Segmentación de profesionales de la salud del sector farmacéutico por Machine Learning para la optimización de frecuencia de visitas.

Santiago Reyes Zabaleta

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

Director(a)
Genaro Cortez Aguilar

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, ENERO DE 2025

TABLA DE CONTENIDO

INTRODUCCIÓN 5
1. DEFINICIÓN DEL PROBLEMA 6
1.1. PLANTEAMIENTO DEL PROBLEMA 6
1.2. FORMULACIÓN DEL PROBLEMA 8
2. OBJETIVOS DEL PROYECTO 9
2.1. OBJETIVO GENERAL 9
2.2. OBJETIVOS ESPECÍFICOS 9
3. MARCO TEÓRICO Y ANTECEDENTES 10
3.1. MARCO TEÓRICO 10
3.1.1. Segmentación de mercados 10
3.1.2. Machine Learning y Ciencia de Datos 12
3.1.3. Optimización de Recursos Comerciales 14
3.2. ANTECEDENTES 16
3.2.1. Métodos Tradicionales de Segmentación 16
3.2.2. Evolución hacia modelos basados en datos 18
3.2.3. Contexto de la industria Farmacéutica 20
3.2.4. Contexto Sanofi Colombia 22
4. ESTRUCTURACIÓN Y ANÁLISIS DE DATA 23
4.1. CUESTIONARIO EN CRM (VEEVA) 23
4.2. DESCARGA DE DATA Y ANONIMIZACIÓN DE LA INFORMACION 24
4.3. ANÁLISIS EXPLORATORIO DE LA DATA 25
4.3.1. Estadística Descriptiva 25
4.3.2. Distribución de la cantidad de Pacientes 26
4.3.3. Proporción por tipo de consulta: 27
4.3.4. Distribución del porcentaje de consulta 27
5. MODELOS DE CLUSTERIZACIÓN 31
5.1. K-MEANS 31
5.1.1. Análisis del método del Codo 32

5.1.2. Resultados K-Means 33
5.2. DBSCAN 34
5.3. CLUSTERIZACIÓN JERARQUICA 35
5.3.1. Dendrogramas 35
5.3.2. Clusterización 35
5.4. BISECTING K-MEANS: 36
5.5. GAUSSIAN MIXTURE MODEL 37
5.7.1. Regresión Logística 40
5.7.2. K-Nearest Neighbors (KNN) 41
5.7.3. Bosques Aleatorios (Random Forest) 42
5.7.4. Support Vector Machine (SVM) 42
5.7.5. Comparativa modelos 43
6. SEGMENTACIÓN FINAL – ENTREGABLE SANOFI 46
7. ESTRATEGÍA Y PRIORIZACION DE VISITA 50
7.1. SEGMENTO A (ALTO POTENCIAL) 51
7.2. SEGMENTO B (POTENCIAL MEDIO) 51
8. CONCLUSIONES Y TRABAJOS FUTUROS 53
8.1. CONCLUSIONES 53
8.2. TRABAJOS FUTUROS: 54
9. REFERENCIAS 56
10. ANEXOS 57
10.1. Carta de Autorización uso de Datos 57
10.2. Limpieza y Análisis Exploratorio (GitHub) 58
10.3. Resumen Modelos de cauterización (GitHub) 58
10.4. K-Means (GitHub) 58
10.5. DBSCAN (GitHub) 58
10.6. Clusterización Jerárquica (GitHub) 58
10.7. K-Means Bisecting (GitHub) 58
10.8. Gaussian Mixture Model (GitHub) 58
10.9. Modelos ML Supervisados (GitHub) 58
10.10. Entregable Final Sanofi (GitHub) 58

LISTA DE TABLAS E ILUSTRACIONES

Tabla 1. Distribución de panel Médico por Gerencia de Distrito y producto 6
Tabla 2. Cuestionario de perfilamiento Gaucher.....23
Tabla 3. Descripción del Dataset.....24
Tabla 4. Estadística Descriptiva 26
Tabla 5. Métricas Kmeans.....33
Tabla 6. Resumen Coeficientes de Silueta 38
Tabla 7. Comparativa modelos supervisados 44
Tabla 8. Frecuencias Trimestrales de interacción.....50

Ilustración 1. Modelo actual de Segmentación de Sanofi 22
Ilustración 2. Distribución de Cantidad de Pacientes.....26
Ilustración 3. Proporción tipo de consulta.....27
Ilustración 4. Distribución del porcentaje de consulta 27
Ilustración 5. Top 10 especialidades.....28
Ilustración 6. Distribución de pacientes por tipo de consulta 29
Ilustración 7. Distribución del porcentaje de consulta por tipo de consulta 29
Ilustración 8. Análisis del Método del Codo 32
Ilustración 9. Resultados Kmeans 33
Ilustración 10. DBSCAN.....34
Ilustración 11. Dendrograma.....35
Ilustración 12. Clusterización Jerárquica 36
Ilustración 13. Bisecting K-Means.....36
Ilustración 14. Gaussian Mixture Model.....37
Ilustración 15. Matrices de Confusión de los modelos 45
Ilustración 16. Distribución de Clusteres 47
Ilustración 17. Distribución cruzada por cluster y especialidad.....48

INTRODUCCIÓN

El presente proyecto aborda una problemática actual en Sanofi Colombia, relacionada con la segmentación del panel médico, la cual hasta ahora ha sido un proceso arbitrario y subjetivo. Esta situación ha generado la necesidad de explorar cómo optimizar dicho proceso mediante un modelo analítico y basado en datos objetivos, que permita mejorar la toma de decisiones estratégicas. La implementación de un enfoque más riguroso en la segmentación busca impactar directamente en la planeación del departamento de marketing, optimizando actividades clave como la entrega de muestras médicas, visitas personalizadas, desarrollo y uso de materiales promocionales (ayudas visuales, videos, correos electrónicos), así como la gestión de eventos, congresos y webinars.

Ante este desafío, se investiga y aplica técnicas de machine learning, ampliamente utilizadas en diversos sectores por su capacidad de segmentar poblaciones objetivo de manera eficiente y precisa. Estas herramientas han demostrado ser soluciones efectivas en la generación de modelos analíticos que no solo eliminan la subjetividad en los procesos, sino que también facilitan la personalización y priorización de estrategias promocionales.

El proyecto se enfoca en desarrollar un modelo de segmentación que cumpla con estándares objetivos y proporcione resultados confiables, los cuales serán utilizados por la empresa para mejorar la asignación de recursos y maximizar el impacto promocional. Este enfoque no solo ayudará a Sanofi Colombia a optimizar su relación con los profesionales de la salud, sino también a sentar las bases para la implementación de estrategias basadas en datos en el futuro, posicionando a la organización como líder en la gestión estratégica del mercado de la salud en el país.

1. DEFINICIÓN DEL PROBLEMA

A continuación, se planteará el problema de segmentación actual de la empresa Sanofi.

1.1. PLANTEAMIENTO DEL PROBLEMA

Actualmente, la empresa emplea un modelo de segmentación basado fundamentalmente en criterios subjetivos. Este modelo otorga a los representantes de ventas la capacidad de ajustar, de manera discrecional, la clasificación de los profesionales de la salud que visitan, siempre que cuenten con la aprobación del Gerente de Distrito correspondiente. En este proceso, se asume que los representantes de ventas actúan con objetividad al realizar las segmentaciones y que los Gerentes de Distrito poseen un conocimiento integral y detallado de los paneles médicos en sus territorios asignados. Sin embargo, esta metodología presenta varios desafíos que impactan la eficacia de las operaciones comerciales.

El problema principal radica en la subjetividad del proceso de segmentación, que depende en gran medida del juicio personal de los representantes y de la validación de los Gerentes de Distrito. Este enfoque no solo introduce un alto nivel de variabilidad en las decisiones, sino que también se ve limitado por la incapacidad práctica de los Gerentes de Distrito para tener un conocimiento exhaustivo de todos los médicos que forman parte de sus paneles. Dado el volumen de profesionales involucrados, el proceso carece de consistencia y objetividad, lo que puede comprometer tanto la asignación de recursos como la efectividad de las estrategias comerciales.

Sanofi Colombia, en particular, enfrenta este desafío en un contexto operativo compuesto por 40 representantes de ventas y 4 Gerencias de Distrito que supervisan 8 líneas de negocio diferentes. Estas líneas abarcan áreas terapéuticas diversas y específicas, incluyendo asma, pólipos nasales, dermatitis atópica, púrpura trombocitopénica trombótica adquirida (PTTa), esclerosis múltiple, y enfermedades raras como Gaucher, Fabry y Pompe. Cada una de estas líneas está asociada a un producto específico, y el número de médicos en los paneles varía significativamente según la especialidad y el producto, como se detalla en la Tabla 1.

Tabla 1. Distribución de panel Médico por Gerencia de Distrito y producto

Gerente de Distrito	Product	# Médicos Approx
Gaucher	Cerezyme	500
Pompe	Myozyme	500
Fabry	Fabrazyme	600
PTTa	Cablivi	500
Asma	Dupixent	400
Polipos nasales	Dupixent	500

Dermatitis atópica	Dupixent	900
Esclerosis múltiple	Aubagio/Lemtrada	80
Total		3.980

El panel médico, compuesto por un total de 3,980 profesionales, representa un desafío considerable para la gestión eficiente de los recursos de ventas. Las diferentes líneas de negocio, como dermatitis atópica y enfermedades raras, presentan variaciones significativas en el número de médicos que deben ser atendidos, lo que demanda estrategias específicas y adaptadas para garantizar una cobertura adecuada. Por ejemplo, la línea de dermatitis atópica, con 900 médicos asignados, requiere un esfuerzo considerable para asegurar que los representantes puedan atender de manera efectiva a los profesionales clave. En contraste, la línea de esclerosis múltiple, con solo 80 médicos, demanda un enfoque más especializado y menos exigente en términos de alcance territorial.

El proceso actual requiere segmentar más de 3,980 combinaciones, cada una de las cuales debe ser revisada y aprobada por los Gerentes de Distrito. En casos extremos, como el de la línea de dermatitis atópica, un solo gerente tendría que aprobar cerca de 900 combinaciones, lo que convierte este procedimiento en una tarea altamente compleja, tediosa e inviable. Es prácticamente imposible que un solo Gerente de Distrito tenga un conocimiento exhaustivo de todos los médicos dentro de su territorio.

Esta limitación hace que el proceso de segmentación termine siendo subjetivo, ya que los representantes suelen proponer ajustes en la segmentación, y los Gerentes de Distrito, ante la imposibilidad de validar exhaustivamente cada propuesta, terminan aprobando la totalidad de las sugerencias presentadas por los representantes.

En respuesta a esta problemática, se ha comenzado a plantear un nuevo modelo de segmentación basado en criterios más objetivos. Este nuevo enfoque incluye la incorporación de preguntas estructuradas que evalúen el conocimiento de los representantes sobre cada médico, permitiendo así construir una base de datos más sólida y fundamentada. Con esta herramienta, se busca reducir la subjetividad en el proceso, optimizar la asignación de recursos y garantizar una segmentación más estratégica y eficiente.

1.2. FORMULACIÓN DEL PROBLEMA

De acuerdo con el anterior numeral, el problema podría formularse de la siguiente manera:

- Subjetividad en la Segmentación: La falta de un modelo de segmentación estructurado introduce riesgos significativos, como inconsistencias en la clasificación de los médicos y la posibilidad de priorizar incorrectamente a ciertos profesionales.

- Sobrecarga de Responsabilidades en los Gerentes de Distrito: Se espera que los Gerentes de Distrito tengan un conocimiento detallado de todos los médicos en sus territorios. Sin embargo, dada la magnitud del panel médico, esta expectativa resulta poco realista y genera un cuello de botella en el proceso de toma de decisiones.

- Desigualdad en la Distribución de Recursos: La variabilidad en el tamaño de los paneles médicos por la línea de negocio sugiere que algunas áreas terapéuticas podrían no estar recibiendo la atención necesaria para maximizar su potencial de mercado.

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Desarrollar un modelo de segmentación para el panel médico de Cerezyme, una marca focal de la empresa, utilizando técnicas de Machine Learning, con el propósito de generar resultados completamente objetivos que respalden y optimicen la toma de decisiones estratégicas.

2.2. OBJETIVOS ESPECÍFICOS

OE1: Estructurar la base de datos del panel médico, asegurando la calidad y relevancia de las variables clave necesarias para el desarrollo del modelo de segmentación.

OE2: Desarrollar e implementar múltiples modelos de segmentación basados en técnicas de Machine Learning, que clasifiquen de manera objetiva a los profesionales de la salud según su relevancia estratégica y potencial de prescripción, evaluando y comparando su desempeño para seleccionar el modelo más preciso, eficiente y aplicable.

OE3: Entregar un panel médico segmentado en formato visual y tabular, optimizado para su interpretación y con formato listo para ser cargado en el CRM de la empresa.

OE4: Optimizar la frecuencia y priorización de visitas a los médicos, utilizando los resultados generados por el modelo para maximizar la eficiencia y el impacto comercial.

3. MARCO TEÓRICO Y ANTECEDENTES

A continuación, se muestra el marco teórico investigado para el presente proyecto, adicionalmente, se explica los antecedentes globales e internos con que Sanofi Colombia ha intentado atacar este problema.

3.1. MARCO TEÓRICO

3.1.1. Segmentación de mercados

La segmentación de mercados es un proceso esencial en marketing estratégico que permite dividir un mercado amplio y heterogéneo en grupos más pequeños y homogéneos basados en características compartidas. Según Kotler y Keller, esta estrategia identifica oportunidades de mercado, ajusta productos a las necesidades específicas de cada segmento y optimiza el uso de los recursos comerciales, maximizando el impacto de las campañas [1].

En la industria farmacéutica, la segmentación adquiere particular relevancia debido a la diversidad de especialidades médicas, patrones de prescripción y dinámicas regulatorias. Cada médico, dependiendo de su especialidad y práctica, tiene necesidades y comportamientos específicos que influyen en su decisión de prescripción. Por ejemplo, un oncólogo que trata cáncer metastásico necesita información detallada sobre terapias avanzadas y basadas en biomarcadores, mientras que un médico generalista podría priorizar medicamentos de uso generalizado, como antihipertensivos. Estas diferencias subrayan la necesidad de una segmentación que permita a las empresas farmacéuticas adaptar sus estrategias comerciales a las particularidades de cada grupo [2].

La segmentación de mercados también es crucial en un entorno donde los recursos son limitados. Las empresas farmacéuticas suelen enfrentarse a restricciones presupuestarias y limitaciones en el tiempo disponible para las visitas de los representantes médicos. La segmentación ayuda a priorizar médicos clave, asegurando que los recursos se concentren en aquellos que pueden generar el mayor impacto comercial. Además, los ciclos de vida de los productos, como la introducción, el crecimiento, la madurez y el declive, influyen directamente en la forma en que se deben segmentar los mercados. En la fase de introducción, por ejemplo, los médicos innovadores y "adoptadores tempranos" son un objetivo clave, mientras que en etapas posteriores se requiere una estrategia más amplia para maximizar la adopción generalizada [3].

Las normativas y restricciones legales también juegan un papel importante en la segmentación farmacéutica. Las regulaciones internacionales limitan las interacciones entre las empresas farmacéuticas y los profesionales de la salud, lo que aumenta la importancia de priorizar a los médicos más relevantes dentro de las restricciones legales. Por ejemplo, los códigos de conducta

como el de la EFPIA (European Federation of Pharmaceutical Industries and Associations) establecen lineamientos estrictos sobre la promoción de medicamentos, lo que obliga a las empresas a ser más estratégicas en su enfoque [4].

Históricamente, la segmentación de mercados en la industria farmacéutica se ha basado en enfoques tradicionales, como criterios demográficos, geográficos, psicográficos y conductuales. Los criterios demográficos incluyen variables como la especialidad médica, la edad y el nivel de experiencia del médico. Los aspectos geográficos clasifican a los médicos según la ubicación de sus consultorios, clínicas u hospitales, permitiendo identificar diferencias entre regiones urbanas y rurales. Los elementos psicográficos se centran en las actitudes y valores de los médicos hacia nuevas terapias, mientras que los conductuales analizan patrones históricos de prescripción y participación en campañas promocionales [5].

A pesar de su utilidad, estos enfoques presentan importantes limitaciones. La subjetividad es una de las principales debilidades de los métodos tradicionales, ya que muchas decisiones de segmentación dependen de la experiencia personal de los representantes médicos, lo que puede introducir sesgos significativos. Además, estos enfoques no son fácilmente escalables en mercados grandes y diversificados, y a menudo son reactivos, basándose únicamente en datos históricos sin considerar cambios futuros en el mercado [6].

La evolución hacia una segmentación moderna basada en datos ha permitido superar muchas de estas limitaciones. Con el auge del análisis predictivo y el Machine Learning, las empresas farmacéuticas ahora pueden utilizar herramientas avanzadas para segmentar a los médicos de manera más precisa y dinámica. Los algoritmos predictivos permiten clasificar a los médicos según su probabilidad de prescripción o adopción de nuevas terapias. Además, las técnicas avanzadas de clustering, como K-Means, DBSCAN y Modelos de Mezclas Gaussianas (GMM), identifican patrones complejos en los datos que los métodos tradicionales no pueden detectar [7].

El uso de sistemas de gestión de relaciones con el cliente (CRM), como Salesforce o Veeva, ha revolucionado la segmentación farmacéutica. Estas plataformas integran datos de múltiples fuentes y aplican análisis avanzados para identificar médicos clave, priorizar segmentos y personalizar estrategias comerciales. A diferencia de los enfoques tradicionales, los métodos basados en datos eliminan la subjetividad, mejoran la escalabilidad y permiten una actualización dinámica de los modelos para reflejar cambios en el comportamiento del mercado [8].

La segmentación de mercados también tiene aplicaciones prácticas significativas en la industria farmacéutica. Permite priorizar a los médicos más influyentes según su impacto potencial en términos de volumen de prescripción y relevancia estratégica. Por ejemplo, un médico generalista puede ser más relevante para un medicamento ampliamente utilizado, mientras que un especialista en reumatología será clave para una terapia avanzada contra la artritis reumatoide. Asimismo, la segmentación facilita el diseño de estrategias adaptadas a cada segmento, como

proporcionar materiales técnicos detallados a especialistas y estudios de caso más generales a médicos generalistas [9].

Finalmente, la segmentación moderna permite a las empresas farmacéuticas optimizar la cobertura territorial, utilizando herramientas de análisis para diseñar rutas de visitas más eficientes. Este enfoque no solo reduce costos operativos, sino que también maximiza el impacto comercial al garantizar que los representantes médicos enfoquen sus esfuerzos en los médicos adecuados. En el contexto de la introducción de nuevas terapias, la segmentación dirigida a médicos "adoptadores tempranos" acelera significativamente la penetración de mercado, mientras que en etapas posteriores facilita una adopción más amplia [10].

De esta forma, la segmentación de mercados es un componente esencial en la estrategia de las empresas farmacéuticas. A medida que las herramientas avanzadas de análisis y Machine Learning se integran en los procesos comerciales, la segmentación no solo mejora la eficiencia y la personalización de las estrategias, sino que también se convierte en un diferenciador competitivo clave en un mercado dinámico y altamente regulado.

3.1.2. Machine Learning y Ciencia de Datos

El Machine Learning (ML) y la Ciencia de Datos han revolucionado la forma en que las empresas analizan y extraen valor de los datos en diversos sectores, incluida la industria farmacéutica. Según Bishop, el ML es una subdisciplina de la inteligencia artificial que utiliza algoritmos y modelos matemáticos para identificar patrones y realizar predicciones, mientras que la Ciencia de Datos integra estadística, programación y conocimiento específico del dominio para procesar grandes volúmenes de datos y convertirlos en información útil y accionable [4][5].

En la industria farmacéutica, estas tecnologías han adquirido una importancia estratégica debido al volumen, la complejidad y la diversidad de datos disponibles. La información generada por patrones de prescripción, datos de ventas, registros clínicos y resultados de campañas promocionales ofrece un vasto campo de análisis. Estas herramientas permiten optimizar la segmentación de mercados, personalizar estrategias comerciales, prever comportamientos médicos y aumentar la eficiencia en la asignación de recursos. La integración de estas tecnologías en los procesos comerciales no solo mejora la toma de decisiones, sino que también ofrece una ventaja competitiva en un mercado cada vez más complejo [6][7].

El Machine Learning se clasifica en tres categorías principales: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. El aprendizaje supervisado utiliza datos etiquetados para entrenar modelos capaces de predecir o clasificar salidas basadas en nuevas entradas. Este enfoque es particularmente útil en la industria farmacéutica para tareas como clasificar médicos en segmentos de "alto", "medio" o "bajo potencial", o para predecir la adopción

de nuevas terapias. Algunos de los algoritmos más utilizados incluyen la regresión logística, que permite clasificaciones binarias o multiclase; los árboles de decisión, que dividen los datos en subconjuntos según las variables más discriminantes; y los métodos avanzados como Gradient Boosting Machines (GBM), que corrigen errores secuencialmente para mejorar la precisión en problemas complejos [8][9][10].

Por otro lado, el aprendizaje no supervisado es una herramienta poderosa para descubrir patrones ocultos en datos no etiquetados, como la segmentación de médicos basada en similitudes en su comportamiento de prescripción. Técnicas como K-Means, que agrupa los datos en clusters predefinidos, y el Modelo de Mezclas Gaussianas (GMM), que asigna probabilidades de pertenencia a cada grupo, son ampliamente utilizadas en esta área. Además, métodos como DBSCAN son útiles para identificar médicos con patrones de prescripción únicos o atípicos. Estas técnicas permiten a las empresas identificar grupos homogéneos de médicos y personalizar estrategias comerciales para cada segmento [11][12][13].

El aprendizaje por refuerzo, aunque menos común en la industria farmacéutica, tiene aplicaciones interesantes, como la optimización de rutas de visita de los representantes médicos. Este enfoque se basa en un modelo que aprende mediante prueba y error, ajustando sus decisiones en función de recompensas o penalizaciones recibidas. Es particularmente útil para diseñar estrategias adaptativas en mercados dinámicos o cuando se necesita tomar decisiones en tiempo real [14].

La Ciencia de Datos complementa al ML al proporcionar un marco integral para el análisis, la interpretación y la visualización de datos. Antes de aplicar cualquier modelo, es esencial realizar una limpieza y preparación de los datos, corrigiendo errores, eliminando valores atípicos y normalizando las variables para garantizar su calidad. Posteriormente, el análisis exploratorio permite identificar patrones generales y relaciones entre las variables, facilitando la selección de características relevantes para el modelo. Herramientas como Tableau y Power BI son fundamentales para la visualización de los resultados, ya que representan gráficamente patrones complejos y ayudan a los equipos comerciales a identificar oportunidades clave de manera eficiente [15].

En el contexto farmacéutico, la implementación de modelos predictivos y sistemas de gestión como CRMs (e.g., Salesforce, Veeva) ha permitido automatizar decisiones comerciales y personalizar las interacciones con los médicos. Por ejemplo, un modelo puede recomendar a un representante médico cuáles son los profesionales prioritarios para visitar, basándose en su potencial de prescripción y su proximidad territorial. Esta integración de tecnología no solo mejora la eficiencia operativa, sino que también optimiza la experiencia de los médicos al recibir información relevante y adaptada a sus necesidades específicas [16].

El impacto del Machine Learning y la Ciencia de Datos en la industria farmacéutica es significativo. Estas tecnologías han permitido a las empresas identificar médicos clave para sus estrategias, optimizar recursos y predecir tendencias en la adopción de nuevas terapias. Además, su capacidad para procesar grandes volúmenes de datos de manera objetiva elimina el sesgo humano y aumenta la precisión en la toma de decisiones. Sin embargo, también enfrentan desafíos importantes, como la necesidad de contar con datos de alta calidad, la complejidad computacional de algunos modelos y las dificultades para interpretar los resultados de algoritmos avanzados, como las redes neuronales profundas. A pesar de estos retos, el uso estratégico de estas tecnologías continúa transformando la forma en que las empresas farmacéuticas gestionan y aprovechan sus datos [17][18].

En conclusión, el Machine Learning y la Ciencia de Datos ofrecen herramientas indispensables para la innovación y la competitividad en la industria farmacéutica. Al integrarse en procesos como la segmentación de médicos, la optimización territorial y la predicción de adopción de terapias, estas tecnologías permiten a las empresas tomar decisiones más informadas y eficaces, maximizando su impacto en un entorno comercial cada vez más complejo y dinámico.

3.1.3. Optimización de Recursos Comerciales

La optimización de recursos comerciales es un elemento clave en la estrategia de marketing, especialmente en industrias como la farmacéutica, donde los costos asociados a las actividades promocionales son altos y los recursos, como el tiempo de los representantes de ventas y el presupuesto de marketing, son limitados. Este proceso busca maximizar el retorno de inversión (ROI) al asignar eficientemente los recursos hacia áreas, segmentos o actividades con mayor potencial de impacto. Según Bertsimas et al., la optimización de recursos comerciales no solo mejora la eficiencia operativa, sino que también permite a las empresas responder de manera ágil a los cambios en el mercado [16].

En el sector farmacéutico, la optimización de recursos comerciales se enfrenta a desafíos únicos. La diversidad de especialidades médicas, el comportamiento heterogéneo de los médicos en términos de prescripción y las estrictas normativas que regulan las interacciones entre empresas y profesionales de la salud exigen una priorización precisa y fundamentada. Por ejemplo, las leyes que restringen la cantidad y naturaleza de las interacciones comerciales en muchos países obligan a las empresas farmacéuticas a maximizar el impacto de cada visita o campaña promocional [4]. Esto hace que la optimización sea más que una cuestión de eficiencia operativa; es una necesidad estratégica.

Uno de los aspectos fundamentales en la optimización de recursos comerciales es la determinación de la frecuencia de visitas a los profesionales de la salud. El concepto de frecuencia óptima establece que cada médico tiene un número ideal de interacciones necesarias para

maximizar su probabilidad de adopción de un medicamento o terapia específica. Estudios como los de Shankar et al. han demostrado que un exceso de visitas puede generar saturación y efectos adversos, mientras que una frecuencia insuficiente puede resultar en un impacto limitado en la intención de prescripción [15]. La determinación de esta frecuencia óptima requiere un análisis cuidadoso de variables como el comportamiento histórico de prescripción, la especialidad médica, el nivel de influencia del médico en su comunidad y las características del medicamento promovido.

Además de la frecuencia de visitas, la priorización de médicos clave es otro pilar en la optimización de recursos comerciales. No todos los médicos tienen el mismo nivel de impacto en el mercado, y las empresas farmacéuticas deben concentrar sus esfuerzos en aquellos con mayor potencial estratégico. Esto incluye identificar a los médicos que tratan un volumen significativo de pacientes, aquellos que son líderes de opinión en su área terapéutica o aquellos más propensos a adoptar nuevas terapias. En este contexto, los modelos predictivos y las técnicas de segmentación, como las basadas en Machine Learning, juegan un papel crucial al clasificar a los médicos en segmentos de alto, medio y bajo impacto [7][9].

La optimización territorial es otro componente esencial. Este proceso implica diseñar rutas eficientes para los representantes médicos, asegurando que puedan visitar a los médicos prioritarios en el menor tiempo posible y con costos operativos mínimos. Según Topaloglu, la planificación territorial no solo reduce los costos de transporte y tiempo de viaje, sino que también garantiza una cobertura más equitativa y efectiva de las áreas asignadas [14]. Los modelos de programación lineal y heurísticas avanzadas, como los algoritmos de optimización basados en redes, se utilizan comúnmente para resolver estos problemas. Estas herramientas permiten considerar múltiples variables, como la ubicación geográfica de los médicos, la frecuencia óptima de visitas y las restricciones logísticas.

En la práctica, la optimización de recursos comerciales también incluye la evaluación continua del desempeño de las estrategias implementadas. El análisis de datos en tiempo real, combinado con herramientas de visualización como Tableau y Power BI, permite a las empresas farmacéuticas ajustar sus tácticas según los resultados obtenidos. Por ejemplo, si una campaña promocional en una región específica no genera el impacto esperado, los recursos pueden redistribuirse hacia áreas con mayor potencial o hacia actividades más efectivas [15].

Un aspecto crítico en la optimización de recursos comerciales es la alineación con los ciclos de vida de los productos farmacéuticos. Durante la fase de introducción de un medicamento, los esfuerzos promocionales suelen concentrarse en médicos innovadores o "adoptadores tempranos", quienes tienen mayor probabilidad de influir en otros profesionales. En etapas de crecimiento y madurez, la estrategia puede expandirse hacia médicos más conservadores o hacia aquellos que manejan un mayor volumen de pacientes. En la fase de declive, los recursos suelen redirigirse hacia productos más nuevos o hacia estrategias que mantengan la participación de mercado del medicamento en cuestión [3][6].

Finalmente, la optimización de recursos comerciales no solo tiene implicaciones financieras, sino también estratégicas. En un mercado saturado y altamente competitivo como el farmacéutico, la capacidad de asignar recursos de manera eficiente puede marcar la diferencia entre el éxito y el fracaso de una estrategia comercial. Además, en un entorno regulado, donde cada interacción con los profesionales de la salud debe justificarse, la optimización asegura que las empresas cumplan con las normativas mientras maximizan el impacto de sus iniciativas.

La optimización de recursos comerciales es un componente esencial en la estrategia de marketing farmacéutico. Al integrar análisis avanzados, modelos predictivos y herramientas de planificación territorial, las empresas pueden maximizar su ROI, mejorar la eficacia de sus campañas promocionales y adaptarse rápidamente a las dinámicas cambiantes del mercado. Este enfoque no solo mejora la eficiencia operativa, sino que también refuerza la posición competitiva de las empresas en un entorno altamente exigente y regulado.

3.2. ANTECEDENTES

3.2.1. Métodos Tradicionales de Segmentación

La segmentación de mercados en la industria farmacéutica, durante gran parte de su historia, se ha basado en métodos tradicionales que emplean datos simples y criterios predefinidos. Estos métodos, aunque útiles en sus inicios, han mostrado limitaciones significativas al enfrentar la complejidad y dinamismo de los mercados actuales. Según Kotler y Keller, los enfoques tradicionales de segmentación emplean variables demográficas, geográficas, psicográficas y conductuales para clasificar a los clientes en grupos relativamente homogéneos [1].

En el contexto farmacéutico, la segmentación tradicional se centró inicialmente en variables demográficas, como la especialidad médica, la edad o el género de los profesionales de la salud. Este enfoque simplificado facilitaba la identificación de médicos que podrían ser más relevantes para ciertos productos o terapias. Por ejemplo, los pediatras serían segmentados como prioritarios para vacunas infantiles, mientras que los cardiólogos se enfocaban en terapias para enfermedades cardiovasculares. Sin embargo, este tipo de segmentación rara vez consideraba factores más complejos, como las preferencias individuales de los médicos o sus patrones de adopción de terapias [2][4].

La segmentación geográfica también fue ampliamente utilizada, clasificando a los médicos según la ubicación de sus consultorios, clínicas u hospitales. Esto permitió a las empresas priorizar áreas con alta densidad de pacientes o necesidades específicas. Por ejemplo, en regiones urbanas donde el acceso a especialistas es más fácil, las estrategias promocionales podrían diferir de las

aplicadas en áreas rurales con recursos limitados. Aunque útil, este enfoque no abordaba adecuadamente las diferencias individuales entre los médicos de una misma región [3].

Otro enfoque común fue la segmentación psicográfica, que buscaba clasificar a los médicos según sus actitudes, intereses y valores. Este tipo de segmentación intentaba identificar médicos innovadores que estuvieran dispuestos a adoptar nuevas terapias más rápidamente. Sin embargo, los datos necesarios para implementar este enfoque no siempre estaban disponibles o eran difíciles de obtener, lo que limitaba su aplicabilidad en comparación con los métodos más directos, como los demográficos o geográficos [6].

La segmentación conductual, por otro lado, utilizaba datos históricos de prescripción y participación en eventos promocionales para clasificar a los médicos. Este enfoque se centraba en el comportamiento pasado como un indicador del potencial futuro. Por ejemplo, los médicos que recetaban con frecuencia un medicamento específico eran clasificados como prioritarios para recibir información sobre productos similares. Aunque este enfoque era más sofisticado que los anteriores, también presentaba problemas. La dependencia de datos históricos limitaba su capacidad para predecir cambios en el comportamiento de los médicos, especialmente en mercados dinámicos donde las condiciones pueden cambiar rápidamente [7].

A pesar de sus ventajas iniciales, los métodos tradicionales de segmentación presentan limitaciones importantes. En primer lugar, muchos de ellos dependen en gran medida del juicio y la experiencia de los representantes médicos, lo que introduce subjetividad y variabilidad en los resultados. Por ejemplo, dos representantes podrían clasificar al mismo médico de manera diferente, dependiendo de su percepción personal. Además, estos enfoques a menudo no consideran múltiples variables al mismo tiempo, lo que resulta en una visión fragmentada del mercado. Esto se traduce en decisiones subóptimas y en una asignación ineficiente de los recursos comerciales [8].

Otra limitación significativa es la falta de adaptabilidad de los métodos tradicionales. Estos enfoques se basan en datos estáticos que no reflejan cambios en tiempo real, como nuevas preferencias de los médicos o la entrada de competidores en el mercado. Esto hace que los modelos tradicionales sean reactivos en lugar de proactivos, limitando su capacidad para anticipar y adaptarse a nuevas dinámicas del mercado [9][14].

En mercados con paneles médicos grandes y diversos, la escalabilidad de los métodos tradicionales también es un desafío. Procesar manualmente grandes cantidades de datos para clasificar a miles de médicos es una tarea ardua y propensa a errores. Además, la creciente especialización médica y la introducción de nuevas terapias han aumentado la necesidad de métodos más precisos y sofisticados que puedan integrar múltiples variables y manejar grandes volúmenes de datos [11].

En conclusión, los métodos tradicionales de segmentación en la industria farmacéutica fueron una

base fundamental en sus primeros años, permitiendo a las empresas identificar y priorizar a los médicos de manera básica. Sin embargo, con el aumento de la complejidad del mercado, estas estrategias han mostrado importantes limitaciones, como la subjetividad, la falta de escalabilidad y su incapacidad para adaptarse a cambios dinámicos. La evolución hacia métodos basados en datos y tecnologías avanzadas, como el Machine Learning, ha permitido superar estas barreras y ofrece a las empresas herramientas más precisas y eficientes para la segmentación de mercados en la actualidad [5][8][16].

3.2.2. Evolución hacia modelos basados en datos

La evolución hacia modelos basados en datos en la industria farmacéutica ha transformado significativamente la forma en que se segmentan los mercados y se toman decisiones comerciales. Este cambio se ha visto impulsado por el incremento en la disponibilidad de grandes volúmenes de datos, el avance en tecnologías de análisis y la necesidad de superar las limitaciones inherentes a los métodos tradicionales de segmentación. Según Han et al., el análisis avanzado de datos permite integrar múltiples variables y descubrir patrones complejos que los enfoques tradicionales no son capaces de identificar [3].

En sus etapas iniciales, la segmentación basada en datos se limitaba a análisis estadísticos básicos. Herramientas como tablas de frecuencia y análisis de correlación se utilizaban para identificar tendencias y relaciones entre variables demográficas, geográficas o conductuales. Aunque estos métodos aportaban un nivel de objetividad mayor que la subjetividad de los representantes médicos, carecían de la profundidad necesaria para abordar la creciente complejidad de los mercados farmacéuticos. Además, estos análisis no eran dinámicos y no podían adaptarse rápidamente a cambios en el comportamiento de los médicos o a nuevas condiciones del mercado [1][6].

El desarrollo de tecnologías más avanzadas y la adopción de herramientas de ciencia de datos marcaron un punto de inflexión en la evolución hacia modelos más sofisticados. Las técnicas de análisis predictivo, basadas en algoritmos de Machine Learning, permitieron a las empresas farmacéuticas predecir comportamientos médicos y clasificar a los profesionales de la salud en función de su probabilidad de prescribir ciertos medicamentos. Según Bishop, los algoritmos supervisados, como la regresión logística y los árboles de decisión, se convirtieron en herramientas clave para estimar la adopción de nuevas terapias y priorizar a los médicos con mayor potencial de influencia [4].

La incorporación de técnicas no supervisadas, como el clustering, agregó otra dimensión a la segmentación basada en datos. Métodos como K-Means y los Modelos de Mezclas Gaussianas (GMM) permitieron agrupar médicos con comportamientos similares en clusters homogéneos, independientemente de que estos estuvieran etiquetados previamente. Estas técnicas identificaron patrones ocultos en los datos, revelando relaciones que no eran evidentes mediante

los enfoques tradicionales. Por ejemplo, el clustering permitió identificar subgrupos de médicos que, aunque pertenecían a una misma especialidad, tenían enfoques terapéuticos distintos en función de su ubicación geográfica o perfil de pacientes [11][13].

Un aspecto crítico en esta evolución fue la integración de datos provenientes de múltiples fuentes. Las empresas comenzaron a combinar información de prescripciones, registros clínicos, participación en eventos científicos y resultados de campañas promocionales. La interoperabilidad entre sistemas, como los CRM (Customer Relationship Management) y las plataformas de análisis de datos, facilitó una visión más integral del comportamiento de los médicos. Esto permitió no solo segmentar a los profesionales de manera más precisa, sino también personalizar las estrategias comerciales en función de sus necesidades específicas. Según Topaloglu, este enfoque integrador mejoró significativamente la capacidad de las empresas para responder a las dinámicas del mercado [14].

La evolución hacia modelos basados en datos también estuvo impulsada por la necesidad de mejorar la eficiencia operativa. En un entorno donde los recursos comerciales son limitados, la segmentación avanzada ayudó a las empresas farmacéuticas a priorizar sus esfuerzos en los médicos más relevantes. Esto incluyó la optimización de la frecuencia de visitas, el diseño de rutas más eficientes y la asignación de presupuestos promocionales basados en el impacto potencial. Estudios como los de Bertsimas et al. han demostrado que el uso de modelos predictivos y algoritmos de optimización puede aumentar significativamente el retorno de inversión (ROI) en campañas farmacéuticas, al reducir el desperdicio de recursos y maximizar el impacto en los médicos prioritarios [16].

Otro avance significativo fue el uso de visualización avanzada de datos para facilitar la interpretación de los resultados de segmentación. Herramientas como Tableau y Power BI permitieron representar gráficamente patrones complejos, haciendo que los equipos comerciales pudieran identificar rápidamente oportunidades y ajustar sus estrategias en tiempo real. Estas visualizaciones también mejoraron la comunicación entre los equipos técnicos y las áreas comerciales, haciendo que los insights generados por los modelos fueran más accesibles y accionables [15].

Sin embargo, la transición hacia modelos basados en datos no estuvo exenta de desafíos. Uno de los principales retos fue garantizar la calidad y consistencia de los datos. Datos incompletos, inconsistentes o desactualizados podían comprometer la precisión de los modelos predictivos. Además, la complejidad de algunos algoritmos avanzados, como las redes neuronales, planteó dificultades en términos de interpretabilidad y aceptación en un entorno altamente regulado como el farmacéutico. A pesar de estos desafíos, el impacto positivo de la adopción de modelos basados en datos ha sido innegable, transformando la segmentación médica en un proceso más objetivo, dinámico y efectivo [8][10].

En conclusión, la evolución hacia modelos basados en datos ha revolucionado la segmentación de

mercados en la industria farmacéutica. Al integrar técnicas avanzadas de análisis predictivo, clustering y visualización de datos, las empresas han logrado superar las limitaciones de los métodos tradicionales y adaptarse a las complejidades del mercado moderno. Esta transformación no solo ha mejorado la eficiencia operativa, sino que también ha permitido estrategias más personalizadas, impactantes y alineadas con las necesidades específicas de los profesionales de la salud.

3.2.3. Contexto de la industria Farmacéutica

La industria farmacéutica es un sector estratégico que combina innovación científica con estrategias comerciales complejas para desarrollar, producir y comercializar medicamentos que mejoren la salud de las personas. Este sector, considerado uno de los más regulados y competitivos a nivel global, enfrenta desafíos únicos derivados de las altas inversiones en investigación y desarrollo (I+D), el cumplimiento de normativas estrictas y la necesidad de adaptarse a las demandas cambiantes del mercado y los pacientes [2][4].

Uno de los aspectos más característicos de la industria farmacéutica es el alto costo asociado al desarrollo de nuevos medicamentos. Según Thomas, el proceso de desarrollo de un fármaco desde la investigación inicial hasta su aprobación comercial puede tardar más de una década y costar miles de millones de dólares. Esto se debe a los múltiples ensayos preclínicos y clínicos necesarios para garantizar la seguridad, eficacia y calidad del medicamento. Además, solo una pequeña fracción de los compuestos investigados llega al mercado, lo que hace que las empresas dependan en gran medida de unos pocos productos exitosos para recuperar sus inversiones [2][5].

La dinámica regulatoria también juega un papel central en la configuración de las estrategias comerciales en este sector. Normas internacionales, como las establecidas por la FDA (Food and Drug Administration) en Estados Unidos, la EMA (European Medicines Agency) en Europa y otras agencias regulatorias en el mundo, establecen estrictos controles sobre la aprobación, comercialización y promoción de medicamentos. Estas regulaciones, aunque necesarias para proteger la salud pública, limitan las interacciones entre las empresas farmacéuticas y los profesionales de la salud. Por ejemplo, las empresas deben proporcionar información precisa y respaldada científicamente sobre sus productos, lo que requiere campañas promocionales cuidadosamente diseñadas para cumplir con estas normativas [14].

Además de los desafíos regulatorios, la industria farmacéutica opera en un entorno altamente competitivo y dinámico. La entrada de medicamentos genéricos, tras la expiración de las patentes, representa una amenaza significativa para los ingresos de las empresas. Por ello, la diferenciación a través de la innovación en productos y servicios es crucial. Las empresas invierten constantemente en el desarrollo de nuevas terapias dirigidas, medicamentos biotecnológicos y tratamientos personalizados que ofrecen ventajas clínicas y comerciales. Según Bishop, estas innovaciones no solo buscan mejorar los resultados clínicos, sino también aumentar la eficiencia

en términos de costos y acceso [4][6].

En este contexto, la segmentación de mercados y la optimización de recursos comerciales se convierten en herramientas esenciales para mantener la competitividad. La identificación y priorización de médicos clave permite a las empresas maximizar el impacto de sus esfuerzos promocionales, mientras que la optimización territorial ayuda a garantizar una cobertura más eficiente de las áreas geográficas. Sin embargo, el sector también enfrenta la presión de demostrar valor no solo a los médicos, sino también a los pagadores, como sistemas de salud públicos y privados. Esto ha llevado a un enfoque creciente en los análisis de costo-efectividad y la implementación de estrategias basadas en datos para justificar el precio y el uso de nuevos medicamentos [3][15].

Otro aspecto importante del contexto farmacéutico es la especialización creciente en áreas terapéuticas específicas. Con el avance de la ciencia médica, las empresas han pasado de desarrollar medicamentos de uso generalizado a enfocarse en terapias dirigidas para condiciones complejas o enfermedades raras. Estas áreas, aunque representan un mercado más pequeño, ofrecen oportunidades significativas debido a la falta de alternativas terapéuticas y la posibilidad de precios premium. Este cambio ha intensificado la necesidad de estrategias promocionales precisas que aborden las necesidades específicas de cada grupo de médicos y pacientes [16].

La digitalización y la transformación tecnológica también han impactado profundamente en la industria farmacéutica. La adopción de tecnologías avanzadas, como la inteligencia artificial (IA), el análisis de big data y los sistemas de gestión de relaciones con el cliente (CRM), ha permitido a las empresas analizar grandes volúmenes de datos en tiempo real, identificar patrones de comportamiento en los médicos y personalizar sus estrategias comerciales. Por ejemplo, la integración de herramientas de análisis predictivo ha permitido anticipar cambios en el mercado y diseñar campañas más efectivas, mientras que la visualización de datos facilita la interpretación y comunicación de insights clave [3][8].

Por lo tanto, la industria farmacéutica opera en un entorno complejo donde confluyen desafíos regulatorios, presiones competitivas, la necesidad de innovación constante y la creciente demanda de evidencia de valor por parte de médicos y pagadores. Este contexto ha llevado a una transformación significativa en las estrategias comerciales, destacando la importancia de la segmentación de mercados, la optimización de recursos y el uso de herramientas avanzadas de análisis. Estas tendencias no solo buscan maximizar el retorno de inversión, sino también garantizar que los esfuerzos promocionales se alineen con las necesidades de los profesionales de la salud y, en última instancia, de los pacientes.

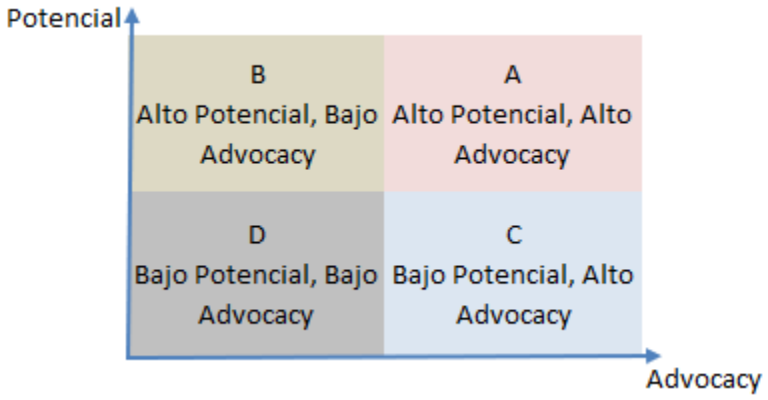
3.2.4. Contexto Sanofi Colombia

Hablando directamente de Sanofi Colombia, la empresa tiene muy presente que el targeting y segmentación es de vital importancia para una adecuada planeación estratégica a nivel de marketing. Es por ello por lo que se ha estado implementando modelos de segmentación empíricos basados únicamente en conceptos de los representantes de ventas y Gerentes de Distrito. El modelo actual solo obedece a identificar dos variables binarias que se detallan a continuación:

- Potencial: ¿El Médico es potencial para la marca? (SI/NO)
- Adopción: ¿El Médico prescribe para el paciente adecuado el producto de Sanofi como primera opción? (SI/NO)

A partir de las respuestas se agrupan a los HCPs dentro de los 4 cuadrantes posibles de la siguiente manera:

Ilustración 1. Modelo actual de Segmentación de Sanofi



Y es aquí donde la empresa considera necesario dar un salto a una forma más analítica y objetiva de encontrar efectivamente ese potencial y advocacy del HCP, sin entrar en subjetividades por parte de los visitadores Médicos. Para ello, en una primera instancia tendremos en consideración la recolección de nueva información mediante formularios.

4. ESTRUCTURACIÓN Y ANÁLISIS DE DATA

El análisis estadístico descriptivo de los datos permitió identificar tendencias clave y características generales del conjunto de datos, como la distribución de la cantidad de pacientes, el porcentaje de consulta y las especialidades más comunes. Este análisis inicial proporciona una visión clara de la variabilidad y los patrones predominantes, lo que facilita la identificación de áreas de interés para un análisis más profundo. El código utilizado para llevar a cabo la limpieza y el análisis exploratorio se encuentra documentado en el Anexo 10.2. Limpieza y análisis exploratorio.

4.1. CUESTIONARIO EN CRM (VEEVA)

Se desarrolló un cuestionario específico para la marca Gaucher, diseñado e integrado dentro del sistema de gestión de relaciones con el cliente (CRM) Veeva utilizado por la empresa. Este cuestionario tiene como objetivo principal recopilar información clave del mercado y obtener una comprensión más profunda de las preferencias, necesidades y comportamiento del panel médico enfocado en esta enfermedad rara.

Tabla 2. Cuestionario de perfilamiento Gaucher

No	Pregunta	Tipo respuesta
1	¿Cuántos pacientes al mes ve con esplenomegalia, trombocitopenia y/o hepatomegalia (al menos una) ?	Número entero
2	¿Cuál es el tipo de consulta que prefiere atender?	Benigno/Maligno
3	¿En que porcentaje atiende ese tipo de consulta?	Porcentaje

La creación del cuestionario contó con la participación de expertos en la enfermedad de Gaucher provenientes de diversas áreas estratégicas de la organización. Desde el área médica, se aportó conocimiento técnico y clínico que permitió formular preguntas relevantes para el entendimiento de las decisiones terapéuticas y la práctica clínica asociada a esta condición. Por su parte, el equipo de Customer Engagement contribuyó con su experiencia en interacción con los profesionales de la salud, asegurando que el cuestionario aborde los aspectos que influyen en la relación médico-empresa. Finalmente, el área de Sales Force Effectiveness (SFE) integró su perspectiva analítica y operativa, garantizando que las preguntas estuvieran alineadas con los objetivos comerciales y fueran fáciles de gestionar y analizar en el sistema Veeva.

El cuestionario fue diseñado con un enfoque estructurado y estratégico para recopilar datos relevantes sobre el comportamiento clínico, las preferencias de tratamiento y las necesidades del panel médico. Su integración en el CRM Veeva asegura un proceso eficiente de captura de datos, facilitando el análisis posterior y la generación de insights en tiempo real. Además, esta herramienta permite a los representantes médicos documentar las respuestas de manera inmediata durante sus interacciones, asegurando la precisión y actualidad de la información recopilada.

Este proyecto es un esfuerzo colaborativo que refleja la importancia de alinear diferentes áreas de la organización para abordar los retos asociados con la gestión de enfermedades raras. El cuestionario no solo constituye una herramienta clave para optimizar las estrategias comerciales de la marca Gaucher, sino que también fortalece la capacidad de la empresa para satisfacer las necesidades específicas del panel médico y mejorar su posicionamiento en este segmento altamente especializado.

Adicionalmente, se estableció un período de tres meses, correspondiente a un ciclo promocional completo, para que la fuerza de ventas capturara toda la información relevante mediante el cuestionario integrado en el CRM Veeva. Este tiempo fue considerado adecuado para garantizar una cobertura completa y uniforme del panel médico asociado a la enfermedad de Gaucher, permitiendo a los representantes médicos recopilar datos de manera detallada y sistemática durante sus interacciones.

4.2. DESCARGA DE DATA Y ANONIMIZACIÓN DE LA INFORMACION

Tras la finalización del período de captura de información definido para el cuestionario de la marca **Gaucher**, se procedió a la extracción de los datos recopilados desde el sistema **Veeva CRM**. Este proceso estuvo orientado a consolidar una base de datos que permitiera el análisis profundo de la información obtenida, garantizando tanto la calidad como la integridad de los datos. El data set tiene un tamaño de 547 filas y 8 Columnas. De los cuales se dividio en dos sub data sets (Benigno y maligno) de acuerdo con el tipo de consulta.

Tabla 3. Descripción del Dataset

Columna	Descripción
ONE KEY	Identificador único del cliente.
REPRESENTANTE	Nombre del representante encargado del cliente.
CUSTOMER_ID	Identificación numérica única del cliente.
CLIENTE	Nombre del cliente (persona u organización).
Cantidad_pacientes	Número de pacientes atendidos por el cliente.
tipo_consulta	Clasificación de la consulta: `benigna` o `maligna`.
porcentaje_consulta	Porcentaje relativo asociado a la consulta.
especialidad	Especialidad médica del cliente (por ejemplo, Genetista, Hematólogo).

Antes de proceder con el análisis, se implementó un riguroso proceso de **anonimización** de la data, en cumplimiento con las políticas internas de la empresa (ver anexo 9.1 Carta de Autorización uso de Datos) y las normativas regulatorias aplicables para la protección de datos personales y confidenciales. Este proceso incluyó la eliminación de toda la información identificable de los profesionales de la salud (HCP) y representantes médicos, la cual se encontraba en las columnas:

- CLIENTE
- CUSTOMER ID
- REPRESENTANTE

En su lugar, se asignó un identificador único a cada médico, diseñado específicamente para garantizar que no pudiera ser relacionado con ninguna identidad personal. Este enfoque aseguró que la privacidad de todos los participantes se mantuviera intacta y que el análisis posterior fuera completamente anónimo y conforme a los estándares éticos y legales.

4.3. ANÁLISIS EXPLORATORIO DE LA DATA

Se procedió a hacer todo el análisis exploratorio de esta data para un entendimiento inicial de esta, la cual es soportada por los expertos del negocio de **Gaucher** en Sanofi.

4.3.1. Estadística Descriptiva

Los resultados estadísticos descriptivos revelan interesantes tendencias en las variables analizadas. Para la variable Cantidad_pacientes, se observa que, después de la limpieza de datos, se cuenta con 546 registros completos. El promedio de pacientes registrados es de aproximadamente 70, lo que indica que, en general, los médicos manejan una cantidad moderada de pacientes. Sin embargo, la desviación estándar de 44.99 refleja una variabilidad considerable, lo que sugiere que algunos clientes atienden significativamente más pacientes que otros. El mínimo registrado es de solo 5 pacientes, mientras que el percentil 25 muestra que el 25% de los clientes tienen 30 pacientes o menos, apuntando a una concentración hacia valores bajos, aunque con excepciones que probablemente representen outliers con cantidades mucho mayores.

En cuanto a la variable porcentaje_consulta, también se cuentan con 546 registros válidos, lo que garantiza una base completa para su análisis. El promedio de esta variable es del 83.31%, lo que refleja que, en términos generales, los clientes tienen una alta efectividad en sus consultas. Sin embargo, la desviación estándar de 20.28% muestra que hay variaciones significativas entre los clientes. El mínimo registrado es un 5%, lo que indica que algunos clientes tienen tasas de consulta extremadamente bajas. A pesar de esto, el percentil 25 nos dice que el 25% de los registros se encuentran por debajo del 80%, lo que aún está cerca del promedio general.

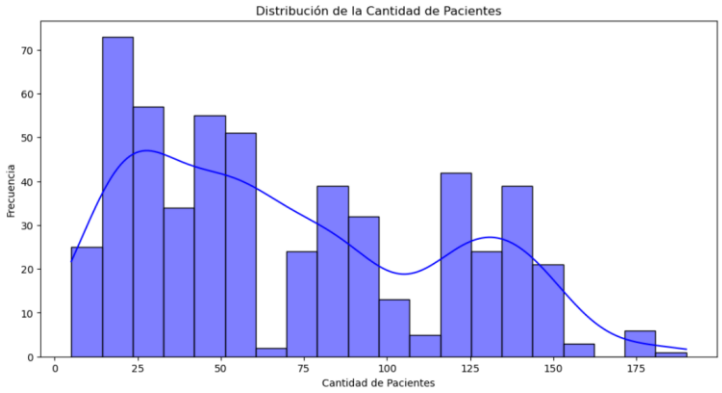
En síntesis, los datos muestran un panorama positivo en la mayoría de los registros, con un desempeño general sólido en términos de consultas realizadas. No obstante, las variaciones significativas en ambos indicadores sugieren la necesidad de investigar los casos extremos para entender mejor las dinámicas particulares de estos clientes. Esto podría incluir el análisis de outliers en la cantidad de pacientes o la identificación de factores que afectan la efectividad de las consultas en los casos más bajos.

Tabla 4. Estadística Descriptiva

	Cantidad_pacientes	porcentaje_consulta
count	546.000000	546.000000
mean	70.476190	0.833114
std	44.996354	0.202750
min	5.000000	0.050000
25%	30.000000	0.800000
50%	60.000000	0.900000
75%	110.000000	1.000000
max	190.000000	1.000000

4.3.2. Distribución de la cantidad de Pacientes

Ilustración 2. Distribución de Cantidad de Pacientes

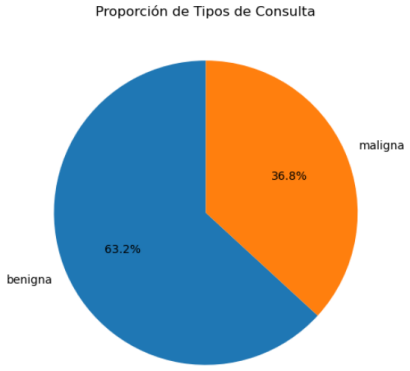


La gráfica de distribución de la cantidad de pacientes muestra una fuerte concentración de valores en los rangos bajos, con un descenso gradual a medida que los valores aumentan. Esto indica que la mayoría de los clientes manejan un número moderado o bajo de pacientes. Sin embargo, también se observan algunos casos de valores significativamente altos que destacan como outliers, representando clientes con un volumen excepcionalmente grande de pacientes.

La curva de densidad suaviza la distribución y refuerza esta observación, mostrando una clara asimetría hacia la derecha. Esto sugiere que la distribución no es completamente uniforme y está sesgada hacia valores más bajos, pero con una cola extendida hacia valores mayores.

4.3.3. Proporción por tipo de consulta:

Ilustración 3. Proporción tipo de consulta

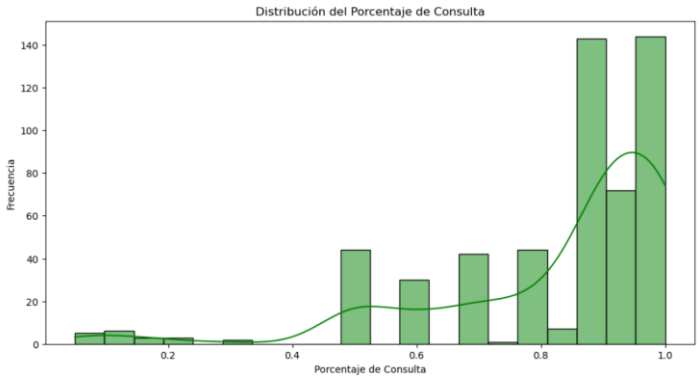


La gráfica muestra que más del 60% de las consultas corresponden a casos benignos, lo que indica un claro predominio sobre las consultas malignas. Este resultado es consistente con la naturaleza de la enfermedad de Gaucher, que se clasifica principalmente como una afección benigna en la mayoría de los casos. Esto explica por qué una mayor proporción de consultas está dirigida hacia esta categoría.

El predominio de consultas benignas refleja la naturaleza de la población médica atendida, donde la mayoría de los pacientes están relacionados con el manejo y seguimiento de esta enfermedad. Sin embargo, es importante destacar que aunque las consultas malignas representan una menor proporción, siguen siendo significativas y pueden requerir una atención especializada debido a la gravedad potencial de estas condiciones.

4.3.4. Distribución del porcentaje de consulta

Ilustración 4. Distribución del porcentaje de consulta

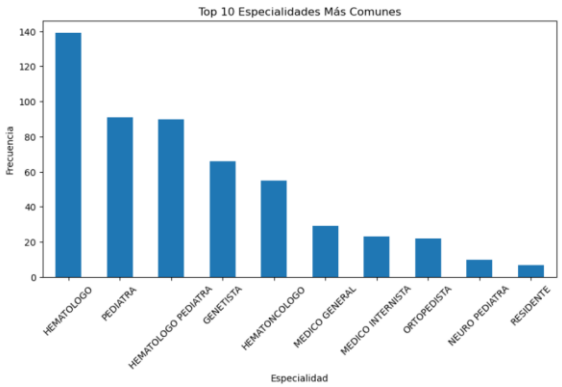


La gráfica muestra la distribución del porcentaje de consulta entre los clientes. La mayoría de los

valores están concentrados en el rango entre 50% y 100%, con un pico notable cerca del 100%. Esto indica que muchos clientes están alcanzando niveles altos de consulta, lo que sugiere un compromiso generalizado con la atención médica de su preferencia (benigno o maligno).

4.3.5. Top 10 especialidades más comunes

Ilustración 5. Top 10 especialidades



El gráfico destaca las 10 especialidades más comunes en los datos, con una clara predominancia de hematólogos y genetistas. Esta tendencia tiene mucho sentido en el contexto de la enfermedad de Gaucher, ya que estas especialidades están directamente relacionadas con el diagnóstico y tratamiento de esta condición.

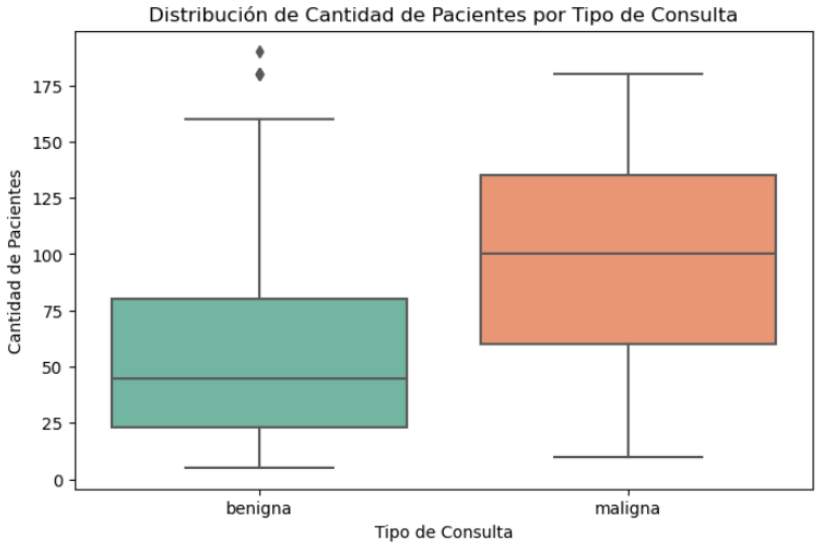
Los hematólogos desempeñan un papel crucial porque la enfermedad de Gaucher afecta principalmente al sistema hematológico. La acumulación de glucocerebrósidos en las células, característica de la enfermedad, puede provocar anemia, trombocitopenia (bajo recuento de plaquetas) y otras complicaciones hematológicas. Por esta razón, los pacientes con Gaucher suelen estar bajo el cuidado continuo de hematólogos, quienes manejan estas manifestaciones clínicas y monitorean los niveles sanguíneos.

Por otro lado, los genetistas son fundamentales en el diagnóstico y manejo de la enfermedad de Gaucher, ya que se trata de una condición hereditaria causada por mutaciones en el gen GBA. Los genetistas se encargan de realizar pruebas genéticas para confirmar el diagnóstico, identificar mutaciones específicas y asesorar a las familias sobre el riesgo de herencia y las opciones reproductivas. Además, su participación es clave en la identificación de portadores y en los estudios poblacionales que ayudan a entender la prevalencia de la enfermedad.

El predominio de estas dos especialidades en los datos refleja su papel central en la gestión de la enfermedad de Gaucher.

4.3.6. Análisis Cruzado

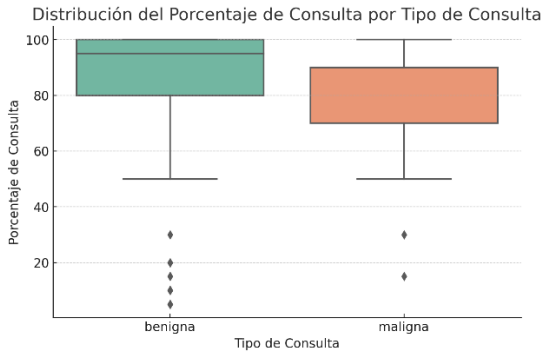
Ilustración 6. Distribución de pacientes por tipo de consulta



El boxplot muestra la distribución de la cantidad de pacientes por tipo de consulta (benigna y maligna), revelando diferencias significativas en su comportamiento. Las consultas malignas tienden a tener una mayor cantidad promedio de pacientes y presentan una mayor variabilidad, como lo indica la amplitud de los valores y la presencia de outliers. Esto podría reflejar que los clientes relacionados con consultas malignas tratan con una mayor diversidad de pacientes o casos más complejos.

En contraste, las consultas benignas muestran un rango más controlado y concentrado de pacientes, con menos variabilidad en comparación con las consultas malignas. Esto es consistente con el enfoque clínico de la enfermedad de Gaucher, en la que los casos benignos, aunque numerosos, tienden a ser más homogéneos en términos de manejo y seguimiento.

Ilustración 7. Distribución del porcentaje de consulta por tipo de consulta



Se muestra la distribución del porcentaje de consulta por tipo de consulta, revelando diferencias entre consultas benignas y malignas. En general, las consultas benignas tienden a tener un porcentaje de consulta más elevado y una menor variabilidad, indicando que estas consultas suelen realizarse de manera más consistente y eficiente. Esto puede reflejar que los casos benignos están mejor estructurados en términos de seguimiento y cumplimiento de consultas.

Por otro lado, las consultas malignas presentan un rango más amplio en los porcentajes de consulta, lo que indica una mayor variabilidad en la realización de estas consultas. Esto podría deberse a factores como la complejidad de los casos o la irregularidad en el seguimiento de los pacientes con condiciones malignas.

5. MODELOS DE CLUSTERIZACIÓN

Para abordar de manera integral la segmentación de los médicos, se emplearon modelos de aprendizaje no supervisado como K-means, DBSCAN, Clustering Jerárquico, Bisecting K-means y Modelos de Mezclas Gaussianas (GMM). Estos enfoques permitieron identificar patrones subyacentes en los datos y agrupar a los clientes según características clave, estableciendo segmentos claros que reflejan su comportamiento y potencial. Este proceso inicial es crucial para entender las dinámicas del mercado y crear una base sólida para la toma de decisiones estratégicas. Ver anexo 10.3 Resumen Modelos de cauterización.

Una vez obtenida la segmentación, se plantearon modelos de aprendizaje supervisado, como Regresión Logística, Random Forest, K-Nearest Neighbors (KNN) y Support Vector Machine (SVM). Estos modelos se entrenaron utilizando la segmentación generada como referencia, con el objetivo de construir una rutina automatizada que facilite la asignación de nuevos clientes a los segmentos previamente definidos. Este enfoque no solo asegura la escalabilidad del análisis, sino que también permite evaluar y optimizar el rendimiento de cada modelo, seleccionando aquel que ofrezca los mejores resultados en términos de precisión y adaptabilidad. Ver Anexo 10.9. Modelos ML Supervisados

El flujo de trabajo desarrollado tiene como finalidad no solo ser replicable en el tiempo, sino también adaptarse a diferentes geografías clave, como México y Argentina, donde las dinámicas del mercado pueden variar. De esta manera, se construye un proceso robusto y escalable que permite a Sanofi implementar una estrategia de segmentación consistente y eficiente en diferentes contextos, maximizando así el impacto de sus iniciativas comerciales y de marketing.

5.1. K-MEANS

Primero que todo se decide dividir la data en dos data frames (Maligno y Benigno). La división del DataFrame en consultas benignas y malignas se basa en sus diferencias clave en especialidades, cantidad de pacientes y porcentajes de consulta. Este enfoque permite realizar clusterizaciones específicas para cada grupo, identificando patrones únicos y relevantes. Separar los datos asegura un análisis más preciso y facilita estrategias personalizadas que optimizan recursos y mejoran la efectividad de las decisiones. Es una práctica recomendada antes de realizar cualquier análisis avanzado.

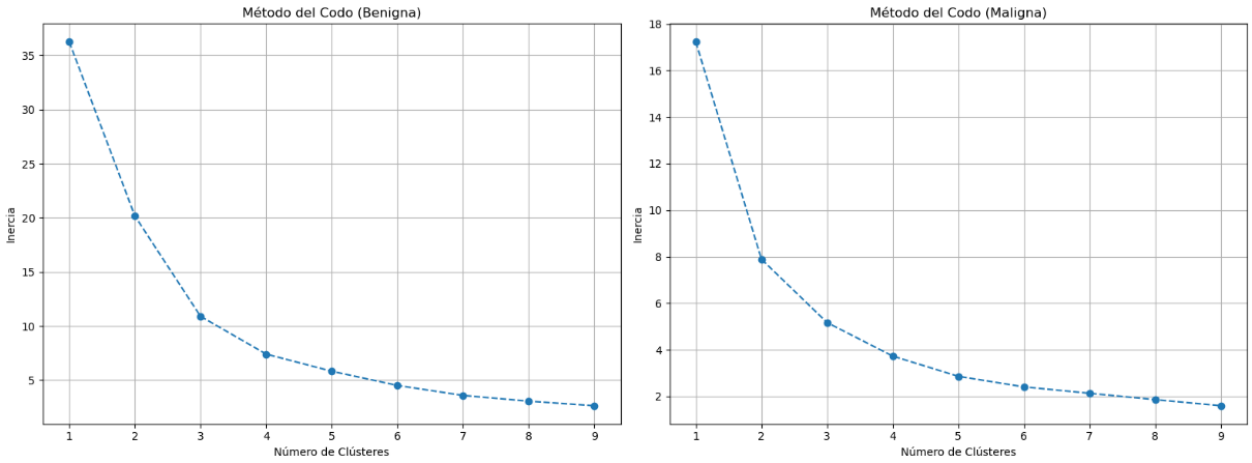
El proceso de clusterización incluye varios pasos clave. Primero, se realiza el preprocesamiento de datos seleccionando variables relevantes como cantidad de pacientes, porcentaje de consulta y especialidad (codificada si es necesario), y normalizándolas para asegurar una contribución equitativa. Ver anexo 10.4. K-Means.

El número óptimo de clústeres se determina utilizando métodos como el codo o el coeficiente de

silueta. Una vez elegido, se entrena el modelo, se asignan registros a los clústeres y se evalúan los resultados mediante visualizaciones y análisis de métricas como la inercia o el coeficiente de silueta. Finalmente, se interpretan los resultados para identificar patrones y ajustar estrategias según las necesidades de cada grupo, asegurando la escalabilidad y ajustes iterativos según sea necesario.

5.1.1. Análisis del método del Codo

Ilustración 8. Análisis del Método del Codo



El análisis del método del codo revela que la inercia disminuye rápidamente al pasar de 1 a 3 clústeres, lo que indica una mejora significativa en la cohesión de los grupos iniciales. A partir de 3 o 4 clústeres, la reducción de la inercia se vuelve menos pronunciada, lo que sugiere que agregar más clústeres no aporta beneficios significativos en términos de cohesión.

El número recomendado de clústeres es 3, ya que el "codo" del gráfico es más evidente en este punto. Esta elección equilibra la simplicidad del modelo con una agrupación efectiva de los datos. Como alternativa, 4 clústeres podrían considerarse si existen razones específicas desde el negocio o el dominio que justifiquen una segmentación más detallada.

En conclusión, dividir los datos en 3 clústeres parece ser la mejor opción, ya que captura la estructura principal de los datos sin introducir complejidad innecesaria, proporcionando un modelo eficiente y práctico para la segmentación.

5.1.2. Resultados K-Means

Tabla 5. Métricas Kmeans

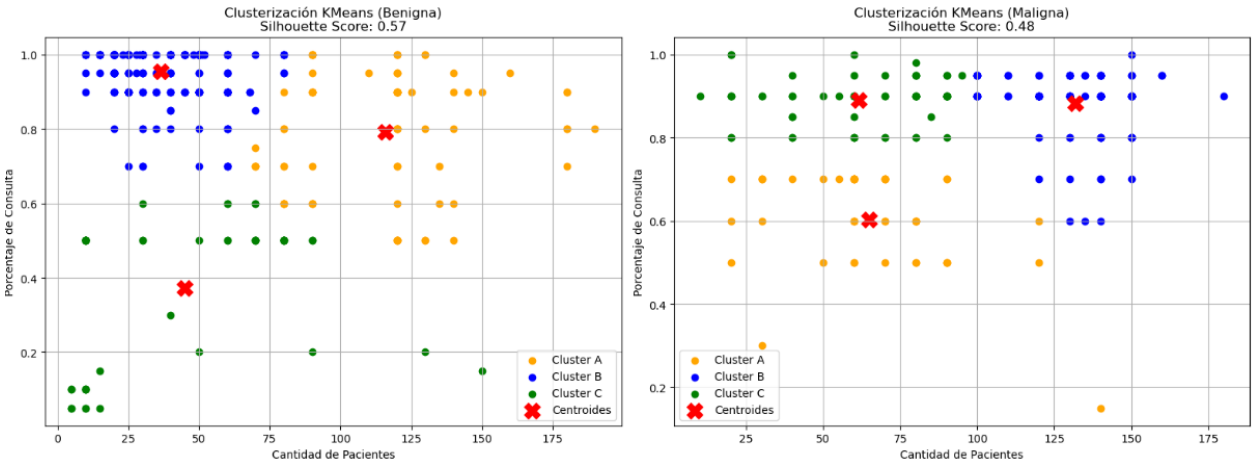
Métricas	Benigna	Maligna
Inercia	10.90	5.17
Coeficiente de Silueta	0.57	0.48

El análisis de las métricas de clusterización proporciona información clave sobre las características y la calidad de los clústeres formados para los datos benignos y malignos. En el caso de las consultas benignas, la inercia es de 10.90, lo que sugiere una mayor dispersión o variabilidad dentro de los clústeres. Sin embargo, el coeficiente de silueta de 0.57 indica que estos clústeres están razonablemente bien separados, con una agrupación interna sólida, lo cual es coherente con una buena separación visual en los gráficos generados.

Por otro lado, para las consultas malignas, la inercia es significativamente más baja, con un valor de 5.17. Esto indica que los clústeres malignos son más compactos, reflejando una menor dispersión interna. Sin embargo, el coeficiente de silueta de 0.48 revela que la separación entre los clústeres no es tan clara como en los datos benignos, lo que sugiere una mayor superposición entre los grupos. Este resultado se alinea con la observación gráfica, donde los clústeres malignos muestran una mayor mezcla.

En términos comparativos, los datos benignos presentan clústeres mejor definidos, con una separación más clara y métricas de silueta superiores. En contraste, los datos malignos muestran clústeres más compactos, pero menos diferenciados entre sí, probablemente debido a características compartidas que generan mayor solapamiento. Este análisis refuerza la importancia de adaptar las estrategias de segmentación según las particularidades de cada tipo de consulta.

Ilustración 9. Resultados Kmeans

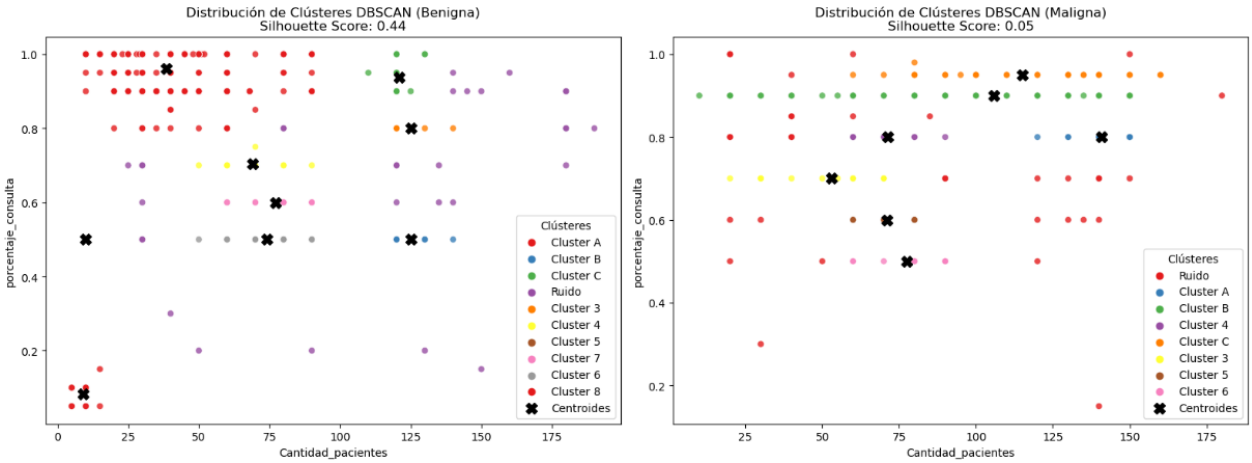


5.2. DBSCAN

El uso de DBSCAN para la clusterización permitió analizar la estructura de los datos en consultas benignas y malignas, arrojando resultados interesantes. Para las consultas benignas, el algoritmo logró identificar clústeres con una separación moderada, como lo indica un coeficiente de silueta de 0.44. Esto significa que los puntos dentro de cada clúster están agrupados de manera coherente y separados de otros clústeres, aunque con algo de variabilidad interna. Estos clústeres ayudan a descubrir patrones importantes, como la conexión entre la cantidad de pacientes y el porcentaje de consultas. Ver anexo 10.5. DBSCAN

En el caso de las consultas malignas, el modelo tuvo mayores dificultades para definir clústeres claros, evidenciado por un coeficiente de silueta bajo de 0.05. Esto señala una mayor superposición entre los clústeres y una asignación más frecuente de puntos como ruido. La menor definición podría atribuirse a la complejidad de los datos malignos y a la similitud entre los grupos, lo que sugiere que DBSCAN podría no ser el método más adecuado para segmentar este tipo de datos de manera efectiva.

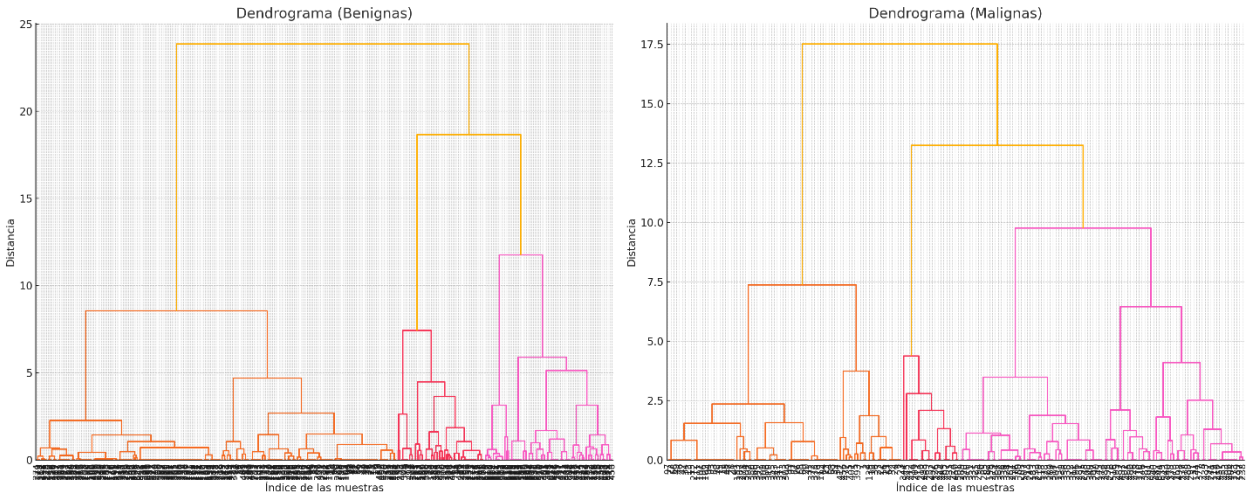
Ilustración 10. DBSCAN



5.3. CLUSTERIZACIÓN JERARQUICA

5.3.1. Dendrogramas

Ilustración 11. Dendrograma



El modelo de clusterización jerárquica permite analizar los dendrogramas generados, los cuales brindan una representación visual de las relaciones jerárquicas entre los datos. En el caso de las consultas benignas, el dendrograma muestra una estructura clara, con divisiones bien definidas entre los grupos. Esto indica que los datos benignos tienen patrones homogéneos que favorecen la formación de clústeres coherentes, facilitando su interpretación y uso en estrategias específicas.

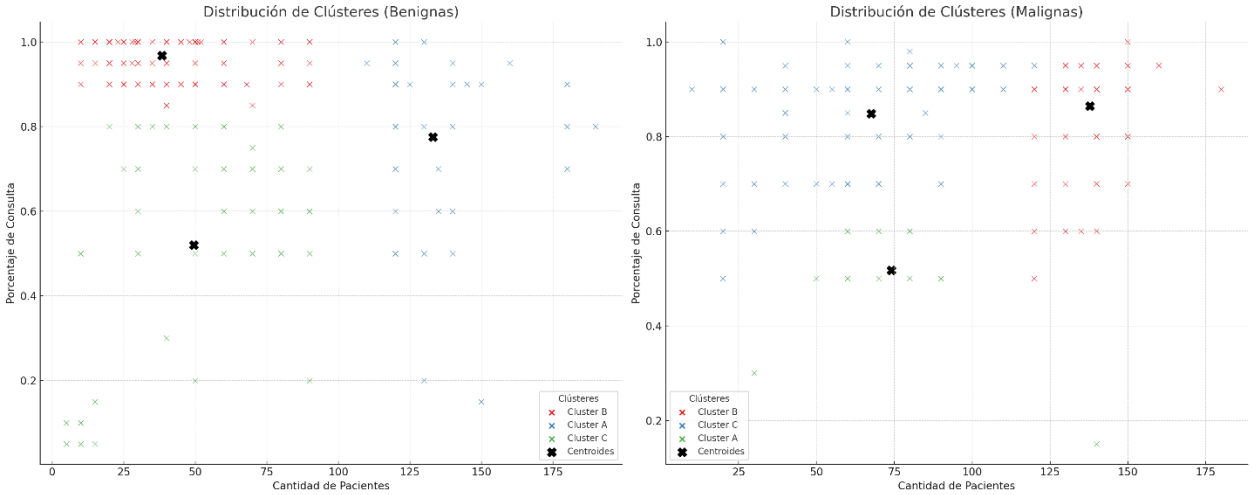
Por otro lado, el dendrograma correspondiente a las consultas malignas revela una estructura más compleja, con un mayor grado de proximidad entre los puntos. Esta mayor densidad sugiere una mayor similitud entre las observaciones y hace que las separaciones entre los clústeres sean menos evidentes. Como resultado, los grupos generados tienden a mostrar más superposición, lo que complica la segmentación y la interpretación directa.

5.3.2. Clusterización

El modelo de clusterización jerárquica aplicado a los datos de consultas benignas y malignas brinda una perspectiva organizada sobre la segmentación en función de las características analizadas. En el caso de las consultas benignas, el modelo demostró una adecuada cohesión interna y una separación clara entre los clústeres, con un coeficiente de silueta de 0.56. Esto evidencia que los grupos formados son bien diferenciados, permitiendo identificar patrones relevantes, como la relación entre la cantidad de pacientes y el porcentaje de consultas realizadas. Ver Anexo 10.6. Clusterización Jerárquica

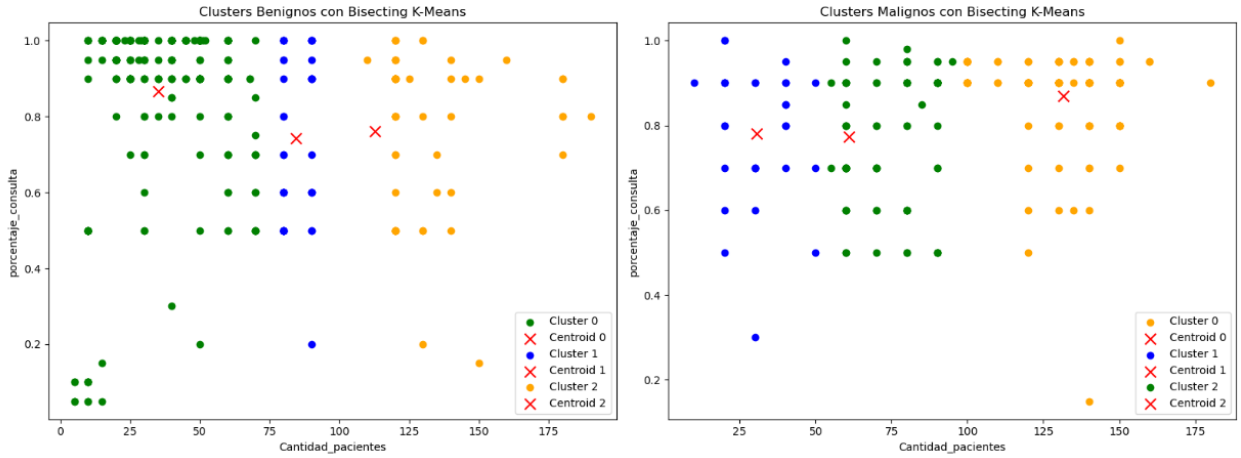
En contraste, para las consultas malignas, el modelo presentó dificultades para lograr una separación nítida entre los clústeres, lo cual se refleja en un coeficiente de silueta de 0.40. Esto sugiere una mayor superposición entre los grupos, probablemente debido a la complejidad intrínseca de los datos malignos y la similitud entre las observaciones. Si bien el modelo ofrece una segmentación inicial valiosa, puede no ser el enfoque más eficaz para datos con alta heterogeneidad o relaciones complejas.

Ilustración 12. Clusterización Jerárquica



5.4. BISECTING K-MEANS:

Ilustración 13. Bisecting K-Means




```
Coeficiente de Silhouette para datos benignos: 0.56
Coeficiente de Silhouette para datos malignos: 0.62
```

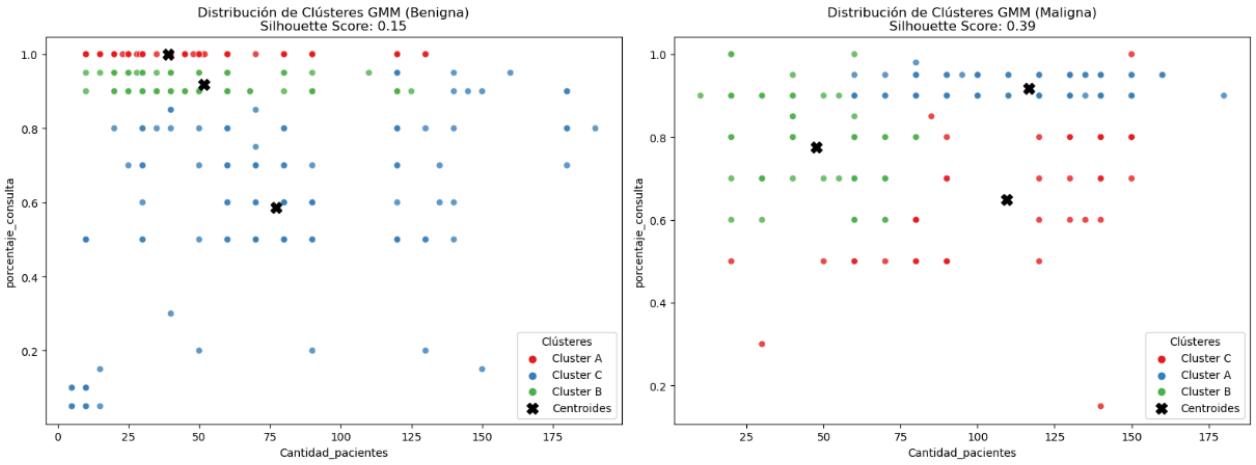
La técnica de Bisecting K-Means no cumplió con las expectativas según el análisis realizado junto a expertos en la materia. Aunque se obtuvieron clústeres con coeficientes de silueta razonables (0.56 para consultas benignas y 0.62 para consultas malignas), los resultados no reflejan las necesidades clínicas específicas del caso. Ver anexo 10.7. K-Means Bisecting

El modelo generó tres clústeres para cada tipo de consulta (benigna y maligna). Si bien los coeficientes de silueta indican una cohesión interna aceptable en los clústeres, las diferencias clave entre ellos se centraron en el número de pacientes. Por otro lado, no se logró una diferenciación significativa en el porcentaje de consulta, una variable crítica para la segmentación en este contexto. Además, la distribución de los clústeres fue predominantemente vertical, lo cual no corresponde con los patrones esperados. Según los expertos en la enfermedad de Gaucher, los clústeres deberían estar definidos por variaciones claras en el porcentaje de consulta, dado que este es un indicador fundamental para clasificar y segmentar a los pacientes de manera clínica y significativa.

A pesar de los coeficientes de silueta moderadamente altos, los resultados no cumplieron con las expectativas debido a la falta de diferenciación en dimensiones relevantes como el porcentaje de consulta. Esto sugiere que Bisecting K-Means podría no ser la técnica de clustering más adecuada para este problema.

5.5. GAUSSIAN MIXTURE MODEL

Ilustración 14. Gaussian Mixture Model



El análisis de los resultados obtenidos mediante el modelo de GMM (Gaussian Mixture Model) revela diferencias significativas en la calidad de la clusterización para consultas benignas y malignas. Para las consultas benignas, el coeficiente de silueta fue de 0.15, lo que sugiere una baja cohesión y separación entre los clústeres. Este valor indica que los puntos dentro de los

clústeres no están suficientemente agrupados y que existe una considerable superposición entre ellos, lo cual dificulta la identificación de patrones claros y consistentes.

Por otro lado, las consultas malignas mostraron un coeficiente de silueta de 0.39. Aunque este valor es más alto que el de las benignas, todavía refleja una separación moderada entre los clústeres. Esto indica que, si bien los clústeres malignos tienen una mejor cohesión interna y separación, la segmentación no alcanza el nivel de diferenciación esperado, especialmente para datos clínicos donde el porcentaje de consulta es una variable crítica.

En ambos casos, los coeficientes de silueta sugieren que el modelo de GMM no logró capturar de manera efectiva las estructuras latentes en los datos. Los gráficos confirman que los clústeres generados tienen una superposición considerable, lo que limita su utilidad para segmentar a los médicos de manera significativa. Ver anexo 10.8. Gaussian Mixture Model.

5.6. RESUMEN MODELOS CLUSTERIZACIÓN NO SUPERVISADOS

Tabla 6. Resumen Coeficientes de Silueta

Modelo	Coeficiente de Silueta	
	Benigna	Maligna
K-Means	0.57	0.48
DBSCAN	0.44	0.05
Jerárquica	0.56	0.40
Bisecting K-Means	0.56	0.62
GMM	0.15	0.39

El modelo K-Means mostró el mejor desempeño general, con coeficientes de silueta de 0.57 para consultas benignas y 0.48 para consultas malignas. Estos valores reflejan una buena cohesión y separación dentro de los clústeres, indicando que el modelo logró agrupar los datos de manera relativamente clara, especialmente en las consultas benignas. No obstante, la menor silueta en las consultas malignas sugiere que la delimitación entre clústeres en esta categoría no fue tan precisa, lo que podría deberse a una mayor variabilidad en los datos o a una posible superposición entre subgrupos.

Por otro lado, Bisecting K-Means obtuvo coeficientes de silueta de 0.56 en benignas y 0.62 en malignas, lo que en términos cuantitativos lo posiciona como el modelo con la mejor cohesión dentro de las consultas malignas. Sin embargo, un análisis cualitativo realizado en conjunto con expertos clínicos reveló que, a pesar de estos valores prometedores, los clústeres generados no lograron diferenciarse significativamente en el porcentaje de consulta, una variable clave para la segmentación en este estudio. Este hallazgo pone en evidencia la importancia de no depender exclusivamente de métricas como el coeficiente de silueta, sino de validar la utilidad práctica de

los clústeres en el contexto clínico.

Modelos como DBSCAN y GMM presentaron los coeficientes de silueta más bajos, especialmente en las consultas malignas. DBSCAN, con valores de 0.44 para benignas y apenas 0.05 para malignas, mostró una alta dispersión y falta de estructura en los clústeres malignos, lo que indica que el modelo tuvo dificultades para identificar grupos bien definidos en esta categoría. Este resultado es consistente con la naturaleza de DBSCAN, que tiende a ser más efectivo en conjuntos de datos con diferencias de densidad más marcadas, lo que podría no haber sido el caso en este análisis.

Por su parte, GMM (Gaussian Mixture Model) también mostró un desempeño deficiente, con coeficientes de 0.15 para benignas y 0.39 para malignas. La baja cohesión en la categoría benigna sugiere que el modelo no logró una segmentación clara, posiblemente debido a la suposición de distribuciones gaussianas que no se ajustaron bien a la estructura de los datos. Aunque su desempeño en la categoría maligna fue ligeramente mejor, sigue estando por debajo de otros modelos, lo que indica problemas de superposición y falta de definición en los límites entre clústeres.

En cuanto a la clusterización jerárquica, su desempeño fue intermedio, con coeficientes de 0.56 para benignas y 0.40 para malignas. Esto la sitúa como una alternativa más robusta que DBSCAN y GMM, pero aún menos efectiva que K-Means y Bisecting K-Means en términos de cohesión. Su ventaja principal radica en su capacidad para representar relaciones entre clústeres a diferentes niveles de granularidad, lo que podría ser útil para exploraciones iniciales. Sin embargo, su menor cohesión en las consultas malignas sugiere que la estructura de estos datos no se ajusta bien al enfoque jerárquico utilizado.

En general, estos hallazgos sugieren que, aunque K-Means sigue siendo la opción más adecuada en términos de métricas de silueta, es esencial complementar este análisis con evaluaciones cualitativas y expertas para garantizar que los clústeres obtenidos sean clínicamente significativos y alineados con las necesidades del estudio. La diferencia entre las métricas y la aplicabilidad real del modelo en la segmentación clínica resalta la importancia de ajustar cuidadosamente las variables de análisis y considerar enfoques híbridos o ajustes en la parametrización de los algoritmos para optimizar los resultados.

5.7. MODELOS ML SUPERVISADOS

En este análisis, se partió de un enfoque no supervisado utilizando el algoritmo K-Means, el cual demostró ser el mejor modelo para identificar patrones significativos en los datos. Los clústeres generados por K-Means fueron validados con la colaboración de expertos en la enfermedad de Gaucher, quienes confirmaron que la agrupación representaba adecuadamente las características clave y las diferencias clínicas entre los pacientes. Esta validación experta respaldó la selección de

K-Means como el modelo no supervisado más adecuado para este caso de estudio.

Posteriormente, los clústeres generados por K-Means se utilizaron como etiquetas para entrenar varios modelos supervisados de clasificación. El objetivo principal fue evaluar y comparar el desempeño de los modelos supervisados en términos de precisión, con el fin de identificar el modelo que mejor se ajuste a los datos y ofrezca un rendimiento óptimo.

Esta metodología no solo permitió aprovechar la robustez del modelo K-Means en la etapa inicial, sino también seleccionar el mejor modelo supervisado para su implementación en la toma de decisiones clínicas y estratégicas. Además, se busca que este enfoque sea escalable, permitiendo su aplicación en otras regiones como México y Argentina. Esto garantiza que los modelos desarrollados puedan adaptarse a las particularidades de cada región y mejorar la gestión de los pacientes en un contexto más amplio.

A continuación, se presenta un análisis detallado del desempeño de cada modelo supervisado, entrenado utilizando los clústeres generados por el modelo K-Means como etiquetas. Cada modelo fue evaluado en términos de precisión, recall, F1-score y matriz de confusión, para identificar el mejor modelo para replicar en otras regiones. Ver anexo 10.9. Modelos ML Supervisados.

5.7.1. Regresión Logística

La regresión logística, como modelo de clasificación lineal, mostró un desempeño aceptable pero limitado en comparación con otros modelos más avanzados. Con una precisión global del 68.90%, fue capaz de clasificar correctamente una parte significativa de los clústeres principales, pero enfrentó dificultades en los casos más complejos, especialmente aquellos con fronteras de decisión no lineales o una distribución de datos más dispersa.

Uno de los aspectos más notables en su desempeño fue su capacidad de identificar correctamente el clúster dominante, reflejado en su alto recall del 90% en esta categoría. Sin embargo, en los clústeres más pequeños y menos representados, la regresión logística presentó debilidades marcadas, con valores de recall de solo 55% y 25% para los clústeres 1 y 2, respectivamente. Estos resultados sugieren que el modelo tiene dificultades para capturar patrones en clases con menor frecuencia, un problema común en algoritmos lineales cuando los datos presentan distribuciones desbalanceadas o relaciones no lineales entre variables.

Uno de los factores que explica estas limitaciones es la naturaleza lineal de la regresión logística, que asume que la relación entre las variables independientes y la probabilidad de pertenencia a una clase es lineal. En conjuntos de datos más complejos, como el presente análisis, donde las diferencias entre clústeres pueden depender de interacciones no lineales o características con relaciones más intrincadas, este modelo se ve superado por alternativas más flexibles, como Random Forest, SVM o Redes Neuronales.

Además, la regresión logística puede verse afectada por multicolinealidad entre las variables predictoras, lo que impacta su estabilidad y la interpretabilidad de los coeficientes. En este caso, aunque se aplicaron técnicas de preprocesamiento para minimizar estos efectos, el modelo aún mostró dificultades en la clasificación de los clústeres más pequeños, lo que reduce su aplicabilidad en escenarios clínicos donde la identificación precisa de cada grupo es fundamental para el diagnóstico y la toma de decisiones médicas.

A pesar de sus limitaciones, la regresión logística sigue siendo una opción rápida, interpretable y fácil de implementar, lo que la hace útil en aplicaciones donde la interpretabilidad es clave y los datos presentan relaciones aproximadamente lineales. No obstante, para problemas más complejos y con mayor heterogeneidad en los datos, otros modelos más avanzados ofrecen un desempeño significativamente superior.

5.7.2. K-Nearest Neighbors (KNN)

El modelo K-Nearest Neighbors (KNN) destacó por su simplicidad y alto rendimiento, alcanzando una precisión del 97.56%. A pesar de ser un algoritmo basado en instancias y no en modelos paramétricos, logró un equilibrio notable entre precisión y recall en todos los clústeres, lo que sugiere que fue capaz de clasificar tanto las clases mayoritarias como las minoritarias con una efectividad destacable. Esta capacidad es especialmente valiosa en problemas donde la distribución de los datos puede ser desigual, asegurando que ningún clúster quede subrepresentado.

Uno de los factores clave que contribuyó al desempeño de KNN fue la escala estandarizada de las variables, ya que este modelo basa sus predicciones en la distancia entre puntos en un espacio multidimensional. La normalización de los datos permitió que KNN identificara vecinos cercanos de manera más precisa, evitando que variables con rangos numéricos más amplios dominaran la métrica de distancia utilizada (por ejemplo, la euclidiana o Manhattan).

Sin embargo, KNN también presenta ciertos desafíos. Una de sus principales limitaciones es su sensibilidad a la selección del valor de k, ya que un valor demasiado pequeño puede hacer que el modelo sea susceptible a ruido y sobreajuste, mientras que un valor demasiado grande puede suavizar en exceso la clasificación, reduciendo su capacidad de detectar patrones específicos. En este análisis, la elección óptima de k fue fundamental para mantener el balance entre sesgo y varianza.

Otro aspecto a considerar es la naturaleza computacionalmente intensiva de KNN, especialmente cuando se aplica a conjuntos de datos de gran tamaño. A diferencia de modelos como Random Forest o Support Vector Machine, que pueden entrenarse previamente y luego realizar inferencias rápidamente, KNN requiere calcular distancias en tiempo real para cada nueva instancia, lo que puede volverse prohibitivo en bases de datos masivas. Esta limitación es

particularmente relevante en aplicaciones con grandes volúmenes de información, como en estudios poblacionales en regiones extensas como México o Argentina, donde la cantidad de registros puede ser significativamente mayor.

A pesar de estos desafíos, el alto rendimiento de KNN en este análisis demuestra que sigue siendo una opción robusta y efectiva para tareas de clasificación, especialmente cuando se cuenta con un conjunto de datos bien preprocesado y de tamaño manejable. Su facilidad de implementación y la ausencia de una fase de entrenamiento lo hacen un modelo atractivo para aplicaciones donde se requiere rapidez en el ajuste y adaptabilidad a nuevos datos.

5.7.3. Bosques Aleatorios (Random Forest)

El modelo Random Forest fue el que mostró el mejor desempeño global, alcanzando una precisión del 99.39%. Su rendimiento excepcional se debe a su capacidad para manejar relaciones no lineales y datos heterogéneos, permitiéndole identificar con alta precisión todos los clústeres y lograr valores cercanos al 100% en métricas clave como precisión, recall y F1-score. Este resultado evidencia la eficacia de Random Forest en escenarios donde las fronteras de decisión son complejas y las clases pueden solaparse parcialmente.

Uno de los principales atributos de este modelo es su capacidad para manejar datos desbalanceados, un desafío común en problemas de clasificación. Al construir múltiples árboles de decisión y combinar sus resultados, Random Forest reduce el riesgo de sobreajuste, asegurando una generalización sólida incluso en conjuntos de datos con una distribución desigual de clases. Además, su enfoque basado en muestreo aleatorio (bagging) lo hace menos susceptible a la varianza de los datos, mejorando su estabilidad y confiabilidad.

Otra ventaja clave es que proporciona información sobre la importancia de las características, lo que permite identificar qué variables tienen mayor peso en la clasificación. Esta capacidad es especialmente útil en aplicaciones donde la interpretación del modelo es crucial, como en el ámbito clínico, donde conocer los factores más relevantes en el diagnóstico puede contribuir a la toma de decisiones médicas más informadas.

Dado su desempeño sobresaliente y su capacidad para manejar datos ruidosos y complejos, Random Forest se posiciona como una de las mejores opciones para problemas de clasificación robusta y confiable. En el caso particular de la enfermedad de Gaucher, donde la detección precisa es fundamental para un diagnóstico temprano y un tratamiento adecuado, este modelo se convierte en una herramienta valiosa al proporcionar resultados altamente precisos y fácilmente interpretables.

5.7.4. Support Vector Machine (SVM)

El modelo Support Vector Machine (SVM) demostró un desempeño sobresaliente, alcanzando

una precisión del 98.17%. Este resultado destaca la capacidad del modelo para clasificar los datos con gran exactitud, manteniendo un equilibrio notable entre precisión y recall en todos los clústeres. Esto indica que SVM maneja de manera efectiva fronteras de decisión complejas, lo que resulta especialmente útil en problemas donde los datos son bien definidos y, en gran medida, separables.

Uno de los factores clave en el desempeño del modelo fue la selección del kernel, que influye directamente en la capacidad de SVM para modelar relaciones no lineales entre las clases. En este caso, la elección del kernel adecuado permitió optimizar la separación de los datos, reduciendo el riesgo de sobreajuste y mejorando la generalización del modelo. No obstante, SVM puede ser computacionalmente costoso, especialmente en conjuntos de datos de gran tamaño, lo que puede afectar su escalabilidad en aplicaciones más extensas.

Si bien su desempeño fue similar al de Random Forest, existen diferencias clave entre ambos modelos. Mientras que Random Forest destaca por su interpretabilidad y capacidad de manejar datos con ruido sin necesidad de un preprocesamiento exhaustivo, SVM requiere una selección cuidadosa de hiperparámetros (como el parámetro C y el tipo de kernel) para garantizar un rendimiento óptimo. Además, su sensibilidad a outliers puede afectar la clasificación si no se emplean técnicas adecuadas de normalización y preprocesamiento de datos.

A pesar de estos desafíos, el alto nivel de precisión y balance en las métricas de desempeño hacen que SVM sea una opción sólida para este problema, especialmente en escenarios donde se requiere un modelo robusto para manejar límites de decisión no triviales con alta exactitud.

5.7.5. Comparativa modelos

La tabla comparativa destaca a **Random Forest** como el modelo más robusto, con una precisión del 99.39%, gracias a su capacidad para manejar datos complejos y heterogéneos, aunque con mayores requerimientos computacionales. **SVM**, con un 98.17% de precisión, también mostró un desempeño sobresaliente, ideal para datos bien definidos, pero requiere un ajuste cuidadoso de parámetros. **KNN** alcanzó el 97.56% de precisión, equilibrando simplicidad y rendimiento, aunque su escalabilidad puede verse limitada en grandes volúmenes de datos. Por último, la **Regresión Logística**, con una precisión del 68.90%, es una opción interpretativa y sencilla, pero inadecuada para datos complejos. En general, Random Forest es la opción más adecuada, seguido de SVM y KNN, dependiendo del contexto y los recursos disponibles.

Tabla 7. Comparativa modelos supervisados

Modelo	Precisión	Ventajas	Desventajas	Recomendación
Regresión Logística	0.689	Interpretable y eficiente para problemas lineales	Limitado para datos complejos o no lineales	Útil en problemas simples y cuando la interpretabilidad es clave
K-Nearest Neighbors (KNN)	0.9756	Buen equilibrio entre precisión y recall; simplicidad	Computacionalmente costoso para grandes conjuntos de datos	Adecuado para datos bien escalados y de tamaño moderado
Random Forest	0.9939	Máxima precisión; robustez y manejo de datos no lineales	Requiere más recursos; menor interpretabilidad	Mejor opción para problemas complejos y datos heterogeneos
Support Vector Machine (SVM)	0.9817	Capaz de manejar fronteras de decisión complejas; alto rendimiento	Ajuste complejo de parámetros; dependiente del kernel	Viable en escenarios donde se necesitan fronteras no lineales

Hablando netamente de las matrices de confusión, en el caso de la **regresión logística**, se observó un buen desempeño en la clasificación del clúster dominante (clúster 0), con la mayoría de las instancias correctamente asignadas a este grupo. Sin embargo, la capacidad del modelo para identificar correctamente los clústeres más pequeños (1 y 2) fue limitada. Esto se reflejó en un número significativo de falsos positivos y falsos negativos en estos clústeres, lo que indica que la regresión logística tiene dificultades para separar casos menos frecuentes o más complejos en el espacio de características.

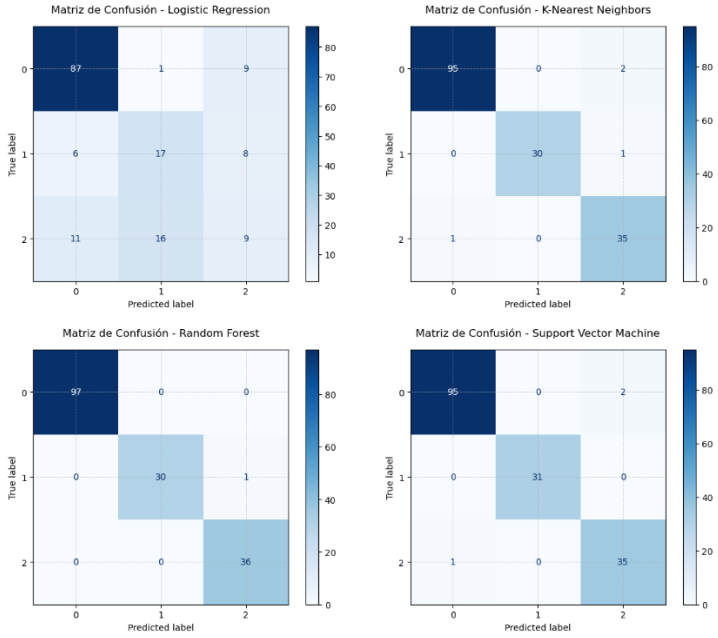
El modelo **K-Nearest Neighbors (KNN)** mostró un excelente equilibrio en la clasificación de los tres clústeres. Las instancias de los clústeres minoritarios (1 y 2) se identificaron con alta precisión y bajo error, lo que refleja la capacidad de KNN para capturar patrones locales en los datos. Esto sugiere que el modelo aprovecha eficazmente la relación entre los vecinos más cercanos para clasificar con precisión, aunque puede ser más sensible al ruido en los datos.

En el caso de **Random Forest**, la matriz de confusión demostró que este modelo sobresale al clasificar correctamente casi todas las instancias en los tres clústeres. Los falsos positivos y falsos negativos fueron prácticamente inexistentes, reflejando una excelente capacidad para manejar relaciones no lineales y complejas en los datos. Además, su robustez en la clasificación de clústeres minoritarios refuerza su idoneidad para aplicaciones donde la precisión es crítica, como en la gestión clínica.

Por su parte, el modelo **Support Vector Machine (SVM)** también mostró un desempeño sobresaliente, con una matriz de confusión que evidencia pocos errores en la clasificación de los tres clústeres. Aunque SVM no logró la perfección observada en Random Forest, su precisión en

los clústeres minoritarios fue notablemente alta. Esto refuerza la idea de que SVM es una opción poderosa en escenarios donde las fronteras entre clústeres son complejas y no lineales.

Ilustración 15. Matrices de Confusión de los modelos



En resumen, las matrices de confusión reflejan las diferencias clave entre los modelos supervisados. Mientras que Random Forest mostró el mejor desempeño global, KNN y SVM también se destacaron por su capacidad para manejar la complejidad de los datos. En contraste, la regresión logística tuvo dificultades en los clústeres menos representados, lo que limita su aplicabilidad en este contexto.

6. SEGMENTACIÓN FINAL – ENTREGABLE SANOFI

El entregable para la empresa es el código que se encuentra en el Anexo 10.10. Entregable Final Sanofi , que constituye una solución integral diseñada para abordar las necesidades de análisis y predicción basadas en clústeres generados previamente mediante técnicas no supervisadas, como K-Means. Este código permite entrenar un modelo de aprendizaje supervisado, en este caso un Random Forest, utilizando los datos etiquetados con clústeres validados por expertos. Además, incluye un flujo automatizado para realizar predicciones sobre nuevos datos y generar análisis detallados para respaldar la toma de decisiones estratégicas.

El código se compone de dos funcionalidades principales: el entrenamiento del modelo y la predicción. Durante el entrenamiento, el código preprocesa los datos (incluyendo la codificación de variables categóricas y la estandarización de características), entrena el modelo Random Forest, y evalúa su desempeño mediante métricas clave como precisión, matriz de confusión y reporte de clasificación. El modelo entrenado, junto con el escalador y el codificador, se guarda como un archivo reutilizable para facilitar su implementación y uso futuro.

En la etapa de predicción, el código permite cargar datos nuevos, preprocesarlos según las mismas transformaciones aplicadas durante el entrenamiento, y realizar predicciones sobre los clústeres a los que pertenece cada registro. Además, genera análisis estadísticos y visualizaciones, como tablas de conteos por clúster, gráficos circulares que muestran la distribución de los clústeres, análisis cruzados entre especialidades y clústeres, y gráficos de barras apiladas. Estas visualizaciones ayudan a comprender los resultados y respaldan la interpretación de los datos en un contexto clínico y estratégico.

Una característica clave de este entregable es que genera una salida en formato CSV que contiene las predicciones realizadas por el modelo, incluyendo los clústeres asignados y las etiquetas correspondientes. Este archivo CSV está diseñado para integrarse fácilmente con el CRM de la empresa, Veeva, permitiendo una actualización automatizada del sistema con la información más reciente. Esto asegura que los datos procesados y analizados estén disponibles de manera inmediata para los equipos operativos y estratégicos, optimizando así la gestión de los Médicos y la toma de decisiones en tiempo real.

Este entregable está diseñado para ser modular y escalable, lo que facilita su uso en distintas regiones, como México y Argentina, permitiendo la replicación de los análisis con datos regionales. Además, su estructura clara y el uso de herramientas estándar aseguran que sea fácil de mantener, ajustar y expandir según las necesidades futuras de la empresa. En esencia, este código proporciona una solución robusta y automatizada para integrar análisis de datos, predicción supervisada y generación de reportes, asegurando además una integración fluida con las herramientas operativas existentes de la empresa.

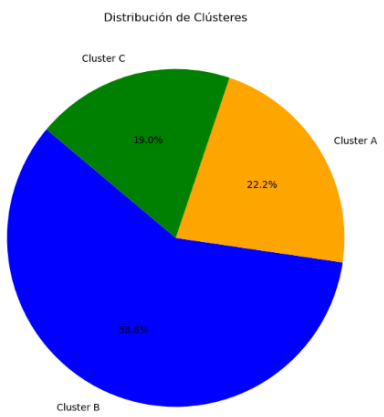
6.1. ANALISIS DE LOS RESULTADOS DE SEGMENTACIÓN

Los clústeres generados por K-Means representan grupos diferenciados dentro de los datos, que fueron validados por expertos en la enfermedad de Gaucher. Este paso asegura que los grupos no solo tienen sentido estadístico, sino también relevancia clínica. Cada clúster refleja un segmento específico de pacientes basado en características clave como:

- Cantidad de pacientes: Esto puede indicar grupos de alta o baja prevalencia de la enfermedad.
- Tipo de consulta: Diferencia entre consultas benignas y maligna.
- Porcentaje de consultas: Representa el peso relativo de cada grupo dentro del total.

El proceso de validación garantiza que los clústeres identificados son consistentes con el conocimiento médico, permitiendo una segmentación significativa para acciones clínicas y estratégicas.

Ilustración 16. Distribución de Clusteres



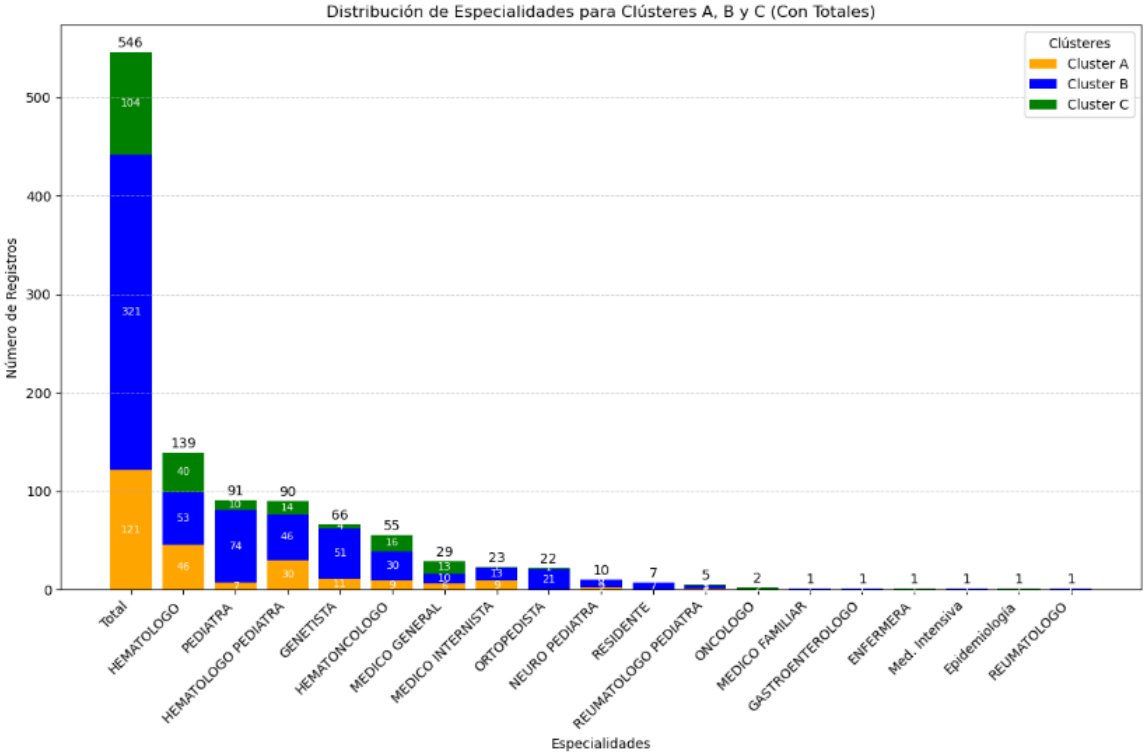
El diagrama circular ilustra claramente la distribución proporcional de los registros en cada segmento generado mediante la segmentación por K-Means, clasificados según el potencial estratégico definido por los expertos. El **Segmento A (Alto Potencial)**, representa el grupo más relevante para la estrategia de atención, debido a su impacto clínico y económico. Este segmento debe recibir la prioridad en términos de asignación de recursos y diseño de estrategias personalizadas.

El **Segmento B (Potencial Medio)** ocupa la mayor proporción. Aunque no es tan crítico como el Segmento A, sigue siendo un grupo relevante para estrategias de mantenimiento y crecimiento, con enfoque en optimización de recursos y seguimiento constante.

Finalmente, el **Segmento C (Bajo Potencial)**, con la menor proporción, refleja un grupo de menor impacto estratégico. Este segmento puede gestionarse con estrategias automatizadas o de bajo

costo, asegurando que los recursos principales se dirijan hacia los segmentos de mayor relevancia.

Ilustración 17. Distribución cruzada por cluster y especialidad



En cuanto a la distribución general de los clústeres, se observa que el Cluster B es el predominante, representando un 59% del total de registros. Este clúster domina ampliamente sobre los demás, lo que sugiere que su comportamiento característico está más presente en los datos. El Cluster A sigue con un 22%, mientras que el Cluster C abarca el 19% restante, indicando una menor prevalencia pero aún significativa en términos estratégicos.

En la distribución de especialidades por clúster, se destaca que Hematólogo es la especialidad más representada, con un total de 139 registros, seguido de Pediatra con 91 registros y Hematólogo Pediatra con 90 registros. El Cluster B domina en casi todas las especialidades principales, lo que refuerza su importancia como agrupación que engloba patrones comunes en estas especialidades clave.

Finalmente, el gráfico de barras apiladas que muestra la distribución de especialidades indica que Hematólogos y Pediatras tienen una representación destacada en todos los clústeres. El Cluster B mantiene su significancia en la mayoría de las especialidades, mientras que las menos representadas, como Enfermera, Reumatólogo Pediatra y Gastroenterólogo, tienen una

representación más equilibrada en los distintos clústeres, reflejando un comportamiento más homogéneo en estos casos. Estas conclusiones proporcionan una visión integral para priorizar estrategias en función de los clústeres y especialidades clave.

7. ESTRATEGÍA Y PRIORIZACION DE VISITA

Con base en la Segmentación y tneiendo en cuenta los comentarios de los expertos en Gaucher los clusters se pueden interpretarse en términos prácticos como sigue:

- **Segmento A:** Representa el grupo de mayor potencial, asociado a características como mayor número de pacientes y un porcentaje significativo de consultas. Este segmento probablemente incluye médicos con casos más complejos o críticos que requieren un seguimiento prioritario y estrategias personalizadas.

- **Segmento B**: Considerado de potencial medio, este grupo agrupa pacientes con métricas intermedias. Es posible que estos médicos necesiten atención constante, pero con menos urgencia o recursos comparado con el Segmento A.

- **Segmento C**: Clasificado como de bajo potencial, este segmento incluye médicos con menor relevancia estratégica. Es posible que los casos en este grupo sean consultas simples o con menor impacto clínico y económico.

Ahora, teniendo en cuenta los análisis, se desarrolló junto con el Customer Engament Lead de Gaucher la siguiente estrategia de marketing y planificación de frecuencias trimestrales para cada segmento:

Tabla 8. Frecuencias Trimestrales de interacción

Frecuencias Trimestrales por Segmento			
Canal	**A**	**B**	**C**
F2F	4	3	1
Teléfono	0	1	1
Whatsapp	0	1	1
Email personalizado	1	1	1

La estrategia propuesta se basa en una segmentación por prioridad, donde la frecuencia y los canales de contacto varían según la relevancia estratégica de cada segmento, maximizando el impacto en los grupos clave. Se emplea un enfoque de canales mixtos, que combina interacciones presenciales, digitales y remotas para garantizar una estrategia omnicanal efectiva que se adapta a las necesidades específicas de cada grupo. Los correos electrónicos personalizados actúan como un soporte transversal en todos los segmentos, reforzando la comunicación y brindando un canal constante de información. Además, la efectividad de la estrategia será monitoreada trimestralmente, lo que permitirá realizar ajustes en las frecuencias y enfoques según los

resultados obtenidos y las necesidades emergentes.

7.1. SEGMENTO A (ALTO POTENCIAL)

Este grupo incluye los casos de mayor relevancia estratégica, con pacientes críticos y complejos que requieren un seguimiento cercano.

Objetivo Estratégico: Establecer relaciones sólidas y de confianza con los médicos de este segmento, reforzando la presencia de la marca y brindando soporte constante para la gestión de los casos más complejos.

Frecuencia de Contacto:

- F2F (Face-to-Face): 4 visitas trimestrales. Este canal principal asegura una interacción cercana y personalizada, ideal para abordar temas estratégicos y generar un impacto directo.
- Teléfono: No se programan llamadas telefónicas, ya que el contacto presencial se prioriza.
- Whatsapp: No se utilizan interacciones por este canal, dado que la prioridad es mantener un trato personalizado.
- Email Personalizado: 1 correo trimestral diseñado específicamente para reforzar la comunicación sobre novedades, actualizaciones y materiales de soporte técnico y clínico.

7.2. SEGMENTO B (POTENCIAL MEDIO)

Este segmento engloba pacientes con relevancia intermedia, que requieren una combinación de contacto presencial y canales de soporte adicionales.

Objetivo Estratégico: Mantener una relación activa y de soporte con los médicos de este segmento, asegurando la continuidad en la gestión de casos relevantes y su alineación con los objetivos estratégicos de la marca.

Frecuencia de Contacto:

- F2F (Face-to-Face): 3 visitas trimestrales, asegurando un contacto presencial por mes
- Teléfono: 1 llamada trimestral para mantener la comunicación en los intervalos entre visitas presenciales y resolver dudas puntuales.
- Whatsapp: 1 interacción trimestral, que sirve como un canal práctico y directo para actualizaciones rápidas o consultas breves.
- Email Personalizado: 1 correo trimestral con contenido educativo, actualizaciones relevantes y recursos diseñados para apoyar la gestión de los pacientes.

7.3. SEGMENTO C (BAJO POTENCIAL)

Este segmento representa el menor impacto estratégico, lo que permite implementar una estrategia más optimizada en términos de recursos.

Objetivo Estratégico: Garantizar un nivel básico de contacto para mantener una relación activa con los médicos, asegurando que estén informados y alineados con los lineamientos básicos de la marca.

Frecuencia de Contacto:

- F2F (Face-to-Face): 1 visita trimestral para mantener el vínculo y resolver consultas básicas.
- Teléfono: 1 llamada trimestral para asegurar un nivel mínimo de comunicación y soporte.
- Whatsapp: 1 interacción trimestral, útil para mantener un contacto eficiente y no invasivo.
- Email Personalizado: 1 correo trimestral enfocado en proporcionar información clave, novedades y materiales de apoyo generales.

8. CONCLUSIONES Y TRABAJOS FUTUROS

8.1. CONCLUSIONES

El proyecto desarrollado representa un avance significativo en la optimización de la segmentación de los profesionales de la salud del panel médico de Sanofi Colombia, aportando un enfoque más eficiente, objetivo y basado en datos. La implementación de técnicas de machine learning, particularmente el modelo de K-Means, permitió identificar patrones en las características y comportamientos de los profesionales, segmentándolos en grupos estratégicos según su potencial (alto, medio, bajo). Este enfoque no solo redujo la subjetividad en el proceso de segmentación, sino que también mejoró la precisión y consistencia en la clasificación, creando una base sólida para la toma de decisiones.

Uno de los principales logros fue la segmentación diferenciada entre consultas benignas y malignas, una decisión estratégica que permitió personalizar los análisis y maximizar la relevancia de los resultados. Los datos fueron procesados y analizados para identificar cómo los médicos se distribuyen entre los distintos segmentos, proporcionando información clave para asignar recursos promocionales de manera más efectiva y estratégica. Esto no solo optimiza el impacto de las visitas médicas, sino también el uso de materiales promocionales, la planeación de eventos, y las actividades de comunicación digital, como correos personalizados y campañas en canales como WhatsApp.

Además, el uso de modelos supervisados como Random Forest y SVM permitió validar la segmentación inicial, demostrando que los grupos definidos por el modelo no supervisado son consistentes y útiles para el enfoque estratégico. Estos modelos también ofrecieron una visión replicable y escalable que puede aplicarse a nuevos profesionales de la salud o extenderse a otros mercados. El análisis de métricas como el Silhouette Score mostró la calidad de las agrupaciones, destacando que el segmento de alto potencial (Cluster A) concentra los médicos más estratégicos, mientras que el segmento de menor potencial (Cluster C) puede gestionarse con estrategias automatizadas o de bajo costo.

El proyecto también destaca por la integración de herramientas prácticas, como la generación de reportes en formatos accesibles (e.g., CSV), que pueden ser fácilmente integrados en el CRM de la empresa (Veeva). Esto asegura una actualización dinámica y automática de la información, facilitando la implementación de las estrategias por parte de los equipos comerciales y de marketing. Además, la propuesta incluye una planificación de frecuencias de contacto optimizada para cada segmento, alineada con los objetivos estratégicos de la empresa.

En términos globales, este proyecto no solo resuelve una problemática específica de Sanofi Colombia, sino que también sienta las bases para una transformación más amplia en la gestión estratégica de los médicos en mercados farmacéuticos. La creación de un modelo estándar, replicable y basado en datos permite una asignación de recursos más eficiente, un impacto

promocional optimizado y una alineación clara con las necesidades del mercado.

Finalmente, el éxito del proyecto abre la posibilidad de implementarlo en otros mercados, como México y Argentina, adaptando los parámetros a las características locales. Además, la metodología y los resultados obtenidos posicionan a Sanofi como líder en el uso de la inteligencia artificial y el análisis de datos para la toma de decisiones estratégicas en el sector farmacéutico. Este enfoque no solo mejora los resultados comerciales, sino que también refuerza el compromiso de la empresa con la innovación y la excelencia operativa.

8.2. TRABAJOS FUTUROS:

El presente proyecto abre múltiples oportunidades para mejorar, escalar y diversificar la implementación del modelo de segmentación basado en machine learning en Sanofi Colombia y otros mercados. A continuación, se detallan áreas clave para trabajos futuros:

- Replicación en Nuevas Regiones: Una de las prioridades es implementar el modelo de segmentación en mercados internacionales como México y Argentina. Cada país tiene dinámicas de mercado diferentes, por lo que sería necesario ajustar parámetros clave del modelo, como las variables utilizadas, los algoritmos de segmentación y las estrategias de frecuencia de contacto. Esta expansión permitirá evaluar la versatilidad del modelo y su capacidad para adaptarse a nuevos contextos, consolidando un enfoque global para la gestión estratégica de los profesionales de la salud.

- Integración en Tiempo Real: Incorporar datos en tiempo real desde el CRM (Veeva) y otros sistemas internos permitirá actualizar constantemente la segmentación. Esto posibilitará que el modelo responda de manera dinámica a los cambios en el comportamiento de los médicos o las condiciones del mercado, asegurando que las estrategias se mantengan relevantes y oportunas.

- Ampliación de Variables: Actualmente, el modelo utiliza variables clave relacionadas con el comportamiento del médico y las consultas. En el futuro, se podrían incorporar variables adicionales como indicadores socioeconómicos, datos epidemiológicos, resultados clínicos de pacientes atendidos o incluso datos externos, como la localización geográfica de los médicos, para enriquecer el modelo y hacerlo más robusto.

- Automatización y Herramientas de Usuario Final: Desarrollar herramientas automatizadas y accesibles para los equipos de marketing y comerciales permitirá que estos puedan aplicar el modelo sin necesidad de conocimientos técnicos avanzados. Un ejemplo sería la creación de dashboards interactivos que visualicen en tiempo real la segmentación, los resultados de las estrategias y las recomendaciones automatizadas para las visitas médicas.

- Personalización de Estrategias Promocionales: Basado en la segmentación, se podrían desarrollar estrategias de marketing altamente personalizadas que incluyan no solo frecuencias de contacto, sino también contenido promocional ajustado a las necesidades específicas de cada segmento. Por ejemplo, el uso de sistemas de recomendación para sugerir contenidos educativos, muestras médicas y eventos relevantes para cada médico.

- Estudio Comparativo con Competencia: Realizar análisis comparativos para evaluar cómo la implementación de este modelo posiciona a Sanofi frente a sus competidores. Esto permitirá identificar áreas de ventaja competitiva y oportunidades de mejora para fortalecer la posición de la empresa en el mercado.

- Capacitación Interna: Diseñar programas de formación para los equipos de marketing y ventas sobre cómo interpretar y aplicar los resultados del modelo de segmentación. Esto garantizará que el personal esté alineado con las herramientas y metodologías utilizadas, maximizando el impacto de las estrategias diseñadas.

Estos trabajos futuros no solo permitirán perfeccionar el modelo y sus aplicaciones, sino que también posicionarán a Sanofi como una empresa innovadora en el uso de datos para la gestión estratégica, consolidando su liderazgo en el sector farmacéutico.

9. REFERENCIAS

1. P. Kotler and K. Keller, *Marketing Management*, 15th ed. Pearson Education, 2020.
2. R. J. Thomas, *Pharmaceutical Marketing: Strategy and Cases*, 3rd ed. Routledge, 2020.
3. J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Elsevier, 2011.
4. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
5. Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452-459, May 2015.
6. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.
7. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.
8. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785-794.
9. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
10. J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp.*, 1967, pp. 281-297.
11. M. Ester et al., "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf.*, 1996, pp. 226-231.
12. I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.
13. U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395-416, 2007.
14. H. Topaloglu, "Territory planning in the pharmaceutical industry," *European Journal of Operational Research*, vol. 253, no. 2, pp. 1-11, 2016.
15. K. Shankar et al., "Impact of frequency optimization on pharmaceutical sales," *Journal of Business Research*, vol. 104, pp. 123-134, 2019.
16. D. Bertsimas et al., "Optimal resource allocation for pharmaceutical marketing," *Operations Research*, vol. 58, no. 4, pp. 1-15, 2010.
17. P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, 2012.
18. Apache Software Foundation, "Apache Hadoop Project." [Online]. Available: https://hadoop.apache.org/. [Accessed: Jan. 19, 2025].

10.ANEXOS

10.1. Carta de Autorización uso de Datos

28 de junio de 2024

Señores
PONTIFICIA UNIVERSIDAD JAVERIANA DE CALI

Apreciado,

Yo **Maria Isabel Viramontes**, en mi calidad de Head of Go-To-Market Capabilities LATAM de la empresa Sanofi, autorizo a **Santiago Reyes Zabaleta**, identificado con CC 1.018.465.371 de la ciudad de Bogotá, estudiante de la maestría de Ciencias de Datos, de la Pontificia Universidad Javeriana de Cali, a utilizar información anonimizada del panel médico para el proyecto denominado **"Segmentación de profesionales de la salud del sector farmacéutico por Machine Learning".**

Como condiciones, el estudiante se obliga a PRIMERO no divulgar ni usar para fines personales la información (documentos, expedientes, escritos, artículos, contratos, estados de cuenta y demás materiales) que, con objeto de la relación de trabajo, le fue suministrada; SEGUNDO no proporcionar a terceras personas, verbalmente o por escrito, directa o indirectamente, información alguna de las actividades y/o procesos de cualquier clase que fuesen observadas en la empresa durante la duración del proyecto. **El estudiante asume que toda información y el resultado del proyecto serán de uso exclusivamente académico.**

Atentamente,

Isabel Viramontes

Maria Isabel Viramontes
Head of Go-To-Market Capabilities LATAM

10.2. Limpieza y Análisis Exploratorio (GitHub)

10.3. Resumen Modelos de cauterización (GitHub)

10.4. K-Means (GitHub)

10.5. DBSCAN (GitHub)

10.6. Clusterización Jerárquica (GitHub)

10.7. K-Means Bisecting (GitHub)

10.8. Gaussian Mixture Model (GitHub)

10.9. Modelos ML Supervisados (GitHub)

10.10. Entregable Final Sanofi (GitHub)