Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

EXPERIENCIA DEL CLIENTE EN E-COMMERCE, UN ENFOQUE DESDE LA MINERÍA DE DATOS. EL CASO DE OLIST

AUTOR: JUAN IGNACIO BIGOURDAN

MENTOR: NELIDA MONICA CANTONI RABOLINI

DICIEMBRE 2020

Resumen

El presente trabajo propone analizar bases de datos de una compañía de comercio electrónico utilizando herramientas de análisis del tipo cuantitativo en un marco de trabajo ágil. Haciendo foco en las necesidades del cliente, se plantea encontrar variables que incidan en la experiencia de compra en la empresa brasileña Olist, de E-Commerce. Para ello, se utilizará el lenguaje SQL para unificar la información e incorporar variables de análisis. Luego, se utilizará un algoritmo llamado *Stacking* para combinar tres modelos predictivos que buscarán predecir la satisfacción de los clientes, a saber, *Random Forest, Rule Induction* y Regresión Logística. Finalmente, se hará una introducción a las metodologías ágiles y se seleccionarán dos, Kanban y Scrum, para llevar adelante este proyecto en Olist. Este estudio, por tanto, buscar poner en evidencia que la experiencia del cliente no debe valerse únicamente de técnicas cualitativas para priorizar acciones y que contar con un marco de trabajo ágil y recursos para hacer análisis desde la minería de datos es de vital importancia para aumentar los ingresos en las plataformas de E-Commerce y en el resto de las organizaciones.

Palabras clave: Experiencia del cliente, E-Commerce, Olist, Modelos predictivos, Minería de datos, Metodologías ágiles

Índice

1. Introducción

En un mercado como el actual, competitivo y saturado, contar con ventajas sobre el resto de la competencia se traduce en mayores beneficios para las organizaciones. La puja entre los oferentes de productos y servicios por ampliar su cartera de consumidores y retener los clientes ha sido una constante a lo largo de toda la historia. Esta competencia empujó a las empresas a crear un sinfín de estrategias, acciones y modelos, derivando en el surgimiento de disciplinas enteras estudiadas en universidades de todo el mundo. Es también gracias a esta competencia, que cada vez es más usual encontrar técnicas y métodos estadísticos aplicados en las empresas ya que con ellas es posible llevar adelante un correcto manejo de grandes volúmenes de datos llegando a conclusiones más acertadas y permitiendo incluso anticiparse a eventos con cierto grado de probabilidad.

Los modelos predictivos son el resultado de agrupar técnicas estadísticas y utilizarlas para analizar datos históricos y actuales para, como su palabra lo indica, realizar predicciones sobre sucesos inciertos que aún no ocurrieron. De su aplicación, se obtiene una predicción para cada sujeto analizado y una probabilidad de ocurrencia. Son de gran utilidad en las organizaciones ya aportan información valiosa para la toma de decisiones. Con niveles altos de precisión se puede confirmar, corregir y establecer campos de acción futuros para lograr mayores beneficios.

Hoy en día, escuchar el término *Customer Experience* (Experiencia del Cliente) es habitual en compañías. Sin embargo, lo que se entiende por Experiencia del Cliente tiene una historia reciente y, por tal motivo, aún no posee una única acepción puertas adentro de las organizaciones. Las interpretaciones implican distintas formas de gestionar la experiencia de los clientes. Esto puede apreciarse en el siguiente ejemplo: si la experiencia es entendida como el resultado entre la expectativa de atención del cliente y la que obtiene realmente, entonces las tareas van a estar centradas en las áreas de servicio al cliente; por otra parte, si se la considera más ampliamente y pesa, no sólo la atención recibida, sino también todo su *Customer Journey* (Viaje de Cliente), las acciones van a estar orientadas en generar al cliente una emoción positiva que se guarde en el recuerdo y posicione a la marca por encima del resto.

En este contexto de evolución constante en lo referido a experiencia, también se dieron cambios en el ámbito de las investigaciones. Hace no mucho tiempo (incluso en la actualidad sigue sucediendo), cuando las empresas precisaban respuestas o información sobre los clientes acudían a institutos o compañías de investigación. Estos institutos utilizaban encuestas, se apoyaban en datos internos de la propia empresa y realizaban estudios cualitativos para dar respuestas a las inquietudes y con ello se tomaban decisiones. Lo que ocurre con este tipo de estudios es, en primer lugar, que uno debe descansar en que la metodología utilizada es correcta. Además, los datos internos utilizados no los conocen a detalle como los empleados de la misma organización. Por último, conserva la dependencia con un tercero en caso de surgir dudas lo que provoca mayores costos. Por todas estas contras, algunas compañías incorporaron dentro de sus estructuras, las gerencias de Voz del Cliente o Experiencia del Cliente que se encargan de llevar las necesidades del cliente a toda la organización y tratan de dar una respuesta diferenciadora a sus consumidores.

Este informe nace en una etapa de expansión de lo entendido como *Customer Experience* y cómo abordar su análisis, y en un período en el que el E-Commerce se volvió de vital importancia en Argentina (y en el mundo) por la pandemia mundial. Es que el comercio electrónico, a diferencia de la mayoría de las empresas con relación directa con clientes, nace en la época de expansión del negocio centrado en las necesidades del cliente. Por esta razón, incorporar la minería de datos al área de experiencia del cliente es fundamental para el crecimiento del negocio ya que permite priorizar variables que predicen la experiencia que los usuarios, generar análisis posteriores para profundizar con otras herramientas cuantitativas o cualitativas, derribar mitos que no pueden falsearse hasta no mostrar datos concretos y, por último, establecer jerarquías en las mesas de decisión para que los gerentes tomen partida por los proyectos con mejores resultados en términos de experiencia del cliente y, al mismo tiempo, más rentables.

Anteriormente se mencionó que definir experiencia del cliente presenta cierta complejidad por sus múltiples aseveraciones, sin embargo, es una tarea necesaria para el desarrollo de esta investigación. El contar con tantas acepciones se origina en el hecho de que cada persona experimenta de forma distinta a la otra. Nuestros sentidos, nuestras percepciones no son lineales ni unidimensionales y, por lo tanto, nuestra experiencia como

clientes tampoco lo es. La forma más clara de ver esta complejidad es en las encuestas donde para un mismo *customer journey*[1] (viaje del cliente) dos clientes puntuar distinto en el *CSAT*[2].

Para comprender la experiencia del cliente y cómo fueron integrándose todos sus conceptos, se utilizarán las etapas históricas del marketing propuestas por Lemon, Katherine N. y Verhoef, Peter C. (Lemon & Verhoef, 2016, págs. 71-73). Estas permitirán lograr un mayor entendimiento de los orígenes del estudio a los clientes y cómo estos análisis influyeron en la disciplina en cuestión para generar más valor en las organizaciones.

De acuerdo a estos autores, entre los años 1960 y principios de 1970, surgieron los primeros modelos orientados a entender mejor a los clientes, los modelos sobre el comportamiento en el proceso de compra del cliente. En esta etapa comenzó a estudiarse en los clientes cómo funcionaba el proceso de identificar una necesidad, satisfacerla con la compra y el rol de la publicidad en la generación de nuevas necesidades. Más tarde, a mediados de 1970, surgió el concepto de satisfacción del cliente y con ello la medición orientada a conocer qué tan satisfecho se encontraba el cliente con lo consumido. Lo más importante de este momento fue que, por primera vez, aparecieron diferencias entre las expectativas del cliente y la oferta de las firmas. Seguidamente, comenzó la etapa de poner foco en la calidad de servicio. Era la década de 1980 y nacía el concepto de momento de la verdad3 para el cliente, y comenzaron los primeros mapeos del viaje del cliente para mejorar su experiencia.

Los noventa fueron el escenario del marketing relacional. En esta etapa, se profundizó el entendimiento del servicio al cliente, se incorporaron las emociones y las distintas percepciones de los clientes al concepto de experiencia del cliente. Luego, sobrevino el nuevo mileno y, con él, el quinto foco del marketing: el Customer Relationship Management (CRM). Este enfoque se caracteriza por relacionar elementos que componen la experiencia de los clientes y aprovecharlos para generar mayores beneficios. Durante esta década también

[1] Todas las interacciones, los canales disponibles y elementos por los que atraviesa un cliente durante todo el ciclo de compra.
[2] Customer Satisfaction (CSAT) es un tipo de encuesta donde se le pregunta al cliente que tan satisfecho se encuentra con el canal, producto o servicio utilizado, comprado o recibido.
[3] El instante en que el cliente se pone en contacto con la firma y forma una opinión acerca de la calidad del mismo.

emergió el marketing centrado en el cliente, en el que el objetivo pasó a ser, mediante el aprovechamiento de los grandes volúmenes de datos, acercarse a la los clientes para entenderlos individualmente. Por último, en el año 2010, comenzó la etapa de poner la atención sobre la experiencia del cliente para fidelizarlo. En esta época, el cliente pasó a tener mayor poder en la relación con la empresa ya que la revolución de las redes sociales le dio otro lugar, un lugar de co-creador o de destructor de valor. La experiencia pasó a estar más allá del proceso de compra, el generar una experiencia diferenciadora precisa comprender las actitudes, los comportamientos y extraer valor de los clientes de forma personalizada, para esto la identificación con la marca se volvió fundamental.

La historia del marketing da cuenta de la complejidad de definir la experiencia del cliente. El paso de los años incorporó más conceptos que permitieron definir a la experiencia del cliente. De esta forma, se llegó a que es una construcción multidimensional basada en todas interacciones que el cliente tiene con la organización durante la duración de la relación comercial. Siendo estas interacciones generadoras de distintas reacciones por pate de los clientes desde el punto de vista cognitivo, emocional, sensorial y social, ya que en cada una de las interacciones, el cliente llega con una cierta expectativa. Por tanto, la experiencia de un individuo se mide durante todos los puntos de contacto como la diferencia entre sus expectativas y el servicio recibido de parte de la empresa. (WEREDA & GRZYBOWSKA, 2016, pág. 199)

En la actualidad, las empresas se enfrentan a una competencia diferente a décadas atrás. Las masividad de la disponibilidad de internet y del uso redes sociales dio a los clientes la posibilidad de elegir entre muchas más opciones, en menor tiempo y con la posibilidad de comparar precios en el momento. La importancia del precio en la elección del producto o servicio perdió peso porque las compañías entendieron que si competían por precio iban a desaparecer. En este contexto, el ofrecer experiencias memorables que puedan generen un aumento sostenido de las ventas se volvió vital para las empresas. Aquí es donde emerge el concepto de *Customer Experience Management* definido como:

> (…) customer experience management es más que atender a sus clientes en línea. Se trata de conocer a sus clientes tan bien que pueda crear y proporcionar experiencias personalizadas que los lleven, no solo a permanecer leales a su

> marca, sino también a atraer a otros a acercarse a ella, y esa es la forma de publicidad más valiosa que existe. (ANETCOM, 2013)

El repentino aumento del interés por la experiencia del cliente responde entonces a los réditos que trae aparejados una correcta gestión de la misma y, es en este punto, donde las técnicas de la minería de datos se incorporaran en el sector de la experiencia del cliente para brindar más información para tomar decisiones. La minería de datos, de acuerdo a Hung, S., Yen, D.C., & Wang, H., es el proceso que abarca la exploración, selección, análisis y modelado de grandes volúmenes de datos para identificar patrones que permitan extraer información que pueda ser utilizada en las empresas para obtener una ventaja que genere nuevas oportunidades (Hung, Yen, & Wang, 2006, págs. 1125-1126). Incorporar técnicas de la minería de datos en *customer experience* trae aparejados una serie de beneficios entre los que se destacan: fortalecer la fidelización con la marca, aumentar los ingresos con ventas incrementales de clientes actuales y de las nuevas producto del boca a boca, generar lealtad en el cliente y, con ello, crear "defensores" de la marca en las redes sociales y, por último, reducir los costos ya que disminuye la imprevisibilidad en la cantidad de clientes y las ventas. Esto resuelta en lograr un mejor conocimiento de los clientes lo que permite crear y entregar experiencias personalizadas que los harán no solo fieles a la organización, sino también promotores de sus productos generando comentarios positivos en redes sociales que es, actualmente, la forma más valiosa y eficaz de hacer marketing para cualquier organización.

Durante la pandemia, son de público conocimiento los problemas económicos que enfrentaron la mayoría de las empresas de distintos rubros. Sin embargo, un sector se vio beneficiado por esta crisis sanitaria, el E-Commerce. Ejemplo de esto es Mercado Libre, que más que duplicó su facturación en el segundo trimestre del 2020 respecto al mismo trimestre el año pasado (Lafuente, 2020). Para las empresas enfocadas en el comercio electrónico, la experiencia del cliente juega un rol central en la toma de decisiones de negocio. Tal es así, que destinan cuantiosos recursos para: acercar los productos a los usuarios de forma personalizada, dar previsibilidad sobre la fecha de llegada, comunicar simple los procesos de devolución y simplificar los costos asociados al servicio.

El objetivo del presente informe es predecir la experiencia de los clientes que realizan compras online con la plataforma Olist empleando técnicas de la minería de datos en una organización que podría utilizar una metodología de trabajo ágil. Para lograrlo, se

perseguirán los siguientes objetivos específicos: Describir la organización y el contexto en el cual se desarrolla, hallar el modelo o combinación de modelos que prediga con mayor precisión la satisfacción de los clientes con el proceso de compra identificando las variables más relevantes para la experiencia de los clientes y, por último, describir cómo podría implementarse ese modelo en un marco de trabajo ágil. Se muestra en cuatro secciones. En primer lugar, realiza la presentación de la organización, cómo es la gestión de los datos y las problemáticas con las que lidia. En segundo lugar se realiza la descripción de los aspectos metodológicos del trabajo: se presenta la base de datos, la estructura del conjunto de datos de la base de datos original y las variables que se incluyeron en el análisis. Luego, se presentan las herramientas utilizadas para el tratamiento de los datos y se detalla el proceder en la limpieza de la base de entrenamiento y de testeo para la aplicación de modelos. Una vez detallado lo anterior, se presentar un análisis descriptivo de las variables seguido por la explicación de los tres modelos aplicados y métricas elegidas para predecir, el detalle de los resultados obtenidos comparando los modelos aplicados y su interpretación. En tercer lugar, se definen las metodologías de trabajo ágiles y cómo llevarlas adelante en el contexto de la organización bajo estudio. Finalmente, se explican las conclusiones del análisis previo, el conocimiento generado, las principales limitaciones y las posibles mejoras que pueden aplicarse al modelo.

2. Gestión de datos en contextos organizacionales

Para poder comenzar con cualquier análisis es necesario el entendimiento de los datos con los que se trabajar. Entenderlos conlleva entender la organización, el contexto en el que se desarrolla, su arquitectura de datos y el sector en el que se desarrolla. El presente capítulo permite entender el contexto de la organización que es objeto de estudio y la gestión que realiza de sus datos. Se realizará, en primer lugar, una descripción de la organización donde entenderemos su historia, su misión, su negocio y el contexto que atraviesa. Luego, se procederá a describir los datos con los que Olist cuenta, cómo los estructura y cómo los utiliza. Finalmente se presentarán las principales problemáticas con las que se enfrenta la organización en la gestión de sus datos.

2.1. Descripción de la Organización

El presente trabajo se desarrolla sobre una base de datos de una organización llamada Olist. Esta compañía de *E-Commerce*, radicada en Curitiba, Brasil, tiene como actividad principal ofrecer soluciones de ventas y brindar servicios a comerciantes y empresas que desean vender sus productos por internet. En su sitio web, Olist revela su misión (Olist;, 2020): fortalecer el comercio en el mundo. Para ello, ofrecen una cuatro servicios: importar los productos fácilmente, lo que implica que ofrecen un *onboarding* simple a su plataforma de venta online, utilizar anuncios registrados previamente, que permite a los nuevos usuarios tomar modelos exitosos de publicidad pre armados para comenzar a publicitar los productos desde el momento en el que se contratan los servicios, identificar los mejores precios, que se realiza a través de una herramienta desarrollada por la compañía y permite comparar los precios del mismo producto que vende el cliente en otras plataformas, y por último, ofrece la buena reputación de la marca como motor de ventas.

Lo que distingue a Olist es su historia. Antes de ser lo que es, Olist fue una tienda física, luego un distribuidor y, más tarde, un *marketplace*, hasta convertirse en una solución para el *retail* online. En el año 2007, nació con el nombre de Solidarium con la misión de convertirse en el mayor mercado de artesanos de Brasil. Sin embargo, el fundador de Olist, Tiago Dalvi, advirtió que el desafío de acceder a las principales cadenas minoristas no era sólo un

problema de los artesanos, así fue que en 2005 creó Olist con la misión de ayudar a todos y cada uno de los comerciantes a llegar a los mejores y más grandes mercados nacionales e internacionales. Por tanto, Olist posee un gran conocimiento de las problemáticas de los pequeños comerciantes y lo difícil que es iniciarse en el comercio electrónico.

Para el año 2019, de acuerdo al sitio web Contxto (Contxto, 2019), la compañía ya contaba con 7 mil usuarios utilizando el software. De estos, los clientes de mayor cuantía eran artesanos y pequeños comerciantes que no estaban en línea y contrataron los servicios para vender online. Sin embargo, también había comercios que ya vendían por la web y buscaban un mejor posicionamiento de producto, inquilinos que buscaban nuevos canales de venta o clientes, consumidores finales, importadores y distribuidores.

El objetivo de Olist es que los usuarios pueden registrar sus productos y vender en las principales tiendas minoristas en Brasil. Pero también, Olist incorpora productos en otros *marketplaces* importantes. Algunos de estos incluyen a Amazon, Mercado Libre y Viar Varejo. Al contratar los servicios, reciben un panel operativo para el control de una sola pantalla, además de numerosas funciones de sistema de planificación de recursos empresariales y se les ofrece una categorización automática de artículos a través de IA (Inteligencia Artificial), así como comparaciones de precios de mercado para tener los precios más competitivos.

Atendiendo las necesidades de los comerciantes Olist también busca satisfacer la de los consumidores. La empresa ofrece soluciones de envío y trabaja mejorando los flujos dentro de su plataforma de ventas para simplificar lo máximo posible el proceso de compra. De esta forma, Olist desarrolla un negocio basado en la tecnología y los datos (de consumidores y vendedores) para aumentar las ventas de sus clientes y generar más ingresos que son un porcentaje sobre cada venta de los clientes.

Durante la pandemia, la plataforma de mercado vio cómo su crecimiento se disparó a la vez del aumento del comercio electrónico brasileño. Hoy en día, ya son más de 90 mil minoristas que utilizan los servicios para la gestión de ventas online y la logística (Gazzeta Do Povo, 2020). De acuerdo al fundador Dalvi, en el mercado actual, de gran dinamismo y con muchas empresas bien capitalizadas buscando innovar con soluciones digitales, lo que hizo posible que Olist pudiera crecer adaptarse en este contexto es el equipo. Por ello es que para Olist desafía a sus empleados, los motiva, y los contrata sólo si ellos son capaces de

adaptarse a la cultura de la compañía, a los principios que ellos denominan “valores olistas” (Olist;, 2020):

Somos personas reales

Todos somos propietarios

Somos un equipo

Lo hacemos realidad

Ganamos de la manera correcta

Estamos dedicados al éxito de nuestros clientes

El contexto actual para una empresa con la experiencia y el conocimiento del negocio como Olist es envidiable. De acuerdo a Exam, las ventas online en Brasil crecieron 47% en el primer semestre y sumó a 7.3 millones de nuevos consumidores (Exame, 2020). Esto incluso superó las proyecciones previas a la pandemia que auguraban un incremento de 18% en el total del comercio. El aislamiento social modificó el comportamiento de compra de los consumidores finales ya que no aumentó sólo la cantidad gastada por las personas sino también el monto promedio gastado por los brasileros. El cambio es tal, que según el estudio de Ebit/Nielsen mencionado en el sitio web, el 93,4% de los consumidores tenía la intención de volver a comprar online en los tres meses posteriores a su compra. Pero los cambios no fueron sólo de los consumidores, los comercios también se adaptaron:

> (…) Según la Asociación Brasileña de Comercio Electrónico (ABComm), de marzo a julio se crearon 150.000 nuevas tiendas en línea en Brasil (Exame, 2020).

En los primeros seis meses de 2020, los *marketplaces* fueron responsables de 30 mil millones de reales de ingresos por comercio electrónico, un crecimiento del 56% con respecto al mismo período en 2019 y llegaron a representar el 78% del total del comercio electrónico. Miles de comerciantes que, hasta ese momento, trabajaban solo con la tienda física, decidieron darle una oportunidad al mercado online para sobrevivir. Así, las tiendas que funcionan tanto online como offline representan hoy el 73,1% de las ventas digitales en Brasil y se llegó al registro de 57 millones de pedidos en el primer semestre del año, un 54% más que en el mismo período de 2019 (Exame, 2020).

2.2. Gestión de Datos por parte de la Organización

En este contexto de crecimiento exponencial del comercio electrónico, con incrementos en la cantidad de vendedores y de consumidores que participan de este mercado, las empresas buscan posicionarse y captar la mayor cantidad de clientes posible antes que sus competidores. Este problema no le atañe sólo a Olist, el MIT en el año 2011 en un artículo de su Management Review ya adelantaba que, en todas partes del mundo, los líderes se preguntaban si están extrayendo todo el valor de las enormes cantidades de información que ya tenían dentro de sus organizaciones. Esto quiere decir, que ya hace casi 10 años se hablaba de que las nuevas tecnologías tenían una gran capacidad de recopilar más datos pero que lo más desafiante para las organizaciones es encontrar las formas de obtener valor de sus datos y competir en el mercado (Lavalle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011, pág. 21). Hay una necesidad imperante de saber qué está sucediendo ahora, qué es lo más probable que suceda a continuación y qué debe hacerse.

Las dificultades existentes para realizar una correcta gestión de los datos por parte de las organizaciones son transversales a todos los sectores y Olist no es la excepción. En su arquitectura de datos la firma cuenta, por un lado, de datos estructurados donde almacena toda la información transaccional: tipo de productos, especificidades técnicas de los mismos, información de los clientes (vendedores), información de los compradores, ubicaciones geográficas, estados de las órdenes de compra, información de los pagos, datos de los envíos y reviews de los clientes. Por otra parte, trabaja con datos semi estructurados que viajan a la base de datos a través de archivos de proveedores externos a la firma y de los vendedores que suben imágenes que deben procesar en el momento para saber si se tratan de productos falsos y estafas. Por último, posee datos no estructurados que contienen información en tiempo real de la página de internet (mapas de calor, cantidad de visitas, favoritos, cantidad de *clicks*, entre otros) donde se encuentra montada su operatoria tanto en su plataforma como en la de los *marketplaces* que utiliza para publicitar el resto de los productos. Esta compleja trama de datos está ideada para incorporar los datos en los procesos de toma de decisión. Olist utiliza los datos para expandir su negocio y aumentar las ventas de sus clientes. Por tanto, podemos

catalogarla como una organización *data driven* (en español, guiada por datos) ya que la toma de decisiones se realiza a partir de información objetiva.

Llevar los datos en la toma de decisiones es algo que ocurre con cada vez mayor frecuencia en las organizaciones. Hoy en día, escuchar el término transformación digital, organizaciones *data driven*, es habitual y esto se debe a la irrupción del *Big Data* (grandes volúmenes de datos en castellano) a las áreas de negocios. El término *Big Data* refiere a un conjunto de técnicas y tecnologías que permiten almacenar, clasificar y analizar grandes volúmenes de datos. Otra forma en la que suele definirse este término es a partir de las características que tienen los grandes volúmenes de información conocidas tambien como las 6V: volumen, velocidad, variedad, veracidad, valor y visibilidad. El volumen refiere a la capacidad de almacenar grandes bancos de información que existe en la actualidad. Por su parte, la velocidad implica que la generación de nueva información se realiza en menor tiempo. La variedad pretende soslayar que, en la actualidad, ya no se trabaja solamente con un tipo de datos. Por el contrario, las organizaciones como Olist trabajan con datos del tipo estructurado, semi estructurados y no estructurados. En cuanto a la veracidad, está definida por la elección de los registros de datos que verdaderamente aporten valor y también se encuentra ligada a la variedad ya que los tratamientos de los datos para obtener una mayor calidad son distintos de acuerdo al tipo de dato. En quinto lugar, mencionamos el valor como característica de los grandes volúmenes de información, los datos deben llevarnos a descubrir lo que no sabe, a obtener conocimiento predictivo y comunicar historias de datos relevantes. Por último, la visibilidad hace referencia al conjunto de herramientas que permiten plasmar en gráficas toda la información de mayor importancia (El blog de John A. Carvajal, 2016).

Que Olist sea una organización *data driven* siguiendo el camino de otras grandes compañías del sector como Amazon, Alibaba y Mercado Libre no es casualidad. Un estudio realizado por Harvard a 330 compañías de Estados Unidos en el año 2012 reveló:

> (…) companies in the top third of their industry in the use of data-driven decision making were, on average, 5% more productive and 6% more profitable than their competitors. This performance difference remained robust after accounting for the contributions of labor, capital, purchased services, and traditional IT investment. It was statistically significant and economically important. and was reflected in measurable

> increases in stock market valuations. (McAfee & Brynjolfsson, 2012, págs. 5-6)
>
> [Las empresas en el top 3 de su industria en el uso de la toma de decisiones basada en datos fueron, en promedio, un 5% más productivas y un 6% más rentables que sus competidoras. Esta diferencia de rendimiento se mantuvo incluso después de considerar las contribuciones de mano de obra, capital, servicios adquiridos e inversión tradicional en tecnología. La diferencia fue estadísticamente significativa y económicamente importante, y se reflejó en aumentos en las valuaciones en el mercado de capitales]

Esto pone de manifiesto esta compañía de *retail* se encuentra a la vanguardia en lo que respecta a su arquitectura y uso de datos ya que tiene un profundo entendimiento del beneficio de utilizar los datos para la toma de decisiones.

2.3. Problemática de la Organización y la Gestión de los Datos

El camino a hacer un uso intensivo de los datos fue más simple para Olist (por ser una *start up*) que para las organizaciones tradicionales que están pasando por un proceso transformación digital y cultural en su organización. Sin embargo, esto no quiere decir que se encuentren exentos de algunas problemáticas en términos de datos. En la entrevista concedida a Gazeta Do Povo, el fundador de Olist expresó que dentro de sus objetivos actuales se encuentran formar a sus empleados, realizar inversiones en formación, buscar alianzas con establecimientos educativas y atraer talentos de todo Brasil . Esta problemática de la escasa oferta de especialistas en datos es la realidad de muchas empresas que, ante la dificultad, deben formar a sus empleados perdiendo tiempo valioso o atraer recursos de otras firmas o regiones geográficas a un mayor costo (Gazzeta Do Povo, 2020). Este problema no se da sólo con los empleados, hoy los que toman las decisiones, ya no tienen el mismo perfil profesional que los de hace unos años. Las empresas que hoy se destacan no son por tener mejores o más datos, sobresalen porque tienen líderes con objetivos claros y cuestionan los datos correctamente. Esto demuestra que contar con grandes volúmenes de datos no implica dejar de necesitar la visión de un humano sino más bien lo opuesto. Los líderes que sepan

gestionar el talento, logren comprender el mercado, pensar creativamente, detectar oportunidades, proponer novedades y extraer el máximo valor a los datos son los que van a llevar a la firma a mayores y mejores resultados (McAfee & Brynjolfsson, 2012, pág. 8). En este sentido, una empresa como Olist que atraviesa un proceso de expansión debe asegurarse de promover al capital humano correcto para generar el mayor valor posible.

Otra de las problemáticas de Olist en este contexto es la tecnología para llevar adelante los cambios. Es necesario mantener la inversión en tecnología para tener una gran capacidad de procesamiento de datos y en este rubro compite contra jugadores de tamaños colosales que se reinventan continuamente como lo son Amazon y Mercado Libre. Pero no sólo es importante la tecnología, invertir en la formación de los empleados es igualmente trascendental ya que, con la revolución de los datos, la información se crea y comparte rápidamente pero el conocimiento no está necesariamente donde antes solía estar. Por esta razón, es importante que se utilice la información correcta para responder a los problemas pero también el profesional adecuado que sepa trabajar con esos datos y que tenga espacio para aportar ideas y valor en la mesa. Esto que se encuentra íntimamente ligado al pinto anterior es vital no sólo para el aprovechamiento de los datos sino también de la tecnología con que la compañía cuenta.

Una cuarta cuestión a tener en cuenta para esta organización y toda aquella que tenga datos de clientes es el correcto manejo de la información. Se debe trabajar con minuciosa cautela y debe haber un equipo experto en seguridad de la informática para estar siempre cuidando de no poner demasiadas trabas al negocio, concientizando a los empleados para estar alertas de las amenazas y, por último, estando alerta a las nuevas herramientas de los *hackers* que se reinventan continuamente. Sin embargo, esta problemática no afecta sólo a los clientes que quieren vender sus productos, también afecta a los compradores que utilizan la plataforma y realizan transacciones monetarias. Este problema es tan actual que el Banco Galicia el 16 de septiembre de este año tuvo que cerrar su cuenta oficial de Instagram por las continuas estafas que sufrían sus clientes. Cuando desde BAE le consultaron los motivos, el CISO (Chief Information Security Officer) de la compañía expresó: “Desde que empezó la pandemia empezamos a tener muchas estafa con perfiles falsos, particularmente en Instagram, el número era cada vez mayor. Los contactaban y le daban usuario, contraseña, token y accedían al homebanking y hasta pedían préstamos haciéndose pasar por nuestros

clientes. Hicimos una fuerte campaña de concientización y seguían creyéndoles" (Moreno, 2020).

En quinto lugar podemos mencionar la necesidad de utilizar herramientas de visualización para traducir la complejidad de los datos en información simple al alcance de todos. Esta problemática es habitual cuando existe mucha variedad de datos, de distintos tipos, de distintas áreas, y se presenta cuando éstas deben interactuar entre sí. El resultado de no contar con herramientas de BI que simplifiquen la lectura de la información es la pérdida de tiempos cuando alguna área precisa de la información de otro sector. Esto se resuelve poniendo a disposición de todas las áreas reportes, tableros interactivos, armando vistas simplificadas y haciendo foco en la comunicación para concentrarse más en los resultados que en el cómo se llegó a los mismos.

Por último, Olíst no está exento de la problemática de los recursos. Estos no son ilimitados y por esto es necesaria una correcta asignación de parte de los líderes. Este punto es relevante para presente trabajo ya que su objeto es predecir la experiencia de los compradores de la plataforma de Olist a partir del análisis de sus datos transaccionales. Para poder llevar acabo esta tarea se precisan *datascientists* (científicos de datos) o *data engineers* (ingenieros de datos), activos con los que las gerencias de experiencia del cliente no siempre cuentan ya que muchas veces se las separa del análisis cuantitativo de los datos. Esto atenta contra la posibilidad de realizar un CRM (Customer Relationship Management) que se traduzca en iniciativas de negocio y mayores beneficios.

3. Descripción Metodológica

En el presente apartado se expondrán una serie de pasos implementados en este análisis para poder alcanzar los resultados. Sin dudas es la sección más importante desde la perspectiva de la minería de datos ya que se muestra la técnica con la que se trabajó con los datos. La técnica es la que permite traducir la realidad en datos reproducibles y objetivos, por tanto es necesario detenernos en este punto para que el análisis tenga la rigurosidad necesaria para ser implementada en otras organizaciones (Lam Diaz, 2005, pág. 1).

Para comenzar con la descripción de la metodología y la técnica empleadas se procederá a describir la base de datos. Luego de desarrollará la etapa de procesamiento de datos que dará lugar a la etapa de análisis. Consecutivamente se presentarán los modelos aplicados y

se seleccionará la métrica con la que se medirán los resultados. Finalmente, se mostrarán los resultados obtenidos.

3.1. Descripción de la Base de Datos

El siguiente trabajo de desarrollará utilizando la base de datos disponible en el sitio web Kaggle. El nombre de esta base es Brazilian E-Commerce Public Dataset by Olist y contiene información sobre transacciones realizadas por los clientes de la firma Olist, una plataforma de comercio electrónico situada en Brasil, entre el 2016 y 2018 (Olist & Sionek, 2018).

El data set fue publicado el 21 de septiembre de 2018 por André Sionek y Olist y se compone de información de 99.941 compras realizadas entre el 3 de octubre de 2016 y el 29 de agosto de 2018 a través de la plataforma. La última actualización realizada por el autor data del 29 de noviembre de 2018.

En cuanto a sus características, esta base se divide en ocho conjuntos de datos en formato CSV (valores separados por comas por sus siglas en inglés) estructurados como se ve en la ilustración 1 con el fin de simplificar el análisis. Se estructuran de forma matricial y contiene 36 variables mixtas (categóricas y numéricas) de las cuales se puede extraer información detallada de cada compra: estado final del pedido, monto total abonado, costo de envío, precio del producto, atributos del producto (tipo de producto, largo de la descripción, cantidad de fotos, largo, alto, ancho y peso) y, por último, las reseñas de los clientes sobre esa experiencia. En el anexo 1 se puede ver a detalle el nombre de cada variable y su tipo.

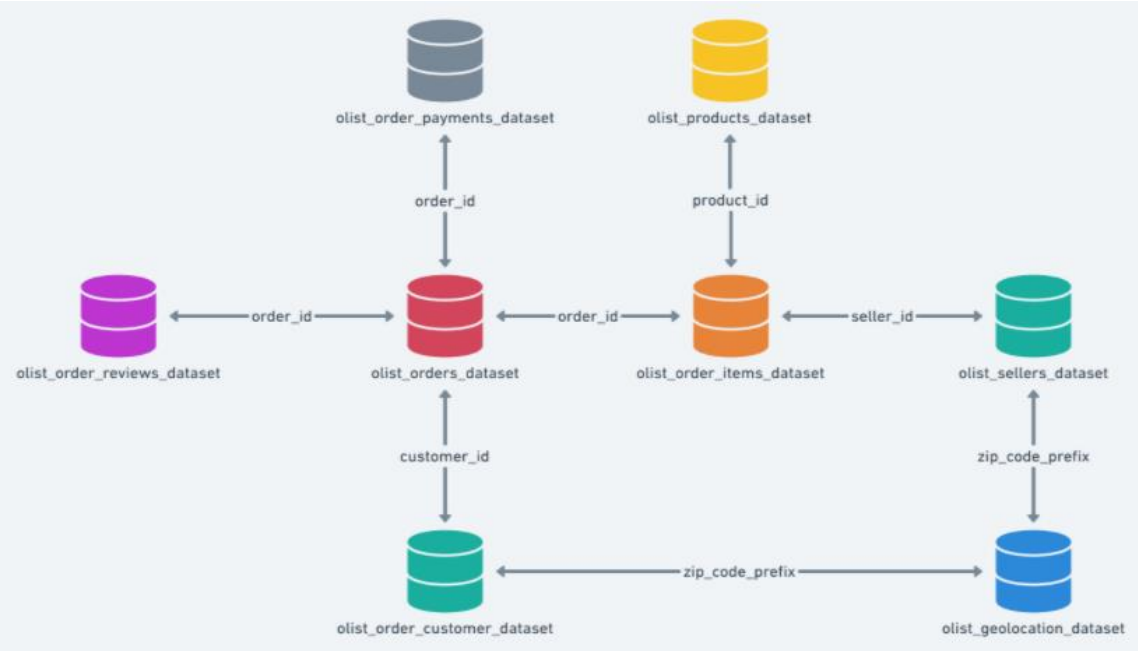


Ilustración 1: Estructura relacional de los subconjuntos de datos de Olist. Fuente: (Olist & Sionek, 2018)

3.2. Procesamiento de Datos

La etapa del procesamiento de datos es crucial para los trabajos de minería de datos. En esta sección se describirán las tareas desarrolladas y los programas utilizados para poner de manifiesto todos los criterios tomados para el análisis.

El primer paso de esta investigación consistió en cargar los distintos conjuntos de datos en un software que permitiera unirlos y trabajar más fácilmente los datos estructurados de las matrices. En el proceso, se descartó uno de los conjuntos de datos llamado *olist_geolocalization_dataset* ya que contenía datos irrelevantes para los análisis en cuestión quedando, por tanto, siete archivos en vez de los ocho iniciales para formar la base de datos. Con esto último como objetivo, se utilizó el programa libre y gratito de Oracle llamado MySQL Workbench en su versión 8.0.20 y se subieron los archivos que estaban en formato xls como bases de datos. (Oracle Corporation, 2020).

Luego, se procedió a unificar las bases de datos utilizando el lenguaje *Structured Query Language* (SQL) (en el Apéndice 1 se pueden encontrar las capturas de pantalla de los scripts que se corrieron y los códigos) y ello dio como resultado una tabla madre con 25 variables que pueden apreciarse en la tabla que se muestra a continuación (Ilustración 2).

order_id	costo_envio
fecha_orden	customer_city
fecha_orden	seller_city
fecha_despacho	misma_ciudad
fecha_entrega	product_category_name
fecha_estimada_envio	product_description_lenght
dias_envio_estimados	product_photos_qty
dias_entre_pedido_entrega	product_weight_g
dias_entrega_estimacion	product_volumen_cm3
Envio_tardio	Costo Relativo Delivery
payment_type	Cantidad_compras
payment_value	Experiencia
Price	

Tabla 1 Variables resultantes de unión del conjunto de datos.

Al comparar las variables del Anexo 1 con las de la Tabla 1 se aprecian diferencias en la cantidad y nombres de las variables, esto es producto de que hay variables extraídas de la unión de bases de datos y otras que fueron creadas a partir de la transformación de otras.

Para comenzar, se generaron variables relacionadas al tiempo de espera entre la compra y la entrega, en este grupo encontramos las columnas días_envio_estimado y días_entrega_estimacion, ellas muestran la cantidad de días transcurridos entre el día que los clientes cargaron la orden de compra, los que la plataforma dijo como día de llegada estimada (días_envio_estimado) y los que realmente fueron (días_entrega_estimacion). De la diferencia entre la fecha estimada y la efectiva de entrega final se crearon otras dos: una cuantitativa llamada "dias_entre_pedido_entrega", que indica la cantidad de días de diferencia en número, y una *dummy* (Envio_tardio) con valor igual 1 si la entrega se realizó más tarde de lo informado al cliente (como fecha estimada) y 0 si, por el contrario, fue de acuerdo al plazo estipulado.

Por otro lado, se incorporó "misma_ciudad" como variable binomial para evidenciar (con 1 y 0) si el cliente y el vendedor son de la misma ciudad o no. Esta busca identificar si la distancia entre la oferta y la demanda es un *driver* relevante para la experiencia de los usuarios. También, se investigó si el tamaño del producto era un factor relevante para la

calificar la experiencia por eso se incluyó la variable "product_volumen_cm3" que surge de la multiplicación de la altura, el ancho y la profundidad en centímetros del producto comprado. Posteriormente, se agregaron dos variables "Costo Relativo Delivery" para corroborar si el costo del delivery sobre el total del monto pagado era relevante y, "Cantidad_compras" para saber si los consumidores que compraron más veces tienen mejor experiencia que el resto. Y, por último se creó la variable a predecir: "Experiencia". Esta puede tomar dos valores: promotor o detractor, los promotores, son los clientes con notas 4 y 5 en las encuesta; los detractores, por consiguiente, son los usuarios que puntuaron entre 1 y 3.

Una vez unificadas las bases y seleccionadas las variables a considerar, comenzó la etapa de quita de los valores no válidos. La gran mayoría de ellos se extrajeron en el proceso de unificación de bases con el lenguaje SQL, no obstante, otra parte se trabajó en planilla de cálculos de Microsoft Excel. Allí se eliminaron más de 2000 órdenes que poseían valores nulos en el campo "orden_id". Además, se consideraron únicamente las órdenes de pagos con estado *delivered* (enviado) para tener la mayor cantidad de fechas con valores no vacíos y, en los casos en que la fecha de aprobación de la orden fuera nula, se completó con la misma fecha de orden.

Una de las consecuencias del filtro para incluir únicamente los envíos realizados y de la remoción de valores nulos es que se eliminaron los valores atípicos o *outliers* que venían dados por fechas mal imputadas.

Existieron, también, clientes que optaban por más de un medio de pago o utilizaban *vouchers* de descuentos. Esto provocaba que se duplicaran las órdenes y, con ellas, las notas de las encuestas de esos clientes. Este inconveniente se solucionó optando por el medio de pago con mayor peso en el monto total abonado. Realizado este cambio, se reajustaron los pagos totales para que figure que los clientes pagaron el total con ese único medio de pago y, de esta manera, la variable Costo Relativo Delivery no se viera afectada.

Los resultados de la limpieza de los datos arrojaron un saldo de 95.109 órdenes para clasificar de acuerdo a la experiencia obtenida y 21 variables que alimentarán los modelos seleccionados para alcanzar los mejores resultados.

El proceso para obtener resultados satisfactorios consta de muchas iteraciones en las que se procura realizar cambios pequeños y así tener mayor certeza acerca de si las implementaciones mejoraron o no la exactitud. En primer lugar, debe medirse el punto de partida o *baseline*. Este resultado muestra lo que obtendríamos en caso de no aplicar ningún modelo, por tanto, es nuestra referencia a partir de la cual medimos las mejoras. En este análisis, al ser una variable binomial la predicha, el *baseline* está dado por el valor más repetido que son los promotores cuya presencia relativa sobre el total de registros es de 78,59%.

Para probar las variables y modelos, primero se tomó una muestra de 5.400 órdenes que mantuviera la misma proporción de promotores y detractores para realizar las pruebas de optimización sobre los modelos y hacer una correcta selección de variables. El software utilizado para este fin se denomina RapidMiner y se utilizó la versión 9.7. (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020). Se incorporó la base de datos en archivo en formato xls en el software RapidMiner mediante un operador llamado Read Excel. Sobre él, se aplicó un filtro para seleccionar únicamente las variables de interés para el proceso iterativo de testeo.

Seguidamente, se normalizaron las variables y se separó la muestra tomada en una base de entrenamiento (80% del total de la muestra) y una base de testeo mediante un operador llamado *Split Data*. Finalizados estos primeros pasos, se aplicó un operador llamado *Optimaze Parameters*, el cual permite realizar pruebas a los modelos de forma tal de obtener los parámetros que permitan alcanzar los mejores resultados. Es así cómo comenzó la etapa de prueba y selección de algoritmos.

En cuanto a los atributos a considerar, el proceso de selección consistió en encontrar el modelo que derivara en el mayor *accuracy* para luego ir agregando o quitando variables y corriendo los algoritmos para observar las variaciones en la métrica. De esta forma es como se definieron las variables de entrada para los modelos y los parámetros que garantizaran mayores resultados.

En último lugar, una vez seleccionados loe modelos y variables con mejores resultados, se cargó la base de datos con todos los registros y se utilizó el operador de validación cruzada (o *cross validation*, por siglas en inglés). El operador en cuestión permite separar la base total

en conjuntos de datos de acuerdo a la cantidad. En este caso, se tomó como base el valor por default que es diez separaciones. El algoritmo detrás del operador separa la base en 10 partes iguales, toma una muestra aleatoria que represente 10% del total de los datos y prueba los modelos en los datos restantes. Luego, toma otra muestra de 10% de forma aleatoria con reposición, es decir, que podría seleccionar algunos de los seleccionados en la muestra anterior, y repite el procedimiento 10 veces.

El procedimiento iterativo comenzó utilizando el total de las variables de la muestra tomada y, con el operador *Optimize* se iba encontrando la mejor composición de los atributos. El primer modelo optimizado fue *K-NN*, sin embargo, sus resultados no fueron los esperados. Por tal motivo, se utilizó otro modelo llamado *Random Forest*. En este se fueron ajustando los parámetros más importantes (número de árboles generados, el criterio de medición y la máxima profundidad) en forma individual y luego agregada ya que se comprobó una mejor precisión.

A continuación, se probó el algoritmo llamado *Decision Trees* y se lo comparó con los resultados obtenidos por *Rule Induction*. El resultado del ajuste de los parámetros demostró que el segundo modelo mencionado poseía mayor poder predictivo y, por tanto, se avanzó con la optimización de los parámetros.

Finalmente, se ajustaron los parámetros del algoritmo *Logistic Regression*. Este modelo obtuvo buenos resultados y no hubo otro de los no probados que mejorara los resultados como para ingresar dentro de los mejores modelos. Así fue que se seleccionaron estos tres modelos que luego se ensamblaron con el *Stacking*. El ensamblado también requirió de optimización ya que había que hallar el modelo que mayor precisión y varianza generaran. Fue así que se optó por ensamblar esos modelos con otro *Random Forest*.

Recién explicamos el desarrollo de la selección los modelos. Por lo tanto, es momento de explicar el de las variables. A medida que un modelo lograba avances considerables en la precisión, se lo utilizaba para probar diversas combinaciones de variables con el fin de obtener las más preponderantes. Por ejemplo, una prueba consistía en tomar todas las variables relacionadas a fecha de la orden y el envío del producto, correr el modelo y observar los resultados. Si eran mejores se los separaba para incluirlos, sino se los descartaba y

utilizaba nuevos. El resultado del proceso fue hallar las variables con mejor resultado para cada uno de los modelos.

3.3. Análisis de datos

Los resultados del procesamiento de datos llevaron a contar con más 95.109 órdenes a predecir de acuerdo a la experiencia obtenida y 17 variables que alimentarán los modelos seleccionados. Sin embargo, la base de datos cuenta con 25 variables de origen mixto. En las próximas líneas, se presentará un análisis descriptivo de las variables cuantitativas y cualitativas consideradas para el análisis.

La primera variable a definir es order_id. Esta funciona como clave primaria en la base de datos ya que identifica con un nombre único a cada orden de compra enviada a los clientes. Esta variable se utilizó como *ID* en RapidMiner para hacer las predicciones y es de tipo texto.

Por otra parte, presentamos el atributo a predecir: Experiencia. Como mencionamos en la sección anterior, esta nace como resultado de la transformación de la variable numérica *score* (nota) y es de tipo binomial ya que puede tomar dos valores, promotor y detractor. En RapidMiner, esta variable fue definida como *label* para indicar que era sobre la cual debía calcularse la performance de los modelos. La Ilustración 3 muestra la frecuencia y los porcentajes que fueron la base de la cual se partieron los algoritmos para realizar las predicciones.

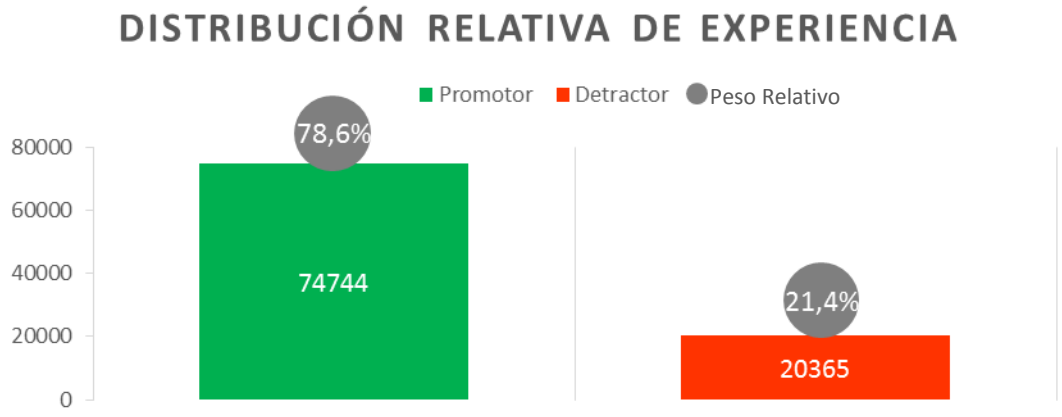


Ilustración 2: Distribución relativa de Experiencia. Fuente: Elaboración propia.

Continuando con el detallado de los atributos, nos encontramos con cuatro variables que son del tipo fecha: fecha_orden, fecha_aprobacion_orden, fecha_despacho y fecha_entrega. Ninguna de estas variables fue tenida en cuenta como variables de entrada de los modelos ya que no mejoraban los resultados. En cuanto a su descripción en la Tabla 2 se detalla el valor mínimo, el máximo, el rango, la moda y la mediana de estas variables.

Variable	Tipo de variable	Min	Max	Rango (en días)	Moda	Mediana
fecha_orden	Date	03/10/2016	29/08/2018	695	22/11/2017	21/01/2018
fecha_aprobacion_orden	Date	04/10/2016	29/08/2018	694	24/04/2018	22/01/2018
fecha_despacho	Date	08/10/2016	11/09/2018	703	28/11/2017	24/01/2018
fecha_entrega	Date	11/10/2016	17/10/2018	736	27/08/2018	03/02/2018
fecha_estimada_envio	Date	27/10/2016	25/10/2018	728	20/12/2017	16/02/2018

Tabla 2: Caracterización de variables de tipo fecha. Fuente: Elaboración propia.

Los clientes de este dataset realizaron sus compras con cuatro distintas formas de pago y esto quedó registrado en la variable payment_type (forma de pago). Las formas de pago utilizadas fueron: tarjeta de crédito, boleto bancario, tarjeta de crédito y cupones de descuento. La participación de cada medio de pago sobre el total de las compras puede apreciarse a continuación en la Ilustración 3.

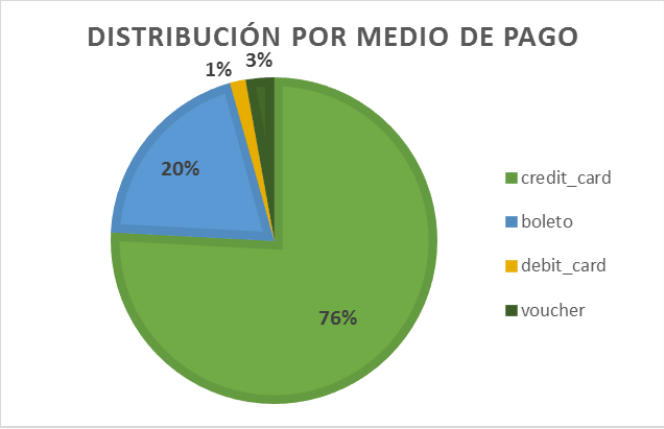


Ilustración 3: Distribución por medio de pago. Fuente: Elaboración propia

Además de la variable a predecir, existen otras dos variables binomiales que toman un valor cuando una condición se cumple y otro cuando no. En este caso las variables tomaron valores numéricos, por lo tanto, el valor es 1 si la condición se cumple y si no, 0. Estas variables son: envio_tardio y misma_ciudad. En la primera de ellas, se verifica que la cantidad de envíos tardíos, envíos entregados luego de la fecha estimada al cliente, totalizaron

6.436 casos (6,8% sobre el total de compras). Respecto a la cantidad de clientes que compraron a vendedores de la misma ciudad, los registros suman 4.878 ventas que en términos relativos es un 5,1% sobre el total.

En último lugar, se encuentras las variables cuantitativas. Para realizar el análisis de las mismas se calculó el valor mínimo, el valor máximo, el rango (diferencia entre valor máximo y mínimo), el promedio simple, la mediana, la moda, la asimetría y la curtosis.

Variable	Tipo de variable	Min	Max	Rango	Promedio	Desvío Estandar	Moda	Mediana	Asimetría	Curtosis
dias_envio_estimados	Integer	3	156	153	24,38	8,77	22	24	0,91	4,33
dias_entre_pedido_entrega	Integer	0	210	210	12,49	9,55	7	10	3,82	39,25
dias_entrega_estimacion	Integer	-188	147	335	11,88	10,18	14	12	-2,01	28,08
payment_value	Real	6	6929,31	6923,23	145,48	196,63	77,57	95,85	7,61	113,62
price	Real	1	6735,00	6734,15	125,30	189,42	59,90	79,00	7,83	118,98
costo_envio	Real	0	409,68	409,68	20,18	15,83	15,10	16,39	5,59	58,77
product_description_lenght	Integer	4	3992	3988	793,39	653,37	1893	608	1,99	4,84
product_photos_qty	Integer	1	20	19	2,25	1,75	1	2	1,85	4,49
product_weight_g	Integer	2	40425	40423	2104,68	3757,39	200	700	3,60	16,37
product_volumen_cm3	Integer	168	296208	296040	15205,06	23307,96	8000	6426	4,07	25,61
Costo Relativo Delivery	Real	0	0,96	0,96	0,21	0,13	0,00	0,18	1,06	1,19
Cantidad_compras	Integer	1	15	14	1,10	0,37	1	1	13,29	365,78

Tabla 3: Estadística descriptiva variables cuantitativas. Fuente: Elaboración propia.

En la tabla 3 se detallan las variables cuantitativas seleccionadas como *inputs* para los algoritmos. El cálculo y definición de los primeros dos registros fue explicado en la sección anterior.

Payment_value, por su parte, es el pago total que realizó el cliente, explicado por la suma del precio del producto (price) y el costo de delivery (costo_envío). Esta desagregación del precio total era importante hacerla ya que los costos de envío varían y, por lo tanto, también el peso de ese costo sobre el total pagado (Costo Relativo Delivery) y este era una de las variables que se deseaba estudiar. El objetivo de detallar estos atributos relacionados al costo era comprobar si son un factor determinante para predecir la experiencia y, de serlo, entender si era el precio del producto, el de delivery o el costo total lo que influía más.

La relación entre las características del producto y la experiencia se realizó a través del agregado de las siguientes variables: product_descripcion_lenght, que indica cuántos caracteres se utilizaron para describir al producto en la publicación, product_photos_qty, muestra la cantidad de fotos que contenía la publicación sobre el producto,

product_weight_g, que expresa el peso en gramos del ítem publicado, y product_volumen_cm3 que expresa el tamaño total del producto basándose en las

La última variable numérica incorporada fue cantidad de compras. Esta muestra la cantidad de ventas que realizó el vendedor de ese producto. Sin embargo, como veremos más adelante esta variable luego terminó no contemplándose como *input* de los modelos.

3.4. Modelos aplicados

En la presente sección se definirán, en primer lugar, los modelos utilizados para realizar las predicciones sobre la experiencia de los clientes. Luego, se presentarán las variables de entrada seleccionadas para esos modelos.

Los modelos aplicados para predecir la experiencia de los clientes fueron los siguientes: *Rule Induction, Random Forest, Logistic Regression y Stacking*.

Rule Induction es un proceso de minería de datos en el que sea crean condiciones denominadas *IF-THEN* en inglés sobre el conjunto de datos. Este modelo va generando reglas de decisión para explicar la relación existente entre los atributos y las etiquetas de los datos (Kotu & Deshpande, 2015). Esto lo hace incorporando condicionales (si se cumple esta condición, el resultado es el siguiente) que son simples de entender y por esta razón fue uno de los primeros en ser elegidos para esta investigación.

El segundo método utilizado fue regresión logística. Como su nombre lo indica, se encuentra dentro de los modelos de regresión y el objetivo es representar la variable a predecir en términos de otras variables (Kotu & Deshpande, 2015, pág. 23). Esta clase de modelos realizan una regresión para cada clase, estableciendo la salida igual a uno para las instancias de entrenamiento que pertenecen a la clase y cero para las que no. Sin embargo, los modelos de regresión lineal o multivariado no sirven para predecir variables no numéricas y por eso se utilizó regresión logística. Este método aplica una función de transformación denominada logit a la variable objetivo que no puede aproximarse utilizando una función lineal. Los valores resultantes de este método ya no están restringidos al intervalo de 0 a 1, sino que pueden encontrarse entre menos y más infinito permitiendo obtener resultados ajustados al propósito de este análisis. (Witten, Frank, Hall, & Pal, 2017, págs. 129-130).

Por último, se empleó *Random Forest* ya que este algoritmo, para disminuir el error, toma aleatoriamente un subconjunto de atributos y observaciones de la base de entrenamiento, crea los árboles de decisión y utiliza esos árboles para predecir la clase a cada registro. (Kotu & Deshpande, 2015, pág. 162).

Con el propósito de alcanzar los mejores resultados se recurrió a dos técnicas adicionales a la hora de aplicar los modelos. En primer lugar, se utilizaron distintos atributos para los modelos, para *Random Forest* y *Rule Induction* se utilizaron las siguientes variables extraídas de la tabla 1 y definidas en el anexo 1. En cambio, para el algoritmo de regresión logística se quitaron las variables payment_type y customer_city ya que disminuía la precisión y aumentaba la varianza en los resultados.

La otra técnica empleada fue recurrir a un cuarto algoritmo que permitiera combinar los modelos anteriormente descriptos, *stacking*. *Stacking* intenta aprender cuales modelos son los confiables utilizando un algoritmo de aprendizaje llamado *metalerner*. Con este, descubre la mejor manera de combinar los resultados (Witten, Frank, Hall, & Pal, 2017, pág. 497). El hecho de que *stacking* permitiera combinar modelos diferentes entre sí y para llegar a las conclusiones llevó a considerarlo dentro de los algoritmos plausibles de aplicación.

3.5. Métricas

La forma en la que el analista puede contrastar los resultados de las diferentes pruebas es a través de una métrica. Esta debe permitir comparar los resultados obtenidos no sólo con otros modelos realizados por otras personas, sino también para tener una historia de los resultados alcanzados y estar seguros de que sean mejores que los anteriores.

En el presente trabajo se seleccionó la exactitud (*accuracy* en inglés) como medida que permitiera comparar resultados. El *accuracy* mide la proporción de casos realmente bien predichos por el algoritmo sobre el total de los casos que fueron outputs del modelo. El cálculo surge de la matriz de confusión (Ilustración 2) y es el siguiente:

$$Accuracy = \frac{Valores\ Positivos + Valores\ Negativos}{Valores\ Positivos + Falsos\ Positivos + Valores\ Negativos + \ Falsos\ Negativos}$$

Ecuación 1: Fórmula de cálculo de Accuracy. Fuente: Elaboración Propia

Ilustración 4: Matriz de confusión (Barrios Arce, 2019)

Esta métrica se usó para comparar los modelos y nos permitió conocer cuál era el valor inicial que debíamos mejorar a partir de la minería de datos. También se utilizó esa medida para comparar los resultados de las distintas iteraciones permitiendo así realizar la selección de los atributos y parámetros que mejoraron la performance de los modelos. Por consiguiente, seleccionada la mejor combinación de atributos, modelos y parámetros, también se utilizó esta medida para compartir el resultado final.

3.6. Resultados

En las secciones del anterior capítulo se precedió a describir la base de datos, cómo se la unificó, el tratamiento que se le dio a los *outliers* y a los valores faltantes. Luego, se explicaron los modelos y variables utilizadas, cuales fueron descartándose y con cuales se decidió avanzar mediante la iteración y optimización. Finalmente se definió la métrica con la cual se iban a mostrar los resultados y con la que se compararon las distintas alternativas. En este capítulo se presentaran los resultados de los distintos algoritmos probados, se expondrán los motivos por los cuales se optó por el modelo final y, por último, se interpretarán los resultados obtenidos.

Los modelos escogidos para realizar este análisis sobre los compradores de Olist fueron *random forest*, *rule induction*, regresión logística y *stacking*. Al ser algoritmos distintos, los resultados son también lo son y por lo tanto veremos los resultados por separado.

El modelo de regresión logística arrojó un 81,85% de *accuracy* con una varianza de más menos 0,25%. Profundizando en el principal problema de este modelo, lo que podemos encontrar es que predice bien la cantidad de promotores, mejorando el punto de partida, a

saber 78,59%, en 5,8 puntos porcentuales como se aprecia en la columna de *class precision* en la Ilustración 5.

accuracy: 81.85% +/- 0.25% (micro average: 81.85%)

	true Promotor	true Detractor	class precision
pred. Promotor	73107	15623	82.39%
pred. Detractor	1637	4742	74.34%
class recall	97.81%	23.29%	

Ilustración 5: Resultados en RapidMiner del modelo de Regresión Logística. Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

Por su parte, el algoritmo que crea reglas de decisión, *rule induction*, fue superior al modelo de regresión logística ya que, si bien logró una menor precisión de promotores, registró una precisión final superior (81,81%) y una varianza inferior (0,22%). En la salida de RapidMiner (ver Ilustración 6) se aprecian las diferencias mencionadas.

accuracy: 81.81% +/- 0.22% (micro average: 81.81%)

	true Promotor	true Detractor	class precision
pred. Promotor	73370	15922	82.17%
pred. Detractor	1374	4443	76.38%
class recall	98.16%	21.82%	

Ilustración 6 Resultados en RapidMiner del modelo de Rule Induction. Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

El tercer modelo, *random forest*, superó a los anteriormente mencionados en *accuracy*. Este algoritmo que selecciona aleatoriamente muestras de observaciones y atributos para armar árboles de decisión y los utiliza para predecir, mejoró la precisión de los detractores y disminuyó muy poco la de los promotores logrando 79,70% y 82,16% respectivamente (ver Ilustración 7). Con los resultados anteriormente mencionados logró alcanzar 82,02% de precisión con una varianza de 0,25%.

accuracy: 82.02% +/- 0.25% (micro average: 82.02%)

	true Promotor	true Detractor	class precision
pred. Promotor	73630	15991	82.16%
pred. Detractor	1114	4374	79.70%
class recall	98.51%	21.48%	

Ilustración 7 Resultados en RapidMiner del modelo de Random Forest. Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

Por último, se realizó el ensamble de los modelos previamente detallados mediante el algoritmo *stacking*. Si bien la precisión general del modelo no mejoró, la precisión de los detractores aumentó comparando con *Random Forest* y también mejoró un poco la varianza, bajando un 0,01%. De esta forma se llegaron a los resultados que se aprecian en la Ilustración 8 y que colocaron a este modelo por encima de los anteriormente mencionados.

accuracy: 82.02% +/- 0.24% (micro average: 82.02%)

	true Promotor	true Detractor	class precision
pred. Promotor	73644	15996	82.16%
pred. Detractor	1100	4369	79.89%
class recall	98.53%	21.45%	

Ilustración 8 Resultados en RapidMiner del modelo de Stacking. Fuente: (Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7, 2020)

Luego de haber hallado el modelo de mayor precisión se realizaron pruebas con grupos de variables para entender mejor qué tipo de variables incidían más. Los grupos que se tuvieron en cuenta fueron los siguientes: Costos, Delivery, Características Producto y Fechas. Esta clasificación buscó poner de manifiesto los principales *drivers* de la experiencia de los clientes. Los resultados se pueden apreciar en la tabla que se muestra a continuación:

Grupo	Costos	Delivery	Características Producto	Fechas
Variables por cada grupo	payment_type	dias_envio_estimados	product_category_name	fecha_orden
	payment_value	dias_entre_pedido_entrega	product_description_lenght	fecha_aprobacion_orden
	price	dias_entrega_estimacion	product_photos_qty	fecha_despacho
	costo_envio	Envio_tardio	product_weight_g	fecha_entrega
	Costo Relativo Delivery	misma_ciudad	product_volumen_cm3	fecha_estimada_envio
		Cantidad_compras		
Accuracy	78,30%	81,98%	78,56%	81,27%

Tabla 4: Grupos de variables y predicción con los modelos seleccionados. Fuente: Elaboración propia en base.

La Tabla 5 (ver abajo) muestra un resumen de los resultados obtenidos en los modelos habiéndolos corrido individualmente en comparación con el modelo ensamblado que finalmente se utilizó. En el Apéndice 2 se encuentran a disposición las capturas del flujo de cada uno de los modelos implementados.

Modelos Individuales			Ensamble con Stacking			
Nombre del Modelo	Accuracy	Varianza	Nombre del Modelos ensamblados	Modelo de Ensamble	Accuracy	Varianza
Random Forest	82,02%	+- 0,25%	Random Forest	Random Forest	82,02%	+- 0,24%
Regresión Logistica	81,85%	+- 0,25%	Regresión Logistica			
Rule Induction	81,81%	+- 0,22%	Rule Induction			

Tabla 5: Tabla Comparativa Modelos Aplicados. Fuente: Elaboración propia.

De lo expuesto previamente se desprende que el modelo con mejores resultados en términos predictivos en *stacking*. Este modelo supera el *baseline* en 5,43 pp (puntos porcentuales) demostrando que es de cierta utilidad para predecir resultados, especialmente si se trata de clientes promotores. Por otra parte, quedó en evidencia que las variables con mayor peso predictivo con las que componen los grupos “Fechas” y “Delivery”, es decir, todas aquellas relacionadas con el día en que se realizó la orden y la fecha de entrega efectiva y estimada, y las relativas a los datos y condiciones del envío (demora, misma ubicación del vendedor y comprador y cantidad de compras por vendedor, entre otras).

4. Implementación de Modelos

En los tiempos actuales la tecnología, la cultura, las personas y las preferencias cambian constantemente, los beneficios de las empresas se están reduciendo por la alta competencia y los clientes quieren resultados más rápidos y efectivos. En este contexto, se creó la necesidad de contar con un método de gestión de proyectos que se adapte a estos nuevos

requisitos del mercado con rapidez y flexibilidad, y donde los métodos tradicionales de gestión se han demostrado ineficientes a la hora de gestionarlos (TiThink, 2018).

Este capítulo del informe tiene por objeto mostrar cómo se podría implementar el análisis predictivo en organizaciones como Olist. Para ello se definirá el concepto de metodologías ágiles. Luego se mostrará cuál metodología se puede adaptar mejor a la organización y las ventajas que trae aparejada su implementación. Finalmente, se hará un reconto de los recursos y tiempos necesarios para llevar adelante un proyecto de esta magnitud en Olist.

4.1. Metodologías Ágiles

Adaptarse a los nuevos tiempos ya no es opcional para las empresas en una sociedad cada vez más cambiante. Hoy todas las empresas que tienen intenciones de sobrevivir y tener éxito en el mercado actual deben adoptar y probar nuevos rumbos que les permitan mantenerse en la corriente. Por ello, cada vez más empresarios apuestan por una transformación digital completa para su empresa.

La transformación es tal que exige cambiar desde la manera en la que las compañías se relacionan con sus clientes y se comunican con ellos, hasta la forma en la que se organizaba la producción. La necesidad de un cambio se debe a que los procedimientos actuales están haciendo lentos los procesos, son estructurados, burocráticos, no se adaptan a las necesidades de los clientes ni a las circunstancias cambiantes de la sociedad y los clientes. Las metodologías ágiles son las que nacieron para resolver estos problemas (LN Creatividad y Tecnología, 2019).

Michael Cusmano, profesor del MIT define a la agilidad como “la habilidad de adaptarse rápidamente, o incluso anticiparse, al contexto y liderar un cambio. En el sentido más amplio, afecta al diseño estratégico, las operaciones, la tecnología y la innovación” (ContentLab, 2019). En línea con la definición del académico, el blog Kezmo define las metodologías ágiles así:

> Por metodologías ágiles entendemos a aquellas metodologías de gestión que permiten adaptar la forma de trabajo al contexto y naturaleza de un proyecto, basándose en la flexibilidad y la inmediatez, y teniendo en cuenta las exigencias del

> mercado y los clientes. Los pilares fundamentales de las metodologías ágiles son el trabajo colaborativo y en equipo. (Kezmo, 2017)

Por lo tanto, las metodologías ágiles permiten adaptar la forma de trabajar a las condiciones del proyecto, aportando flexibilidad, eficiencia y, por lo tanto, logrando un mejor producto a menor costo. Esto en los tiempos modernos es fundamental para las empresas ya que representa una mejora en todos los sentidos posibles. Además, cuentan con la ventaja de que existe más de una metodología ágil e incluso se pueden entrelazar para adaptarse a las necesidades de cada firma. Entre las metodologías más utilizadas encontramos: Lean, Kanban y Scrum.

Lean surgió en el ámbito de la manufactura impulsada por Taiichi Ono, director y consultor de la empresa Toyota. Se basa en lograr una mejora continua para poder generar un flujo de producción que asegure entregar el máximo valor para los clientes con los mínimos recursos necesarios. Aplicar esta metodología de trabajo implica que todos los miembros involucrados en el proceso productivo tengan una responsabilidad asignada y sean capaces de tomar decisiones. De esta manera, el equipo de mantiene motivado y genera un sentido de pertenencia con la organización (Kezmo, 2017).

La metodología Kanban es un sistema que funciona con posts it o tarjetas visuales. Es un marco muy usado que consiste en la elaboración de diagramas en los que se anotan las tareas pendientes, en proceso o terminadas de un equipo (en inglés, To Do, Doing, Done) y de esta forma apela a la comunicación simple entre los miembros del equipo para trabajar en forma coordinada y transparente entre todos los miembros. Su principal ventaja es que mejora el trabajo en equipo manteniendo a todos los miembros en el proyecto que se está llevando adelante y que, a diferencia de Scrum, no requiere cambios organizacionales. (ContentLab, 2019)

En relación a Scrum, es un marco de trabajo diseñado para facilitar el desarrollo ágil de un proyecto. Se basa en una filosofía colaborativa y en dividir el trabajo en ciclos temporales. Ponerlo en marcha es un proceso complejo ya que requiere una transformación organizacional. Sin embargo, Henrik Kinberg y Mattias Skaring (Kinberg & Skaring, 2010, págs. 19-21) lo explican de forma muy simple: en primer lugar, hay que dividir la organización en equipos pequeños, multifunción e independientes entre sí. Después, hay que

separar el trabajo en pequeñas entregas incrementales y ordenarlas en base al valor que agregan contemplando los costos económicos y de tiempo. Luego hay que establecer períodos de tiempo iguales donde se hagan entregas de valor, suelen elegirse lapsos con una duración entre una a cuatro semanas. En cuarto lugar se va a ir optimizando las entregas de valor en base a las prioridades que va definiendo el cliente. Por último, se va optimizando el proceso teniendo una retrospectiva después del fin de cada período de trabajo De esta manera queda un equipo chico que dedica poco tiempo a construir algo pequeño (teniendo que regularmente integrar todo) en vez de un grupo grande que pasa mucho tiempo construyendo algo enorme y que no posee flexibilidad alguna.

4.2. Adaptación de Metodologías Ágiles a la Organización y Ventajas

Para llevar adelante el proyecto en Olist se va a recurrir a una mezcla de las metodologías ágiles Scrum y Kanban anteriormente descriptas ya que estas permiten organizar el trabajo en el tiempo priorizando por el valor que agrega y el esfuerzo que conlleva y, a la vez, dejarlo plasmado en la diaria de forma tal de transparentar los procesos y mejorar el trabajo en equipo.

Las ventajas de utilizar estas metodologías respecto a la tradicional que trabaja por silos son muchas. En primer lugar, en el trabajo por silos los equipos están conformados por personas que comparten conocimiento específico, es decir, los equipos de negocios conocen los productos, el pricing, tienen contacto con los clientes finales y proponen ideas que tienen que desarrollar e implementar los equipos de sistemas y de seguridad. Estos últimos, por otro lado, conocen los sistemas, requisitos tecnológicos necesarios y herramientas disponibles para poder hacer desarrollos. En este simple intercambio de información ya se producen las primeras ineficiencias en la comunicación ya que existe una brecha entre las expectativas de las áreas de negocio y sus tiempos, y la realidad y tiempos del sector de IT. En las organizaciones, este cortocircuito se presenta de la siguiente manera: los de áreas de negocio llevan la idea a implementar a IT, hacen la solicitud y esperan que vuelva resuelta. Por su parte, los equipos de sistema deben interpretar esa idea y desarrollarla sin tener toda la tecnología necesaria y con el surgimiento de retrasos en el medio. El resultado de esta metodología de trabajo es un loop (ciclo repetitivo) de intercambios ya que sistemas debe

entender qué quiere el negocio y el negocio las limitaciones. Mientras tanto, el tiempo no deja de transcurrir y se producen tiempos muertos por las dependencias que van apareciendo. En las metodologías ágiles estas pérdidas de tiempo no se dan o se resuelven más rápido ya que los equipos son multidisciplinarios. En ellos participan, perfiles profesionales variados que permiten acercar y compartir los conocimientos, de modo tal que si alguna de las partes tiene una duda o retraso lo puede resolver e identificar sin perder tiempo. Este pequeño cambio en la forma de organizar los equipos tiene un impacto grande. Ya no hay tiempos muertos y por tanto se mantiene la intensidad de trabajo en el tiempo

Otra ventaja que poseen las metodologías ágiles es que permiten visualizar en todo momento los cuellos de botella, los procesos que traban avances quedan expuestos y deben resolverse para cumplir con los objetivos propuestos. Ya sea sumando personas de equipo o involucrando gente externa para destrabar los problemas, el hecho de identificar rápidamente la necesidad desde la mirada de sistemas y de negocio hace que la resolución sea no sólo más rápida sino también más precisa de cara a las necesidades del cliente. Otras externalidades que se obtienen, además de mejores resultados, es que fomenta la colaboración entre los empleados y genera transparencia en los procesos.

En tercer lugar, en las metodologías ágiles, además de establecerse los objetivos anuales llamados KPIs (por sus siglas en inglés, Key Performance Indicators), se trabaja con OKRs (Objective Key Results) que son objetivos trimestrales medibles que van a velar por el cumplimiento de los objetivos anuales pero que son de naturaleza operacional, es decir, que se espera se cumplan al 70%. Para cumplir los Objective Key Results, se piensan y establecen durante el trimestre distintas iniciativas que, de cumplirlas, aumentan el porcentaje de cumplimiento de los objetivos. Este cambio en la estrategia teniendo distintos horizontes, objetivos cortos y más largos, permite mantener la motivación del equipo ya que los objetivos cambian más rápidamente y las iniciativas también permitiendo ajustar si no dieron el resultado esperado. Esta flexibilidad otorga mayores instancias para evaluar las acciones por lo que también ahorra costos a las empresas de inmiscuirse en proyectos eternos sin conocer su resultado hasta el final.

La cuarta ventaja de las metodologías ágiles viene dada por la manera en la que se fijan objetivos. A diferencia de las metodologías tradicionales en la que los gerentes decidían los

objetivos de forma *top down*, en las metodologías ágiles los equipos proponen las iniciativas y objetivos trimestrales que van a buscar cumplir con el objetivo anual, es decir, son *buttom up* (de abajo hacia arriba). Esto representa una ventaja ya que los dueños de los productos, los que lidian con los problemas diarios, saben mejor qué cosas son más prioritarias para los usuarios finales y, sobretodo, conocen los plazos que llevan las implementaciones. Por esto, una vez propuestos las iniciativas y objetivos trimestrales, ordenan en una línea de tiempo para darle todo el tiempo visibilidad a los avances logrados en pos del cumplimiento de los objetivos propuestos.

Por último, a diferencia de la metodología por silos donde el desarrollo se veía al final, la Scrum conlleva entrega de valor constante. Esto se asegura mediante la organización del trabajo en *sprints*. Los *sprints* son períodos de tiempo (una semana, quince días, un mes son los más usuales) en los que se cumplen una serie de ceremonias tendientes a no perder de vista los objetivos, identificar a tiempo los problemas, resolverlos, entregar MVPs (Producto Mínimo Viable en inglés) y hacer modificaciones en base al *feedback* obtenido. Las ceremonias que componen los *sprints* son: el refinamiento, que es una instancia en la que se detallan todas las tareas a llevar acabo para cumplir una tarjeta que es el objetivo semanal y para que todo el equipo aprenda sobre el trabajo que conlleva cada iniciativa, la *daily* sincro, que es una reunión diaria en la que se muestra lo que se realizó el día anterior, se cuentan impedimentos y logros de la tarjeta y se cuenta en qué se va a avanzar en el día, la *planning*, que es la ceremonia en la que cada colaborador toma las tarjetas (que fueron previamente refinadas) de las cuales se va a ser responsable durante ese sprint para cumplirlas. Por último, se encuentra la *review* que es la ceremonia en la que se le lleva al usuario final el MVP para recibir *feedback*. Este orden en sprints y ceremonias permite tener mayores instancias con el usuario final de los productos y estar más cerca de las necesidades que tiene, agregando mayor valor con cada iteración y construyendo valor con reseñas previas de forma tal que la empresa, en este caso Olist, tiene la seguridad de seguir avanzando sobre lo ya validado.

4.3. Aplicación para Olist

Para cumplir con la implementación de este modelo predictivo en Olist se precisará de un ingeniero de datos, un científico de datos, un scrum master, un *product owner* o *product manager* y un analista de negocio. El ingeniero de datos va a ser el encargado de integrar

todos los datos transaccionales que se relevan de distintos canales y estructurarlos en tablas para ser explotadas en el *data warehouse*[4]. El científico de datos va a ser el encargado de trabajar con los datos integrados por el ingeniero de datos para desarrollar el modelo en algún programa o nube que permite implementar modelos predictivos o de aprendizaje automáticos. El científico de datos debe tener conocimientos estadísticos que le permitan encontrar la mejor métrica para el objetivo propuesto y el modelo que mejor performance alcance con los mejores parámetros. El analista será el encargado de aportar la visión de negocio al modelo desarrollado por el científico de datos para que los atributos incluidos, los resultados obtenidos y los gráficos sean fáciles de comprender para áreas sin grandes conocimientos de datos. El scrum master, por su parte, será el encargado de presidir las ceremonias para garantizar que los objetivos de cada una se cumplan y que los impedimentos salgan a la luz, fomentando la comunicación y la transparencia entre los colaboradores. Por último, el *product owner* o *product manager* será el encargado de planificar y priorizar las tareas ordenándolas en *sprints* para garantizar el cumplimiento de los objetivos propuestos teniendo en consideración el valor que aporta cada objetivo y el esfuerzo que requiere.

Este proyecto en particular se podría desarrollar en seis *sprints* con duración quincenal. El primer sprint va a tener como objetivo estructurar la información necesaria para poder iniciar con la preparación de los datos del modelo. El segundo sprint y el tercero van a estar destinados a la preparación de los datos. En el segundo, el objetivo será realizar un *brain storming* (lluvia de ideas, en castellano) con gente del equipo de experiencia del cliente en conjunto con las áreas de delivery, *pricing*, marketing, canales y facturación para incorporar la mayor cantidad de variables al modelo. Con esto, se procederá al mapeo de las variables. En la tercera quincena, se llevará a cabo la limpieza de las variables: se eliminarán los valores anómalos, se estudiaran las distribuciones, se hará la estadística descriptiva, se trabajará sobre los valores nulos y se estandarizarán las variables. En la cuarta etapa, se implementarán los distintos modelos, se ajustarán los parámetros para obtener las mejores resultados, se priorizarán las variables y se definirá la métrica que sea mejor para comparar los resultados obtenidos. Más adelante, en el quinto sprint, se analizarán los resultados obtenidos con el mejor modelo, se expondrán las variables de mayor relevancia para los usuarios finales. En

[4] Un Data Warehouse es un almacén electrónico donde generalmente una empresa u organización mantiene una gran cantidad de información

el sprint final se documentarán los procesos realizados y se crearán las visualizaciones en una presentación con formato PPT, PDF o tablero de Power BI para compartir los hallazgos con las distintas áreas de la organización.

Para conseguir mayor feedback de los usuarios finales, cada sprint contará con 2 reviews, de forma tal de validar con el cliente final si los resultados obtenidos a la fecha están en línea con lo diseñado por nosotros. De esta manera quedaría implementado el modelo en la organización seleccionada, listo para su uso por parte del resto de la organización para priorizar las listas de pendientes en otras áreas.

5. Conclusiones

Al comienzo de este trabajo se explicitó el objetivo de este informe: predecir la experiencia de los consumidores finales de Olist utilizando las metodologías ágiles para implementarlo. El haber podido utilizar modelos predictivos en Olist abre el juego a poder aplicarlos a otras organizaciones e incluso sectores. Esto es gracias a que la empresa posee una arquitectura de datos compleja, que integra distintos tipos de datos (estructurados, semi estructurados y no estructurados), y a que pertenece a uno de los rubros con mayor crecimiento económico del año. Implementar esta predicción en una empresa que se encuentra a la vanguardia de tecnología y datos y en un mercado muy dinámico, permite demostrar la importante de incluir análisis cuantitativos basados en la minería de datos en las áreas de experiencia del cliente de todas las organizaciones.

Para alcanzar la meta propuesta, se trabajó con información transaccional de los usuarios ya que esto permitía identificar los principales *drivers* de experiencia generando focos de trabajo concretos para los decisores de las empresas. Cabe señalar que para llegar a estos resultados se tomó la decisión de analizar la experiencia de aquellos usuarios a los que se les había enviado su compra. Esta decisión se fundó en la necesidad de simplificar el trabajo de limpieza de datos y el de formar una única base de datos.

El procesamiento de los datos arrojó, por un lado, una base de datos con variables plausibles de agrupar en cuatro grandes grupos: variables relacionadas con el costo, las relacionadas con el delivery, las referidas a las características del producto comprado y las relativas a la fecha en la cual se realizó la compra. Por otro, los modelos, los parámetros y los atributos a ser incluidas para los modelos seleccionados. En cuanto a los algoritmos utilizados, estos fueron previamente ajustados con un operador llamado *optimize parameters* que, como su nombre en inglés lo indica, permite realizar iteraciones en los parámetros para alcanzar el mejor ajuste para la métrica elegida. De esta forma se seleccionaron los modelos y parámetros con mayor precisión predictiva: regresión logística, *random forest*, *rule induction* y *stacking*.

Los resultados del análisis cuantitativo se obtuvieron utilizando el operador llamado *stacking* que combina modelos para mejorar la precisión final. Esto lo hace recurriendo otro

algoritmo, *random forest,* y que permitió alcanzar 82,02% de precisión cuando el valor de partida era 78,59%.

Las conclusiones extraídas de este análisis, a pesar de los modestos resultados del modelo predictivo, son relevantes. El primer punto a destacar es que se obtuvieron los factores más relevantes para explicar la experiencia de los clientes. Las variables relacionadas al delivery como la demora en el envío y la experiencia del vendedor en el comercio electrónico son las que mayor valor aportaron para la predicción. Luego se ubicaron las variables relacionadas a la fecha de compra del producto. En otro orden de importancia se encuentran los atributos relacionados a los costos (precio del producto, costo del delivery, medio de pago, entre otros) y los relativos a las características del producto. Esta información es muy importante para la gestión de la experiencia ya que se pueden priorizar acciones para mejorar los tiempos de entrega de los productos y la previsibilidad de los mismos y, con ello, más rápidamente la experiencia. Además, permite abordar análisis posteriores con ventanas temporales para determinar por qué en determinados períodos la experiencia de los clientes fue menor o mayor que en otros. Por último, permite cambiar las prioridades en las tareas a realizar como puede ser, revisar la estrategia de *pricing* de los envíos o realizar una campaña para dejar de vender productos con determinadas característica de tamaño o peso.

En segundo lugar, se mostró cuál era la mejor manera de llevar adelante la implementación de este modelo predictivo en una organización a la que le espera mucho crecimiento. En este sentido se definió el concepto de metodologías ágiles y se explicitaron las bondades que la convierten en la ideal para desarrollar este tipo de proyectos. Entre ellas encontramos que: permiten acelerar los tiempos de trabajo, garantizan una mayor calidad en el resultado final, dan mayor transparencia a los procesos, fomentan la unión del equipo porque generan sentido de pertenencia y, por último, aminoran los costos y los riesgos de los proyectos.

De lo expuesto anteriormente se desprende que incorporar herramientas de la estadística multivariada y la minería de datos a los análisis de experiencia del cliente sumado al empleo de un marco metodológico ágil mejoran la calidad de los informes, ahorra recursos porque disminuye la probabilidad de malas decisiones de negocio y, por último, ayudan a ofrecer

experiencias diferenciadoras y centradas en los clientes que se traducen en mayor recompra de los clientes, más cantidad de usuarios y, por tanto, en mayores beneficios para la empresa.

6. Bibliografía

ANETCOM. (2013). *Estrategias de marketing para las pymes.* Valencia: ANETCOM.

Barrios Arce, J. (26 de Julio de 2019). *Hearth Big Data*. Obtenido de https://www.juanbarrios.com/matriz-de-confusion-y-sus-metricas/

ContentLab. (5 de Junio de 2019). *Gestión*. Obtenido de Gestión: https://gestion.pe/especial/businessstyle/innovacion/metodologias-agiles-que-son-y-que-son-importantes-noticia-1994291

Contxto. (23 de Octubre de 2019). *Contxto*. Obtenido de https://contxto.com/es/brasil/softbank-lidera-ronda-de-us46-65-millones-para-solucion-brasilena-de-e-commerce-olist/

El blog de John A. Carvajal. (25 de Mayo de 2016). Obtenido de http://blog.jacagudelo.com/las-6-v-del-big-data/

Exame. (28 de Agosto de 2020). *Exame*. Obtenido de https://exame.com/pme/e-commerce-brasil-cresce-47-primeiro-semestre-alta-20-anos/

Gazzeta Do Povo. (27 de Noviembre de 2020). *Gazzeta Do Povo*. Obtenido de https://www.gazetadopovo.com.br/gazz-conecta/expansao-internacional-e-aquisicao-de-startups-os-planos-da-olist-apos-aporte-de-r-310-milhoes/

Hung, S., Yen, D., & Wang, H. (2006). Applying data mining to telecom churn management. *Expert Syst. Appl. 31*.

Kezmo. (20 de Marzo de 2017). *Kezmo Blog*. Obtenido de https://blog.kezmo.com/qu%C3%A9-son-las-metodolog%C3%ADas-%C3%A1giles-y-por-qu%C3%A9-debes-implementarlas-en-tu-organizaci%C3%B3n-484a510e5b0

Kinberg, H., & Skaring, M. (2010). *Kanban and Scrum making the most of both.* United States of America: C4Media Inc.

Kotu, V., & Deshpande, B. (2015). En V. Kotu, & B. Deshpande, *Predictive Analytics and Data Mining.* Waltham, Massachusetts, USA: Elsevier Inc.

Lafuente, E. (10 de Agosto de 2020). *La Nacion*. Obtenido de La Nacion: https://www.lanacion.com.ar/economia/negocios/mercado-libre-facturo-us8784-millones-segundo-trimestre-nid2417375

Lam Diaz, R. M. (2005). *Metodología para la confección de un proyecto de investigación*. La Habana, Cuba.

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insight to Value. *MIT Sloan Management Review, 52*(2).

Lemon, K. N., & Verhoef, P. C. (Noviembre de 2016). Understanding Customer Experience Throughout the Customer Journey. *Journal of Marketing: AMA/MSI Special Issue*, 69-96.

Lemon, K., & Verhoef, P. (2016). Understanding Customer Experience. *Journal of Marketing: AMA/MSI Special Issue*.

LN Creatividad y Tecnología. (19 de Agosto de 2019). *LN Creatividad y Tecnología*. Obtenido de LN Creatividad y Tecnología: https://www.luisan.net/blog/transformacion-digital/que-son-las-metodologias-agiles

McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*.

Mierswa, , I.; Klinkenberg, R.; RapidMiner 9.7. (2020). *https://rapidminer.com/*. Obtenido de https://rapidminer.com/

Moreno, G. (11 de Septiembre de 2020). Banco Galicia se va de Instagram por las estafas virtuales. *BAE Negocios*.

Olist, & Sionek, A. (2018). *https://www.kaggle.com/*. Obtenido de https://olist.com/: https://www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_order_reviews_dataset.csv

Olist;. (Diciembre de 2020). *Olist*. Obtenido de https://olist.com/sobre-o-olist/

Oracle Corporation. (2020). *https://www.mysql.com/*. Obtenido de https://dev.mysql.com/downloads/installer/

TiThink. (16 de Octubre de 2018). *TiThink*. Recuperado el 2020, de TiThink: https://www.tithink.com/es/2018/10/16/metodologias-agiles-que-son-y-para-que-sirven/

WEREDA, W., & GRZYBOWSKA, M. (2016). CUSTOMER EXPERIENCE – DOES IT MATTER? *MODERN MANAGEMENT REVIEW*, 199-207.

Witten, I., Frank, E., Hall, M., & Pal, C. (2017). Data Mining - Practical Machine Learning Tools and Techniches. En I. H. Witten, E. Frank, M. A. Hall, & C. J. Pal, *Data Mining - Practical Machine Learning Tools and Techniches.* Cambridge: Elseiver Inc.

7. Anexo

7.1. Descripción de la base de datos

customer_dataset	
Customer_id	string
Customer_unique_date	string
Customer_Zip_Cod	Numeric
Customer_City	String
Customer_State	string
geo_location_dataset	
Geolocation_Zip_Cod	Numeric
geolocation_lat	Numeric
geolocation_ing	Numeric
geolocation_city	string
geolocation_state	String
orders_dataset	
Order_id	String
Customer_id	String
order_status	String
order_purchase_timestamp	Date
order_approved_at	Date
order_delivered_carried_date	Date
order_delivered_customer_date	Date
order_estimated_delivy_date	Date
order_reviews_dataset	
review_id	String
order_id	String
review_score	Numeric
review_comment_title	String
review_comment_message	String
review_creation_date	Date
review_answer_timestamp	Date
order_items_dataset	
order_id	String
order_item_id	String
product_id	String
seller_id	String
shipping_limit_date	Date
price	Numeric
freight_value	Numeric
order_payments_dataset	
payment_sequential	Numeric
payment_type	String
payment_installments	Numeric
payment_value	Numeric
sellers_dataset	
seller_id	String
seller_zip_code_prefix	Numeric
seller_city	String
seller_state	String

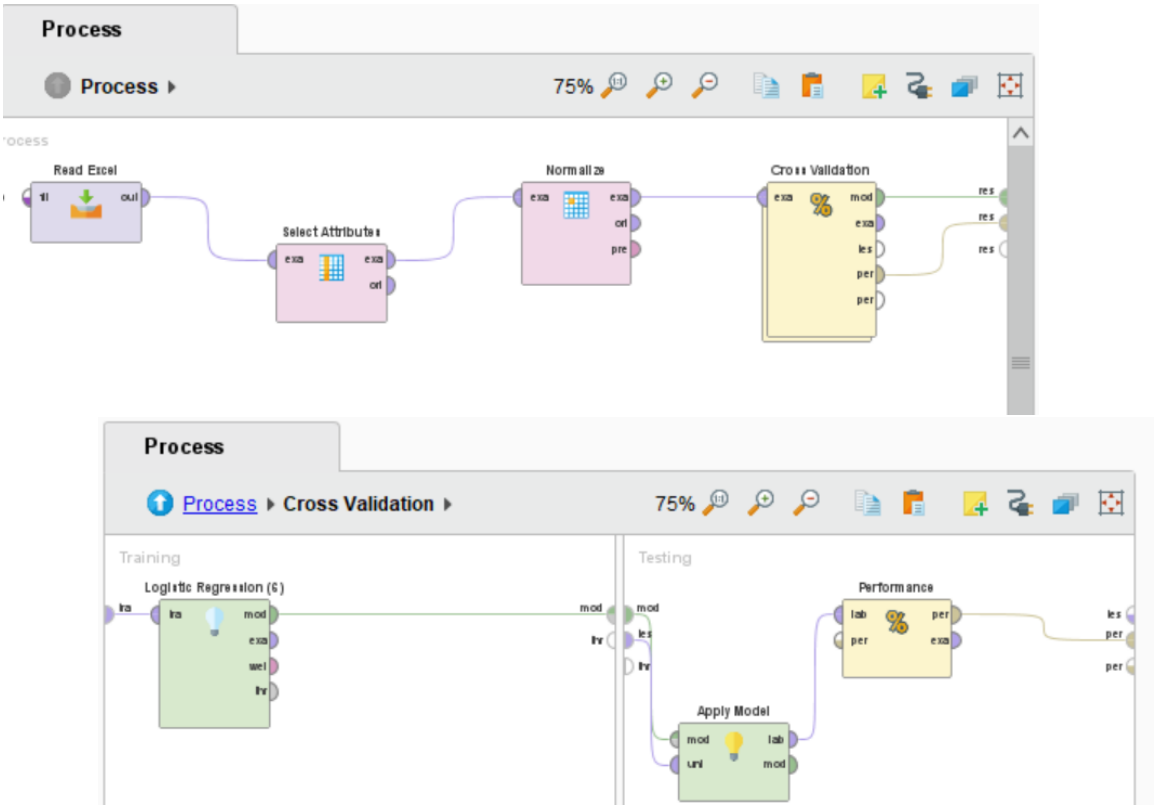
7.2. Modelos utilizados y parámetros

Modelo	Parámetros	Valor
Regresión Logistica	Solver	Auto
	Reproducible	√
	Maximimim number of threads	1
	Use regularization	√
	Lambda search	√
	Number of lambdas	0
	Early stopping	√
	Standarize	√
Rule Induction	Criterion	Information_gain
	Sample ratio	0.5
	Pureness	0.8
	Minimal prune benefit	0.6
Random Forest	Number of trees	670
	Criterion	Accuracy
	Maximal depth	30
	Guess subset ratio	√
Random Forest (para Stacking)	Number of trees	460
	Criterion	Information_gain
	Maximal depth	3
	Guess subset ratio	√

8. Apéndices

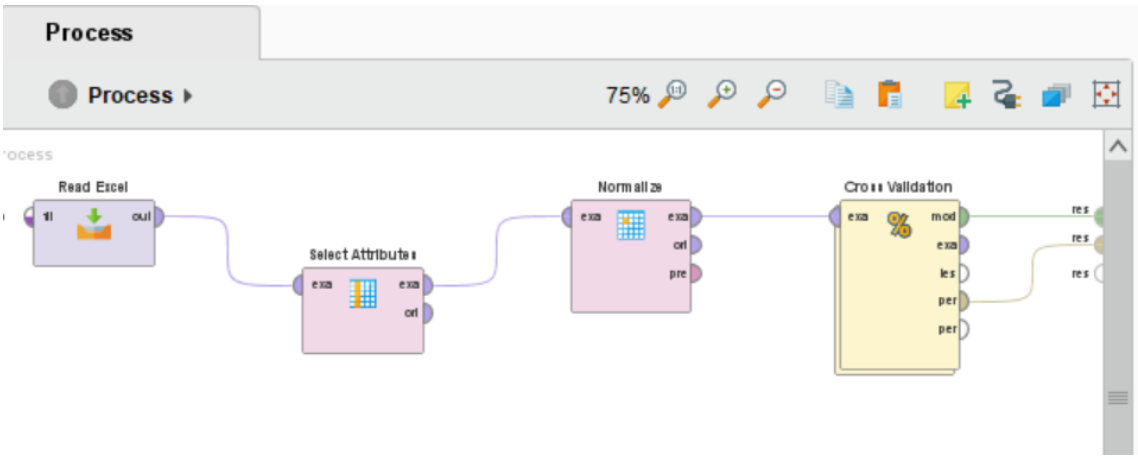
8.1. Código y/o capturas de pantalla de procedimientos

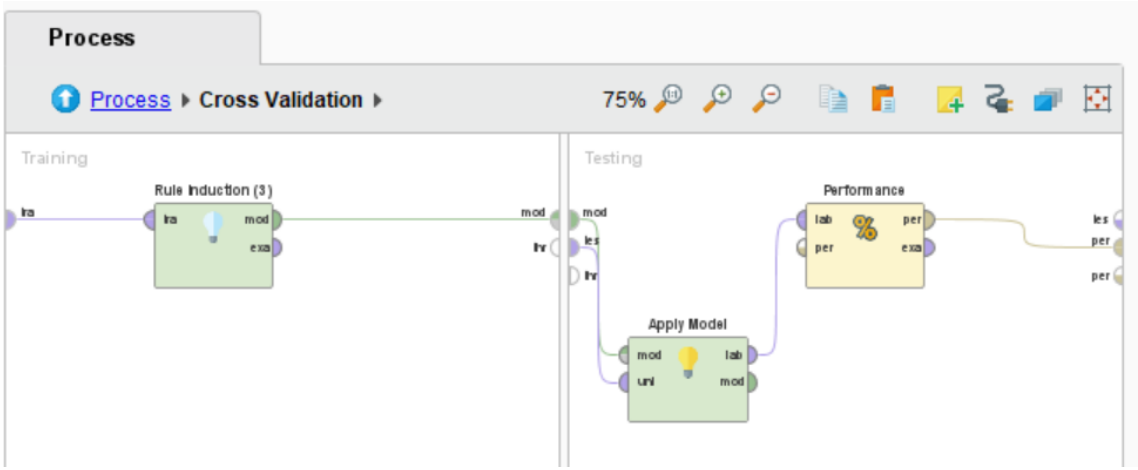
Regresión Logística



Parámetros	Valor
Solver	Auto
Reproducible	√
Maximimim number of threads	1
Use regularization	√
Lambda search	√
Number of lambdas	0
Early stopping	√
Standarize	√

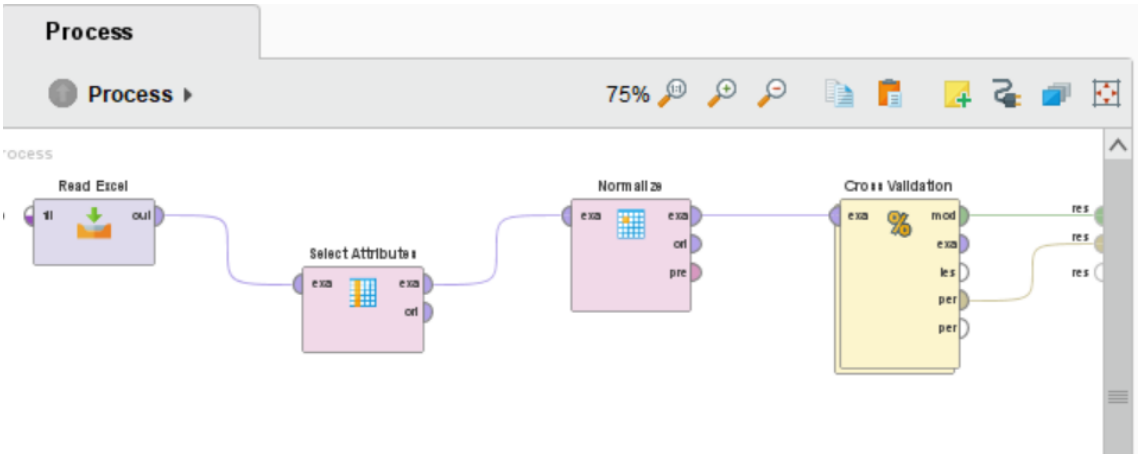
Rule Induction

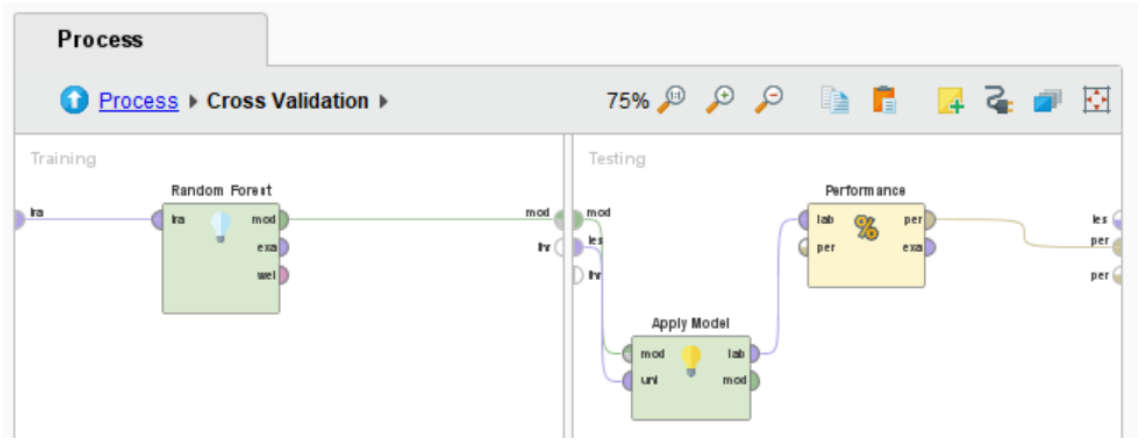




Parametros	Valor
Criterion	Information_gain
Sample ratio	0.5
Pureness	0.8
Minimal prune benefit	0.6

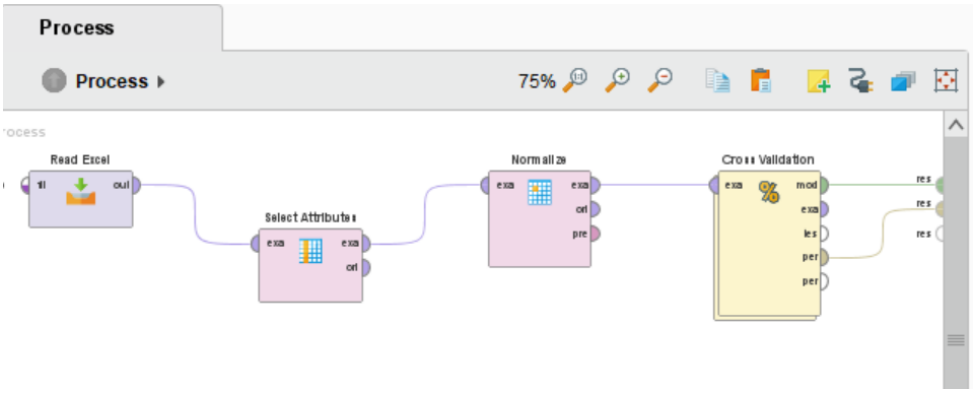
Random Forest

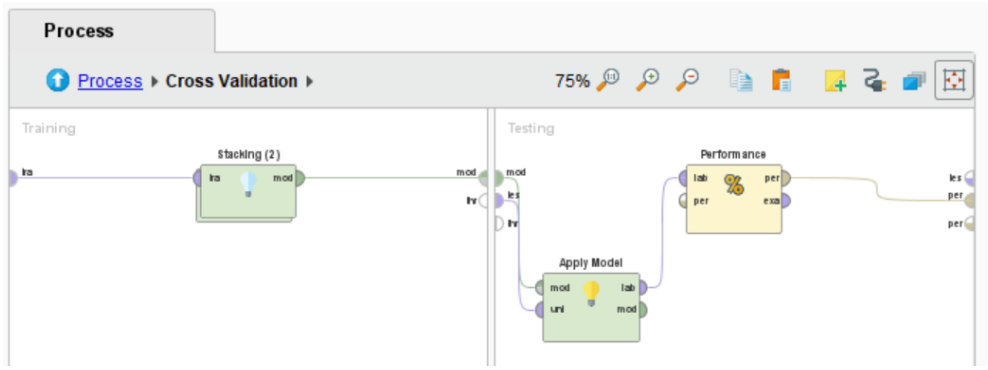




Parametros	Valor
Number of trees	670
Criterion	Accuracy
Maximal depth	30
Guess subset ratio	√

Stacking





Parametros	Valor
Number of trees	460
Criterion	Information_gain
Maximal depth	3
Guess subset ratio	√