Escuela de Negocios. Master in Management + Analytics

Modelo Probit en la Industria Farmacéutica Argentina

Autoría: Ko, Gastón Nicolás
Año: *2024*

UNIVERSIDAD TORCUATO DI TELLA

Master in Management + Analytics

Modelo Probit en la Industria Farmacéutica Argentina

TESIS

Gastón Nicolás Ko

Abril 2024

Tutor: M. Florencia Gabrielli

Resumen

El objetivo de esta tesis es modelar la probabilidad de compra de un producto específico dentro de una orden. Para lograr esto, es necesario comprender las variables a incluir en el Modelo Probit que se utiliza, el cual se emplea para predecir la probabilidad de demanda de algún producto.

El proyecto se divide en dos etapas principales. En la primera etapa, se utiliza un modelo de k-means para agrupar las farmacias según sus características, lo cual genera una variable relevante para analizar la probabilidad de compra en la siguiente etapa. En la segunda etapa, se utiliza un Modelo Probit para predecir la probabilidad de demanda del producto.

Uno de los aspectos relevantes de este proyecto es ilustrar el uso de técnicas de aprendizaje automático, en conjunto con el uso de modelos econométricos más tradicionales en la toma de decisiones, por ejemplo en temas relacionados con políticas de precios, ventas, promociones y marketing. Esta tesis muestra cómo puede llevarse a cabo dicho análisis en el contexto de la compra de un producto de consumo masivo que no requiere receta médica dentro de una plataforma electrónica utilizada por las farmacias.

Índice

1. Introducción

1.1 Contexto

En el contexto actual de la industria farmacéutica argentina, se utiliza un enfoque tradicional para establecer los precios de venta de los fármacos. Las regulaciones son acordadas entre la Cámara Industrial de Laboratorios Farmacéuticos Argentinos (CILFA) y el Estado. Sin embargo, debido al avance de las tecnologías digitales, ha surgido un nuevo paradigma en forma de ventas y compras digitales.

Esta tesis se centra en el estudio del comportamiento comercial de la industria farmacéutica argentina en el ámbito digital, específicamente a través de la plataforma Labbi, una startup B2B que ha revolucionado la relación entre laboratorios y farmacias. Labbi ofrece una alternativa que optimiza las ventas y mejora la rentabilidad para ambas partes.

El objetivo principal de este trabajo es modelar la probabilidad de demanda de productos en la plataforma Labbi. Para lograrlo, se propone desarrollar un modelo estadístico basado en el análisis de ventas históricas y la aplicación de diversas técnicas de aprendizaje automático. A lo largo de la tesis, se utilizan distintos algoritmos (ej: k-nearest neighbour) con el objetivo de optimizar los resultados de un modelo estadístico más tradicional que busca predecir la probabilidad de compra de una farmacia. Este análisis permite obtener una visión más clara y eficiente de la demanda en el mercado farmacéutico digital.

Comportamiento comercial de la industria farmacéutica argentina:

En el contexto actual de la industria farmacéutica argentina, se aplican medidas como topes máximos de incremento de precios sobre los productos medicinales para todos los laboratorios de la Nación. Según el informe de CILFA (2022) La industria farmacéutica argentina su carácter estratégico y perspectivas, "*el sector farmacéutico tiene una estrategia de precios acorde con las necesidades del país y alineada con la situación socioeconómica nacional y que facilita el acceso a los medicamentos.*"

En el Gráfico 1 a continuación, se presenta la evolución del precio de un medicamento, junto con sus respectivos márgenes de ganancia para cada agente involucrado en la cadena de comercialización. El proceso inicia con el precio de salida establecido por el laboratorio, continúa con el precio de venta hacia una droguería y finaliza con el precio al que la farmacia adquiere el medicamento, concluyendo en el precio ofrecido al público. (Subsecretaría de Programación Regional y Sectorial, 2022)

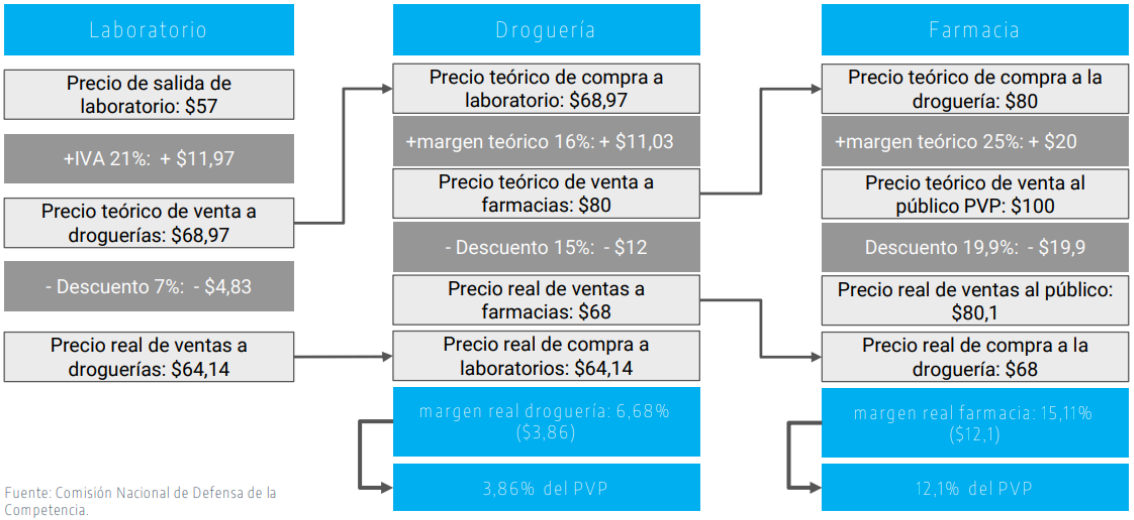

Gráfico 1: Precios de productos dentro de la industria farmacéutica argentina

Existen ciertos mecanismos de incentivos para fomentar las compras de productos de las farmacias. Entre ellos, la oferta de un descuento sobre un precio base a través de un intermediario (droguería), encargado de la distribución y facturación de los productos del laboratorio hacia la farmacia, también conocido como “transfer”.

Según un artículo de la Confederación Farmacéutica Argentina (2022), “la búsqueda por mejorar la rentabilidad de las farmacias tiene varias aristas. Uno de los caminos centrales a recorrer es el de optimizar la compra, y es aquí donde la presencia de los transfers es una alternativa eficiente para alcanzar ese objetivo”. En los últimos años, han ganado importancia debido a que los laboratorios buscan que sus productos sean recomendados en el punto de venta. Los farmacéuticos tienen un gran poder de influencia sobre los consumidores, ya que tienen la capacidad de sugerir ciertas marcas en lugar de otras a la hora de la venta al consumidor final. El transfer es la mejor herramienta para facilitar la relación entre el laboratorio y el punto de venta. El personal de ventas del laboratorio visita la farmacia y presenta los transfers activos detallando los valores porcentuales de los descuentos sobre los productos, unidades mínimas de compra, plazos de pago, etc. Si se realiza una transacción entre el laboratorio y la farmacia sobre productos con un descuento, esta se canaliza a través de la droguería que la farmacia utiliza.

La estrategia descrita de venta es empleada con el objetivo de mejorar la porción de mercado de un laboratorio en una industria concentrada pero con varios productos sustitutos. Según el reporte de la industria farmacéutica argentina de CILFA (2022), se identifican "211 principios activos", que corresponden a los ingredientes principales de los medicamentos. Estos principios activos representan el 50% del mercado de las drogas comercializadas en Argentina y cuentan con una oferta superior a diez productos distintos cada uno. En otras palabras, cada uno de estos principios activos, que abarcan la mitad de la facturación del país en el sector farmacéutico, tiene disponibles diez tipos de productos sustitutos, es decir, el grado de sustitución es significativo.

La estrategia de descuentos se diferencia por grupos de farmacias. El área comercial de los laboratorios utiliza la variable de unidades compradas históricas por farmacia para definir el tamaño y la importancia comercial de cada farmacia. De esta manera, los laboratorios establecen las siguientes categorías ordenadas de mayor a menor en términos de capacidad de compra: cadena de farmacias, farmacias independientes gestionadas por un vendedor físico del laboratorio y farmacias

chicas independientes que compran al laboratorio a través de un intermediario, la droguería. A su vez, se destaca que las últimas, obtienen el nivel más bajo de descuentos en esta estructura de categorización.

A pesar de que las farmacias más pequeñas representan más del 70% del total, reciben una gestión de ventas indirecta, ya que los laboratorios no pueden gestionar cada una de ellas sin incurrir en altos costos de personal, lo cual afectaría la rentabilidad del laboratorio. Ante esta situación, la solución estándar en la industria farmacéutica es priorizar las farmacias de mayor tamaño mediante la asignación de Gerentes de Cuentas Claves que las visitan mensualmente, mientras que las farmacias más pequeñas son atendidas de manera indirecta a través de la droguería sin un contacto directo con el laboratorio y en la mayoría de los casos, sin un descuento comercial beneficioso para ellas.

La droguería es un actor fundamental de la industria farmacéutica argentina, la cual tiene inventario completo de todos los productos comercializados en el mercado. La empresa cuenta con un grupo de clientes farmacéuticos a los cuales ofrece una línea de crédito y les proporciona un portal en línea que les permite acceder a los productos farmacéuticos. El factor de relevancia que obtiene la droguería en la industria farmacéutica es debido a la eficiente capacidad de logística de productos. En algunos casos, una entrega de productos puede realizarse en un mismo día y con un tiempo estimado de tardanza de 1 a 3 días. En Argentina, existen alrededor de 445 droguerías registradas. De estas, cuatro droguerías concentran cerca del 70% del mercado: Droguería del Sud, Droguería Monroe Americana (Grupo Gomer), Droguería Suizo Argentina y Droguería Barracas. Subsecretaría de Programación Microeconómica (2018). Las droguerías gestionan las compras de mercadería de las farmacias de manera física o a través del portal en línea utilizando los precios y niveles de descuentos comerciales por productos establecidos por el laboratorio.

Aunque este proceso comercial genere ventas para los laboratorios, la venta indirecta hacia farmacias independientes a través de un intermediario tiene algunas desventajas. Una de ellas es la falta de relación y comunicación directa con las farmacias produciendo de esta manera, un proceso tardío de promoción de nuevas estrategias comerciales como descuentos limitados y lanzamientos de nuevos productos. Además, los laboratorios carecen de acceso a la información de venta generada por cada transacción de la droguería hacia una farmacia, ya que esa información es propiedad exclusiva de la droguería. Los dos factores mencionados son los causantes de un proceso ineficiente de venta del laboratorio hacia más del 70% de las farmacias del país al no poseer un canal directo y controlado de comunicación y al mismo tiempo, no saber qué, cuándo o por qué compran.

La necesidad de entender estadísticamente la demanda de un producto en la industria farmacéutica

En 2019, nace la startup Labbi, un marketplace B2B que conecta laboratorios con farmacias a través de la venta de productos online con la misión de ayudar a las farmacias independientes del país. Gracias a la herramienta, cada farmacia chica que no es visitada por un vendedor físico de un laboratorio, puede acceder a un nivel superior de descuentos que antes no poseía. El ecosistema Labbi es resultado de la colaboración entre el Grupo Bagó, el grupo empresarial farmacéutico más grande de Latinoamérica, y YOPLABS, una incubadora digital especializada en el desarrollo y ejecución de iniciativas tecnológicas. Este emprendimiento conjunto busca aprovechar el

conocimiento del Grupo Bagó en el sector farmacéutico y combinarlo con la experiencia tecnológica de YOPLABS. Hoy en día Labbi ya está presente en Argentina, Bolivia, Uruguay y Ecuador. En Argentina, tiene más de 7000 farmacias registradas, lo que representa más del 60% de las farmacias del país, además de una facturación mensual de más de $150.000 dólares a fines del 2022, creciendo a doble dígito mes a mes.

Labbi brinda beneficios tanto a las farmacias como a los laboratorios. Para las farmacias, esto implica acceder a mejores condiciones comerciales y obtener mayores descuentos, lo que aumenta su rentabilidad. Por otro lado, los laboratorios tienen la capacidad de segmentar a cada usuario de manera individual, ofreciéndoles descuentos diferenciales que pueden variar según el usuario y el momento de la compra. Esta capacidad de segmentación era previamente inalcanzable. Gracias a estas ventajas, se abre la posibilidad de optimizar las ventas del laboratorio en términos de cobertura y penetración.

La plataforma actual respeta la cadena farmacéutica vigente. En ella, el laboratorio carga un catálogo de productos medicinales éticos, que exigen receta para su adquisición, y productos considerados como de consumo masivo, que pueden ser vendidos libremente. Los precios de los productos del catálogo son los estándares de la industria. Estos precios se actualizan mensualmente por los diferentes laboratorios. Los descuentos, correspondientes a cada uno de los productos, son designados por el equipo gerencial del laboratorio y aplicados en la plataforma. Estos transfers han sido previamente comunicados a la droguería. Estos descuentos superan el descuento base mínimo disponible para compras a través de droguerías, pero se sitúan por debajo de los descuentos ofrecidos por los ejecutivos de cuentas físicos de los laboratorios. Esta medida busca evitar conflictos de interés en la industria farmacéutica.

Aunque las condiciones comerciales propuestas para el universo de farmacias registradas en Labbi son más favorables que las disponibles a través de la droguería, estas no surgen de una estrategia óptima que maximice los resultados de los laboratorios. En cambio, buscan proporcionar una ventaja competitiva frente a la compra indirecta mediante la droguería.

Oportunidad y propuesta de mejora

La propuesta de valor más importante para el farmacéutico es la posibilidad de mejorar su rentabilidad al conseguir productos que son demandados constantemente por los consumidores finales que visitan sus farmacias, a un menor precio a través de descuentos más altos.

Al poder ofrecer descuentos diferenciados a cada farmacia, un laboratorio participante en la plataforma argentina encuentra una oportunidad de expandir su propuesta de valor hacia farmacias. Esta expansión no solo abarca a las pequeñas farmacias independientes, a las cuales antes no llegaba directamente, sino también a las farmacias con un nivel de venta intermedio. Anteriormente, estas últimas eran visitadas por ejecutivos de cuentas físicos mensualmente, pero ahora son atendidas por ejecutivos virtuales de venta dentro de Labbi, lo que mejora significativamente la contribución marginal del laboratorio al aumentar la venta de esas farmacias. Las farmacias que forman parte de la plataforma pueden acceder a sus mismos niveles de descuentos comerciales, tal como lo hacían previamente cuando un ejecutivo las visitaba para generar una venta. La única distinción radica en que ahora las compras pueden efectuarse en cualquier momento del día sin limitarse por las horas laborales del ejecutivo fisico. Además, el proceso de compra cuenta con una sólida gestión y

comunicación por parte de vendedores virtuales, quienes se basan en datos como análisis de recurrencia de compra de farmacias y de productos más comprados, para su enfoque y estrategia de comunicación y venta. Este grupo de farmacias son conocidas como farmacias "Migración", ya que migraron de la gestión física a la virtual. El proyecto hoy en día es exitoso con más de 3000 farmacias gestionadas, las cuales aumentaron la cantidad de unidades compradas, al mismo tiempo que se redujo el descuento medio comercial aplicando el porcentaje de descuento adecuado para cada farmacia. En consecuencia, se generó una reducción de costos para el laboratorio.

El proyecto "Migración" es un claro ejemplo de la oportunidad que existe de mejorar la rentabilidad de las farmacias y los laboratorios mediante la oferta de productos altamente demandados con descuentos más atractivos a las farmacias con mayor capacidad de compra. Dentro del desarrollo del proyecto, es necesario analizar ciertas características de las farmacias para la creación de los planes comerciales: los tipos de productos que compran a través de la plataforma en línea, los días que compran, las unidades promedios que tienden a comprar o también, la zona en la que se localizan. Estas variables son utilizadas con el objetivo de tomar mejores decisiones de negocios y aumentar mensualmente las ventas en la plataforma. Sin embargo, muchas de esas acciones implementadas, no tienen un resultado significativo. Este escenario, es un claro ejemplo, de la oportunidad de utilización de metodologías estadísticas tradicionales combinadas con técnicas de aprendizaje automático para analizar la información histórica de las farmacias en el contexto de la compra de medicamentos por parte de las farmacias que utilizan una plataforma electrónica y de esa forma, obtener resultados que ayuden en el proceso de toma de decisión para el aumento de ventas del laboratorio.

Para que el laboratorio capitalice plenamente esta oportunidad, es crucial dentro de la toma de decisión, predecir de manera efectiva la demanda de un producto con el fin de crear y gestionar planes comerciales con mayor efectividad al entender mejor la probabilidad de compra de una farmacia sobre un producto en particular. Principalmente, entender la demanda es uno de los factores más importantes a la hora de crear cualquier estrategia comercial efectiva. El presente trabajo propone un modelo estadístico basado en las ventas históricas de las farmacias en la plataforma electrónica Labbi, con el objetivo de modelar la probabilidad de compra de un producto específico. A través de un enfoque metodológico de dos pasos, se busca tomar decisiones informadas, descartando hipótesis infundadas y optimizando el modelo para obtener la máxima eficiencia.

Este trabajo consta de dos pasos.

1. Primero, se agrupan las farmacias según sus características mediante un modelo de k-means, método de aprendizaje no supervisado, que se utiliza para agrupar conjuntos de datos en diferentes categorías, creando una variable explicativa a la hora de estudiar la probabilidad de la compra de un producto.

2. Segundo, se utiliza un Modelo Probit para predecir la probabilidad de demanda de un producto específico e identificar cuáles son las variables relevantes a analizar para predecir la compra del mismo.

En resumen, el objetivo de esta tesis es modelar la probabilidad de compra de un producto dentro de una orden mediante el uso de un Modelo Probit acompañado de técnicas de aprendizaje automático para optimizar los resultados. Se busca comprender de manera más precisa los factores que pueden ayudar a mejorar la eficientización de las ventas de los laboratorios en la plataforma en línea.

1.2 Problema

La incapacidad de modificar libremente los precios de los productos, sumado a la ausencia de una plan eficaz de descuentos y la necesidad constante de los laboratorios de mejorar su rentabilidad, resulta en la creación de estrategias de ventas ineficientes. Estas estrategias consisten en ofrecer un mismo descuento fijo a las 8000 farmacias dentro de la plataforma, sin realizar distinciones entre ellas. Esto genera altos costos comerciales día a día para el laboratorio al no saber con exactitud dónde invertir sus esfuerzos comerciales para aumentar sus ventas. Por ejemplo, hoy en día se destina mucho esfuerzo en la modificación intuitiva de niveles de descuentos o las fechas de promociones especiales, sin generar ninguna mejora en las ventas.

Este problema se evidencia principalmente en productos de consumo masivo, como por ejemplo los analgésicos. En este contexto, un consumidor acude a una farmacia en busca de un producto OTC (Over The Counter) sin requerir una receta médica. En ocasiones, el farmacéutico puede ofrecerle un analgésico de una marca que tiene un valor de compra inferior al que el consumidor está buscando con el objetivo de mejorar su rentabilidad. Por lo tanto, los equipos comerciales de los laboratorios se concentran en ofrecer descuentos agresivos sobre productos de consumo masivo para incentivar la compra creando un ambiente de fuerte competencia promocional entre diferentes laboratorios pero sin ningún resultado diferencial. El cambio de las palabras dentro de la comunicación de la promoción, la utilización de otra audiencia de farmacias a la hora de comunicar y otros mecanismos, son claros ejemplos de acciones sin ningún aumento positivo de las ventas para los laboratorios. En este contexto, existe una gran incertidumbre de datos relevantes que ayuden a los equipo de negocios del laboratorio para incentivar la venta de esos productos.

En este trabajo, se analiza un producto de consumo masivo, en contraste con los medicamentos recetados por un profesional médico que tienen una relación entre la demanda y el precio menos flexible debido a las leyes dentro del sistema de salud del Estado. Este factor genera poco espacio de gestión a los laboratorios para modificar el descuento a ofrecer. Además, a fines del año 2022, la Cámara de Laboratorios de la Argentina (CILFA) ha estado trabajando en un acuerdo, al cual se han sumado los principales laboratorios del país. Este acuerdo tiene como objetivo regular los descuentos comerciales aplicados a los productos bajo receta médica, con el fin de equilibrar el mercado de medicamentos. Según se detalla en una nota periodística dirigida a las farmacias, “los laboratorios acordaron reducir “sensiblemente el porcentaje” de descuento que ofrecen a las farmacias a través de las droguerías. Hasta ahora ese descuento sobre los productos medicinales se ubica entre el 12% y 18%, pero disminuiría hasta un 7% a partir de diciembre del 2022. Por este motivo, la rentabilidad de las farmacias disminuiría de 3-5% a 2%.”

1.3 Objetivo

El propósito de este trabajo es aplicar un método basado en análisis de datos que permita modelar la probabilidad de comprar un producto dentro de la plataforma mediante la aplicación del Modelo Probit y la utilización de técnicas de aprendizaje automático para optimizar los resultados obtenidos.

En la industria farmacéutica argentina, se aplican acciones ineficientes para fomentar la venta de productos sin ninguna justificación basada en fundamentos estadísticos, más allá de la mera intuición del equipo a cargo de las ventas. Por ende, el objetivo del proyecto es desarrollar un modelo

probabilístico que proporcione una variable que calcule la probabilidad de compra de un producto perteneciente a la categoría de consumo masivo, como analgésicos, aspirinas y cremas corporales, dentro de una plataforma electrónica. Este enfoque se aplica específicamente para uno de los laboratorios participantes en la plataforma Labbi en Argentina, al que llamaremos “Laboratorio 1”. La razón principal es que este laboratorio tiene una gran necesidad de optimizar sus ventas y además, es el que mayor cantidad de datos históricos posee en la plataforma. A partir de los resultados del Modelo Probit y de la predicción de la probabilidad de compra, el laboratorio podrá utilizar este análisis para mejorar el proceso de toma de decisiones en sus acciones comerciales dentro de la plataforma.

La elección de la categoría de consumo masivo como foco de análisis se fundamenta en que tiene una gran cantidad de productos sustitutos y en la relevancia de los esfuerzos comerciales y de comunicación para su posicionamiento en el mercado. A diferencia de la categoría de productos bajo receta médica, los productos de consumo masivo, presentan una mayor dependencia en estrategias de marketing y ventas. Específicamente, se analiza la venta del producto con ID 2657 dentro de la plataforma, al que llamaremos “Analgésico A”, un analgésico antiinflamatorio indicado para aliviar y desinflamar de forma efectiva los dolores de espalda, dolores musculares y de las articulaciones. Es el producto más comprado dentro de la plataforma en la categoría de consumo masivo, lo que brinda mayor información para generar un mejor análisis de su venta. Además, se utiliza el producto con ID 1978, al que llamaremos “Crema A”, el cual es un producto corporal del mismo Laboratorio 1 que, al mismo tiempo, forma parte de los 3 productos de consumo masivo más comprados en la plataforma. El análisis de otro producto es favorable a la hora de validar los resultados obtenidos para el Analgésico A.

Primero, se emplea un modelo de k-means, un algoritmo de clustering no supervisado. En el aprendizaje no supervisado, no existe una medida de resultados y el objetivo es describir las asociaciones y patrones entre un conjunto de medidas de entrada. (Hastie, T., Tibshirani, R., & Friedman, J., 2009). Este algoritmo se encarga de segmentar objetos en "k" grupos, basándose en la distancia entre ellas, representadas por variables seleccionadas. Estas variables incluyen la fecha de registro, la fecha de primera compra de la farmacia en la plataforma, la cantidad total de unidades compradas, el valor monetario total de los productos adquiridos y la cantidad total de órdenes realizadas. El modelo tiene la función de segmentar las farmacias en grupos homogéneos, creando una variable única llamada “Segmento” para cada farmacia. Esta variable tiene un gran valor al sintetizar varias características de una farmacia en un único valor. Esto ayuda en la modelación de la probabilidad de compra de un producto.

Una vez definidas las variables que mejor explican la compra del producto, se utiliza un Modelo Probit para analizar las variables que influyen en la demanda de las farmacias en la plataforma. El modelo se basa en la información de ventas del año 2021, 2022 y comienzos del 2023 en la plataforma. Esto incluye tanto a las farmacias que forman parte del Proyecto Migración como a aquellas que no participan con el objetivo de utilizar la máxima cantidad de datos sobre las órdenes generadas.

Este es un breve listado de preguntas que el método a analizar intenta resolver:

- Dado un determinado nivel de descuento del producto y de las unidades dentro de una orden de una farmacia en la plataforma, ¿cuál es la probabilidad de que la farmacia compre el Analgésico A en una orden?

- ¿En cuanto mejoran las chances de que el Analgésico A sea comprado en una orden si aumenta el porcentaje de descuento?

- ¿Cuán confiables son las respuestas a las preguntas anteriores?

Este trabajo se estructura de la siguiente manera: en primer lugar, se presentan los datos empleados para llevar a cabo el análisis. Luego, se plantea la metodología empleada, detallando el desarrollo y los resultados obtenidos tanto del modelo k-means como del Modelo Probit. Por último, se incluye una sección con las principales conclusiones y una breve discusión de los resultados obtenidos.

2. Datos

2.1 Base de datos

Para el armado del modelo de segmentación k-means y el Modelo Probit, se emplean dos bases de datos recopiladas de la plataforma, con la debida autorización del Laboratorio 1. La primera base, denominada "base farmacias", almacena información cualitativa y cuantitativa acerca de las farmacias presentes en la plataforma. En la Tabla 1, se observa la información que incluye el identificador único de cada farmacia, la ubicación geográfica, la fecha de registro en la plataforma, la fecha de la primera compra en la plataforma, la fecha de la última compra, si pertenece al proyecto migración, las órdenes totales, las unidades totales y la facturación total. La base de datos está compuesta de 3866 usuarios repartidos en 24 provincias y 737 localidades que compraron en la plataforma productos de consumo masivo del Laboratorio 1 entre julio de 2021 hasta el 6 de marzo de 2023. Con el objetivo de lograr mayor precisión en el análisis de segmentación, se emplea únicamente la información de aquellas farmacias que han adquirido al menos una vez un producto de consumo masivo. Esto se debe a que ciertas farmacias solo compran productos bajo receta médica en la plataforma y no están dispuestos a comprar productos masivos, por lo tanto, su inclusión no aporta al análisis. Las variables que componen la base son:

- ID_Farmacia: número de identificación único de la farmacia en la plataforma.
- Localidad: localidad o ciudad donde está ubicada la farmacia.
- Provincia: provincia donde está ubicada la farmacia.
- Fecha_Registro: fecha en la cual la farmacia se registró.
- Fecha_Primera_Compra: fecha en la cual la farmacia hizo su primera compra en la plataforma.
- Fecha_Ultima_Compra: fecha en la cual la farmacia hizo su última compra en la plataforma.
- Farmacia_Migracion: “1” si la farmacia pertenece al grupo de farmacias de migración con un mayor nivel adquisitivo y “0”, en caso de que no pertenezca.
- Ordenes_Totales: sumatoria de pedidos finalizados generados por la farmacia en la plataforma.
- Unidades_Totales: sumatoria de unidades de productos comprados por la farmacia en la plataforma.
- Facturación_Total: sumatoria de valor monetario de compras generadas de la farmacia en la plataforma.

ID_Farmacia	Localidad	Provincia	Fecha_Registro	Fecha_Primera_Compra	Fecha_Ultima_Compra	Farmacia_Migracion	Ordenes_Totales	Unidades_Totales	Facturacion_Total
22	LA PLATA	BUENOS AIRES	19/7/2020	13/7/2021	10/11/2021	0	7	179	$ 155.074,90
27	PUNTA ALTA	BUENOS AIRES	20/7/2020	1/7/2020	4/10/2021	0	3	43	$ 15.781,44
28	ARROYO DULCE	BUENOS AIRES	20/7/2020	21/7/2020	27/10/2022	1	29	227	$ 110.580,03
29	MAR DEL PLATA	BUENOS AIRES	20/7/2020	2/7/2020	14/9/2022	1	26	1355	$ 919.725,82
30	PERGAMINO	BUENOS AIRES	20/7/2020	25/7/2020	5/10/2020	1	4	8	$ 5.013,20
31	POSADAS	MISIONES	20/7/2020	1/7/2020	7/4/2021	0	8	34	$ 17.099,78
32	PERGAMINO	BUENOS AIRES	20/7/2020	21/7/2020	1/11/2022	1	5	25	$ 7.983,71
33	ZEBALLOS	BUENOS AIRES	21/7/2020	7/7/2020	12/11/2020	0	3	10	$ 4.445,22
35	LUJAN	BUENOS AIRES	21/7/2020	26/10/2021	7/11/2022	1	4	209	$ 291.724,79
36	SAN PEDRO	BUENOS AIRES	21/7/2020	2/7/2020	10/3/2021	1	17	137	$ 56.398,53

Tabla 1: Base farmacias

En la Tabla 2, es posible observar que las órdenes totales de farmacias que tienen al menos una órden de productos de consumo masivo en la plataforma del laboratorio analizado son 33.415 generando así 1.519.912 unidades vendidas con una facturación de $1.694.537 de pesos argentinos en abril del 2023. Además, en la base de datos se puede comprobar la diversidad de actividad de las farmacias dentro de la plataforma. Al observar los registros, se encuentran farmacias que han realizado más de 114 órdenes y adquirido más de 36.740 unidades, mientras que otras presentan únicamente una orden y una unidad comprada. En promedio, cada farmacia genera alrededor de 9 órdenes, con un total de 393 unidades compradas, lo que representa una facturación media de $438.318 a valores de abril del 2023. Por último, es relevante mencionar que el proceso de registración en la plataforma ha sido sostenido en los últimos 3 años. Se registra la presencia de una farmacia con una cuenta registrada desde julio del año 2020 y se observa la última registración generada de una farmacia en enero de 2023. Este comportamiento similar se refleja también en las fechas de la primera y última compra de las farmacias, lo que indica un movimiento activo de registración y compras en la plataforma desde el año 2020.

	ID_Farmacia	Localidad	Provincia	Fecha_Registro	Fecha_Primera_Compra	Fecha_Ultima_Compra	Farmacia_Migracion	Ordenes_Totales	Unidades_Totales	Facturacion_Total
Total	3866	738	25	-	-	-	2498	33415	1519912	$ 1.694.537.284
Max	-	-	-	12/1/2023	19/1/2023	26/1/2023	1	114	36740	$ 31.660.666
Min	-	-	-	19/7/2020	1/7/2020	17/7/2020	0	1	1	$ 168
Media	-	-	-	18/5/2021	21/8/2021	29/6/2022	1	9	393	$ 438.318
Desvio Estandar	-	-	-	-	-	-	0	10	1125	$ 1.206.698

Tabla 2: Análisis base farmacias

La ingeniería de atributos implica la extracción y transformación de variables dentro de un conjunto de datos con el objetivo de aprovechar los mismos para un mejor entrenamiento y predicción en los modelos utilizados. Aplicando este mecanismo de transformación de las variables originales, es posible generar nuevas variables que resultan útiles para obtener un mejor conocimiento de las farmacias. Una de estas variables, Días_Registros, consiste en detallar la antigüedad de una farmacia en la plataforma. Esta variable se obtiene al calcular la diferencia en días entre la fecha de extracción de datos, en este caso 6 de marzo de 2023, y la fecha que la farmacia se dio de alta en la plataforma. Otra variable relevante calcula la diferencia entre la fecha de extracción de los datos y la fecha de la última compra realizada por la farmacia. Estas variables son importantes ya que de esta forma, se posee un número representado en días con el cual se puede medir la última actividad generada por la farmacia y compararlo con farmacias que pueden haber estado inactivas por más de 1 o 2 años en la plataforma. Además, se crea la variable Unidades_Por_Orden la cual se calcula dividiendo la cantidad de unidades totales sobre la cantidad de órdenes totales de cada farmacia. Esta variable es relevante ya que nos indica el promedio de unidades que compra una farmacia por orden. Utilizando la ingeniería de atributos, las variables creadas son:

1. Días_Registros: cantidad de días desde que una farmacia se registró en la plataforma.

2. Días_Primera_Compra: cantidad de días que una farmacia tardó desde que se registró para hacer su primera compra.

3. Días_Última_Compra: cantidad de días desde la última compra de una farmacia hasta el día de la extracción de los datos (6/2/2023).

4. Unidades_Por_Orden: cantidad de unidades compradas por una farmacia sobre la cantidad de órdenes generadas por una farmacia.

Con esta información, es posible identificar y segmentar farmacias según su antigüedad en la plataforma, la rapidez con la que efectuaron su primera transacción, su relación con la última compra realizada en la misma y las unidades promedio que compran por orden.

Por otro lado, se cuenta con una segunda base de datos denominada "base órdenes". Esta base contiene la información de las ventas realizadas por cada farmacia en productos de consumo masivo, analizados desde julio de 2020 hasta febrero de 2023 en la plataforma sobre productos del Laboratorio 1. En total, esta base de datos está compuesta por 329.164 líneas de productos comercializados que componen múltiples órdenes de compra:

- ID_Farmacia: número de identificación único de la farmacia en la plataforma.
- Fecha_de_Compra: fecha en la cual se efectivizó dicha compra.
- ID_Producto: identificador numérico único del producto.
- ID_Marca: número de identificación de las 18 marcas de productos de consumo masivo del Laboratorio 1 de los productos comprados.
- ID_Categoría: número de identificación asignado a cada una de las 9 posibles categorías de productos de consumo masivo del Laboratorio 1 del producto adquirido.
- Unidades: cantidad de unidades compradas de un mismo producto.
- Descuento: porcentaje descuento a aplicar sobre el precio original del producto comprado.
- Unidades_Mínima_Descuento: cantidades de unidades mínima a comprar para poder acceder al descuento del producto.
- Precio_Producto: valor monetario del producto comprado antes de la aplicación del descuento.
- Facturación_Total: valor monetario de la suma de unidades compradas multiplicado por el precio del producto aplicado el descuento.

En la Tabla 3, se observa que cada línea de la base de órdenes contiene información sobre un producto dentro de una orden de compra realizada por una farmacia con un total de 3866 farmacias. Los comportamientos de compra de las farmacias son únicos, evidenciándose que algunas realizan compras de 1 producto por orden, mientras que otras tienden a adquirir 5 productos en una misma orden. Esto se observa viendo la fecha de compra y el ID de la farmacia. Por ejemplo, en la columna 1 de la tabla, la farmacia de ID 22 compra 5 productos dentro de la misma orden. Asimismo, se ve que cada producto tiene un descuento aplicado sobre su precio y que, para acceder a dicho descuento, se requiere alcanzar una cantidad mínima de unidades en la compra. Dicho valor se observa en la columna “Unidades_Mínimas_Descuento” donde por ejemplo, para acceder al descuento del 4% sobre el precio del producto con ID 2664, se deben comprar un mínimo de 2 unidades del producto.

ID_Farmacia	Fecha_De_Compra	ID_Producto	ID_Marca	ID_Categoría	Unidades	Descuento	Unidades_Mínimas_Descuento	Precio_Producto	Facturación_Total
22	Jul 19, 2020, 3:00 AM	2664	158	84	2	4%	2	$ 626	$ 661
22	Jul 19, 2020, 3:00 AM	2668	158	84	2	4%	2	$ 626	$ 661
22	Jul 19, 2020, 3:00 AM	2667	158	84	2	4%	2	$ 626	$ 661
22	Jul 19, 2020, 3:00 AM	2680	158	84	2	4%	2	$ 626	$ 661
22	Jul 19, 2020, 3:00 AM	2811	158	84	2	4%	2	$ 626	$ 551
27	Jul 20, 2020, 3:00 AM	1844	38	30	4	8%	2	$ 113	$ 257
28	Jul 20, 2020, 3:00 AM	2216	49	17	2	8%	2	$ 1,260	$ 668
28	Jul 20, 2020, 3:00 AM	2183	49	17	2	8%	2	$ 1,447	$ 840
28	Jul 20, 2020, 3:00 AM	2183	49	17	2	8%	2	$ 1,447	$ 777

Tabla 3: Base órdenes

En la Tabla 4, se observan marcadas diferencias en el comportamiento de compra de productos por parte de las farmacias. Se constata que las unidades máximas adquiridas en una sola orden, llegan a 2000, mientras que el mínimo registrado en una orden, es de solo 1 unidad. Además, se aprecian significativas disparidades en los descuentos comerciales. En un extremo, se encuentra una oferta sin descuento alguno para compras mínimas de 1 unidad. Por otro lado, se ofrece un descuento del 50% para una cantidad mínima de 70 unidades.

Estos hallazgos evidencian la implementación de diversas estrategias comerciales por parte del Laboratorio 1 en esta plataforma. Se destaca la presencia de productos con atractivas condiciones comerciales, caracterizados por descuentos elevados, aunque requieren una cantidad considerable de unidades mínimas para acceder a dichos beneficios. Esta táctica tiene como objetivo impulsar la venta de una mayor cantidad de productos. En el promedio de las órdenes, las unidades presentan un descuento del 10% y establecen una compra mínima de 2 unidades, lo que facilita el acceso a los descuentos ofrecidos.

	ID_Farmacia	Fecha_De_Compra	ID_Producto	ID_Marca	ID_Categoría	Unidades	Descuento	Unidades_Minimas_Descuento	Precio_Producto	Facturación_Total
Total	10763	0	25	18	9	499533	21%	15	106	$ 344,880
Max	-	12/01/2023	-	-	-	2000	50%	70	3962	$ 1,428,773
Min	-	19/07/2020	-	-	-	1	0%	1	113	$ 101
Media	-	-	-	-	-	5	10%	2	1340	$ 3,204
Desvación Estándar	-	-	-	-	-	18	5%	2	894	$ 14,650

Tabla 4: Análisis base órdenes

Con estos datos, se analiza para cada una de las 3866 farmacias, la cantidad de compras en valor monetario o cantidad de unidades por producto, por marca, por categoría, por oferta aplicada. De esta manera, se obtiene información variada de las operaciones generadas por las farmacias en la plataforma que luego, se analizan en el modelo estadístico.

2.2 Análisis cuantitativo:

En esta sección, se presenta un análisis estadístico-descriptivo de los datos. Dicho análisis, se lleva a cabo utilizando la información correspondiente a los meses de julio de 2020 a octubre de 2022.[1] El objetivo es observar el comportamiento de compra de las distintas farmacias durante periodos de funcionamiento habitual en la comercialización de productos en la plataforma.

El Gráfico 2 muestra la evolución de las ventas de los 25 productos de la categoría de consumo masivo del laboratorio seleccionado para este análisis, el cual fue elegido por ser uno de los

[1] Los meses de noviembre y diciembre de 2022 y enero y febrero de 2023 se excluyen ya que presentan comportamientos atípicos, caracterizados por un cambio de comportamiento de las farmacias en la plataforma motivado por una regulación sobre los descuentos.

laboratorios con mayor venta dentro de la plataforma. En este gráfico, se observan los valores máximos de unidades vendidas identificados en ciertos meses específicos. Por ejemplo, durante el mes de mayo de 2021, se llevaron a cabo diversas acciones de marketing dirigidas a los consumidores finales por parte de las distintas marcas de consumo masivo. Como resultado de estas estrategias, se generó un crecimiento en la demanda de esas marcas en las farmacias, lo que llevó a las mismas, aumentar su compra de productos con el fin de abastecer la demanda.

Además, se observa un nivel constante de venta de unidades de los 25 productos de consumo masivo, a lo largo de los meses de marzo a diciembre, aunque se presentan ciertas caídas en enero y febrero del año 2022, debido al bajo nivel de consumo en la industria farmacéutica durante esos meses específicos por ser temporada vacacional. Por otro lado, existen meses con altas ventas, como mayo de 2021, impulsadas por eventos comerciales con varias promociones atractivas para las farmacias que estimularon múltiples compras en la plataforma.

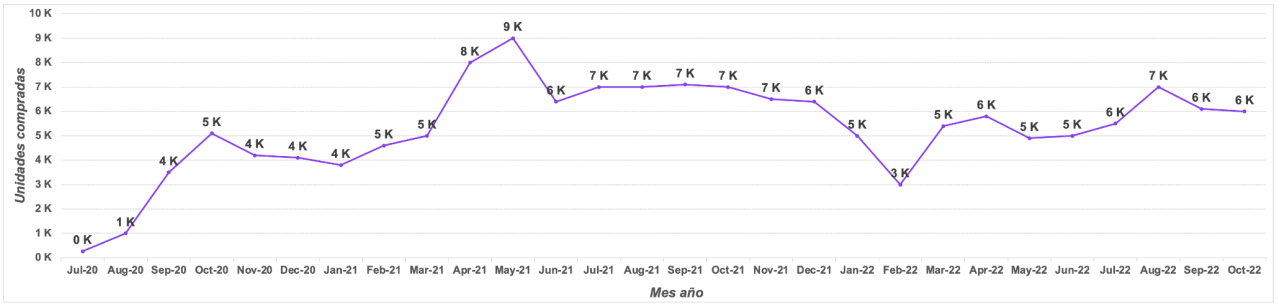

Gráfico 2: Evolución mensual de unidades vendidas

El Gráfico 3 muestra la evolución de las farmacias compradoras únicas en cada uno de los meses. Se aprecia un crecimiento que se inicia en octubre de 2022, seguido de valores constantes en los meses subsiguientes. De forma similar al Gráfico 2, se observan algunas caídas en enero 2022 y febrero 2022 debido a la reducida actividad en la industria farmacéutica argentina durante esos meses. En contraste, en agosto del 2022 se registra un alto número de farmacias compradoras debido a una estrategia dirigida a los consumidores finales de productos de consumo masivo. Esta acción conlleva un aumento en la demanda de los productos por parte de las farmacias, generando un correspondiente incremento en las compras realizadas por ellas en la plataforma.

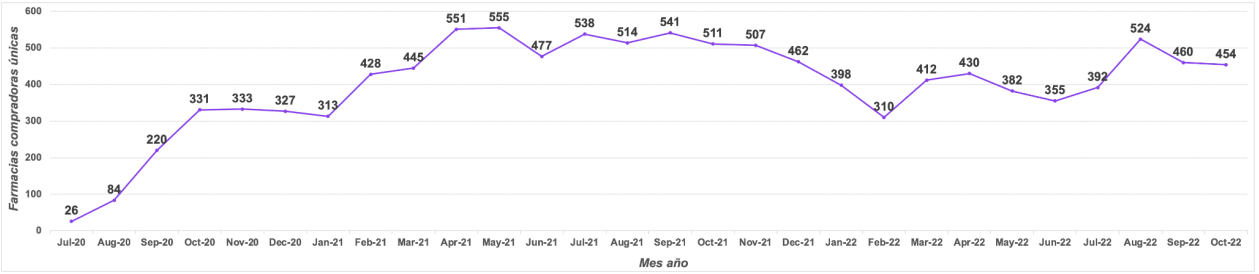

Gráfico 3: Evolución mensual de farmacias compradoras únicas

El Gráfico 4 detalla la distribución porcentual de las ventas semanales de unidades de la plataforma a lo largo de un mes, desde junio hasta octubre de 2022. Se destaca que el máximo de ventas mensuales ocurre en la tercera semana, mientras que la menor actividad se registra en la última semana. Este comportamiento refleja la tendencia de las farmacias a incrementar sus compras en la tercera semana para ajustar sus inventarios antes de enfrentar los procesos administrativos típicos

del cierre de mes en la cuarta semana. Por otro lado, se registran algunas farmacias con un comportamiento de compra diferente, realizando sus compras en la primera semana del mes. La distribución semanal de unidades vendidas se mantiene constante a lo largo de las cuatro semanas. Además, es relevante resaltar un conjunto de farmacias que emplea la plataforma de manera consistente y recurrente, variando desde el uso diario hasta semanal, lo cual resulta en una distribución homogénea de las compras a lo largo del mes. Es importante remarcar que ninguna semana supera el 30% de participación del total de unidades vendidas, evitando así problemas de gran concentración temporal en las compras de la plataforma.

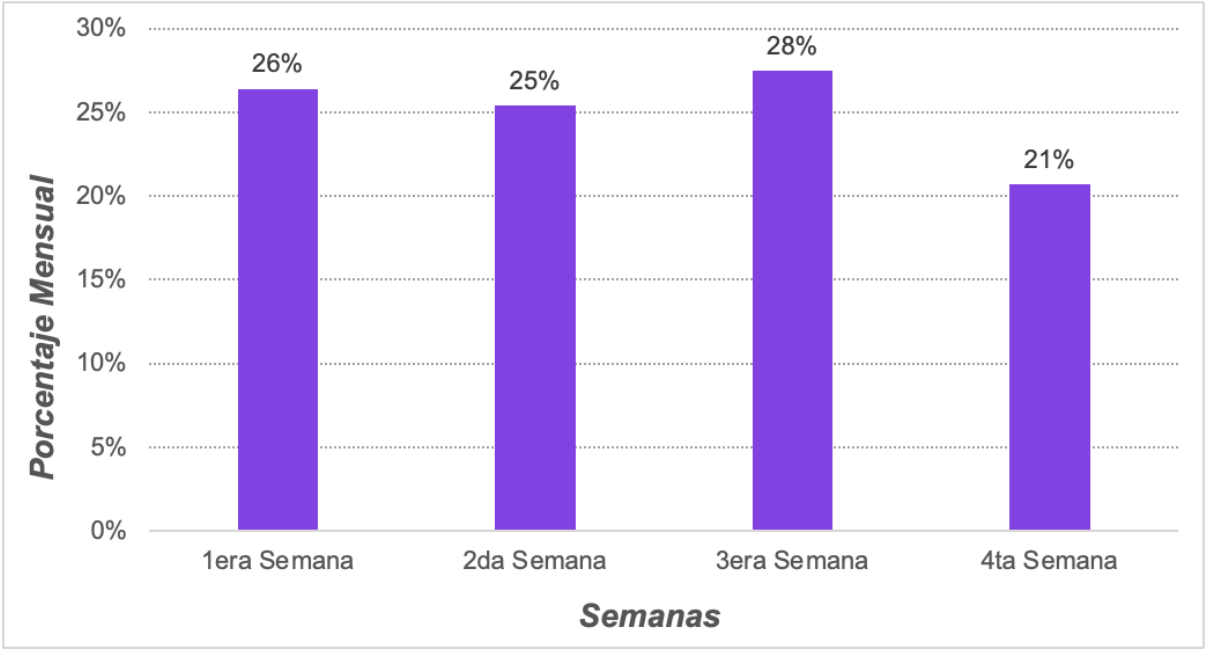

Gráfico 4: Porcentaje de unidades vendidas por semana de los últimos 6 meses

Por último, en el Gráfico 5, se muestra la distribución de las farmacias compradoras únicas mensualmente en septiembre y octubre del 2022. Se observa que un gran número de farmacias adquiere productos con un descuento del 8%. Esta variable no solo se ve influenciada por la cantidad de farmacias que compran con dicho descuento, sino también por la cantidad de productos dentro de la plataforma que disponen un descuento del 8% en comparación con aquellos que ofrecen un descuento del 12%.

Es importante señalar que un mismo producto puede tener diferentes niveles de descuento según la cantidad de unidades compradas. Además, el Laboratorio 1 tiene la capacidad de ofrecer a cada farmacia distintos descuentos por producto. Se destaca la presencia de un grupo significativo de farmacias compradoras que acceden a niveles de descuentos del 5%, 8% y 12%. Este hecho puede sugerir que el laboratorio no ha experimentado con otros porcentajes de descuento, como un 7% o un 11%, por ejemplo.

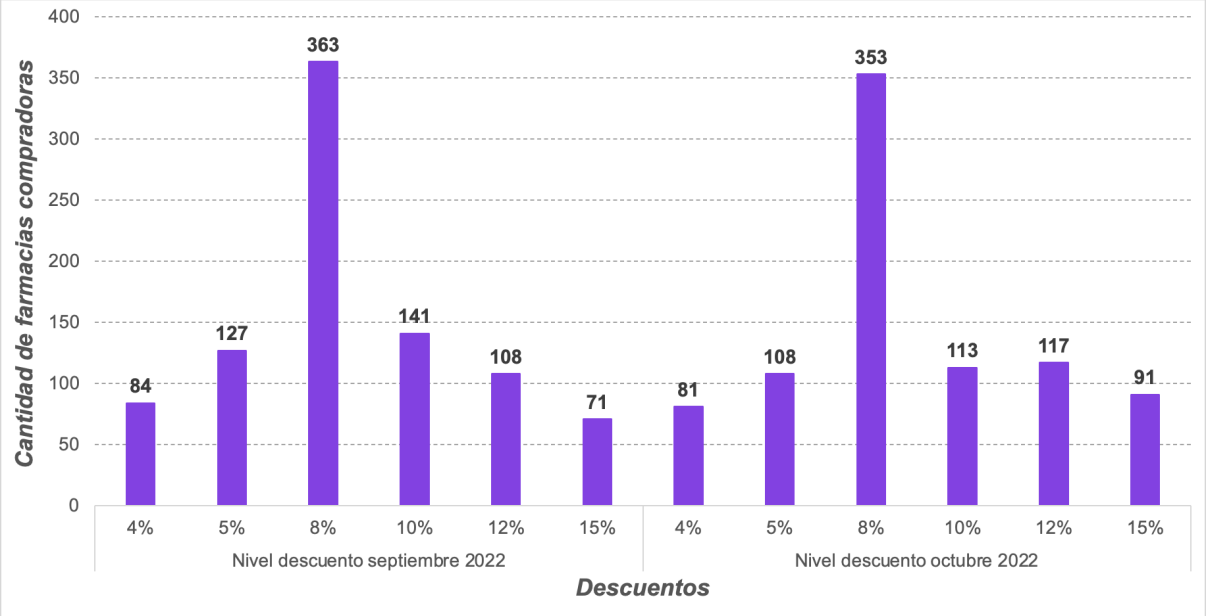

Gráfico 5: Evolución mensual de farmacias compradoras únicas por Descuento (%)

3. Metodología

3.1 Agrupación por k-means

El algoritmo de k-means es un método de aprendizaje no supervisado que se utiliza para agrupar conjuntos de datos en diferentes categorías (Alpaydin, 2010). El proceso comienza por definir un número de grupos (k) y seleccionar k puntos aleatorios con una distribución normal estandar como centros de los grupos. Luego, se asigna cada punto del conjunto de datos al grupo cuyo centro está más cercano, y se calcula nuevamente el centro del grupo en función de los puntos que le han sido asignados (Bishop, 2006). Este proceso se repite hasta que no se produzcan cambios significativos en la asignación de elementos a grupos o hasta que se alcance un número máximo de iteraciones.

En el caso de la predicción de demanda, el algoritmo de k-means puede ser utilizado para agrupar a los usuarios según sus características de consumo, preferencias y comportamientos. De esta forma, el resultado de esa categorización, crea un índice que simplifica los patrones de consumo de una farmacia, y es utilizado como variable a la hora de predecir la demanda de productos con un modelo probabilistico tradicional. Este análisis también permite optimizar las estrategias de pricing y marketing del Laboratorio 1, enfocándose en las necesidades de grupos de farmacias con características similares en lugar de analizar individualmente a cada farmacia de manera manual e intuitiva.

Según Alpaydin (2010), la lógica matemática detrás del algoritmo es la siguiente:

1. Selección inicial de k centroides: el algoritmo comienza seleccionando k centroides de forma aleatoria.

2. Asignación de elementos a centroides: a continuación, cada elemento del conjunto de datos se asigna al centroide más cercano en función de una métrica de distancia, típicamente la distancia euclidiana.

3. Cálculo de nuevos centroides: una vez que todos los puntos han sido asignados a los centroides, se calcula el nuevo centroide para cada grupo como la media aritmética de todos los puntos asignados a ese centroide.

4. Repetición del proceso: se repiten los pasos 2 y 3 hasta que los centroides no cambien significativamente o se alcance un número máximo de iteraciones. En el caso del trabajo se pone un máximo de 20 iteraciones.

El objetivo del algoritmo es minimizar la suma del cuadrado de las distancias entre cada punto y su centroide correspondiente, lo que se conoce como la función objetivo. El algoritmo de k-means tiene varias ventajas para la segmentación de usuarios en el contexto de la predicción de demanda y en este trabajo, se plantea el mismo razonamiento con las farmacias como consumidores, entonces es apropiado pensar que las ventajas que se han observado en el contexto de predicción de demanda en otros ámbitos de estudio, también se pueden ver en este trabajo. Las ventajas de la segmentación a través del método k-means son (Aggarwal, 2015):

1. Agrupamiento no supervisado: el algoritmo no requiere etiquetas previas de los grupos, lo que lo hace útil para encontrar patrones y estructuras ocultas.

2. Flexibilidad: el número de grupos o centroides puede ser ajustado para obtener diferentes niveles de granularidad en la segmentación.

3. Identificación de patrones: el algoritmo puede identificar patrones de "consumos" y "preferencias" de las farmacias que pueden ser utilizados para predecir la demanda futura.

4. Robustez: el algoritmo de k-means puede manejar datos con ruido y valores atípicos, lo que lo hace útil para conjuntos de datos con poca información de venta.

Es importante tener en cuenta que la calidad de los resultados depende de la selección adecuada de los parámetros y de la calidad de los datos disponibles. Cuando hay una cantidad limitada de datos, el uso del algoritmo de k-means puede ser desafiante, ya que el algoritmo se basa en encontrar estructuras y patrones en los datos, lo que puede ser difícil si la cantidad de datos es pequeña. Además, es recomendable utilizar técnicas de preprocesamiento y transformación de datos para mejorar la calidad de los resultados. Algunas técnicas comunes que se utilizan en este trabajo incluyen:

1. Normalización de datos: escalar los datos para que tengan una media de cero y una desviación estándar de uno puede mejorar la eficiencia del algoritmo de k-means. [2]

2. Selección de características: seleccionar solo las características más relevantes y descartar aquellas que no contribuyen significativamente a la segmentación puede mejorar la calidad de los resultados.

[2] Esta técnica es usual en ciencia de datos para mejorar la eficiencia y precisión de los algoritmos. Al escalar y transformar los datos, se logra que estén en un rango común independientemente de la escala original. De esta manera, se busca reducir el impacto de las diferencias en la escala y la magnitud de los atributos de los datos.

La segmentación de datos es una técnica ampliamente utilizada para identificar patrones de comportamiento similares en grupos de observaciones. En este contexto, se utiliza el algoritmo k-means para segmentar las farmacias en función de sus características y comportamiento de compra en la plataforma. En la operación diaria de la plataforma, este modelo es útil para que los responsables de marketing del Laboratorio 1, diseñen estrategias personalizadas para cada segmento, adaptándose a sus necesidades y características específicas, lo que a su vez puede mejorar la experiencia de compra y fidelizar a los clientes.

La segmentación realizada proporciona información valiosa para mejorar la asignación de recursos de marketing y ventas en la plataforma. En este sentido, la variable resultante de la segmentación, se utiliza en un Modelo Probit, en la segunda etapa de este trabajo. El resultado del modelo k-means se convertirá en una variable explicativa para predecir la probabilidad de que una farmacia realice una compra en la plataforma. Con esta propuesta, se espera mejorar la capacidad de predicción del modelo y aumentar las ventas del Laboratorio 1.

A continuación, se proporciona una explicación detallada del algoritmo k-means, cómo se utiliza para llevar a cabo la segmentación de las farmacias en la plataforma y un análisis del resultado obtenido.

Selección de variables e ingeniería de atributos

La ingeniería de atributos o extracción de features es un proceso de selección y transformación de variables relevantes de un conjunto de datos para mejorar el rendimiento de un modelo de aprendizaje automático. Se condiciona el modelo con un algoritmo para aprender patrones específicos en conjuntos de datos y proporcionar información y predicciones a partir de ellos. (Aggarwal, C. C., 2015)

En el contexto del análisis de segmentación de farmacias, se utiliza la ingeniería de atributos para crear nuevas variables basadas en la antigüedad de la farmacia en la plataforma, la frecuencia de compra, el tiempo hasta la primera compra de la farmacia y el promedio de unidades por orden para mejorar la precisión del modelo k-means.

En particular, estas nuevas variables permiten analizar a las farmacias de una manera más completa y detallada, tomando en cuenta su comportamiento en la plataforma. La antigüedad de una farmacia en la plataforma, por ejemplo, indica su grado de experiencia y fidelidad a la plataforma. Por otro lado, el tiempo de conversión aporta una idea del tiempo que tarda una farmacia en realizar una compra y la frecuencia de compra indica el hábito de compra del cliente.

En consecuencia, al agregar estas nuevas variables al análisis, se obtiene una visión más completa y precisa del comportamiento de las farmacias y, por ende, una segmentación más eficiente. Las nuevas variables creadas son:

- Días_Registrado
- Días_Primera_Compra
- Días_Ultima_Compra
- Unidades_Por_Orden

Luego, se agregan otras variables las cuales ya fueron mencionadas en la sección de descripción de datos:

- Ordenes_Totales
- Unidades_Totales
- Facturación_Total
- Ticket_Promedio
- Farmacia_Migración (indicador binario)

Al realizar el análisis cualitativo de la muestra, se observan similitudes y diferencias entre las farmacias registradas en la plataforma. Es evidente y esperable que la muestra de usuarios no sea homogénea, lo que nos permite identificar diversos grupos con diferentes comportamientos y características, enriqueciendo así el estudio.

En la base de datos analizada, una importante proporción de usuarios cuenta con una antigüedad de más de un año. En el Gráfico 6, se observa que a partir de los 400 días de registro, existe una tendencia creciente de farmacias que tienen más de 400 días registrados en la plataforma. Esta tendencia puede atribuirse al alto número de registros durante los primeros meses de lanzamiento de la plataforma, impulsado por campañas exitosas de adquisición de usuarios. Posteriormente, se observa una disminución en los meses subsiguientes debido al cambio de estrategia, pasando de enfocarse en el registro a priorizar la activación de los uspuarios, es decir, a la generación de su primera compra en la plataforma.

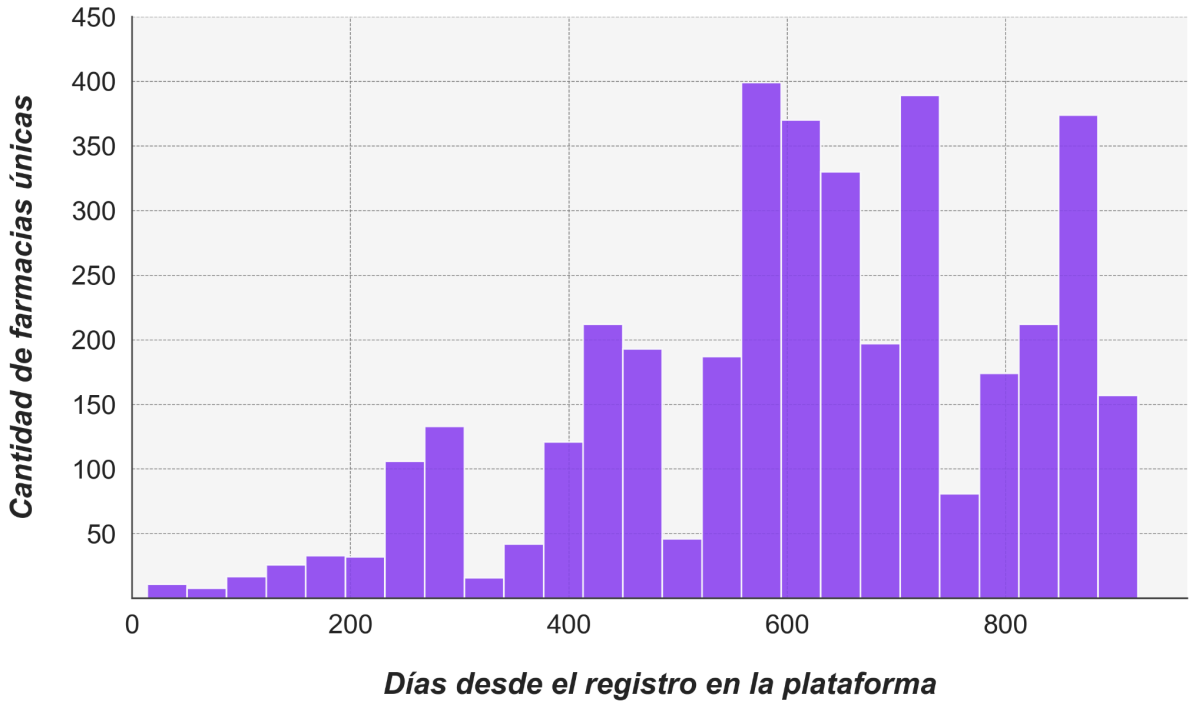

Gráfico 6: Distribución de farmacias únicas por días registradas en la plataforma

En el Gráfico 7, se observa que más de 1750 farmacias generan su primera compra desde su registración en la plataforma, en los primeros 33 días como lo indica la barra de mayor altura en el histograma. No obstante, es importante también analizar aquellas entidades que no tienen este mismo comportamiento. Las razones por las que esto sucede es un punto interesante a profundizar, ya que pueden denotar diferencias en la primera interacción de la farmacia con la plataforma y su interés por los productos y descuentos ofrecidos por la misma.

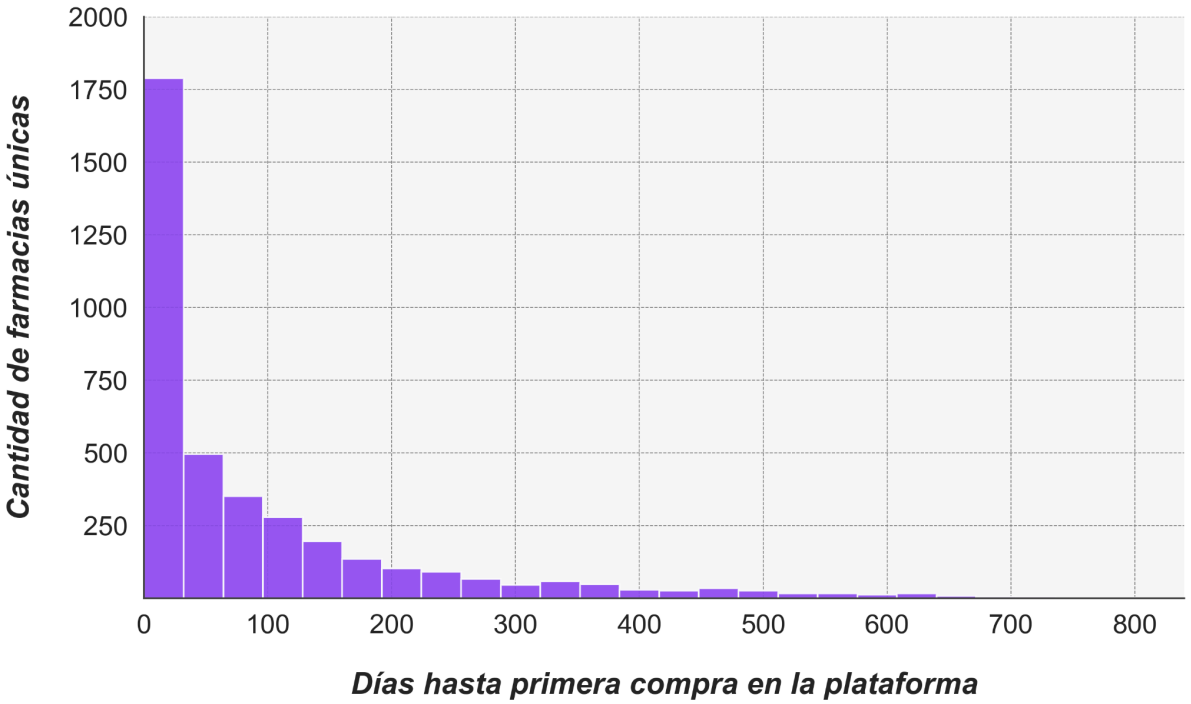

Gráfico 7: Distribución de farmacias únicas por días que pasaron hasta su primera compra

Propuesta de trabajo y resultados del modelo utilizado

1. Preproceso de datos:

La normalización de datos es una técnica común en el preprocesamiento de datos utilizada para escalar los valores de diferentes características de los datos en un rango común. Se realiza para evitar que una característica que tiene valores numéricos grandes o pequeños, tenga más peso en la segmentación que otras características. Esta metodología mejora la eficiencia del algoritmo de k-means en la segmentación de datos, ya que reduce la influencia de las características con valores numéricos grandes y permite que el algoritmo identifique patrones y estructuras más precisas en los datos. Además, al estandarizar, se eliminan las unidades de medida de las distintas variables, lo que hace más sencilla la comparación entre las mismas.

A través de una función logarítmica se ejecuta la normalización. La normalización logarítmica es útil cuando los datos tienen una amplia gama de valores y están distribuidos de manera desigual. Para normalizar los datos utilizando una función logarítmica, se sigue el siguiente proceso:

1. Se calcula el logaritmo de cada valor en el conjunto de datos.
2. Se calcula la media y la desviación estándar de los valores logarítmicos transformados.
3. Se escala cada valor restando la media y dividiendo por la desviación estándar.
4. Los datos normalizados se utilizan para alimentar el algoritmo de k-means.

2. Utilización del método del codo:

La idea básica de los algoritmos de agrupación es la minimización de la varianza intra-grupo y la maximización de la varianza intergrupo. Es decir, se busca que cada observación se encuentre muy cerca a las observaciones de su mismo grupo y los grupos lo más lejos posible entre ellos. El método del codo utiliza la distancia media de las observaciones a su centroide. Es decir, se fija en las distancias intra-grupo. Cuanto más grande es el número de grupos k, la varianza intra-grupo tiende a disminuir. Cuanto menor sea la distancia intra-grupo mejor, ya que significa que los clústers, es decir, los grupos, son más compactos. El método del codo busca el valor k que satisfaga que un incremento de k, no mejore sustancialmente la distancia media intra-cluster. Este es uno de los varios algoritmos utilizados a lo largo del trabajo, para eficientizar la implementación de los modelos utilizados.

El método del codo es una herramienta que permite determinar el número óptimo de grupos en un conjunto de datos utilizando el algoritmo de k-means. El gráfico generado por el método del codo, muestra la relación entre la variación total explicada por el modelo y el número de grupos. El objetivo del algoritmo es encontrar el "codo" en el gráfico, es decir, el punto en el que la adición de más grupos no proporciona una mejora significativa en la variación total explicada. El número de grupos en el codo se considera el número óptimo de grupos para el modelo. Para utilizar el algoritmo del método del codo, se ejecuta el algoritmo de k-means para diferentes valores de k, el número de grupos, y se registra la suma de los cuadrados de las distancias (SSE) de cada punto de datos al centroide de su grupo más cercano. SSE es una medida de la variación total dentro de cada cluster y se puede calcular utilizando la fórmula:

$$SSE = \sum(xi - ci)^2$$

donde xi es un punto de los datos, ci es el centroide de su cluster y $\sum$ representa la suma de todos los puntos de datos en el cluster (Alpaydin, 2010). Los centroides se inicializan en coordenadas aleatorias.

Después de ejecutar el algoritmo de k-means para diferentes valores de k, se grafica el valor de SSE para cada valor de k. El gráfico muestra una curva descendente, ya que agregar más clústeres siempre reduce la SSE. El codo en el gráfico es el punto en el que la disminución en SSE se desacelera significativamente y el modelo comienza a sufrir de sobreajuste.

El método del codo es una técnica útil para determinar el número óptimo de clusters en un conjunto de datos puesto que permite identificar el punto en el que el modelo proporciona una segmentación adecuada sin sobreajuste a los datos. Cuando se utilizan modelos de aprendizaje automático para hacer predicciones, primero se entrena el modelo en un conjunto de datos conocido. Luego, basándose en esta información, el modelo intenta predecir los resultados para los nuevos conjuntos de datos. El sobreajuste es un comportamiento de aprendizaje automático no deseado que se produce cuando el modelo de aprendizaje automático proporciona predicciones precisas para los datos de entrenamiento, pero no para los datos nuevos. Un modelo con sobreajuste puede proporcionar predicciones inexactas. (Alpaydin, 2010).

Los parámetros fundamentales utilizados en el algoritmo de k-means son el número de clusters, el método de iniciación, y la semilla aleatoria. El algoritmo de k-means se itera un total de 20 veces. Recorre valores de k de 1 a 20 donde dentro de cada iteración, se ajusta el modelo de k-means y se calcula la suma de los errores al cuadrado (SSE) para ese valor de k. Por lo tanto, el modelo busca determinar el número óptimo de clusters evaluando el SSE para 20 valores diferentes de k. Además, se define como parámetro el método de iniciación a utilizar, es decir, uno que elija los centroides iniciales de forma eficiente lo que ayuda a la convergencia del algoritmo. En este trabajo se utilizo el

metodo “kmeans++”. Por último, se fija la semilla aleatoria en 1 para garantizar la reproducibilidad de los resultados, lo que asegura que el algoritmo produzca los mismos resultados para la misma entrada de datos y los mismos parámetros.

A continuación, se presenta el gráfico del codo. El algoritmo correspondiente calcula el punto llamado “elbow point” que ocurre cuando la suma de los errores al cuadrado (SSE) comienza a decrecer a una tasa más lenta. En éste caso particular, el algoritmo nos indica que el punto se encuentra en k = 5.

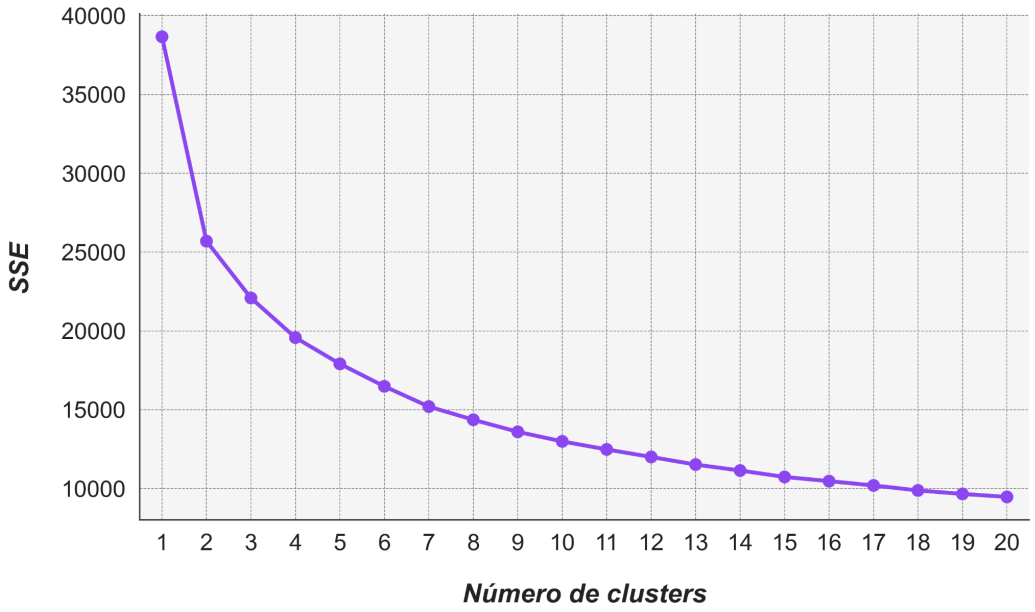

Gráfico 8: Gráfico del codo para selección de número de clusters

3. Reducción de grupos:

Reducir la cantidad de grupos sugerida por el algoritmo del método del codo, puede ayudar a hacer la segmentación más óptima en algunos casos, ya que una mayor cantidad de grupos no siempre se traduce en una mejor segmentación. Si k es demasiado grande, es posible que el algoritmo k-means esté sobre ajustando los datos, creando pequeños grupos sin significado. Por otro lado, si k es demasiado pequeño, el algoritmo podría no ser lo suficientemente expresivo para capturar las estructuras subyacentes en los datos.

Sin embargo, para este trabajo, es relevante tener en cuenta que todo análisis creado va a ser utilizado por los equipos comerciales del Laboratorio 1. La creación de muchos grupos puede dificultar la interpretación de los resultados y hacer que sea complejo tomar decisiones basadas en ellos. Por lo tanto, una segmentación con un número reducido de grupos puede ser fácil de interpretar y utilizar para la toma de decisiones para el laboratorio. Se proponen 4 grupos k, como un número ideal teniendo en cuenta el resultado del método del codo y la práctica usual de la gestión comercial del Laboratorio 1 que tiende a reducir el número de grupos a controlar.

4. Aplicación del modelo de k-means:

Por último, se corre el modelo de k-means con 4 grupos de segmentos para agrupar a cada farmacia en sus diferentes clusters utilizando las variables analizadas anteriormente: Días_Registrado, Días_Primera_Compra, Días_Última_Compra, Ordenes_Totales, Unidades_Totales, Facturación_Total,

Farmacia_Migración. Cada una de las 3866 farmacias, es asignada a uno de los grupos creando la variable Segmento dentro de la base de datos.

Es interesante analizar la distribución de los 4 grupos de farmacias identificados a partir de k-means en términos de las variables “Días_Ultima_Compra” y “Días_Registrados”. Para esto se utiliza un gráfico de dispersión (ver Gráfico 9). Los colores asignados a cada grupo permiten una fácil identificación de los diferentes segmentos y cómo se distribuyen dentro del gráfico. Se puede apreciar que los cuatro grupos se distinguen claramente, con el Grupo 1 (violeta) ubicado en la esquina inferior izquierda, el Grupo 2 (verde) a lo largos de la parte inferior del gráfico, el Grupo 3 (rosa) en la parte superior derecha y el Grupo 4 (naranja) en la parte inferior derecha. Esta visualización permite una fácil identificación de las diferentes características de cada segmento y cómo se relacionan entre sí, lo que ayuda a entender mejor cómo funcionan las diferentes variables. Esto es útil para la toma de decisiones y la segmentación de clientes y permite una mayor focalización de las estrategias de marketing y ventas en base a los 4 grupos identificados.

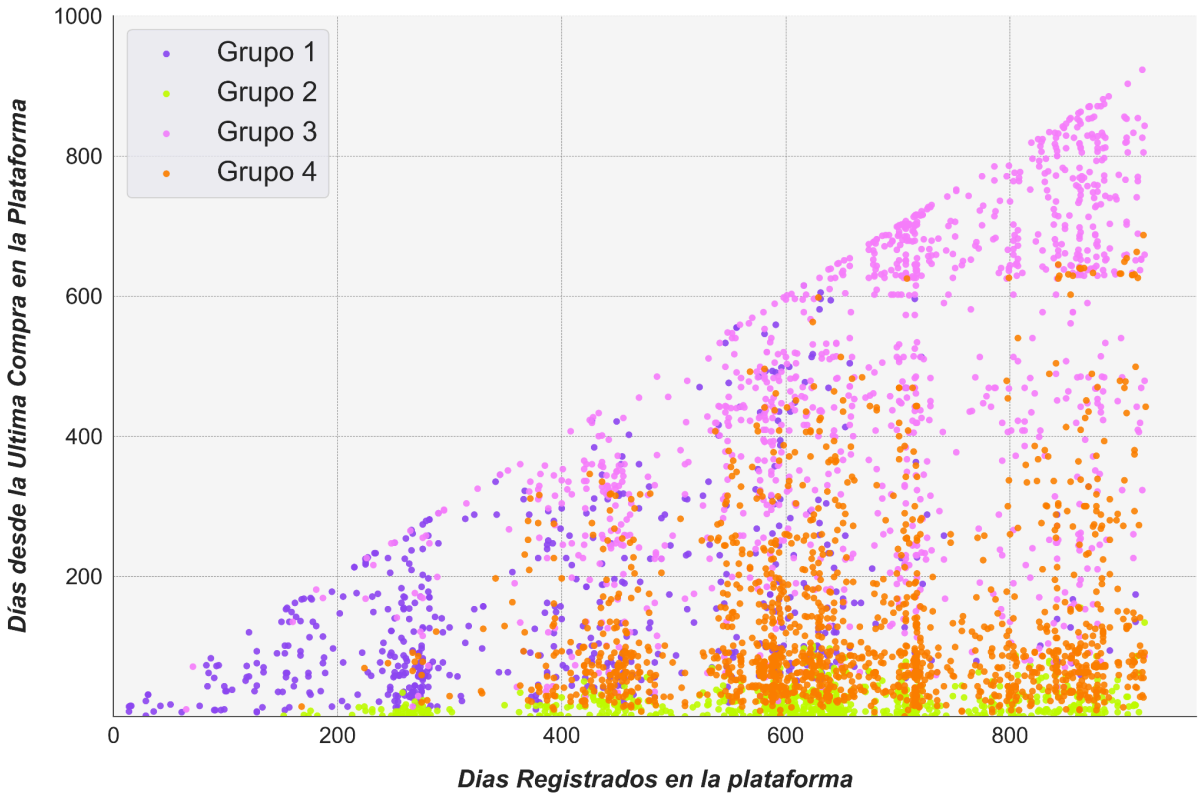

Gráfico 9: Dispersión entre Días Registrados y Días Última Compra

Con el objetivo de analizar los resultados obtenidos y comparar los grupos, se calculan los valores medios de las variables utilizadas en cada grupo. Cada segmento presenta características distintivas en cuanto a su tiempo de registro en la plataforma, la frecuencia de compra, la cantidad de unidades compradas y la facturación generada.

- Grupo 1: farmacias con más de 1 año registradas en la plataforma Son las farmacias con la mayor cantidad de unidades promedio en una orden, generando la segunda facturación total promedio más alta y el ticket promedio más alto.

- Grupo 2: farmacias con más de 1 año y medio registradas en la plataforma, con la mayor frecuencia de compra, siendo las que tienen la mayor facturación total promedio superando los $1.400.000 y el segundo ticket promedio más elevado.

- Grupo 3: farmacias con más de 1 año y medio registradas en la plataforma. Son las que más tardan en realizar su primera compra y presentan menor frecuencia de compra que los primeros 2 grupos, junto con la menor cantidad promedio de unidades totales, facturación total y ticket promedio.

- Grupo 4: farmacias con la mayor antigüedad, con 681 días en la plataforma, con la menor frecuencia de compra. En promedio, hace más de 1 año y 3 meses que no compran en la plataforma, y tienen los niveles más bajos de órdenes totales, de unidades totales, unidades promedio, facturación total y ticket promedio.

A continuación en la Tabla 5, se presenta que resume la media y desvío estándar de cada variable en cada grupo de farmacias.

		Días			Órdenes Totales	Unidades		Facturación Total ($)	Ticket Promedio ($)
		Registrados	Primera Compra	Última Compra		Totales	Por Orden		
Grupo 1	Media	400	54	150	5	639	118	$ 762.872	$ 138.335
	(Desvío Estándar)	(191)	(101)	(127)	(4)	(1.068)	(159)	($ 1.443.870)	($ 219.498)
Grupo 2	Media	610	77	19	20	1.238	71	$ 1.442.220	$ 86.830
	(Desvío Estándar)	(172)	(118)	(20)	(14)	(2.189)	(118)	($ 2.136.170)	($ 123.016)
Grupo 3	Media	661	122	133	9	157	22	$ 134.341	$ 19.951
	(Desvío Estándar)	(150)	(150)	(128)	(8)	(125)	(15)	($ 116.104)	($ 17.202)
Grupo 4	Media	682	95	465	3	25	10	$ 14.445	$ 5.950
	(Desvío Estándar)	(162)	(129)	(223)	(3)	(24)	(9)	($ 14.969)	($ 6.043)

Tabla 5: Estadísticas descriptivas por grupo

En general, se observa que los 4 grupos tienen un tamaño similar de cantidad de farmacias pero con diferencias significativas en los valores promedio de sus variables. El Grupo 4, es el de rendimiento más bajo. Aquí se encuentran las farmacias más antiguas en la plataforma, quienes han tardado más de 464 días en realizar su primera compra. Además, sus métricas promedio en todas las demás áreas son las más bajas entre los grupos. El Grupo 3 representa un nivel superior al Grupo 4, es decir, un mejor comportamiento en la plataforma, pero con áreas de mejora. Si bien su frecuencia de compra es menor que en el Grupo 4 y tiene un alto nivel de órdenes totales, todavía mantiene niveles bajos de unidades totales y facturación total. A medida que ascendemos al Grupo 2, se denota una mejora significativa en ciertos aspectos. Las farmacias pertenecientes a este grupo tienen una gran frecuencia de órdenes, hasta cuatro veces más que en el Grupo 1, aunque con un menor número de unidades por pedido y un ticket promedio más bajo. Sin embargo, la facturación total en este grupo es la más alta siendo 10 veces más grande que la del Grupo 3.

Por último, el Grupo 1 y el Grupo 2 son los grupos que utilizan más frecuentemente la plataforma, pero con comportamientos de compra distintos. El Grupo 1 efectúa órdenes más grandes con el mayor ticket promedio y la mayor cantidad de unidades por pedido con menor frecuencia. Se puede deducir que en el Grupo 1 se encuentran las farmacias con un comportamiento de compra mensual o bimensual de gran cantidad de productos para aumentar el inventario aprovechando los descuentos en la plataforma. En contraste, el Grupo 2 se destaca por su gran frecuencia de órdenes y su gran cantidad de órdenes totales, aunque con menos unidades por orden y un ticket promedio más bajo, lo que muestra un comportamiento de compras chicas diarias o semanales. Ambos grupos se desempeñan excepcionalmente bien pero con hábitos de compra diferentes.

En resumen, a medida que se pasa del Grupo 4 al Grupo 1, las métricas de facturación y unidades compradas varían, mostrando cómo diferentes hábitos de compra pueden llevar a un rendimiento destacado en diferentes aspectos del negocio. Este análisis es fundamental para la creación de estrategias de marketing y ventas de la plataforma, ya que permite identificar los patrones de comportamiento de las farmacias y adaptar la oferta de productos y servicios a sus necesidades y preferencias. Por ejemplo, se identifica claramente que es fundamental tener en cuenta a la hora de comunicar nuevas promociones con el objetivo de aumentar ventas a las farmacias del Grupo 1 y 2. Gracias al resultado obtenido por el modelo, se pueden crear estrategias de descuentos más personalizadas dependiendo del hábito de compra de la farmacia con una comunicación segmentada más efectiva. Por ejemplo, comunicar descuentos diarios únicamente a las farmacias del Grupo 2 y combinaciones de productos para abastecer mensualmente el inventario a las farmacias del Grupo 1.

En conclusión, la segmentación de las farmacias basada en sus características y comportamiento de compra en la plataforma es realizada mediante el uso del algoritmo de k-means. Los resultados obtenidos permiten la identificación de patrones de comportamiento de compra en las farmacias, clasificándolas en 4 segmentos distintos. En este trabajo, la variable resultante de la segmentación, denominada "Segmento", es empleada como variable explicativa en el Modelo Probit de la siguiente sección que busca predecir la probabilidad de que una farmacia realice una compra en la plataforma. La inclusión de esta valiosa información sobre el comportamiento y las características de compra de cada farmacia en el modelo puede mejorar significativamente su capacidad de predicción, lo que a su vez permite una mejor asignación de recursos de marketing y ventas para maximizar las oportunidades de negocio.

3.2 Modelo Probit

En esta sección, se muestra el Modelo Probit que se utiliza para modelar y predecir la probabilidad de que un determinado producto sea comprado en la plataforma. El objetivo es determinar la probabilidad de que una farmacia compre el producto Analgésico A dentro de una orden en la plataforma.

Para este propósito, se emplea un subconjunto de variables de la base de datos mencionada anteriormente en la sección “Data” llamada “base órdenes”. La misma abarca las ventas históricas generadas en la plataforma desde julio del 2020 a febrero del 2023 por las 3.866 farmacias utilizadas en el modelo de k-means para los productos de consumo masivo. Con el objetivo de mejorar la representatividad de la base de datos y aumentar la variabilidad en los datos disponibles, se incluyen las compras de productos bajo receta realizadas por las 3.866 farmacias en el conjunto de datos denominado "base órdenes". Esto enriquece la base de datos al proporcionar una mayor diversidad de productos y transacciones, lo que permite obtener un modelo más robusto y preciso para responder a la pregunta bajo análisis. En contraste, si se hubieran considerado únicamente las órdenes relacionadas con productos de consumo masivo, la variabilidad en los datos habría sido limitada, lo que podría haber afectado negativamente la capacidad del modelo para proporcionar resultados significativos.

Esto genera así una base de datos con 82.990 órdenes diferentes compuesta por 431.175 líneas de productos comercializados para las 3866 farmacias. Cada línea dentro de la base utilizada en el modelo representa una orden que contiene a la variable dependiente y las variables explicativas.

Modelo Lineal

Un modelo de probabilidad lineal es una técnica de regresión utilizada para analizar la relación que existe entre una variable dependiente y una o más variables explicativas. En un modelo binario, la variable dependiente toma dos valores posibles, 0 y 1, que pueden ser asociados a la ocurrencia de un evento (1 si ocurre y 0 si no). Se dispone de una muestra aleatoria de n observaciones donde Y_i, i = 1,...,n, toma el valor 1 si la farmacia compra el Analgésico A dentro de una orden y 0, si no. El subíndice i se refiere a la i-ésima farmacia dentro de la base de datos. Un modelo de elección binaria es un modelo de la probabilidad de que ocurra el evento Y_i condicionado por el conjunto de variables explicativas X_i. Una manera de modelizar esto es a través de una relación lineal entre Y_i y X_i. Donde:

$$Y_i = X_i'\beta + u_i$$

$$\text{con } E(u_i|X_i) = 0.$$

- Y_i: la variable dependiente que puede tomar valor 1 y 0, cuando la farmacia "*i*" compra el Analgésico A, respectivamente.
- X_i': las variables explicativas con información de la farmacia *i*.
- β: valor del coeficiente que mide la importancia de cada variable para explicar la compra del producto en la orden.

Esta especificación lineal presenta un serio problema ya que el modelo lineal no impone ninguna restricción sobre los valores de $X_i'\beta$ y por lo tanto, podría predecir valores negativos o mayores que uno para la probabilidad estimada en base a las variables explicativas. Por ende, se utiliza un modelo de índices transformados como el Modelo Probit.

Modelo Probit

Se debe adoptar un tipo de especificación bajo la cual los valores de P_i están restringidos al intervalo [0,1]. Una forma conveniente de restringir la forma funcional es la siguiente (Wooldridge, J. M. 2003):

$$P_i = F(X_i'\beta)$$

En donde $F(X_i'\beta)$ es una función diferenciable monótona creciente con dominio real y rango [0,1]. Nuestro modelo no-lineal sería el siguiente:

$$Y_i = F(X_i'\beta) + u_i$$

$$\text{con } u_i = Y_i - F(X_i'\beta)$$

El Modelo Probit plantea una función de distribución normal la cual resuelve el problema planteado anteriormente donde:

$$P_i = F(X_i'\beta) = \Phi(X_i'\beta)$$

- $F(X_i'\beta)$ — La probabilidad acumulada de compra del producto en la orden.

- $X'_i\beta$ — El valor obtenido al estimar el Modelo Probit
- $\Phi(X'_i\beta)$ — La función de distribución normal estándar

En resumen, este modelo probabilístico no lineal nos permite predecir la probabilidad de que un producto sea comprado en una orden en función de ciertas variables. A través del análisis de los coeficientes, los valores p y los efectos marginales, se determina la influencia de cada variable en la probabilidad de compra del Analgésico A en una orden. De esta manera, se pueden tomar decisiones informadas en términos de estrategias de marketing, precios y promociones.

Variable dependiente y explicativas

En el modelo se utiliza un subconjunto de las variables mencionadas para analizar el producto Analgésico A y que tiene como identificador único el número "2657" en la base de datos. Este producto es un analgésico antiinflamatorio perteneciente a la categoría consumo masivo que no requiere receta médica para ser vendido en las farmacias. Se ha seleccionado este producto debido a su alta demanda y la posibilidad de modificar su descuento a lo largo del tiempo para fomentar su compra. Al ser uno de los productos más vendidos en la plataforma, su análisis es particularmente interesante para el Laboratorio 1.

Finalmente, se crea la variable dependiente "Compra_2657". A continuación, se detallan las variables que se emplean en el modelo:

Variable dependiente:

- Compra_2657: toma el valor 1 si la farmacia compra el Analgésico A con ID 2657 dentro de una orden y 0, si no.

Las variables explicativas utilizadas en el Modelo Probit son:

- Segmento_i, i = 1,2,3,4: es decir un conjunto de variables dummies que identifica el segmento al cual pertenece la farmacia. La categoría base es el Segmento_4.
- Segmento_i_Unidades, i = 1,2,3,4: interacciones entre cada segmento y la cantidad de unidades compradas.
- Segmento_i_Descuento, i = 1,2,3,4: interacciones entre cada segmento y el descuento.

Normalización de Datos

La normalización de los datos conlleva la generación de una nueva base de datos llamada "subdata", que se emplea para el modelo, representando cada orden como una fila. Si la orden incluye el Analgésico A, las variables relacionadas con dicho producto (Unidades y Descuento) toman los valores asociados. En caso contrario, se calcula el promedio de esas mismas variables para los productos que forman la orden, mientras que las demás variables genéricas de la farmacia permanecen sin cambios.

Esta simplificación del modelo permite que cada fila represente una orden, lo que facilita el cálculo de la probabilidad de una orden específica en la farmacia. Al analizar la variable "Compra_2657" en este nuevo conjunto de datos, se observa que del total de productos comprados:

- La cantidad de órdenes que incluyen el Analgésico A es de 4199.

- La cantidad de órdenes que no incluyen el Analgésico A es de 78782.

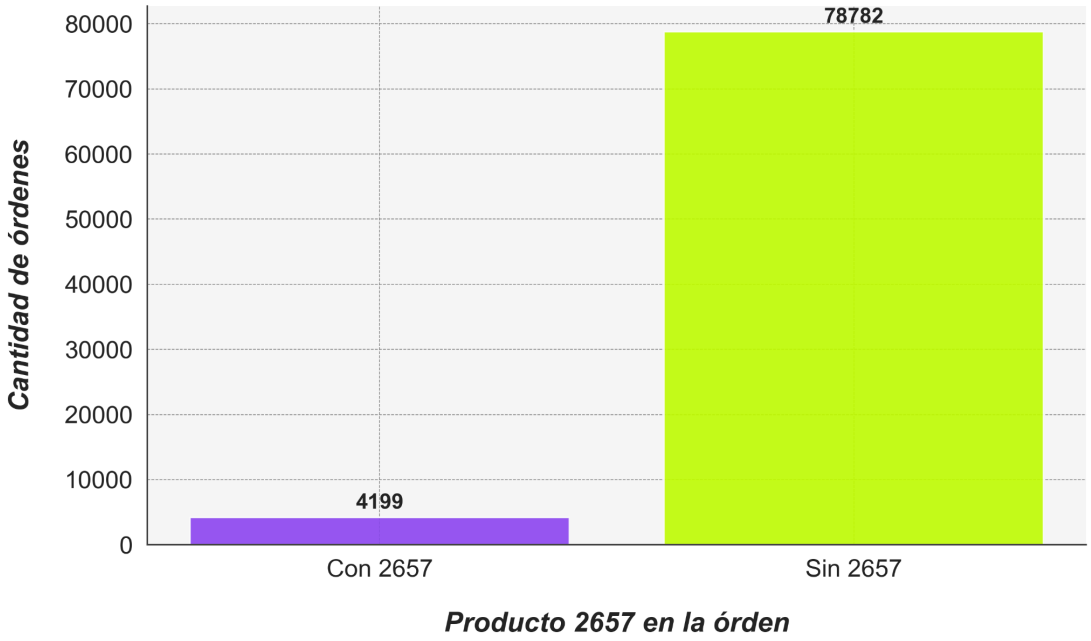

Gráfico 10: Distribución órdenes con Analgésico A

El Gráfico 10 muestra la distribución de las órdenes con Analgésico A, donde se observa que en un 5% de las órdenes se ha comprado el producto, mientras que en un 95%, esto no ha sucedido. Esta discrepancia demuestra el desequilibrio entre las clases de compras, donde se tiene un mayor número de casos en los que el Analgésico A no es comprado en una orden. Por lo tanto, el siguiente paso consiste en hacer una división de la base de datos en entrenamiento y testeo. Finalmente, se procede a balancear los casos usando el algoritmo “SMOTE”.

Muestreo para balancear los casos

En este apartado, se menciona como se ha abordado el hecho que existe un desbalance significativo entre la cantidad de órdenes que contienen el Analgésico A y las que no. Para balancear las clases, se aumenta la cantidad de ejemplos de casos minoritarios. En este contexto, el caso minoritario se refiere a la cantidad de órdenes que contienen el Analgésico A.

En línea con la motivación del presente trabajo, se utiliza el algoritmo SMOTE, también conocido como "Synthetic Minority Over-sampling Technique". El algoritmo SMOTE aborda el problema de desbalance, generando ejemplos "sintéticos" para la clase minoritaria, en este caso, las órdenes de farmacias que contienen el Analgésico A. Para ello, crean de manera aleatoria ejemplos ficticios mediante el algoritmo de “k-nearest neighbours”.

En resumen, SMOTE es una técnica que permite crear ejemplos adicionales para las clases minoritarias, lo cual contribuye a mejorar el equilibrio en el conjunto de datos y evita el sesgo hacia las clases mayoritarias. El proceso del algoritmo se puede describir de la siguiente manera (Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., 2002):

- Elegir de manera aleatoria un ejemplo de la clase minoritaria.
- Calcular la distancia entre ese ejemplo y su vecinos más cercano utilizando el algoritmo de k-nearest neighbours, el cual es un algoritmo de aprendizaje supervisado que en este caso, es utilizado para clasificar diferentes muestras en grupos por la distancia que hay entre las muestras.
- Multiplicar la diferencia de distancia entre el ejemplo y sus vecinos por un número aleatorio entre 0 y 1. Agregar el resultado como una muestra sintética a la clase minoritaria.
- Continuar con el procedimiento hasta generar la cantidad deseada de muestras para la clase minoritaria.

Se utiliza el algoritmo SMOTE y se finaliza con una base de entrenamiento con 105.568 muestras con 52.784 casos con órdenes con la compra del Analgésico A y 52.784 sin la compra del Analgésico A dentro de la orden. De esta manera, se genera una muestra balanceada para ser utilizada en el modelo probabilístico.

Estimación del Modelo

El modelo utiliza las variables explicativas mencionadas en la sección anterior. La Tabla 6 presenta los resultados de la estimación.

	Coeficiente	Error estándar	Estadístico z	Valor P del estadístico z	[0,025	0,975]
const	-2,9342	0,0432	-67,8835	0,0000	-3,0189	-2,8495
Segmento_1	-1,3399	0,058	-23,1022	0,0000	-1,4536	-1,2262
Segmento_1_Unidades	-0,0029	0,0002	-13,4201	0,0000	-0,0034	-0,0025
Segmento_1_Descuento	0,3658	0,0028	128,7491	0,0000	0,3602	0,3713
Segmento_2	0,7742	0,0545	14,1956	0,0000	0,6673	0,8811
Segmento_2_Unidades	0,0245	0,0051	4,823	0,0000	0,0146	0,0345
Segmento_2_Descuento	0,1755	0,0032	54,6457	0,0000	0,1692	0,1818
Segmento_3	0,1233	0,0473	2,6052	0,0092	0,0305	0,2160
Segmento_3_Unidades	0,0032	0,0014	2,3188	0,0204	0,0005	0,0059
Segmento_3_Descuento	0,2474	0,0017	143,9356	0,0000	0,2440	0,2508
Segmento_4_Unidades	0,0010	0,0025	0,3811	0,7031	-0,0040	0,0059
Segmento_4_Descuento	0,2726	0,0035	77,5751	0,0000	0,2657	0,2795

Tabla 6: Resultados del Modelo Probit para Analgésico A

En la Tabla 6, la columna 2 presenta los coeficientes del Modelo Probit. El primer resultado interesante es que las variables explicativas incluidas en el Modelo Probit exhiben coeficientes significativos con valores-p menores a 0,05, con la excepción de la interacción entre segmento 4 y unidades que resulta no significativa. El segmento que representa la categoría base o excluida es el segmento 4. Los resultados indican que pertenecer al segmento 1 reduce la probabilidad de compra en relación a estar en el segmento 4, pero ocurre lo contrario para las farmacias en los segmentos 2 y 3. Es interesante destacar que en contraposición a lo esperado, las farmacias del segmento 1 y 4 tienen un comportamiento similar. La manera usual de interpretar cómo actúan las farmacias del

segmento 1, con un alto nivel de compra y el ticket promedio más alto, es esperar que tengan una mayor probabilidad de compra del producto. Sin embargo, este no es el caso. Este fenómeno puede atribuirse a varias razones. Una posible explicación puede ser que las farmacias con mayor poder adquisitivo tienden a atraer una multitud de ofertas para compras masivas, que vienen con descuentos significativos en productos de consumo masivo, como los analgésicos de diversos laboratorios. Este factor competitivo podría resultar en una baja probabilidad de compra de analgésicos del Laboratorio A a través de la plataforma. Este fenómeno sugiere la necesidad de una estrategia de mercado más sofisticada para aumentar la penetración de productos en estos segmentos.

Por otro lado, salvo para el segmento 1, el hecho de comprar más unidades del producto, aumenta la probabilidad de compra y para todos los segmentos tenemos que mayores descuentos llevan a incrementar la probabilidad de compra, como es de esperar. Dado que este modelo es no lineal, la interpretación de los coeficientes se debe hacer con cautela, ya que las magnitudes de los mismos no tienen la interpretación usual de los modelos lineales. Por lo tanto, es de mayor utilidad analizar los efectos marginales. Estos efectos marginales se calculan en la media, es decir, tomando los valores medios de las otras variables.

	dy/dx	Error estándar	Estadístico z	Valor P del estadístico z	[0,025	0,975]
Segmento_1	-0,5345	0,023	-23.100	0,0000	-0,5800	-0,4890
Segmento_1_Unidades	-0,0012	0,00009	-13.420	0,0000	-0,0010	-0,0010
Segmento_1_Descuento	0,1459	0,001	128.750	0,0000	0,1440	0,1480
Segmento_2	0,3088	0,022	14.196	0,0000	0,2660	0,3510
Segmento_2_Unidades	0,0098	0,002	4.823	0,0000	0,0060	0,0140
Segmento_2_Descuento	0,0700	0,001	54.639	0,0000	0,0680	0,0730
Segmento_3	0,0492	0,019	2.605	0,0090	0,0120	0,0860
Segmento_3_Unidades	0,0013	0,001	2.319	0,0200	0,0000	0,0020
Segmento_3_Descuento	0,0987	0,001	143.850	0,0000	0,0970	0,1000
Segmento_4_Unidades	0,0004	0,001	0,381	0,7030	-0,0020	0,0020
Segmento_4_Descuento	0,1088	0,001	77.583	0,0000	0,1060	0,1110

Tabla 7: Efectos marginales en la media del Modelo Probit para Analgésico A

En la Tabla 7, se observan los efectos marginales. Es interesante resaltar que el impacto del descuento afecta de manera distinta según el segmento al que pertenece la farmacia. Las farmacias que son más sensibles a cambios de los descuentos son las del segmento 1. Mientras que las farmacias que menos reaccionan a cambios en el descuento son las del segmento 2. También, los efectos marginales revelan que la respuesta de las farmacias que demandan más unidades dentro de una orden aumentan la probabilidad de compra independientemente del segmento al que pertenezcan, aunque en distinta magnitud, excepto, por las farmacias del segmento 1 cuya probabilidad de compra se reduce levemente.

El modelo resalta la importancia de mirar las particularidades de las farmacias (resumidas en la variable segmento) a la hora de analizar características que puedan ayudar a la toma de decisión. Es lógico pensar que cuanto más unidades compradas tiene una farmacia sobre otra y mayor ticket promedio tiene, mayor es su probabilidad de compra, aunque esto no ocurre para las farmacias del segmento 1, lo cual es un comportamiento interesante.

Los resultados obtenidos en el Modelo Probit deben ser comparados con las hipótesis generalmente utilizadas en la toma de decisiones en la comercialización de productos. En muchas ocasiones, como mencionamos anteriormente, se tiende a creer que el porcentaje del descuento del producto es el único factor relevante que influye en la decisión de compra. No obstante, este estudio revela que, a través del uso de una segmentación basada en técnicas de aprendizaje no supervisado, como el método k-means, se puede obtener una comprensión más profunda de la probabilidad de compra de un producto en función del segmento al que pertenezca la farmacia.

La combinación del modelo k-means de agrupamiento y el Modelo Probit aporta una perspectiva más completa y granular del comportamiento de compra. El método k-means revela variables significativas sobre las características y comportamientos de las farmacias, las cuales son esenciales para entender cómo la pertenencia a un segmento específico puede aumentar o disminuir la probabilidad de compra de un producto. Este análisis de segmentación provee un marco analítico poderoso para la definición de estrategias comerciales. En conclusión, estos resultados proporcionan un valioso recurso para mejorar las decisiones empresariales, basándose en análisis estadísticos robustos y técnicas de aprendizaje automático. El Modelo Probit complementa el análisis aportando elementos adicionales para la toma de decisiones estratégicas en el campo de la comercialización de productos.

Resultados Modelo Probit con Crema A

A modo de análisis de robustez de los resultados, se realiza una extensión del análisis a un producto adicional dentro de la categoría de consumo masivo, Crema A, identificado con el código 1978 dentro de la plataforma.

En lo que respecta a la Crema A, como se evidencia en la Tabla 8, cada una de las variables explicativas presenta coeficientes estadísticamente significativos, con valores-p inferiores a 0,05. La probabilidad de adquisición de la Crema A por parte de las farmacias, independientemente del segmento al que pertenezcan, únicamente se incrementa con el aumento del descuento. La asociación con cualquier segmento resulta en una disminución de la probabilidad de compra en comparación con la pertenencia al segmento 4. Este es un hallazgo interesante, ya que sugiere que el segmento 4 presenta características que favorecen la probabilidad de compra de la Crema A. Además, se analiza que un aumento en la cantidad de unidades del producto a adquirir no estimula la probabilidad de compra. De hecho, tiene un efecto inverso, reduciendo la probabilidad de compra para cualquier tipo de farmacia.

Este fenómeno podría explicarse en el marco de que, aunque la Crema A es un producto de consumo masivo similar al Analgésico A, es un cosmético, y por ende se enfrenta a un grado de competencia más elevado en el mercado que un analgésico, que se asemeja más a un medicamento de prescripción y cuya demanda por parte de los consumidores es probablemente más inelástica.

	Coeficiente	Error estándar	Estadístico z	Valor P del estadístico z	[0,025	0,975]
const	-0,1053	0,0377	-27,9480	0,0052	-0,1792	-0,0315
Segmento_1	-0,5072	0,0450	-11,2596	0,0000	-0,5954	-0,4189
Segmento_1_Unidades	-0,0361	0,0013	-27,5313	0,0000	-0,0387	-0,0335
Segmento_1_Descuento	0,1276	0,0027	47,9354	0,0000	0,1224	0,1329
Segmento_2	-0,2120	0,0521	-40,6720	0,0000	-0,3142	-0,1098
Segmento_2_Unidades	-0,2483	0,0101	-24,5612	0,0000	-0,2681	-0,2284
Segmento_2_Descuento	0,0791	0,0038	20,7326	0,0000	0,0717	0,0866
Segmento_3	-0,3903	0,0415	-9,4014	0,0000	-0,4717	-0,3089
Segmento_3_Unidades	-0,1083	0,0033	-33,2685	0,0000	-0,1147	-0,1019
Segmento_3_Descuento	0,0858	0,0020	42,8896	0,0000	0,0819	0,0897
Segmento_4_Unidades	-0,0602	0,0041	-14,7883	0,0000	-0,0682	-0,0523
Segmento_4_Descuento	0,0468	0,0041	11,2974	0,0000	0,0387	0,0549

Tabla 8: Resultados del Modelo Probit para Crema A

Un aspecto clave a considerar es el descuento ofrecido a las farmacias. Al analizar los efectos marginales en la Tabla 9, se observa que las farmacias pertenecientes al segmento 1 son las que tienen una mayor probabilidad de comprar el producto al aumentar el descuento ofrecido. Sin embargo, es importante destacar que el aumento en la probabilidad de compra nunca supera el 5%, lo cual indica que la Crema A es un producto de difícil colocación en el mercado, independientemente de la estrategia comercial empleada, debido a la elevada competencia dentro del sector cosmético.

	dy/dx	Error estándar	Estadístico z	Valor P del estadístico z	[0,025	0,975]
Segmento_1	-0,2023	0,0180	-11,2600	0,0000	-0,2370	-0,1670
Segmento_1_Unidades	-0,0144	0,0010	-27,5360	0,0000	-0,0150	-0,0130
Segmento_1_Descuento	0,0509	0,0010	47,9370	0,0000	0,0490	0,0530
Segmento_2	-0,0846	0,0210	-4,0670	0,0000	-0,1250	-0,0440
Segmento_2_Unidades	-0,0990	0,0040	-24,5660	0,0000	-0,1070	-0,0910
Segmento_2_Descuento	0,0316	0,0020	20,7340	0,0000	0,0290	0,0350
Segmento_3	-0,1557	0,0170	-9,4010	0,0000	-0,1880	-0,1230
Segmento_3_Unidades	-0,0432	0,0010	-33,2760	0,0000	-0,0460	-0,0410
Segmento_3_Descuento	0,0342	0,0010	42,8980	0,0000	0,0330	0,0360
Segmento_4_Unidades	-0,0240	0,0020	-14,7890	0,0000	-0,0270	-0,0210
Segmento_4_Descuento	0,0187	0,0020	11,2980	0,0000	0,0150	0,0220

Tabla 9: Efectos marginales en la media (Crema A)

En conclusión, los resultados obtenidos para la Crema A muestran el impacto que tienen las características de este producto y de las farmacias en la variabilidad de la probabilidad de compra. Es importante destacar la importancia de la heterogeneidad entre farmacias, capturada en la variable "segmento", para explicar los distintos patrones de comportamiento observados.

En contraste con los analgésicos, un producto cosmético como la Crema A demuestra que el descuento emerge como el único factor significativo capaz de modificar la probabilidad de compra. Este hallazgo, sin embargo, se encuentra acompañado por un desafío inherente en la

comercialización del producto, señalado por la presencia de un signo negativo en todas las variables de segmentos. La complejidad de esta situación plantea la oportunidad para revisar y reajustar la estrategia de distribución de esfuerzos comerciales, orientándose hacia productos con mayor facilidad de venta a la hora de optimizar el volumen de ventas del portafolio completo de consumo masivo del laboratorio.

Por otro lado, para los equipos comerciales que buscan impulsar las ventas de productos cosméticos como una crema corporal, es crucial enfocarse en estrategias de descuento para incrementar la probabilidad de venta en las farmacias. Aprovechar una estrategia de descuentos diferenciados basada en el segmento de cada farmacia puede ser particularmente efectivo, dada la variabilidad en la sensibilidad al descuento existente entre los distintos segmentos. No obstante, es crucial evitar la dependencia exclusiva en el aumento del descuento para potenciar la probabilidad de compra del producto, como se evidencia en los valores de los efectos marginales para la variable de descuento. Resulta fundamental complementar la estrategia comercial con otras acciones como campañas de comunicación de marca, entre otras, para fortalecer la presencia del producto en el mercado.

Este análisis subraya la diversidad existente entre los productos, proporcionando una metodología para comprender qué variables afectan la probabilidad de compra. Los resultados demuestran la relevancia de la pertenencia a distintos segmentos según el tipo de producto analizado y arrojan luz sobre hallazgos que, desde una perspectiva de negocios convencional, podrían considerarse contradictorios; por ejemplo, la asociación entre pertenecer a un grupo de farmacias con alto nivel de compras y una reducida probabilidad de adquirir el producto. Además, el modelo facilita la creación de estrategias de descuentos adaptadas a cada segmento, teniendo en cuenta la sensibilidad diferenciada al descuento.

Estos resultados no solo validan la eficacia del modelo en la orientación de decisiones comerciales y estratégicas para impulsar la venta de productos dentro de la categoría analgésica del laboratorio, sino que también demuestran su aplicabilidad en un contexto más amplio. Concretamente, establecen un marco metodológico innovador que puede generalizarse al análisis de otros productos de la misma categoría. La capacidad de adaptarse a diferentes contextos de productos aporta un valor añadido al modelo, ya que permite una comprensión más profunda de las dinámicas de compra y ayuda a optimizar las estrategias de marketing para diferentes productos.

El Modelo Probit, complementado con técnicas de aprendizaje no supervisado como el método k-means, ofrece una visión interesante de los patrones de compra, permitiendo identificar los factores que aumentan o disminuyen la probabilidad de adquisición de un producto. Al proporcionar una herramienta para entender la segmentación del mercado y las variables que inciden en la probabilidad de compra, esta metodología apoya una toma de decisiones más informada y basada en datos, esencial para el éxito en el competitivo mercado de consumo masivo.

División de la base de datos en entrenamiento y testeo

La siguiente etapa consiste en la división de la base de datos del modelo en dos partes. Una base de entrenamiento y otra de testeo. La base de entrenamiento debe contener una gran porción de los datos del total de muestras disponibles, en este caso el 66% del total. Este conjunto se utiliza para entrenar el modelo de predicción, lo que significa que aprenderá a partir de estos datos para hacer predicciones futuras de la compra del Analgésico A. Por otro lado, la base de testeo debe contener el

porcentaje restante de datos del total, en este caso el 33% del total. Este conjunto se reserva exclusivamente para evaluar el rendimiento del modelo entrenado. Después de haber entrenado el modelo con la base de entrenamiento, se emplea la base de testeo para realizar las predicciones sobre datos no vistos previamente por el modelo. Al comparar estas predicciones con los valores reales de la muestra de la compra del producto Analgésico A, se evalúa la precisión general del modelo.

La base de testeo, que representa un tercio del total, proporciona una muestra representativa para evaluar el modelo. De esta manera, obtenemos una base de entrenamiento con 70.730 casos y una base de testeo con 34.838, que se utiliza para medir la capacidad de predicción del modelo. Para abordar el desequilibrio de clases, es necesario aumentar la cantidad de ejemplos de los casos minoritarios, en este caso, las órdenes que contienen el Analgésico A en la base de entrenamiento. El desequilibrio de clases puede afectar negativamente el rendimiento del modelo, ya que los algoritmos de aprendizaje automático pueden estar sesgados hacia la clase mayoritaria, en este caso el de la no compra del Analgésico A, y no aprender adecuadamente los patrones de la clase minoritaria. Esto puede conducir a un modelo poco preciso y sesgado hacia la clase dominante.

Es importante tener en cuenta que el modelo de k-means se realiza sobre la base total de farmacias de la plataforma. Una vez completado el clustering, se utiliza la base de órdenes históricas de las farmacias segmentadas para construir el Modelo Probit. Este modelo utiliza la base de datos de las farmacias y se divide en conjuntos de entrenamiento y prueba. Aunque no se incorporan muestras nuevas durante el proceso de modelado es relevante considerar que los clusters pueden cambiar con la adición de nuevas farmacias, lo que podría afectar el modelo. Sin embargo, en este caso, el enfoque se centra en utilizar la base de datos históricos de manera eficiente para construir y evaluar el Modelo Probit. Es cierto que la incorporación de nuevas farmacias podría afectar la clusterización y por ende, el análisis subsiguiente. Una alternativa sería dividir la base de farmacias en entrenamiento y testeo. De esta manera, con la base de entrenamiento se realizaría el clustering y luego, con la base de testeo, se asignaría la farmacia al cluster más cercano.[3] En principio, los resultados no cambiarían de manera significativa pero un análisis más profundo excede el objetivo de esta tesis y se dejará para futuras investigaciones.

Modelo Probit con la base de datos de entrenamiento

El modelo estadístico consiste en encontrar los valores óptimos de los coeficientes del modelo que mejor se ajusten a los datos de entrenamiento. Se utiliza el algoritmo de máxima verosimilitud (Maximum Likelihood Estimation, MLE). El algoritmo busca encontrar los valores de los coeficientes que maximizan la función de verosimilitud, que representa la probabilidad de observar los datos reales bajo el Modelo Probit. El proceso de ajuste del Modelo Probit implica iterativamente probar diferentes valores para los coeficientes y ajustarlos para maximizar la función de verosimilitud. Una vez que se encuentra el conjunto óptimo de coeficientes, el algoritmo proporciona los resultados, incluyendo los coeficientes estimados, errores estándar, valores p y otras estadísticas relevantes para interpretar y evaluar el modelo que se encuentran en la Tabla 10. Los valores p indican la mínima probabilidad requerida para que el coeficiente sea igual a cero, por lo tanto valores p más bajos indican coeficientes más significativos. En contraste con los resultados del modelo con la base entera, se observan valores diferentes en los coeficientes, los errores estándar y los estadísticos z “lo que

[3] Agradezco esta valiosa sugerencia realizada por uno de los revisores.

garantiza mejores estimaciones de los parámetros porque su estructura matemática se ajusta al comportamiento real de los datos."(Gómez-Mejia, 2022).

	Coeficiente	Error estándar	Estadístico z	Valor P del estadístico z	[0,025	0,975]
const	-2,9420	0,0540	-54,2900	0,0000	-3,0480	-2,8360
Segmento_1	-1,3805	0,0720	-19,2690	0,0000	-1,5210	-1,2400
Segmento_1_Unidades	-0,0031	0,0000	-11,9120	0,0000	-0,0040	-0,0030
Segmento_1_Descuento	0,3690	0,0030	107,5510	0,0000	0,3620	0,3760
Segmento_2	0,7022	0,0680	10,2560	0,0000	0,5680	0,8360
Segmento_2_Unidades	0,0282	0,0070	4,2970	0,0000	0,0150	0,0410
Segmento_2_Descuento	0,1837	0,0040	45,0950	0,0000	0,1760	0,1920
Segmento_3	0,1438	0,0590	2,4350	0,0150	0,0280	0,2600
Segmento_3_Unidades	0,0043	0,0020	2,5200	0,0120	0,0010	0,0080
Segmento_3_Descuento	0,2466	0,0020	117,5160	0,0000	0,2420	0,2510
Segmento_4_Unidades	-0,0011	0,0030	-0,3650	0,7150	-0,0070	0,0050
Segmento_4_Descuento	0,2735	0,0040	61,7920	0,0000	0,2650	0,2820

Tabla 10: Resultados del Modelo Probit para el Analgésico A con la base de entrenamiento

Predicción con la base de datos de testeo

Una vez entrenado el modelo con la base de entrenamiento, se utiliza el modelo creado para predecir la probabilidad de que se compre el Analgésico A dentro de una orden. Para ello se utilizan los coeficientes que se calcularon previamente a través del modelo que fue entrenado por la base de entrenamiento y se aplica a los valores de la base de testeo.

La predicción de la probabilidad de compra utilizando el Modelo Probit se basa en el cálculo de un índice lineal, $X'_i\beta$, para cada observación. Las variables son las que indican a qué segmentos pertenece cada farmacia y las respectivas interacciones con las variables "descuento" y "cantidad de unidades en la orden".

El primer paso consiste en obtener el índice lineal. Una vez que se ha calculado el índice lineal para cada observación, se utiliza la función normal estándar para obtener la probabilidad predicha de compra. Estas probabilidades se reportan en la Tabla 11. Por ejemplo, para una farmacia que pertenece al Segmento 1 y en una orden genera una compra de 9 unidades con un descuento del 8% sobre el Analgésico A, el modelo predice una probabilidad de compra del 96% (ver la fila 2 de la Tabla 11). Estos resultados son utilizados para evaluar el rendimiento del modelo a través de calcular la matriz de confusión, los valores de precisión, recall, f-1 score, support y la curva ROC.

const	Segmento_1	Segmento_1_Unidades	Segmento_1_Descuento	...	Segmento_4_Unidades	Segmento_4_Descuento	Predicción
1	1	9	8%	...	0	0%	9%
1	1	17	16%	...	0	0%	96%
1	1	2	5%	...	0	0%	1%
1	0	0	0%	...	5	9%	97%

Tabla 11: Predicciones del Modelo Probit (Analgésico A)

Exactitud de la muestra de testeo

El propósito del Modelo Probit es determinar la probabilidad de compra del Analgésico A para todas las farmacias en el conjunto de pruebas. Después de aplicar el modelo a la base de testeo, se ha obtenido una precisión del 87%. Esta precisión se refiere a la capacidad del modelo para predecir correctamente la probabilidad de compra en un 87% de los casos donde las farmacias realmente compraron el Analgésico A. En otras palabras, al evaluar el rendimiento del modelo sobre el conjunto de testeo, se compararon las probabilidades de compra predichas por el modelo con los resultados reales de compra.

Matriz de Confusión

La matriz de confusión es una herramienta útil para evaluar el rendimiento de un modelo. En este caso, se comparan las predicciones del Modelo Probit con los resultados reales de compras históricas del Analgésico A en la base de testeo.

El Gráfico 11 muestra cuatro posibles resultados de las predicciones del modelo en función de la realidad de las compras de las farmacias. Los resultados se dividen en cuatro categorías:

- Verdaderos Positivos (VP): Estas son las farmacias que compraron el Analgésico A y el modelo predijo acertadamente que lo harían. El Gráfico 11 muestra 15.589 verdaderos positivos.
- Verdaderos Negativos (VN): Representa las farmacias que realmente no compraron el Analgésico A y el modelo predijo correctamente que no lo comprarían. En el caso mencionado, el Gráfico 11 muestra que hay 14.650 verdaderos negativos.
- Falsos Positivos (FP): Estas son las farmacias que en realidad no compraron el Analgésico A, pero el modelo predijo erróneamente que sí lo comprarían. En el Gráfico 11, se registran 2769 falsos positivos.
- Falsos Negativos (FN): Corresponde a las farmacias que compraron el Analgésico A, pero el modelo predijo de manera incorrecta que no lo harían. Según el Gráfico 11, se tienen 1830 falsos negativos.

En base a estos resultados, se puede concluir que:

- La cantidad de verdaderos positivos y verdaderos negativos (15.589 y 14.650 respectivamente) es alta en comparación con los falsos positivos y falsos negativos (2769 y 1830 respectivamente). Esto indica que el modelo tiene una buena capacidad para predecir correctamente tanto las farmacias que comprarán el Analgésico A como las que no lo comprarán.
- La precisión del modelo (calculada como la suma de los verdaderos positivos y verdaderos negativos dividida por el total de predicciones) es alta debido a la gran cantidad de aciertos.

		Predicción	
		Positivo	**Negativo**
Observación	**Positivo**	15589	1830
	Negativo	2769	14650

Gráfico 11: Matriz de confusión

En resumen, la matriz de confusión proporciona una visión del desempeño del Modelo Probit, permitiendo analizar su capacidad para predecir correctamente las compras del Analgésico A en las farmacias del conjunto de testeo.

Precisión computada, exhaustividad, valor-f y soporte

La Tabla 12 proporciona información detallada sobre el rendimiento del modelo predictivo. Se calcula la precisión, la recuperación (también conocida como sensibilidad), el valor-f y el soporte para cada clase de la variable de respuesta (0 para farmacias que no compran el Analgésico A, 1 para farmacias que compran el Analgésico A). La precisión es la proporción de verdaderos positivos sobre verdaderos positivos más falsos positivos, es decir, la tasa de instancias clasificadas correctamente de una clase en particular en relación con todas las instancias clasificadas en esa clase. La recuperación es la proporción de verdaderos positivos sobre verdaderos positivos más falsos negativos, es decir, la tasa de instancias positivas que se identifican correctamente. El valor-f es una medida combinada de la precisión y la recuperación, que se calcula como la media armónica de ambas medidas. El soporte es el número de instancias de cada clase en el conjunto de prueba.

Los resultados muestran que la precisión, la recuperación y el valor-f son más altas para la clase 1 (farmacias que compran el Analgésico A), lo que indica que el modelo tiene una mejor capacidad para detectar farmacias que compran el Analgésico A que para detectar aquellas que no lo hacen. Además, la precisión global del modelo es del 87%, lo que indica que el modelo es preciso en su trabajo de clasificación.

	Precisión	Exhaustividad	Valor-F	Soporte
0	0,85	0,89	0,87	17419
1	0,89	0,84	0,86	17419
Exactitud			0,87	34838
Promedio macro	0,87	0,87	0,87	34838
Promedio de peso	0,87	0,87	0,87	34838

Tabla 12: Valores de precisión, exhaustividad, valor-f y soporte (Analgésico A)

Curva ROC

El gráfico ROC (Receiver Operating Characteristic) es una representación visual del rendimiento de un modelo de clasificación binaria. En el Gráfico 12, la curva violeta representa la tasa de verdaderos positivos (TPR) en función de la tasa de falsos positivos (FPR) en diferentes umbrales de probabilidad para clasificar la observación en la clase positiva. La línea rosa representa la línea base, que se obtiene si se clasifica aleatoriamente la observación en la clase positiva o negativa.

La curva ROC muestra como el modelo está realizando la tarea de clasificación a diferentes niveles de sensibilidad y especificidad. La sensibilidad se refiere a la tasa de verdaderos positivos, es decir, la proporción de observaciones de la clase positiva que se clasifican correctamente como positivas. La especificidad se refiere a la tasa de verdaderos negativos, es decir, la proporción de observaciones de la clase negativa que se clasifican correctamente como negativas.

Un modelo de clasificación perfecto tendría una curva ROC que se ajustaría completamente al borde superior izquierdo del gráfico, lo que significa que tendría una tasa de verdaderos positivos del 100% y una tasa de falsos positivos del 0% para todos los umbrales de probabilidad. En la práctica, cuanto más se aleje la curva ROC de la línea base, mejor será el rendimiento del modelo.

El valor del área bajo la curva (AUC) es una medida de la capacidad de discriminación del modelo. El valor del AUC varía entre 0 y 1, donde un valor de 0,5 indica que el modelo no es mejor que una clasificación aleatoria y un valor de 1 indica una clasificación perfecta. En este caso, el valor del AUC es 0,87, lo que indica que el modelo tiene una buena capacidad de discriminación.

En resumen, la curva ROC y el valor del AUC proporcionan una evaluación adicional del rendimiento del modelo de clasificación binaria y ayudan a seleccionar el umbral de probabilidad adecuado para clasificar las observaciones en una clase determinada.

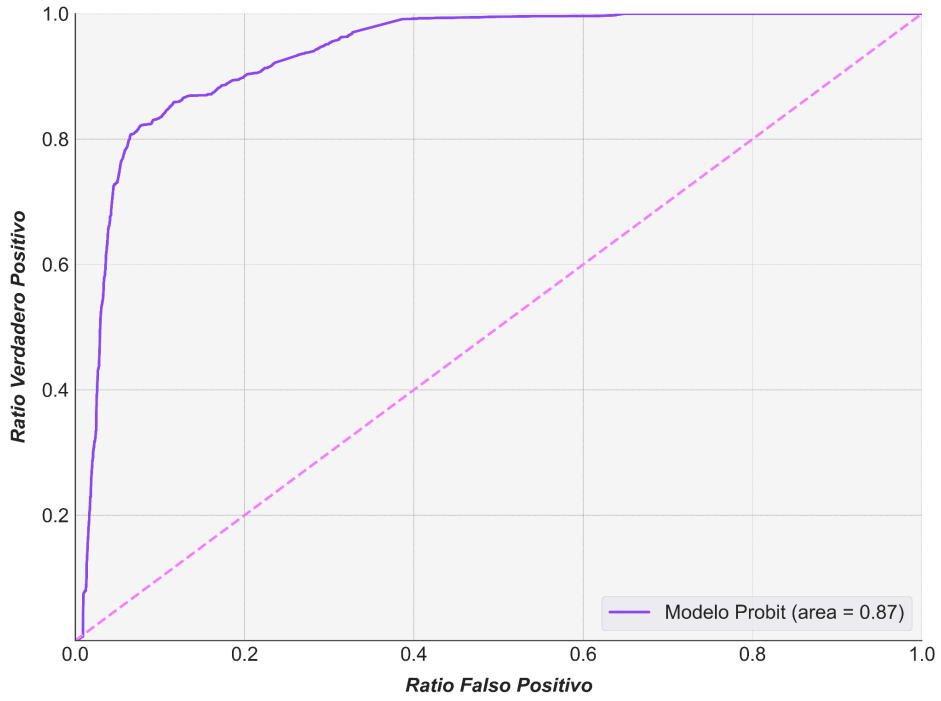

Gráfico 12: Curva ROC

4. Conclusiones

Este trabajo muestra cómo puede utilizarse un modelo probabilístico para predecir la probabilidad de compra de un producto en una plataforma electrónica y entender cuáles son las variables que pueden explicar estadísticamente ese suceso. Esto permite, entre otras cosas, que se puedan tomar decisiones basadas en datos para mejorar la rentabilidad y reducir costos comerciales de los laboratorios que operan en la plataforma. La propuesta de valor desarrollada en esta tesis tiene un gran potencial para mejorar la eficiencia de las estrategias de ventas en la industria farmacéutica y contribuir al crecimiento sostenible de los negocios en este sector.

El presente trabajo ilustra cómo se deben combinar técnicas de machine learning junto con modelos clásicos de probabilidades, como el Modelo Probit, para entender y analizar el comportamiento de las farmacias a la hora de usar una plataforma electrónica para comprar productos de consumo masivo. Para ello, se desarrolla un modelo estadístico que utiliza las ventas históricas de las farmacias para comprender cuáles son los factores que influyen en la probabilidad de que un usuario compre o no un producto en la plataforma, lo que ayuda a mitigar decisiones sin fundamentos empíricos.

La metodología consta de dos pasos: en el primer paso se agrupan las farmacias según sus características mediante un modelo de k-means, y en el segundo, se utiliza un Modelo Probit para predecir la probabilidad de demanda de un producto específico. Se han utilizado diferentes técnicas de machine learning, como el algoritmo SMOTE y k-nearest neighbours, para optimizar el modelo, creando, transformando y seleccionando las variables más significativas. El problema que se aborda en esta tesis es la ineficiencia de las estrategias de ventas de un laboratorio en la industria farmacéutica argentina. Un ejemplo claro es el plan de acción de ofrecer un mismo descuento fijo a todas las farmacias sin distinción alguna. Otro ejemplo de esta ineficiencia, es el alto costo de inversión en canales de comunicación donde se promocionan productos a todo el universo de farmacias de la plataforma, tratando a todas por igual. Estas acciones acompañadas con información valiosa como la segmentación de farmacias, el análisis de los coeficientes de las variables explicativas del Modelo Probit y la probabilidad de compra del producto, pueden llevar a resultados más atractivos para el laboratorio. Esta problemática ha sido evidente en productos de consumo masivo, como un analgésico y una crema corporal, donde existe una gran incertidumbre sobre cuáles son las farmacias más propensas a comprar estos productos y qué debe tener en cuenta el equipo de negocios del laboratorio para incentivar sus ventas. La propuesta de valor desarrollada en esta tesis ofrece una primera solución a este problema al permitir a los operadores de la plataforma y laboratorios tomar decisiones basadas en datos para mejorar sus ventas.

En conclusión, esta tesis de maestría demuestra el potencial de la propuesta desarrollada para mejorar la eficiencia de las estrategias de ventas en la industria farmacéutica y contribuir al crecimiento sostenible de los negocios en este sector. La aplicación del Modelo estadístico Probit brinda una ventaja competitiva al proporcionar una comprensión sólida de los factores que influyen en las decisiones de compra.

Es importante reconocer las limitaciones de este estudio. Primero, el análisis se centra en solo productos de consumo masivo (Analgésico A y Crema A) y un solo laboratorio participante en la plataforma. Los resultados pueden variar para otros productos y laboratorios. Además, aunque el Modelo Probit muestra una alta precisión en el conjunto de prueba, existen posibles sesgos y limitaciones en la muestra que deben ser considerados.

Para futuras investigaciones, se sugiere ampliar el análisis a otros productos y laboratorios, con el objetivo de obtener una visión más completa del comportamiento de las farmacias en la plataforma. Además, sería valioso explorar otras técnicas de aprendizaje automático y comparar sus resultados con el Modelo Probit utilizado en este estudio.

Referencias

- Aggarwal, C. C. (2015). Data mining: The textbook. Springer.
- Alpaydin, E. (2010). Introduction to machine learning. 2nd Edition, MIT Press, Cambridge.
- Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357.
- CILFA (Cámara Industrial de Laboratorios Farmacéuticos Argentinos). (2022). La industria farmacéutica argentina 2022. https://cilfa.org.ar/wp1/wp-content/uploads/2022/07/CILFA-La-industria-farmaceutica-argentina-2022.pdf
- Confederación Farmacéutica Argentina. (2022). Transfers SIAFAR: Acciones para mejorar la rentabilidad de las farmacias. https://www.cofa.org.ar/?p=40631
- Gómez-Mejia Alberto. (2020). Modelo de máxima verosimilitud. Libre Empresa vol. 17, No. 2.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
- Subsecretaría de Programación Microeconómica (2018). Industria Farmacéutica Argentina. https://www.argentina.gob.ar/sites/default/files/sspmicro_cadenas_de_valor_farmacia_0.pdf
- Subsecretaría de Programación Regional y Sectorial (2022). Industria Farmacéutica Argentina. https://www.argentina.gob.ar/sites/default/files/industria_farmaceutica_-_version_web_febrero_2022.pdf
- Wooldridge, J. M. (2003). Introductory Econometrics: A Modern Approach, ed. South-Western, 2nd edition.

Apéndice

Eliminación Recursiva de Características

A los efectos de estimar un modelo parsimonioso se podría implementar un algoritmo de optimización que identifica el subconjunto de variables dentro de todas las variables explicativas del modelo con mayor rendimiento. Esto se conoce como eliminación recursiva de características. Este es otro ejemplo de cómo en este trabajo se podrían emplear diversos algoritmos para eficientizar el análisis de las compras en las plataformas.

El objetivo principal de la eliminación recursiva de características es reducir la complejidad y mejorar el rendimiento del modelo de predicción. Este método funciona seleccionando atributos de forma recursiva, comenzando con todas las variables explicativas del modelo y progresivamente eliminando las menos importantes.

El proceso se inicia entrenando el estimador, que en este caso es un modelo de regresión logística, con todas las variables explicativas. Luego, se calcula la importancia de cada característica mediante algún atributo específico provisto por el estimador. Las variables menos relevantes se identifican y se eliminan del conjunto actual. El algoritmo se repite iterativamente, ajustando el modelo en cada interacción con el conjunto reducido de características, hasta que se alcanza el número deseado de características seleccionadas. En este caso, al tener un número reducido de variables, no es necesario utilizar este algoritmo.

Matriz de Correlación de Variables

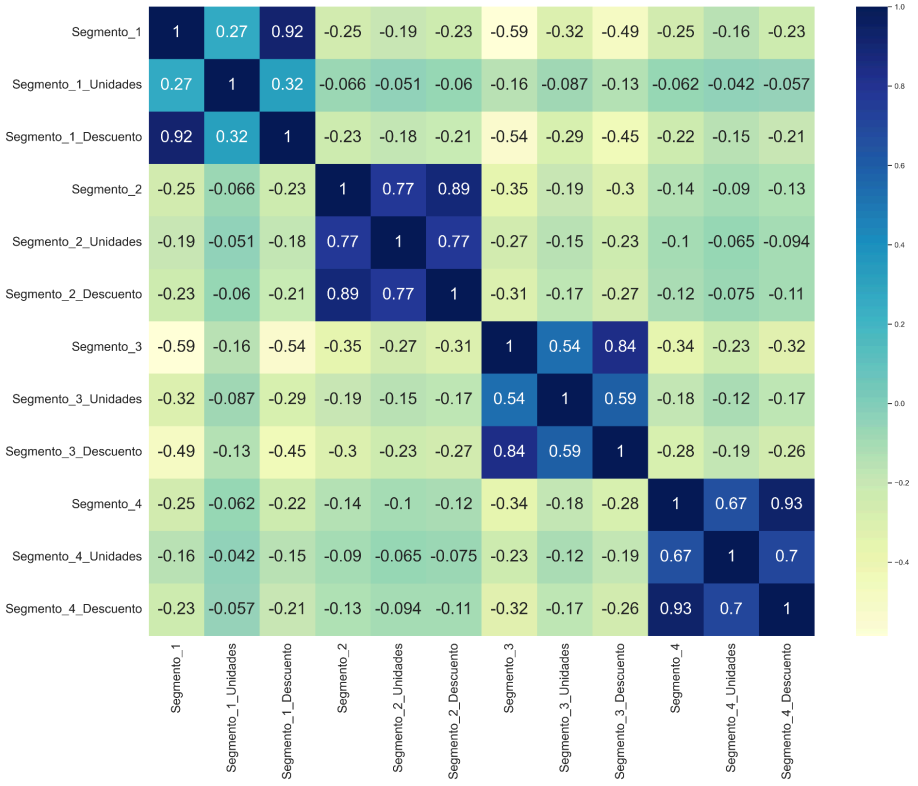

Gráfico 12: Matriz de Correlación Variables