ANALÍTICA DE VARIABLES ASOCIADAS A LA GENERACIÓN DE RECLAMOS EN LA DISTRIBUCIÓN DIRECTA.

Presentado por:
Cristian Felipe Albarracín Riveros
Ingeniero Industrial

Director:

PhD Gonzalo Enrique Mejía

MAESTRÍA EN DISEÑO Y GESTIÓN DE PROCESOS
ÉNFASIS EN SISTEMAS LOGÍSTICOS, MODALIDAD PROFUNDIZACIÓN
FACULTAD DE INGENIERÍA
Bogotá, 2021

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

RESUMEN

Uno de los actores principales en el modelo de venta directa son las promotoras comerciales. Las empresas dedicadas a este modelo prestan especial atención a las reclamaciones por productos faltantes posterior al proceso de entrega de sus órdenes de pedido. Identificar las variables que ocasionan estas reclamaciones por parte de la promotora permiten obtener un activo valioso y competitivo, de donde se podrán generar análisis y acciones de analítica predictiva para evitar el reclamo y mejorar el nivel de servicio. Esta investigación presenta un panorama general de la logística del proceso de entrega de los productos a la promotora comercial en una empresa de venta directa y propone un modelo predictivo de clasificación supervisada para encontrar las futuras promotoras reclamantes. La implementación de este modelo permitió identificar que los días disponibles para venta es la variable protagonista en el comportamiento de la promotora y genera información beneficiosa para mitigar los reclamos y los costos que estos conllevan.

ABSTRACT

One of the main actors in the direct sales model is commercial promoters. The companies dedicated to this model take special attention to claims for missing products after the delivery process of their order forms. Identifying the variables that cause these claims allows obtaining a valuable and competitive asset, from which analyzes, and predictive analytical actions can be generated to avoid the claim and improve the level of service. This research presents an overview of the logistics of the product delivery process to the commercial promoter at the company and proposes a predictive model of supervised classification to find future claimant promoters. The implementation of this model allowed us to identify that the days available for sale is the leading variable in the behavior of the developer and generates beneficial information to mitigate the claims and the costs.

Keywords: *Venta Directa, Regresión Logística, Arboles de decisión, Analítica de datos, Bayes Naive, predicción, pronósticos.*

1. INTRODUCCIÓN

El modelo de venta directa es aquel en el cual se vende un producto o un servicio, de una persona a otra, en un lugar que no tiene un propósito comercial. El vendedor es considerado un representante, agente comercial, promotor o cliente de la compañía que le proporciona instrumentos y material a promocionar conociendo de que obtendrá un margen de rentabilidad de acuerdo con la gestión comercial que realice (Luis Fernández Martínez, Cobo Quesada, and Sánchez-Bayón 2017).

En este modelo de negocio se realiza la distribución de los productos al promotor por medio de compañías logísticas generalmente tercerizadas que entregan directamente en el domicilio o sitio acordado. De esta manera el promotor o agente comercial de la compañía espera recibir su pedido en la fecha de promesa establecida y con las cantidades completas de productos solicitados. Esta parte del nivel de servicio es crítico y fundamental en las empresas de venta directa y se denomina proceso de logística de “Última Milla”, este proceso puede ser el más determinante de toda la cadena logística, cuyo objetivo es entregar al domicilio del cliente en el menor tiempo, alineando los procesos digitales de realización de pedidos y logística de distribución (Janjevic and Winkenbach 2020).

De acuerdo con lo anterior los procesos de última milla son relevantes en el desempeño del negocio de venta directa y estos procesos enfrentan numerosos desafíos como: i) el cumplimiento de los pedidos de acuerdo a la expectativa del cliente, ii) las fuertes presiones competitivas hacia los servicios de entrega gratuitos, iii) el aumento de las expectativas del cliente con respecto a plazos de entrega cortos y iv) plazos de entrega programados individualmente (Janjevic and Winkenbach 2020). La fuerte conexión entre el promotor y los procesos de última milla genera que las compañías de venta directa tomen decisiones enfocadas a favorecer siempre al promotor. Un ejemplo de estas decisiones es la atención de toda reclamación sin objeción alguna. Las falencias en los procesos de última milla

y las políticas flexibles de las compañías permiten al promotor realizar reclamos falsos para facilitar el cumplimiento de metas comerciales. Esta problemática reduce el nivel de servicio esperado por el departamento de distribución.

La reacción de las empresas de venta directa, para atender cada una de las reclamaciones en la entrega de los pedidos, ocasiona un incremento de los costos dentro de la cadena de suministro. Identificar las variables que abordan esta problemática y que relaciona los procesos comerciales y de última milla, serán el soporte para generar una importante base de datos apta para ser tratada con Analítica de Datos, esta es una herramienta de la cuarta revolución industrial utilizada para inferir o predecir causales de reclamaciones a evitar, disminuyendo así los costos respectivos e incentivar a las promotoras para mejorar sus expectativas comerciales (Witkowski 2017).

Para aplicar las metodologías de inteligencia artificial y ciencia de datos en el proceso de última milla de las empresas de venta directa, se utilizaron los datos de una empresa líder del sector cosmético. Para desarrollar un algoritmo de clasificación supervisada que permite anticipar reclamos por parte de las promotoras y facilitar el cumplimiento de las metas comerciales, las técnicas de analítica basadas en modelos de clasificación supervisada, otorgan un sistema de caracterización de variables de conducta de las promotoras, que potencialmente realizarán un reclamo con beneficio propio y que implicara sobrecostos (Kleyner and Sandborn 2004).

Existen diversas técnicas de clasificación supervisada aptas para desarrollar problemas de asignación predictiva, cada técnica explora los datos de maneras diferentes y predice el resultado con un grado de precisión que es acorde a la naturaleza de la técnica. Teniendo en cuenta lo expuesto, se optó por estudiar 4 métodos de clasificación supervisada (Arboles de decisión, Bayes Naive, K vecinos más cercanos y Regresión Logística) y aplicarlas a los datos disponibles. La información resultante permitirá tomar contramedidas a las reclamaciones y

generará eficiencia dentro de la cadena de suministro de las empresas de venta directa.

No obstante, dentro de las distintas áreas que componen la cadena de suministro, se conocen pocos casos en donde las técnicas analíticas han desempeñado un papel que involucre tanto el área logística de entregas como el área comercial. En las empresas de venta directa, normalmente se subcontrata el proceso de última milla y esto ha limitado las investigaciones analíticas dentro de esta área (Trkman et al. 2010).

En consecuencia, las actividades ejecutadas para contener las reclamaciones son de naturaleza reactiva, ante esta situación la empresa no tiene control sobre las reclamaciones y las causas de la problemática siguen vigentes. Implementar un método de clasificación supervisada permite cambiar el sistema reactivo de contramedidas a un sistema predictivo basado en la ponderación de relevancia de las variables que definen la decisión dicotómica (Sí/No realiza reclamo) de la promotora que recibe su pedido.

2. JUSTIFICACIÓN

El modelo de comercialización a través de la venta por catálogo o venta directa es un modelo de negocio muy común en el entorno colombiano, ofrece beneficios al productor al evitar el pago de establecimientos comerciales, este modelo otorga la oportunidad de crecimiento al promotor, quien al promocionar los productos tiene un mayor margen de ganancia. El modelo implica la entrega de artículos directamente al domicilio del promotor.

En el primer semestre del 2020 el mercado colombiano de la venta directa a experimentado una contracción del 7% debido a la pandemia, en contraste de los años anteriores donde el canal de venta directa presentaba un crecimiento anual

del 10% promedio en América Latina, desde el 2018 el 25% de las ventas de productos cosméticos y cuidado personal se han realizado a través de este canal.

En el 2017, los productos provenientes de la venta directa correspondían a un 4% de la canasta familiar; para 2018 este porcentaje se redujo a 3% en el ámbito colombiano.

La categoría estrella de este canal de ventas son los productos de tocador y aseo, considerándose los cosméticos dentro de esta categoría, actualmente el 79% de los hogares colombianos están familiarizados con los productos cosméticos de la venta directa. Los hogares numerosos considerados entre 5 y 6 miembros y en estratos entre 1 y 3 son el grueso en la participación de este canal. (Portafolio 2019).

La fidelización del cliente es el "santo grial" de la venta directa. Las estrategias de fidelización garantizan ingresos constantes a las compañías, y permiten trabajar en el acaparamiento de más clientes y mercados. Como consecuencia de esto, las empresas han construido un nutrido portafolio de clientes y bases históricas de ventas enfocando sus inversiones en mantener calidad en sus productos y red operativa de cara al cliente.

En Colombia existen alrededor de 2.276.225 personas que ejercen una actividad comercial a través del canal de venta directa o venta por catálogo **Tabla 1**. Según el informe de la Asociación Nacional de Industriales (ANDI 2016), solo en el segmento de venta de cosméticos por catálogo se generan 1.200.000 empleos indirectos y 35.000 directos, de los que las mujeres son la principal fuerza laboral (Gutierrez Arenas 2019).

Tabla 1. Número de vendedoras por catálogo en Latinoamérica. (May 2016).

Región/País	Ventas en Retail 2018		% Incremento año anterior	% de incremento promedio desde el 2015 - 2018	Representación independiente
	Moneda Local (Millones)	USD (Millones)			
Sur y Centro América	na	25,140	2.00%	2.30%	13,194,615
Argentina	45,323	1,613	23.10%	33.40%	898,000
Bolivia	2,466	357	3.40%	2.30%	329,310
Brasil	37,260	10,198	-1.50%	-1.30%	3,820,000
Chile	442,148	689	6.70%	5.60%	432,545
Colombia	7,357,000	2,489	4.30%	1.70%	2,276,225
Ecuador	1,197	1,197	0.40%	7.80%	913,280
México	112,864	5,865	1.30%	2.30%	2,770,000
Perú	6,130	1,880	8.10%	5.20%	762,030
Uruguay	2,666	87	9.60%	5.60%	102,595
Venezuela	na	na	na	Na	na
Centro América y Caribe	na	736	-5.00%	-2.50%	862,730
Otros	na	30	2.00%	3.00%	27,900

Fuente: WFDSA

En 2018 América Latina contribuyó con US$ 23,4 billones y más de 14 millones de vendedores por catálogo, fue la región que registró el segundo mayor crecimiento (5,3%) después de África y Oriente Medio a nivel mundial, de esta región latinoamericana, Colombia y Perú fueron los países que generaron el mayor crecimiento (Cabrejos-burga et al. 2020).

Las compañías de venta directa en Colombia se han visto beneficiadas por los altos índices de fidelización. Los incentivos atractivos, la rentabilidad inmediata, la independencia económica y la flexibilidad han atraído a una diversidad de clientes con culturas y hábitos característicos. La multiculturalidad de las promotoras exige a las compañías de venta directa controles generalizados, especialmente en el área de distribución.

Habitualmente las empresas de venta directa carecen de recursos y capacidades que les permitan operar de manera independiente los procesos de distribución última milla. Ante estas, deben buscar un proveedor externo para superar esa debilidad, la selección de proveedores para esta actividad requiere un estudio

destacado que contemple todas las aristas del negocio y garantice una ventaja competitiva sostenible (Tayauova 2012).

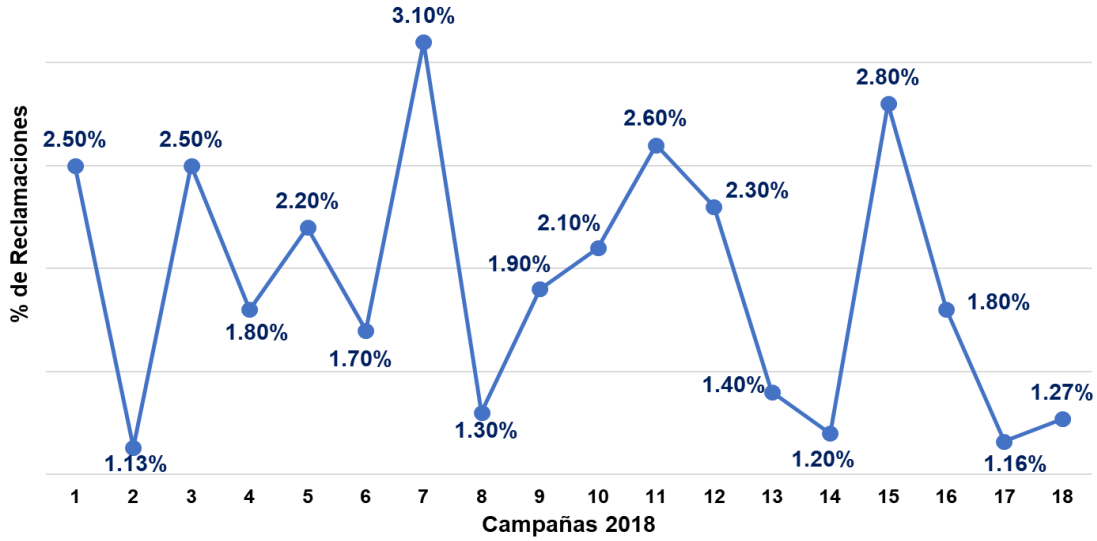
En consecuencia, la tercerización está limitada por los recursos y acuerdos exigidos al proveedor, esto ha permitido la filtración de prácticas indebidas por parte de la promotora, tal es el caso de las reclamaciones falsas.

Las prácticas indebidas es el motivo de investigación, en el cual se ha tomado en contexto una compañía de venta directa que opera en el territorio colombiano comercializando y distribuyendo productos cosméticos a través de su red de promotoras hasta el consumidor final. La compañía determinó que el 15% de los sobrecostos en el área de distribución se generaban por las reclamaciones a causa de productos no enviados dentro de sus pedidos.

El canal de ventas de la empresa cubre todo el territorio colombiano y está comprendido por 24 regiones, en las cuales cada región comprende 10 zonas y cada zona 10 territorios, en promedio la empresa cuenta con 23.500 promotoras activas distribuidas en sus 2.400 territorios.

Considerando que cada promotora realiza pedido de manera frecuente, la compañía genera en promedio 23.500 pedidos vendidos cada 20 días, equivalente a una campaña de ventas, de los cuales se ha reportado que en promedio el 1,93% de los pedidos son reclamados por faltante de productos. La **Figura 1** muestra el promedio porcentual de pedidos con faltante en requerimientos por el promotor, seccionados cada 20 días del año 2018, de esta forma se toman 18 campañas al año.

Figura 1. Tendencia de productos faltantes en pedido por campaña en 2018.

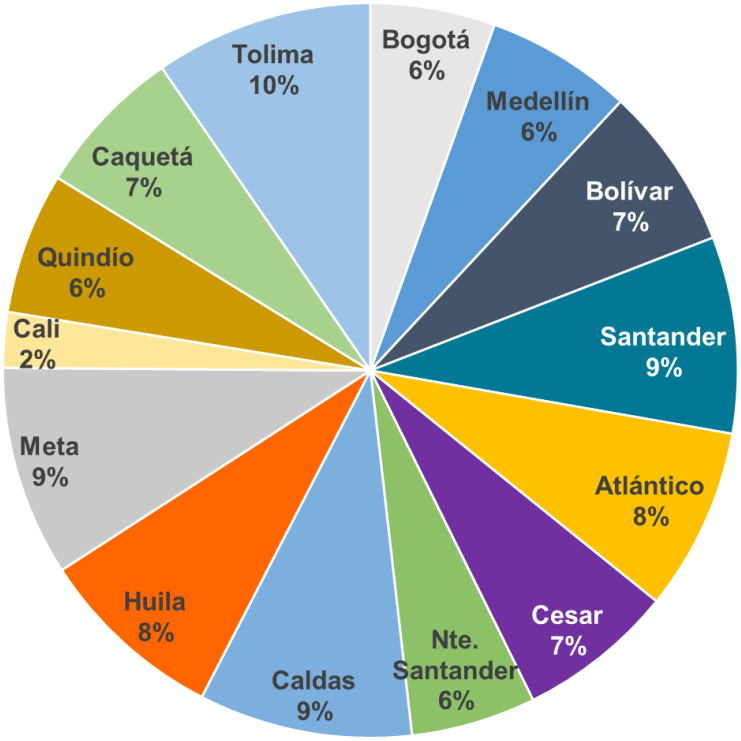


Fuente: Elaborada por el autor

Este porcentaje de productos faltantes traducido a nivel monetario en dólares corresponde a un promedio de USD 29,690 anuales o un promedio de 103,5 millones de pesos colombianos anules, causados por la logística de reembolsar el producto faltante y atender nuevamente al cliente reclamante, así como de reponer un producto que se entiende como extraviado pero que se contempla como fabricado.

Esta cifra porcentual de productos carentes en los pedidos se puede dividir en cada una de las zonas donde opera la compañía, de esta forma, durante el año 2018 los pedidos con productos faltantes se distribuyen de acuerdo con la siguiente gráfica.

Figura 2. Reclamos por regiones 2018.



Fuente: Elaborada por el autor

La compañía es consciente del impacto de estas reclamaciones y por lo tanto opta por la automatización del proceso de armado. Garantizando desde el armado de los pedidos la exactitud en cada caja, se reducen las reclamaciones al 0%, sin embargo, las reclamaciones continuaron su promedio del 2%.

La empresa ha ejecutado medidas adicionales para contener las reclamaciones; entre éstas, se encuentra la verificación de la exactitud de los pedidos directamente en la entrega a la promotora en donde el transportador realiza un chequeo producto por producto acorde a la factura de la promotora. No obstante, esta medida es ejecutada por decisión del transportador y por ende no valora ningún tipo de información que asocie a las promotoras reclamantes. Esto implica que esta acción es de carácter aleatorio y correctivo, lo cual, no elimina la reclamación al considerar que las promotoras realizan la petición a conciencia de su retribución. Por consiguiente, surge la necesidad de implementar ciencia de datos e inteligencia

artificial para predecir reclamaciones falsas, lo cual garantiza una reducción en los costos asociados a la logística y un aumento en el nivel de servicio al promotor. Con esta implementación se remplaza estadísticamente la aleatoriedad y se ataca la causa de las reclamaciones sin afectar la operatividad del proveedor (Liliana 2016).

2.1. DESCRIPCIÓN DEL PROCESO

En la venta directa, el promotor capta clientes (consumidores) ofreciendo los productos que comercializa la empresa mediante un catálogo virtual o físico. En la medida que el promotor incremente sus ventas las ganancias subirán, el promotor también puede obtener mayor ganancia si adquiere promociones ofrecidas por la compañía, lo que le permite obtener productos adicionales a menor precio, adicional, las personas que venden por catálogo son comúnmente motivadas con premios que incentivan la constancia de solicitud de pedidos y los montos del pedido, los premios varían desde artículos básicos para el hogar hasta viajes y vehículos.

Esta promesa de valor permite a la compañía de venta directa trabajar con promotoras quienes, sin necesidad de un contrato laboral, operan como fuerza de ventas. Sin embargo, muchas promotoras realizan prácticas indebidas en la recepción de sus pedidos, alegando que los pedidos no llegaron completos, por lo cual, reciben un producto adicional o no se les cobra el producto reportado, según desee la promotora.

Las promotoras tienen la opción de realizar uno o varios pedidos cada campaña de ventas, la cual está comprendida por 20 días calendario. Es importante resaltar que cada promotora adquiere sus productos a un precio diferencial al catálogo, por lo que se espera que cada promotora obtenga un margen de rentabilidad que oscila entre el 30% y 40% del monto solicitado. Adicional cada promotora maneja un sistema de puntos asociado a la frecuencia con la cual realiza pedidos, si una

promotora realiza pedido durante tres campañas consecutivas, en la cuarta recibirá un premio por su labor de ventas.

Cada promotora es afiliada por un agente comercial de la empresa. El agente comercial se encarga de captar nuevas promotoras y garantizar la continuidad de éstas. Existe un agente comercial por cada zona y su labor es evaluada de acuerdo con las ventas que realicen y la cantidad de ingresos y egresos que mantengan.

Las promotoras reciben sus pedidos directamente en su domicilio considerando el tiempo que transitó el pedido desde el centro de distribución ubicado en Bogotá hasta el punto de destino. Esto implica que no todas las promotoras tienen el mismo tiempo disponible de ventas considerando los tránsitos de transporte. Estas particularidades llevan a investigar la cadena de suministro de la empresa en busca de las variables que inciden directamente con la conducta de las promotoras ante la generación de un reclamo.

2.2. DESCRIPCIÓN DEL PROCESO ESPECÍFICO

La empresa contempla todas las etapas de la cadena de suministro desde la recepción, almacenamiento y calidad de producto entrante hasta la distribución del pedido directamente con el cliente, el mapa de relaciones detallado en el Anexo 1, describe la estructura de la empresa y especifica el proceso distribución como área enfoque en esta investigación.

A continuación, se describen los agentes claves para el desempeño del negocio y los cuales van relacionados con el nivel de servicio que se le promete a la promotora.

Pronósticos: El área de pronósticos realiza un análisis de la demanda de acuerdo con la intensidad de las campañas publicitarias y el promedio de ventas por periodo y producto.

Proveedores: Los proveedores siguen el requerimiento de la demanda pronosticada y envían la materia prima estimando un tiempo de entrega entre el pedido de la materia prima y la entrega respectiva, cada materia prima por proveedor contempla un tiempo específico.

Almacenamiento: almacenar materias primas e insumos utilizando los métodos de organización de inventario como FIFO (First In – First Out), o LIFO (Last In – First Out) de acuerdo con el comportamiento de la rotación de la mercancía. (Bu and Potop-Butucaru 2018)

Producción: transforma las materias primas e insumos en producto terminado aplicando los diferentes métodos de producción, producción en línea, producción en lotes o células y producción por trabajo o pedido.

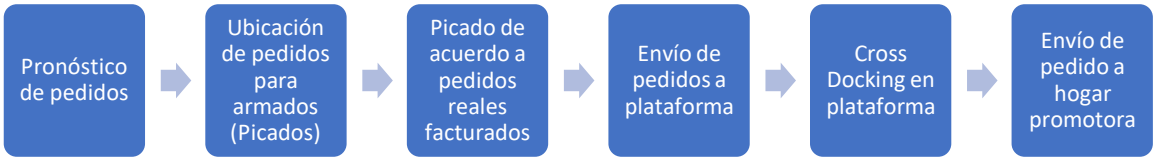
Almacenamiento producto terminado: almacenar los productos terminados de acuerdo con los requerimientos de facturación o demanda. Por otra parte, este inventario puede ser ajustado basándose en los pronósticos de venta, estimados de demanda o a la rotación que la compañía considere adecuada.

Despacho y entrega final: El producto terminado es distribuido según solicitud del cliente con un Lead Time que varía dependiendo de la zona a la que debe llegar el pedido. Los pedidos son enviados inicialmente a bodegas que operan como pequeños centros de distribución ubicados en cada región del país, estos pequeños centros de distribución serán llamados plataformas, los pedidos al llegar a cada plataforma son descargados y cargados nuevamente bajo la estrategia *Cross-Docking*.

El proceso de *Cross-Docking* es un proceso relevante dentro de la cadena de suministro, ya que corresponde al proceso de conexión entre el centro de distribución y las promotoras. Los flujos entrantes desde el centro de distribución se consolidan en envíos únicos hasta las promotoras y se entregan directamente, omitiendo así la fase de almacenamiento (Van Belle, Valckenaers, and Cattrysse 2012).

El proceso de distribución está basado en los pronósticos de armado y en las ventas, con estas variables realiza la planeación de vehículos para entrega de la mercancía, la figura 3 detalla el flujo de actividades dentro de la cadena de suministro hasta llegar a la promotora.

Figura 3. Diagrama de proceso Distribución pronóstico



Fuente: Elaborado por el autor

En el proceso de entrega de pedidos participan dos actores que son protagonistas en la trazabilidad de las entregas (Anexo 2):

i. El transportador es quien tiene contacto con la promotora al momento de la entrega del pedido, de igual manera se comunica con ella en caso de requerir información adicional sobre el pedido o dirección de entrega.
ii. La agente comercial es la persona responsable de vincular promotoras acordes a una zona previamente asignada, mostrando y evidenciando los beneficios de la comercialización de productos de la empresa.

Cada campaña tiene un tiempo disponible de facturación de 5 días hábiles. Si las promotoras no generan el pedido durante este tiempo, no obtendrán los beneficios mencionados, esta restricción otorga celeridad a la agente comercial, quien contacta a todas las promotoras activas y las motiva a realizar pedido.

Una vez la promotora genera el pedido, la empresa inicia el proceso de armado y entregas respectivas. A continuación, se detalla el paso a paso hasta la entrega a la promotora.

Paso 1. El pedido es armado el día posterior al día realizado por la solicitud.

Paso2. Dependiendo de la región, el pedido es despachado. En las zonas más alejadas, la mercancía tarda en llegar 3 días calendario (Costas y Amazonas); en las regiones cercanas (Bogotá), la mercancía llega a la bodega logística en 1 día calendario.

Paso 3. Mediante el proceso de *Cross-Docking* el pedido es asignado a un transportador quien normalmente realiza la distribución en la zona determinada.

Paso 4. El transportador direcciona el pedido según la dirección especificada por la promotora.

Paso 5. El transportador contacta previamente a la promotora para acordar la entrega y la visita al domicilio especificado.

Paso 6. Durante la entrega el transportador decide si realiza verificación de la exactitud del pedido.

Paso 7. La promotora firma la prueba de entrega y recibe el desprendible de entrega correspondiente. El transportador mantiene la prueba de entrega.

Paso 8. El transportador confirma la entrega en un dispositivo móvil que captura las coordenadas de entrega, la imagen de la prueba de entrega, fecha y hora de entrega.

Cabe resaltar que las reclamaciones son atendidas siempre y cuando no exista una prueba física que la desmienta, por lo cual, si la promotora no se le realizó verificación del pedido al momento de la entrega, tendrá la oportunidad de generar una reclamación por productos faltantes.

3. MARCO TEORICO.

Los procesos de última milla a menudo representan los procesos más complejos y costosos en la logística de distribución B2C (Bussines-to-consumer), se ha considerado que la entrega final al cliente desde los centros de distribución representa cerca del 28% de los costos totales logísticos, así mismo, es un proceso crítico, pues la eficiencia y desempeño del proceso de última milla puede impactar de manera negativa o positiva el cliente (Bergmann, Wagner, and Winkenbach 2020). El nivel de servicio ofrecido al cliente va directamente relacionado con el cumplimiento de los objetivos comerciales.

Diferentes estudios han mostrado que la ciencia de datos y específicamente los modelos de clasificación logran detectar patrones de conducta que llevaran a predecir con un grado de probabilidad una actividad. Tal es el caso de Chile, donde se aplicó un modelo de detección de contribuyentes con facturas falsas por medio de los algoritmos Neural Gas y SOM que permitió detectar previamente clientes con pagos fraudulentos (Castellón González and Velásquez 2013). En Filipinas se caracterizaron clientes de una empresa de venta directa para detectar su comportamiento de fidelidad al negocio a lo largo del tiempo, en dicho modelo se aplicaron cadenas de Markov y series binomiales de regresión logística, lo cual, resultó un beneficio al calcular el valor del cliente en el mercado y disponer recursos programados a la actividad de inclusión (Mauricio,Payawal, J.M.M.Dela Cueva, Quevedo, V.C 2016) y en el ambito bancario se han logrado detectar fraudes en el uso de tarjetas de credito aplicando Redes Neuronales y redes de creencia bayesiana, en los cuales se comparó el rendimiento de los modelos con datos de entrenamiento y demostrando que las redes de creencia Bayesiana presenta un rendimiento superior al detectar el fraude (Ravelin 2002).

Los modelos de clasificación supervisada han contribuido a la competitividad de las empresas y al fortalecimiento de las áreas comerciales y logísticas. Un estudio en Mexico demostró a partir de datos de la plataforma de comercio electrónico de

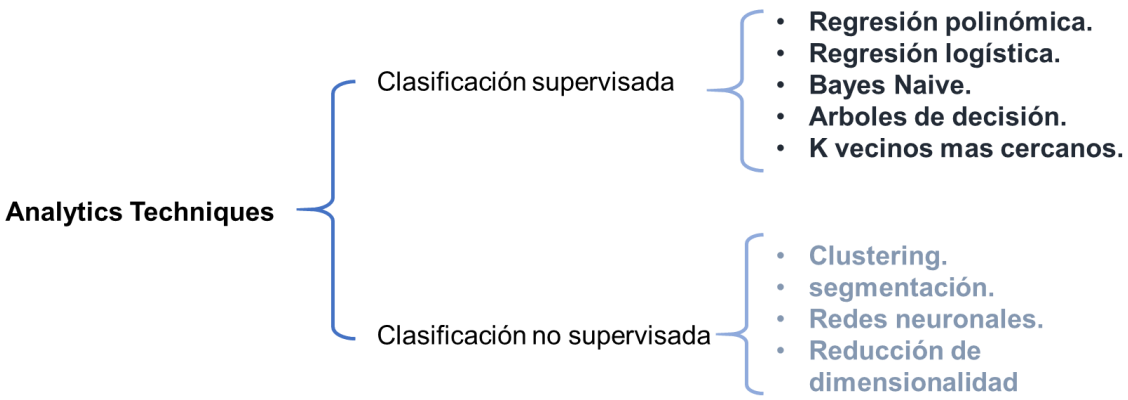
MercadoLibre el comportamiento futuro de la competencia aplicando modelos de clasificación supervisada, el resultado premitió generar estrategias enfocadas al cliente (Miriam and Arroyo 2019). En el sector logístico un estudio de la universidad Guandgong logró estimar los indices de inventario de manera adecuada y eficiente mediante la implementación de modelos de regresión logística y arboles de decisión, el resultado presentó una mejora en el S&OP de las empresas y conllevó a una reducción de inventario obsoleto (Zhou et al. 2019). En logistica de ultima milla tambien existen estudios en donde se han implementado arboles de decision para mejorar el proceso de entregas en ciudades mediante drones, optimizando las caracteristicas del drone al considerar las condiciones climaticas, distancia hasta el destino y peso de la carga (Shahzaad et al. 2021).

Los modelos de clasificación se encuentran divididos en dos familias al momento de abordar los datos, por una parte, están los problemas de clasificación no supervisada, estos tratan a la clasificación como el descubrimiento de las clases de un determinado problema. Es decir, un conjunto de elementos descritos por un conjunto de características, sin conocer a que clase pertenece cada uno de ellos. En contraste, se encuentran los problemas de clasificación supervisada, que corresponden a un conjunto de elementos descritos por un conjunto de características de las cuales conocemos la clase a la que pertenece (Corso 2009).

3.1. METODOS DE ESTIMACIÓN

Enfocados en el objetivo del proyecto, se determinó que el problema está alineado con las técnicas analítica de datos. Dentro de esta familia la **Figura 4** destaca las técnicas usadas para el desarrollo de la problemática.

Figura 4 Técnicas de clasificación



Fuente: Elaborado por el autor

Existen muchos paquetes informáticos aptos para el tratamiento de datos y aplicación de técnicas de clasificación, sin embargo, la investigación realiza su análisis en el Software open-source} KNIME Analytics platform, este software utiliza un sistema basado en los flujos de trabajo para la integración, análisis y exploración de los datos, fue diseñado para manejar grandes cantidades de datos heterogéneos en un entorno informático independiente de la plataforma y ha tenido éxito en satisfacer las complejas demandas de extremo a extremo en varias comunidades (Dietz et al. 2020).

Las técnicas de clasificación supervisada exploradas en la KNIME con fin de encontrar la asignación de promotoras que realizaran reclamos son descritas a continuación:

3.1.1. Regresión logística.

La regresión logística es un modelo asociado a la regresión polinómica en el cual se detalla el resultado de una variable predictora del tipo binario a través de sus variables explicativas o independientes, la función predictora detalla el resultado en términos dicotómicos y se explica mediante la siguiente formula:

$$Log\left[\frac{p}{1-p}\right] = B_0 + B_1x_1 + B_2x_2 + \cdots + B_ix_i$$

Donde:

- p es la probabilidad de respuesta de la variable de interés.
- B_0 es un término de intercepción.
- $B_1 \ldots . B_i$ Son los coeficientes asociados a cada variable X.
- $x_1 \ldots x_i$ Son las variables independientes.

Al ser derivada de la regresión polinómica, la regresión logística presenta las mismas restricciones al ser vinculada con los modelos, sin embargo, como ventaja la regresión logística es de fácil interpretación a diferencia de otras técnicas, por lo cual es de amplio uso en el ámbito médico y financiero (Khoshgoftaar and Allen 1999).

3.1.2. Naïve Bayes.

Es una familia de clasificadores probabilísticos simples basados en la aplicación del teorema de Bayes con fuertes supuestos de independencia entre las características (Naik and Samant 2016).

Supone que el efecto de un valor de atributo en una clase dada es independiente de los valores de los otros atributos (rara vez cierto), se conoce como "independencia condicional". No obstante, funciona bien en la práctica y la clasificación no necesita una estimación de probabilidad exacta si se da la probabilidad máxima a la clase correcta (Ricciardi et al. 2020).

3.1.3. Arboles de decisión.

Un Árbol de decisión es un diagrama de flujo con estructura ramificada, donde cada nodo interno corresponde a la prueba de un atributo y cada rama representa la salida resultante en la prueba en el nodo (Erkut 2016).

Las pruebas sobre cada atributo se ejecutan mediante una relación de ganancia. Este proceso permite obtener el nodo más puro de acuerdo con una diferencia

basada en la fragmentación del conjunto de datos iniciales. La pureza está definida como la probabilidad de que un dato abarque el mayor porcentaje posible dentro de una clase (Cohen, Rokach, and Maimon 2007).

Su estructura facilita mucho el entendimiento del modelo y permite asociar las variables de manera independiente al momento de encontrar una respuesta, no obstante, las reglas de asignación son bastante sencillas y pueden generar distorsión en los datos.

3.1.4. K-vecinos más cercanos.

Este algoritmo es considerado uno de los más simples dentro de la familia de algoritmos para clasificación supervisada. Es un algoritmo del estilo no paramétrico y basa su desarrollo en el cálculo de distancias. En consecuencia, esto implica que realiza suposiciones sobre la funcionalidad de los datos, lo que implica que el algoritmo no aprende explícitamente del modelo si no que a su vez memoriza las distancias para la predicción.

Su esencia se centra en buscar datos con tendencias y características semejantes que permitan obtener información relevante para predecir el resultado de acuerdo a la situación problema especificada, el éxito del algoritmo depende de la recolección de las variables y la afinidad con los datos (Romero, Riveros, and Herrera 2017).

3.2. VALIDACIÓN CRUZADA Y METRICAS DE EVALUACIÓN

La validación cruzada es un procedimiento estadístico implementado desde 1995 para encontrar la precisión real de un modelo predictivo. En la validación cruzada un conjunto de datos se divide aleatoriamente en k subconjuntos mutuamente excluyentes (también llamados "pliegues") de un número casi igual de instancias. El

modelo está entrenado y probado k veces, cada vez que se entrena en diferentes pliegues k-1 y se prueba en uno. (Ricciardi et al. 2020).

La curva (ROC) describe el rendimiento de una prueba de diagnóstico, que clasifica los resultados en una de dos categorías. Se define como una gráfica de la tasa de verdaderos positivos contra la tasa de falsos positivos, o la sensibilidad frente a la especificidad para varios valores de un umbral (Jokiel-Rokita and Topolnicki 2020).

3.3. CLASE MINORITARIA

La clase minoritaria se presenta en los problemas de clasificación en donde la variable objetivo tiene baja proporcionalidad con respecto al total de observaciones, lo que ocasiona medidas imperfectas en la estimación, dado que el modelo no encuentra patrones singulares que se atribuyan a la respuesta (Lin and Nguyen 2020), una de las técnicas que resuelve este problema es la técnica SMOTE.

Para balancear los datos con respecto a las clases de la base, se aplicó la técnica SMOTE (*Synthetic Minority Oversampling Technique*). La idea básica del algoritmo SMOTE es analizar las muestras de clase minoritaria y agregar nuevas muestras que se sintetizan artificialmente de acuerdo con las muestras de clase minoritaria al conjunto de datos. Los pasos del algoritmo son los siguientes:

- Para cada muestra en la clase minoritaria, la distancia desde X_i a todas las muestras en la clase de muestra minoritaria se calcula por distancia euclidiana.
- Se establece el aumento de muestreo N de acuerdo con la relación de desequilibrio muestral, para cada muestra minoritaria de clase, se selecciona aleatoriamente varias muestras de sus k vecinos asumiendo que la selección de la muestra vecina es $X_i´$.

- Para cada vecino seleccionado aleatoriamente$X_i{}'$, muestra una construcción completa acorde a la siguiente ecuación.

$$x_{New} = x_i + Random(0,1) * (x_i{}' - x_i)$$

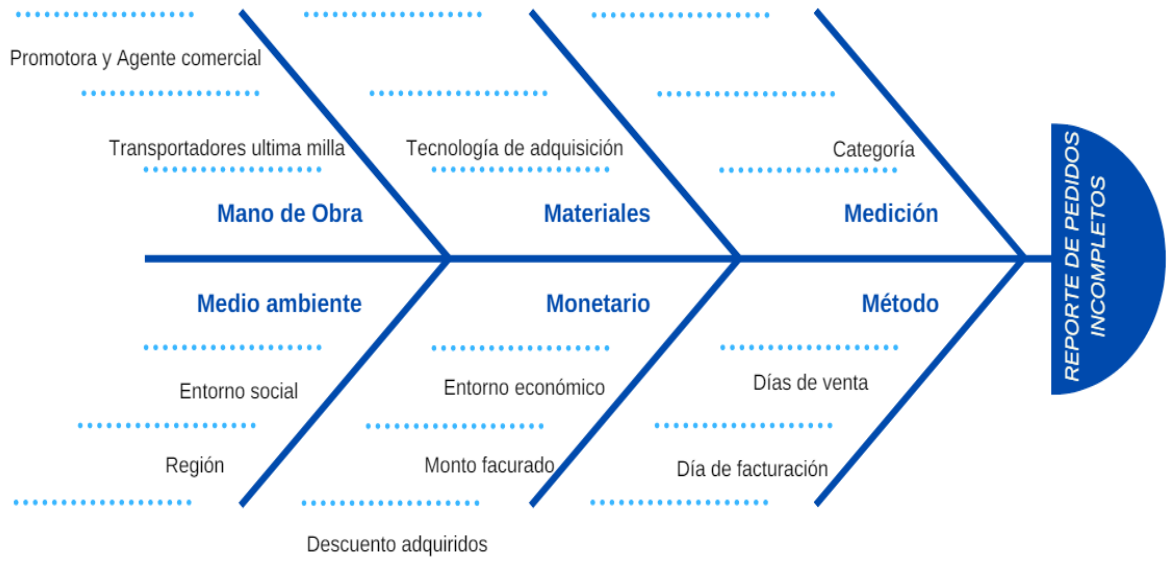
Donde:

- x_{New} es la nueva muestra generada,
- x_i es la muestra actual de datos y
- $x_i{}'$ es cada muestra de los datos de la clase minoritaria (Liang et al. 2020).

4. IDENTIFICACIÓN DE OPORTUNIDADES DE MEJORA

Las reclamaciones por parte de las promotoras no tienen una medida de control de causa raíz, pese a que se han realizado actividades correctivas y preventivas, no se ha generado una acción que permita mitigar el error encontrado desde la raíz, el diagrama de Ishikawa relata todas las posibles causas que enlazan la problemática (Liliana 2016).

Figura 5. Diagrama Ishikawa



Fuente: Elaborado por el autor

El diagrama de Ishikawa hace un análisis de las 6 M (Método, Material, Mano de Obra, Medición, Medio Ambiente y Monetario) que vinculan las reclamaciones con las causas, con fin de encontrar las variables que enlazan el comportamiento de la promotora con la acción de realizar una reclamación con beneficio propio.

A continuación, se describen las variables asociadas a cada rama del diagrama de Ishikawa.

Mano de obra: involucra las acciones realizadas por el transportador de entrega puerta a puerta quien en momentos dados puede hurtar la mercancía por lo que la misma llega incompleta a la promotora.

La agente comercial puede inducir a la promotora a realizar una reclamación con fin de que la promotora se mantenga activa.

Materiales: El proceso de solicitud de pedido se realiza de manera virtual, ya sea mediante una página web o una aplicación móvil, sin embargo, se han evidenciado casos en donde las promotoras realizan erradamente el pedido, escogiendo el producto que no desean o pagando por promociones que no les benefician, al obtener el pedido evidenciar el error, algunas promotoras tienden a solucionarlo realizando un reclamo falso que les evite el pago de los productos facturados por error.

Medición: La promotora es categorizada de acuerdo con su nivel de ventas, esta puede ser definida como Diamante, Esmeralda o Perla.

Medio ambiente: El entorno socioeconómico del país considera las zonas de conflicto donde la promotora está expuesta a influencias que habitúen realizar prácticas indebidas.

Monetario: El monto facturado, el monto descontado por la adquisición de promociones son variables que relacionan la actividad de ventas de la promotora y con la adquisición de incentivos.

Método: Corresponde a la metodología de ventas establecida por la empresa para la actividad comercial, destacando los días programados para

que la promotora realice su facturación y los días disponibles para que genere sus ventas.

De acuerdo con el análisis, la causa de los reclamos no está distribuida en una sola condición. En consecuencia, es necesario abordar el análisis tomando como base todas las variables que puedan interferir en la generación de reclamos por parte de las promotoras.

Con un pronóstico preciso de las reclamaciones generar cuantiosos beneficios para la empresa, los cuales estarán medidos en los siguientes indicadores de gestión:

- **Porcentaje de reclamaciones por campaña:** Acorde al estimado se verificará campaña a campaña el porcentaje de reclamaciones con respecto al total de pedidos, el porcentaje estará alineado con la precisión del modelo.

- **Sobrecostos de reclamaciones:** indica el costo de los productos reportados como faltantes, sin un modelo predictivo, las reclamaciones oscilan un promedio de COP 5.750.000 por campaña.

El implementar el modelo predictivo estimará una reducción del costo de las reclamaciones en una tasa equivalente a la reducción porcentual.

5. OBJETIVOS

5.1. OBJETIVO GENERAL

Desarrollar un modelo predictivo que permita identificar mediante un algoritmo de Inteligencia Artificial las reclamaciones de las promotoras en el proceso de distribución directa.

5.2. OBJETIVOS ESPECIFICOS

- Crear una base de datos que disponga toda la información relevante para distinguir las promotoras que generaran reclamos.
- Ponderar las variables de acuerdo con su relevancia en la estimación de reclamaciones.
- Proponer contramedidas asociadas a la predicción que reduzcan las reclamaciones y mitiguen los sobrecostos que conllevan.

6. METODOLOGÍA

A continuación, se detalla de forma estructurada las herramientas, variables y análisis requeridos para el desarrollo de un modelo analítico que predice los reclamos falsos de clientes que alegan pedidos incompletos en una compañía de venta directa que opera en Colombia.

6.1. RECOPILACIÓN DE VARIABLES

En primera instancia, considerando las causas asociadas a las reclamaciones se recopila un conjunto de datos que involucra todas las posibles variables que inciden en su realización, las siguientes son algunas de las variables consideradas: costo del pedido, antecedentes en reclamaciones, fecha de entrega, fecha de pago, territorio y estado socio económico. Cada variable se relaciona con todos los clientes de la empresa, la esta base de datos es el principal *Input* del modelo, cuyo objetivo es identificar las variables con mayor participación en la conducta de la promotora al realizar una reclamación.

Aplicando un modelo analítico de clasificación supervisada asociado a un conjunto de datos con variables significantes, se predicen patrones comportamentales que permitirán identificar aquellos promotores con alta probabilidad de realizar un reclamo falso.

6.2. CARACTERIZACIÓN DE DATOS

Se consideró la trayectoria de la empresa, evaluando 13 variables ordenadas en columnas a cada una de las 23.500 promotoras activas, con esta información se analizó el tipo de variable y la relevancia sobre la respuesta dicotómica que corresponde a la ejecución o no de un reclamo por productos faltantes.

El número de variables y la definición de estas es fundamental para lograr el desempeño adecuado del modelo. La empresa suministró una base de datos que contiene 52 columnas con diversa información sobre cada promotora, por lo cual, se excluyeron las variables redundantes, las variables con bajo poblamiento y las variables que no aportan información que esta alienada al contexto de la problemática. A continuación, se detallan las variables numéricas y categóricas dispuestas en el modelo.

Variables numéricas

I. **Monto monetario del pedido:** Corresponde al monto pagado por la promotora al adquirir el producto, esta variable varía entre COP 180.000 hasta COP 10.000.000 por campaña.

II. **Número de artículos y categoría:** Corresponde al número de productos solicitados por pedidos y su clasificación con las categorías que los representan, las categorías van relacionadas con el tipo de producto adquirido y se dividen en tres; i) fragancias, ii) cremas y iii) maquillajes.

III. **Descuento:** Es el valor monetario por pedido que la promotora evita pagar al adherirse a las promociones dispuestas por la empresa. Los descuentos generalmente están asociados a la adquisición de paquetes promocionales (Pague 1 lleve 2), por lo que el descuento equivale a la diferencia evitada.

IV. **Días disponibles para la venta:** Indica el número de días que la promotora tiene desde que recibe su pedido hasta que inicia la facturación de la siguiente campaña.

Las variables categóricas encontradas para el análisis del modelo son las siguientes:

I. **Región:** Corresponde al departamento de Colombia en el cual la promotora se encuentra ubicada.

II. **Zona**: es una subdivisión de la región y comprende territorios municipales o barrios con un casco urbano de 500.000 personas aproximadamente.

III. **Sección**: son subdivisiones de la zona y comprende territorios de 10000 personas que incluye zonas veredales, corregimientos o caseríos.

IV. **Día calendario de compra:** Corresponde al día calendario entre lunes y viernes en donde el cliente está habilitado para realizar su compra.

V. **Día de facturación:** Indica el día en el cual la promotora realiza la compra, de acuerdo con la empresa la promotora está habilitada para realizar pedido en 5 días calendarios categorizados del 1 hasta el 5 y ocasionalmente se extiende hasta 7.

VI. **Estado Socioeconómico de la promotora:** Corresponde al estrato social del domicilio de la promotora categorizado de 1 hasta 6.

VII. **Antecedentes:** Identifica a las promotoras que han reclamado en campañas previas a la analizada, esta categorizado con un valor binario en donde 1 son las promotoras con antecedentes y 0 las que no han reclamado.

VIII. **Tipo de cliente:** Corresponde a la clasificación que la compañía le otorga al cliente al considerarse relevante dentro del ejercicio comercial de la venta; a) si el cliente cuenta con más de 3 campañas seguidas realizando compra es clasificado como cliente frecuente, b) es un cliente inconstante si no ha facturado durante 2 campañas consecutivas, c) es cliente egresado si acumula 3 campañas consecutivas sin realizar un pedido y d) es un cliente reingresado si ha facturado después de 2 campañas sin realizar pedido.

IX. **Promociones:** Indica si la promotora ha adquirido alguna promoción en la facturación previa.

La tabla 2 y la tabla 3 detallan la clasificación de las variables categóricas y el resumen estadístico de las variables numéricas.

Tabla 2. Clasificación de variables categóricas.

	Dicotómicas	Politómicas	Ordinales	Intervalo	Nominal
Región		X			
Zona		X			
Sección		X			
Día calendario				X	
Día de facturación				X	
Estrato			X		
Antecedentes	X				
Tipo de cliente					X
Promociones	X				

Fuente: Elaborado por el autor.

Tabla 3. Indicadores resumen variables numéricas.

	Media	Mediana	Moda	Varianza	Rango
Monto del pedido	$ 181,356	$ 156,001	$ 113,447	15,308,343,841	$ 6,867,456
Número de productos	19.97	17	14	11.86	82
Descuento	$ 35,568	$ 7,025	NA	$ 100,371	$ 890,320
Días disponibles para la venta	14.76	16	16	1.85	12

Fuente: Elaborado por el autor.

La identificación de las variables asignadas correspondió en parte a la experiencia del elaborador. Existen en la literatura casos en los cuales se han construido bases de datos en donde coexisten variables cualitativas y cuantitativas, generando bases no homogéneas. Esta clase de data requiere especial atención dependiendo del modelo al cual se va a integrar (Gibert and Cortés 1997).

6.3. INTEGRACIÓN DE LOS DATOS

Con la base de datos disponible, se realizó un análisis enfocado en la variable objetivo, en donde se verificó la proporción de casos positivos (Reclamaciones) respecto a los casos negativos (No reclamaciones), comprobando que las reclamaciones presentan la menor proporción de datos en referencia a las observaciones totales, por lo cual, el conjunto de datos de reclamaciones comprende la clase minoritaria de los datos. La desproporción entre casos positivos a predecir frente a los negativos es de 98%, lo que implica que las técnicas predictivas implementadas trabajarán con esta desproporcionalidad y deberán realizar el análisis en función de todas las variables que suministran información.

La base de datos cuenta con 23.500 observaciones que corresponden a todas las promotoras activas de la empresa a las cuales se asoció con las 13 variables escogidas, asociando un registro único para cada observación.

6.4. SELECCIÓN DE LA TÉCNICA DE MODELAMIENTO

A partir de la base de datos desarrollada, se inició la fase exploratoria de la minería de datos, en la cual, se identificaron técnicas analíticas para la clasificación de las observaciones. Cada técnica fue probada mediante validación cruzada, con 10 iteraciones por modelo y una proporción de 70% en prueba y 30 % de predicción.

Los indicadores de evaluación de las técnicas con validación cruzada permitieron identificar la técnica que presentará el menor error en la estimación con el 100% de los datos, el resultado de cada técnica se expone en una matriz de confusión resultante que muestra los indicadores de especificidad y sensibilidad, al igual que la relación en la tasa de verdaderos positivos contra falsos positivos mostrada en la curva ROC.

El resultado de la curva ROC permitió identificar la técnica adecuada para la modelación, de esta manera se procedió a ejecutar el modelo con el 100% de los datos y encontrar la predicción.

Los primeros resultados fueron evaluados en vivo mediante un plan piloto el cual comparaba la predicción con el resultado real. Para la segunda campaña de ventas el modelo trabajó en vivo y los resultados fueron comparados con los indicadores de gestión del área de distribución, los cuales detallan el número de promotoras que realizan reclamos contra el total de promotoras que reciben pedidos.

7. DATOS, ANÁLISIS Y RESULTADOS

El resultado de la implementación permitió conocer con base en un análisis estadístico los clientes que no desean pagar los productos despachados a causa de que no lograran venderlo o a causa de obtener una ganancia adicional.

Aplicando el proceso de caracterización de las variables previamente vistas se logró encontrar un modelo acorde a una predicción viable de promotoras que realizaran un reclamo falso, el modelo condujo una respuesta dicotómica de acuerdo con la interacción de sus variables independientes con la variable respuesta, por tal motivo el primer resultado corresponde al tratamiento de la data numérica y la construcción de la base de datos.

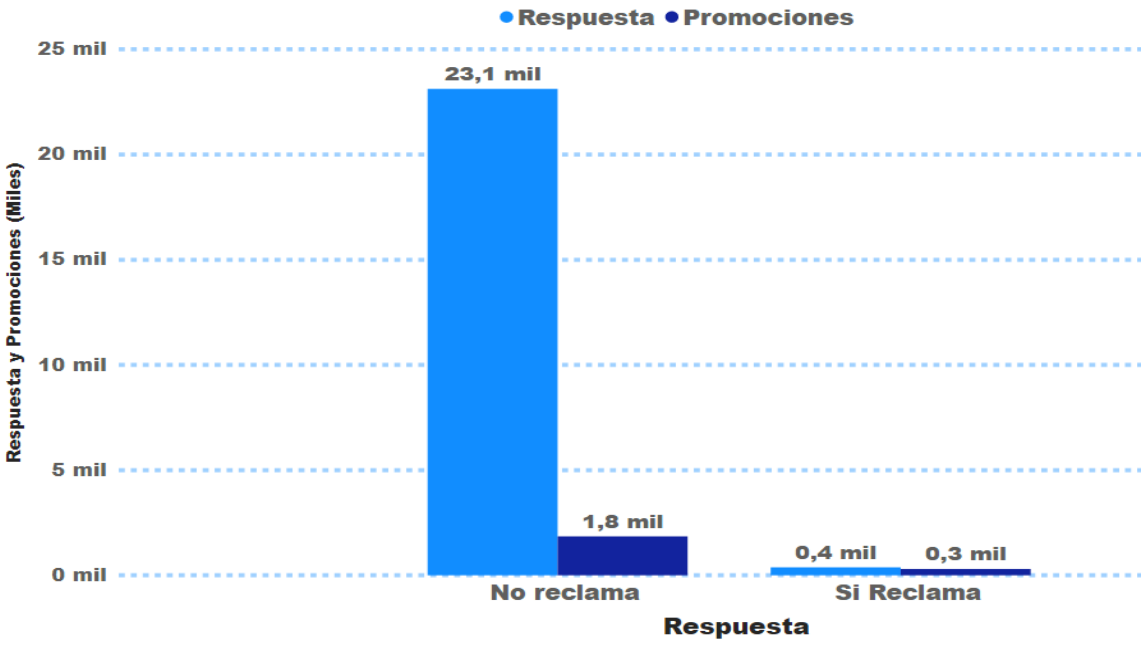
7.1. TRATAMIENTO DE DATOS

Los datos fueron tratados en función de la variable objetivo (reclamaciones), escogiendo las variables que aportan información relevante al modelo y excluyendo las variables redundantes y que no contienen la información apta para el modelamiento.

7.1.1. Promoción.

Esta variable tiene alta significancia en los modelos desarrollados, concretamente influye de manera directa en las respuestas, la magnitud de la comparación de la columna respuesta y variable promoción está representada en el siguiente BarChart.

Figura 5. Bar Chart Promociones.



Fuente: Elaborado por el autor.

La gráfica identifica la relación que existe entre las personas que realizan reclamo y las personas que se vinculan a una promoción. Identificando que dentro del grupo de reclamantes el 80% adquirió su pedido escogiendo una promoción.

7.1.2. Costo del pedido.

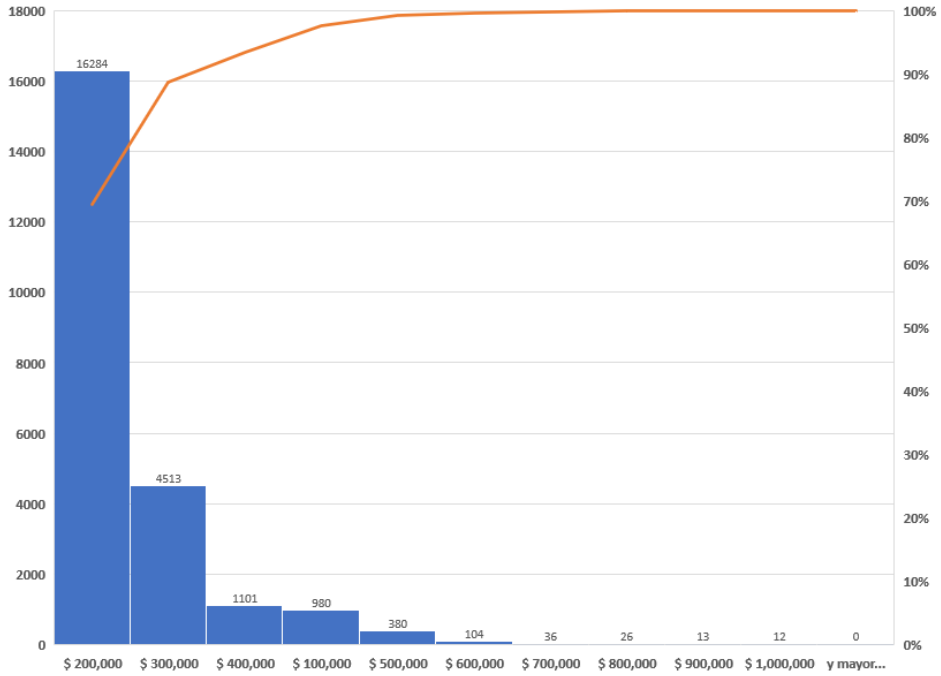
El monto del pedido aporta información importante al modelo, en donde se detalla una relación porcentual con respecto a las clases del histograma, estableciendo que un 58,09% de las reclamaciones se encuentran distribuidos en los pedidos que se encuentran entre $ 100.000 y $ 200.000. Lo cual es información importante para las técnicas analíticas implementadas.

Tabla 4. Tabla de frecuencias Monto del pedido.

Monto pedido	*Frecuencia*	*% Acumulado*	*Reclamaciones*	*% Participación*
$ 100,000	980	4.18%	19	5.04%
$ 200,000	16,284	73.62%	219	58.09%
$ 300,000	4,513	92.87%	86	22.81%
$ 400,000	1,101	97.56%	32	8.49%
$ 500,000	380	99.19%	14	3.71%
$ 600,000	104	99.63%	3	0.80%
$ 700,000	36	99.78%	1	0.27%
$ 800,000	26	99.89%	1	0.27%
$ 900,000	13	99.95%	1	0.27%
$ 1,000,000	12	100.00%	1	0.27%
y mayor...	0	100.00%		

Fuente: Elaborado por el autor.

Figura 6. Diagrama de Pareto Monto del pedido.



Fuente: Elaborado por el autor.

7.1.3. Número de productos por pedido.

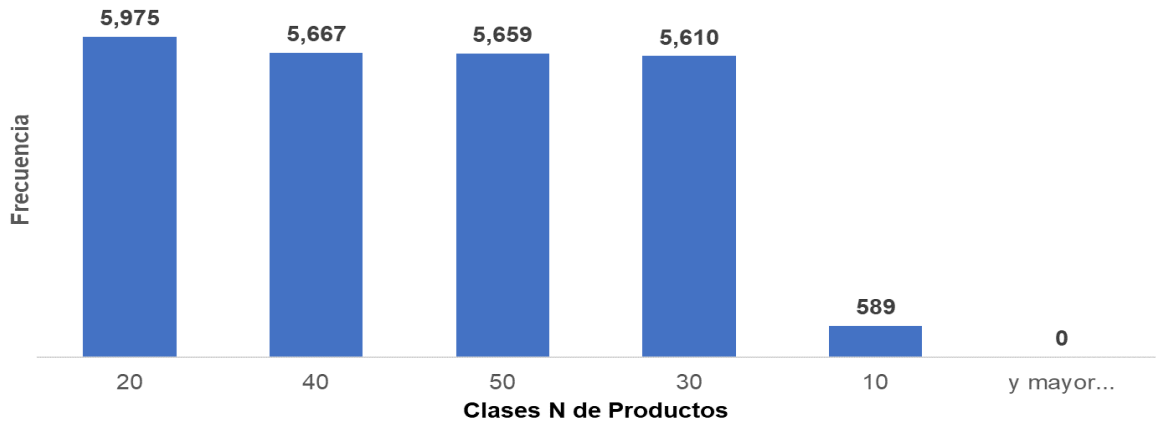
Al igual que el monto del pedido esta variable resume su información con una tabla de frecuencias y un histograma.

Tabla 5. Tabla de frecuencias N de productos.

N Productos	*Frecuencia*	*% Acumulado*	*N Reclamaciones*	*% Reclamaciones*
20	5975	25.43%	39	10.34%
40	5,667	49.54%	338	89.66%
50	5659	73.62%	0	0.00%
30	5,610	97.49%	0	0.00%
10	589	100.00%	0	0.00%
y mayor...	0	100.00%	0	0.00%

Fuente: Elaborado por el autor.

Figura 7. Histograma Número de productos.

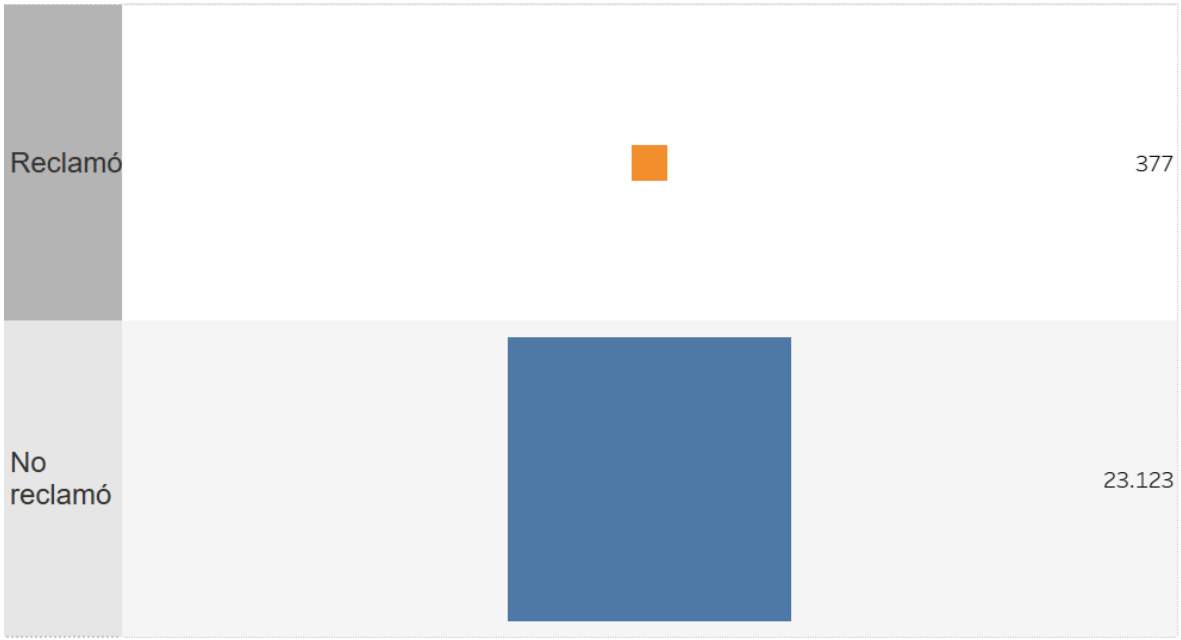


Fuente: Elaborado por el autor.

7.1.4. Descuentos adquiridos.

Es la tercera variable nominal que brinda información al modelo, esta correlacionada con la variable respuesta en medida de la magnitud de los descuentos por compra, el siguiente Scatter Plot muestra la relación.

Figura 8. Diagrama de barras Descuentos.



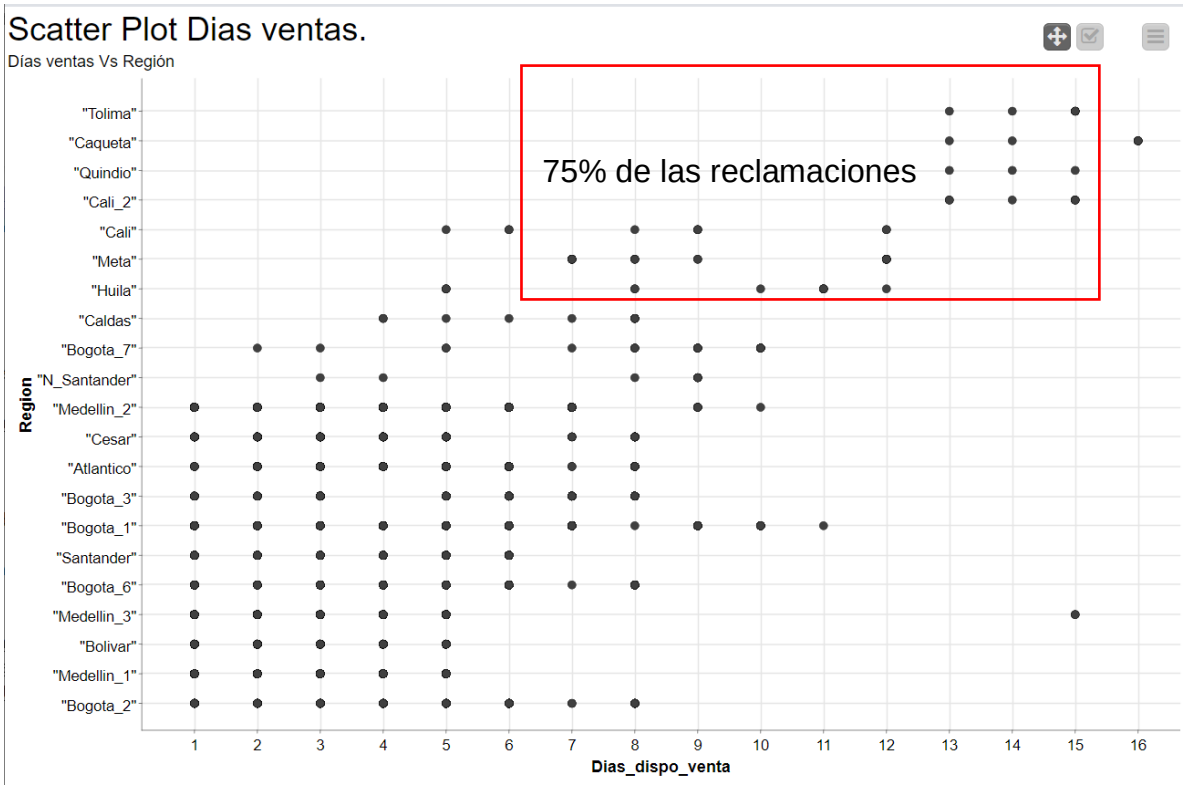
Fuente: Programa Knime Analytics Platfom

Existe asociación entre las reclamaciones y el monto de los descuentos adquiridos, sesgando la respuesta a los descuentos inferiores a $ 12.000.

7.1.5. Días disponibles de ventas y Regional.

Estas dos variables brindan información conjunta para el modelo, dada la relevancia y su información independiente no se encontró factible generar una variable derivada de ambas. El diagrama de dispersión a continuación detalla su relación con la respuesta y en conjunto.

Figura 9. Diagrama de dispersión Regional/Días Ventas.



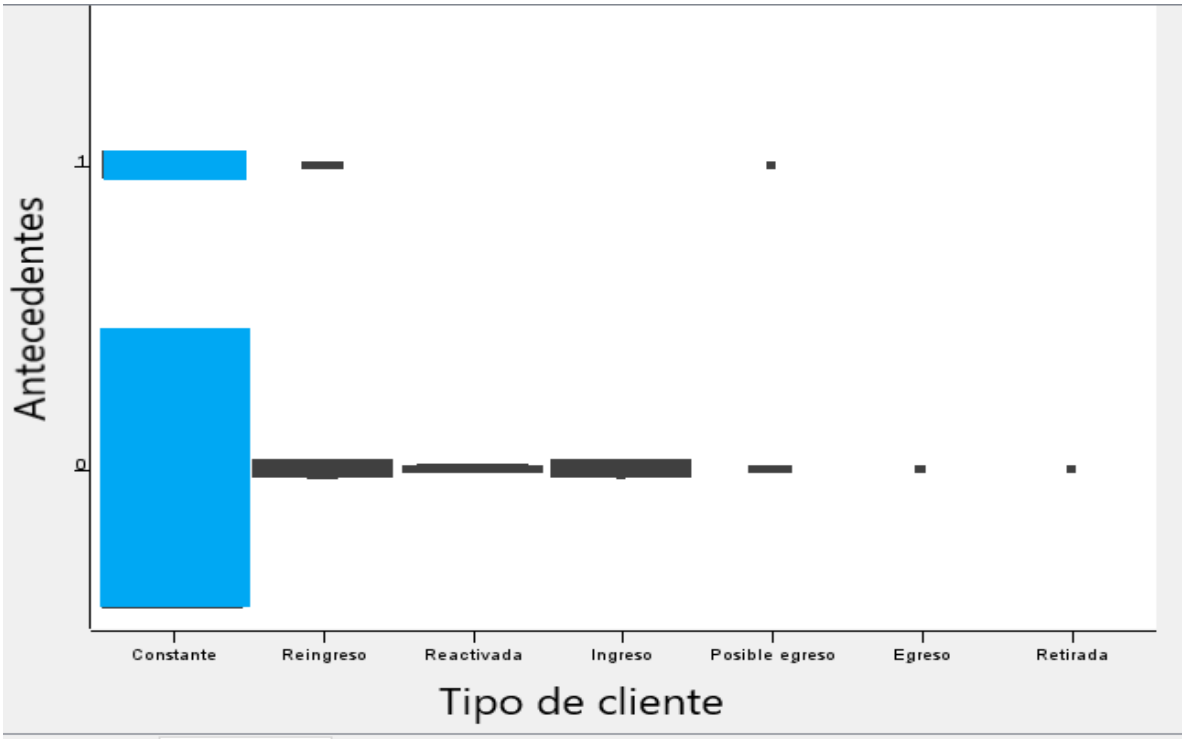
Fuente: Programa Knime Analytics Platfom

7.1.6. Antecedentes y tipo de cliente.

Estas dos variables brindan información en conjunto y por sí solas, al igual que las anteriores se detalla su información mediante un diagrama de dispersión. El área sombreada en azul indica el 90% de las reclamaciones, por lo cual se determina la relación respectiva al igual que la relación entre los antecedentes y el tipo de cliente.

Las variables adicionales del tipo categórica brindan información características al negocio, como la fecha de facturación y la zona sección. Estas variables clasifican a las promotoras y le proveen al modelo rutas de asignación.

Figura 6. Tipo de cliente y Antecedentes.



Fuente: Programa Knime Analytics Platfom

7.1.7. Técnica SMOTE.

Considerando las variables y el objetivo del proyecto, se realizó la modelación aplicando la comparación de técnicas de clasificación supervisada, sin embargo, para el problema en cuestión, el objetivo representa un porcentaje pequeño con respecto a la población, teniendo en cuenta que las reclamaciones corresponden al 2% del total de las observaciones. Por definición, cuando el número de datos en las clases son distintas en proporción con su tamaño, se considera que se está trabajando bajo datos desbalanceados.

Al aplicar la función SMOTE dentro de las 23.500 observaciones en la cual la clase minoritaria corresponde al 1%, se obtuvo una data de 87.761 observaciones con la clase minoritaria incremento a un 42% del total de datos.

7.2. ANÁLISIS DE LOS DATOS

Los datos sobre-muestreados por la técnica SMOTE adecuaron los datos para que las técnicas analíticas presentaran un desempeño apropiado sobre la respuesta. No obstante, en medida que el sobre-muestreo interfiera en los datos, la sensibilidad del modelo varía. Por esta razón, esta técnica sólo se aplica para los datos de entrenamiento y en cada una de las diferentes técnicas de predicción.

7.2.1. Resultados entrenamiento de los datos.

Contemplando el universo de datos y la proporción que mantiene la variable respuesta, se escogió de manera aleatoria el 70% de los datos como datos de entrenamiento, de donde la respuesta fue evaluada con el 30% restante corroborando el asertividad de los modelos previstos

A partir de la muestra que contempla el 70% de los datos equivalente a 61.433 registros, se evaluó el mejor modelo generando iteraciones de entrenamiento para cada modelo, de esta manera, mediante 10 repeticiones aleatorias ejecutadas por el Software KNIME, el programa escogía el mejor modelo posible para cada una las técnicas.

Cada técnica analítica fue evaluada inicialmente mediante una matriz de confusión, identificando las clases pronosticadas contra el pronóstico asociado a los datos de prueba. Las columnas de una matriz de confusión representan los resultados de la clase de predicción, y las filas representan los resultados reales de la clase. Esta enumera todos los casos posibles de un problema de clasificación (Xu, Zhang, and Miao 2020).

La variable respuesta representada como una variable binaria requiere de una matriz de confusión de dimensión 2 × 2, la cual provee una serie de medidas de rendimiento del algoritmo como la tasa de verificación de región positiva y la tasa

de recuerdo de clase negativa. Estas medidas fueron evaluadas en las cuatro técnicas de clasificación trabajadas.

Tabla 6. Indicadores matrices de confusión

	Arboles de decisión	Bayes Naive	Regresión Logistica	K - Vecinos más cercanos
Clasificados correctamente	49.146	48.563	45.631	38.703
Clasificados erradamente	12.287	12.870	15.802	22.730
Precisión	80%	79%	65%	63%
Error	20%	21%	35%	37%
Cohens Kappa	0,46	0,42	0,28	0,00
Verdaderos Positivos	23.095	22.458	22.997	21.753
Falsos Positivos	7.949	6.422	12.565	9.984
Falsos Negativos	4.337	6.448	3.237	12.746
Verdaderos Negativos	26.051	26.105	22.634	16.950

Fuente: Elaborado por el autor.

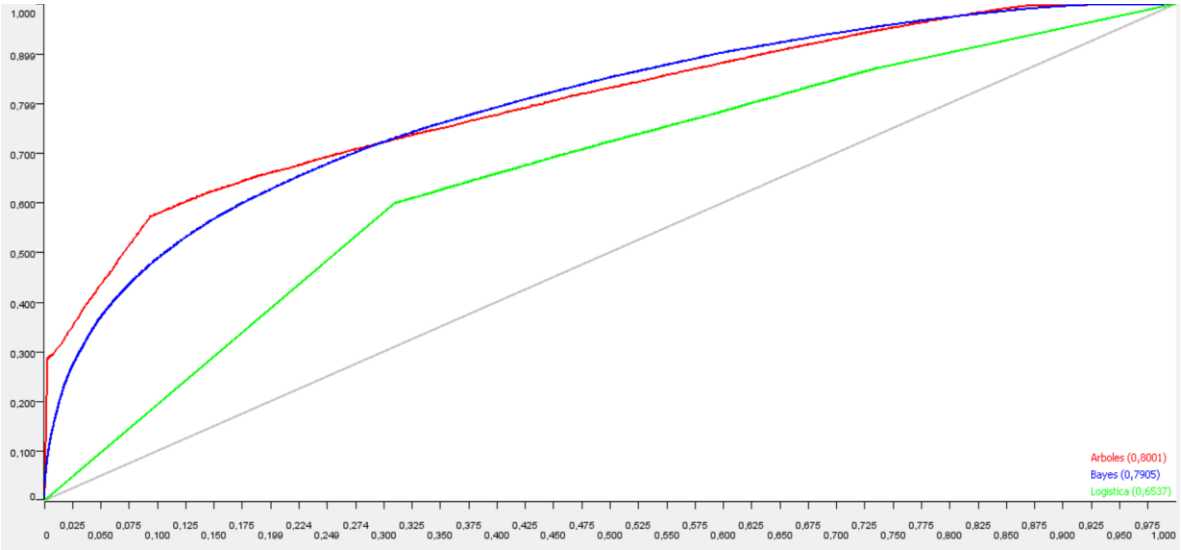
El resultado de cada algoritmo es acorde para el análisis. Esto se ve reflejado en la precisión de verdaderos positivos y falsos negativos que se estimaron frente a los datos de prueba. Si bien, el algoritmo desempeñará el mejor rendimiento en los datos en vivo, es necesario realizar la comparación de los modelos en paralelo con respecto a su sensibilidad y especificidad.

El coeficiente Cohens Kappa, determina el efecto del azar sobre la predicción, este indicador está relacionado con la probabilidad de un evento según los observadores, en el caso concreto de los reclamos ejecutados por las promotoras, entre más alto sea el coeficiente Kappa, indica que existe menos sesgo entre el resultado real y la predicción, denotando que Arboles de Decisión presenta la mejor predicción.

Las curvas ROC, que grafican la sensibilidad en función de la especificidad para todos los umbrales posibles, ilustran la compensación de un clasificador entre verdaderos positivos y falsos negativos. Un mayor valor de sensibilidad para un determinado valor de especificidad indica un mejor rendimiento. El área bajo la

curva ROC (AUC) es una métrica de uso común para evaluar el rendimiento de un clasificador (Jokiel-Rokita and Topolnicki 2020).

Figura 7. Curva ROC



Fuente: Programa Knime Analytics Platfom

La estimación comparativa de los modelos aplicando los modelos bajo datos de entrenamiento que comprenden el 70% del total de los datos indicó que con una precisión del 80% asociada a la sensibilidad y especificad de los datos en relación con la tasa de verdaderos positivos y falsos positivos. La técnica arboles de decisión fue la que obtuvo los mejores resultados modelar el problema de clasificación que busca identificar promotoras con falsos reclamos sin generar sobre ajuste en el modelo.

7.3. RESULTADOS E IMPLEMENTACIÓN

Una vez se identificó mediante datos de entrenamiento que el modelo más adecuado es arboles de decisión, se probó y estudió el modelo con el 30% de datos no utilizados en el entrenamiento, el desarrollo de estos datos corroboraría la precisión del modelo y a la viabilidad de la implementación en vivo.

El resultado del modelo es la vista ramificada de las decisiones en base a su probabilidad, mostrando los nodos que identifican las características de división, las ramas que identifican los valores de división de cada nodo. Al igual que el resultante de la predicción con la matriz de confusión detallada en el Anexo 3.

Dentro de modelo de Arboles de Decisión se encontró que la variable que predomina en la promotora al momento de realizar un reclamo es el día de ventas, con un resultado una precisión de del 73% sobre los datos evaluados, encontrando que de cada 10.817 promotoras cuyo día de venta es el viernes, 6.634 realizaran un reclamo de productos faltantes.

La variable estado socioeconómico también indicó relevancia en el comportamiento de la promotora, encontrando una correlación del 82%, asociando que de cada 7.499 promotoras cuyo estrato es inferior a 3, generaran reclamo 4.415

Alineando los resultados del algoritmo con el objetivo a conocer, se propuso implementar la evaluación del algoritmo en vivo para la campa de ventas subsiguiente, en la cual se estimará, más de 170.000 registros a evaluar.

8. CONCLUSIÓN

En esta investigación se evaluaron los modelos de clasificación supervisada para predecir reclamaciones en el caso de un modelo de ventas directas, logrando

detallar que arboles de decisión provee la mejor predicción en términos de precisión y ajuste.

El algoritmo de árboles de decisión encontró las variables relevantes que inciden en el comportamiento de las promotoras de una empresa de venta directa para la ejecución de un reclamo falso. El proceso de predicción ejecutado implica la caracterización previa de las promotoras mediante las variables asociadas al comportamiento, determinando que los días disponibles de ventas y el estado socioeconómico son las variables que afectan con mayor incidencia en el comportamiento de la promotora con respecto a la generación de un reclamo.

La consistencia entre los indicadores de especificidad y sensibilidad de todos los algoritmos trabajados concuerda con los resultados en los datos de entrenamiento, en donde arboles de decisión obtuvo el menor error en la predicción y el modelo se desempeñó correctamente con el 30% de los datos no probados.

Con los resultados se propone una metodología de control de reclamaciones que están en función de la predicción y del chequeo realizado por el transportador en el momento de la entrega del pedido. Esta metodología pasa a ser una acción predictiva basada en la estadística aplicada y en la analítica, dejando atrás la aleatoriedad y la disponibilidad del transportador. El control de las reclamaciones se practicó en las campañas subsiguientes, en donde se resaltaron las promotoras que reclamarían y se le otorgó esta información al transportador, quien recibió una capacitación para la realización de inventario del pedido al momento de la entrega.

A nivel empresarial esta metodología tiene proyectado vincularse con las aplicaciones digitales ya existentes en la empresa, con tal de que la predicción de errores y fraudes a partir de modelos de clasificación supervisada se puedan realizar de manera ágil a lo largo de la cadena de suministro. Específicamente en el área de Distribución se vinculará con el software de ruteo inteligente, de esta manera los transportadores integraran las entregas con el chequeo de posibles reclamos bajo la misma aplicación y en el mismo dispositivo.

9. BIBLIOGRAFÍA

Van Belle, Jan, Paul Valckenaers, and Dirk Cattrysse. 2012. “Cross-Docking: State of the Art.” *Omega* 40(6): 827–46. http://dx.doi.org/10.1016/j.omega.2012.01.005.

Bergmann, Felix M., Stephan M. Wagner, and Matthias Winkenbach. 2020. “Integrating First-Mile Pickup and Last-Mile Delivery on Shared Vehicle Routes for Efficient Urban e-Commerce Distribution.” *Transportation Research Part B: Methodological* 131: 26–62.

Bu, Gewu, and Maria Potop-Butucaru. 2018. “FIFO Order Reliable Convergecast in WBAN.” *Computer Networks* 146: 200–216.

Cabrejos-burga, Raúl, César A Bernal-torres, Tamara Pando-ezcurra, and Edgar Y Mayorga. 2020. “Una Visión Integral de Personas Con Trayectoria Laboral En Venta Multinivel En Bogotá (Colombia) y Lima (Perú) A Comprehensive View of People with Career Paths in Multilevel Sales in Bogota (Colombia) and Lima (Peru).” 31: 117–32.

Castellón González, Pamela, and Juan D. Velásquez. 2013. “Characterization and Detection of Taxpayers with False Invoices Using Data Mining Techniques.” *Expert Systems with Applications* 40(5): 1427–36.

Cohen, Shahar, Lior Rokach, and Oded Maimon. 2007. “Decision-Tree Instance-Space Decomposition with Grouped Gain-Ratio.” *Information Sciences* 177(17): 3592–3612.

Corso, Cynthia Lorena. 2009. “Aplicación de Algoritmos de Clasificación Supervisada Usando Weka.” *Universidad Tecnológica Nacional, Facultad Regional Córdoba*: 11. http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf.

Dietz, Christian et al. 2020. “Integration of the ImageJ Ecosystem in KNIME Analytics Platform.” *Frontiers in Computer Science* 2(March): 1–17.

Erkut, Burak. 2016. “Structural Similarities of Economies for Innovation and Competitiveness - a Decision Tree Based Approach.” *Studia Oeconomica Posnaniensia* 4(5): 85–104.

Gibert, Karina, and Ulises Cortés. 1997. “Weighting Quantitative and Qualitative Variables in Clustering Methods.” *Mathware & soft computing* 4(3): 251–66.

Gutierrez Arenas, Andrea. 2019. “Las Tecnologías de La Información y Las Comunicaciones En El Proceso de Ventas Por Catálogo Como Un Apoyo Para La Gestión Del Vendedor: Caso de Estudio En Medellín, Antioquia.” *JSR Funlam Journal of Students’ Research* (4): 146–58.

Janjevic, Milena, and Matthias Winkenbach. 2020. “Characterizing Urban Last-Mile

Distribution Strategies in Mature and Emerging e-Commerce Markets." *Transportation Research Part A: Policy and Practice* 133(January): 164–96. https://doi.org/10.1016/j.tra.2020.01.003.

Jokiel-Rokita, Alicja, and Rafał Topolnicki. 2020. "Estimation of the ROC Curve from the Lehmann Family." *Computational Statistics and Data Analysis* 142: 106820. https://doi.org/10.1016/j.csda.2019.106820.

Khoshgoftaar, Taghi M., and Edward B. Allen. 1999. "Logistic Regression Modeling of Software Quality." *International Journal of Reliability, Quality and Safety Engineering* 6(4): 303–17.

Kleyner, Andre, and Peter Sandborn. 2004. "A Warranty Forecasting Model Based on Piecewise Statistical Distributions and Stochastic Simulation." : 1–8.

Liang, X. W. et al. 2020. "LR-SMOTE — An Improved Unbalanced Data Set Oversampling Based on K-Means and SVM." *Knowledge-Based Systems* 196.

Liliana, Luca. 2016. "A New Model of Ishikawa Diagram for Quality Assessment." *IOP Conference Series: Materials Science and Engineering* 161(1): 0–6.

Lin, Hsien I., and Mihn Cong Nguyen. 2020. "Boosting Minority Class Prediction on Imbalanced Point Cloud Data." *Applied Sciences (Switzerland)* 10(3).

Luis Fernández Martínez, José, Francisco B Cobo Quesada, and Antonio Sánchez-Bayón. 2017. "Régimen Jurídico-Económico De La Venta Directa: Estudio Histórico-Comparado Y Su Situación Actual En España." : 1–38. www.derechoycambiosocial.com│.

May, Published. 2016. "Global Direct Selling - 2014 Retail Sales." (5): 2014–15.

Miriam, Dra, and Martínez Arroyo. 2019. "MERCADOLIBRE." 11(6): 1326–32.

Naik, Amrita, and Lilavati Samant. 2016. "Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime." *Procedia Computer Science* 85(Cms): 662–68. http://dx.doi.org/10.1016/j.procs.2016.05.251.

Portafolio. 2019. "Alimentos y Bebidas, En La Conquista de La Venta Directa." *Portafolio*. https://www.portafolio.co/negocios/empresas/alimentos-y-bebidas-en-la-conquista-de-la-venta-directa-525074.

Ravelin. 2002. "Machine Learning for Fraud Detection." (May). https://www.ravelin.com/insights/machine-learning-for-fraud-detection?utm_term=fraud machine learning&utm_campaign=Awareness+-+Machine+learning&utm_source=adwords&utm_medium=ppc&hsa_ver=3&hsa_grp=58391109750&hsa_kw=fraud machine learning&hsa_tgt=kwd-30262.

Ricciardi, Carlo et al. 2020. "Application of Data Mining in a Cohort of Italian Subjects Undergoing Myocardial Perfusion Imaging at an Academic Medical Center."

Computer Methods and Programs in Biomedicine 189: 105343. https://doi.org/10.1016/j.cmpb.2020.105343.

Romero, Juan Guillermo, Oscar Arley Riveros, and Jhon Francined Herrera. 2017. “Implementación de La Técnica de Los K-Vecinos En Un Algoritmo Recomendador Para Un Sistema de Compras Utilizando NFC y Android.” *Inge Cuc* 13(1): 9–18.

Shahzaad, Babar, Athman Bouguettaya, Sajib Mistry, and Azadeh Ghari Neiat. 2021. “Resilient Composition of Drone Services for Delivery.” *Future Generation Computer Systems* 115: 335–50. https://doi.org/10.1016/j.future.2020.09.023.

Tayauova, Gulzhanat. 2012. “Advantages and Disadvantages of Outsourcing: Analysis of Outsourcing Practices of Kazakhstan Banks.” *Procedia - Social and Behavioral Sciences* 41: 188–95. http://dx.doi.org/10.1016/j.sbspro.2012.04.023.

Trkman, Peter, Kevin McCormack, Marcos Paulo Valadares De Oliveira, and Marcelo Bronzo Ladeira. 2010. “The Impact of Business Analytics on Supply Chain Performance.” *Decision Support Systems* 49(3): 318–27.

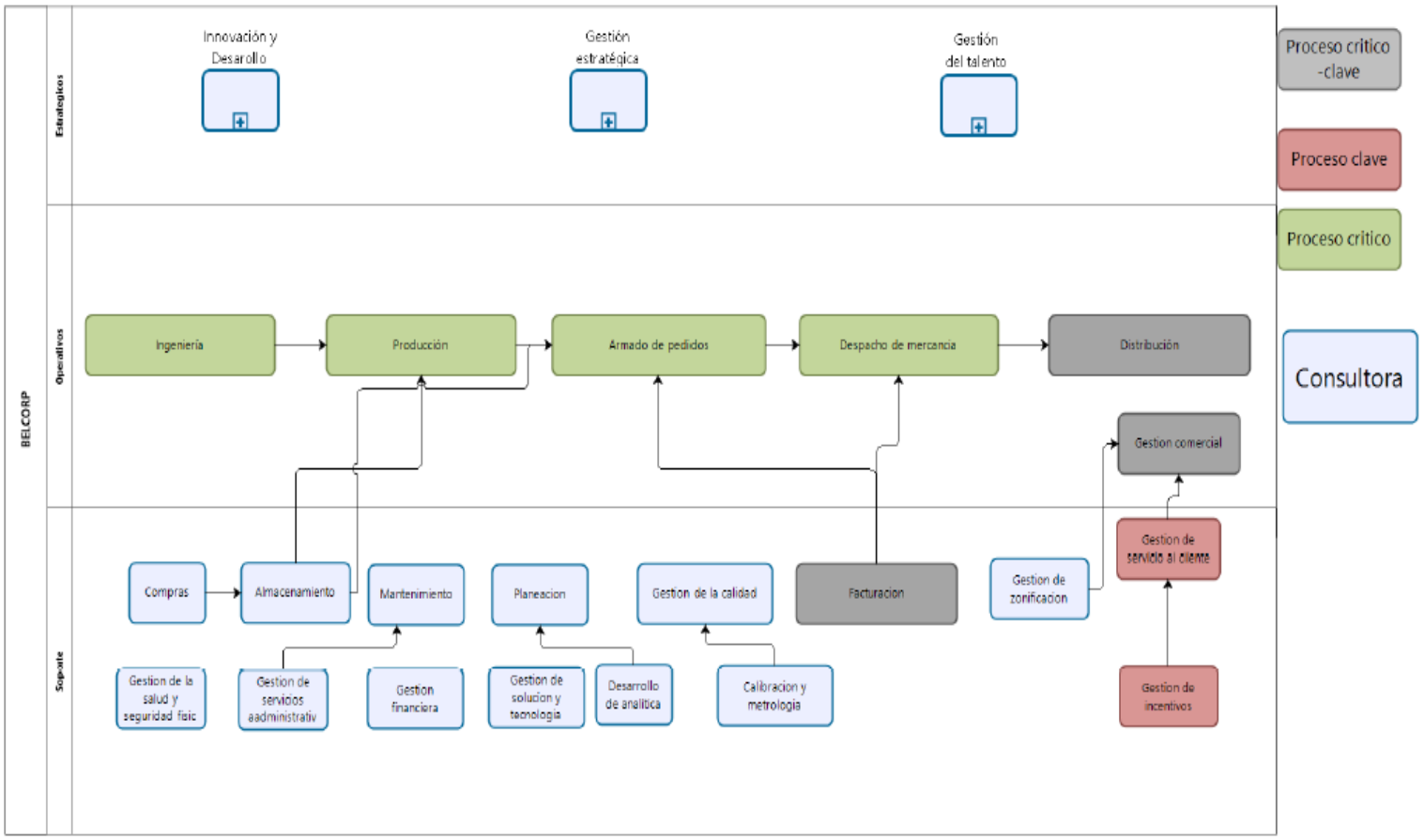
Witkowski, Krzysztof. 2017. “Internet of Things, Big Data, Industry 4.0 - Innovative Solutions in Logistics and Supply Chains Management.” *Procedia Engineering* 182: 763–69. http://dx.doi.org/10.1016/j.proeng.2017.03.197.

Xu, Jianfeng, Yuanjian Zhang, and Duoqian Miao. 2020. “Three-Way Confusion Matrix for Classification: A Measure Driven View.” *Information Sciences* 507: 772–94. https://doi.org/10.1016/j.ins.2019.06.064.

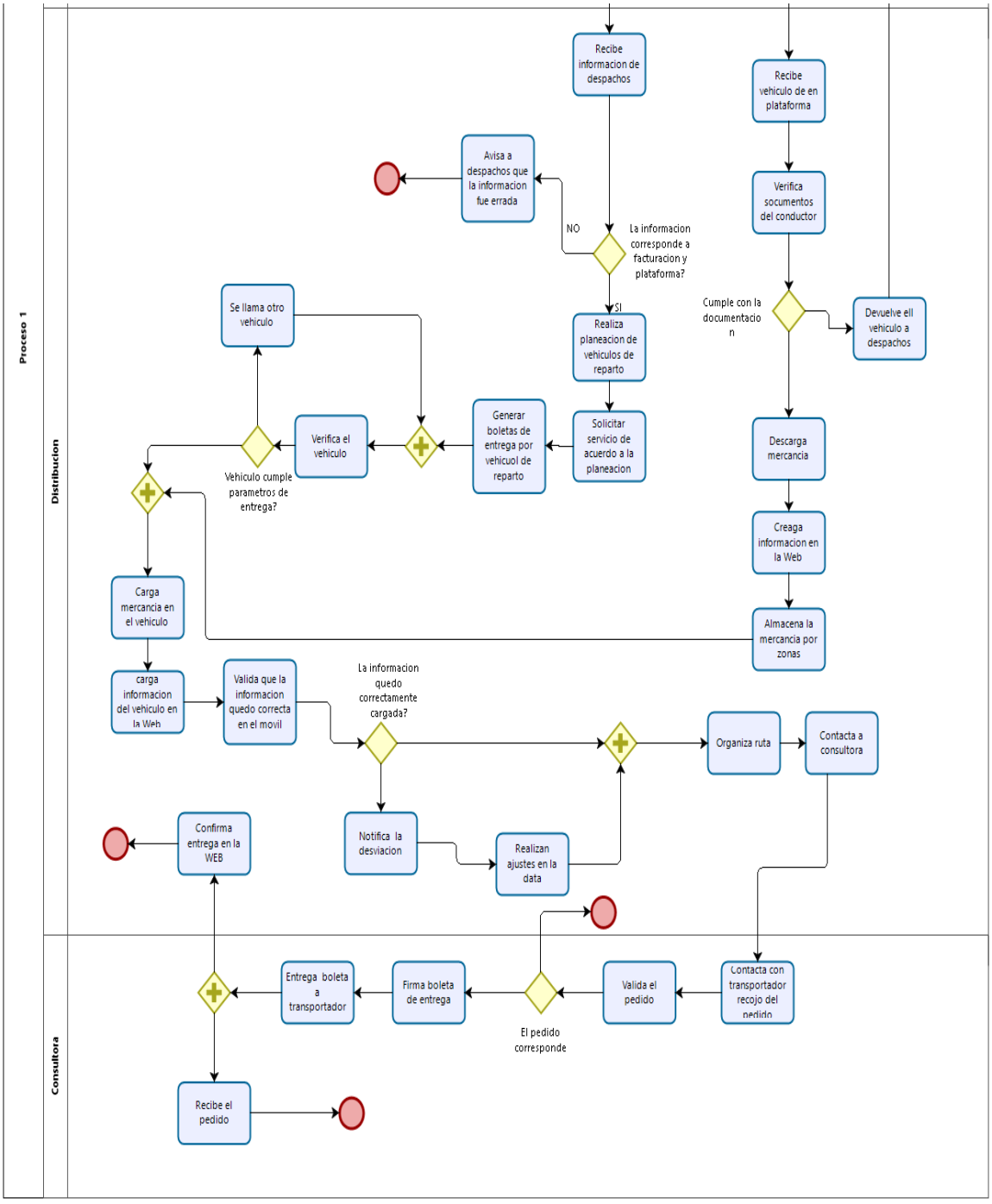
Zhou, Feng, Qun Zhang, Didier Sornette, and Liu Jiang. 2019. “Cascading Logistic Regression onto Gradient Boosted Decision Trees for Forecasting and Trading Stock Indices.” *Applied Soft Computing Journal* 84: 105747. https://doi.org/10.1016/j.asoc.2019.105747.

ANEXOS

Anexo 1. Diagrama de relaciones de una empresa que contienen todos los eslabones de la cadena de abastecimiento (Recepción. Fabricación, Distribución y comercialización) en el enfoque de productos cosméticos bajo el modelo de venta directa.



Anexo 2. Diagrama de proceso, proceso Distribución



Anexo 3. Matriz de confusión, modelo con el 100% de los datos.

Fila	P reclamantes	P No reclamantes
P reclamantes	357	2.802
P No reclamantes	2.464	20.705