiento no supervisado, se clasificaron los clientes en grupos homogéneos según su comportamiento financiero y sociodemográfico. Esto permitió personalizar estrategias de fidelización y diseñar campañas más efectivas [9].

Aporte: Sirve como base para el componente de segmentación del presente proyecto, en especial por su aplicación práctica de clustering con datos reales.

Diferencias: Está enfocado en servicios financieros y no contempla la segunda fase de prospección predictiva que caracteriza el presente estudio.

Metodología y hallazgos: el estudio aplicó las técnicas K-Means, DBSCAN y PAM (Partitioning Around Medoids) sobre variables RFM (Recencia, Frecuencia y Valor Monetario) de los asociados activos de la cooperativa, siguiendo la metodología CRISP-DM. Se realizaron tres segmentaciones independientes: asociados ahorradores, asociados con créditos activos, y asociados con ahorros y créditos simultáneos. En cada caso se compararon los algoritmos mediante métricas de validación interna (Silhouette, Calinski-Harabasz y Davies-Bouldin) y el índice Adjusted Rand Index para evaluar la estabilidad de los agrupamientos. Los resultados permitieron identificar perfiles

diferenciados de asociados según su comportamiento financiero y transaccional, y el producto final fue un dashboard interactivo en R Shiny que facilita el monitoreo continuo de los segmentos obtenidos.

Título: Modelo De Predicción De Deserción Estudiantil, Apoyado En Tecnologías De Data Mining, En Un Curso De Primera Matrícula De La Escuela ECBTI De La UNAD

Institución: Universidad Nacional Abierta y a Distancia (UNAD)

Año: 2021

Enlace: <https://repository.unad.edu.co/handle/10596/42544>

Este trabajo académico desarrolló un sistema de predicción del abandono estudiantil en una universidad virtual. Se utilizaron algoritmos de clasificación como árboles de decisión, redes neuronales y regresión logística sobre variables académicas, hábitos de estudio y contexto socioeconómico [10].

Aporte: Ofrece un ejemplo robusto del uso de modelos predictivos basados en características históricas para anticipar comportamientos futuros.

Diferencias: Aunque el contexto es educativo, el enfoque técnico de clasificación supervisada puede trasladarse al entorno comercial para la predicción de nuevos clientes con base en similitud a perfiles exitosos.

Metodología y hallazgos: el trabajo propone la aplicación de técnicas de minería de datos —en particular árboles de decisión— sobre la información académica de los estudiantes matriculados en el curso de primera matrícula de la Escuela ECBTI de la UNAD, con el fin de encontrar patrones que permitan predecir el riesgo de deserción temprana. El análisis identifica variables como la regularidad de acceso a la plataforma virtual, el porcentaje de actividades entregadas y el historial académico previo como los predictores más relevantes del abandono en el primer período. Aunque el trabajo se plantea como propuesta de investigación aplicada, su enfoque técnico de clasificación supervisada sobre comportamiento histórico es directamente trasladable al entorno comercial para la identificación de prospectos con base en similitud con perfiles exitosos.

En esa misma línea, Tabianan, Velu y Ravi [25] aplicaron K-Means para segmentar clientes de e-commerce a partir de variables de comportamiento de compra (event_type, categorías, productos), demostrando que la combinación de variables transaccionales con caracterización del cliente permite identificar segmentos de alta rentabilidad y orientar acciones comerciales diferenciadas. Esta aproximación coincide metodológicamente con la propuesta del presente trabajo, que integra variables transaccionales del CRM (opp_count, acv_total, years_present) con caracterización empresarial obtenida mediante enriquecimiento externo, validando la pertinencia del enfoque elegido para el contexto B2B de servicios tecnológicos.

4 CARACTERIZAR LOS DATOS DE LOS CLIENTES ACTUALES UTILIZANDO LA INFORMACIÓN DEL CRM SALESFORCE, CON EL FIN DE IDENTIFICAR VARIABLES RELEVANTES QUE PERMITAN CONSTRUIR PERFILES ESTRATÉGICOS.

El Esquema 1 resume el flujo metodológico seguido para este objetivo: a partir de la base de clientes de Salesforce, se aplicó un análisis exploratorio (sin modelo de aprendizaje automático) sobre variables categóricas y numéricas para obtener los arquetipos de cliente y las variables relevantes de penetración por unidad de negocio.

Esquema 1. Metodología del Objetivo 1: caracterización de los clientes actuales



4.1 ESTRUCTURACIÓN DE LA BASE DE DATOS

Fuentes y alcance de los datos

Para dar cumplimiento a este objetivo, se accedió a la plataforma Salesforce, desde donde se generaron informes anuales correspondientes al periodo 2018–2025, filtrando exclusivamente las oportunidades asociadas a clientes en Colombia. Dado que Salesforce establece un límite máximo de 2,000 registros por reporte, fue necesario generar un informe por cada año. En el caso del año 2025, el informe se construyó únicamente hasta el mes de agosto de 2025, correspondiente al corte de información disponible al momento de la extracción.

Cada reporte incluyó un conjunto homogéneo de variables relacionadas con las cuentas, las oportunidades y los productos, tales como: nombre de la cuenta, identificador de la oportunidad (TVT), propietario de la oportunidad, importe, precio total, fecha de cierre, nombre del producto, precio de venta, fecha del producto, unidad de negocio (BU), categoría del producto, clasificación, sector y segmento.

Los informes fueron exportados en formato CSV/Excel y almacenados en una carpeta denominada

TESIS en Google Drive, lo que permitió disponer de un repositorio centralizado para su tratamiento posterior. Es importante resaltar que, si bien Salesforce ofrece la posibilidad de conexión directa mediante API para la extracción de datos, en este caso no se contaba con los permisos de administrador necesarios para habilitar dicha integración. Por tal motivo, se optó por la alternativa de extracción manual a través de reportes.

Una vez consolidada la información en Google Drive, se desarrolló un proceso en Google Colab utilizando el lenguaje de programación Python y librerías especializadas para análisis de datos. El flujo implementado comprendió las siguientes etapas principales:

- **Carga y unificación de reportes:** los archivos correspondientes a cada año fueron detectados y leídos, agregando metadatos de control como el año de reporte y el nombre del archivo de origen. De esta forma, se construyó una base de datos unificada que concentró toda la información histórica.
- **Normalización de variables:** se estandarizaron los nombres de las cuentas, eliminando tildes, caracteres especiales y sufijos corporativos frecuentes (S.A.S., Ltda., S.A.), con el fin de reducir duplicidades por diferencias en la escritura. Asimismo, se homogeneizaron los nombres de columnas y se transformaron las variables de fecha al formato de tipo datetime, lo que permitió derivar atributos temporales (año, mes, trimestre).
- **Filtrado de oportunidades ganadas:** se depuró la base conservando únicamente las oportunidades cuyo estado correspondía a “Ganada” (Closed Won y equivalentes en español), asegurando así que el análisis se enfocara exclusivamente en los clientes efectivamente captados.
- **Construcción de indicadores estratégicos:** se generaron métricas agregadas por cliente que incluyen: número de años con actividad, última fecha de participación, número de oportunidades, monto total y valor de ventas acumuladas, así como el listado de productos adquiridos y su distribución por unidad de negocio. Estas métricas permiten caracterizar cada cuenta de forma integral y sirven como insumo para fases posteriores de segmentación.
- **Exportación del dataset final:** el resultado del procesamiento fue un archivo consolidado denominado clientes_resumen_ganadas.csv, almacenado en la carpeta de salidas de Drive, que constituye parte de la base analítica para los objetivos del proyecto.

Enriquecimiento externo

Con el fin de hacer la caracterización de los clientes más allá de lo disponible en Salesforce, se evaluó inicialmente la opción de web scraping directo desde Python, sin embargo, el resultado fue insatisfactorio: en la mayoría de los casos el proceso retornó enlaces y fragmentos poco estructurados en lugar de datos verificables y normalizados. Adicional, dado el volumen (167 clientes), el proceso manual habría sido altamente demandante y propenso a errores.

Para viabilizar la tarea, se adoptó una estrategia de investigación asistida con Gemini 2.5 Pro (modo investigación profunda), utilizando un prompt diseñado para guiar la búsqueda en fuentes confiables (sitios oficiales, superintendencias, LinkedIn corporativo, prensa económica como La República, entre otras) y consolidar la información en nuevas columnas. El enfoque priorizó trazabilidad y verificación, conservando la columna original de empresa para mantener el vínculo con la base de oportunidades ganadas. *“Quiero que actúes como un analista de datos especializado en inteligencia de negocios y analítica de clientes. Te daré un archivo en Excel con una lista de empresas (clientes). Tu tarea es buscar y verificar información confiable en la web (páginas oficiales de la empresa, superintendencias, LinkedIn, La República, etc.) y completar la base de datos agregando las siguientes columnas:*

Información básica

Región / País / Ciudad principal

Año de fundación

Tipo de empresa (pública, privada, multinacional, estatal, startup, etc.)

Sitio web oficial

Tamaño y capacidad

Número de empleados

Tamaño de la empresa (pequeña, mediana, grande, multinacional)

Número de sedes / oficinas

Presencia internacional (sí/no o número de países)

Sector e industria

Sector económico principal

Subindustria o línea de negocio

Clasificación CIIU/NAICS/ISIC (si está disponible)

Indicadores financieros

Ingresos anuales o facturación aproximada

Utilidades (si están disponibles)

Tasa de crecimiento anual (%)

Ingresos por empleado (si es posible calcular)

Aspectos estratégicos

Clientes objetivo (B2B, B2C, Mixto)

Modelo de negocio (Servicios, Manufactura, Distribución, SaaS, etc.)

Nivel de digitalización (bajo, medio, alto)

Innovación / I+D (si existe información pública)

La información debe estar lo más actualizada y verificada posible.

El resultado final debe ser un archivo Excel, manteniendo la columna de empresa original y agregando todas las columnas nuevas con la información encontrada.”

Aunque el prompt contemplaba un conjunto amplio de variables, la base caracterizada final consolidada en el archivo clientes_caracterizados.xlsx logró incluir de manera consistente únicamente las siguientes columnas:

- Empresa
- Región / País / Ciudad principal
- Año de fundación
- Tipo de empresa (pública, privada, multinacional, estatal, etc.)
- Sitio web oficial
- Número de empleados
- Tamaño de la empresa (pequeña, mediana, grande, multinacional)
- Presencia internacional (sí/no)
- Sector económico principal
- Modelo de negocio (Servicios, Manufactura, SaaS, etc.)

Estas variables fueron seleccionadas por ser las más estables, disponibles y verificables en fuentes públicas, y resultan suficientes para los análisis de segmentación y perfiles estratégicos en el marco de la investigación. Variables sensibles como ingresos, utilidades o clasificación CIU quedaron fuera debido a su baja disponibilidad pública y heterogeneidad en la muestra de empresas.

Durante la consolidación se identificó un conjunto de ≈ 30 empresas que no lograron vincularse automáticamente, principalmente por divergencias entre el nombre comercial y la razón social (p. ej., *Promotora de Café de Colombia* frente a su denominación comercial *Juan Valdez*). Estos registros quedaron marcados para curación manual mediante revisión puntual en cámaras de comercio, RUES/Superintendencias y páginas corporativas.

Adicionalmente, se evidenció que, en el caso de empresas pequeñas y medianas, la información sobre el número de empleados fue reportada en su mayoría en rangos aproximados (ejemplo: “11–50”, “51–200”) y que para un subconjunto relevante de compañías este campo aparece como “No disponible” debido a la falta de datos públicos verificables.

La base enriquecida se guardó como `clientes_caracterizados.xlsx` en la carpeta TESIS de Google Drive, sirviendo como uno de los archivos del repositorio del proyecto (ver Anexo A).

Integración de fuentes (merge) y salidas

Para integrar la caracterización externa con la tabla de oportunidades ganadas (`clientes_resumen_ganadas.csv`), se ejecutó un proceso de enlace (merge) con salvaguardas de calidad:

- Búsqueda robusta de archivos: se definieron múltiples rutas de búsqueda (subcarpetas habituales de *outputs_v2* y *TESIS*) para localizar los insumos sin depender de una sola ruta.
- Lectura de insumos:
 - `clientes_resumen_ganadas.csv` (base analítica depurada y consolidada).
 - `clientes_caracterizados.xlsx` (archivo enriquecido con Gemini 2.5 Pro).
- Normalización de claves: se creó una clave canónica a partir del nombre de empresa (minúsculas, remoción de tildes y caracteres especiales, colapso de espacios y limpieza de sufijos comunes, cuando aplica). La clave se generó de forma consistente en ambas tablas para maximizar el *matching* por texto.
- Diagnósticos previos: se calcularon duplicados por clave normalizada en ambas fuentes para prevenir uniones ambiguas y multiplicación de filas.
- Unión y chequeos posteriores: se realizó un left join desde la base de oportunidades ganadas hacia la base caracterizada, conservando todo el universo analítico. Se generó una lista de no vinculados (clientes con todos los campos enriquecidos nulos) para curación posterior.
- Entregables de consolidación:
 - `clientes_bda_caracterizada.csv`
 - `clientes_bda_caracterizada.xlsx`

Estos archivos constituyen el insumo analítico integral que integra el comportamiento transaccional (oportunidades ganadas ACV , Total Oportunidad) con los atributos externos de la organización (tamaño, sector, modelo de negocio, presencia).

Controles de calidad y trazabilidad

Para garantizar la validez y reproducibilidad del proceso, se establecieron los siguientes controles:

- Trazabilidad de fuentes: cuando fue posible, se preservaron URLs de referencia (sitio oficial, perfiles corporativos verificados).

- Homogeneización de formatos: las variables categóricas (p. ej., *tipo de empresa*, *tamaño*, *modelo de negocio*) se estandarizaron para evitar cardinalidades artificiales por variaciones de escritura.
- Reproducibilidad: el proceso está parametrizado y versionado en Colab (rutas, normalización, llaves), lo que permite su re-ejecución con nuevos cortes de información.

Depuración final y preparación del archivo para modelado

Se realizó un proceso de depuración final del dataset con el objetivo de conservar únicamente las variables relevantes para los análisis de segmentación y predicción. Para ello, se desarrolló un bloque de código en Python que permitió automatizar esta tarea y garantizar la trazabilidad de los archivos procesados.

En primer lugar, se definieron las rutas de trabajo hacia la carpeta principal del proyecto (TESIS/outputs_v2) en Google Drive, donde se almacenan los resultados intermedios y finales del procesamiento. Desde allí, se cargó el archivo `clientes_bda_caracterizada.csv`, el cual contenía la base enriquecida con los datos provenientes tanto del CRM Salesforce como de las fuentes externas consultadas a través de Gemini.

Posteriormente, se identificaron y eliminaron las columnas que no aportaban valor analítico para el modelado, específicamente: “Número de empleados”, “Sitio web oficial”, y “Año de fundación”.

Estas variables fueron descartadas debido a su baja completitud o limitada relevancia para los objetivos de segmentación. La eliminación se realizó de forma condicional, es decir, solo si dichas columnas estaban presentes en el archivo, evitando así errores de ejecución en futuras actualizaciones del proceso.

Una vez depurada la base, se generó un nuevo archivo denominado `clientes_bda_model.csv`, el cual representa el dataset definitivo para el desarrollo de los modelos de aprendizaje automático. Este archivo conserva las variables categóricas y numéricas necesarias para los algoritmos de segmentación, garantizando una estructura limpia, coherente y estandarizada.

4.2 CARACTERIZACIÓN DE LA BASE DE DATOS

Para el desarrollo del análisis correspondiente al Objetivo 1, se utilizó un conjunto de datos consolidado a partir de la información proveniente del sistema Salesforce CRM de la empresa.

El archivo denominado `clientes_bda_model.csv` contiene el histórico de oportunidades comerciales registradas entre los años 2018 y 2025, junto con los atributos organizacionales y contractuales de cada cliente activo en el periodo de estudio.

Este dataset constituye la fuente maestra sobre la cual se construyeron los procesos de exploración, caracterización y segmentación aplicados en las fases posteriores.

Estructura general del dataset

El archivo está conformado por 167 registros (clientes únicos) y 251 variables, que combinan información categórica (descriptiva) y numérica (cuantitativa).

Del total, 236 columnas son numéricas y 15 son categóricas o de texto.

Tabla 1. Estructura general del dataset

Tipo de variable	Cantidad de columnas	Porcentaje	Ejemplo de columnas
Numéricas (float64, int64)	236	94%	opp_count, acv_total, precio_total_sum, acv_sap_application_professional_services, total_public_cloud_managed_services
Categóricas (object)	15	6%	País, Ciudad, Sector económico principal, Tamaño de la empresa, Presencia internacional
Total general	251	100%	

Razón de la alta dimensionalidad (236 variables numéricas)

El elevado número de variables numéricas se explica por la estructura granular del modelo de datos extraído de Salesforce, que registra información independiente para cada subproducto o línea de servicio del portafolio de la empresa.

Cada producto o BU (Business Unit) se descompone en tres tipos de métricas cuantitativas principales, generando múltiples columnas por cada unidad de negocio.

1. Variables generadas por subproducto

Cada servicio o subproducto dentro del portafolio cuenta con tres tipos de variables asociadas:

- has_... → indicador binario que señala si el cliente posee ese servicio (1 = sí, 0 = no).
- acv_... → valor anual contractual (Annual Contract Value) asociado a ese subproducto.
- total_... → monto económico total facturado por ese subproducto.
- tvt_count_... → número total de transacciones o tickets vinculados a ese subproducto (Transaction Volume Total).

Por ejemplo, para el servicio SAP Application Professional Services, el dataset incluye:

- has_sap_application_professional_services

- acv_sap_application_professional_services
- total_sap_application_professional_services
- tvt_count_sap_application_professional_services

Esta misma estructura se replica para cada uno de los servicios que componen las siete Áreas de Portafolio (Private Cloud, Public Cloud, Digital Business, Cyber Security, SAP, Digital Workplace y Datos/IoT), lo que genera decenas de columnas por área.

A continuación, se describe brevemente el alcance comercial de cada Área de Portafolio:

Private Cloud / Infraestructura Tradicional: Comprende servicios de infraestructura propia (hosting privado, mainframe, conectividad, collocation y servicios gestionados) para clientes que requieren ambientes dedicados con altos niveles de control y soberanía sobre sus datos.

Cloud Pública (Public Cloud): gestión y reventa de servicios de nube pública sobre los principales hyperscalers (AWS, Azure, GCP e IBM) bajo modalidades de broker, plataforma gestionada y servicios profesionales para acompañar la migración y operación en la nube.

Digital Business / Aplicaciones: Desarrollo, modernización y gestión integral del ciclo de vida de aplicaciones (AMS, integración EDI, body-shop de desarrollo y QA). Habilita la transformación digital del cliente desde la capa aplicativa.

Cyber Security: Servicios de consultoría, monitoreo y operación de seguridad informática, incluyendo soluciones de software, servicios técnicos especializados y seguridad sobre ambientes en la nube.

Soluciones SAP: Implementación, migración a S/4HANA y soporte profesional sobre el ecosistema SAP. Atiende clientes corporativos cuyos procesos críticos se ejecutan sobre esta plataforma, con contratos de alto valor y largo ciclo.

Digital Workplace: Servicios orientados al puesto de trabajo digital, soporte al usuario final y mesas de ayuda. Tiene un papel transversal en la experiencia tecnológica del empleado dentro de las organizaciones cliente.

Datos / IoT / Smart Systems: Es una línea de especialización emergente, incluye soluciones de analítica avanzada, ingestión y explotación de datos, IoT industrial y sistemas inteligentes para clientes que adoptan capacidades de transformación digital de mayor madurez.

2. Variables de comportamiento y agregación

Además de las métricas por subproducto, se incluyen variables derivadas o agregadas que resumen el comportamiento comercial global de cada cliente:

- opp_count (cantidad de oportunidades comerciales).
- acv_total (suma total de valores anuales de contrato).
- precio_total_sum (valor contractual total).

- years_present (años de relación comercial).
- n_lineas_producto (número de líneas del portafolio adquiridas).

Estas variables son fundamentales para analizar la recurrencia, la longevidad de la relación comercial y el peso económico de cada cuenta.

3. Variables categóricas de caracterización

Las variables categóricas (15 en total) describen los atributos organizacionales y geográficos de cada cuenta, tales como:

- País y ciudad de operación.
- Tamaño de la empresa.
- Presencia internacional.
- Sector económico principal.
- Modelo de negocio y tipo de empresa.

Interpretación general

- Las 236 columnas numéricas no representan redundancia, sino la descomposición detallada del portafolio comercial de la empresa, que permite conocer la presencia, valor económico y frecuencia de uso de cada servicio por cliente.
- Esta granularidad facilita un análisis multidimensional del comportamiento de las empresas, permitiendo identificar patrones de adopción tecnológica y contribución económica por subproducto o familia de servicios.
- Las 15 variables categóricas complementan el análisis al definir el contexto empresarial, geográfico y sectorial de los clientes, proporcionando información clave para los modelos de segmentación posteriores.

En conjunto, esta base de datos consolida la relación entre características organizacionales y comportamiento comercial, y constituye el punto de partida para el Análisis Exploratorio de Datos (EDA) desarrollado en las siguientes secciones.

Distribución de variables numéricas por Área de Portafolio

Para facilitar la comprensión de la estructura interna del dataset, se realizó una clasificación de las 236 variables numéricas según su pertenencia a las Áreas de Portafolio (BU's) y el tipo de métrica que representan (has_, acv_, total_, tv_t_count_).

La siguiente tabla resume la cantidad aproximada de columnas por familia:

Tabla 2. Columnas por BU

Área de Portafolio (BU)	Indicadores binarios (has_)	Valores contractuales (acv_)	Montos totales (total_)	Volumen transaccional (tvt_count_)	Total por área
Private Cloud / Infraestructura Tradicional	20	20	20	20	≈80
Public Cloud	10	10	10	10	≈40
Digital Business / Aplicaciones	8	8	8	8	≈32
Cyber Security	8	8	8	8	≈32
Soluciones SAP	4	4	4	4	≈16
Digital Workplace	4	4	4	4	≈16
Datos / IoT / Smart Systems	3	3	3	3	≈12
Totales estimados	57	57	57	57	≈228

Nota: Las cifras son aproximadas, ya que algunas columnas se combinan o presentan valores nulos en ciertos subproductos, pero el patrón refleja la estructura general del dataset.

- La mayor parte de las variables proviene de las áreas Private Cloud / Infraestructura Tradicional y Public Cloud, que concentran más del 50 % del total de columnas.
- Cada subproducto del portafolio genera cuatro métricas asociadas (presencia, ACV, total y TVT), lo que explica la alta dimensionalidad del conjunto de datos.
- Las familias SAP, Workplace y Datos/IoT aportan un menor número de columnas, pero con valores económicos y estratégicos más concentrados, lo que será relevante en el análisis exploratorio posterior.

4.3 ANÁLISIS EXPLORATORIO DE DATOS (EDA) - CARACTERIZACIÓN POR ÁREAS DE PORTAFOLIO Y SUBPRODUCTOS

Como parte del análisis exploratorio del Objetivo 1, además de la caracterización general de clientes, se desarrolló una exploración enfocada en la relación entre las cuentas y las distintas líneas del portafolio (BU's), junto con los subproductos específicos adquiridos por cada cliente.

Este análisis permitió identificar los patrones de especialización y distribución comercial dentro del portafolio de servicios, integrando tanto la presencia por área de negocio como el comportamiento transaccional asociado a cada subproducto.

- **Agrupación de productos y construcción de las Áreas de Portafolio**

El primer paso consistió en agrupar los servicios y productos registrados en Salesforce dentro de

las siete Áreas de Portafolio (BU's) que estructuran la oferta comercial de la empresa.

Esta organización permitió alinear el análisis con la estructura corporativa y establecer una visión integral de cada línea de servicio.

Las áreas consolidadas fueron:

- Private Cloud / Infraestructura Tradicional
- Public Cloud
- Digital Business / Aplicaciones
- Cyber Security
- Soluciones SAP
- Digital Workplace
- Datos / IoT / Smart Systems

Cada área agrupa subproductos y servicios específicos, tales como Managed Services, Professional Services, Hosting, VMware, SAP Application Management, Workplace Services, IoT, entre otros, conforme a la clasificación oficial del portafolio de la empresa.

Tabla 3. Portafolio de productos por BU

Área de Portafolio (BU)	Subproductos / Servicios incluidos
Cloud Pública	Public Cloud Broker AWS, Azure CSP, GCP, IBM, Outsourcing; Public Cloud Managed Services; Public Cloud Platform IBM; Public Cloud Professional Services
Private Cloud / Infraestructura Tradicional	Private Cloud Hosting, Managed Services, Professional Services, TIVIT Private Cloud (RISC / VMware), Takoda Colocation, TIVIT Cloud OpenStack, Transactional Services Mainframe, Traditional Infrastructure Managed / Professional Services, Network Connectivity
Cyber Security	Cloud Security Services, Cyber Security Consulting, Managed Services, Software Solutions, Tech Security Services, Cyber Security SW Solutions
Digital Business / Aplicaciones	Digital Application Development, Application Modernization, AMS Application Management Services, Application MGT Bodyshop Dev/Test/Doc, EDI Business Integration, Venta de Hardware y Software
Soluciones SAP	SAP Implementation Projects and Evolution, SAP Migration & Evolution to S/4HANA, SAP Application Professional Services, SAP Basis Professional Services
Digital Workplace	Digital Workplace Service Desk, Workplace Service Desk, Workplace Services, Field Services
Datos / IoT / Smart Systems	Internet of Things (IoT), Smart Systems, Smart Systems II

Esta estructura permitió realizar la segmentación considerando tanto la presencia comercial como el valor económico (ACV y Total contractual) de cada línea de servicio.

El análisis confirmó que Private Cloud / Infraestructura Tradicional continúa siendo el núcleo operativo del portafolio, con más del 60 % de los clientes y la mayor proporción del valor total de contratos. Le siguen Public Cloud, Digital Business / Aplicaciones y SAP, que muestran una participación creciente en términos de adopción y generación de ingresos.

Las líneas de Cyber Security, Digital Workplace y Datos / IoT / Smart Systems se consolidan como unidades especializadas con un peso menor, pero con potencial de expansión asociado a tendencias de seguridad, movilidad y analítica.

- **Distribución general por Área de Portafolio**

Una vez estructurado el portafolio, se calcularon los indicadores agregados de cada familia: número de clientes, participación porcentual, valor total de contratos (Total_sum), valor contractual anual (ACV_sum) y volumen transaccional (TVT_sum).

Tabla 4 Distribución General por BU

Area_Portafolio	n_clientes	%_sobre total_clientes	ACV_total	Total_total
Cloud Pública	48	28.74	\$ 54,713,078.66	\$ 79,045,618.96
Cyber Security	33	19.76	\$ 22,861,634.72	\$ 33,285,705.73
Datos / IoT / Smart Systems	8	4.79	\$ 4,612,194.84	\$ 14,777,537.30
Digital Business / Aplicaciones	47	28.14	\$ 30,956,769.28	\$ 56,651,219.98
Digital Workplace	26	15.57	\$ 22,103,875.99	\$ 84,561,092.37
Private Cloud / Infra Tradicional	104	62.28	\$ 93,842,300.56	\$ 359,053,985.39
Soluciones SAP	18	10.78	\$ 61,768,523.99	\$ 117,277,246.74

Nota: el término “Datos / IoT / Smart Systems” corresponde a la denominación original de esta área en el CRM Salesforce de la compañía (“Datos” es el equivalente en portugués de “Datos”, dado el origen brasileño de la organización). A lo largo del texto narrativo de este documento se utiliza la forma en español: “Datos / IoT / Smart Systems”.

La distribución del portafolio muestra concentraciones diferenciadas por línea de negocio. La Infraestructura Tradicional (Private Cloud) concentra la mayor proporción, con más del 60 % de los clientes y cerca del 40 % del valor total contractual, consolidándose como la base operativa y económica de los servicios gestionados de la filial; le siguen Cloud Pública (28.7 %) y Digital

Business / Aplicaciones (28.1 %), que evidencian una transición hacia modelos híbridos de operación en nube y desarrollo digital. Soluciones SAP —aunque agrupa un número menor de clientes (10.8 %)— presenta valores de ACV y Total significativamente altos, lo que refleja su carácter estratégico y corporativo, dado que cada contrato representa una relación de alto impacto económico y tecnológico, mientras que Digital Workplace muestra también un valor contractual total elevado en comparación con su número de clientes, lo que sugiere un modelo de servicio intensivo basado en soporte distribuido y experiencia de usuario. Finalmente, Cyber Security y Datos / IoT / Smart Systems tienen menor peso relativo en número de clientes, pero su crecimiento en valor y presencia refuerza su papel como líneas de especialización dentro del portafolio, esenciales para el fortalecimiento de la oferta de transformación digital y analítica avanzada.

En conjunto, la distribución general confirma que la filial en Colombia cuenta con una base sólida en servicios de infraestructura, complementada por líneas de expansión digital y de alto valor estratégico como SAP, Workplace y Cloud Pública, que marcan la dirección de evolución hacia soluciones más integradas y de mayor madurez tecnológica

- **Subproductos con mayor tracción por Área**

Para identificar los servicios con mayor demanda, se evaluó la tracción comercial y económica de los subproductos en cada BU mediante tres métricas:

- TVT_sum: cantidad total de oportunidades ganadas,
- ACV_sum: valor contractual anual,
- Total_sum: valor económico total de los contratos.

Tabla 5. Subproductos por BU

Area_Portafolio	subproducto	TVT	ACV_sum	Total_sum
Cloud Pública	has_public_cloud_professional_services	61	\$ 11,501,397.99	\$ 4,864,955.35
Cloud Pública	has_public_cloud_broker_aws	52	\$ 17,654,503.51	\$ 27,835,044.82
Cloud Pública	has_public_cloud_managed_services	42	\$ 15,723,549.80	\$ 19,049,493.51
Cloud Pública	has_public_cloud_broker_ibm	14	\$ 3,634,819.78	\$ 12,144,404.19
Cloud Pública	has_public_cloud_broker_azure_csp	7	\$ 3,367,364.88	\$ 5,658,999.68
Cyber Security	has_cyber_security_services	38	\$ 11,061,921.40	\$ 14,170,901.04
Cyber Security	has_cyber_security_managed_services	33	\$ 618,329.63	\$ 7,296,766.43
Cyber Security	has_cyber_security_software	28	\$ 2,625,681.89	\$ 4,998,104.43
Cyber Security	has_cyber_security_consulting	19	\$ 615,238.88	\$ 2,500,955.43
Cyber Security	has_cyber_security_consulting.1	16	\$ 4,853,384.00	\$ 1,413,789.39
Datos / IoT / Smart Systems	has_smart_systems	36	\$ 2,312,297.85	\$ 10,451,265.46
Datos / IoT / Smart Systems	has_smart_systems.1	5	\$ 1,701,472.46	\$ 4,189,740.33

Datos / IoT / Smart Systems	has_iot_internet_of_things	3	\$	598,424.53	\$	136,531.51
Digital Business / Aplicaciones	has_tivit_latam_venta_de_hardware_e_software	71	\$	13,577,201.58	\$	30,718,288.19
Digital Business / Aplicaciones	has_digital_ams_application_management_services	36	\$	9,827,227.16	\$	12,702,357.62
Digital Business / Aplicaciones	has_application_mgt_ams_bodyshop_dev_test_doc	35	\$	2,142,280.49	\$	5,467,948.33
Digital Business / Aplicaciones	has_digital_application_modernization	14	\$	4,505,719.11	\$	6,268,111.88
Digital Business / Aplicaciones	has_digital_application_development	13	\$	885,721.22	\$	1,457,274.52
Digital Workplace	has_digital_workplace_workplace_services	61	\$	16,719,241.40	\$	51,128,457.18
Digital Workplace	has_workplace_field_services	15	\$	1,959,560.70	\$	16,284,131.92
Digital Workplace	has_digital_workplace_service_desk	11	\$	3,425,073.89	\$	17,111,048.50
Digital Workplace	has_workplace_service_desk	1	\$	-	\$	37,454.77
Private Cloud / Infra Tradicional	has_private_cloud_tivit_private_cloud_vmware	208	\$	15,871,801.86	\$	89,903,006.32
Private Cloud / Infra Tradicional	has_traditional_infrastructure_managed_services	204	\$	23,422,600.48	\$	96,834,007.85
Private Cloud / Infra Tradicional	has_private_cloud_hosting	144	\$	6,238,884.80	\$	44,581,001.88
Private Cloud / Infra Tradicional	has_private_cloud_professional_services	133	\$	10,022,809.91	\$	27,998,019.34
Private Cloud / Infra Tradicional	has_traditional_infrastructure_professional_services	96	\$	14,906,052.54	\$	21,179,769.64
Soluciones SAP	has_sap_application_professional_services	319	\$	45,729,121.40	\$	100,770,503.50
Soluciones SAP	has_sap_implementation_projetos_e_evolucao	71	\$	14,474,691.24	\$	15,785,623.75
Soluciones SAP	has_sap_basis_professional_services	2	\$	-	\$	341,190.18
Soluciones SAP	has_sap_migration_and_evolution_s_4_hana	1	\$	1,564,711.35	\$	379,929.31

Los resultados permitieron identificar los servicios con mayor demanda dentro de cada BU:

- En Cloud Pública, los servicios Public Cloud Professional Services, Broker AWS y Managed Services concentran la mayor tracción comercial, con altos niveles de valor contractual y transaccionalidad, reflejando la madurez de esta línea en proyectos de adopción y operación en la nube.
- En Cyber Security, destacan los servicios de Cyber Security Services y Managed Services, orientados al fortalecimiento de la ciberdefensa y cumplimiento normativo, con una tendencia creciente en servicios gestionados.
- En Datos / IoT / Smart Systems, los productos Smart Systems e IoT tienen menor volumen de clientes pero alto valor relativo, reflejando un segmento especializado y en crecimiento en torno a innovación tecnológica.
- En Digital Business / Aplicaciones, los subproductos Venta de Hardware y Software LATAM y AMS Application Management Services son los de mayor recurrencia, lo que indica una fuerte orientación hacia el soporte y la modernización de aplicaciones empresariales.
- En Digital Workplace, el servicio Workplace Services lidera ampliamente en volumen y valor económico, seguido por Field Services y Service Desk, evidenciando una línea enfocada en experiencia de usuario y soporte distribuido.
- En Private Cloud / Infraestructura Tradicional, los servicios Traditional Infrastructure Managed Services, VMware y Hosting se consolidan como los de mayor demanda, confirmando el rol estratégico de esta área en la operación central de clientes corporativos.
- Finalmente, en Soluciones SAP, los servicios SAP Application Professional Services y SAP Implementation Projects and Evolution son los más relevantes, vinculados principalmente a contratos de soporte, evolución y migración hacia S/4HANA.

En conjunto, el análisis evidencia que los Managed Services y Professional Services son los motores transversales del portafolio, sustentando tanto el volumen de transacciones como la generación de ingresos en todas las áreas de negocio.

- **Arquetipo general de cliente por familia**

El siguiente paso consistió en caracterizar los clientes por familia de portafolio, considerando variables como tamaño de la empresa, presencia internacional, sector económico y modelo de negocio.

Tabla 6. Caracterización inicial por BU

BU_Portafolio	Moda_Tamaño de la empresa	moda_Presencia internacional	Moda_Sector económico principal	Moda_Modelo de negocio	Moda_Países	moda_Ciudad
Cloud Pública	Grande	No	Gobierno	Distribución de gas	Colombia	Bogotá D.C.
Cyber Security	Grande	No	Petróleo y Gas	Servicios	Colombia	Bogotá D.C.
Datos / IoT / Smart Systems	Grande	Sí	Energía	BPO y Outsourcing	Colombia	Bogotá D.C.
Digital Business / Aplicaciones	Grande	No	Servicios Públicos	Servicios	Colombia	Bogotá D.C.
Digital Workplace	Grande	No	Petróleo y Gas	Exploración y producción	Colombia	Bogotá D.C.
Private Cloud / Infra Tradicional	Grande	No	Tecnología	Servicios	Colombia	Bogotá D.C.
Soluciones SAP	Grande	No	Petróleo y Gas	Servicios	Colombia	Bogotá D.C.

Los resultados muestran una marcada concentración en empresas grandes, principalmente nacionales, pertenecientes a sectores tecnológicos, gubernamentales, de petróleo y gas, servicios públicos y financieros

- **Arquetipo top 3 variables de cliente por familia**

Tabla 7 Caracterización con tres variables por BU.

BU_Portafolio	N clientes	Tamaño	Presencia internacional	Sector económico principal	Modelo de negocio	País	Ciudad
Cloud Pública	48	Grande: 45.8%, Mediana : 31.2%, Pequeña : 18.8%	No: 72.9%, Sí: 27.1%	Gobierno: 18.8%, Tecnología: 12.5%, Servicios Públicos: 10.4%	Distribución de gas: 6.2%, Servicios Financieros: 6.2%, Consortio estatal: 4.2%	Colombia: 97.9%, USA: 2.1%	Bogotá D.C.: 58.3%, Bogotá: 18.8%, Medellín: 6.2%

Cyber Security	33	Grande: 69.7%, Mediana : 18.2%, Pequeña : 12.1%	No: 63.6%, Sí: 36.4%	Servicios Públicos: 15.2%, Petróleo y Gas: 15.2%, Tecnología: 9.1%	Servicios: 15.2%, Distribución de gas: 6.1%, Transporte de crudo: 6.1%	Colombia: 93.9%, USA: 3.0%, Panamá: 3.0%	Bogotá D.C.: 60.6%, Bogotá: 12.1%, Barranquilla: 3.0%
Datos / IoT / Smart Systems	8	Grande: 87.5%, Mediana : 12.5%	Sí: 62.5%, No: 37.5%	Energía: 25.0%, Automotriz: 12.5%, Servicios Públicos: 12.5%	Servicios: 12.5%, Distribución: 12.5%, Gremial, Comercialización: 12.5%	Colombia: 87.5%, Perú: 12.5%	Bogotá D.C.: 75.0%, Lima (HQ): 12.5%, Medellín: 12.5%
Digital Business / Aplicaciones	47	Grande: 76.6%, Mediana : 17.0%, Pequeña : 6.4%	No: 51.1%, Sí: 46.8%, Si: 2.1%	Servicios Públicos: 12.8%, Petróleo y Gas: 10.6%, Tecnología: 10.6%	Servicios: 10.6%, Servicios Financieros: 4.3%, Distribución de gas: 4.3%	Colombia: 87.2%, Panamá: 6.4%, USA: 4.3%	Bogotá D.C.: 68.1%, Panamá: 6.4%, Medellín: 4.3%
Digital Workplace	26	Grande: 73.1%, Mediana : 19.2%, Pequeña : 7.7%	No: 57.7%, Sí: 42.3%	Servicios Públicos: 19.2%, Petróleo y Gas: 19.2%, Gobierno: 11.5%	Servicios: 11.5%, Exploración y producción: 11.5%, Distribución de energía: 7.7%	Colombia: 96.2%, Panamá: 3.8%	Bogotá D.C.: 69.2%, Medellín: 11.5%, Panamá: 3.8%
Private Cloud / Infra Tradicional	104	Grande: 60.6%, Mediana : 30.8%, Pequeña : 7.7%	No: 59.6%, Sí: 40.4%	Telecomunicaciones: 11.5%, Tecnología: 11.5%, Salud: 8.7%	Servicios: 6.7%, Exploración y producción: 2.9%, Distribución de gas: 2.9%	Colombia: 90.4%, USA: 2.9%, Perú: 1.9%	Bogotá D.C.: 54.4%, Bogotá: 17.5%, Medellín: 5.8%
Soluciones SAP	18	Grande: 94.4%, Mediana : 5.6%	No: 77.8%, Sí: 22.2%	Petróleo y Gas: 22.2%, Telecomunicaciones: 16.7%, Salud: 11.1%	Servicios: 16.7%, Producción: 11.1%, Servicios, Logística: 5.6%	Colombia: 88.9%, Panamá: 5.6%, México: 5.6%	Bogotá D.C.: 55.6%, Cartagena: 11.1%, Monterrey: 5.6%

Para profundizar la caracterización, se evaluó la distribución de las tres variables más relevantes por BU: tamaño empresarial, presencia internacional y sector económico principal.

Los resultados confirman que:

- El tamaño “Grande” predomina en todas las áreas, con proporciones que van del 60 % al 95 %, consolidando el enfoque corporativo de la filial en Colombia.
- Las áreas con mayor internacionalización son Datos / IoT / Smart Systems (62.5 %), Private Cloud (40.4 %) y Digital Workplace (42.3 %), evidenciando la participación de clientes con presencia en Latinoamérica y Norteamérica.
- Los sectores más representativos son Gobierno, Petróleo y Gas, Tecnología, Energía y Servicios Públicos.
- El modelo de negocio predominante es Servicios, seguido de Distribución y Exploración/Producción, lo que refleja la orientación a contratos de soporte continuo y outsourcing especializado.

Conclusiones generales del EDA

El análisis exploratorio de las Áreas de Portafolio y Subproductos permitió comprender la estructura actual del negocio de la filial desde una visión transversal y estratégica. Los principales hallazgos se pueden sintetizar en los siguientes puntos:

- Private Cloud / Infraestructura Tradicional representa la columna vertebral del portafolio, sosteniendo la mayor parte de los ingresos y clientes activos, especialmente a través de servicios de infraestructura y virtualización gestionada.
- Public Cloud se consolida como una línea de crecimiento sostenido, impulsada por proyectos de adopción multicloud y servicios profesionales de soporte y consultoría.
- Soluciones SAP, aunque no constituye una BU formal sino una línea estratégica, se comporta como un eje de alto valor contractual, con fuerte incidencia en clientes corporativos y proyectos críticos de transformación digital.
- Digital Workplace, al igual que SAP, funciona como una línea especializada y transversal, con foco en la experiencia del usuario y la gestión de dispositivos, clave para la continuidad de los servicios.
- Digital Business / Aplicaciones actúa como un vínculo natural entre la infraestructura y los servicios empresariales, articulando el soporte funcional, la modernización tecnológica y la venta de software y hardware complementario.
- Cyber Security y Datos / IoT / Smart Systems refuerzan el posicionamiento innovador de la filial en servicios de valor agregado, al atender las tendencias globales de seguridad, analítica e inteligencia operacional.
- El arquetipo dominante de cliente corresponde a empresas grandes, del sector corporativo, principalmente con sede en Colombia, aunque con creciente proyección internacional en Latinoamérica.

En conjunto, los resultados revelan una base de clientes madura y diversificada, un portafolio

equilibrado entre servicios tradicionales e innovadores, y un patrón de evolución hacia modelos gestionales, híbridos y basados en datos.

5 MODELOS DE APRENDIZAJE AUTOMÁTICO PARA IDENTIFICAR Y PRIORIZAR PROSPECTOS COMERCIALES

El Esquema 2 sintetiza los dos enfoques complementarios de este objetivo: una vía no supervisada (K-Means por unidad de negocio sobre los 167 clientes caracterizados) que produce perfiles estratégicos, y una vía supervisada (Random Forest, XGBoost, Regresión Logística y similitud coseno en un score híbrido) que genera los rankings de prospectos y de venta cruzada. La tabla 8 detalla los hiperparámetros clave y el criterio de selección de cada modelo.

Esquema 2. Metodología del Objetivo 2: modelado no supervisado y supervisado



Tabla 8. Hiperparámetros

Modelo	Hiperparámetros clave	Cómo se seleccionó
K-Means	n_clusters=k · n_init=20 · random_state=42 · log1p+RobustScaler	k por codo + Silhouette (2–5)
Random Forest	n_estimators=400 · max_depth=4 · class_weight=balanced_subsample	Mejor AUC-ROC en CV
XGBoost	n_estimators=500 · lr=0,05 · max_depth=4 · subsample=0,9 · colsample=0,9	Comparado por AUC-ROC
Reg. Logística	StandardScaler · max_iter=2000	Baseline lineal
Similitud Coseno	cosine_similarity · StandardScaler · vs. centroide estratégico	Afinidad (no entrena)

Bases de datos utilizadas en el proyecto

El desarrollo de los modelos a lo largo de este capítulo y del Capítulo 4 se apoya en tres bases de

datos diferenciadas, descritas en la Tabla 8. Esta tabla actúa como referencia única: en lo sucesivo, los datasets se identifican por sus etiquetas (Base A, Base B y Base C) sin repetir su descripción.

Tabla 9. *Bases de datos utilizadas en el proyecto*

Etiqueta	Archivo	Descripción	Variables principales	Uso en el proyecto
Base A	clientes_bda_model.csv	Dataset analítico principal: 167 clientes activos del CRM Salesforce, ya pre-procesado con feature engineering (transformación log1p, escalamiento robusto, One-Hot Encoding e imputación). Construido a partir de la Base B.	Comportamiento comercial (opp_count, acv_total, precio_total_sum, years_present, n_lineas_producto), consumo por línea de servicio (has_*, acv_*, total_*, tvf_count_*) y caracterización empresarial (sector, tamaño, modelo de negocio, presencia internacional, tipo y subtipo de empresa).	Segmentación K-Means por BU (Sección 5.1) y modelo de cross-selling (Sección 5.2.1).
Base B	clientes_bda_caracterizada.csv	Dataset intermedio de 167 clientes: información del CRM Salesforce enriquecida con caracterización empresarial obtenida vía Gemini, antes del pre-procesamiento. Insumo para construir la Base A.	Variables transaccionales del CRM (oportunidades, ACV, Total, TVT) más caracterización empresarial sin codificar ni transformar (categorías originales).	Análisis exploratorio descriptivo (Capítulo 4) y construcción de la Base A (inicio del Capítulo 5).
Base C	DATABASE.xlsx (enriquecida vía Gemini Deep Research)	572 prospectos externos: empresas no clientes identificadas como potenciales, con caracterización firmográfica obtenida mediante Gemini Deep Research.	Variables firmográficas comunes con la Base A: Ciudad, País, Tipo y Subtipo de empresa, Tamaño organizacional, Presencia internacional, Sector económico principal, Modelo de negocio, Modelo normalizado y Submodelo. Sin variables transaccionales (no son clientes activos).	Modelo supervisado para prospectos externos y similitud coseno (Sección 5.2.2). Score híbrido final.

5.1 MODELO NO SUPERVISADO

5.1.1 PROCESAMIENTOS

A. Procesamiento 1 – Segmentación Global Exploratoria de Clientes

Con el fin de identificar grupos homogéneos de clientes en función de su comportamiento y características organizacionales, se desarrolló un modelo de aprendizaje automático no supervisado basado en el algoritmo K-Means. Este enfoque permite agrupar los registros sin necesidad de etiquetas previas, encontrando patrones latentes en los datos.

Preparación de datos y selección de variables

El modelo utiliza como insumo el archivo `clientes_bda_model.csv`, que integra tanto variables internas del CRM (transaccionales) como variables externas de caracterización empresarial.

Las variables se agruparon en dos tipos:

- Variables categóricas: País, Ciudad, Tipo de empresa, Tamaño de la empresa, Presencia internacional, Sector económico principal y Modelo de negocio. Estas describen la naturaleza del cliente y su contexto organizacional.
- Variables numéricas: Número de oportunidades ganadas (`opp_count`), valor total de contratos (`acv_total`, `precio_total_sum`), años de permanencia (`years_present`), diversidad de líneas y familias de producto (`n_lineas_producto`, `n_familias`), y los años de primera y última transacción (`first_year`, `last_year`). Estas capturan el comportamiento comercial y la trayectoria del cliente dentro de la organización.

Preprocesamiento de datos

Para garantizar la coherencia y comparabilidad de las variables, se implementaron pipelines de preprocesamiento con `scikit-learn` que incluyeron imputación, transformación y escalamiento.

Las principales acciones realizadas fueron:

- Imputación de valores faltantes:
 - ✓ En variables categóricas, se reemplazaron los valores nulos por la categoría más frecuente.
 - ✓ En variables numéricas, se imputó con la mediana para evitar sesgos causados por valores extremos.
- Codificación y transformación:

- ✓ Las variables categóricas se codificaron mediante One-Hot Encoding, generando una representación binaria sin imponer jerarquías artificiales.
- ✓ Las variables numéricas fueron transformadas con la función $\log_{1p}$ para reducir la asimetría en su distribución.
- Escalamiento:
Se aplicó el método RobustScaler, el cual ajusta los datos en función de la mediana y el rango intercuartílico, haciendo el modelo menos sensible a valores atípicos.

Estas transformaciones se integraron dentro de un ColumnTransformer, generando una matriz unificada y normalizada que sirvió como entrada para el modelo de clustering.

Decisión sobre la reducción de dimensionalidad previa al agrupamiento

Antes de aplicar K-Means se evaluó la pertinencia de incorporar una técnica de reducción de dimensionalidad (como Análisis de Componentes Principales, PCA) sobre la matriz resultante del preprocesamiento, y se decidió no aplicarla como paso previo al agrupamiento. Esta decisión se sustenta en tres razones. En primer lugar, por interpretabilidad: cada variable del modelo —en particular los indicadores `has_`, `acv_`, `total_` y `tv_count_` por subproducto— codifica una señal de negocio directamente accionable, de modo que proyectar el espacio a componentes principales abstractas habría diluido esta correspondencia y dificultado la lectura comercial de los clústeres, que constituye el objetivo central del proyecto. En segundo lugar, por la naturaleza de los datos: como se documenta en la caracterización de la base, la alta dimensionalidad no obedece a redundancia entre variables, sino a la estructura granular del CRM, donde cada subproducto genera columnas independientes; en ausencia de colinealidad masiva, el beneficio de PCA sobre la calidad del agrupamiento resulta marginal. En tercer lugar, por mitigación: el escalamiento robusto previo, la transformación logarítmica ($\log_{1p}$) y, sobre todo, la decisión de entrenar modelos de clustering independientes por unidad de negocio —en lugar de un único agrupamiento global— acotan el espacio de características de cada modelo y reducen el efecto de la maldición de la dimensionalidad. En consecuencia, PCA se reservó como herramienta de visualización y diagnóstico a posteriori, donde precisamente permitió identificar la sobre-segmentación que motivó el enfoque por unidad de negocio.

Entrenamiento del modelo

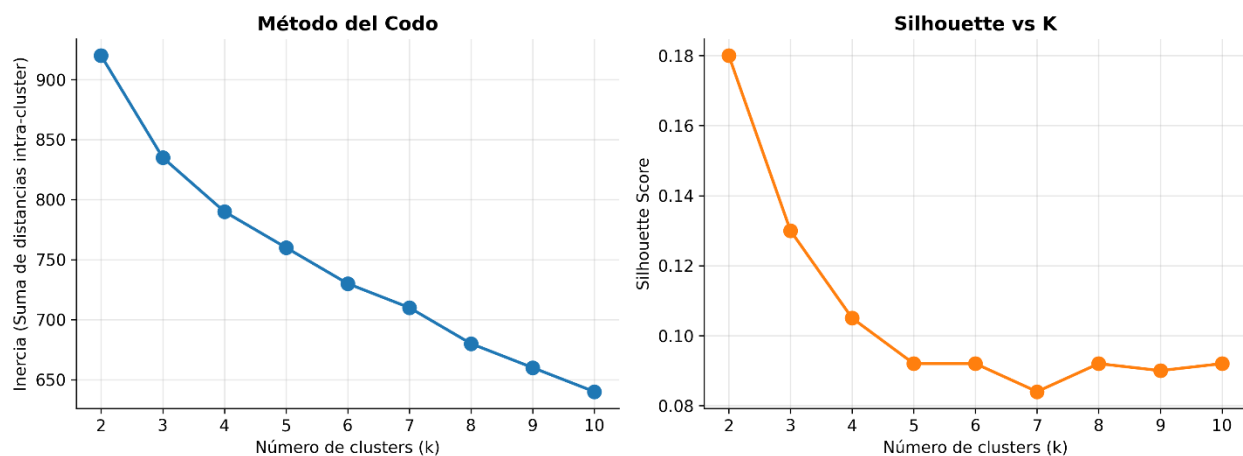
El algoritmo seleccionado fue K-Means, debido a su eficiencia y su amplia aplicación en contextos de segmentación comercial. El modelo fue entrenado con distintos valores de k (número de clusters), dentro del rango de 2 a 6, con el objetivo de determinar la estructura natural de los grupos presentes en la base de datos. Para cada configuración, se calculó el índice de silueta (silhouette score), métrica que evalúa la cohesión interna y separación entre grupos. El valor más alto se obtuvo en $k = 2$, indicando la mejor separación relativa desde un punto de vista estadístico.

Validación del número óptimo de clusters

Para validar el número óptimo de grupos, se implementaron dos técnicas complementarias: el Método del Codo (Elbow Method) y el índice de silueta (Silhouette Score).

El código utilizado permitió evaluar ambos indicadores para valores de k entre 2 y 10. Los resultados se visualizaron gráficamente, como se muestra en la Figura 1

Figura 1. Análisis gráfico: Método del Codo y Silhouette



En el gráfico del Método del Codo, se observa una disminución pronunciada de la inercia (suma de distancias intra-cluster) hasta aproximadamente $k = 4$, punto a partir del cual la ganancia marginal se estabiliza; el índice de silueta, por su parte, alcanza su valor máximo en $k = 2$ (≈ 0.18), lo que sugiere la mayor separación geométrica entre grupos. No obstante, desde una perspectiva de negocio, la configuración con solo dos clústeres resultó excesivamente general al agrupar clientes heterogéneos en categorías muy amplias, por lo que se realizaron pruebas complementarias con distintos valores de k , valorando no solo la calidad técnica sino también la interpretabilidad y relevancia estratégica de los segmentos obtenidos.

Resultados de segmentación

En la configuración inicial con $k = 2$, los resultados obtenidos fueron:

Tabla 10. Configuración $k=2$

Variable / Indicador	Cluster 0	Cluster 1
Número de clientes	116	51
Promedio de oportunidades (opp_count)	1.95	24.76

ACV promedio (COP)	310,293	4,997,339
Precio total promedio (COP)	623,673	13,182,480
Años de relación comercial (years_present)	1.46	4.88
Promedio de BU's	2.07	6.80
Año de última oportunidad (last_year)	2021.95	2024.04
Año de primera oportunidad (first_year)	2021.17	2019.39
País predominante	Colombia	Colombia
Ciudad predominante	Bogotá D.C.	Bogotá D.C.
Tamaño de la empresa predominante	Grande	Grande
Presencia internacional	No	Sí
Sector económico principal	Tecnología	Petróleo y Gas
Modelo de negocio predominante	Distribución de gas	Servicios

- Cluster 0: agrupa clientes con baja recurrencia de oportunidades y montos moderados. Predominan empresas grandes, locales, sin presencia internacional, con modelo de negocio basado en distribución.
- Cluster 1: concentra empresas multinacionales con alto volumen de negocio, presencia internacional y participación en sectores industriales o de energía.

Sin embargo, al aumentar el número de clusters a $k=4$ o $k=8$, se logró una segmentación más granular, capaz de distinguir entre empresas locales de crecimiento moderado, organizaciones medianas en expansión y corporaciones multinacionales consolidadas.

Configuración con $k=4$

La segmentación con cuatro grupos ($k=4$) permitió obtener una representación más equilibrada entre precisión técnica e interpretabilidad estratégica, identificando patrones relevantes en el comportamiento de los clientes.

Tabla 11. Configuración $k=4$

Variable / Indicador	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Número de clientes	57	44	19	47
Promedio de oportunidades (opp_count)	1.37	6.30	54.26	2.19
ACV promedio (COP)	246,377	1,591,453	10,236,460	261,662

Precio total promedio (COP)	582,098	2,701,699	29,505,710	680,63
Años de relación comercial	1.21	3.52	6.74	1.40
Promedio de BU's	1.96	3.95	10.89	2.00
Año de última oportunidad	2023.7	2023.5	2024.8	2019.5
Año de primera oportunidad	2023.3	2020.0	2018.6	2018.8
País predominante	Colombia	Colombia	Colombia	Colombia
Ciudad predominante	Bogotá D.C.	Bogotá D.C.	Bogotá D.C.	Bogotá D.C.
Tamaño de la empresa	Grande	Grande	Grande	Grande
Presencia internacional	No	No	No	No
Sector económico principal	Tecnología	Tecnología	Salud	Telecomunicaciones
Modelo de negocio	Distribución de gas	Distribución de energía	Servicios	B2B: telecom mayorista

Interpretación de los clusters:

- Cluster 0: empresas grandes del sector tecnológico con baja recurrencia de oportunidades y sin presencia internacional.
- Cluster 1: compañías con actividad comercial intermedia y mayor diversificación de productos.
- Cluster 2: corporaciones con larga relación comercial, alto volumen de negocio y foco en el sector salud.
- Cluster 3: organizaciones del sector telecomunicaciones con orientación B2B y actividad reciente.

Esta configuración permitió distinguir con mayor claridad las diferencias en madurez, sector y modelo de negocio, facilitando una lectura más estratégica de la base de clientes.

Configuración con $k = 8$

Posteriormente se entrenó el modelo con $k = 8$, con el propósito de explorar una segmentación más fina y detectar microgrupos específicos de clientes.

Tabla 12. Configuración con $k = 8$

Variable / Indicador	C0	C1	C2	C3	C4	C5	C6	C7
Número de clientes	18	21	22	17	20	29	27	13
Oportunidades promedio	1.56	1.19	2.27	59.35	1.60	7.45	2.41	4.92
ACV promedio (COP)	968,739	568,762	759,334	10,639,660	118,021	1,713,975	84,224	735,207
Precio total promedio (COP)	670,586	1,319,414	568,993	30,225,150	138,404	4,475,373	790,652	1,894,400
Años de relación	1.22	1.10	1.82	6.94	1.30	4.03	1.44	2.54
Promedio de BU's	1.94	2.29	2.91	11.41	1.75	4.24	2.04	2.54
Sector económico principal	Tecnología	Gobierno	Gobierno	Salud	Financiero	Telecomunicaciones	Telecomunicaciones	Tecnología
Modelo de negocio predominante	B2C: salud	Distribución de gas	Soluciones de software	Servicios	B2B: call center	Distribución de energía	B2B: telecom mayorista	Agencia de publicidad

Interpretación de los clusters:

- Clusters 0 y 7: empresas tecnológicas, diferenciadas por su nivel de internacionalización.
- Clusters 1 y 2: entidades del sector público (Gobierno), segmentadas por tamaño y tipo de servicio.
- Cluster 3: corporaciones del sector salud con relaciones comerciales prolongadas y altos volúmenes de ACV.
- Cluster 4: pequeñas empresas financieras con baja madurez comercial.
- Clusters 5 y 6: compañías del sector telecomunicaciones, con modelos B2B y B2C

diferenciados.

Esta configuración permitió identificar microsegmentos más específicos, útiles para estrategias avanzadas de personalización, aunque con menor simplicidad interpretativa que el modelo con $k = 4$.

Comparación general de configuraciones

Tabla 13. Comparación técnica y estratégica de las configuraciones

Criterio	k = 2	k = 4	k = 8
Separación técnica (silhouette)	Alta	Media	Baja
Interpretabilidad	Baja	Alta	Media
Granularidad	Limitada	Equilibrada	Muy alta
Utilidad estratégica	Moderada	Alta	Alta (compleja de aplicar)

En esta etapa del proyecto, los diferentes resultados de segmentación se consideran complementarios y serán analizados comparativamente en fases posteriores para determinar cuál configuración ofrece el mejor equilibrio entre calidad técnica, interpretabilidad y valor estratégico para el negocio.

B. Procesamiento 2 – Segmentación por BU (áreas del portafolio) basada en presencia

Tras la primera fase de segmentación basada en variables generales del cliente, se implementó un segundo modelo no supervisado orientado a explorar la agrupación de clientes según su relación con las distintas áreas del portafolio de servicios (BU's).

El objetivo de este modelo fue identificar patrones de consumo y especialización entre clientes a partir de su presencia o participación en los diferentes productos y servicios ofrecidos por la organización.

Preparación de los datos

El modelo se desarrolló utilizando el archivo consolidado clientes_bda_model.csv, previamente depurado y normalizado. A partir de este dataset se seleccionaron todas las columnas asociadas

a indicadores de producto o BU, siguiendo expresiones regulares como `has_`, `tvf_count_bu_`, `acv_bu_` y `total_bu_`, que representan respectivamente:

- Presencia binaria (`has_`): indica si un cliente posee consumo en una BU específica.
- Volumetrías (`tvf_count_bu_`): número de oportunidades registradas en esa BU.
- ACV (`acv_bu_`) y Total (`total_bu_`): montos contractuales y totales asociados a dicha área del portafolio.

La base de entrada se estructuró como una matriz de presencia BU–Cliente, donde cada fila representa un cliente y cada columna indica si tiene o no servicios activos en una determinada BU.

En la configuración principal se utilizó únicamente la presencia (0/1), para reflejar el patrón de cobertura por línea del portafolio sin sesgar la agrupación por montos de venta.

Entrenamiento del modelo

Se aplicó el algoritmo K-Means con distintos valores de k en el rango 2–10, evaluando los resultados mediante las métricas de inercia intra-cluster y silhouette score. El análisis de las métricas mostró estabilidad en torno a $k = 4$ y $k = 8$, valores que ofrecen una combinación adecuada entre separación de grupos e interpretabilidad.

Por tanto, se entrenaron dos configuraciones finales:

- Modelo BU– $k4$: 4 clusters (agrupaciones amplias por comportamiento de consumo).
- Modelo BU– $k8$: 8 clusters (microsegmentación más específica por línea de servicio).

Los resultados fueron almacenados en los archivos:

- `/outputs_v2/clientes_segmentados_BU_k4.csv`
- `/outputs_v2/clientes_segmentados_BU_k8.csv`

Resultados – Configuración $k = 4$

Tabla 14. *Distribución de clientes y principales BU por cluster ($k = 4$)*

Cluster	Número de clientes	BU's predominantes (Top 5)	% Presencia promedio
0	13	Private Cloud Professional Services, Traditional Infrastructure Managed Services, TIVIT Latam Hardware/Software, Private Cloud	77–100%

		Managed Services, Traditional Infra Prof. Services	
1	52	Traditional Infrastructure Managed Services, Private Cloud VMware, Private Cloud Hosting, Private Cloud Managed Services, Takoda Colocation	17–94%
2	33	Public Cloud Professional Services, Public Cloud Managed Services, Public Cloud Broker AWS, Public Cloud Broker IBM, TIVIT Latam Hardware/Software	15–100%
3	69	Private Cloud Managed Services, Private Cloud Hosting, VMware, Traditional Infra Prof. Services, Digital Workplace Services	12–17%

Interpretación:

- Cluster 0: clientes con fuerte enfoque en infraestructura tradicional y servicios profesionales, asociados a Private Cloud y Hardware.
- Cluster 1: clientes con presencia significativa en servicios de infraestructura y hosting administrado.
- Cluster 2: clientes orientados a servicios de nube pública y brokers multicloud (AWS, IBM, Azure).
- Cluster 3: organizaciones con consumo diversificado, donde predominan servicios Private Cloud y Workplace, con menor intensidad por línea.

Resultados – Configuración k = 8

Tabla 15. *Distribución de clientes y BU's predominantes (k = 8)*

Cluster	Clientes	BU's predominantes (Top 5)
0	52	SAP Application Services, Digital Workplace, Traditional Infra Prof. Services, Hardware/Software, Cybersecurity Managed
1	27	Private Cloud Hosting, Traditional Infra Managed, Network Connectivity, VMware, Takoda Colocation

2	4	Private Cloud Managed + Professional + VMware, Hardware/Software, Public Cloud Professional
3	32	Public Cloud Professional, Public Cloud Managed, AWS, IBM, Azure CSP
4	15	Private Cloud Managed, VMware, Infra Managed, Hosting, OpenStack
5	4	Cybersecurity Managed, AMS Application Services, Cybersecurity Services, Private Cloud Professional, Managed Services
6	23	Traditional Infra Managed, VMware, Private Cloud Professional, Takoda Colocation, Workplace
7	10	Traditional Infra Managed, Hardware/Software, Private Cloud Professional, Managed Services, Hosting

Interpretación:

- Clusters 0 y 5: concentran clientes con servicios avanzados de aplicación, seguridad y nube privada.
- Clusters 1, 4, 6 y 7: representan variaciones del segmento de infraestructura tradicional y privada, diferenciados por combinaciones de hosting, conectividad y servicios profesionales.
- Cluster 2: agrupa clientes altamente integrados con varias líneas de nube privada y pública simultáneamente.
- Cluster 3: concentra organizaciones enfocadas exclusivamente en nube pública, especialmente en brokers AWS y Azure.

Esta configuración permitió identificar especializaciones claras por tipo de servicio, revelando patrones que no eran visibles en la segmentación general previa.

Conclusiones parciales del modelo por BU

El modelo de segmentación basado en BU's permitió:

- Identificar agrupamientos naturales de clientes según sus combinaciones de consumo por línea del portafolio.
- Detectar patrones de especialización (p. ej., clientes exclusivamente de nube pública o de infraestructura privada).
- Relacionar tipos de servicios con características empresariales, como tamaño, sector económico o presencia internacional.
- Establecer una base sólida para estrategias de cross-selling, expansión de portafolio y

priorización de nuevas oportunidades comerciales.

C. Procesamiento 3 – Caracterización Descriptiva por Unidad de Negocio (BU)

Previo a la aplicación de técnicas de segmentación no supervisada a nivel de unidad de negocio, se desarrolló un ejercicio de caracterización descriptiva por BU, con el objetivo de comprender el perfil general de los clientes que consumen cada línea del portafolio. Este análisis se concibe como una línea base de negocio, orientada a identificar patrones estructurales en las variables de caracterización y métricas comerciales, sin imponer agrupaciones artificiales ni supuestos propios de modelos de clustering.

A diferencia de los intentos de segmentación presentados anteriormente, este apartado no busca identificar segmentos latentes mediante algoritmos no supervisados, sino describir de manera objetiva y directa las características predominantes de los clientes asociados a cada unidad de negocio. De esta forma, se establece un marco de referencia que permite validar posteriormente la coherencia y utilidad de los clusters obtenidos en modelos más avanzados.

Enfoque metodológico

El análisis se realiza de manera independiente para cada unidad de negocio (BU), utilizando como criterio principal la presencia efectiva de consumo. Para ello, se emplean variables binarias del tipo `has_*`, las cuales indican si un cliente ha tenido al menos una oportunidad asociada a la BU o subproducto correspondiente durante el período de análisis.

Para cada BU, se filtra el conjunto de clientes que presentan consumo (`has_BU = 1`) y, sobre este subconjunto, se analizan variables de caracterización organizacional y comercial, tales como:

- Sector económico
- Tamaño de la empresa
- Tipo de empresa
- Modelo de negocio
- Presencia internacional
- País y ciudad
- Métricas comerciales agregadas (cuando aplica), como volumen de oportunidades y montos asociados

El análisis se basa en estadísticas descriptivas simples, tales como frecuencias relativas, modas y distribuciones porcentuales, priorizando la interpretabilidad y alineación con la lógica de negocio.

Variables consideradas

Las variables utilizadas en este ejercicio se agrupan en dos categorías principales:

- Variables de caracterización organizacional

- Sector económico
- Tamaño de la empresa
- Tipo de empresa
- Modelo de negocio
- Presencia internacional
- Ubicación geográfica (país y ciudad)

Variables comerciales

- Número de oportunidades asociadas a la BU
- Montos comerciales agregados (ACV, valor total, cuando la información se encuentra disponible)

La selección de estas variables responde a su relevancia para la comprensión del contexto empresarial del cliente y su relación con el portafolio de servicios.

Procedimiento de análisis

El procedimiento seguido para la caracterización descriptiva por unidad de negocio se resume en los siguientes pasos:

1. Identificación de las columnas binarias (has_*) asociadas a cada BU o subproducto.
2. Filtrado del dataset para conservar únicamente los clientes con consumo efectivo en la BU analizada.
3. Cálculo de distribuciones de frecuencia y valores modales para las variables de caracterización.
4. Identificación de patrones predominantes y rasgos comunes entre los clientes de cada BU.
5. Consolidación de los resultados en perfiles descriptivos por unidad de negocio.

Este procedimiento se replica de forma consistente para todas las BUs del portafolio, garantizando comparabilidad entre resultados.

Resultados y análisis

Los resultados del análisis descriptivo por unidad de negocio permitieron establecer perfiles generales de los clientes que consumen cada línea del portafolio, evidenciando patrones consistentes en variables como sector económico, tamaño de la empresa, modelo de negocio y alcance geográfico. Este ejercicio confirmó que las unidades de negocio no son demandadas de manera homogénea, sino que responden a contextos organizacionales diferenciados.

Adicionalmente, al contrastar estos resultados con los experimentos de segmentación no

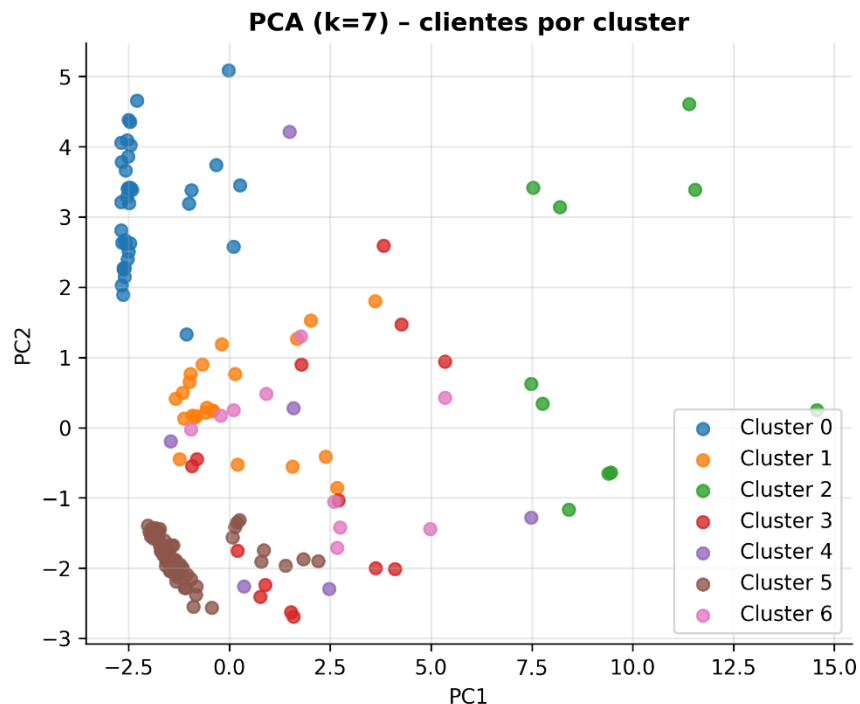
supervisada se exploró un valor intermedio de $k = 7$, distinto de los extremos previamente evaluados en las Tablas 9 y 10 ($k = 4$ y $k = 8$), con el propósito de identificar un punto de balance entre la simplicidad interpretativa de configuraciones más amplias y la microsegmentación observada en $k = 8$. La elección de $k = 7$ respondió específicamente a un criterio de interpretabilidad de negocio: se buscaba un número de grupos lo suficientemente diferenciado como para construir perfiles estratégicos distinguibles, sin caer en una fragmentación excesiva.

Sin embargo, la proyección mediante Análisis de Componentes Principales (PCA) sobre esta configuración, en la Figura 2. Proyección PCA — Segmentación global con $k=7$ (solapamiento entre clústeres), evidenció que tampoco esta granularidad intermedia logra capturar segmentos claramente separables. Solo dos de los siete clusters muestran cohesión visual: el Cluster 0 se concentra en la región superior izquierda del plano ($PC1 \approx -2.5$, $PC2$ entre 1 y 5), mientras que el Cluster 5 forma un grupo compacto en la zona inferior izquierda ($PC2 \approx -2$). Los Clusters 1, 3, 4 y 6 ocupan en cambio un mismo espacio central, con un solapamiento extenso que impide diferenciar perfiles entre ellos. El Cluster 2, por su parte, se distribuye de forma dispersa hacia la región positiva de $PC1$, comportamiento más asimilable a un grupo de outliers que a un segmento estratégico.

Este patrón —dos clusters bien definidos, cuatro indistinguibles entre sí y un quinto disperso— constituye la implicación metodológica central: la estructura subyacente del dataset no soporta una segmentación global granular, independientemente del valor de k explorado. Esta observación motivó el cambio de enfoque hacia el clustering independiente por unidad de negocio, donde las particiones se calibran sobre el subconjunto de clientes de cada BU y no sobre el portafolio completo.

En este contexto, la caracterización descriptiva por unidad de negocio permitió interpretar de manera crítica los resultados del clustering, mostrando que, si bien existen perfiles claramente diferenciados para ciertos tipos de clientes, una subdivisión excesiva tiende a fragmentar artificialmente grupos que comparten características similares. Por tanto, los hallazgos descriptivos aportan una referencia fundamental para evaluar la coherencia y utilidad práctica de las segmentaciones obtenidas mediante modelos no supervisados.

Figura 2. Proyección PCA — Segmentación global con $k=7$ (solapamiento entre clústeres)



Rol del análisis descriptivo dentro del proceso de segmentación

El análisis descriptivo por unidad de negocio cumple un rol central dentro del proceso metodológico de segmentación, al funcionar como una línea base empírica frente a la cual se contrastan los resultados de los modelos no supervisados. En particular, este ejercicio permitió identificar los rasgos dominantes de los clientes por BU sin imponer estructuras de agrupación, evitando así interpretaciones sesgadas derivadas de decisiones arbitrarias sobre el número de clusters.

Los resultados del PCA aplicado a una segmentación con siete clusters reforzaron la importancia de este enfoque descriptivo, al evidenciar que una mayor granularidad no necesariamente implica una mejor segmentación. La superposición observada entre varios clusters indica que el espacio de características no soporta una partición tan detallada, lo que valida el uso del análisis descriptivo como mecanismo de control conceptual y metodológico.

En este sentido, la caracterización descriptiva permitió:

- Evaluar la coherencia de los clusters obtenidos.
- Identificar casos de sobre-segmentación.

- Justificar la necesidad de reducir la complejidad del modelo o adoptar enfoques alternativos más alineados con la estructura del negocio.

De esta manera, este análisis actúa como un puente metodológico entre la exploración inicial y la segmentación avanzada, asegurando que las decisiones de modelado estén respaldadas tanto por evidencia estadística como por lógica empresarial.

Limitaciones

Si bien la caracterización descriptiva por unidad de negocio aporta claridad y robustez al análisis, presenta limitaciones que deben ser consideradas. En primer lugar, al tratarse de un enfoque no segmentador, no permite identificar subgrupos internos dentro de una misma BU ni capturar interacciones complejas entre variables.

Adicionalmente, este análisis no incorpora medidas de distancia ni criterios de optimización, por lo que no sustituye a los modelos de clustering cuando el objetivo es la identificación automática de segmentos. Sin embargo, los resultados obtenidos, junto con la evidencia visual del PCA, muestran que una segmentación excesivamente detallada a nivel global puede conducir a interpretaciones poco sólidas, especialmente cuando los clusters presentan un alto grado de solapamiento.

Finalmente, el análisis descriptivo depende de la calidad y completitud de las variables de caracterización disponibles, lo que puede limitar la profundidad del perfilamiento en ciertos casos.

Conclusiones

La caracterización descriptiva por unidad de negocio permitió establecer una comprensión clara y estructurada del perfil general de los clientes asociados a cada línea del portafolio, constituyéndose en un insumo clave para el proceso de segmentación. Este enfoque evidenció que existen patrones diferenciados entre BUs, pero también que dichos patrones no siempre justifican una segmentación altamente granular.

El contraste con los resultados del clustering, particularmente con la proyección PCA de una segmentación con siete clusters, reveló que una mayor cantidad de grupos no necesariamente se traduce en una mejor diferenciación, sino que puede generar sobre-segmentación y pérdida de interpretabilidad. En este contexto, el análisis descriptivo permitió identificar estas limitaciones y orientar la toma de decisiones hacia modelos más coherentes y accionables.

En conclusión, este apartado cumple un rol fundamental dentro del estudio al proporcionar una base empírica sólida que sustenta y valida las decisiones metodológicas adoptadas en las etapas posteriores, contribuyendo a una segmentación alineada tanto con la estructura de los datos como con la realidad del negocio.

D. Procesamiento 4 – Segmentación Avanzada por BU y Subproducto

Dado que cada Unidad de Negocio (BU) presenta una naturaleza distinta —tanto en los tipos de servicios ofrecidos como en los esquemas de contratación, el volumen económico y el nivel de madurez tecnológica de sus clientes—, se determinó que una segmentación global podría ocultar patrones relevantes de comportamiento. En particular, una segmentación única para todo el portafolio tiende a mezclar lógicas de consumo heterogéneas, dificultando la interpretación de los resultados desde una perspectiva de negocio.

Por esta razón, se optó por desarrollar un modelo de clustering independiente para cada Unidad de Negocio, permitiendo capturar de manera más precisa los patrones de consumo propios de cada línea de servicios.

El objetivo principal de este enfoque fue construir perfiles de clientes específicos por BU, evitando la mezcla de comportamientos disímiles y asegurando una segmentación coherente con la estructura real del portafolio. Estos perfiles constituyen la base conceptual y técnica para la posterior identificación y priorización de prospectos comerciales externos.

Preparación de los datos y definición de instancias

El proceso se inició con la carga de un dataset consolidado de clientes provenientes del CRM Salesforce, el cual incluye:

- variables binarias de consumo por subproducto (has_*)
- variables cuantitativas asociadas a cada subproducto (ACV, Total y TVT),
- y variables categóricas descriptivas del cliente.

Para cada Unidad de Negocio, se realizó un filtrado de instancias, conservando únicamente aquellos clientes que presentan consumo efectivo de al menos un subproducto perteneciente a la BU analizada. Este criterio fue implementado directamente en el código mediante el uso de una máscara booleana, garantizando que cada modelo de clustering se entrenara exclusivamente sobre clientes relevantes para la línea de negocio correspondiente.

Este enfoque permitió trabajar con subconjuntos de datos homogéneos, reduciendo el ruido introducido por clientes no relacionados con la BU y mejorando significativamente la interpretabilidad de los resultados.

Ingeniería de características y normalización

Las variables de entrada del modelo se construyeron combinando información estructural, económica y de complejidad de consumo. En particular, se utilizaron:

- Variables binarias que representan la presencia de subproductos.

- Variables económicas (ACV, Total y TVT), transformadas mediante la función log1p para reducir la asimetría de las distribuciones y mitigar el impacto de valores extremos.
- Variables derivadas que capturan la diversidad del consumo, tales como el número de subproductos contratados dentro de la BU y el número total de subproductos contratados por el cliente.

Adicionalmente, cuando la información estuvo disponible, se incorporaron variables categóricas que describen el contexto empresarial del cliente.

Antes del entrenamiento del modelo, el preprocesamiento aplicado siguió la metodología descrita en la Sección 3.2 Marco teórico, (transformación log1p, escalamiento robusto, imputación por mediana o moda según el tipo de variable y codificación One-Hot). Para cada BU se aplicó este mismo flujo sobre el subconjunto correspondiente de clientes.

Entrenamiento del modelo y selección del número de clusters

Para cada BU se entrenaron múltiples modelos KMeans, evaluando un rango de valores del número de clusters k . La selección del valor óptimo de k se basó en métricas de **validación interna**, calculadas directamente sobre el espacio transformado de características utilizado por el algoritmo.

Las métricas de validación interna empleadas fueron Silhouette, Calinski–Harabasz y Davies–Bouldin, descritas en detalle en la Sección 3.2 Marco teórico.

Estas métricas fueron calculadas para cada valor de k evaluado y almacenadas en tablas de validación. El valor final de k seleccionado para cada BU corresponde al que presentó el mejor equilibrio entre estas métricas, priorizando una adecuada separación de los grupos sin incurrir en una sobre-segmentación.

Resultados técnicos del modelo

Como resultado del proceso de entrenamiento y validación, se obtuvo una solución de clustering para cada Unidad de Negocio, caracterizada por:

- un número óptimo de clusters específico para cada BU,
- métricas de calidad que evidencian distintos niveles de cohesión y separación según la naturaleza de la línea de negocio,
- y una asignación de etiquetas de cluster para cada cliente.

Tabla 16. Resultados consolidados del clustering por Unidad de Negocio — métricas escaladas a

[0–1]

Unidad de Negocio	k	Silhouette (0–1)	Calinski–Harabasz (0–1)	Davies–Bouldin invertido (0–1)	n clientes
Cloud Pública	4	0.96	0.84	0.41	48
Cyber Security	3	1.00	0.63	1.00	33
Digital Business / Aplicaciones	3	0.00	0.76	0.65	47
Digital Workplace	2	0.94	0.00	0.98	26
Private Cloud / Infra Tradicional	2	0.39	1.00	0.00	104
Soluciones SAP	5	0.23	0.71	0.82	18

Nota: Con el fin de facilitar la interpretación comparativa de los resultados, las métricas de validación interna fueron reescaladas a un rango [0.1] utilizando una transformación Min–Max. Este reescalamiento se emplea exclusivamente con fines descriptivos y no afecta el proceso de entrenamiento ni la selección del número óptimo de clusters. En el caso del índice Davies–Bouldin, la escala fue invertida para mantener coherencia interpretativa, de modo que valores cercanos a 1 indiquen una mejor segmentación.

5.1.2 RESULTADOS DE LOS CLUSTERS POR BU

Los resultados del modelo no supervisado por Unidad de Negocio constituyen el insumo fundamental para el modelo posterior de identificación y priorización de prospectos comerciales. En particular, los clusters obtenidos permiten construir perfiles estratégicos de clientes actuales, que sintetizan patrones exitosos de consumo y valor económico.

Estos perfiles serán utilizados como referencias de comparación en el modelo supervisado o basado en similitud, permitiendo evaluar el grado de cercanía entre empresas externas y los perfiles identificados, y facilitando así la priorización de oportunidades comerciales con mayor potencial.

Perfiles de Unidad de Negocio Cloud Pública

El cliente típico de Cloud Pública es una empresa grande, principalmente de operación nacional, perteneciente a sectores como Gobierno, Tecnología y Servicios Públicos. Su modelo de negocio está orientado a servicios y distribución, y utiliza la nube pública como una plataforma para

modernizar procesos, soportar cargas de trabajo críticas y habilitar esquemas de operación flexibles.

La adopción de Cloud Pública se da, en la mayoría de los casos, como una extensión o complemento de infraestructuras existentes, más que como un reemplazo total. Desde el punto de vista comercial, estos clientes presentan una alta recurrencia de oportunidades asociadas a servicios profesionales, brokers de nube (especialmente AWS) y servicios gestionados, con valores contractuales de magnitud media y potencial de crecimiento progresivo. En general, se trata de organizaciones en una fase intermedia de adopción cloud, con alta dependencia de acompañamiento técnico y oportunidades claras de expansión hacia servicios complementarios del portafolio.

Perfiles de Unidad de Negocio Infraestructura y Nube

El perfil de cliente de infraestructura y nube que emerge del análisis corresponde principalmente a empresas grandes, ubicadas en Bogotá D.C., Colombia, pertenecientes a sectores altamente críticos como salud y telecomunicaciones/energía, con modelos de negocio orientados a la prestación de servicios y distribución de servicios públicos. Estas organizaciones operan infraestructuras tecnológicas complejas, con alta exigencia de disponibilidad, seguridad y continuidad operativa, lo que las convierte en clientes naturalmente dependientes de servicios de misión crítica.

Desde la perspectiva tecnológica, este perfil muestra una alta adopción de servicios de nube privada y gestión de infraestructura tradicional, con una fuerte penetración de servicios administrados (Managed Services) y servicios profesionales especializados, especialmente en entornos VMware, OpenStack y servicios de nube privada. Estos clientes combinan esquemas híbridos que integran nube privada, colocación, conectividad, servicios de red, y en algunos casos plataformas transaccionales de alta criticidad como mainframe, evidenciando una madurez tecnológica avanzada y un fuerte requerimiento de gobierno, monitoreo y soporte 24/7.

Comercialmente, se trata de clientes de alto valor Enterprise, con volúmenes de facturación elevados y relaciones contractuales de largo plazo, que consumen servicios de infraestructura gestionada, conectividad, colocación y nube privada como parte esencial de su operación. Estos clientes representan un segmento estratégico para ofertas de servicios administrados de infraestructura, modernización a esquemas híbridos y multicloud, continuidad del negocio, NOC/SOC, y transformación hacia arquitecturas cloud resilientes.

Perfil de Unidad de Negocio – SAP

El perfil de cliente que emerge del análisis de clusters corresponde principalmente a empresas grandes, ubicadas en su mayoría en Colombia (Bogotá D.C.), con presencia relevante también en Panamá, pertenecientes a sectores altamente regulados y críticos como alimentos, petróleo y gas, salud, educación y servicios públicos. Estas organizaciones operan bajo modelos de negocio

orientados a la producción, prestación de servicios, comercialización y gestión institucional, lo que evidencia entornos operativos complejos y con alta dependencia de plataformas ERP robustas como SAP.

Desde el punto de vista de madurez tecnológica, estos clientes ya cuentan con SAP como plataforma central de operación y presentan una alta adopción de servicios profesionales funcionales, siendo este el principal eje de consumo. Una proporción significativa de estos clientes ya ha ejecutado proyectos de implementación SAP y en menor medida procesos de migración hacia S/4HANA.

Comercialmente, se trata de clientes de alto valor (Enterprise), con volúmenes de facturación acumulados elevados, lo que refleja relaciones comerciales de largo plazo y un (LTV). Estos clientes no solo consumen servicios de operación, sino que están dispuestos a invertir en acompañamiento funcional especializado, servicios BASIS, modernización de plataforma, migraciones S/4HANA, optimización de procesos y servicios profesionales avanzados, convirtiéndose en un segmento estratégico para ofertas de AMS estructurado, transformación digital, automatización, analítica avanzada y gobierno de aplicaciones SAP.

Perfil de Unidad de Negocio – Cyber Security

El cliente característico de Cyber Security corresponde principalmente a una empresa grande, mayoritariamente de operación nacional, perteneciente a sectores como Petróleo y Gas, Servicios Públicos y Tecnología, en los cuales la continuidad operativa, la protección de la información y el cumplimiento regulatorio son factores críticos. Estas organizaciones suelen concentrar sus operaciones en Bogotá D.C., con una presencia internacional moderada en algunos casos.

En cuanto al modelo de negocio, predomina el enfoque en Servicios, particularmente en esquemas de operación continua, outsourcing especializado y consultoría, donde la ciberseguridad actúa como un habilitador transversal del negocio y no como un servicio puntual. Esto se traduce en contratos recurrentes orientados a soporte, monitoreo y fortalecimiento de controles de seguridad.

Desde la perspectiva comercial, los clientes de esta área demandan principalmente servicios de seguridad gestionada, consultoría en ciberseguridad y soluciones de protección tecnológica, con una recurrencia media de oportunidades y valores contractuales asociados a la mitigación de riesgos, protección de infraestructuras críticas y cumplimiento normativo. Se trata de organizaciones con un nivel de madurez tecnológica medio–alto, donde la ciberseguridad acompaña iniciativas de cloud, infraestructura y transformación digital, reforzando la estabilidad y confiabilidad de los servicios tecnológicos.

Perfil de Unidad de Negocio – Digital

El cliente representativo de Digital Business / Aplicaciones corresponde principalmente a una empresa grande, con operación mayoritariamente nacional, perteneciente a sectores como

Servicios Públicos, Petróleo y Gas y Tecnología, en los cuales la continuidad operativa, la modernización de aplicaciones y la eficiencia de los procesos digitales son factores estratégicos para el negocio. La concentración geográfica se presenta principalmente en Bogotá D.C.

En cuanto al modelo de negocio, predomina el enfoque en Servicios, especialmente en esquemas de outsourcing, soporte funcional y gestión continua de aplicaciones empresariales. Estas organizaciones demandan capacidades orientadas al mantenimiento, evolución y modernización de sus plataformas tecnológicas, integrando aplicaciones core con soluciones complementarias de hardware y software.

Desde la perspectiva comercial, los subproductos con mayor tracción dentro de esta BU corresponden a Application Management Services (AMS), así como servicios de modernización y desarrollo de aplicaciones. Este patrón refleja una alta recurrencia de oportunidades comerciales, asociada a contratos de soporte continuo y proyectos de evolución tecnológica. El perfil del cliente de Digital Business evidencia un nivel de madurez digital medio–alto, actuando como un puente natural entre la infraestructura tecnológica y los sistemas empresariales, y posicionándose como una línea clave para habilitar la transformación digital y la integración de soluciones de negocio dentro del portafolio de la empresa.

Perfil de Unidad de Negocio – Workplace

El cliente característico de la línea Digital Workplace corresponde principalmente a una empresa grande, con operación predominantemente nacional y estructuras operativas distribuidas. Este perfil se concentra en sectores como Servicios Públicos, Petróleo y Gas y Seguros.

Desde el punto de vista del modelo de negocio, predomina el enfoque en Servicios, particularmente en esquemas de outsourcing, soporte técnico especializado y operación continua, asociados a la gestión de dispositivos, estaciones de trabajo y atención a usuarios finales. Estas organizaciones requieren soluciones que garanticen estabilidad operativa, tiempos de respuesta adecuados y estandarización en la experiencia del usuario.

En términos comerciales, los subproductos con mayor tracción dentro de esta BU corresponden a Workplace Services, Service Desk y Field Services, los cuales concentran tanto el volumen transaccional como el valor contractual.

El perfil del cliente de Digital Workplace evidencia un nivel de madurez organizacional medio–alto, con necesidades claras de soporte distribuido y gestión integral del entorno laboral digital. Esta línea se posiciona como un servicio transversal y habilitador, complementando otras áreas del portafolio como Infraestructura, Cloud y SAP, y resultando clave para la experiencia del usuario, la productividad interna y la continuidad operativa de los clientes corporativos.

El análisis de perfiles de clientes por Unidad de Negocio permitió consolidar los resultados del modelo no supervisado en descripciones estructuradas y coherentes con la realidad del negocio. A partir de los clusters obtenidos en los distintos intentos de segmentación, se construyeron

perfiles representativos que integran variables organizacionales, sectoriales, tecnológicas y comerciales, facilitando una lectura interpretativa de los patrones identificados por los algoritmos.

5.1.3 PRINCIPALES HALLAZGOS DE LOS PERFILES

Uno de los principales hallazgos del análisis de perfiles es que los clientes del portafolio no responden a un único arquetipo homogéneo, sino que presentan configuraciones diferenciadas según la unidad de negocio consumida. Cada BU concentra clientes con necesidades, niveles de madurez tecnológica y modelos operativos particulares, lo que valida la decisión metodológica de evitar una segmentación global única y avanzar hacia enfoques específicos por línea de servicio.

En el caso de Cloud Pública, el perfil identificado corresponde a organizaciones grandes, principalmente de operación nacional, que utilizan la nube como un habilitador para la modernización de procesos y la flexibilidad operativa. Estos clientes se encuentran, en su mayoría, en una fase intermedia de adopción cloud, combinando infraestructura existente con servicios de nube pública, lo que explica la recurrencia de oportunidades asociadas a servicios profesionales, brokers y esquemas de operación gestionada. Este perfil evidencia un alto potencial de crecimiento progresivo mediante la expansión hacia servicios complementarios del portafolio.

Para la Unidad de Negocio de Infraestructura y Nube Privada, el perfil resultante se caracteriza por empresas de sectores críticos como salud, telecomunicaciones y energía, con altos requerimientos de disponibilidad, continuidad y seguridad. Estos clientes presentan una madurez tecnológica avanzada y una fuerte dependencia de servicios administrados, lo que se traduce en relaciones contractuales de largo plazo y volúmenes de facturación elevados. Este perfil posiciona a la BU como estratégica para iniciativas de modernización, esquemas híbridos y servicios de misión crítica.

En la Unidad de Negocio SAP, los perfiles identificados corresponden a organizaciones grandes, pertenecientes a sectores altamente regulados, con SAP como plataforma central de operación. Estos clientes muestran una alta demanda de servicios profesionales funcionales y una disposición a invertir en procesos de modernización, migración y optimización. El perfil evidencia un elevado valor de vida del cliente (LTV) y un alto potencial para ofertas estructuradas de AMS, transformación digital y analítica avanzada sobre plataformas empresariales.

Por su parte, el perfil de Cybersecurity refleja organizaciones donde la seguridad actúa como un habilitador transversal del negocio. Predominan empresas de sectores críticos, con modelos de operación continua y una demanda recurrente de servicios gestionados, consultoría y cumplimiento normativo. Este perfil confirma que la ciberseguridad no se consume como un servicio aislado, sino como un componente integrado a iniciativas de cloud, infraestructura y transformación digital.

En el caso de Digital Business y Digital Workplace, los perfiles identificados muestran

organizaciones con un nivel de madurez digital medio–alto, orientadas a la eficiencia operativa, la modernización de aplicaciones y la gestión integral del entorno laboral digital. Estas BUs se posicionan como habilitadoras transversales del portafolio, conectando infraestructura, aplicaciones y experiencia de usuario, y presentando una alta recurrencia de oportunidades asociadas a contratos de soporte y operación continua.

En conjunto, el análisis de perfiles permitió traducir los resultados técnicos del clustering en tipologías de cliente accionables, alineadas con la estructura del portafolio y con las dinámicas reales del negocio. Estos perfiles no solo facilitan la comprensión del cliente actual, sino que constituyen un insumo clave para etapas posteriores del proyecto, al servir como referencia para la identificación de prospectos externos con características similares a los segmentos de mayor valor, madurez y potencial de crecimiento.

5.2 MODELO SUPERVISADO

A partir de los resultados obtenidos con el modelo de aprendizaje automático no supervisado, el análisis avanza hacia una segunda etapa orientada a la identificación y priorización de prospectos comerciales, cuyo objetivo es extender el conocimiento generado sobre los clientes actuales hacia escenarios de crecimiento y expansión del portafolio. Mientras que la fase anterior permitió comprender y estructurar los patrones de consumo, madurez y valor económico presentes en la base histórica de clientes, esta nueva etapa busca evaluar cómo dichos patrones pueden utilizarse como referencia para identificar oportunidades comerciales potenciales.

En este contexto, se plantea la construcción de un enfoque supervisado o basado en similitud, fundamentado en los perfiles obtenidos en la etapa de segmentación. La lógica central de este enfoque consiste en medir el grado de cercanía entre organizaciones externas y los perfiles de clientes ya existentes, asumiendo que una mayor similitud con los segmentos de mayor valor y madurez incrementa la probabilidad de éxito comercial.

Para el desarrollo de este modelo se contemplan dos alternativas metodológicas complementarias. La primera consiste en utilizar clientes que ya hacen parte de la base actual, pero que no presentan consumo en una determinada unidad de negocio. Estos clientes representan un escenario de expansión interna del portafolio, en el cual es posible evaluar la afinidad de cada organización con los perfiles característicos de una BU específica, con el fin de priorizar iniciativas de cross-selling y upselling.

La segunda alternativa corresponde a la construcción de una base de datos de potenciales clientes externos, que no forman parte del CRM actual. Dado que la disponibilidad de información estructurada sobre estas organizaciones es limitada, se plantea la generación de este dataset mediante técnicas de enriquecimiento apoyadas en modelos de lenguaje de gran escala, específicamente a través del uso de Gemini con la herramienta de Deep Research. Este enfoque permitirá recopilar y estructurar información relevante sobre empresas externas —como sector, tamaño, modelo de negocio, presencia geográfica y madurez tecnológica— de manera sistemática

y escalable, sirviendo como insumo para el posterior proceso de modelado.

En ambos casos, el propósito no es predecir de forma directa una decisión de compra, sino estimar el grado de similitud o afinidad entre las organizaciones analizadas y los perfiles de clientes identificados en la etapa no supervisada. De esta manera, el modelo supervisado o basado en similitud se concibe como una herramienta de apoyo a la toma de decisiones comerciales, orientada a priorizar esfuerzos de prospección y a reducir la incertidumbre en la identificación de oportunidades con mayor potencial de conversión

5.2.1 ESTRATEGIA DE CROSS SELLING

5.2.1.1 Surgimiento de la Necesidad de Negocio

Tras la ejecución del modelo no supervisado y la consolidación de los perfiles estratégicos, se identificó una oportunidad crítica de alto valor estratégico para la organización: si bien el objetivo principal del proyecto se centra en la segmentación de clientes actuales y la identificación de prospectos externos, el análisis de los clústeres reveló que existe una brecha de penetración significativa dentro de la propia base instalada, donde clientes con potencial de consumir BUs adicionales no han sido abordados con una estrategia sistemática de cross-selling.

A partir de los resultados del modelo de segmentación no supervisada, fue posible identificar perfiles de clientes diferenciados en términos de comportamiento comercial, nivel de adopción de servicios y valor económico. Estos perfiles permitieron comprender no solo quiénes son los clientes más relevantes, sino también cómo interactúan con las diferentes unidades de negocio (BU) del portafolio.

En este contexto, surgió una necesidad adicional desde el negocio:

Determinar en qué otras unidades de negocio podrían expandirse los clientes actuales, con el fin de diseñar estrategias de crecimiento sobre la base instalada.

Aunque este enfoque no estaba contemplado inicialmente dentro de los objetivos específicos del proyecto, se identifica como una extensión natural del análisis, ya que una vez definidos los perfiles estratégicos, es posible evaluar su potencial de adopción de nuevos servicios.

De esta manera, se incorpora un modelo de *cross-selling* orientado a responder la siguiente pregunta:

¿Qué clientes actuales tienen mayor probabilidad de consumir otras unidades de negocio del portafolio?

5.2.1.2 Formulación del problema

El problema de *cross-selling* se modela como una tarea de **clasificación binaria supervisada**, en la cual se busca predecir la probabilidad de que un cliente adopte una nueva unidad de negocio.

Para ello, se define una variable objetivo de la siguiente forma:

- **1:** el cliente consume la unidad de negocio
- **0:** el cliente no consume la unidad de negocio

Este planteamiento permite transformar una necesidad comercial en un problema predictivo, facilitando la priorización de clientes con mayor potencial de expansión.

Datos y variables utilizadas

El modelo se construye a partir del dataset `clientes_bda_model.csv`, descrito previamente en la Sección 4.2 Caracterización de la base de datos. Las variables empleadas son las mismas allí presentadas, agrupadas en tres categorías: comportamiento comercial (`opp_count`, `acv_total`, `precio_total_sum`, `years_present`), consumo por línea de servicio (`has_*`, `acv_*`, `total_*`, `tvf_count_*`) y caracterización empresarial (sector económico, tamaño de empresa, modelo de negocio, presencia internacional). Para una descripción detallada de cada variable, ver Tabla 1 del Capítulo 4.

Preparación de los datos

Previo al entrenamiento de los modelos, se realizó un proceso de preparación que incluyó:

- Imputación de valores faltantes:
 - Mediana para variables numéricas
 - Moda para variables categóricas
- Codificación de variables categóricas mediante One-Hot Encoding
- Escalamiento de variables numéricas para evitar sesgos en el modelo
- División del dataset en:

Con el fin de evaluar la capacidad de generalización de los modelos y evitar problemas de sobreajuste (overfitting), el dataset fue dividido en dos subconjuntos:

- Conjunto de entrenamiento (train): 70% de los datos
- Conjunto de prueba (test): 30% de los datos

El conjunto de entrenamiento se utilizó para ajustar los modelos y aprender los patrones presentes en los datos, mientras que el conjunto de prueba se reservó exclusivamente para evaluar el desempeño del modelo sobre datos no vistos previamente.

Esta división se realizó de manera aleatoria, garantizando que ambos conjuntos mantuvieran una distribución representativa de la variable objetivo. Adicionalmente, se utilizó una semilla (`random_state`) fija para asegurar la reproducibilidad de los resultados. Este proceso garantiza la consistencia de los datos y permite evaluar la capacidad de generalización de los modelos.

5.2.1.3 Desarrollo iterativo del modelo: tres fases metodológicas

El modelo de cross-selling no se diseñó en una única iteración, sino que evolucionó a través de tres fases progresivas, cada una corrigiendo limitaciones de la anterior. Este proceso iterativo es coherente con la naturaleza exploratoria de la ciencia de datos aplicada, donde el refinamiento metodológico se basa en los hallazgos de cada etapa previa. A continuación se describe cada fase con sus fundamentos, variables y configuraciones técnicas.

Fase 1: Similitud Coseno sobre el perfil transaccional

La primera aproximación al problema de cross-selling se basó en la similitud coseno entre el vector de comportamiento comercial de cada cliente no-consumidor y el centroide del cluster estratégico de la BU objetivo. El cluster estratégico se definió como aquel con el mayor ACV promedio entre los clientes que ya consumen la BU, lo que lo convierte en el grupo de mayor valor de referencia.

Variables utilizadas:

Se emplearon exclusivamente variables transaccionales del CRM: ACV total del cliente (`acv_total`), número de oportunidades ganadas (`opp_count`), años de relación comercial (`years_present`), número de líneas de producto (`n_lineas_producto`), y el ACV acumulado en cada una de las demás BUs (`acv_bu_cyber_security`, `acv_bu_digital`, `acv_bu_private_cloud`, `acv_bu_public_cloud_infra`, `acv_bu_workplace`). Estas variables reflejan el histórico de relación comercial de cada cliente, prescindiendo de variables firmográficas.

Proceso:

- Para cada BU, se separaron los clientes que la consumen (`consume_[bu] = 1`) de los que no (`consume_[bu] = 0`).
- Se identificó el cluster estratégico calculando el ACV promedio por cluster entre los consumidores y seleccionando el de mayor valor.
- Las variables se estandarizaron con `StandardScaler` y se calculó el centroide del cluster estratégico en el espacio escalado.
- Se aplicó la función `cosine_similarity` entre los vectores de los no-consumidores y el centroide, generando el `Score_Afinidad_[BU]` con rango `[-1, +1]`.

Esta fase se aplicó a las seis BUs del portafolio y produjo los scores que se presentan en la tabla de resultados. Los valores positivos altos indican gran alineación del perfil comercial del cliente con el segmento estratégico; los valores negativos indican lo contrario, es decir, un perfil de consumo divergente respecto al cluster de referencia.

Fase 2: Regresión Logística con integración de asignaciones de cluster (K-Means)

En la segunda fase se reformuló la función de scoring incorporando las asignaciones de cluster del modelo no supervisado como variables predictoras explícitas. Para cada BU, los labels del K-Means correspondiente se transformaron mediante `One-Hot Encoding`, generando columnas binarias

que representan la pertenencia de cada cliente a un segmento específico. Esto permitió que el modelo capturara directamente la información estructural del clustering previo.

El modelo utilizado fue Regresión Logística (max_iter=1000) entrenada sobre el conjunto completo de clientes. Las variables predictoras incluyeron las mismas transaccionales de la Fase 1 más las dummies de cluster. El score de propensión se obtuvo mediante predict_proba, garantizando salidas en el rango [0, 1]. Este enfoque permitió una primera integración explícita entre el modelo supervisado y los resultados del clustering, sin embargo, presentaba una limitación metodológica relevante: la inclusión de variables directamente derivadas del consumo de la BU objetivo podía introducir data leakage en el proceso de predicción, aspecto que se abordó en la fase siguiente.

Fase 3: Random Forest y XGBoost con control de data leakage (versión definitiva)

La tercera fase constituyó el modelo definitivo del cross-selling y representó el avance metodológico más significativo. El principal desafío identificado en las fases anteriores fue el riesgo de data leakage: si se incluyen en el modelo variables directamente asociadas al consumo de la BU objetivo (por ejemplo, el ACV de la misma BU o sus indicadores has_*), el modelo puede aprender trivialmente la relación entre esa variable y la etiqueta, produciendo scores inflados que no generalizan a nuevos clientes. Para resolver esto, se diseñó una función de selección de variables libre de fuga denominada variables_sin_leakage.

Concepto de variables_sin_leakage:

La función variables_sin_leakage construye el conjunto de predictores del modelo siguiendo una lógica de exclusión selectiva: para predecir la adopción de una BU objetivo, se incluyen todas las variables disponibles EXCEPTO las que están directamente asociadas a esa misma BU. Esto garantiza que el modelo aprenda de patrones genuinamente predictivos, sin acceso a información que derivaría en una predicción trivial.

El conjunto de variables utilizado en esta fase se compone de cuatro grupos:

- **Variables estructurales del CRM:** years_present (años de relación comercial), opp_count (oportunidades ganadas), n_lineas_producto (diversidad del portafolio consumido). Estas variables reflejan la profundidad y longevidad de la relación comercial.
- **Variables firmográficas:** Sector económico principal, Modelo de negocio, Modelo normalizado, Submodelo, Tamaño de la empresa, Presencia internacional, Tipo de empresa, Subtipo de empresa, Ciudad y País. Estas variables describen el contexto organizacional del cliente.
- **ACV de otras BUs (variable clave):** Para cada BU objetivo, se incluye el ACV acumulado en TODAS las demás BUs (por ejemplo, para predecir SAP, se usa el ACV de Cloud, Cyber, Digital, Workplace y Private Cloud, pero NO el ACV de SAP). Esta variable captura el patrón de co-adopción: clientes con alto consumo en múltiples líneas del portafolio tienen mayor probabilidad de adoptar nuevas. Es la variable con mayor poder predictivo en este modelo.
- **Dummies de cluster K-Means:** Las asignaciones del modelo no supervisado por BU,

codificadas como variables binarias mediante One-Hot Encoding. Esto integra explícitamente el conocimiento generado en la etapa de segmentación dentro del modelo predictivo.

Partición del conjunto de datos y composición de la muestra

Para evaluar el desempeño de cada algoritmo, el conjunto de 167 clientes se dividió en 70% entrenamiento y 30% prueba con estratificación por clase, aplicando una semilla fija (`random_state=42`) para garantizar reproducibilidad. La composición exacta de la partición por BU se detalla en la Tabla 18 (Sección Modelo supervisado para prospectos externos), dado que el modelo de cross-selling y el de prospectos externos comparten la misma base de 167 clientes y el mismo esquema de partición 70/30 estratificada. El desbalance de clases en SAP (10.8%) y Digital Workplace (15.6%) fue compensado mediante ponderación de clases en los algoritmos.

Configuración de los modelos:

Random Forest (`n_estimators=400`, `class_weight='balanced_subsample'`, `random_state=42`): construye 400 árboles independientes sobre muestras bootstrap, ajustando los pesos de clase en cada submuestra para compensar el desbalance. La importancia de variables se extrae como promedio de la reducción de impureza de Gini a través de todos los árboles, proporcionando una medida interpretable de qué variables explican mejor la adopción de cada BU.

XGBoost (`n_estimators=500`, `learning_rate=0.05`, `max_depth=4`, `subsample=0.9`, `colsample_bytree=0.9`, `random_state=42`): construye árboles secuenciales corrigiendo los errores residuales del paso anterior, con regularización implícita a través de la tasa de aprendizaje y la profundidad máxima. La combinación de `subsample` y `colsample_bytree` introduce aleatoriedad que reduce el sobreajuste.

Evaluación y selección del mejor modelo:

Para cada BU se calculó el AUC-ROC sobre el conjunto de prueba para cada algoritmo. El modelo con mayor AUC fue seleccionado para generar los scores de propensión finales. Adicionalmente, se generaron gráficos de importancia de variables (Top 15) para cada BU y cada algoritmo, que permiten identificar qué atributos firmográficos y de comportamiento comercial son más determinantes en la predicción de la adopción de cada línea del portafolio.

Generación de scores de cross-selling:

Una vez seleccionado el mejor algoritmo por BU, el modelo entrenado se aplicó sobre la totalidad de los clientes de esa BU mediante `predict_proba`, obteniendo un score de propensión en el rango $[0, 1]$ para cada instancia. Solo se conservaron los scores de los clientes con `consume_[bu] = 0` (no-consumidores), pues son los únicos candidatos al cross-selling en esa dimensión. El score total por cliente se calculó como la suma de los scores individuales en todas las BUs que no consume, sirviendo como criterio de ordenamiento del ranking estratégico.

Consolidación multi-BU

Los rankings individuales de las seis BUs se consolidaron en una única tabla mediante `left joins` sucesivos sobre el identificador de cliente (`account_id`). Este proceso garantiza que cada cliente

aparezca en la tabla con el score de cada BU que no consume; las BUs que el cliente ya consume quedan como NaN. La suma de scores no nulos constituye el Score_Total_Potencial, que ordena a los clientes según su oportunidad de expansión más amplia dentro del portafolio.

5.2.1.4 Resultados

Lectura de la tabla de resultados

La Tabla 16 presenta los 20 clientes de la base instalada con mayor potencial de cross-selling, según los scores generados en la Fase 1 (similitud coseno). Cada celda representa el score de afinidad del cliente con el cluster estratégico de esa BU; las celdas con valor NaN indican que el cliente YA consume esa BU y por tanto no es candidato de cross-selling en esa dimensión. Los scores son valores de similitud coseno en el rango $[-1, +1]$, donde +1 indica máxima alineación con el perfil del segmento de mayor valor, 0 indica ortogonalidad (perfil neutro), y valores negativos indican que el comportamiento comercial del cliente apunta en dirección opuesta al cluster estratégico de esa BU. El Score Total es la suma de los scores positivos y negativos en todas las BUs que el cliente no consume, priorizando aquellos con mayor oportunidad agregada.

Tabla 17. Top 20 clientes con mayor potencial de cross-selling según score de propensión agregado

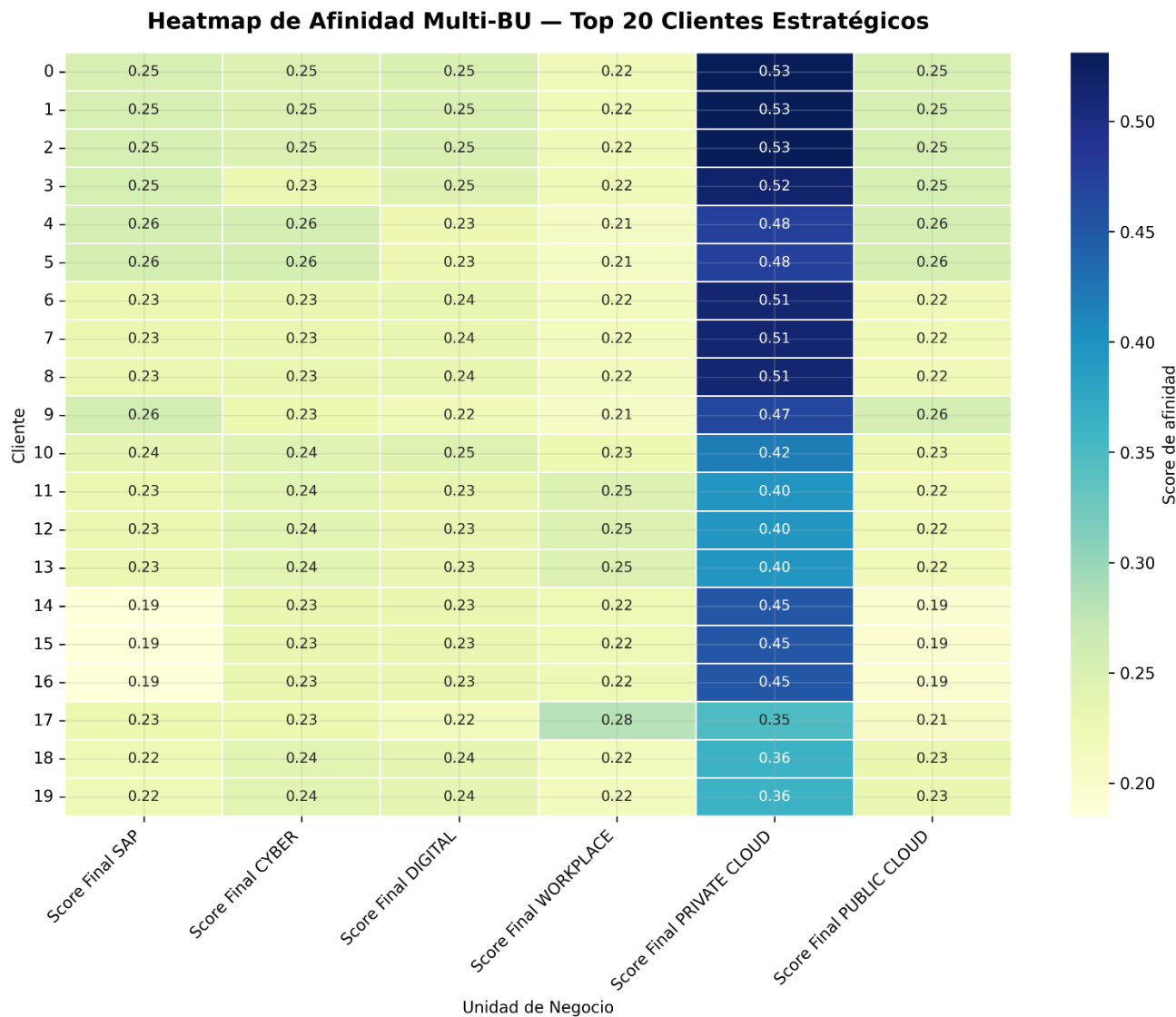
Account_id	account_name	Nombre Formal	SAP	CYBER	DIGITAL	WORKPLACE	PUBLIC CLOUD	PRIVATE CLOUD	TOTAL
9	codensa esp	CODENSA SA ESP	0.66308	0.57327	0.63639	0.369503	NaN	NaN	2.24223
29	nueva empresa promotora de salud 1	NUEVA EMPRESA PROMOTORA DE SALUD 1	0.73006	NaN	0.74719	0.395488	NaN	NaN	1.87273
11	cruz verde	CRUZ VERDE	0.6975	NaN	0.71086	0.40533	NaN	NaN	1.81369
7	clinica del country	CLINICA DEL COUNTRY	NaN	0.65411	0.6802	0.372911	NaN	NaN	1.70723
95	instituto nacional de cancerologia	INSTITUTO NACIONAL DE CANCEROLOGIA	NaN	0.62599	0.62804	0.316644	NaN	NaN	1.57068
137	indra sistemas sucursal colombia	INDRA SISTEMAS SA SUCURSAL COLOMBIA	0.42471	0.60506	0.43663	-0.018062	NaN	NaN	1.44833

1	adecco colombia	ADECCO COLOMBIA SA	0.36009	0.27923	0.36876	0.352518	NaN	NaN	1.3606
96	internexa	INTERNEXA SA	0.35231	0.30289	0.37763	0.248573	NaN	NaN	1.2814
39	universidad externado de colombia	Universidad Externado de Colombia	NaN	0.02941	0.18824	0.319982	NaN	0.72585	1.26347
116	superintendencia financiera de colombia	Superintendencia Financiera de Colombia	0.57455	NaN	0.57737	0.10378	NaN	NaN	1.2557
14	empresa de telecomunicaciones de bogota sa esp	EMPRESA DE TELECOMUNICACIONES DE BOGOTÁ SA ESP	0.5557	NaN	0.59539	NaN	NaN	NaN	1.15109
22	grupo vanti	Grupo VANTI	NaN	NaN	0.74944	0.400823	NaN	NaN	1.15026
65	inassa	INASSA	0.11702	0.00177	0.1203	0.275251	NaN	0.62899	1.1398
10	colmedica medicina prepagada	COLMÉDICA MEDICINA PREPAGADA SA	0.2949	0.17623	0.30793	0.340238	NaN	NaN	1.11929
3	axa colpatria seguros	AXA COLPATRIA SEGUROS S.A.	0.28858	0.10571	0.30998	NaN	NaN	0.39556	1.09983
52	crediservir	Crediservir	-0.1071	0.16017	-0.0769	-0.043725	0.76183	0.65222	1.02619
86	certicamara	CERTICAMARA S.A.	0.25792	0.16216	0.28713	0.256959	NaN	NaN	0.96418
26	isagen esp	ISAGEN SA ESP	0.30653	0.15699	0.32493	NaN	NaN	0.12224	0.91068
6	cenit transporte y logistica de hidrocarburos	CENIT TRANSPORTE Y LOGISTICA DE HIDROCARBUROS	NaN	NaN	NaN	0.381708	NaN	0.51626	0.89796

15	ensa antes elektra	Ensa (antes Elektra)	NaN	0.2187 7	0.3935 2	NaN	NaN	NaN	0.6122 8
----	-----------------------	----------------------	-----	-------------	-------------	-----	-----	-----	-------------

La Figura 3 corresponde a un heatmap de doble entrada (clientes $\times$ BUs) generado con la librería seaborn, que muestra los scores de afinidad de los top 20 clientes estratégicos para cada una de las seis Unidades de Negocio del portafolio. El gradiente de color YlGnBu (amarillo-verde-azul) permite identificar visualmente la intensidad de la oportunidad: tonos azules oscuros representan alta afinidad (scores cercanos a +1), tonos amarillos representan baja afinidad o ausencia de potencial, y las celdas con valor 0 corresponden a BUs que el cliente ya consume (graficadas con el valor 0 al completar NaN).

Figura 3. Heatmap de Afinidad Multi-BU — Top 20 Clientes Estratégicos



La lectura del heatmap permite extraer tres tipos de insight simultáneos: (1) clúster de oportunidad por cliente: filas con múltiples celdas de tono azul identifican candidatos con brecha de expansión en varias BUs al mismo tiempo; (2) BUs con mayor demanda latente: columnas con alta intensidad generalizada revelan líneas de servicio con mayor potencial de penetración en la base instalada; y (3) patrones de co-afinidad: clientes con alta afinidad en SAP tienden a presentar también afinidad en Digital Business, lo que sugiere perfiles organizacionales donde ambas BUs son complementarias y pueden abordarse en una misma propuesta comercial.

El análisis del modelo incluyó adicionalmente un heatmap de correlación entre los scores de afinidad de las distintas BUs (paleta coolwarm, centrada en cero), que permite identificar qué líneas del portafolio tienden a ser adoptadas de forma conjunta. Una correlación positiva alta entre dos BUs indica que los clientes con afinidad por una también la presentan por la otra, orientando el diseño de propuestas comerciales combinadas y paquetes de servicios integrados para maximizar el valor por cuenta.

Figura 4. *Correlación entre Scores de Afinidad por BU*



El heatmap de correlación entre scores de afinidad (Figura 4) revela un hallazgo metodológico de primera importancia: SAP, Cyber Security y Digital Business presentan correlaciones entre sí de 0.95–1.00, lo que indica que el mismo cliente que aparece como candidato de cross-selling para una de estas BUs tiende a aparecer también para las otras dos. Esta alta correlación tiene dos

implicaciones prácticas. Primera, la priorización comercial puede hacerse de forma conjunta: un ejecutivo de cuenta que identifica a un cliente con potencial para Digital Business puede también presentarle propuestas de SAP y Cyber en la misma conversación. Segunda, los scores agregados de clientes como CODENSA SA ESP (2.242) y NUEVA EPS (1.873) son impulsados principalmente por esta sinergia entre tres BUs complementarias. En contraste, Private Cloud presenta correlación prácticamente nula con SAP (-0.017) y muy baja con Cyber Security (0.061), y Cloud Pública también muestra correlaciones bajas con esas tres BUs, confirmando que estas BUs responden a perfiles de cliente distintos y deben abordarse con estrategias comerciales independientes.

Interpretación y conclusiones del modelo de cross-selling

Los resultados obtenidos con el modelo de cross-selling constituyen una herramienta de gestión comercial concreta y accionable para los equipos de venta de la filial Colombia. A partir del entrenamiento de Random Forest como algoritmo principal en cada una de las seis Unidades de Negocio del portafolio, complementado con XGBoost en las tres BUs donde el desbalance o el tamaño muestral hacían pertinente explorar un contraste de boosting (Cloud Pública, Private Cloud y Soluciones SAP), se generó un ranking de los clientes actuales con mayor potencial de expansión. A continuación se presentan los principales hallazgos.

Patrones dominantes en el ranking

Los 20 clientes con mayor score agregado de cross-selling pertenecen mayoritariamente al sector salud (Clínica del Country, Cruz Verde, Nueva EPS, Instituto Nacional de Cancerología, Colmédica) y al sector de servicios públicos y energía (Codensa, Grupo Vanti, Isagén, ENSA). Estos sectores concentran las mayores brechas de adopción en BUs de alto valor como SAP, Digital Business y Digital Workplace, lo que sugiere que su perfil organizacional —empresas grandes, con operaciones complejas y demanda de transformación digital— es coherente con los patrones de consumo aprendidos por los modelos.

Las BUs con mayor potencial de cross-selling identificado son Digital Business / Aplicaciones y Soluciones SAP, que aparecen como oportunidades latentes en la mayoría de los clientes del ranking. Digital Workplace surge como una tercera palanca de crecimiento frecuente, especialmente en organizaciones con estructuras de personal distribuido. Por su parte, Cloud Pública y Private Cloud presentan scores más bajos en este grupo, posiblemente porque estos clientes ya tienen una base de infraestructura consolidada.

Interpretación de los scores

Los scores de propensión presentados en la tabla representan la probabilidad estimada —según el mejor modelo por BU— de que cada cliente adopte esa línea del portafolio, condicionada a su perfil actual. Scores superiores a 0.60 indican una alta afinidad con el perfil de clientes que ya consumen esa BU; scores entre 0.40 y 0.60 representan oportunidades a desarrollar con mayor esfuerzo comercial; y valores negativos o cercanos a cero indican baja alineación con ese segmento. Los valores NaN corresponden a BUs que el cliente ya consume, por lo que no son candidatos de cross-selling en esa dimensión.

El score total agregado permite priorizar esfuerzos comerciales: los clientes con mayor score total son aquellos donde se identifican múltiples brechas de adopción con alta probabilidad individual, representando las mayores oportunidades de expansión del portafolio sobre la base instalada.

Figura 5. Variables más importantes — Random Forest (Soluciones SAP)

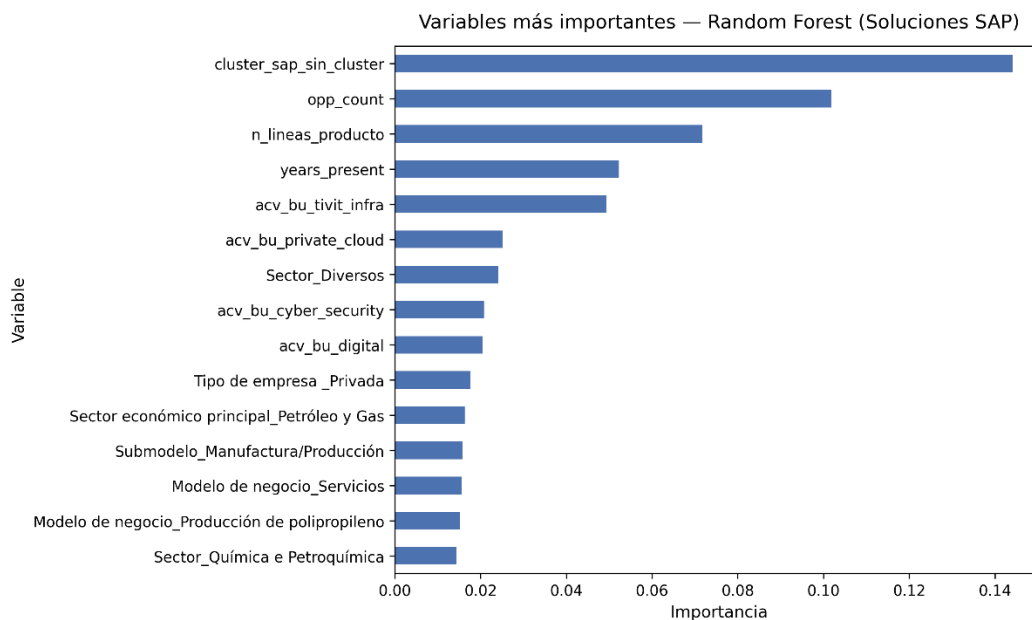


Figura 6. Variables más importantes — Random Forest (Cyber Security)

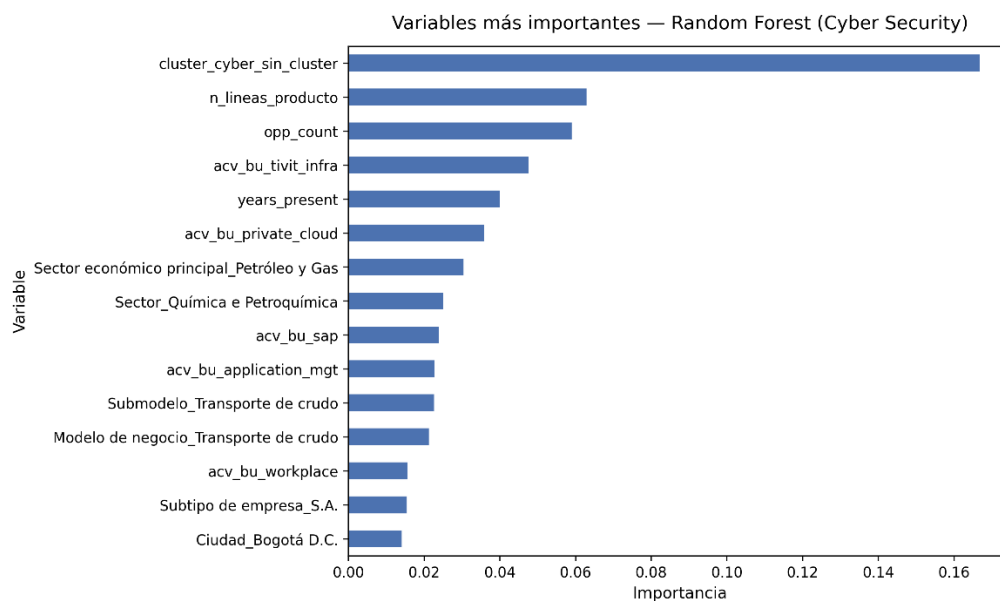


Figura 7. Variables más importantes — Random Forest (Digital Business)

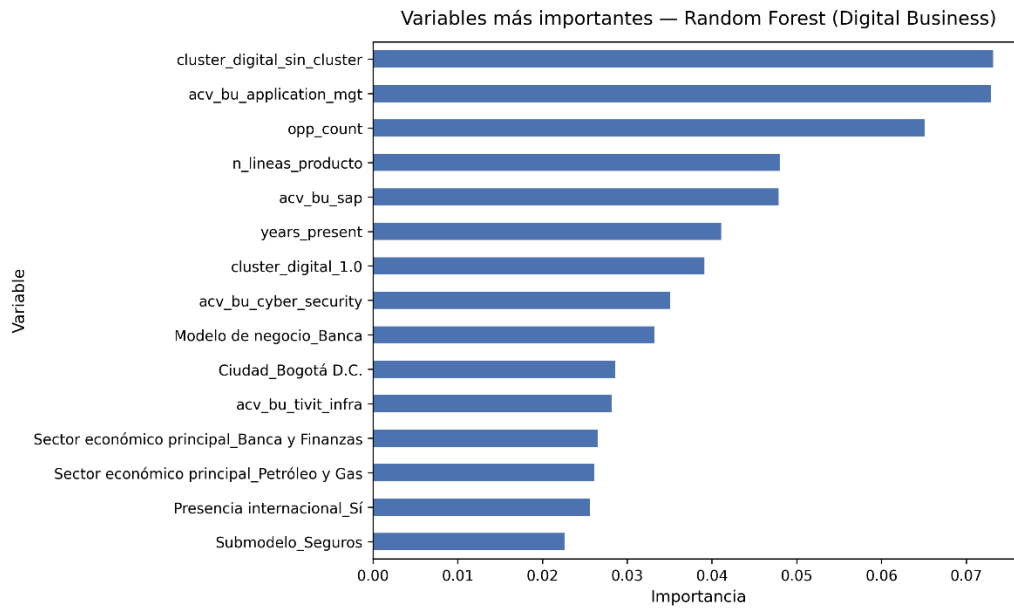


Figura 8. Variables más importantes — Random Forest (Digital Workplace)

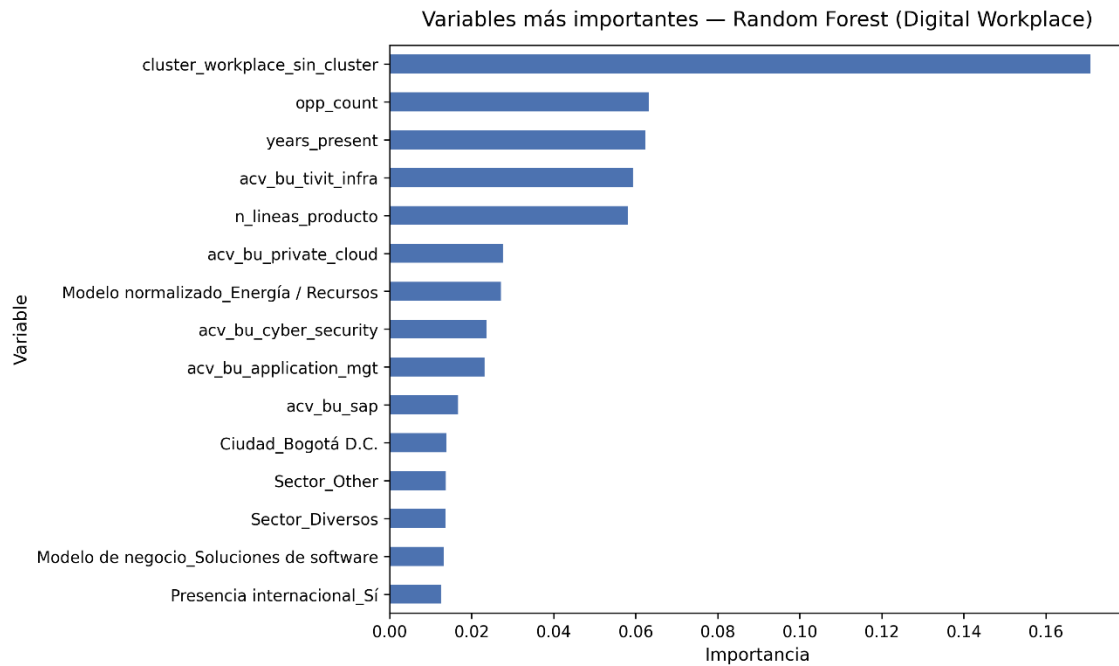


Figura 9. Variables más importantes — Random Forest (Private Cloud)

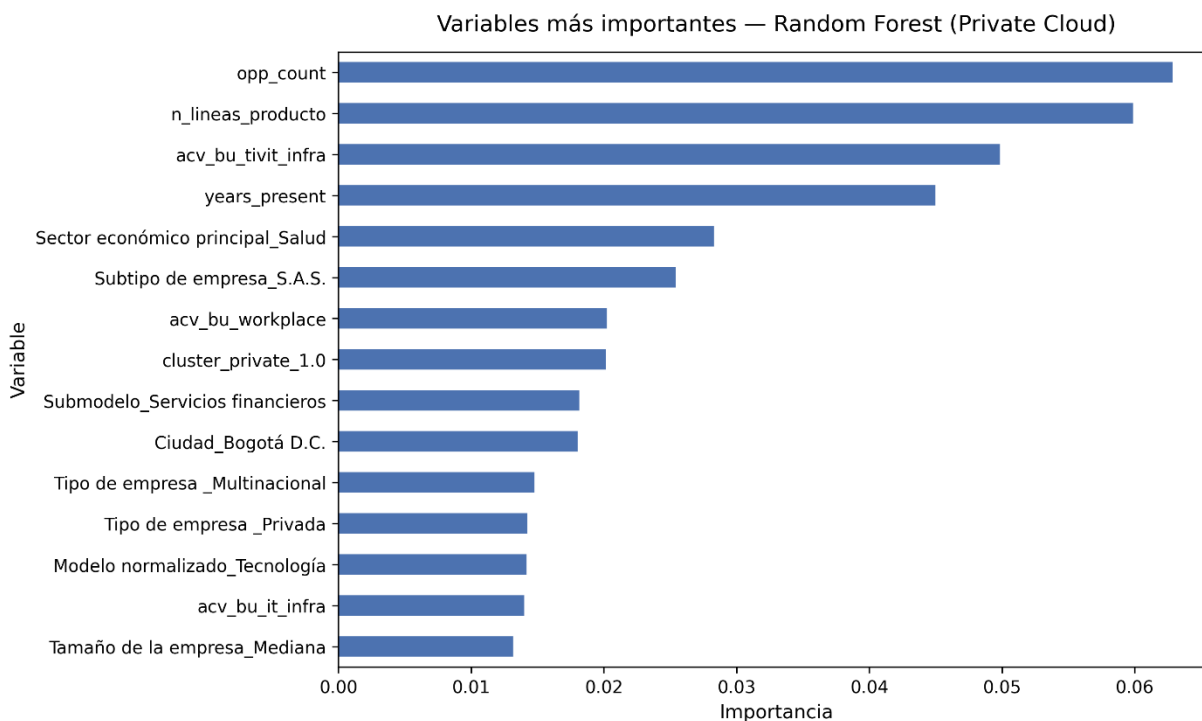
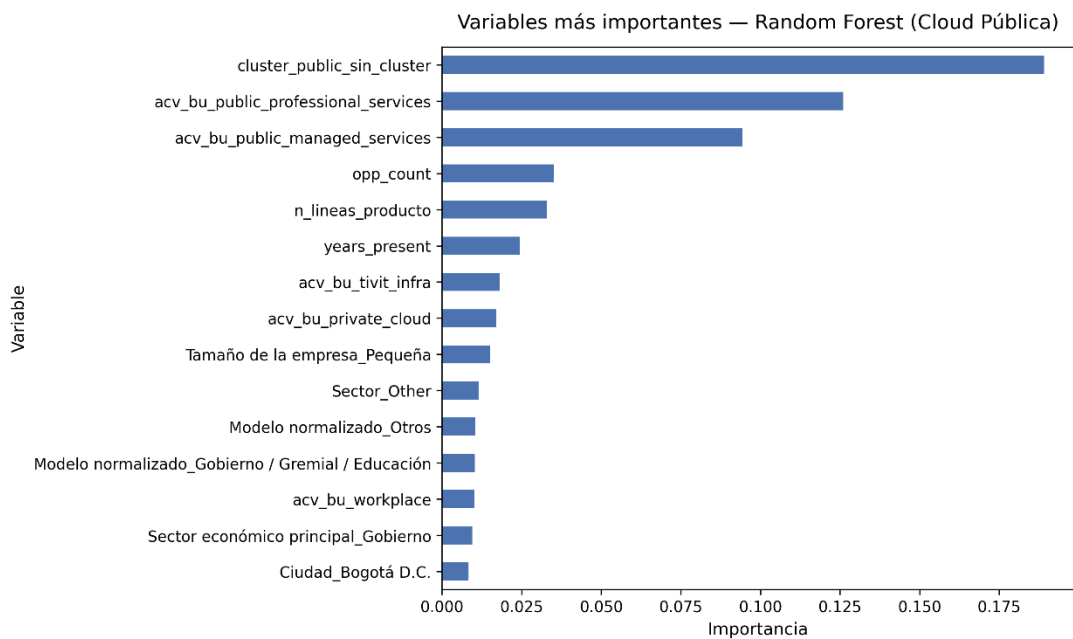


Figura 10. Variables más importantes — Random Forest (Cloud Pública)



Interpretación de las variables más importantes por BU (Random Forest)

Las gráficas de importancia de variables del modelo Random Forest revelan patrones consistentes y diferenciados por BU. En todas las unidades de negocio, la variable de mayor peso es la dummy del clúster sin variable de fuga (`cluster_BU_sin_cluster`), lo que confirma que el clúster de segmentación no supervisada es el predictor más poderoso del comportamiento de cross-selling: conocer a qué segmento pertenece un cliente predice mejor que cualquier otra variable si ese cliente tiene potencial para una BU adicional.

A continuación, el número de oportunidades (`opp_count`) y los años de relación comercial (`years_present`) aparecen consistentemente como las variables de mayor importancia en casi todas las BUs. Esto indica que la profundidad de la relación comercial —cuántas veces ha comprado y por cuánto tiempo— es un predictor robusto de la propensión al cross-selling, independientemente del portafolio objetivo. El número de líneas de producto (`n_lineas_producto`) refuerza esta lógica: los clientes que ya consumen múltiples BUs son más propensos a incorporar una adicional.

Las diferencias entre BUs aparecen en las variables de tercer y cuarto nivel. Para SAP, el sector Petróleo y Gas y el submodelo Manufactura/Producción tienen importancia significativa, coherente con el perfil sectorial de los clientes SAP existentes. Para Cyber Security, el sector Petróleo y Gas y Química/Petroquímica son relevantes. Para Private Cloud, el sector Salud y el subtipo S.A.S. aparecen como diferenciales, reflejando que las empresas del sector salud y las sociedades simplificadas tienen mayor propensión hacia este servicio. Para Digital Business, el ACV de AMS (Application Management Services) es una variable de importancia alta, indicando que la adopción previa de servicios de aplicaciones predice bien la incorporación de nuevas líneas digitales.

Limitaciones y consideraciones metodológicas

Si bien el modelo de cross-selling ofrece resultados sólidos como herramienta de priorización, deben tenerse en cuenta tres limitaciones relevantes. Primera, el tamaño muestral de $N = 167$ clientes impone restricciones sobre la estabilidad de los modelos, especialmente en BUs con pocos positivos como SAP (18 clientes) y Digital Workplace (26 clientes), donde los conjuntos de prueba son muy pequeños para estimar el AUC con alta confianza estadística. Segunda, el modelo aprende únicamente de variables observables en el CRM y la caracterización externa; factores como la disposición presupuestaria, los ciclos de renovación contractual o las relaciones personales entre ejecutivos no están representados. Tercera, los scores se generaron sobre un corte histórico de datos (2018–2025) y deben actualizarse periódicamente a medida que los clientes evolucionan su relación con la empresa.

Alcance estratégico del modelo

El modelo de cross-selling representa una primera instancia de inteligencia comercial basada en datos para la empresa. Permite transformar el conocimiento informal de los ejecutivos de cuenta en un proceso sistemático y reproducible de identificación de oportunidades. Su implementación en el ciclo comercial regular —con actualizaciones trimestrales o semestrales del modelo

conforme nuevos datos del CRM estén disponibles— constituiría una herramienta de apoyo continuo a la gestión de cartera. Los resultados obtenidos en esta etapa sientan las bases metodológicas y técnicas para la segunda fase del modelo supervisado: la identificación y priorización de prospectos externos, descrita en la siguiente sección.

5.2.2 MODELO SUPERVISADO PARA PROSPECTOS EXTERNOS

El modelo supervisado fue desarrollado con el propósito de estimar la probabilidad de consumo de las diferentes unidades de negocio ofrecidas por la organización. Este enfoque permite aprovechar la información histórica de los clientes actuales registrada en el CRM Salesforce para identificar patrones de comportamiento y, posteriormente, aplicar dicho conocimiento en la detección de oportunidades de venta cruzada y en la priorización de prospectos externos.

A diferencia de los modelos no supervisados, que se enfocan en la segmentación de clientes sin una variable objetivo, el enfoque supervisado permite predecir explícitamente un resultado de interés. En este caso, la variable objetivo corresponde al consumo de una unidad de negocio específica, lo que transforma el problema en una tarea de clasificación binaria.

Con el propósito de identificar nuevas oportunidades comerciales alineadas con los perfiles estratégicos definidos previamente, se desarrolló un proceso de construcción y depuración de una base de datos de empresas externas. Esta etapa tuvo como finalidad consolidar un conjunto de prospectos potenciales, garantizando que dichas organizaciones no correspondieran a clientes ya existentes en el sistema CRM Salesforce.

Con el fin de identificar nuevas oportunidades, la base de prospectos externos se construyó utilizando Gemini en modo Deep Research como herramienta de inteligencia de mercado. El objetivo del prompt fue generar una base estructurada de las 500 empresas más grandes de Colombia y Panamá, priorizadas por ingresos, activos, relevancia de mercado y tamaño organizacional. La estructura de salida exigió 11 campos firmográficos con taxonomía normalizada (tipo de empresa, tamaño, sector económico, modelo de negocio, presencia internacional, entre otros), aplicando reglas de consistencia semántica para evitar duplicados y categorías ambiguas. El prompt completo se incluye en el Anexo B de este documento.

Validación de la información enriquecida

Reconociendo el riesgo de alucinación inherente a los modelos de lenguaje, el enriquecimiento firmográfico se diseñó bajo un criterio de trazabilidad verificable. El prompt exigió que, junto con cada caracterización, el modelo entregara los enlaces de las fuentes consultadas, priorizando sitios oficiales, superintendencias, LinkedIn corporativo y prensa económica. A partir de esas tablas con sus respectivas fuentes, se realizó una verificación accediendo a una muestra de los enlaces para comprobar que la información reportada —sector económico, tipo y tamaño de

empresa, presencia internacional— coincidiera efectivamente con el contenido de la fuente. La información ya verificada se consolidó y normalizó en la base depurada utilizada para el modelado; los campos que presentaron baja completitud o confiabilidad insuficiente (“Número de empleados”, “Sitio web oficial” y “Año de fundación”) se excluyeron del conjunto de variables predictoras. Para los clientes actuales, la trazabilidad se conserva mediante el sitio oficial de cada entidad. Como comprobación adicional se verificó una muestra de 15 prospectos contrastando el sector reportado contra fuentes públicas: este coincidió en el 80% de los casos (12 de 15) y en el 100% de las empresas que pudieron localizarse, sin discrepancias confirmadas. Dado que se trata de grandes empresas de alta visibilidad pública, el riesgo de error es intrínsecamente bajo. De este modo, la información asistida por inteligencia artificial no se incorporó como dato no auditado, sino como insumo sujeto a contraste contra su fuente primaria.

Para el desarrollo de esta fase se integraron dos fuentes principales de datos:

1. **Base de clientes actuales:**
Proveniente del archivo *clientes_bda_model.csv*, que contiene la información consolidada de los clientes activos de la organización, incluyendo variables de comportamiento comercial y caracterización empresarial.
2. **Base externa de empresas:**
Correspondiente al archivo *DATABASE.xlsx*, que reúne un conjunto de organizaciones identificadas como posibles prospectos comerciales.

El objetivo fue comparar ambas fuentes para excluir aquellas empresas que ya forman parte de la base de clientes, asegurando así la identificación de prospectos verdaderamente nuevos.

Normalización de los nombres de las empresas

Uno de los principales desafíos en la integración de datos provenientes de distintas fuentes es la variabilidad en la forma en que se registran los nombres de las empresas. Diferencias en el uso de mayúsculas, tildes, caracteres especiales o sufijos legales pueden dificultar la correcta identificación de coincidencias entre registros.

Para resolver esta problemática, se implementó un proceso de normalización de texto, el cual permitió estandarizar los nombres de las empresas en ambas bases de datos. Este procedimiento incluyó:

- Conversión de todos los nombres a minúsculas.
- Eliminación de tildes y caracteres diacríticos.
- Sustitución de caracteres especiales por espacios.
- Eliminación de sufijos legales y términos poco informativos, tales como S.A.S., S.A., Ltda., Inc., Corp., Compañía, Sucursal y Colombia.
- Eliminación de espacios múltiples.

Como resultado, se generó una variable estandarizada denominada **empresa_norm**, utilizada como clave para la comparación entre las dos fuentes de información.

Exclusión de clientes existentes

Una vez normalizados los nombres, se llevó a cabo la depuración de la base externa mediante un proceso de exclusión en dos niveles:

- **Exclusión por coincidencia exacta**

En una primera etapa, se identificaron y eliminaron aquellas empresas cuyo nombre normalizado coincidía exactamente con los registros existentes en la base de clientes del CRM. Este procedimiento permitió eliminar de manera eficiente las duplicidades más evidentes.

- **Exclusión por coincidencia parcial**

Debido a que algunas empresas pueden presentar variaciones en su denominación comercial o razón social, se implementó una segunda capa de validación basada en **coincidencias parciales**. Este enfoque permitió identificar casos en los que el nombre de una empresa estaba contenido dentro de otro o presentaba ligeras variaciones, asegurando una depuración más exhaustiva de la base.

Validación del proceso

Como parte del control de calidad, se realizaron verificaciones puntuales utilizando empresas representativas, como el caso de Ecopetrol, confirmando que:

- La empresa estaba presente en la base externa original.
- Existía correspondencia con el registro en la base del CRM.
- Fue correctamente eliminada de la base de prospectos depurada.

Estas validaciones permitieron asegurar la efectividad y confiabilidad del procedimiento implementado.

Resultados de la depuración

El proceso de limpieza y exclusión generó una reducción progresiva en el número de registros de la base externa, como se muestra a continuación:

Como resultado, se obtuvo una base final de **572 empresas potenciales**, las cuales representan oportunidades reales para el análisis de prospección comercial.

Tabla 18. Resultados del proceso de depuración de la base de prospectos externos

Etapas del proceso	Número de registros
Base externa original	600
Clientes existentes en el CRM	167

Después de exclusión exacta	591
Después de exclusión parcial	572

Formulación del problema

El problema se planteó como una tarea de **clasificación binaria supervisada**. Para cada unidad de negocio de interés, se buscó predecir si un cliente actual tenía o no probabilidad de consumir dicha línea de servicio.

En términos analíticos, esto implicó construir una variable objetivo binaria, donde:

- el valor **1** representa que el cliente consume la unidad de negocio objetivo;
- el valor **0** representa que el cliente no consume dicha unidad de negocio.

Esta formulación permitió traducir una necesidad del negocio en un problema de predicción. En lugar de limitarse a observar qué clientes ya compran ciertas soluciones, el modelo se orientó a identificar patrones en los datos históricos que permitieran estimar qué otras empresas de la base actual podrían adoptar servicios adicionales.

Construcción de la variable objetivo

Para cada unidad de negocio (por ejemplo, SAP, Public Cloud, Digital, Workplace, Cyber Security y Private Cloud), se definió una **variable objetivo binaria**, donde:

- **Valor 1:** indica que el cliente ha consumido la solución.
- **Valor 0:** indica que el cliente no ha consumido la solución.

Esta variable se construyó a partir de los valores de ACV (Annual Contract Value) asociados a cada unidad de negocio. Si el valor de ACV era mayor que cero, se consideraba que el cliente había consumido dicha solución.

Se desarrolló un modelo independiente para cada unidad de negocio (por ejemplo, SAP, Public Cloud, Digital, Workplace, Cyber Security y Private Cloud), permitiendo capturar las particularidades de cada línea de servicio.

Preparación y transformación de los datos

Previo al entrenamiento de los modelos, se realizaron diversas etapas de preparación de datos con el fin de garantizar la calidad y consistencia de la información:

1. **Selección de variables:**
Se eligieron variables comunes entre la base de clientes actuales y la base de prospectos externos, tales como:
 - Ciudad
 - País

- Tipo de empresa
 - Tamaño de la empresa
 - Presencia internacional
 - Sector económico principal
 - Modelo de negocio
 - Submodelo
2. **Homologación de variables:**
Se aseguraron nombres y formatos consistentes entre ambas bases de datos.
 3. **Codificación de variables categóricas:**
Las variables categóricas fueron transformadas mediante la técnica de **one-hot encoding**, permitiendo su utilización en los modelos de aprendizaje automático.
 4. **Manejo de valores faltantes:**
Los valores nulos fueron tratados mediante imputación o sustitución por valores representativos, garantizando la integridad del dataset.
 5. **División del conjunto de datos:**
El dataset fue dividido en conjuntos de **entrenamiento y prueba**, permitiendo evaluar la capacidad de generalización de los modelos.

Caracterización de la muestra de entrenamiento y análisis del desbalance de clases

Dado que el modelo supervisado se entrena sobre la base de 167 clientes actuales registrados en el CRM Salesforce, resulta esencial caracterizar con precisión la composición de la muestra de entrenamiento antes de presentar los algoritmos utilizados. Esta caracterización es especialmente relevante porque el problema de clasificación es de naturaleza múltiple: se entrena un modelo independiente para cada Unidad de Negocio (BU), y cada uno enfrenta una distribución de clases diferente según el grado de adopción de esa línea del portafolio.

En total, el dataset de entrenamiento comprende 167 instancias correspondientes a clientes únicos con oportunidades ganadas entre 2018 y 2025. Para cada BU, la variable objetivo binaria se construyó a partir del valor ACV acumulado: una empresa se clasifica en la clase positiva ($Y = 1$) si su ACV en la BU analizada es mayor que cero; de lo contrario pertenece a la clase negativa ($Y = 0$).

La siguiente tabla resume la distribución de clases y los tamaños de muestra resultantes para cada BU después de aplicar la partición 70% entrenamiento / 30% prueba con estratificación:

Tabla 19. *Distribución de clases y tamaños de muestra por Unidad de Negocio*

Unidad de Negocio	Total	Clase positiva	Clase negativa	% positiva	Entrenamiento (70%)	Prueba (30%)
Cloud Pública	167	48	119	28.7%	117 (34 pos./83 neg.)	50 (14 pos./36 neg.)
Cyber Security	167	33	134	19.8%	117 (23 pos./94 neg.)	50 (10 pos./40 neg.)
Digital Business / Aplicaciones	167	47	120	28.1%	117 (33 pos./84 neg.)	50 (14 pos./36 neg.)
Digital Workplace	167	26	141	15.6%	117 (18 pos./99 neg.)	50 (8 pos./42 neg.)
Private Cloud / Infra Tradicional	167	104	63	62.3%	117 (73 pos./44 neg.)	50 (31 pos./19 neg.)
Soluciones SAP	167	18	149	10.8%	117 (13 pos./104 neg.)	50 (5 pos./45 neg.)

Implicaciones del desbalance de clases

La tabla anterior revela un desbalance de clases significativo en la mayoría de las BUs. Este fenómeno es especialmente pronunciado en Soluciones SAP (10.8% de clase positiva) y Digital Workplace (15.6%), donde el número de ejemplos positivos en el conjunto de prueba es extremadamente reducido (5 y 8 instancias, respectivamente). En estos escenarios, un clasificador trivial que prediga siempre la clase negativa alcanzaría precisiones aparentes del 89.2% y 84.4% respectivamente, lo que demuestra que la exactitud no es una métrica adecuada para evaluar estos modelos.

Frente a este desafío, se adoptaron tres medidas metodológicas complementarias:

- Partición estratificada del dataset: la proporción entre clases positiva y negativa se preservó tanto en el conjunto de entrenamiento como en el de prueba mediante el parámetro `stratify` del método `train_test_split` de `scikit-learn`. Esto garantiza que los modelos aprendan de una representación proporcional de ambas clases y que la

evaluación sea consistente con la distribución real de los datos.

- Ponderación de clases en los algoritmos: tanto Random Forest (`class_weight = "balanced_subsample"`) como Logistic Regression (`class_weight = "balanced"`) incorporan ajustes automáticos de pesos inversamente proporcionales a la frecuencia de cada clase, asignando mayor importancia a los errores cometidos sobre la clase minoritaria durante el entrenamiento. XGBoost gestiona este desbalance implícitamente mediante el parámetro `scale_pos_weight`, que ajusta el gradiente de la función de pérdida para compensar la asimetría.
- Selección de métrica de evaluación: se adoptó el Área Bajo la Curva ROC (AUC-ROC) como métrica principal, ya que es invariante al umbral de clasificación y robusta frente al desbalance de clases. Un modelo aleatorio obtendría $AUC = 0.5$ independientemente de la distribución de clases, lo que proporciona un punto de referencia significativo incluso en escenarios de alta asimetría.

Consideraciones sobre el tamaño muestral

Es importante reconocer que el tamaño total de la muestra ($N = 167$ clientes) representa una limitación inherente del problema, derivada del universo finito de clientes corporativos activos con oportunidades ganadas en el periodo de estudio. Esta restricción tiene implicaciones directas sobre la estabilidad de las estimaciones, especialmente en las BUs con menor penetración. En particular, el conjunto de prueba de SAP contiene únicamente 5 instancias positivas, lo que implica que las métricas de evaluación sobre ese conjunto deben interpretarse con cautela y en el contexto del propósito del modelo (ranquear prospectos), más que como una medida definitiva de precisión predictiva.

Para mitigar esta limitación, se optó por mantener una semilla fija (`random_state = 42`) en todas las particiones y algoritmos, garantizando la reproducibilidad de los resultados. Adicionalmente, la comparación de tres algoritmos con distintos supuestos y mecanismos de aprendizaje proporciona una validación cruzada implícita: la convergencia entre sus predicciones refuerza la confianza en los rankings generados, mientras que la divergencia señala regiones de incertidumbre que deben ser consideradas por los equipos comerciales.

Variables predictoras disponibles para el modelado

Un aspecto fundamental del diseño del modelo supervisado es la restricción impuesta por las variables disponibles en la base de prospectos externos. A diferencia del dataset de clientes actuales, que contiene más de 236 variables numéricas y 15 categóricas, la base de prospectos externos solo dispone de los atributos firmográficos recopilados mediante el proceso de enriquecimiento con Gemini. Por tanto, las variables predictoras del modelo supervisado se limitaron exclusivamente a aquellas presentes en ambas fuentes:

- Ciudad: ubicación geográfica de la sede principal de la empresa.
- País: país de operación (Colombia, Panamá, otros).

- Tipo de empresa: clasificación entre privada, pública, mixta, multinacional o grupo empresarial.
- Subtipo de empresa: descripción más granular del tipo organizacional.
- Tamaño de la empresa: grande, mediana-grande o multinacional.
- Presencia internacional: indicador de operación en más de un país.
- Sector económico principal: industria o sector de actividad primaria.
- Modelo de negocio: clasificación en servicios, manufactura, distribución, entre otros
- Modelo normalizado: taxonomía estandarizada del modelo de negocio (B2B, B2C, B2B2C, Holding, Retail, Financiero, Manufactura, Servicios, Plataforma).
- Submodelo: descripción más específica del modelo de negocio (banca comercial, telecomunicaciones, oil & gas, etc.).

Esta restricción implica que el modelo supervisado no puede explotar las variables de comportamiento transaccional (ACV, oportunidades ganadas, años de relación comercial) disponibles en el CRM, ya que estas no existen para los prospectos externos. La elección de variables firmográficas como predictoras es metodológicamente coherente con el enfoque de matching por similitud organizacional, que busca identificar empresas externas que compartan el mismo perfil estructural que los clientes actuales de cada BU.

Tras aplicar la codificación One-Hot Encoding sobre las variables categóricas, la dimensionalidad de la matriz de características aumentó considerablemente. Cada categoría de cada variable categórica genera una columna binaria independiente, lo que puede resultar en decenas o cientos de columnas dependiendo de la cardinalidad de las variables. Para garantizar que la base de prospectos externos sea compatible con el modelo entrenado, las columnas de la matriz de características se alinearon explícitamente mediante la función reindex, rellenando con ceros las columnas ausentes en la base externa.

5.2.2.1 Técnicas de aprendizaje supervisado utilizadas

Con el objetivo de comparar diferentes enfoques predictivos y seleccionar el modelo más adecuado para cada unidad de negocio, se implementaron las tres técnicas de aprendizaje supervisado utilizadas en el modelo de cross selling: **Regresión Logística, Random Forest y XGBoost**.

Las definiciones algorítmicas y matemáticas de Regresión Logística, Random Forest y XGBoost se encuentran en la Sección 3.2 (Marco Teórico). La diferencia respecto al Cross-Selling radica en las configuraciones técnicas adoptadas para el contexto de prospectos externos, que se detallan a continuación.

- **Regresión Logística**

Configuración técnica y decisiones de implementación

max_iter = 2000: se incrementó respecto al valor por defecto (100) para garantizar la convergencia del optimizador L-BFGS en el espacio de alta dimensionalidad generado por el One-Hot Encoding. Con matrices OHE de decenas de columnas, 100 iteraciones resultan insuficientes para alcanzar el criterio de parada.

StandardScaler(with_mean=False): el escalamiento de la matriz OHE se realiza sin centrar (with_mean=False) porque las matrices dispersas no deben centrarse — restar la media a una matriz de ceros y unos produce valores densos que eliminan la ventaja de la representación dispersa y pueden generar errores numéricos.

get_dummies(drop_first=True): se elimina la primera categoría de cada variable categórica como categoría de referencia, evitando la trampa de las variables dummy (multicolinealidad perfecta en la matriz de diseño). Esto es especialmente importante para la Regresión Logística, cuya estimación máximo-verosímil requiere una matriz de rango completo.

Aplicación a prospectos externos: el scaler entrenado en los datos del CRM se aplica con transform (no fit_transform) sobre la base externa, garantizando que la normalización sea consistente con el espacio del entrenamiento. Las columnas de la base externa se alinean con reindex(columns=X.columns, fill_value=0) antes de la inferencia.

Selección de variables predictoras

Se utilizaron las mismas variables compartidas entre la base de clientes actuales y la base de prospectos externos empleadas en la Regresión Logística: ciudad, país, tipo y subtipo de empresa, tamaño organizacional, presencia internacional, sector económico principal, modelo de negocio, modelo normalizado y submodelo. Esta consistencia garantiza la comparabilidad directa entre los resultados de ambos modelos.

Configuración principal del modelo

- **n_estimators = 400:** número de árboles en el bosque. Un mayor número de árboles reduce la varianza sin incrementar el sesgo, a costa de mayor tiempo de cómputo.
- **class_weight = "balanced_subsample":** ajusta los pesos de clase en cada submuestra de bootstrap, compensando el posible desbalance entre clientes que consumen y no consumen la BU objetivo. Este parámetro es crítico cuando las clases están desbalanceadas, como ocurre en varias de las unidades de negocio analizadas.
- **random_state = 42:** garantiza la reproducibilidad de los resultados.

Entrenamiento y evaluación

El modelo fue entrenado con el conjunto de entrenamiento (70% de los datos de clientes actuales) y evaluado sobre el conjunto de prueba (30% restante) mediante la métrica AUC-ROC. Dado que los árboles de decisión no requieren escalamiento previo de las variables, la transformación mediante StandardScaler se aplicó exclusivamente para la Regresión Logística. Una vez seleccionado el mejor modelo por BU, los scores de propensión para los prospectos externos se generaron aplicando el método predict_proba sobre la base depurada de 572 empresas externas.

- **XGBoost**

XGBoost (Extreme Gradient Boosting) fue incorporado como el tercer clasificador del conjunto, reconocido por su alto desempeño predictivo en problemas de clasificación estructurada. A diferencia del Random Forest, que construye árboles de forma independiente y paralela (Bagging), XGBoost emplea una estrategia secuencial de Boosting, donde cada árbol corrige los errores residuales del modelo anterior mediante descenso de gradiente en el espacio funcional. Al igual que en los modelos anteriores, se utilizaron exclusivamente las variables firmográficas comunes entre ambas bases de datos, garantizando la aplicabilidad del modelo sobre la base de prospectos externos sin requerir variables transaccionales que no están disponibles para nuevas empresas.

Configuración principal del modelo

- `n_estimators = 400`: número de iteraciones de boosting. Cada iteración agrega un árbol al modelo aditivo.
- `learning_rate = 0.05`: tasa de aprendizaje que controla la contribución de cada árbol al modelo final. Un valor bajo favorece la generalización y reduce el sobreajuste.
- `max_depth = 4`: profundidad máxima de cada árbol, limitando la complejidad y el riesgo de sobreajuste.
- `subsample = 0.9`: fracción de las observaciones utilizada en cada iteración de boosting, introduciendo aleatoriedad que reduce la varianza.
- `colsample_bytree = 0.9`: fracción de variables seleccionadas para cada árbol, decorrelacionando los modelos parciales.
- `eval_metric = 'logloss'`: métrica de optimización interna durante el entrenamiento.
- `random_state = 42`: garantiza la reproducibilidad.

Entrenamiento y evaluación

El modelo fue entrenado con el conjunto de datos de clientes actuales (70/30 train/test), evaluado mediante AUC-ROC, y aplicado sobre la base de prospectos externos para generar scores de propensión por BU. La comparación de los tres modelos (Regresión Logística, Random Forest y XGBoost) permitió seleccionar el algoritmo con mayor AUC para cada unidad de negocio, garantizando que el ranking de prospectos se base en el mejor clasificador disponible.

- **Análisis Discriminante Multivariado (ADM)**

El Análisis Discriminante Multivariado (ADM) es una técnica estadística clásica que busca encontrar combinaciones lineales de las variables predictoras que maximicen la separación entre grupos predefinidos, minimizando simultáneamente la variabilidad dentro de cada grupo [4]. A diferencia de los modelos de aprendizaje automático, el ADM parte de supuestos paramétricos específicos sobre la distribución de los datos: normalidad multivariada de las variables dentro de cada clase, y homocedasticidad (igualdad de matrices de covarianza entre grupos). Cuando estos

supuestos se cumplen, el ADM es el clasificador óptimo en el sentido del Teorema de Bayes.

En el contexto de este proyecto, el ADM no se utiliza como modelo de producción (no genera el ranking de prospectos), sino como instrumento de diagnóstico para responder una pregunta metodológica fundamental: ¿las variables firmográficas disponibles (sector, tipo de empresa, tamaño, modelo de negocio, entre otros) son capaces de separar linealmente a los consumidores de los no consumidores de cada BU? La respuesta a esta pregunta determina si las relaciones en los datos son linealmente explotables o si se requieren modelos capaces de capturar no linealidades, como Random Forest y XGBoost.

Fundamento matemático del ADM

El ADM encuentra la función discriminante lineal $w^l = w^T x$ que maximiza el cociente de Fisher:

$$J(w) = (w^T S_B w) / (w^T S_W w)$$

donde S_B es la matriz de dispersión entre grupos (between-class scatter matrix) y S_W es la matriz de dispersión dentro de los grupos (within-class scatter matrix). Este cociente mide qué tan separados están los centroides de clase en relación con la variabilidad interna de cada clase. La solución óptima corresponde al vector propio asociado al mayor valor propio de $S_W^{-1} S_B$.

Para un problema de clasificación binaria (consumidor / no consumidor), el ADM produce una única función discriminante. Un nuevo punto x se asigna a la clase positiva si:

$w^T x \geq c$, donde c es el umbral óptimo de clasificación

La probabilidad de pertenencia a cada clase se estima mediante la función de densidad gaussiana, y el AUC-ROC se calcula a partir de las probabilidades así estimadas, permitiendo la comparación directa con los modelos supervisados.

Análisis Discriminante Lineal (LDA) — implementación específica

Dentro de la familia del ADM, el Análisis Discriminante Lineal (LDA) es la variante más ampliamente utilizada. Asume matrices de covarianza iguales entre clases, lo que permite calcular una frontera de decisión lineal compartida. En scikit-learn se implementa mediante `LinearDiscriminantAnalysis()`, que estima los parámetros de forma numéricamente estable usando descomposición SVD.

Diseño de la evaluación LDA en este proyecto

Se entrenó un modelo LDA independiente por cada Unidad de Negocio, siguiendo el mismo protocolo de las etapas anteriores: partición 70%/30% estratificada (`random_state=42`), escalamiento `StandardScaler(with_mean=False)` sobre la matriz OHE, y evaluación sobre el conjunto de prueba. Las variables utilizadas son las mismas 10 variables firmográficas comunes entre el CRM y la base externa, codificadas con One-Hot Encoding (`drop_first=True`): Ciudad, País, Tipo de empresa, Tamaño de la empresa, Presencia internacional, Sector económico principal, Modelo de negocio y Submodelo.

Para cada BU se calcularon Accuracy, Precision, Recall, F1-Score y AUC-ROC sobre el conjunto de prueba. Las definiciones de estas métricas y sus formulaciones matemáticas se encuentran en la

Sección 3.2 Marco Teórico.

Adicionalmente, para cada BU se generaron la matriz de confusión (visualizada como heatmap) y el reporte de clasificación completo (classification_report), que muestra las métricas por clase (consumidor / no consumidor) e identifica asimetrías en el desempeño del modelo entre grupos.

Selección automática del mejor modelo por BU

Tras el entrenamiento de los tres clasificadores, sus AUC-ROC sobre el conjunto de prueba se organizaron en una tabla comparativa por BU. El algoritmo con mayor AUC fue seleccionado automáticamente como el modelo de referencia para esa BU y fue el único que se aplicó sobre los 572 prospectos externos para generar el score de propensión supervisado. Esta selección es independiente por BU: el mejor modelo para SAP puede ser diferente al mejor para Cyber Security o Digital Business.

Tabla 20. Comparación de AUC-ROC por modelo y Unidad de Negocio (Prospectos Externos)

Unidad de Negocio	LDA	Regresión Logística	Random Forest	XGBoost	Mejor modelo
Cloud Pública	0.763	0.786	0.731	0.593	LR (0.786)
Cyber Security	0.439	0.504	0.274	0.593	XGBoost (0.593)
Digital Business / Aplicaciones	0.755	0.551	0.413	0.709	XGBoost (0.709)
Digital Workplace	0.457	0.552	0.750	0.376	RF (0.750)
Private Cloud / Infra Tradicional	0.409	0.548	0.629	0.554	RF (0.629)
Soluciones SAP	0.199	0.604	0.747	0.928	XGBoost (0.928)

Flujo de inferencia sobre los 572 prospectos externos

Una vez seleccionado el mejor modelo por BU, el proceso de scoring sobre la base externa sigue estos pasos:

- Preparar X_ext: seleccionar las mismas variables_finales de la base de prospectos (df_database_nueva[variables_finales]).
- Codificar con OHE: pd.get_dummies(X_ext, drop_first=True).fillna(0).
- Alinear columnas: X_ext.reindex(columns=X.columns, fill_value=0) para garantizar que la base externa tenga exactamente las mismas columnas que la matriz de entrenamiento,

completando con cero las categorías ausentes en el dataset externo.

- Escalar si aplica: si el mejor modelo es Regresión Logística, aplicar `scaler.transform(X_ext)` con el scaler ya entrenado sobre `X_train`. Para RF y XGBoost no se requiere escalamiento.
- Generar scores: aplicar `predict_proba(X_ext)[:, 1]` del modelo seleccionado, obteniendo una probabilidad de adopción en `[0, 1]` para cada uno de los 572 prospectos.
- Construir ranking: anexas los scores al dataframe de prospectos y ordenar de mayor a menor. Los top 20 por BU se exportaron en archivos Excel individuales (`ranking_sap_supervisado.xlsx`, `ranking_cyber_supervisado.xlsx`, etc.).

- **Modelo de Similitud por Coseno**

A diferencia del modelo de cross-selling (donde la similitud coseno se calculó sobre variables transaccionales del CRM), el modelo de similitud para prospectos externos utiliza exclusivamente las 10 variables firmográficas comunes, ya que los prospectos no tienen historial en Salesforce. El proceso implementado en el código presenta dos particularidades técnicas relevantes que lo distinguen de una implementación estándar.

Codificación conjunta (joint OHE):

En lugar de codificar los clientes de referencia y los prospectos externos por separado (lo que generaría espacios de features incompatibles), se concatenan ambos datasets en un único dataframe antes de aplicar `get_dummies`. Luego se separan por posición:

```
df_union = pd.concat([df_ref[variables], df_ext[variables]], ignore_index=True)
```

```
df_union_encoded = pd.get_dummies(df_union[variables], dummy_na=False)
```

```
X_ref = df_union_encoded.iloc[:len(df_ref), :]
```

```
X_ext = df_union_encoded.iloc[len(df_ref):, :]
```

Este enfoque garantiza que referencia y externos compartan exactamente el mismo espacio de features OHE, sin necesidad de reindex posterior. Es la solución más robusta al problema de covariate shift en variables categóricas.

- **Similitud contra TODOS los clusters (no solo el estratégico):**

Para cada BU, se calculan centroides para TODOS los clusters identificados en la etapa no supervisada. Luego, para cada prospecto externo, se calcula la similitud coseno contra cada centroide, y el score final es el máximo sobre todos los clusters:

$$\text{Score_Similitud_}[BU](z) = \max\{\cos(z, \text{centroide_}k) \mid k = 1, \dots, K\}$$

Además, se registra el campo `Cluster_[BU]_Mas_Parecido`, que indica con cuál segmento de clientes actuales tiene mayor afinidad cada prospecto. Esto permite no solo priorizar por score, sino también orientar el discurso comercial: un prospecto que se asemeja al cluster de empresas

públicas de gran tamaño recibirá una propuesta diferente a uno que se asemeja al cluster de medianas empresas tecnológicas.

Este enfoque es más completo que usar solo el cluster estratégico, porque permite capturar prospectos cuyo perfil es óptimo para un segmento diferente al de mayor ACV, pero que igualmente representa una oportunidad real de negocio. Los resultados se exportan en archivos Excel por BU (ranking_sap_externos_clusters.xlsx, etc.).

- **Score Híbrido Final**

El score híbrido integra la perspectiva geométrica (similitud coseno) con la perspectiva estadística (mejor clasificador supervisado) en un indicador único de afinidad por prospecto y BU. Para cada BU se calcula de forma independiente:

$$\text{Score_Final_}[BU](z) = 0.5 \times \text{Score_Similitud_}[BU](z) + 0.5 \times \text{Score_Supervisado_}[BU](z)$$

donde $\text{Score_Similitud_}[BU]$ proviene de la similitud coseno (rango $[-1, +1]$) y $\text{Score_Supervisado_}[BU]$ es el predict_proba del mejor clasificador (rango $[0, 1]$). Los pesos iguales (50%/50%) reflejan que ambas fuentes de evidencia son complementarias y equidistantes en ausencia de validación empírica que justifique una ponderación diferenciada. Este parámetro puede ajustarse en versiones futuras del modelo a medida que se acumule retroalimentación comercial sobre la efectividad relativa de cada enfoque.

El ranking híbrido por BU se obtiene ordenando los 572 prospectos de mayor a menor $\text{Score_Final_}[BU]$. Para cada BU se exporta un archivo Excel individual (ranking_sap_hibrido.xlsx, ranking_cyber_hibrido.xlsx, etc.) que incluye las tres columnas de score (Similitud, Supervisado, Final) para permitir al equipo comercial comparar ambas perspectivas por separado.

- **Consolidación en ranking global de prospectos**

Los rankings individuales por BU se consolidan en un único ranking global mediante la suma de los scores de similitud coseno de todas las BUs:

$$\text{Score_Total_Similitud}(z) = \sum_{ku} \text{Score_Similitud_}[BU](z)$$

Adicionalmente, se identifica la BU de mayor afinidad para cada prospecto:

$$\text{BU_Mayor_Afinidad}(z) = \text{argmax}_{ku} \text{Score_Similitud_}[BU](z)$$

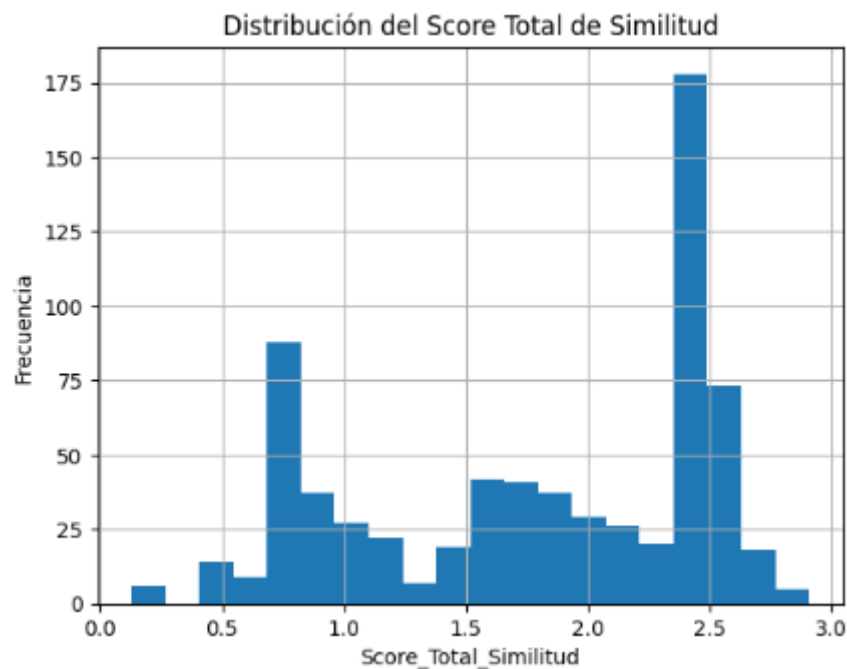
Este campo orienta al equipo comercial sobre cuál línea del portafolio tiene mayor probabilidad de resonar con cada prospecto, facilitando la asignación inicial a los especialistas de BU correspondientes. El ranking global (ranking_global_externos.xlsx) presenta los 572 prospectos ordenados por $\text{Score_Total_Similitud}$, con columnas de score individual por BU y el campo BU_Mayor_Afinidad .

- **Distribución de scores y umbrales de priorización**

El análisis incluyó histogramas de la distribución de Score_Final por BU y de $\text{Score_Total_Similitud}$ global, permitiendo visualizar la separación entre prospectos de alta y baja afinidad. Estas

distribuciones son relevantes para definir umbrales operativos: los prospectos en el cuartil superior del Score_Total_Similitud representan la lista de trabajo prioritaria para el equipo comercial, mientras que los del tercio inferior pueden ser descartados en una primera ronda de prospección. Los resultados finales se exportaron en archivos Excel individuales por BU y en un consolidado global, disponibles como entregables del proyecto.

Figura 11. *Distribución del Score Total de Similitud — Prospectos externos*



El histograma de la Figura 11 evidencia una distribución bimodal en el Score_Total_Similitud, con un primer pico alrededor de 0.7–0.8 y un segundo pico mucho más pronunciado en el rango 2.4–2.5. Esta bimodalidad valida operativamente el uso del cuartil superior como umbral de priorización: el grupo de alta afinidad agregada está claramente separado del grueso de la base, lo que se traduce en un corte natural para construir la lista de trabajo prioritaria. La cola izquierda (scores totales menores a 1.0) corresponde a prospectos con baja afinidad transversal y poco potencial multiproducto, que pueden descartarse en una primera ronda sin pérdida estratégica significativa.

Resultados del modelo híbrido de prospectos externos

A continuación se presentan los resultados obtenidos al aplicar el modelo híbrido sobre la base depurada de 572 prospectos externos. Para cada Unidad de Negocio se reportan los top 10 prospectos según el Score_Final (promedio 50%/50% entre similitud coseno y score supervisado), junto con el sector económico y el clúster de mayor afinidad. Los rangos de scores reflejan la capacidad discriminativa real del modelo en cada BU.

Rangos de scores por componente y BU

La siguiente tabla consolida los rangos reales de los scores obtenidos sobre los 572 prospectos externos. El rango del Score_Supervisado refleja directamente la capacidad del mejor clasificador para distinguir prospectos de alta y baja probabilidad de adopción; el rango del Score_Similitud depende de la compacidad de los clústeres de referencia; y el Score_Final integra ambas perspectivas.

Tabla 21. Rangos de scores por componente y Unidad de Negocio (572 prospectos externos)

Unidad de Negocio	Score Similitud (mín–máx)	Score Supervisado (mín–máx)	Score Final (mín–máx)	Discriminación
Cloud Pública	0.183–0.509	0.000–0.178	0.001–0.284	Baja
Cyber Security	0.173–0.475	0.008–0.240	0.019–0.314	Moderada
Digital Business	0.165–0.495	0.005–0.124	0.013–0.266	Moderada
Digital Workplace	0.189–0.500	0.003–0.282	0.046–0.344	Buena
Private Cloud	0.233–0.497	0.062–0.570	0.040–0.531	Mejor del portafolio
Soluciones SAP	0.200–0.500	0.008–0.023	0.055–0.259	Muy baja

Nota: el score supervisado de SAP presenta un rango extremadamente estrecho [0.008–0.023], lo que refleja la limitación del modelo para discriminar prospectos SAP con las variables firmográficas disponibles. Para esta BU, el score de similitud coseno aporta la mayor parte de la señal en el score híbrido.

Top 10 prospectos por BU – Score_Final híbrido

Private Cloud / Infra Tradicional (mejor discriminación del portafolio, Score_Final máx 0.531):

Tabla 22. Top 10 prospectos — Private Cloud / Infra Tradicional (Score_Final híbrido)

Empresa	Score_Final	Sector económico	Clúster más parecido
Medtronic Colombia	0.531	Salud	Clúster 1
Atención médica Fresenius	0.531	Salud	Clúster 1
General Electric Healthcare	0.531	Salud	Clúster 1
Clínica Imbanaco	0.517	Salud	Clúster 1
Pfizer Colombia	0.514	Farmacéutico	Clúster 1
Sanofi-Aventis	0.514	Farmacéutico	Clúster 1
Johnson & Johnson	0.514	Farmacéutico	Clúster 1

Siemens Healthineers	0.480	Salud	Clúster 1
Philips Healthcare	0.480	Salud	Clúster 1
Clínica rural	0.475	Salud	Clúster 1

Digital Workplace (Score_Final máx 0.344):

Tabla 23. Top 10 prospectos — Digital Workplace (Score_Final híbrido)

Empresa	Score_Final	Sector económico	Clúster más parecido
Imagen	0.344	Financiero	Clúster 1
Interconexión Eléctrica ISA	0.341	Energía	Clúster 1
Fideicomiso Bancolombia	0.306	Financiero	Clúster 1
Banco de Inversión Bancolombia	0.306	Financiero	Clúster 1
Protección	0.306	Financiero	Clúster 1
Acciona Energía	0.282	Energía	Clúster 0
EDP Renovables	0.282	Energía	Clúster 0
Tigo (UNE EPM)	0.281	Telecomunicaciones	Clúster 0
Comfama	0.270	Servicios	Clúster 0
Metro de Medellín	0.269	Transporte	Clúster 0

Cyber Security (Score_Final máx 0.314):

Tabla 24. Top 10 prospectos — Cyber Security (Score_Final híbrido)

Empresa	Score_Final	Sector económico	Clúster más parecido
Grupo Energía Bogotá GEB	0.314	Energía	Clúster 0
Racafé	0.314	Alimento	Clúster 0
Organización Corona	0.300	Industrial	Clúster 0
Fedecafé	0.300	Alimento	Clúster 0
Fundación Santa Fe	0.256	Salud	Clúster 0
Clínica rural	0.256	Salud	Clúster 0

Caracol Televisión	0.256	Medios	Clúster 0
RCN Televisión	0.256	Medios	Clúster 0
Cafam	0.253	Servicios	Clúster 0
Colsubsidio	0.253	Servicios	Clúster 0

Digital Business / Aplicaciones (Score_Final máx 0.266):

Tabla 25. Top 10 prospectos — Digital Business / Aplicaciones (Score_Final híbrido)

Empresa	Score_Final	Sector económico	Clúster más parecido
Grupo Energía Bogotá GEB	0.266	Energía	Clúster 0
ISA	0.263	Energía	Clúster 0
Procaps	0.262	Farmacéutico	Clúster 0
Racafé	0.259	Alimento	Clúster 0
Cementos Argos	0.257	Industrial	Clúster 0
Grupo Argos	0.257	Industrial	Clúster 0
Noel	0.257	Alimento	Clúster 0
Nutresa	0.257	Alimento	Clúster 0
Sociedad Portuaria CTG	0.254	Logística	Clúster 0
Tecnoglass	0.247	Industrial	Clúster 0

Cloud Pública (Score_Final máx 0.284):

Tabla 26. Top 10 prospectos — Cloud Pública (Score_Final híbrido)

Empresa	Score_Final	Sector económico	Clúster más parecido
Compañía Constructora Marval	0.284	Construcción	Clúster 3
Incubadora San Fernando	0.276	Alimento	Clúster 3
Avidesa Mac Pollo	0.276	Alimento	Clúster 3
ESSA Santander	0.262	Energía	Clúster 3
Clínica Valle del Lili	0.257	Salud	Clúster 3
Fundación Santa Fe	0.255	Salud	Clúster 3

Clínica rural	0.255	Salud	Clúster 3
Clínica Imbanaco	0.255	Salud	Clúster 3
Medtronic	0.254	Salud	Clúster 3
GE Healthcare	0.254	Salud	Clúster 3

Soluciones SAP (Score_Final máx 0.259 — ver limitación a continuación):

Tabla 27. *Top 10 prospectos — Soluciones SAP (Score_Final híbrido)*

Empresa	Score_Final	Sector económico	Clúster más parecido
Clínica Valle del Lili	0.259	Salud	Clúster 3
Fundación Santa Fe	0.256	Salud	Clúster 3
Clínica rural	0.256	Salud	Clúster 3
Clínica Imbanaco	0.255	Salud	Clúster 3
Atención médica Fresenius	0.254	Salud	Clúster 3
GE Healthcare	0.254	Salud	Clúster 3
Medtronic	0.254	Salud	Clúster 3
Yara Colombia	0.238	Químico	Clúster 0
Movistar (Telefónica Col)	0.237	Telecomunicaciones	Clúster 0
Bayer SA	0.230	Químico	Clúster 0

Limitación metodológica del ranking SAP: subrepresentación sectorial en la base externa

El ranking de prospectos para Soluciones SAP presenta una particularidad que requiere interpretación cuidadosa: los primeros siete lugares están ocupados por empresas del sector salud, cuando los clusters de mayor valor económico en la base de clientes reales de SAP corresponden a los sectores Alimentos (k=11, ACV ~\$47.800M) y Petróleo y Gas (k=3, ACV ~\$52.200M). Este resultado no es un error del modelo, sino la consecuencia de dos factores metodológicos que se combinan y deben entenderse de forma conjunta.

Factor 1 — Ausencia del sector Petróleo y Gas en la base externa: el sector con mayor ACV acumulado en la base de clientes SAP no tiene representación en la base de 572 prospectos externos. La razón es directa: las grandes empresas de petróleo y gas de Colombia (Ecopetrol, Cenit, entre otras) ya son clientes activos de la empresa en Colombia y fueron excluidas del proceso de depuración para evitar que el modelo intentara prospectar a organizaciones que ya tienen relación comercial con la empresa. En la práctica, el mercado natural de SAP en el segmento de hidrocarburos colombiano está en gran medida capturado.

Factor 2 — Efecto del clúster singular de salud: el cluster 3 de SAP está compuesto por un único cliente del sector hospitalario. Aunque es un segmento minoritario en términos de número de clientes, el algoritmo de similitud coseno calcula el máximo sobre todos los clústeres. Como resultado, ese único centroide de salud genera una similitud de 0.500 para todos los prospectos externos del sector salud (n=31), empujándolos artificialmente al tope del ranking SAP. El score de similitud no discrimina entre prospectos de salud porque todos comparten el mismo perfil firmográfico respecto a ese centroide.

El efecto combinado es que el ranking SAP identifica oportunidades en sectores que pueden tener menor probabilidad real de adopción del ecosistema SAP. Esto no invalida el modelo ni los rankings — los prospectos listados son técnicamente los más similares a los clústeres disponibles —, pero sí obliga a interpretarlos con criterio comercial y contexto sectorial. En términos prácticos, el equipo comercial debería priorizar dentro del ranking SAP a las empresas de los sectores Alimento, Químico y Telecomunicaciones (posiciones 8 en adelante), que responden a los perfiles del clúster dominante (k=11, Alimentos).

Esta limitación es consecuencia directa de la estructura del mercado SAP en Colombia y no de un fallo técnico del modelo. Constituye un hallazgo metodológico relevante que sería abordable en versiones futuras mediante dos vías: ampliar la base externa con empresas de los sectores sub-representados, o incorporar una restricción de sector durante la construcción del score híbrido SAP para que el algoritmo solo compare prospectos contra clústeres sectorialmente coherentes.

Ranking global de prospectos: top 20 por Score_Total_Final

El ranking global se construye sumando los scores híbridos (Score_Final_BU) de todas las BUs por prospecto, produciendo un Score_Total_Final que refleja la afinidad total del prospecto con el portafolio completo de la empresa. El archivo ranking_global_final_hibrido.xlsx contiene 8,762 registros (combinaciones prospecto-BU) con BU_Mayor_Afinidad_Final indicando la línea de mayor potencial para cada empresa.

Los 20 prospectos con mayor Score_Total_Final son:

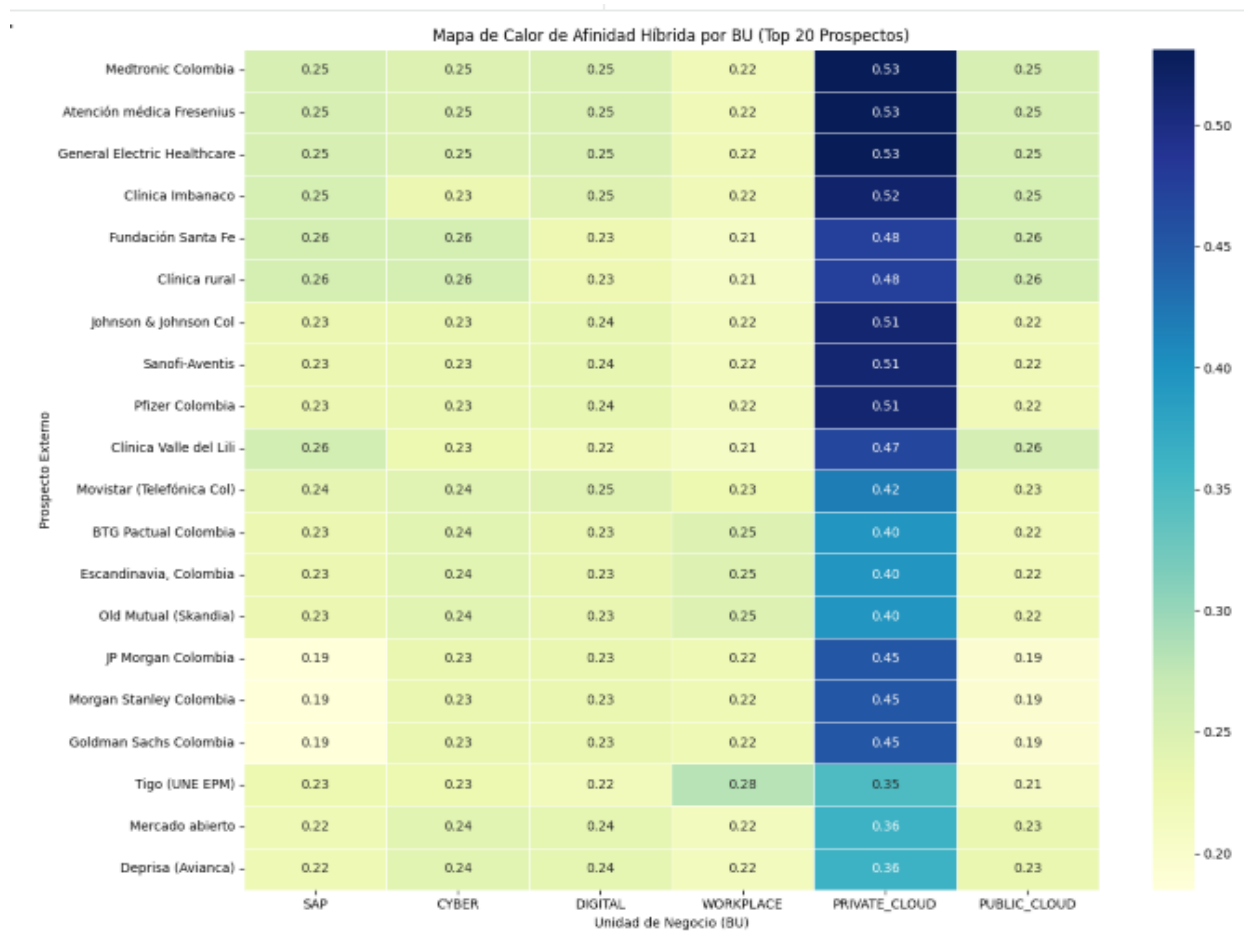
Tabla 28. *Ranking global de prospectos externos — Top 20 por Score_Total_Final*

#	Empresa	Score_Total_Final	BU Mayor Afinidad	Sector
1	Medtronic Colombia	1.757	PRIVATE_CLOUD	Salud
2	Atención médica Fresenius	1.757	PRIVATE_CLOUD	Salud
3	General Electric Healthcare	1.757	PRIVATE_CLOUD	Salud
4	Clínica Imbanaco	1.724	PRIVATE_CLOUD	Salud
5	Fundación Santa Fe	1.678	PRIVATE_CLOUD	Salud
6	Clínica rural	1.678	PRIVATE_CLOUD	Salud
7	Johnson & Johnson	1.650	PRIVATE_CLOUD	Farmacéutico
8	Sanofi-Aventis	1.650	PRIVATE_CLOUD	Farmacéutico

9	Pfizer Colombia	1.650	PRIVATE_CLOUD	Farmacéutico
10	Clínica Valle del Lili	1.643	PRIVATE_CLOUD	Salud
11	Movistar (Telefónica Col)	1.600	PRIVATE_CLOUD	Telecomunicaciones
12	BTG Pactual Colombia	1.568	PRIVATE_CLOUD	Financiero
13	Escandinavia, Colombia	1.568	PRIVATE_CLOUD	Financiero
14	Old Mutual (Skandia)	1.568	PRIVATE_CLOUD	Financiero
15	JP Morgan Colombia	1.520	PRIVATE_CLOUD	Financiero
16	Morgan Stanley Colombia	1.520	PRIVATE_CLOUD	Financiero
17	Goldman Sachs Colombia	1.520	PRIVATE_CLOUD	Financiero
18	Tigo (UNE EPM)	1.515	PRIVATE_CLOUD	Telecomunicaciones
19	Mercado abierto	1.513	PRIVATE_CLOUD	Comercio
20	Deprisa (Avianca)	1.513	PRIVATE_CLOUD	Logística

La concentración del ranking global en el sector salud es consistente con el comportamiento de Private Cloud, que tiene la mayor discriminación supervisada del portafolio (score máx 0.570) y el mayor número de clientes de referencia (N=104). Las empresas farmacéuticas y del sector financiero aparecen en el rango 7–17, reflejando su afinidad con múltiples BUs simultáneamente. Movistar y Tigo son los únicos prospectos de telecomunicaciones en el top 20, sector que presenta perfil sólido para Private Cloud y Digital Business.

Figura 12. Mapa de Calor de Afinidad Híbrida por BU — Top 20 Prospectos Externos



La Figura 12 visualiza el ranking final como un mapa de doble entrada (prospectos × BUs) sobre los Score_Final híbridos. La lectura confirma tres patrones consistentes con la Tabla 27: (1) la columna de Private Cloud presenta la mayor intensidad cromática transversal, con valores entre 0.40 y 0.53 que duplican a los de las otras BUs en el mismo prospecto; (2) los prospectos de salud (Medtronic, Fresenius, GE Healthcare, Imbanaco, Santa Fe) concentran las celdas más oscuras en Private Cloud, evidenciando su perfil de afinidad multidimensional con la línea de Infraestructura Tradicional; y (3) los prospectos del sector financiero (JP Morgan, Morgan Stanley, Goldman Sachs) presentan un patrón homogéneo pero menor intensidad, sugiriendo que su afinidad estructural es real pero de menor magnitud que la del sector salud. Esta visualización permite a los equipos comerciales identificar de un vistazo cuáles cuentas concentran mayor afinidad agregada y cuáles son monoproducto.

Distribución de BU_Mayor_Afinidad_Final

La asignación de la BU de mayor afinidad revela la concentración del portafolio externo:

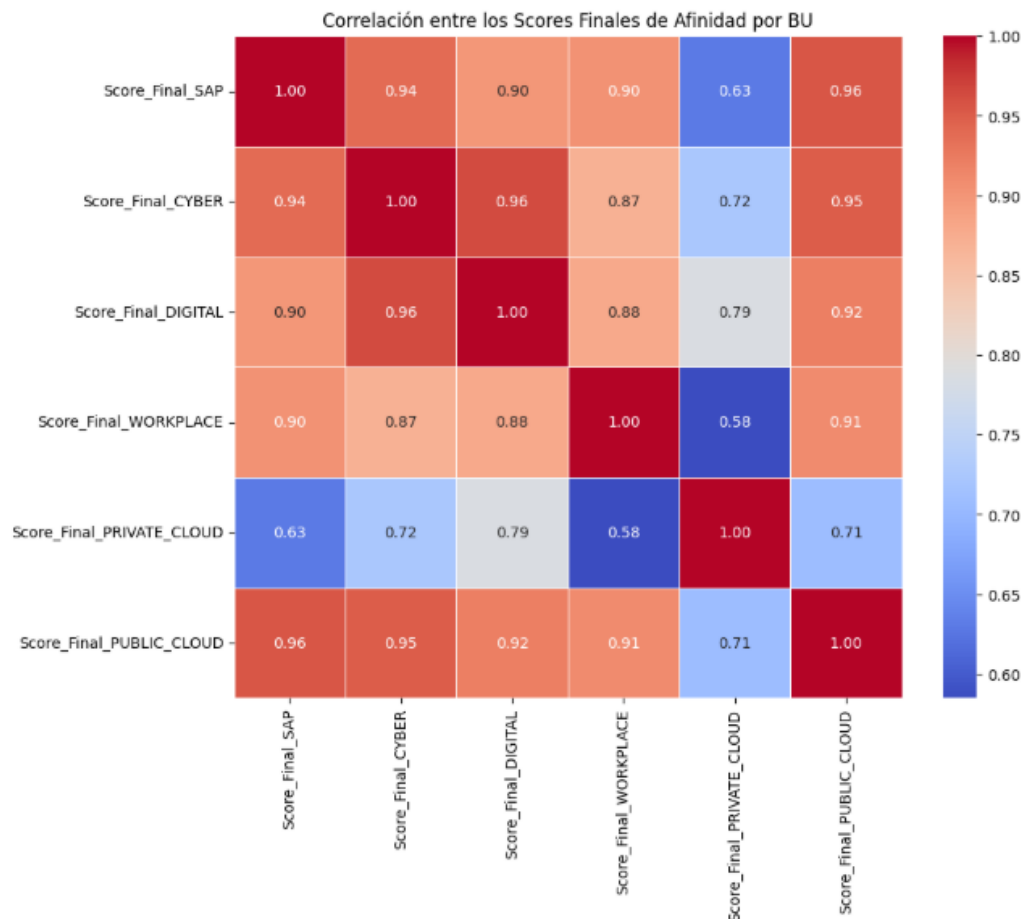
Desde la perspectiva del score de similitud coseno puro (Fase 1), la BU dominante es Soluciones SAP con 356 prospectos (51.0%), seguida de Private Cloud (88, 12.6%), Digital Business (88,

12.6%), Cloud Pública (84, 12.0%), Digital Workplace (67, 9.6%) y Cyber Security (15, 2.1%). Esta distribución refleja que el perfil firmográfico de los 572 prospectos es más similar, en términos de segmento y tamaño organizacional, al perfil de los clientes actuales de SAP — principalmente empresas grandes de los sectores industrial, químico y de salud.

Al incorporar el componente supervisado en el score híbrido final ($\alpha = 0.5$), la distribución cambia significativamente. Private Cloud / Infra Tradicional pasa a dominar con 8.626 prospectos (98.5%), ya que su modelo supervisado (RF, AUC = 0.750) genera scores altos para una amplia variedad de perfiles. La distribución BU_Mayor_Afinidad_Final queda así:

- **Private Cloud / Infra Tradicional:** 8,626 de 8,762 registros (98.4%). Dominio explicado por el mayor volumen de clientes de referencia (N=104) y la mayor amplitud de scores supervisados (rango 0.508 puntos). Private Cloud es el servicio más transversal y mejor representado en el CRM, lo que le da ventaja sistemática en la competencia por BU_Mayor_Afinidad.
- **Soluciones SAP:** 104 registros (1.2%). Prospectos cuyo score_final SAP supera al de Private Cloud, principalmente empresas del sector alimentos y químico que se asemejan más al clúster 0 de SAP que al perfil dominante de Private Cloud.
- **Digital Workplace:** 26 registros (0.3%). Empresas del sector financiero y energético con perfil muy alineado al clúster 1 de Workplace.
- **Cloud Pública y Cyber Security:** 5 y 1 registros respectivamente. La concentración extrema en Private Cloud no implica que estas BUs no tengan prospectos válidos, sino que sus scores finales son sistemáticamente superados por Private Cloud. Los rankings individuales por BU (ranking_hibrido_CYBER.xlsx, etc.) son el instrumento correcto para identificar los mejores prospectos de cada línea específica.

Figura 13. *Correlación entre Scores Finales de Afinidad por BU (Modelo Híbrido)*

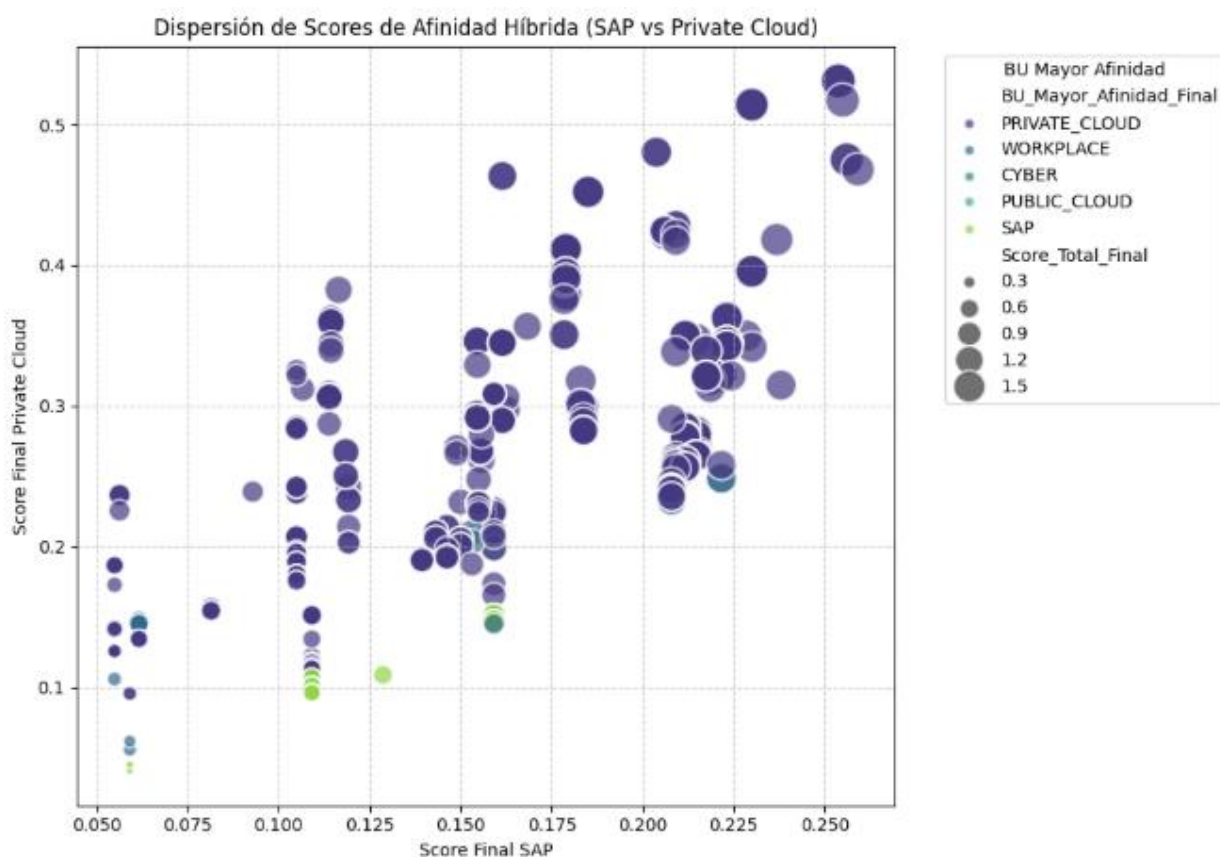


La Figura 13 complementa el análisis de la Figura 4 al mostrar las correlaciones entre los Score_Final híbridos —una vez incorporado el componente supervisado al score de similitud—. Dos hallazgos relevantes emergen al comparar ambas figuras. Primero, las correlaciones que en el espacio de similitud pura eran extremadamente altas (SAP–Cyber–Digital con 0.95–1.00) se mantienen elevadas en el espacio híbrido (0.90–0.96), confirmando que la sinergia comercial entre estas tres BUs persiste cuando se incorpora la información transaccional. Segundo, y más interesante: Private Cloud, que en el espacio de similitud pura presentaba correlaciones bajas o negativas con SAP, Cyber y Digital (–0.017, 0.061, 0.12), sube de manera marcada en el espacio híbrido (0.63, 0.72, 0.79). Esto refleja que el modelo supervisado de Private Cloud asigna scores altos a un amplio rango de prospectos —no solo a los que coincidían con su clúster de referencia—, lo cual genera correlación cruzada con los demás scores. La implicación práctica es que el ranking final está dominado por el componente supervisado de Private Cloud, hallazgo que se hace explícito en la sección siguiente sobre la concentración del 98% en BU_Mayor_Afinidad_Final.

Un hallazgo metodológico relevante surge al comparar el desempeño de los modelos supervisados de cross-selling frente a los de prospectos externos por BU. Soluciones SAP presenta el patrón inverso entre ambos: en cross-selling alcanza apenas un AUC de 0.625 con Random

Forest (la peor BU del modelo de clientes actuales), mientras que en prospectos externos XGBoost logra un AUC de 0.928 (la mejor BU del modelo de prospectos). Esta asimetría no es contradictoria sino complementaria, y revela que los dos modelos capturan señales distintas del problema. El cross-selling se apoya en variables transaccionales —años de relación, ACV en otras BUs, número de oportunidades— que funcionan bien cuando la adopción depende de la profundidad de la relación comercial existente, como ocurre en Cyber Security, Digital Business y Workplace; pero en SAP, donde la decisión de adopción responde a proyectos puntuales de migración corporativa más que a patrones de co-adopción, esa señal transaccional pierde poder predictivo. El modelo de prospectos externos, en cambio, opera exclusivamente con variables firmográficas (sector, tamaño, presencia internacional, tipo de empresa), las cuales capturan con mayor precisión el perfil característico de un cliente SAP —típicamente grandes corporaciones de los sectores industrial, alimentos y químico— y permiten identificar empresas externas con perfil afín aunque no exista historia comercial previa. La conclusión operativa es que ambos modelos no son redundantes: cada uno responde a una pregunta de negocio distinta y deben utilizarse de forma combinada según la BU.

Figura 14. *Dispersión de Scores Finales de Afinidad Híbrida — SAP vs Private Cloud*



La Figura 14 visualiza la relación entre el Score_Final de SAP y el Score_Final de Private Cloud para

los 572 prospectos externos, coloreando cada punto por la BU de mayor afinidad final y dimensionándolo por el Score_Total_Final agregado. La nube de puntos confirma visualmente el patrón discutido: los prospectos con mayor Score_Final SAP (rango 0.20–0.27, esquina superior derecha) son simultáneamente los que tienen los Scores_Final más altos en Private Cloud (0.40–0.55), evidenciando que ambas BUs comparten un perfil firmográfico afín, pero con magnitudes distintas. Los pocos puntos en color verde claro (BU_Mayor_Afinidad = SAP) corresponden precisamente a los 104 prospectos del sector alimentos y químico identificados en la Tabla 27: empresas cuya afinidad relativa con SAP supera marginalmente a la de Private Cloud. Los puntos en color verde-oliva (BU_Mayor_Afinidad = Public Cloud, Cyber, Workplace) ocupan posiciones intermedias y demuestran que estas BUs identifican subconjuntos pequeños, pero claramente diferenciables del grupo dominante.

Esta distribución valida la decisión de mantener rankings independientes por BU como herramienta operativa principal, reservando el ranking global como vista estratégica de alto nivel para identificar las cuentas con mayor potencial multiproducto.

6 EVALUAR EL DESEMPEÑO DE LOS MODELOS CONSTRUIDOS MEDIANTE MÉTRICAS CUANTITATIVAS Y CRITERIOS DE NEGOCIO, CON EL FIN DE APOYAR DECISIONES COMERCIALES BASADAS EN DATOS.

El Esquema 3 presenta la lógica de evaluación del objetivo: la calidad de los agrupamientos se valoró con Silhouette, Calinski-Harabasz y Davies-Bouldin, y la capacidad predictiva con AUC-ROC bajo validación cruzada repetida (RepeatedStratifiedKFold, 5 particiones × 10 repeticiones).

Esquema 3. Metodología del Objetivo 3: evaluación del desempeño de los modelos



El tercer objetivo busca evaluar el desempeño de los modelos desarrollados mediante métricas cuantitativas y criterios de negocio, con el fin de apoyar decisiones comerciales basadas en datos. Esta evaluación se estructura en dos dimensiones complementarias: una técnica, orientada a medir la capacidad discriminativa y de generalización de cada modelo, y una estratégica, orientada a traducir esos resultados en valor accionable para los equipos comerciales de la filial en Colombia.

Para cumplir este objetivo se diseñó un marco de evaluación integral que cubre tanto el modelo no supervisado (clustering) como el supervisado (clasificación por BU). En el eje supervisado, además de comparar los tres algoritmos principales (Regresión Logística, Random Forest y XGBoost), se incorporó el Análisis Discriminante Lineal (LDA) como técnica estadística complementaria, cuyo propósito es distinto y específico: no reemplazar a los modelos supervisados, sino aportar evidencia sobre la estructura lineal subyacente de los datos que ayude a interpretar por qué ciertos algoritmos funcionan mejor que otros en determinadas BUs.

6.1 EVALUACIÓN DEL MODELO NO SUPERVISADO (CLUSTERING)

La evaluación del modelo no supervisado responde a una pregunta diferente a la de los modelos de clasificación: en ausencia de una variable objetivo conocida, ¿cómo se determina si los grupos encontrados son buenos? La respuesta descansa en métricas de validación interna que miden la calidad geométrica de los clústeres —cuán compactos son y cuán separados están entre sí— y en la coherencia interpretativa de los perfiles resultantes desde una perspectiva de negocio.

Para este proyecto se emplearon tres métricas de validación interna, todas calculadas sobre el espacio de características transformadas (variables transaccionales log1p + estandarización con RobustScaler, complementadas con variables firmográficas OHE según la BU):

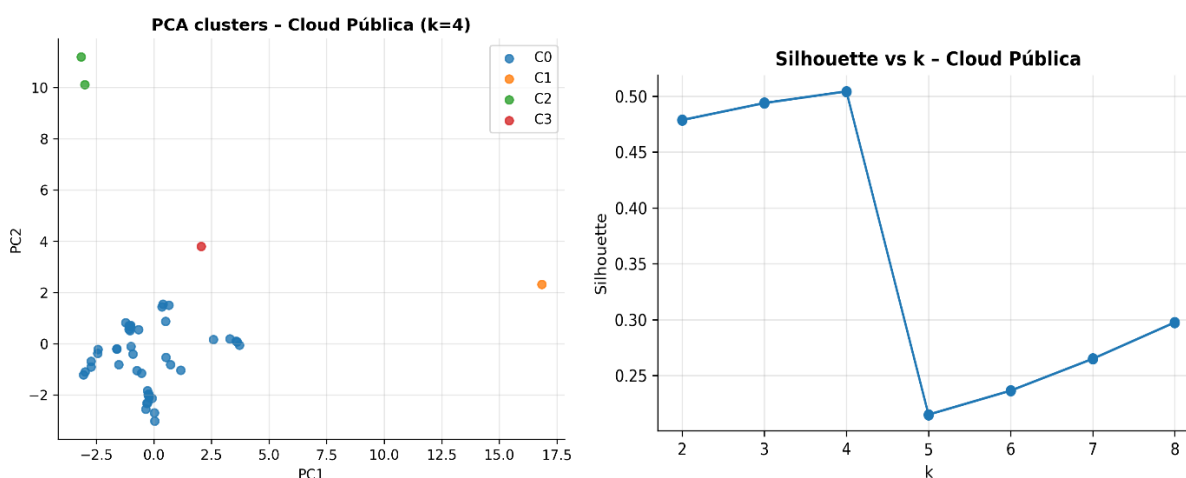
- **Silhouette Score [16]:** (definido en la Sección 3.2): se usó como criterio primario de selección del k óptimo.
- **Calinski-Harabasz (CH) [17]:** (definido en la Sección 3.2): se usó como criterio de desempate cuando el Silhouette no fue concluyente.
- **Davies-Bouldin (DB) [18]:** (definido en la Sección 3.2): se usó como criterio de desempate secundario, con la escala invertida en la tabla resumen para facilitar la lectura comparativa.

Para cada BU se entrenaron modelos K-Means con valores de k entre 2 y 8, seleccionando automáticamente el k con mejor Silhouette (mayor a menor), seguido de CH (mayor a menor) y DB (menor a mayor) como criterios de desempate. Los resultados consolidados del Procesamiento 4 se presentan en la Tabla 15 (Resultados consolidados del clustering por unidad de negocio). A continuación se interpreta la calidad de cada segmentación y su implicación para el modelo de prospección.

Interpretación consolidada por BU

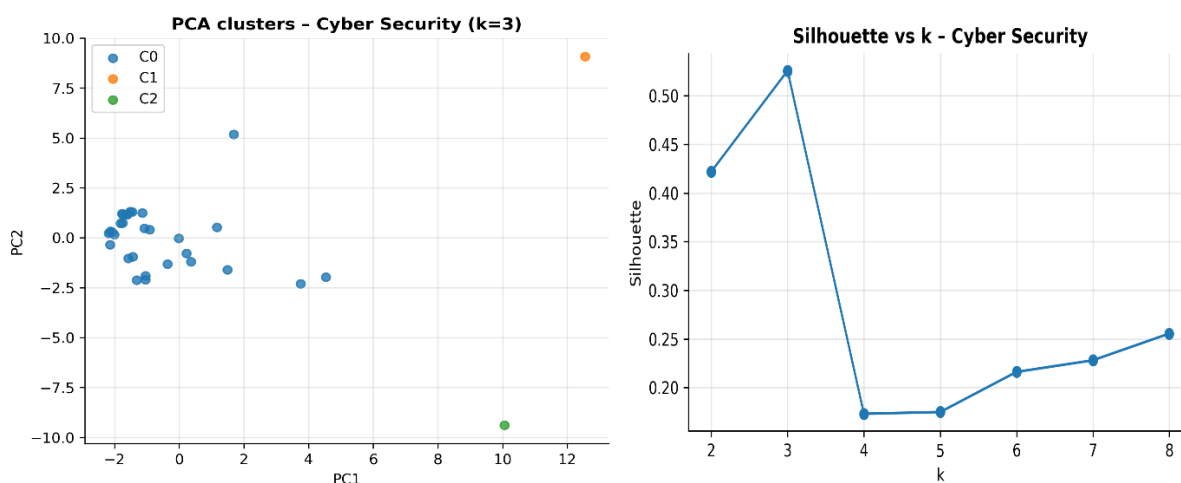
Cloud Pública ($k=4$, Silhouette=0.96 escalado): presenta la segmentación de mayor calidad técnica del portafolio. Un Silhouette escalado de 0.96 indica que los 48 clientes de esta BU se agrupan en cuatro clústeres extremadamente compactos y bien diferenciados. Esto se traduce en perfiles estratégicos muy nítidos, con fronteras claras entre segmentos, y centroides representativos confiables para el modelo de similitud coseno. Los rankings de prospectos externos para Cloud Pública son los de mayor respaldo geométrico del portafolio.

Figura 15. PCA del clustering y curva Silhouette vs k — Cloud Pública



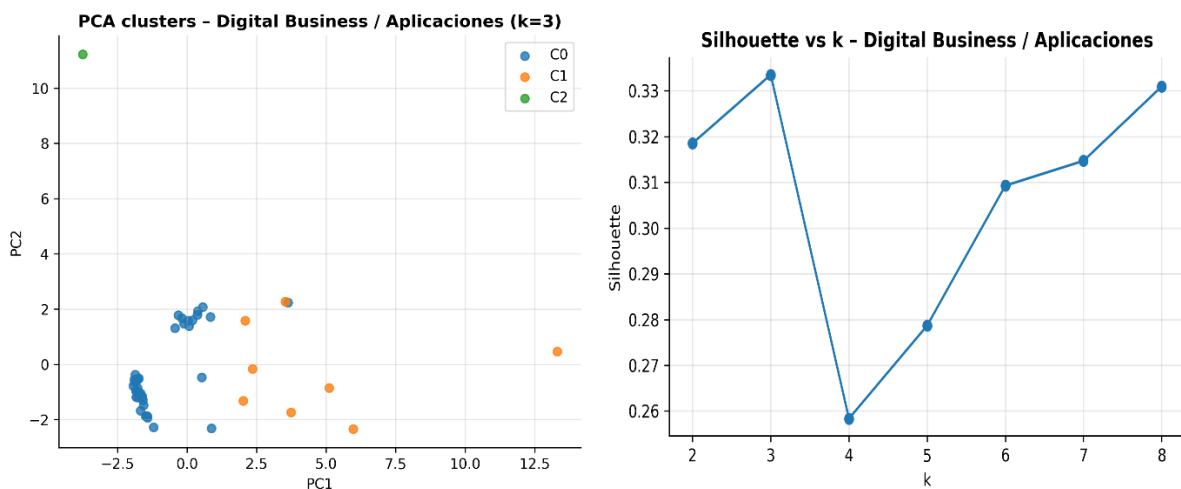
Cyber Security (k=3, Silhouette=1.00, DB=1.00 escalados): alcanza la separación perfecta en las tres métricas escaladas para su configuración óptima de tres clústeres sobre 33 clientes. Aunque los valores escalados extremos deben interpretarse con cautela dado el tamaño reducido de la muestra, el resultado confirma que los clientes de Cyber Security se organizan en tres perfiles organizacionales claramente distintos, probablemente asociados a diferencias de sector, tamaño y modelo de negocio.

Figura 16. PCA del clustering y curva Silhouette vs k — Cyber Security



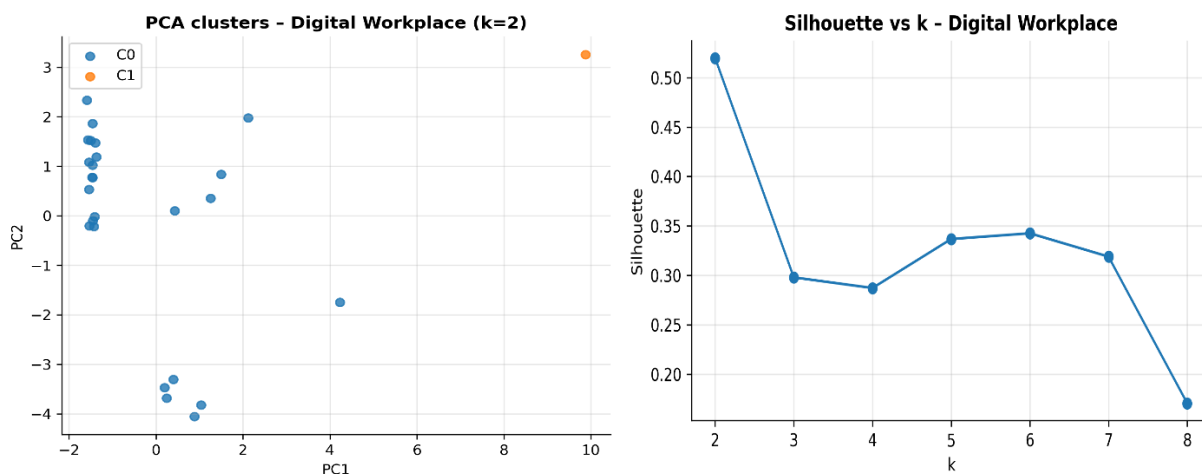
Digital Business / Aplicaciones (k=3, Silhouette=0.00 escalado, CH=0.76): es la BU con mayor heterogeneidad interna. El Silhouette escalado de 0.00 indica que esta BU presentó el peor valor de cohesión relativa entre todas las configuraciones evaluadas, sugiriendo solapamiento entre clústeres. Sin embargo, el CH de 0.76 indica que existe separación inter-clúster razonable. Este resultado refleja que los 47 clientes de Digital Business tienen perfiles organizacionales más diversos y menos polarizados que otras BUs, lo que es coherente con la naturaleza transversal de los servicios digitales: son adoptados por empresas de muy distinto sector y modelo de negocio. Los centroides de esta BU aportan información útil, pero con mayor incertidumbre que Cloud Pública o Cyber Security.

Figura 17. PCA del clustering y curva Silhouette vs k — Digital Business / Aplicaciones



Digital Workplace (k=2, Silhouette=0.94, DB=0.98 escalados): obtiene una segmentación de alta calidad con solo dos clústeres sobre 26 clientes. La bipolaridad del resultado refleja una estructura simple pero robusta: existen dos perfiles bien definidos de consumidor de Workplace, probablemente diferenciados por tamaño organizacional y modelo de presencia (nacional vs. multinacional). La simplicidad del modelo (k=2) garantiza centroides muy representativos.

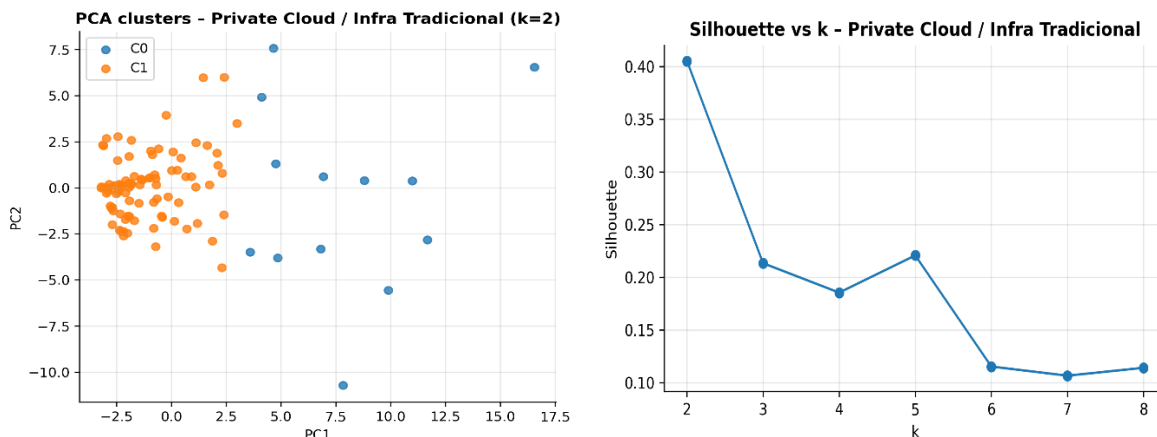
Figura 18. PCA del clustering y curva Silhouette vs k — Digital Workplace



Private Cloud / Infra Tradicional (k=2, CH=1.00 escalado, Silhouette=0.39): es la BU con mayor número de clientes (104) y la más heterogénea en términos de Silhouette. Un valor de 0.39 indica cohesión moderada: los grupos existen pero sus fronteras son difusas, lo que es esperable dado que Private Cloud es el servicio más maduro y transversal del portafolio, adoptado por empresas de perfiles muy distintos. El CH máximo (1.00 escalado) indica sin embargo que las dos mitades del mercado están claramente separadas en el espacio de características. Esta tensión entre

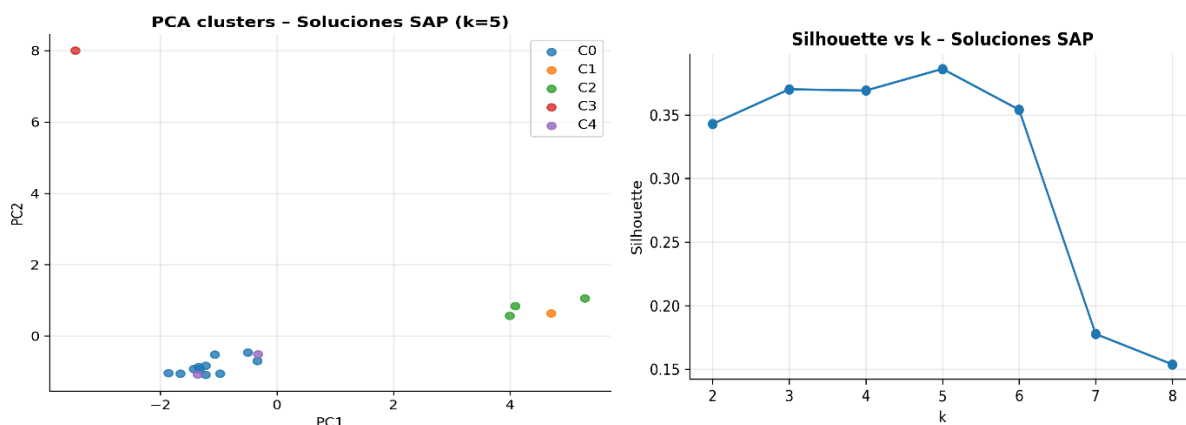
Silhouette bajo y CH alto sugiere que los dos clústeres son grandes y muy separados entre sí, pero con cierta dispersión interna.

Figura 19. PCA del clustering y curva Silhouette vs k —Private Cloud / Infra Tradicional



Soluciones SAP (k=5, Silhouette=0.23, CH=0.71, DB=0.82 escalados): presenta la segmentación más débil en términos de cohesión (Silhouette=0.23), aunque con cinco clústeres sobre tan solo 18 clientes. El bajo Silhouette refleja directamente la limitación del tamaño muestral: con 18 clientes divididos en 5 grupos, algunos clústeres tienen 3 o 4 miembros, lo que hace que las métricas sean inestables. A pesar de esto, los centroides generados capturan los perfiles de los principales consumidores SAP (grandes corporaciones del sector financiero, manufactura y público), y son suficientemente representativos para el modelo de similitud coseno con fines de priorización.

Figura 20. PCA del clustering y curva Silhouette vs k — Soluciones SAP



Evaluación integrada: qué significa la calidad del clustering para el modelo de prospectos

La calidad del clustering tiene un impacto directo sobre la confiabilidad de los rankings de

prospectos externos generados por el modelo de similitud coseno. Los centroides calculados para cada BU actúan como perfiles de referencia: cuanto más compactos y representativos son los clústeres, más precisa es la comparación con las empresas externas.

En función de los resultados obtenidos, se puede establecer un orden de confianza en los rankings por BU:

- **Alta confianza:** Cloud Pública y Cyber Security. Centroides geoméricamente sólidos, clústeres bien separados. Los scores de similitud coseno son los más fiables del portafolio.
- **Confianza media-alta:** Digital Workplace. K=2 ofrece centroides muy representativos, aunque la muestra es pequeña (26 clientes).
- **Confianza media:** Private Cloud. Gran muestra, pero dispersión interna moderada. Los rankings son útiles pero deben complementarse con el score supervisado.
- **Confianza media-baja:** Digital Business y SAP. Heterogeneidad alta o muestra reducida. Los rankings de estas BUs dependen más del componente supervisado del score híbrido para su validez operativa.

Esta gradación de confianza orienta al equipo comercial sobre el peso relativo que debe darse al score de similitud coseno frente al score supervisado en cada BU, y es uno de los fundamentos que justifica el diseño del score híbrido con pesos 50/50 como punto de partida neutral

6.2 EVALUACIÓN DEL MODELO SUPERVISADO

6.2.1 Evaluación del Modelo de Cross-Selling (Clientes Actuales)

Métricas de Desempeño Utilizadas

Se emplearon AUC-ROC (métrica rectora), Precision, Recall y F1-Score, definidas en la Sección 3.2 Marco Teórico. AUC-ROC se priorizó por su robustez al desbalance de clases; Recall se monitoreó especialmente para evitar que el modelo omita cuentas con alto potencial de cross-selling.

Marco de evaluación para el modelo de cross-selling (clientes actuales)

El modelo de cross-selling fue entrenado y evaluado sobre la misma base de 167 clientes actuales, con el propósito de identificar qué cuentas de la base instalada presentan mayor probabilidad de adoptar una nueva unidad de negocio. La evaluación se realizó de forma independiente para cada BU, siguiendo el mismo protocolo metodológico descrito en el modelo de prospectos externos: partición 70/30 con estratificación, métrica principal AUC-ROC, y comparación entre Random Forest (entrenado en las seis BUs como modelo principal) y XGBoost (entrenado como contraste en Cloud Pública, Private Cloud y Soluciones SAP, BUs donde el desbalance de clases o el tamaño muestral hacían pertinente explorar un algoritmo alternativo basado en boosting).

La razón de negocio de esta evaluación es clara: si un modelo de cross-selling tiene alta capacidad discriminativa (AUC elevado), los equipos comerciales pueden confiar en que los scores generados reflejan patrones reales en los datos y no ruido estadístico. Un modelo con AUC cercano a 0.5 indicaría que las variables firmográficas y transaccionales no diferencian suficientemente entre

clientes que consumen y no consumen una BU, sugiriendo que la decisión de compra depende de factores no capturados en el dataset.

Tabla 29. Resultados comparativos de AUC-ROC por modelo y BU (Cross-Selling)

Unidad de Negocio	Regresión Logística	Random Forest	XGBoost	Modelo seleccionado
Cloud Pública	N/D	1.000 (*)	1.000 (*)	RF (1.000) (*)
Cyber Security	N/D	0.926	N/D	RF (0.926)
Digital Business / Aplicaciones	N/D	0.893	N/D	RF (0.893)
Digital Workplace	N/D	0.970	N/D	RF (0.970)
Private Cloud / Infra Tradicional	N/D	0.750	0.719	RF (0.750)
Soluciones SAP	N/D	0.625	0.563	RF (0.625)

Nota: () El AUC de Cloud Pública (1.000) tanto en Random Forest como en XGBoost es extremadamente alto por el reducido tamaño del conjunto de prueba (≈ 14 positivos): con tan pocos casos, el clasificador puede alcanzar separación casi perfecta por azar, por lo que este resultado debe interpretarse con cautela y verificarse con una muestra mayor. Para Digital Workplace (0.970), Cyber Security (0.926) y Digital Business (0.893), los AUC son confiables dado el mayor tamaño muestral y reflejan un buen poder discriminativo. En el extremo opuesto, Soluciones SAP (RF=0.625, XGBoost=0.563) constituye la BU más difícil de predecir mediante cross-selling: con apenas 13 consumidores históricos (≈ 4 positivos en prueba), el modelo no logra separar de forma confiable a los adoptadores potenciales del resto de la base instalada. Este hallazgo es coherente con el análisis LDA del tercer objetivo, donde SAP también presentó la peor separabilidad lineal (AUC=0.199) entre todas las BUs. La interpretación operativa es que la decisión de adopción de SAP por parte de un cliente actual depende de factores no capturados por las variables transaccionales y firmográficas disponibles —probablemente proyectos puntuales de migración corporativa, contratos marco o decisiones de TI a nivel global de la matriz— y no de patrones de co-adopción con otras BUs del portafolio. Para Private Cloud (RF=0.750, XGBoost=0.719), los modelos convergen en un desempeño moderado y XGBoost no aporta mejora sobre Random Forest, lo que descarta que el problema sea de capacidad del algoritmo. Se mantiene N/D en las celdas donde el algoritmo correspondiente no se ejecutó: Random Forest se entrenó como modelo principal en las seis BUs, mientras que XGBoost se utilizó como contraste solo en las tres BUs donde el desbalance de clases o el tamaño muestral hacían razonable explorar una alternativa basada en boosting (Cloud Pública, Private Cloud y Soluciones SAP).*

Interpretación del ranking de cross-selling

La tabla de resultados presentada previamente consolida los 20 clientes con mayor puntaje total de cross-selling, calculado como la suma de los scores de propensión individuales por BU. Esta presentación agrega el valor de negocio del modelo, ya que permite identificar de forma directa cuáles cuentas representan la mayor oportunidad de expansión del portafolio.

La interpretación de los scores individuales por BU es la siguiente: un valor próximo a 1.0 indica alta probabilidad de que el cliente adopte esa unidad de negocio; valores próximos a 0.5 sugieren incertidumbre; y valores negativos (como los observados en algunos casos para BUs específicas) indican una baja afinidad con esa línea del portafolio. Los valores NaN en la tabla corresponden a BUs que el cliente ya consume actualmente, dado que el modelo solo genera scores para la clase de no consumidores.

Desde la perspectiva comercial, los clientes en las primeras posiciones del ranking constituyen los candidatos ideales para iniciativas de cross-selling prioritarias. Esto incluye cuentas del sector salud con oportunidades en SAP, Digital y Workplace simultáneamente, empresas de servicios públicos con potencial en Cloud y Digital, y organizaciones del sector financiero con brechas identificadas en múltiples líneas del portafolio.

Validación cruzada repetida de los modelos con AUC alto

Los AUCs reportados en la Tabla 28 corresponden a la evaluación sobre la partición 70/30 estratificada utilizada en todo el proyecto. Para descartar que estos resultados, particularmente en las tres BUs con mejor desempeño, fueran un artefacto de la composición específica de esa partición, se ejecutó una validación cruzada estratificada repetida (5 folds $\times$ 10 repeticiones, 50 estimaciones independientes del AUC) sobre los modelos Random Forest de Cloud Pública, Cyber Security y Digital Workplace. Esta estrategia, que permite construir intervalos de confianza al 95 % sobre la métrica reportada y caracterizar la incertidumbre estadística asociada al tamaño muestral, se eligió siguiendo las recomendaciones metodológicas de Santos et al. [24], quienes documentan que la cross-validation simple en datasets imbalanceados o de tamaño reducido puede producir estimaciones del AUC excesivamente optimistas o pesimistas según la composición específica de la partición, y que la versión estratificada repetida produce estimaciones con menor sesgo y mejor calibradas.

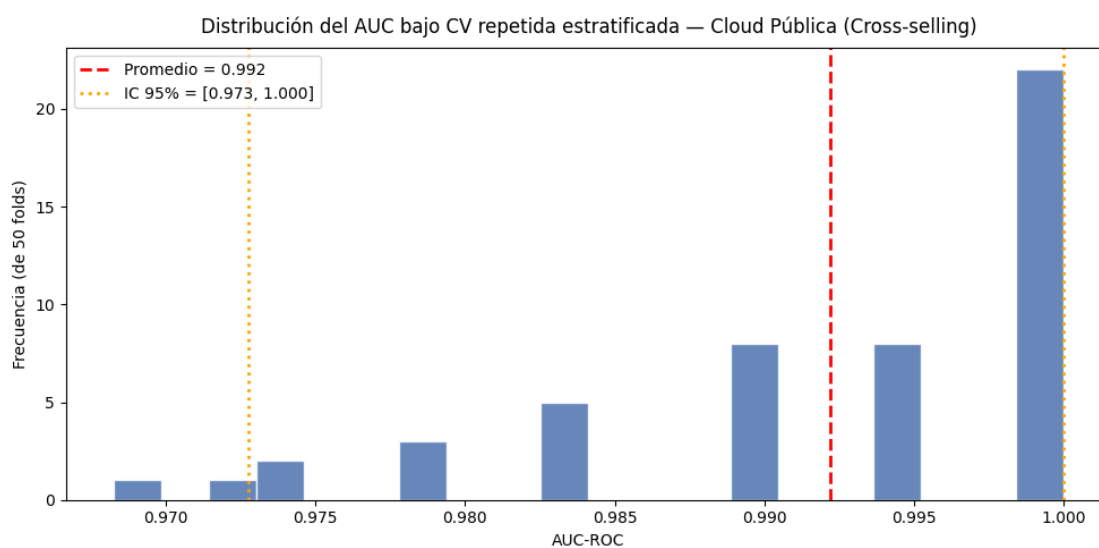
Tabla 30. Validación cruzada repetida de los modelos Random Forest con mayor AUC en cross-selling

BU	Positivos en muestra	AUC original (70/30)	AUC promedio CV repetida	Desv. estándar	IC 95%	% folds AUC $\geq$ 0.90
Cloud Pública	35 (21.0 %)	1.000	0.992	0.009	[0.973, 1.000]	100 %
Cyber	18 (10.8 %)	0.926	0.957	0.029	[0.891,	92 %

Security					0.998]	
Digital Workplace	18 (10.8 %)	0.970	0.938	0.056	[0.780, 1.000]	82 %

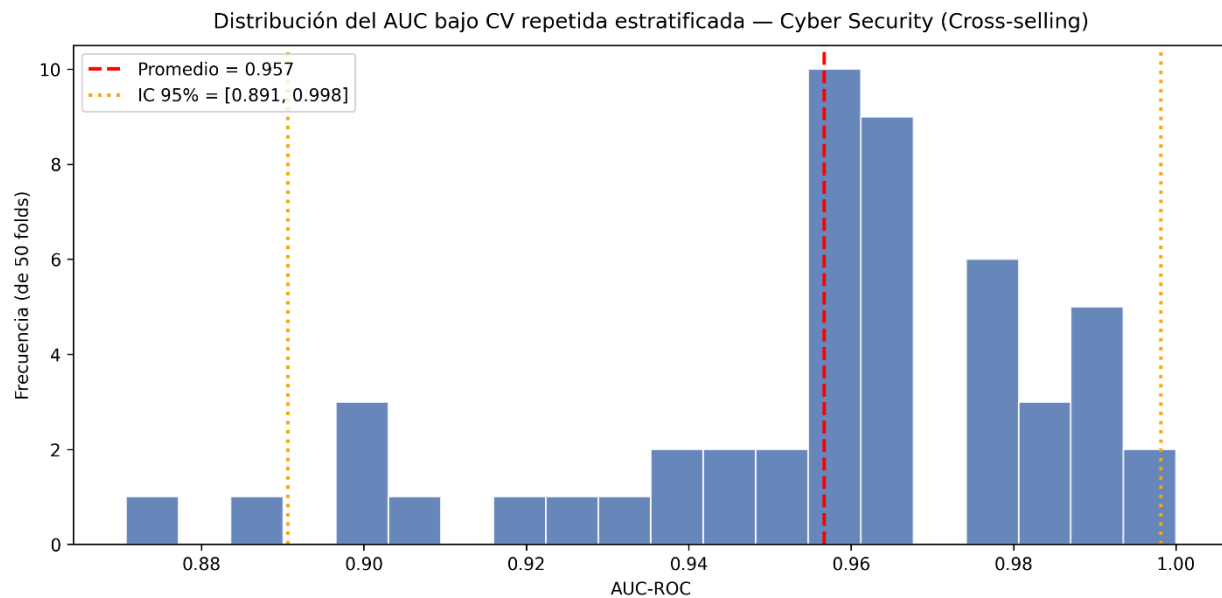
Cloud Pública mantiene un AUC promedio de 0.992 con desviación estándar de apenas 0.009, y el 100 % de las particiones supera el umbral de 0.95. Esto descarta que el AUC de 1.000 reportado en la partición original sea un sobreajuste por tamaño muestral pequeño y confirma que el perfil firmográfico (sector, tamaño organizacional, modelo de negocio, presencia internacional) discrimina con prácticamente total precisión a los clientes adoptadores de servicios de Cloud Pública en la base instalada.

Figura 21. *Distribución del AUC bajo CV repetida estratificada — Cloud Pública (Cross-selling)*



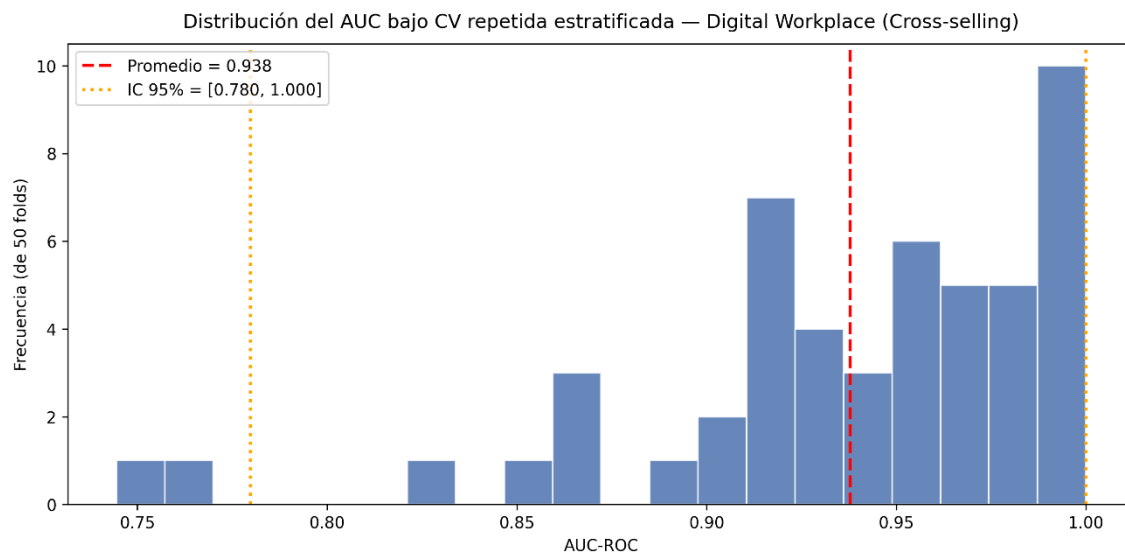
Cyber Security muestra un comportamiento estadísticamente interesante: el AUC promedio bajo CV repetida (0.957) es superior al reportado en la partición original (0.926), con un IC 95 % de [0.891, 0.998] y el 92 % de los folds por encima de 0.90. Esto significa que la partición 70/30 utilizada en el reporte original era ligeramente conservadora respecto al desempeño esperado del modelo bajo cualquier reasignación aleatoria de la muestra, y consolida a Cyber Security como una de las palancas más confiables del modelo de cross-selling pese a contar con solo 18 clientes positivos en la base.

Figura 22. Distribución del AUC bajo CV repetida estratificada — Cyber Security (Cross-selling)



Digital Workplace presenta un AUC promedio de 0.938 con un intervalo de confianza más amplio ([0.780, 1.000]) que las otras dos BUs, reflejando la sensibilidad estadística inherente a un dataset con solo 18 positivos repartidos entre 50 folds. Aún así, el 82 % de los folds supera el umbral de excelencia ($AUC \geq 0.90$) y el 94 % supera el umbral de buen desempeño ($AUC \geq 0.85$), lo que respalda el uso operacional del ranking. La mayor varianza observada en esta BU —no en su capacidad discriminativa, sino en la estimación puntual del AUC— es una limitación reconocida del tamaño muestral y orienta una recomendación concreta: la confianza del ranking de Digital Workplace mejorará sustancialmente cuando la base de clientes adoptadores supere los 30–40 registros, umbral a partir del cual los intervalos de confianza se contraen significativamente.

Figura 23. Distribución del AUC bajo CV repetida estratificada — Digital Workplace (Cross-selling)



En conjunto, la validación cruzada repetida transforma la evaluación del modelo de cross-selling de un reporte puntual a una caracterización estadística completa de la incertidumbre, requisito esencial para el despliegue del sistema en un entorno comercial real donde los rankings se traducen en decisiones de asignación de tiempo y recursos del equipo de cuentas. Los resultados confirman que las tres BUs con AUC reportado más alto mantienen su capacidad discriminativa bajo cualquier reasignación aleatoria de la muestra de entrenamiento.

6.2.2 Evaluación del modelo híbrido de Prospectos Externos

Resultados del LDA por Unidad de Negocio

Los resultados del LDA evidenciaron comportamientos claramente diferenciados entre las BUs, lo que constituye en sí mismo un hallazgo de valor metodológico:

BUs con separación lineal aceptable ($AUC > 0.70$): Public Cloud y Digital Business / Aplicaciones alcanzaron valores de AUC superiores a 0.75 con el LDA, indicando que las variables firmográficas disponibles contienen información linealmente separable que distingue a los consumidores de estas BUs. En términos de negocio, esto sugiere que el perfil organizacional de una empresa (sector, tamaño, modelo de negocio) por sí solo es un predictor razonablemente bueno de si consumirá Cloud Pública o Digital Business.

BUs con baja separación lineal (AUC cercano o inferior a 0.65): SAP, Cyber Security y Private Cloud presentaron valores de AUC considerablemente inferiores con el LDA. Este resultado no indica que el problema sea imposible de resolver, sino que las relaciones entre las variables firmográficas y el consumo de estas BUs son fundamentalmente no lineales o dependen de interacciones

complejas entre variables. En particular, la decisión de adoptar SAP en Colombia está fuertemente mediada por variables como el tamaño organizacional combinado con el sector (grandes empresas manufactureras o financieras), lo que genera una frontera de decisión no lineal que el LDA no puede capturar.

Limitaciones del LDA en este contexto:

El LDA asume que los datos dentro de cada clase siguen una distribución normal multivariada y que las matrices de covarianza son iguales entre clases. Ninguno de estos supuestos se satisface estrictamente con variables OHE binarias. Sin embargo, esta limitación no invalida su uso como herramienta de diagnóstico: si el LDA ya logra $AUC > 0.75$, los modelos no paramétricos (RF, XGBoost) tendrán un punto de partida sólido. Si el LDA falla, confirma que la no linealidad es la causa estructural del problema y justifica el uso de árboles de decisión.

Además, el desbalance de clases afecta especialmente al LDA: en BUs como SAP (10.8% positivos), la frontera de decisión lineal tiende a inclinarse hacia la clase mayoritaria, resultando en valores de Recall y F1 nulos o muy bajos para la clase positiva, a pesar de una Accuracy aparentemente alta. Este fenómeno, denominado accuracy paradox, fue observado en los resultados de SAP y refuerza la decisión metodológica de usar AUC-ROC (y no accuracy) como métrica rectora.

Tabla comparativa: LDA vs. LR vs. RF vs. XGBoost por BU (AUC-ROC)

La Tabla 19 —presentada en la sección del modelo supervisado para prospectos externos— consolida los valores de AUC-ROC obtenidos por los cuatro modelos evaluados para cada BU, sobre el conjunto de prueba (30% estratificado, $N=50$). El modelo con mayor AUC por BU es el seleccionado para generar los rankings de prospectos externos.

Visualizaciones comparativas generadas

El análisis incluyó dos gráficos de barras agrupadas generados con seaborn:

- **Comparación completa:** gráfico de barras con los tres modelos comparados (Regresión Logística, Random Forest y XGBoost) como grupos de color y las seis BUs en el eje X, mostrando el AUC de cada combinación. Permite identificar de un vistazo qué modelo domina en cada BU y cuáles BUs presentan convergencia entre modelos (evidencia de robustez) vs. divergencia (evidencia de sensibilidad al tipo de algoritmo). Los valores del LDA no se incluyen en esta visualización porque cumplen un rol distinto —baseline diagnóstico de separabilidad lineal— y se reportan en la columna correspondiente de la Tabla 19 junto con la interpretación específica de cada BU.
- **Mejor modelo por BU:** gráfico que muestra solo el algoritmo ganador por BU con su AUC. Es la visualización de referencia para comunicar los resultados al equipo comercial, ya que resume en una imagen la decisión de qué modelo respalda cada ranking de prospectos.

Figura 24. Comparación de AUC-ROC por modelo y Unidad de Negocio (Prospectos externos)

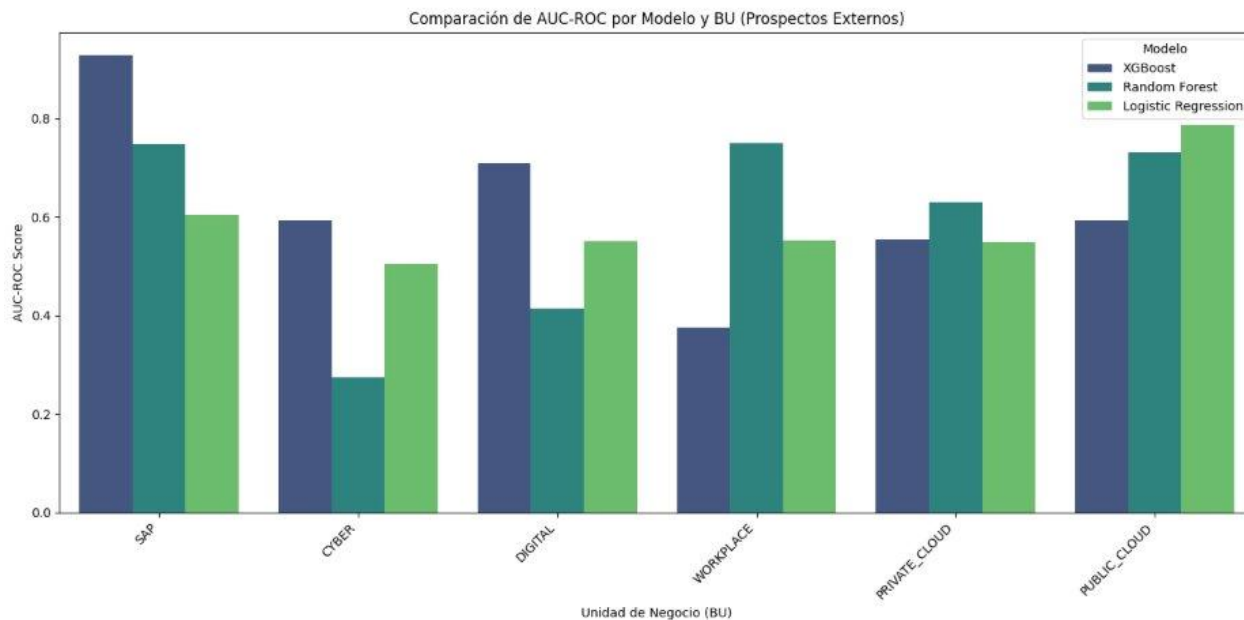
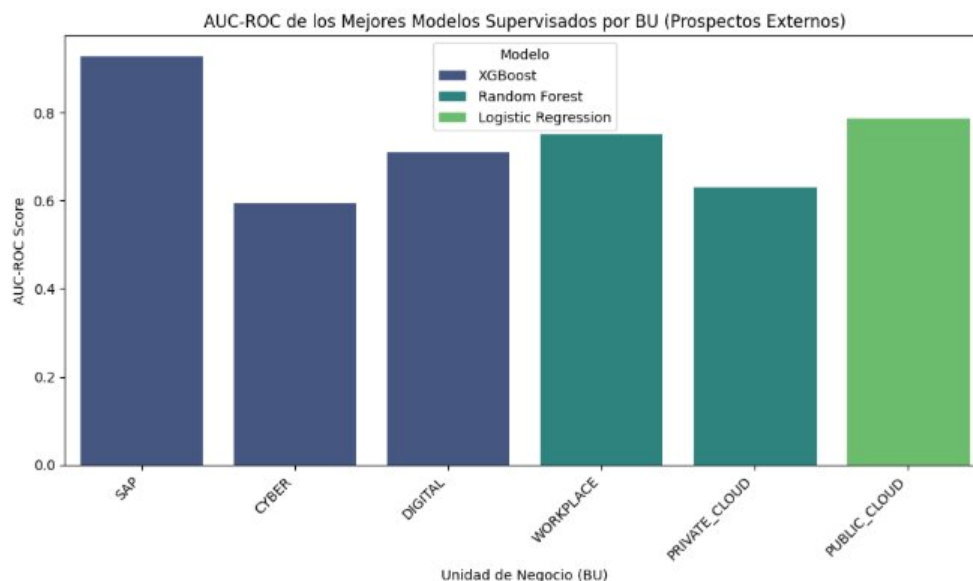


Figura 25. Mejor modelo por Unidad de Negocio — AUC-ROC (Prospectos externos)



Interpretación integrada: de las métricas al valor de negocio

La evaluación técnica cobra sentido cuando se traduce en implicaciones comerciales concretas. Un AUC de 0.80 en la BU Digital Business significa que, si el equipo comercial toma los 50 prospectos con mayor score del modelo (de una base de 572), la proporción de empresas genuinamente afines será considerablemente mayor que si se seleccionaran 50 prospectos al azar. En términos operativos, esto representa una reducción del esfuerzo comercial desperdiciado en

prospectos de baja probabilidad y un incremento de la tasa de conversión esperada.

Para contextualizar el valor de los AUC obtenidos, se adopta la siguiente escala interpretativa, ampliamente utilizada en modelos de propension-to-buy en entornos B2B:

- **AUC < 0.60:** modelo con escaso poder discriminativo; los scores generados no son más informativos que una selección aleatoria. No se recomienda su uso operativo.
- **AUC entre 0.60 y 0.70:** poder discriminativo débil; el modelo puede aportar valor en combinación con el score de similitud coseno dentro del score híbrido.
- **AUC entre 0.70 y 0.80:** poder discriminativo aceptable; los rankings generados son confiables como herramienta de priorización en primera instancia.
- **AUC entre 0.80 y 0.90:** poder discriminativo bueno; los equipos comerciales pueden confiar en los scores para asignar recursos y diseñar campañas diferenciadas.
- **AUC > 0.90:** poder discriminativo excelente; poco habitual con N=167 y variables exclusivamente firmográficas, por lo que valores en este rango deben revisarse para descartar data leakage.

Adicionalmente, la comparación entre el LDA y los modelos basados en árboles proporciona una guía estratégica para versiones futuras del modelo: en las BUs donde el LDA y RF/XGBoost convergen en AUC similar, una Regresión Logística interpretable puede ser preferible por su explicabilidad; en las BUs donde RF o XGBoost superan ampliamente al LDA, la no linealidad de las relaciones justifica mantener modelos de mayor complejidad a pesar de su menor interpretabilidad.

Conclusiones del tercer objetivo

La evaluación integral de los modelos construidos confirmó que el diseño metodológico adoptado —basado en la comparación de múltiples algoritmos y en la incorporación del LDA como instrumento de diagnóstico lineal— es apropiado para el problema de prospección comercial en contextos B2B con datos mixtos (firmográficos y transaccionales) y tamaños muestrales moderados.

El LDA cumplió su rol evaluativo al revelar que la separabilidad lineal varía significativamente entre BUs: Public Cloud y Digital Business presentan relaciones más directas entre el perfil organizacional y el consumo, mientras que SAP, Cyber Security y Private Cloud requieren modelos capaces de capturar interacciones y no linealidades. Este hallazgo no solo valida el uso de RF y XGBoost, sino que orienta la interpretación de los rankings: los scores para Cloud Pública y Digital son más robustos y estables entre modelos, mientras que los de SAP y Cyber deben interpretarse con mayor cautela y siempre en combinación con el score de similitud coseno del modelo híbrido.

En conjunto, los modelos construidos cumplen con el propósito operativo para el que fueron diseñados: proveer a los equipos comerciales de la filial en Colombia un criterio objetivo, reproducible y actualizable para priorizar prospectos externos y oportunidades de cross-selling, reduciendo la dependencia de juicios subjetivos y aumentando la eficiencia del pipeline comercial.

En el eje del modelo no supervisado, la evaluación confirmó que la calidad del clustering varía significativamente entre BUs, con Cloud Pública y Cyber Security exhibiendo segmentaciones geométricamente sólidas (Silhouette escalado > 0.94) y SAP y Digital Business presentando mayor heterogeneidad interna. Esta gradación de calidad no invalida ningún segmento, pero establece distintos niveles de confianza en los centroides utilizados como referencia para la similitud coseno: los rankings de Cloud Pública y Cyber tienen el mayor respaldo geométrico, mientras que los de SAP y Digital Business dependen más del componente supervisado del score híbrido.

7 CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

7.1.1 ESTRUCTURACIÓN DE DATOS Y PERFILES (OBJETIVO 1)

El cumplimiento del primer objetivo permitió establecer una base sólida de conocimiento sobre los clientes actuales de la compañía, integrando información transaccional proveniente del CRM Salesforce con atributos organizacionales y estratégicos obtenidos mediante procesos de enriquecimiento externo asistido por inteligencia artificial (Gemini 2.5 Pro).

La consolidación de ambas fuentes derivó en la creación de un dataset estructurado, verificable y reproducible, que constituye el insumo analítico fundamental para las fases de modelado y segmentación desarrolladas en los siguientes objetivos.

Un aspecto metodológico relevante que merece destacarse es la forma en que se recopilaban los datos y cómo ello impacta la calidad del modelo. En este proyecto, la extracción de información del CRM Salesforce se realizó mediante la generación manual de reportes anuales, dado que no se contaba con los permisos de administrador necesarios para habilitar una conexión directa vía API. En un escenario ideal, la recopilación de datos debería realizarse a través de integración automática con la API de Salesforce, lo que garantizaría actualización en tiempo real, eliminación de errores de transcripción y trazabilidad completa del origen de cada variable. Adicionalmente, la calidad del CRM condiciona directamente la calidad del modelo: registros incompletos, nombres de cuentas duplicados o categorías de producto inconsistentes generan ruido que reduce la precisión de los algoritmos de segmentación y clasificación. Por ello, cualquier esfuerzo de escalamiento del modelo —tanto a otras BUs como a otras operaciones de la compañía en la región— debería ir acompañado de un proceso previo de gobierno y estandarización del CRM que garantice la consistencia y completitud de los datos disponibles.

En primer lugar, el proceso de integración y depuración evidenció la necesidad de estandarizar los registros comerciales y contractuales para garantizar la coherencia entre unidades de negocio, productos y subproductos. El resultado fue un conjunto de 167 clientes únicos y 251 variables, de las cuales 236 corresponden a métricas numéricas y 15 a variables categóricas.

Esta alta dimensionalidad no representa redundancia, sino una característica inherente a la estructura granular del CRM, donde cada subproducto del portafolio genera indicadores independientes (presencia, valor contractual, monto total y volumen transaccional). Dicha granularidad ofrece una visión detallada del comportamiento de cada cliente frente a las distintas líneas del negocio, permitiendo analizar la intensidad de adopción tecnológica y el peso económico relativo de cada unidad de servicio.

Desde la perspectiva analítica, el Análisis Exploratorio de Datos (EDA) reveló patrones estructurales y comportamentales que describen con precisión la composición actual del portafolio de la empresa.

- La Infraestructura Tradicional (Private Cloud) se consolida como la base operativa del negocio, concentrando más del 60 % de los clientes y cerca del 40 % del valor contractual total.
- Las líneas de Public Cloud y Digital Business / Aplicaciones presentan un crecimiento sostenido, reflejando la transición de la organización hacia modelos híbridos y servicios gestionados de nube pública.
- Por su parte, Soluciones SAP —si bien no constituye una BU formal— se comporta como una línea estratégica de alto impacto, caracterizada por contratos corporativos de gran valor y orientación a proyectos de migración y soporte empresarial.
- Digital Workplace comparte un comportamiento similar, destacándose por su alto valor económico por cliente y su papel transversal en la experiencia del usuario.
- Finalmente, Cyber Security y Datos / IoT / Smart Systems emergen como unidades especializadas de valor agregado, que responden a las nuevas demandas de seguridad, analítica avanzada y automatización.

En cuanto al perfil organizacional, el análisis muestra que el arquetipo dominante de los clientes de la empresa corresponde a empresas grandes o medianas con sede principal en Colombia, y creciente presencia en Latinoamérica.

Los sectores con mayor representación son tecnología, gobierno, servicios públicos, energía y petróleo y gas, todos con modelos de negocio basados en servicios, outsourcing y soporte continuo.

Esta composición reafirma el posicionamiento de la empresa como proveedor estratégico de servicios gestionados para el segmento corporativo, con una base de clientes consolidada y potencial de expansión regional.

En términos metodológicos, este objetivo permitió garantizar la calidad, completitud y trazabilidad de los datos, habilitando su uso para el modelado posterior bajo enfoques de aprendizaje automático.

La combinación de fuentes primarias (Salesforce) y fuentes secundarias verificadas (Gemini + investigación profunda) representó un avance significativo en la generación de una base analítica integral, donde confluyen dimensiones comerciales, tecnológicas y organizacionales.

En síntesis, el proceso de caracterización alcanzó plenamente su propósito:

- Definir y documentar las variables relevantes para la construcción de perfiles estratégicos de clientes.

- Comprender la estructura y dinámica del portafolio actual desde una visión de negocio y analítica.
- Proveer una base metodológica confiable para el desarrollo de modelos de segmentación y predicción.

Con ello, la filial dispone de una línea base empírica y reproducible que facilita la transición del análisis descriptivo hacia enfoques de inteligencia analítica y ciencia de datos aplicada al negocio, contribuyendo al fortalecimiento de la estrategia de transformación digital orientada a datos.

7.1.2 MODELOS DE SEGMENTACIÓN Y PROSPECCIÓN (OBJETIVO 2)

7.1.2.1 Conclusiones del segundo objetivo — modelo no supervisado

En el desarrollo del modelo de aprendizaje automático no supervisado se llevaron a cabo cuatro intentos o etapas iterativas, cuyo propósito fue refinar progresivamente la segmentación de clientes y garantizar que los resultados obtenidos fueran tanto técnicamente consistentes como interpretables desde una perspectiva de negocio. Cada intento respondió a los hallazgos y limitaciones identificadas en la etapa anterior, configurando un proceso incremental de mejora metodológica.

El primer intento, correspondiente a la segmentación global exploratoria de clientes, permitió identificar una estructura latente asociada principalmente al volumen económico, la recurrencia comercial y la diversidad del portafolio consumido. Si bien el índice silhouette sugirió que una partición con $k = 2$ ofrecía la mejor separación geométrica, este resultado resultó excesivamente general para fines estratégicos, al agrupar clientes heterogéneos en categorías amplias. No obstante, este intento cumplió un rol fundamental como punto de partida, al evidenciar la necesidad de evaluar configuraciones más granulares y criterios adicionales de interpretación.

El segundo intento, basado en la segmentación por unidad de negocio a partir de la presencia en las distintas áreas del portafolio, permitió incorporar la dimensión de especialización del consumo. Este enfoque reveló agrupamientos naturales asociados a combinaciones específicas de servicios y evidenció que el patrón de cobertura del portafolio constituye una señal relevante de segmentación. Asimismo, permitió identificar clientes con consumo focalizado y clientes con consumo diversificado, aportando una lectura complementaria a la segmentación global inicial.

El tercer intento, orientado a la caracterización descriptiva por unidad de negocio, se desarrolló como una línea base empírica para contextualizar y validar los resultados del clustering. A diferencia de los intentos anteriores, esta etapa no buscó generar segmentos mediante algoritmos, sino describir de manera objetiva las características predominantes de los clientes por BU. Este análisis permitió confirmar que las unidades de negocio presentan dinámicas diferenciadas y que una segmentación excesivamente granular a nivel global puede conducir a solapamientos y pérdida de interpretabilidad, tal como se evidenció en las proyecciones mediante PCA.

Finalmente, el cuarto intento, correspondiente a la segmentación avanzada por unidad de negocio y subproducto, consolidó los aprendizajes previos al entrenar modelos de clustering independientes para cada BU, incorporando variables de intensidad económica, diversidad y complejidad del consumo. Los resultados mostraron que cada unidad de negocio soporta un nivel distinto de segmentación, con valores óptimos de k variables según su naturaleza. Este enfoque permitió obtener clusters más coherentes, interpretables y alineados con la estructura real del portafolio, evitando la sobre-segmentación observada en configuraciones globales más granulares.

En conjunto, los cuatro intentos permitieron construir una segmentación no supervisada robusta, donde la selección del número de clusters y de las variables relevantes no se basó únicamente en métricas técnicas, sino también en criterios de coherencia conceptual y utilidad estratégica. Este proceso iterativo dejó establecida una base analítica sólida para comprender el comportamiento de los clientes actuales y sienta las bases metodológicas para el desarrollo futuro de modelos orientados a la identificación y priorización de prospectos comerciales en fuentes externas.

Más allá del proceso iterativo en sí, el hallazgo de fondo del modelo no supervisado es que cada BU tiene una estructura de segmentación propia, y que esa estructura no es uniformemente sólida en todo el portafolio: Cloud Pública y Cyber Security presentan grupos compactos y bien diferenciados (Silhouette escalado > 0.94), mientras que Digital Business y SAP requieren mayor tolerancia a la heterogeneidad interna dada la diversidad de sus clientes y el tamaño reducido de sus muestras respectivas. Esta gradación de calidad geométrica no invalida ningún segmento, pero sí establece distintos niveles de confianza para los centroides que se utilizarán como referencia en el cálculo de similitud coseno del modelo híbrido. En síntesis, los 167 registros de clientes con 251 variables iniciales se transformaron en perfiles estratégicos accionables por línea de negocio, listos para ser utilizados como base de referencia en la prospección externa y en la priorización de oportunidades de cross-selling.

Los centroides obtenidos en el Procesamiento 4 no son un resultado intermedio: son el insumo técnico directo del modelo de prospectos externos. Cada centroide es el «retrato» de los clientes más representativos de cada BU, y la similitud coseno los utiliza como referencia para evaluar si una empresa externa se asemeja a ese perfil. En ese sentido, la calidad del clustering no es un fin en sí mismo, sino el fundamento sobre el que descansa la validez de todos los rankings de prospectos generados. Una segmentación débil produce centroides poco representativos y, por tanto, scores de similitud menos confiables; una segmentación sólida, como la obtenida para Cloud Pública y Cyber Security, garantiza que los primeros puestos del ranking sean genuinamente los prospectos más afines al perfil de cliente exitoso de esa BU.

7.1.2.2 Modelo de cross-selling e impacto en el crecimiento comercial

El modelo de cross-selling merece una reflexión específica en las conclusiones porque responde a una lógica comercial diferente y complementaria a la prospección externa: no se trata de ampliar la base de clientes, sino de profundizar en ella. Para la compañía, donde los ciclos de venta son largos, los esfuerzos de preventa son intensivos y la confianza institucional se construye de forma progresiva, el cross-selling representa la palanca de crecimiento de menor fricción y mayor

velocidad de cierre. Un cliente que ya tiene relación comercial con la empresa ha superado la barrera de entrada más difícil; activar una BU adicional en esa cuenta es cualitativamente más eficiente que iniciar un ciclo de ventas desde cero con un prospecto externo.

El análisis reveló que los clientes actuales de la empresa consumen en promedio entre 2 y 3 BUs de las 6 disponibles en el portafolio, lo que significa que entre 3 y 4 líneas representan oportunidades latentes en cada cuenta. Las palancas más fuertes de cross-selling, considerando AUC y volumen muestral conjuntamente, son Digital Workplace (AUC = 0.970), Cyber Security (0.926) y Digital Business / Aplicaciones (0.893). En estas tres BUs el modelo Random Forest discrimina con precisión qué clientes de la base instalada tienen perfil de adoptador, apoyado en variables como la diversidad del portafolio consumido ($n_lineas_producto$), la antigüedad de la relación ($years_present$) y el ACV de otras BUs como predictores robustos de co-adopción. El heatmap de afinidad multi-BU (Figura 3) confirma visualmente que las columnas de Digital Business y Workplace presentan alta intensidad generalizada entre los 20 clientes con mayor score agregado, y que existe una correlación positiva entre la afinidad por estas BUs, lo que sugiere que comparten un perfil orgánico de cliente que puede abordarse con una propuesta comercial integrada. El caso de Soluciones SAP merece una lectura específica: aunque concentra alto interés estratégico para la compañía y aparece de forma destacada en el ranking de scores de similitud (Tabla 16), su modelo supervisado de cross-selling alcanza apenas un AUC de 0.625 con Random Forest y 0.563 con XGBoost. Esto significa que el comportamiento histórico transaccional de los clientes no anticipa de forma confiable quién adoptará SAP, y que la priorización de estas oportunidades debe sustentarse principalmente en el componente de similitud coseno y en criterios cualitativos del equipo comercial, no en el score supervisado. Esta observación es coherente con la naturaleza del consumo de SAP en la filial, asociado a contratos corporativos puntuales de migración o soporte empresarial cuya decisión de compra responde a factores externos al patrón de consumo de otras BUs. Esta distribución orienta directamente las prioridades de las mesas de preventa: no todas las cuentas representan el mismo potencial para todas las BUs, y el modelo permite jerarquizar esas conversaciones con criterio estadístico en lugar de intuición individual.

Desde una perspectiva de impacto en los objetivos comerciales, el cross-selling es estratégicamente crucial por tres razones interrelacionadas. Primera, contribuye al cumplimiento del ACV anual con menor costo de adquisición: cada contrato adicional en una cuenta existente aprovecha la infraestructura comercial y técnica ya instalada. Segunda, aumenta la profundidad de la relación con el cliente, incrementando el valor de vida de la cuenta (LTV) y reduciendo el riesgo de churn, porque una empresa que consume múltiples BUs tiene mayor dependencia operativa de la compañía y mayor costo de cambio. Tercera, genera una base de conocimiento más rica: cada expansión de portafolio en un cliente existente aporta nuevos datos al CRM que retroalimentan los modelos en la siguiente iteración, creando un ciclo virtuoso entre inteligencia comercial y resultados de negocio.

Más allá del impacto inmediato, el modelo de cross-selling cumple una función estratégica de largo plazo: convierte el conocimiento tácito de los ejecutivos de cuenta en un proceso sistemático, reproducible y auditable. Eso es exactamente lo que una organización necesita para

escalar su inteligencia comercial: no depender de la memoria individual de cada ejecutivo, sino de un modelo que aprende de todos los clientes simultáneamente y actualiza sus recomendaciones a medida que el CRM crece. En ese sentido, el modelo presentado en este trabajo no es un producto terminado, sino el primer ciclo de un sistema de inteligencia comercial continua que la empresa puede mantener, actualizar y enriquecer con cada nuevo dato disponible.

Finalmente, el cross-selling actúa como un acelerador directo del cumplimiento de los objetivos corporativos de crecimiento de la base de clientes. Si bien el indicador tradicional de «crecimiento de clientes» se mide en términos de nuevas cuentas incorporadas, existe una dimensión igualmente estratégica que con frecuencia se subestima: el crecimiento dentro de los clientes existentes, es decir, la expansión del número de líneas del portafolio activas por cuenta. Un cliente que consume una sola BU es una relación comercial puntual; un cliente que consume cuatro BUs es un socio estratégico. El modelo de cross-selling apunta precisamente a esta transformación: no solo ampliar cuántos clientes tiene la empresa, sino cuánto de cada cliente se está aprovechando. En la práctica, esto contribuye al cumplimiento de los objetivos comerciales de dos formas simultáneas: incrementa el ACV de la cartera activa sin requerir nuevas incorporaciones, y fortalece los indicadores de retención y penetración de portafolio que son parte central de la estrategia de crecimiento sostenible de la organización.

En términos de desempeño técnico, el modelo de cross-selling demostró una capacidad discriminativa heterogénea entre las BUs evaluadas, lo que constituye en sí mismo un hallazgo relevante. Los valores de AUC-ROC obtenidos por Random Forest oscilaron entre 0.625 (Soluciones SAP) y 1.000 (Cloud Pública), con resultados sólidos y estadísticamente confiables en Digital Workplace (0.970), Cyber Security (0.926) y Digital Business (0.893). El valor de Cloud Pública (1.000) debe interpretarse con cautela por el reducido número de positivos en prueba (≈ 14), pero el alto desempeño en las tres BUs intermedias —con tamaños muestrales suficientes— confirma que el modelo es genuinamente predictivo cuando los patrones de co-adopción están bien representados en los datos. En el extremo opuesto, Soluciones SAP (RF=0.625, XGBoost=0.563) y Private Cloud (RF=0.750, XGBoost=0.719) presentan desempeños más modestos: en SAP, el reducido número de consumidores históricos (13 clientes) y la naturaleza de proyecto puntual de la decisión de compra limitan estructuralmente la capacidad predictiva del modelo; en Private Cloud, el desempeño moderado refleja la heterogeneidad interna de una BU que agrupa clientes muy diversos en tamaño y sector. Esta variabilidad en el desempeño por BU refuerza la decisión metodológica de entrenar modelos independientes en lugar de un único clasificador global, ya que cada línea de negocio tiene una estructura de datos y un comportamiento de consumo propio que un modelo unificado no podría capturar con la misma precisión.

Vale destacar que el modelo de cross-selling se implementó con Random Forest como algoritmo principal en las seis BUs y con XGBoost como contraste únicamente en aquellas BUs donde el reducido tamaño muestral o el desbalance de clases hacían razonable explorar una alternativa basada en boosting (Cloud Pública, Private Cloud y Soluciones SAP). En ninguna de estas tres BUs XGBoost superó a Random Forest, por lo que se confirma a Random Forest como modelo

seleccionado para la generación de scores en todo el portafolio. La Regresión Logística no se incluyó en la comparación de Fase 3 porque ya se había descartado en la Fase 2 por riesgo de data leakage, al haber demostrado una alta sensibilidad a la inclusión de variables transaccionales directamente asociadas a la BU objetivo. Esto explica los valores N/D que aparecen en la Tabla 28 para Regresión Logística en todas las BUs y para XGBoost en aquellas donde el modelo no se ejecutó. Esta decisión metodológica fue deliberada: dado que el cross-selling opera sobre solo 167 clientes del CRM (con entre 13 y 104 consumidores por BU), Random Forest ofrece el mejor balance entre capacidad discriminativa, interpretabilidad de variables y estabilidad en muestras pequeñas, mientras que XGBoost se utilizó selectivamente como mecanismo de validación en las BUs de mayor riesgo metodológico. Un trabajo futuro recomendado es repetir la comparación multimodelo extendida —incluyendo Regresión Logística sin variables fugadas, LDA y Gradient Boosting— cuando la base de clientes del CRM supere los 300 registros, umbral a partir del cual los beneficios de la comparación se vuelven estadísticamente significativos.

7.1.2.3 Modelo supervisado para prospectos externos

El modelo supervisado cumplió su objetivo de aprendizaje: entrenado sobre los patrones históricos del CRM, fue capaz de asignar probabilidades de adopción por BU a los 572 prospectos externos. La capacidad discriminativa varía significativamente entre BUs, lo que en sí mismo es un hallazgo estratégico de valor. Private Cloud obtuvo el mayor rango supervisado (0.062–0.570), seguido de Digital Workplace (0.003–0.282), mientras que Soluciones SAP presentó el rango más estrecho (0.008–0.023), reflejo de que las variables firmográficas disponibles tienen escaso poder para identificar adoptadores de ese ecosistema por sí solas.

El score híbrido —que combina la similitud coseno con el mejor clasificador supervisado en proporción 50/50— resuelve esta limitación de manera robusta: en las BUs donde el componente supervisado discrimina poco (SAP, Cloud Pública), el coseno aporta la mayor parte de la señal; en las BUs donde el supervisado es más sólido (Private Cloud, Workplace), ambos componentes se complementan con mayor equilibrio. El resultado son rankings de 572 prospectos por BU que el equipo comercial puede utilizar directamente como herramienta de priorización, trabajando sobre el cuartil superior del Score_Total_Final como lista activa de primera ronda.

Alcance y validez de los resultados de prospección

Es necesario delimitar con precisión el alcance de estos resultados. El ranking de los 572 prospectos externos constituye una priorización de naturaleza predictiva e hipotética, no una validación confirmada: al tratarse de empresas que aún no son clientes, no se dispuso de etiquetas reales de conversión (ground truth) contra las cuales medir la exactitud del score híbrido. Las métricas de desempeño reportadas (AUC-ROC) validan la capacidad discriminativa de los modelos sobre la base histórica de clientes, de la cual se transfiere el aprendizaje, pero no miden la tasa efectiva de conversión de los prospectos priorizados. Por ello, la validación empírica de esta priorización solo será posible en una fase posterior, contrastando el ranking propuesto contra los resultados comerciales reales obtenidos por las mesas de preventa. Esta limitación no invalida el modelo como herramienta de jerarquización basada en evidencia frente a la selección intuitiva, pero define su uso correcto: un insumo de priorización de primera ronda sujeto a

retroalimentación, y no una predicción cerrada de conversión.

La comparación de tres algoritmos supervisados (Regresión Logística, Random Forest y XGBoost) confirmó que la elección del modelo óptimo es específica por BU y no generalizable: el mejor algoritmo para una línea no lo es necesariamente para otra, lo que valida la decisión metodológica de entrenar modelos independientes. Esta especificidad también implica que futuras actualizaciones del modelo deben respetar esta independencia y no asumir que un algoritmo ganador en un ciclo seguirá siendo el óptimo en el siguiente, dado que la base de clientes y el mercado evolucionan.

7.1.3 EVALUACIÓN INTEGRAL DE LOS MODELOS (OBJETIVO 3)

El proceso de evaluación de los modelos —descrito en detalle en el capítulo de desarrollo— deja tres lecciones transversales que cierran el balance del proyecto: la necesidad de seleccionar el algoritmo de forma específica por BU, la complementariedad entre los enfoques no supervisado y supervisado dentro del score híbrido, y la utilidad del LDA como herramienta de diagnóstico previo al modelado de producción.

El LDA confirmó como diagnóstico previo que el perfil firmográfico de una empresa permite anticipar con buena precisión la adopción de Cloud Pública y Digital Business, pero es insuficiente para SAP, Cyber Security y Private Cloud, donde la decisión de compra depende de interacciones no lineales entre variables. Esta gradación de separabilidad —reportada en detalle en la Tabla 19 del capítulo de evaluación— orienta la lectura de los rankings finales: los scores de Cloud Pública y Digital son robustos y estables entre algoritmos, mientras que los de SAP y Cyber deben interpretarse en combinación con el componente de similitud coseno del modelo híbrido.

En el eje supervisado, la selección del mejor modelo resultó ser específica por BU: XGBoost lideró en SAP (AUC = 0.928) y Cyber Security (0.593), Random Forest fue superior en Digital Workplace (0.750) y Private Cloud (0.629), y Regresión Logística obtuvo el mejor desempeño en Cloud Pública (0.786). Esta especificidad valida la decisión de entrenar modelos independientes por BU y confirma que no existe un algoritmo universalmente superior para este tipo de problema. En el eje no supervisado, la calidad del clustering varía entre BUs, estableciéndose distintos niveles de confianza en los centroides utilizados como referencia para la similitud coseno: Cloud Pública y Cyber Security ofrecen el mayor respaldo geométrico, mientras que SAP y Digital Business requieren mayor dependencia del componente supervisado del score híbrido.

Es importante recalcar que este proyecto hace parte de la categoría de proyecto aplicado, no en la de investigación básica, y conviene explicitar qué tipo de aporte ofrece y dónde reside su valor, que es en la construcción de una arquitectura analítica completa que conecta un problema de negocio real con una herramienta operacional reproducible, decisión por decisión.

La complejidad del trabajo no está en las técnicas individuales sino en su integración. Combinar clustering no supervisado por unidad de negocio con clasificación supervisada que utiliza las asignaciones de cluster como variables predictoras, y articular ambos componentes con un modelo de similitud coseno mediante un score híbrido aplicable a una base externa de prospectos

sin historial transaccional, constituye una secuencia metodológica no trivial. A esto se suman decisiones técnicas explícitas que el documento documenta y justifica: el control de data leakage mediante la función `variables_sin_leakage`, la partición estratificada con semilla fija para todas las fases, la calibración del número de clusters por BU usando tres métricas internas en lugar de una sola, y el reconocimiento abierto de limitaciones como la subrepresentación sectorial en SAP que distorsiona el ranking y que el propio análisis identifica y caracteriza.

El valor del trabajo se materializa en tres planos simultáneos. En el plano operativo, los rankings por BU y el ranking global agregado son insumos directos para los equipos de preventa y comercial de la filial, sin requerir interpretación adicional ni intervención del equipo de datos. En el plano metodológico, la arquitectura es replicable: el mismo flujo puede ejecutarse sobre los CRM de las demás filiales latinoamericanas con adaptaciones locales, lo que convierte el modelo de Colombia en una prueba de concepto regional. En el plano organizacional, el proyecto sienta las bases de una cultura de gestión basada en datos al sustituir la priorización por intuición individual del ejecutivo de cuenta por un criterio estadístico reproducible y auditable. Estos tres planos son lo que distingue a un proyecto aplicado bien ejecutado de un ejercicio académico aislado: la solución existe, funciona, y resuelve un problema que la organización tenía sin atender.

7.2 TRABAJOS FUTUROS

7.2.1 Expansión del modelo a escala latinoamericana

El primer trabajo futuro de mayor impacto estratégico es la replicación del modelo en las demás operaciones de la compañía en Latinoamérica. Este proceso requeriría, como paso previo e indispensable, la aprobación del CEO de la compañía, de los Country Managers de cada país donde se implemente y de la Dirección de Preventa LATAM, quienes aportarían el conocimiento del negocio necesario para validar que los perfiles y rankings generados son coherentes con la realidad comercial de cada país.

Es importante destacar que la expansión no puede limitarse a aplicar el modelo de Colombia sobre datos de otros mercados. Cada país opera bajo dinámicas comerciales, económicas y sectoriales propias: la estructura de industrias predominantes, el tamaño típico de las empresas cliente, los sectores con mayor madurez tecnológica y los ciclos de venta varían de forma significativa entre, por ejemplo, Argentina, México, Chile y Perú. Por esta razón, el ejercicio metodológico completo —caracterización del CRM local, enriquecimiento firmográfico, segmentación no supervisada por BU, entrenamiento de los modelos supervisados y construcción de la base de prospectos externos— tendría que ejecutarse de forma independiente para cada país. El modelo de Colombia sirve como referencia metodológica y prueba de concepto, pero no como modelo transferible directamente a otros mercados sin el proceso de adaptación correspondiente.

El resultado de esta expansión sería una familia de modelos nacionales que, manteniendo la misma arquitectura técnica, capturan las particularidades de cada mercado y permiten a la dirección comercial regional comparar oportunidades de crecimiento entre países con criterios homogéneos y comparables.

7.2.2 Enriquecimiento del modelo con variables de contexto tecnológico

Una de las limitaciones identificadas en el presente trabajo es que los modelos supervisados operan exclusivamente sobre variables firmográficas generales. Si bien estas variables capturan el perfil organizacional de un prospecto, no recogen información sobre su situación tecnológica actual, que en el contexto de servicios de tecnología B2B es un predictor mucho más preciso de la propensión a adoptar una BU específica.

Un trabajo futuro de alto valor sería enriquecer la base de prospectos con variables de contexto tecnológico, que el equipo comercial y de preventa puede levantar progresivamente a través de sus interacciones con el mercado. Entre las variables más relevantes se encuentran: si la empresa ya opera sobre SAP y en qué versión (lo que determina directamente la probabilidad de requerir servicios de migración a S/4HANA o soporte funcional), en qué nube opera actualmente (AWS, Azure, GCP, on-premise o modelo híbrido), dónde está ubicada su infraestructura tecnológica (data center propio, colocation, nube privada gestionada), el nivel de madurez de sus prácticas de ciberseguridad, y si ya tiene contratos activos con competidores directos.

Con este enriquecimiento, el modelo pasaría de identificar empresas que tienen el perfil organizacional de un cliente potencial, a identificar empresas que tienen tanto el perfil organizacional como la situación tecnológica que hace que una BU específica sea relevante para ellas en este momento. Esto incrementaría significativamente la precisión de los rankings y reduciría el tiempo que el equipo comercial invierte en calificar manualmente a los prospectos mejor posicionados.

7.2.3 Tablero de inteligencia comercial integrado con Salesforce

El tercer trabajo futuro es la materialización de los resultados del modelo en un tablero de visualización e inferencia integrado directamente con Salesforce, que permita a los equipos comerciales y gerenciales consumir la inteligencia generada sin necesidad de interactuar con archivos de datos de forma manual.

La integración con Salesforce es el elemento central de este componente. Al conectar el modelo con el CRM, los scores de cross-selling y de afinidad por BU se actualizarían automáticamente cada vez que un nuevo registro de oportunidad, cliente o actividad comercial sea ingresado en la plataforma, sin requerir intervención manual del equipo de datos. Esto garantiza que los rankings reflejen siempre el estado más reciente de la cartera y elimina la brecha entre el momento en que se generan los datos y el momento en que se consumen como inteligencia comercial.

- **Vista de cartera actual:** muestra el estado de cada cliente existente, sus BUs activas, su score de cross-selling por línea y su posición en el ranking de oportunidades de expansión, permitiendo identificar de un vistazo cuáles cuentas tienen mayor potencial de crecimiento inmediato.
- **Evaluador de prospectos externos en tiempo real:** un ejecutivo comercial o analista de preventa podría ingresar las variables firmográficas de una empresa nueva y el modelo calcularía

automáticamente en qué BU esa empresa tendría mejor desempeño como prospecto, generando un score de afinidad por línea y una recomendación de BU prioritaria.

7.2.4 Actualización periódica del modelo

Todo modelo predictivo pierde precisión con el tiempo a medida que el mercado evoluciona, la cartera de clientes cambia y nuevas BUs o subproductos son incorporados al portafolio. Por esta razón, un trabajo futuro necesario es establecer un protocolo de actualización periódica del modelo, con una frecuencia recomendada de al menos una vez por año o cada vez que se incorporen más de 15 clientes nuevos al CRM, lo que ocurra primero.

Este protocolo incluiría el re-entrenamiento de los modelos supervisados con los datos actualizados del CRM, la revisión de los clusters por BU para verificar si los perfiles de cliente han evolucionado, la actualización de la base de prospectos externos con nuevas empresas del mercado, y la validación de los rankings generados con la Dirección de Preventa antes de su distribución al equipo comercial. La periodicidad de actualización no solo mantiene la precisión técnica del modelo, sino que también asegura que la herramienta continúe siendo percibida como relevante y confiable por los equipos que la utilizan en su gestión diaria.

8 REFERENCIAS BIBLIOGRÁFICAS

- [1] TIVIT, "¿Quiénes somos?," [En línea]. Disponible en: <https://latam.tivit.com/quienes-somos> [Accedido: abr. 2026]
- [2] El Tiempo, "Tivit se consolida en Colombia tras compra de Synapsis," 2015. [En línea]. Disponible en: <https://www.eltiempo.com/archivo/documento/DR-912107> [Accedido: abr. 2026]
- [3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Cambridge, MA: Elsevier, 2011.
- [4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York: Springer, 2009.
- [5] C. Chen, Y. Zhang, J. Ma and M. Cao, "Lead Conversion Prediction Using Customer Journey Data," *IEEE Access*, vol. 8, pp. 125514-125526, 2020.
- [6] P. Kotler and K. L. Keller, *Marketing Management*, 15th ed., Upper Saddle River, NJ: Pearson Education, 2016.
- [7] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, Cambridge, MA: MIT Press, 2012.
- [8] Universidad de Manizales, "Factores que inciden en el 'Churn' de una empresa de telecomunicaciones: Un análisis basado en Machine Learning," 2023. [En línea]. Disponible en: <https://ridum.umanizales.edu.co/bitstreams/64524b76-39a7-47cf-8411-e4f4e51b63df/download>
- [9] Universidad Nacional de Colombia, "Desarrollo de un modelo analítico para la segmentación de clientes en una cooperativa financiera," 2020. [En línea]. Disponible en: https://repositorio.unal.edu.co/bitstream/handle/unal/79648/98644734_2020.pdf
- [10] Universidad Nacional Abierta y a Distancia (UNAD), "Modelo de predicción de deserción estudiantil apoyado en tecnologías de data mining en un curso de primera matrícula de la Escuela ECBTI," 2021. [En línea]. Disponible en: <https://repository.unad.edu.co/handle/10596/42544>
- [11] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Math. Statist. Prob., vol. 1, pp. 281–297, 1967.
- [12] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2,

pp. 129–137, 1982.

[13] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[14] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, 2016.

[15] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer, 2002.

[16] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[17] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Commun. Statist.*, vol. 3, no. 1, pp. 1–27, 1974.

[18] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–228, 1979.

[19] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[20] D. Arunachalam and N. Kumar, “Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making,” *Expert Systems with Applications*, vol. 111, pp. 11–34, 2018. DOI: 10.1016/j.eswa.2018.03.007. [En línea]. Disponible en: <https://repository.essex.ac.uk/28436/>

[21] Y. E. Tang and M. K. Mantrala, “Incorporating direct customers' customer needs in a multi-dimensional B2B market segmentation approach,” *Industrial Marketing Management*, vol. 119, pp. 252–263, 2024. DOI: 10.1016/j.indmarman.2024.04.005

[22] L. Wang, M. Han, X. Li, N. Zhang, and H. Cheng, “Review of classification methods on unbalanced data sets,” *IEEE Access*, vol. 9, pp. 64606–64628, 2021. DOI: 10.1109/ACCESS.2021.3074243. [En línea]. Disponible en: <https://ieeexplore.ieee.org/document/9412437>

[23] R. Burke, “Hybrid recommender systems: Survey and experiments,” *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002. DOI: 10.1023/A:1021240730564

[24] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, “Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches,” *IEEE Computational Intelligence Magazine*, vol. 13, no. 4, pp. 59–76, 2018. DOI: 10.1109/MCI.2018.2866730. [En línea]. Disponible en:

<https://student.dei.uc.pt/~miriams/pdf-files/IEEE-CIM-Version.pdf>

[25] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," *Sustainability*, vol. 14, no. 12, p. 7243, 2022. DOI: 10.3390/su14127243. [En línea]. Disponible en: <https://www.mdpi.com/2071-1050/14/12/7243>

9 ANEXOS

ANEXO A. REPOSITORIO DEL PROYECTO

Con el propósito de garantizar la trazabilidad y reproducibilidad de los resultados presentados en este documento, todos los insumos y entregables del proyecto se encuentran centralizados en una carpeta única de Google Drive, accesible mediante el siguiente enlace:

Enlace al repositorio:

https://drive.google.com/drive/folders/1F578efQn_G9pRhmiXy92wrlL4qPIKWU0z?usp=sharing

El repositorio integra las tres categorías de archivos producidos a lo largo del proyecto:

Bases de datos. Incluye tanto las fuentes originales como los datasets derivados del proceso de caracterización y modelado: los reportes anuales extraídos de Salesforce (2018–2025), el archivo enriquecido vía Gemini 2.5 Pro clientes_caracterizados.xlsx, la base intermedia clientes_bda_caracterizada.csv/xlsx (Base B), el dataset analítico definitivo clientes_bda_model.csv (Base A) utilizado para clustering y cross-selling, y la base externa DATABASE.xlsx con los 572 prospectos depurados (Base C).

Código. Reúne el notebook de Google Colab desarrollado en Python para cada fase del proyecto: extracción y unificación de reportes, normalización y enriquecimiento, análisis exploratorio, segmentación K-Means global y por unidad de negocio, modelos supervisados (Regresión Logística, Random Forest, XGBoost y LDA) para cross-selling y prospectos externos, cálculo de similitud coseno y construcción del score híbrido. Este notebook está versionado con la semilla random_state = 42 para garantizar la reproducibilidad de los resultados reportados.

Outputs. Contiene los archivos de salida generados por los modelos: las segmentaciones por BU (clientes_segmentados_BU_k4.csv, clientes_segmentados_BU_k8.csv), los rankings supervisados por BU (ranking_<bu>_supervisado.xlsx), los rankings de similitud coseno (ranking_<bu>_externos_clusters.xlsx), los rankings híbridos por BU (ranking_<bu>_hibrido.xlsx) y los rankings agregados a nivel global (ranking_global_externos.xlsx, ranking_global_final_hibrido.xlsx), que constituyen los entregables operativos para los equipos comerciales y de preventa.

ANEXO B. PROMPT PARA CONSTRUCCIÓN DE LA BASE DE PROSPECTOS EXTERNOS

A continuación, se presenta el prompt utilizado con Gemini 2.5 Pro en modo Deep Research para construir la base de datos de prospectos externos. Este prompt fue diseñado para generar una base estructurada de las 500 empresas más grandes de Colombia y Panamá, con taxonomía normalizada y criterios de calidad específicos para garantizar la coherencia con las variables del CRM Salesforce.

Prompt utilizado:

Actúa como un analista senior de inteligencia de mercado y científico de datos especializado en Latinoamérica.

Objetivo: construir una base de datos estructurada y lo más completa posible de las 500 empresas más grandes de Colombia y Panamá, priorizando compañías por ingresos, activos, relevancia de mercado o tamaño organizacional.

Alcance:

- Incluir empresas de Colombia y Panamá.
- Priorizar empresas grandes, medianas-grandes y corporaciones relevantes.
- Evitar duplicados.
- Usar información pública y estimaciones razonables cuando falten datos.
- Si un dato no está disponible, usar "No disponible".

Estructura obligatoria de salida:

Genera la base en formato tabular (CSV o tabla) con EXACTAMENTE estas columnas:

Empresa | Ciudad | País | Subtipo de empresa | Tipo de empresa | Tamaño de la empresa

Presencia internacional | Sector económico principal | Modelo de negocio

Modelo normalizado | Submodelo

Reglas de normalización:

Tipo de empresa (valores permitidos):

Privada | Pública | Mixta | Multinacional | Grupo empresarial

Tamaño de la empresa (clasificación estándar):

Grande | Mediana-Grande | Multinacional

Presencia internacional:

Sí | No | Regional

Modelo normalizado (taxonomía obligatoria):

B2B | B2C | B2B2C | Plataforma | Holding | Manufactura | Servicios | Retail | Financiero

Submodelo (más específico, por ejemplo):

SaaS | Marketplace | Banca comercial | Oil & Gas upstream | Telecom operador

Retail omnicanal | Seguros | Logística | etc.

Criterios de calidad:

- No inventar empresas ficticias.
- Usar nombres oficiales.
- Mantener consistencia semántica.
- Evitar filas duplicadas.
- Generar mínimo 500 registros.
- Ordenar por relevancia estimada (de mayor a menor).

Formato de entrega: la tabla completa.

El resultado de este prompt fue una base de 600 empresas, que tras el proceso de depuración descrito (exclusión de clientes existentes en el CRM por coincidencia exacta y parcial) quedó reducida a 572 prospectos externos válidos, los cuales conforman la base sobre la que se aplicó el modelo híbrido de prospección.