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

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

EL DESARROLLO DE UN MODELO PREDICTIVO EN UN COMERCIO DE INDUMENTARIA

ANÁLISIS Y UTILIZACIÓN DEL BIG DATA PARA PREDECIR EL MÉTODO DE PAGO

AUTOR: ALAN DAVID CASSIN

MENTOR: MELISA ELFENBAUM

MARZO 2021

Resumen

El uso del *Big Data* en las organizaciones es una modalidad que crece diariamente. Su explotación permite conocer información que ayuda a tomar mejores decisiones. En el siguiente trabajo se desarrollará un modelo predictivo que será aplicado a un comercio de indumentaria ubicado en AMBA, Argentina. El objetivo general del trabajo se centrará en generar una nueva herramienta para predecir el método de pago elegido por el cliente a la hora de realizar la compra. Las opciones de clasificación serán efectivo u otros métodos de pago. Ante este fin, se deberá responder la siguiente pregunta: ¿puede una PyME argentina desarrollar un método predictivo que ayude a afrontar alguna de sus problemáticas de su operativa diaria? Durante el transcurso del trabajo final, la consulta será respondida mediante la creación de 3 modelos que utilizan diferentes técnicas: *KNN*, Redes Neuronales *y Gradient Boosted Trees*. Previamente a la creación de estos modelos, se procesó la base de datos y se optimizaron los parámetros de cada uno de los modelos. Estos fueron medidos y comparados con la métrica *accuracy* (exactitud). En base a los resultados obtenidos, se desprende que el mejor modelo para predecir el método de pago es el que utiliza el algoritmo de *Gradient Boosted Trees*. También se categorizaron las diferentes variables, y se determinó, que la categoría más influyente en el modelo corresponde a los atributos relacionados con el precio. Por lo tanto, se puede afirmar que es posible realizar un modelo que prediga el método de pago, y de esta forma, que pueda contribuir a una mejor toma de decisiones dentro de un ámbito organizacional, generando como consecuencia la obtención de mejores rendimientos.

Palabras clave: Big Data, Comercio de Indumentaria, Método de Pago, Modelo Predictivo, *Gradient Boosted Trees.*

Índice

Introducción

Los métodos de pago han ido evolucionando e incrementándose a través del tiempo. El primer método de pago utilizado fue el trueque, este prevaleció hasta la creación de las monedas acuñadas con diferentes metales preciosos (aproximadamente en el año 700 A.C.). Casi un milenio después, los bancos comenzaron a instaurar sus monedas y otros documentos como cheques con el mismo propósito. Luego de un período de estancamiento con relación a la creación de nuevos métodos de pago, a mediados del siglo XX, se crean las primeras formas de pago que no involucran la transferencia física de ninguna unidad: las tarjetas de crédito y débito. A fines de este siglo, con el desarrollo de Internet, se introdujeron los primeros pagos electrónicos. En los últimos años, la creación de diversas criptomonedas y aplicaciones móviles que permiten realizar pagos han incrementado los métodos de pago que pueden definirse como virtuales o que no involucran efectivo. Éstas diversas opciones resultan cada vez más simples y ágiles. Y se espera que, en los próximos años, desaparezca el dinero físico (Tafra, 2016).

El método de pago se considera de gran importancia para las organizaciones ya que influyen en su *cashflow*[1] o flujo de caja. También, cabe mencionar que, dependiendo del método de pago elegido por el cliente, se incurren en distintos costos relacionados con la disponibilidad del dinero para las empresas. Por lo tanto, resulta de vital importancia estimar esta información de manera precisa ya que en caso de contar con ella se podrían tomar mejores decisiones de cara al negocio.

Según la Encuesta Nacional de Gastos de los Hogares realizada por el Instituto Nacional de Estadística y Censos (INDEC) para el período transcurrido entre 2017 y 2018, los gastos realizados con finalidad en prendas de vestir y calzado en el GBA, son pagados principalmente en efectivo, seguido por compras con tarjetas de crédito y en tercer lugar se encuentran los pagos con tarjetas de débito (Instituto Nacional de Estadística y Censos (INDEC), 2019). Estos datos reflejan otra problemática que acontece en la Argentina la cual es el grado de informalidad (Donza, Poy, & Salvia, 2019, pág. 11) y el carecimiento en la bancarización de la población.

[1] Es un término utilizado en finanzas. Hace referencia a la acumulación neta de activos líquidos en un periodo determinado.

De esta manera, se presenta la problemática relacionada a la forma de pago que utilizan los clientes, y se plantea el interrogante acerca de si es posible adelantarse a los hechos y conocer el método de pago antes de que este sea realizado.

La solución a este problema operativo puede ser abordado mediante una herramienta que Carlos Espino Timón introduce como modelo predictivo:

> (…) son modelos de la relación entre el rendimiento específico de una unidad en una muestra y uno o más atributos o características conocidos de la unidad. El objeto del modelo es evaluar la probabilidad de que una unidad similar en una muestra diferente exhiba un comportamiento específico. (…)
>
> El análisis predictivo construye un modelo estadístico que utiliza los datos existentes para predecir datos de los cuales no se dispone. Como ejemplo del análisis predictivo se incluyen las líneas de tendencia o la puntuación de la influencia. Para la creación del modelo predictivo se utilizan unidades de muestra disponibles con atributos conocidos y un comportamiento conocido, a este conjunto de datos se le denomina conjunto de entrenamiento. Por otro lado, se utilizará una series de unidades de otra muestra con atributos similares, pero de las cuales no se conoce su comportamiento, a este conjunto de datos se le denomina conjunto de prueba. (Espino Timón, 2017)

Entonces, ¿se puede predecir la forma de pago que utilizarán los clientes al efectuar una compra en el comercio en estudio? A través del desarrollo del modelo predictivo que se explicará en este informe, se responderá la pregunta en cuestión la cual define el objetivo general del trabajo final de integración.

El trabajo final presentado constará de 4 secciones. Éstas, irán cumpliendo con los objetivos específicos establecidos: describir la importancia del *Big Data* en las organizaciones, preparar los datos y seleccionar la técnica que mejores resultados arroje en el modelo, y por último, analizar y evaluar los beneficios que apareja el desarrollo y uso de un modelo predictivo. En la primera sección, se presentará a la organización y la importancia de contar con grandes volúmenes de datos en las organizaciones, luego se analizará la gestión de los datos por parte de la empresa y se presentarán ciertas problemáticas que tendrá que enfrentar la compañía relacionados a ellos. En segundo lugar, se describirá la metodología que se utilizará con el fin de crear un modelo predictivo que pueda pronosticar el método de pago que utilizarán los clientes al realizar sus compras en tiendas físicas de indumentaria

ubicadas en el Gran Buenos Aires dentro de la República Argentina en el mes de septiembre del año 2019. Para ello, se describirá la base de datos utilizada y los atributos que esta posee (tales como la descripción de los artículos comprados, las fechas, el lugar de la compra y el vendedor), la "limpieza" realizada y el procesamiento requerido, las diferentes técnicas aplicadas con sus diferentes modos de entrenamiento y sus métricas y finalmente los resultados de las pruebas. En tercer lugar, se explicarán los diferentes pasos que se deben llevar a cabo para una correcta implementación de un modelo de aprendizaje automático, ejemplificando sobre un sistema de recomendación que podría ser integrado a la página web de la organización. También se expondrán diferentes reportes que ayudarían al directorio para lograr una mejor gestión y toma de decisiones. Asimismo, se comentará acerca de la metodología ágil que resulta más adecuada para llevar adelante este proyecto en esta organización. En la última sección, se analizarán los resultados y las conclusiones obtenidas. También se mencionarán otras aplicaciones del *Big Data* que podrían ser útiles para la organización y que podrán ser desarrolladas en futuros trabajos.

Apartado 1. Gestión de Datos en Contextos Organizacionales

Contar con una cantidad innumerable de datos que no son gestionados o que son gestionados de manera incorrecta, se puede comparar con contar con un destornillador guardado en un armario de una casa, la cual necesita ajustar muchos tornillos, o a usar esa herramienta para intentar martillar un clavo.

El costo y el tiempo relacionado a la generación de esos datos, el almacenamiento y su protección debería justificar una correcta utilización de los mismos. Mientras que su uso debería estar enfocado en un claro propósito y objetivo.

Para comprender la importancia de los datos, en el siguiente apartado, se contextualizará a la organización en cuestión, se describirá la gestión sobre los datos que genera y, por último, se indicarán ciertas problemáticas relacionadas que enfrenta la compañía.

1.1. Descripción de la Organización

El presente trabajo final se desarrolla sobre una organización con fines de lucro que se dedica a la venta de indumentaria femenina. Esta empresa es originaria de Argentina, más precisamente de Buenos Aires, y fue fundada hace más de 30 años. Con el paso del tiempo, logró expandirse y abrir diferentes tiendas en el país. También, logró afianzar su marca y aumentar sus ventas mayoristas a través de distintas franquicias. Entrado el nuevo siglo, logró desarrollar su página de internet y junto con ella la venta electrónica. Las nuevas tecnologías, las redes sociales y otras nuevas herramientas lograron hacer crecer la venta online año tras año. Por cuestiones de privacidad y seguridad, el nombre de la organización se mantendrá en el anonimato.

Según el Ministerio de Producción y Trabajo, un comercio es considerado PyME[2] según el promedio de los tres últimos ejercicios comerciales o años fiscales cerrados de su facturación o de cantidad de empleo generado (Estado Argentino, 2019). Por lo tanto, al encontrarse dentro de estos parámetros definidos, es considerada una PyME. La compañía emplea cerca de 200 personas, cuenta con más de 10 locales propios y numerosas franquicias. Además, como se mencionó anteriormente, cuenta con su página web que

[2] Pequeña y Mediana Empresa.

permite realizar compras por ese medio. En cuanto al producto de venta, cuenta con producción propia y reventa de productos terminados.

Al explorar su estructura organizacional, se observa que la misma se divide en 3 jerarquías que se encuentran bien delimitadas por sus mandos: altos, medios y bajos. Esto se puede visualizar en la siguiente Figura.

Figura 1

Estructura Organizacional

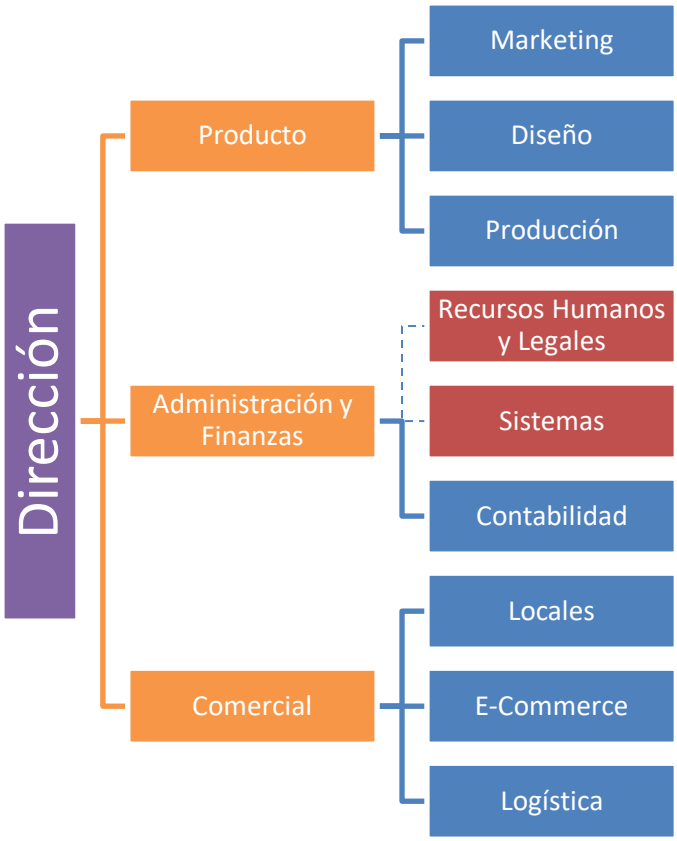


Figura 1: Estructura organizacional. Fuente: *elaboración propia.*

La dirección se compone de una mesa directiva con varias personas que toman las decisiones más importantes y estratégicas de la compañía. Luego, se divide en 3 principales departamentos que tienen sus respectivos gerentes como responsables de los mismos: Producto, Administración y Finanzas y Comercial. El primero se enfoca en sus productos de venta, este es el departamento más creativo e innovador de la empresa. El segundo, abarca la facturación, los pagos, la contabilidad, la creación de reportes, entre otros. Pero también cuenta con la responsabilidad sobre las áreas de Recursos Humanos y Legales y de Sistemas

o comúnmente llamado *IT*[3]. Estas áreas no se desenvuelven dentro de la empresa, sino que se encuentran tercerizadas y su desarrollo se realiza de manera externa. Por último, se encuentra el departamento Comercial, el cual se enfoca tanto en las ventas minoristas y mayoristas como las de forma *online*. Este departamento también debe coordinar la logística de los productos.

Dado que la empresa nació originalmente como una PyME y fue creciendo a pasos agigantados, la organización fue incorporando personal por su grado de confianza. Actualmente, la compañía intentará modificar esta cultura organizacional e incorporar nuevos empleados que se destaquen por sus capacidades y por su profesionalismo.

Durante 2019 (previo a la pandemia por COVID-19), la proporción de facturación se distribuía de la siguiente manera: 47% por ventas generadas en locales propios, 50% por ventas mayoristas a franquicias y 3% por ventas *e-commerce*. Mientras que, en 2020, el porcentaje se modificó drásticamente: 32%, 40% y 28% respectivamente.

El contexto mundial, y particularmente el argentino, amplían un horizonte de incertidumbre y recesión que preocupa a los directivos de la compañía (Página 12, 2021). Es por eso, que más que nunca se encuentran en la búsqueda de incrementar sus ventas y disminuir sus costos con tal de lograr cierta rentabilidad. Ante esta premisa (la cual no dista de la mayoría de los empresarios), y ante el crecimiento en este tipo de ventas experimentado en los últimos meses en el país (Ceurvels, 2020), se intentará aprovechar al máximo la oportunidad de vender de manera *online.* Además, esta modalidad no cuenta con grandes costos operacionales (como los que posee un comercio a la calle) y puede captar ventas provenientes de diversos lugares alrededor del mundo. Para ello, es indispensable proveerle al usuario una buena experiencia en su recorrido virtual, que lo motive a permanecer en la página web y facilite su compra.

1.2. Gestión de Datos por Parte de la Organización

El concepto *Big Data* resuena en innumerables autores. En el comienzo del siglo, el analista Douglas Laney, lo definió como la gestión de datos en la que se cumplan 3 condiciones: volumen, velocidad y variedad (Laney, 2001). El volumen hace referencia a

[3] El acrónimo *IT* son las siglas en inglés de *Information Technology*, cuyo significado en español se traduce como Tecnología de Información.

numerosos datos, la velocidad al tiempo real en que estos fluyen y la variedad a los diferentes tipos de datos que pueden ser capturados (pueden clasificarse de manera simple como estructurados, semi-estructurados o no estructurados). Mientras que según John Akred — quien fuera el fundador y CTO de SVDS (una empresa especializada en *Data Science*) — lo describe como la combinación de un enfoque orientado a guiar la toma de decisiones con descubrimientos analíticos que se extraen de los datos (Akred, 2014). En resumen, Sosa Escudero lo define como "la copiosa cantidad de datos producidos espontáneamente por la interacción con dispositivos interconectados" (Sosa Escudero, 2020). Con un buen aprovechamiento de los mismos, estos datos pueden usarse para el beneficio de quienes los poseen. Ya sea para su análisis para luego tomar mejores decisiones, como para la creación de un modelo predictivo o clasificatorio entre otros ejemplos.

La explotación de estos datos de manera adecuada proporciona una gran ventaja competitiva a las organizaciones y empresas, mientras que, su ignorancia produce grandes riesgos y las hará cada vez menos competitivas (Joyanes Aguilar, 2013).

Desde la creación de la internet a la actualidad, la generación de datos ha crecido de manera exponencial. Tal como se expone en un estudio realizado por International Data Corporation (IDC), los datos crecen de la mano de la tecnología y de las herramientas informáticas desarrolladas:

> Much of today's economy relies on data, and this reliance will only increase in the future as companies capture, catalog, and cash in on data in every step of their supply chain; enterprises collect vast sums of customer data to provide greater levels of personalization; and consumers integrate social media, entertainment, cloud storage, and real-time personalized services into their streams of life.
>
> The consequence of this increasing reliance on data will be a never-ending expansion in the size of the Global Datasphere. Estimated to be 33 ZB in 2018, IDC forecasts the Global Datasphere to grow to 175 ZB by 2025 (Reinsel, Gantz, & Rydning, 2018).
>
> [Gran parte de la economía actual se basa en datos, y esta dependencia solo aumentará en el futuro a medida que las empresas capturen, cataloguen y saquen provecho de los datos en cada paso de su cadena de suministro; las empresas recopilan grandes sumas de datos de clientes para proporcionar mayores niveles de personalización; y los

consumidores integran las redes sociales, el entretenimiento, el almacenamiento en la nube y los servicios personalizados en tiempo real en sus corrientes de vida.

La consecuencia de esta creciente dependencia de los datos será una expansión sin fin en el tamaño de la esfera de datos global. Estimado en 33 ZB en 2018, IDC pronostica que la esfera de datos global crecerá a 175 ZB para 2025.]

En este contexto, la organización en cuestión genera datos de manera masiva, de diferente tipo y provenientes de distintas herramientas y tecnologías. Algunos ejemplos representativos son: las ventas generadas de manera presencial en los locales; los movimientos de stock; las interacciones en sus redes sociales; las interacciones de los usuarios en su página web; y las ventas generadas por este medio. Aun así, la gestión de estos datos no es aprovechada al máximo por la compañía ya que la mayoría de estos datos no son gestionados ni extraídos por personal de la empresa para su procesamiento y posterior análisis.

Como se mencionó en la sección anterior, el área encargada para el mantenimiento de los sistemas y herramientas informáticas de la empresa se encuentra tercerizado. Esto significa que la organización en sí posee una arquitectura de datos gestionada y restringida por terceros. Como consecuencia, los datos que finalmente se visualizan solo se reducen a los estructurados, como la información transaccional que se procesa mediante un ERP[4] interno y un sistema contable: detalle de los productos, información de la mercadería, información de sus materias primas, información de las ventas generadas, datos de los envíos, reportes de costos, entre otros reportes disponibles. Todos estos datos se actualizan mediante diferentes procesos (algunos manuales y otros automatizados) por empleados de la organización. Mientras que el mantenimiento y el soporte de estos sistemas y programas lo realiza la empresa contratada. En cuanto al almacenamiento de los datos, se realiza en servidores físicos que son propiedad de la organización.

Con respecto a la página web y los datos provenientes de la interacción de los usuarios con el dominio, estos datos no estructurados no son extraídos para su posterior análisis. Los únicos datos que son extraídos (estructurados), y que luego son integrados al sistema ERP de la empresa, son los relacionados a las ventas realizadas por este medio. Por lo que

[4] Siglas en inglés de *Enterprise Resource Planning*, Planificación de Recursos Empresariales; son sistemas informáticos destinados a la administración de recursos en una organización.

corresponde a los datos generados a través de las diferentes redes sociales, estos son revisados por su *Community Manager* mediante la herramienta de Google Analytics para visualizar y analizar sus métricas. El desarrollo, mantenimiento y soporte de la página web, es realizado por otra empresa contratada (externa a la organización), que también gestiona y almacena los datos que se generan mediante este medio utilizando el servicio de Amazon. Dentro de la organización, tanto la página web como las redes sociales se encuentran bajo la responsabilidad del área de *e-commerce*.

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

Ninguna organización se encuentra exenta a las problemáticas de la gestión de los datos. Aunque las complejidades que tienen que afrontar pueden ser muy disímiles entre sí, existen ciertas cuestiones que toda organización tiene que resolver y decisiones que sus responsables deben tomar al respecto. Por ejemplo, algunas de las preguntas que la mayoría de los involucrados debe responder son: ¿cómo deben resguardar los datos personales y privados de sus empleados, usuarios o clientes?; ¿qué servicio de almacenamiento es más provechoso?; o ¿resulta más conveniente realizar una ingesta de datos en tiempo real o en períodos definidos y por lotes?

Tal como se mencionó anteriormente, la gestión de los datos por parte de la organización en estudio puede ser calificada como pobre o nula dado que en su mayoría se encuentra en manos de un tercero. Esto no significa necesariamente que influya negativamente en sus resultados, sino que la empresa no se siente preparada para gestionarlos de mejor manera que la organización contratada, no cuenta con los recursos y/o no le da la importancia suficiente a los mismos.

Como cualquier problema en cuestión, la resolución del mismo comienza con su reconocimiento y aceptación. La oficialización y conformación de la dirección general de la compañía, ha sido el primer paso y en los próximos meses se trabajará gradualmente en una mejor utilización de los datos. Por lo tanto, la actual cultura y mentalidad organizacional puede ser considerada como la primera problemática a resolver. Históricamente, la toma de decisiones dentro de la organización fue mayoritariamente basada en impresiones y presentimientos que en datos precisos y estadísticas comprobables. Este cambio de cultura que se intenta adoptar podría ser catalogado como *data-driven* (EY, 2015). Tal como se

demostró en un estudio realizado en el MIT Center, las empresas que utilizan los datos para tomar todo tipo de decisiones, son un 5% más productivas y logran una rentabilidad del 6% mayor a sus competidores que no los utilizan (McAfee & Brynjolfsson, 2012).

Una segunda problemática a tratar sería la reestructuración del área de Sistemas con el fin de crear un departamento propio que pueda gestionar y dar soporte a las demás áreas de la empresa en los sistemas y herramientas que utiliza. Adicionalmente, también sería de suma utilidad la contratación de personal que pueda extraer, procesar, gestionar y crear reportes o modelos de aprendizaje automático a implementar en base a los datos con los que cuenta la compañía. Estos reportes podrían ser visualizados mediante herramientas de *Business Intelligence (BI)*, compartiéndolos con los sectores afectados y permitiendo el análisis de la información con mayor frecuencia y rapidez para una posterior toma de decisiones (Tovar, 2017). Mientras que los modelos de aprendizaje automático servirían para predecir ciertos sucesos que pueden ocurrir en el funcionamiento de la empresa. Como, por ejemplo, determinar la forma de pago que utilizará un cliente al realizar su compra de manera presencial o la creación de un sistema de recomendación que ayude al usuario en su recorrido virtual mostrándole productos que resulten de su interés.

La tercera problemática se relaciona directamente con la segunda, la cual redunda sobre los recursos que debería invertir la organización en la gestión de los datos. Desistir de subcontratar el servicio provisto para el área de Sistemas y la creación de uno propio que pueda contar con personal calificado y tecnologías apropiadas para el correcto desarrollo de sus tareas implica una gran inversión a realizar por parte de la empresa la cual no toda organización está dispuesta a efectuar.

Otra problemática, es el tratamiento de los datos confidenciales y la seguridad de su información. Esto comprende un desafío que debe afrontar la organización y que requiere de cuidado extremo para no comprometer su confianza e imagen pública.

Por último, se puede mencionar como una gran problemática a la competencia con otras organizaciones las cuales explotan los datos que generan. La globalización y el *e-commerce* han derribado la barrera de competencia por ubicación geográfica. Esta conectividad mundial que se ha logrado permite extender el *target* de la compañía y así captar nuevos clientes que se traducen en nuevas ventas. O también, perder los compradores existentes en manos de compañías que antiguamente no hubiesen podido competir debido a su locación.

Es por eso, que resulta indispensable desarrollar una buena y placentera experiencia en el usuario que utiliza la página web de la organización. Además, el proceso y la experiencia de post venta deben ser satisfactorios para así poder retener y fidelizar al cliente, generando potenciales ventas. En cuanto a los datos, estos pueden ser utilizados de diversas formas, ya sea para desarrollar un *chatbot*[5] que pueda responder en tiempo real ciertas preguntas frecuentes, como para analizar, qué les interesa más o menos a los usuarios (basándose en los clics o el tiempo que transcurren en cada pantalla). Existen múltiples utilidades que se pueden realizar con la gestión de los datos, y depende de la organización sacar provecho de ellos de manera anticipada, antes que deban adoptarlos forzosamente luego de que la mayoría de la competencia lo haya hecho y se encuentre en una posición desventajosa.

[5] Son aplicaciones informáticas basadas en la inteligencia artificial que permiten simular la conversación con una persona, dándole respuestas automatizadas a sus dudas o preguntas más comunes.

Apartado 2. Descripción Metodológica: Material y Métodos

A continuación, se presentarán y describirán los datos utilizados en el siguiente trabajo y el desarrollo efectuado para procesarlos.

La utilización de diferentes técnicas nos indicará si la creación de un modelo predictivo para clasificar el método de pago de los clientes puede ser confirmado o contrariado, así como también cuál de estas técnicas es la de mejor desempeño.

La creación y optimización de cada uno de estos tres modelos cumplirá un objetivo específico. Cada modelo utilizará diferentes técnicas: *KNN*, Redes Neuronales y *Gradient Boosted Trees*. Mientras que la comparación entre modelos y la elección del modelo final, en base a los resultados obtenidos, cumplirá con el objetivo general del trabajo final: la creación de un modelo que pueda predecir la elección del método de pago del cliente.

2.1. Descripción de la Base de Datos

La base de datos utilizada contiene datos sobre la venta de indumentaria en diferentes tiendas de comercio minorista ubicados dentro del AMBA (Argentina).

La base de datos fue proporcionada por el directorio de la compañía de manera privada y confidencial[6]. Los resultados obtenidos serán anonimizados en pos de mantener la confidencialidad de los datos. La fecha de relevamiento de los datos data del mes de septiembre del año 2019.

Los datos originales corresponden a información recopilada a través de los sistemas internos de la empresa. Los archivos compartidos son de tipo matriz y sus formatos son xlsx. Los datos son susceptibles de contener errores provenientes de la carga manual de datos o de datos faltantes por contener artículos de temporadas anteriores a la implementación del nuevo sistema de gestión (efectuado a mitad del 2019).

El tipo de datos es mixto de análisis multivariado, los atributos son mixtos (numéricos y categóricos), el número de instancias es de 43,970 y la cantidad de atributos son 33, de los cuales 16 son cualitativos y 17 son cuantitativos. En el Anexo se presenta la estructura del conjunto de datos en mayor detalle.

[6] Se ostenta la autorización correspondiente para utilizar la base de datos y publicar los resultados del trabajo final.

2.2. Procesamiento de Datos

Para el procesamiento de los datos, se utilizaron diferentes softwares. En una primera instancia (la cual puede ser definida como la más trabajosa), los datos fueron procesados mediante el programa Microsoft Excel, luego se utilizó el software RStudio (Team, RStudio, 2020) para realizar el análisis descriptivo y, por último, fueron cargados y procesados en RapidMiner Studio (Mierswa, I.; Klinkenberg, R.; RapidMiner 9.7, 2020) para proceder con el uso de las diferentes técnicas y la creación de los modelos.

La "limpieza" efectuada consistió en la eliminación de ciertos atributos, la creación de nuevos en base a los datos conocidos y no conocidos, el removimiento de ciertas instancias y la suposición de algunos valores faltantes. En el Anexo se detallan cada uno de los atributos.

Resulta importante remarcar que una vez definidos todos los atributos que no contenían datos numéricos, fueron clasificadas en diferentes números enteros con el propósito de contar con una base de datos plenamente numérica y que sea posible de utilizar con ciertas técnicas de procesamiento de datos.

El entrenamiento y la prueba fueron configurados mediante el uso de la técnica de validación cruzada. Esta técnica para evaluar los resultados consiste en partir la base de datos en la cantidad deseada de veces que se quieren realizar diferentes pruebas. Los datos de entrenamiento y de prueba irán variando en cada simulación. Por lo tanto, se obtendrán diferentes errores para cada prueba realizada y el error "final" será el promedio de los errores obtenidos en cada simulación. Para entrenar y evaluar los modelos creados se realizaron 5 y 10 iteraciones, los cuales dividieron la base en 80% y 20% y en 90% y 10% respectivamente. Luego de estas modificaciones, la base de datos cuenta con un total de 36,434 instancias y 31 atributos.

2.3. Análisis de Datos

A continuación, en la Tabla 1, se presentará el análisis descriptivo de todos los atributos que contiene la base de datos y los gráficos de tipo *boxplot* correspondientes (Fig. 2). Como ha sido mencionado anteriormente, se debe tener en cuenta que todas las variables han sido

convertidas en cuantitativas para facilitar el procesamiento de los datos en los modelos creados.

Tabla 1

Descripción de los atributos

Descripción de Atributos	Promedio	Desviación Estándar	Mediana	Media Truncada	Desviación Media Absoluta	Mín.
SUCURSAL	8.52	4.16	9	8.52	4.45	1
ART_COD_MODELO	5398.57	6333.95	1925	4460.8	2720.57	1
CPBTE_FECHA	16.08	8.41	16	16.21	10.38	1
Cot_Dolar	54.54	0.76	55	54.53	1.48	53
FECHA_DIA	4.75	1.88	5	4.9	1.48	1
CPBTE_TURNO	0.9	0.3	1	1	0	0
CPBTE_HORA	15.81	3.01	16	15.88	2.97	8
VENDEDOR_COD	58.61	25.5	61	59.81	31.13	1
Predicción	0.71	0.45	1	0.77	0	0
MONTO_COBRADO	3194.64	3567.79	2146.55	2547.9	2003.66	16.8
Cant_Art_Comp	3.18	3.17	2	2.61	1.48	1
Total_Ventas_Local	3333530	1432883	2753786.92	3329120	1804908.43	1022072.65
Cant_Ventas_Local	1679.44	816.02	1711	1637.61	853.98	470
Ticket_Prom_Local	2177.72	737.93	2450.48	2190.89	787.14	1186.67
Poder_Adq	0.3	0.46	0	0.25	0	0
Nivel_FacturaciÃ³n	1.26	0.7	1	1.32	1.48	0
TEMPO_COD	24.18	4.9	26	25.41	1.48	10
COLOR_COD	16.89	23.6	4	11.67	5.93	0
TALLE_COD	3.44	4.42	3	2.47	1.48	0
ITEM_RUBRO	1.22	0.67	1	1.08	0	0
ITEM_SRUBRO	15.19	15.86	8	12.7	10.38	0
Compra_Con_Cambio	0.05	0.23	0	0	0	0
TK_CANTIDAD	1.04	0.26	1	1	0	1
TK_PRECIO_UNIT	1370.94	1125.01	999	1202.94	858.43	14
TK_MONTO_CON_IVA	1266.87	1091.27	899	1099.08	769.47	0
TK_DESCUENTOS	113.06	243.06	0	54.41	0	0
TK_SENIA	1.53	54.94	0	0	0	0
PRODUCCION	0.35	1.04	0	0.04	0	0

Descripción de Atributos	Máx.	Rango	Asimetría	Curtosis	Error Estándar
SUCURSAL	16	15	0.01	-0.84	0.02
ART_COD_MODELO	43543	43542	1.39	2.52	33.18
CPBTE_FECHA	30	29	-0.05	-1.13	0.04
Cot_Dolar	57	4	0.7	2.05	0

Descripción de Atributos	Máx.	Rango	Asimetría	Curtosis	Error Estándar
FECHA_DIA	7	6	-0.53	-0.91	0.01
CPBTE_TURNO	1	1	-2.72	5.38	0
CPBTE_HORA	22	14	-0.18	-0.91	0.02
VENDEDOR_COD	115	114	-0.33	-0.84	0.13
TOTAL_COBRANZA	44335	44318.2	3.5	20.8	19.76
TIPO_CPBTE_NOMBRE	2	2	-12.55	172.6	0
Predicción	1	1	-0.94	-1.11	0
MONTO_COBRADO	44335	44318.2	3.6	23.38	18.69
Cant_Art_Comp	43	42	5.05	43.56	0.02
Total_Ventas_Local	5489950.4	4467877.75	0.22	-1.38	7506.84
Cant_Ventas_Local	3101	2631	0.42	-0.86	4.28
Ticket_Prom_Local	3224.19	2037.51	-0.15	-1.74	3.87
Poder_Adq	1	1	0.88	-1.23	0
TEMPO_COD	28	18	-1.99	2.73	0.03
COLOR_COD	99	99	1.69	1.9	0.12
TALLE_COD	26	26	3.04	9.63	0.02
ITEM_RUBRO	7	7	5.22	37.16	0
ITEM_SRUBRO	66	66	1.14	0.42	0.08
Compra_Con_Cambio	1	1	3.94	13.54	0
TK_CANTIDAD	10	9	13.06	264.2	0
TK_PRECIO_UNIT	12800	12786	1.49	2.83	5.89
TK_MONTO_CON_IVA	12800	12800	1.57	3.42	5.72
TK_DESCUENTOS	4600.8	4600.8	3.91	26.79	1.27
TK_SENIA	6799	6799	69.38	7011.58	0.29
PRODUCCION	7	7	3.18	10.57	0.01

Tabla 1. Resultado del análisis descriptivo realizado en RStudio sobre todos los atributos presentados en la Base de Datos.

Fuente: *elaboración propia*.

Figura 2

Gráfico Boxplot de los atributos

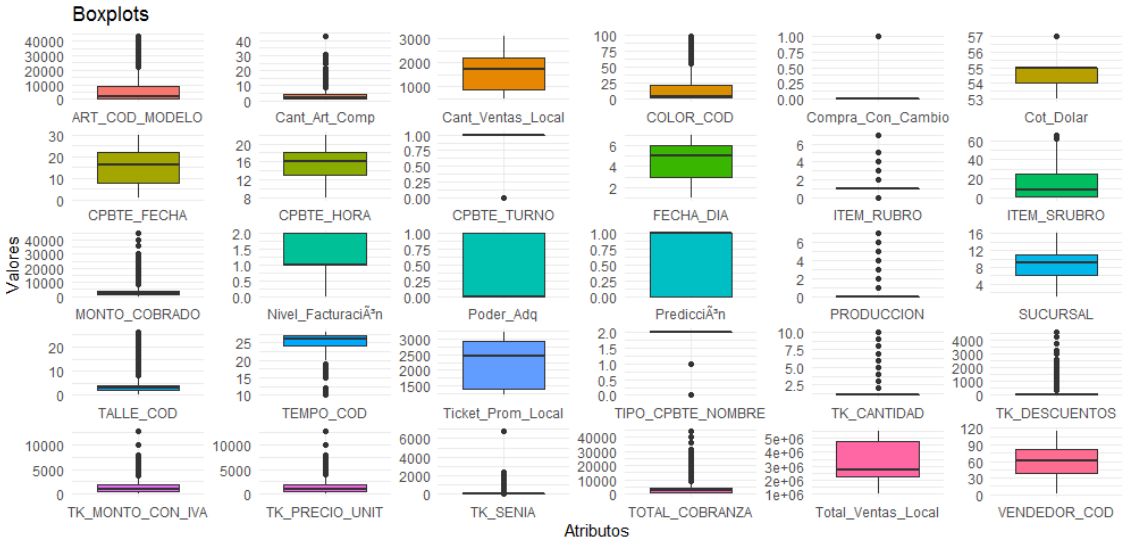


Figura 2. Boxplot de los atributos realizado mediante RStudio. Fuente: *elaboración propia.*

Con el fin conocer el punto de partida de la base de datos y de la variable a predecir, se realizó una exploración de las clasificaciones determinadas para el atributo "Predicción". El método de pago mayoritario y su porcentaje sobre el total de las compras realizadas por los clientes definen el *baseline*[7] del modelo. Como se observa en la Figura 3, el método de pago mayoritario es "Otros Métodos de Pago" y el *baseline* es de 71.30%.

Figura 3

Métodos de pago utilizados por los clientes en la base de datos

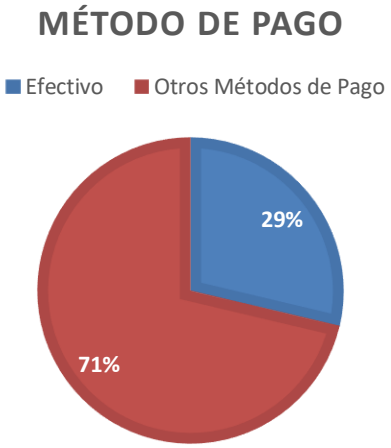


Figura 3. Gráfico de torta que muestra el *baseline* del modelo predictivo. Fuente: *elaboración propia.*

[7] Valor conocido o inicial a partir del cual pueden compararse valores posteriores de lo que se está midiendo.

2.4. Modelos Aplicados

Los modelos se aplicaron utilizando diferentes técnicas. Estos modelos, a partir del aprendizaje supervisado, buscan cumplir con el objetivo de clasificar de la mejor manera posible el método de pago utilizado por los clientes al realizar una compra en alguna de las sucursales de la compañía. El método de pago está representado en el atributo creado "Predicción", el cual puede ser clasificado como 0 (*efectivo*) u 1 (*otros métodos de pago*).

Las variables utilizadas en estos modelos para realizar la clasificación pueden ser encontradas en el Anexo. Estos atributos específicos fueron escogidos al comparar los resultados obtenidos con y sin ellos.

Luego de probar diferentes técnicas para la elección de los 3 modelos, se seleccionaron las que obtenían una mayor precisión en sus resultados, estas son:

1. *KNN*[8]: es un algoritmo de aprendizaje supervisado en donde en base a la cantidad de grupos (K) definidos inicialmente se clasificarán las observaciones teniendo en cuenta su distancia y cercanía con los distintos grupos (Witten, Frank, Hall, & Pal, 2016).
2. Redes Neuronales: es un modelo computacional de clasificación y regresión que está inspirado en las redes neuronales humanas. Este algoritmo trabaja utilizando diferentes capas de neuronas conectadas entre sí que aprenden de sí mismas (Tablada & Torres, 2009).
3. *Gradient Boosted Trees*: es una familia de algoritmos usados tanto en clasificación como en regresión basados en la combinación de árboles de decisión para crear un modelo predictivo más robusto. La generación de los árboles de decisión se realiza de forma secuencial, creándose cada árbol de forma que corrija los errores del árbol anterior (Bosco Mendoza Vega, 2020).

2.5. Métricas

El procedimiento realizado una vez concluida la manipulación y "limpieza" de la base de datos constó de los siguientes pasos:

[8] Vecinos más cercanos.

1. Normalización de los datos.
2. Selección de los atributos que mejoran el rendimiento del modelo.
3. Optimización de los parámetros y selección de métricas.
4. Validación cruzada y obtención de resultados.

Tal como se mencionó anteriormente, el entrenamiento y la prueba para validar los modelos fueron realizados mediante la utilización de la técnica de validación cruzada. Los resultados de los modelos fueron comparados utilizando 5 y 10 iteraciones.

La métrica elegida para evaluar y comparar los modelos es la exactitud (*accuracy*). Esta métrica mide la cantidad de predicciones correctamente realizadas sobre la totalidad de los casos predichos. La varianza también es otra métrica utilizada para comparar los distintos modelos a fin de conocer cuán sobre ajustados se encuentran. Los resultados de estas comparaciones podrán ser visualizados en una tabla comparativa en la sección Resultados.

Finalmente, con el fin de validar la elección del mejor modelo, se mostrará su curva *ROC*[9] y el *AUC*[10] (Kotu & Deshpande, 2014).

2.6. Resultados

Tal como se mencionó anteriormente, se utilizaron tres técnicas diferentes para la creación de los diferentes modelos: *KNN*, Redes Neuronales y *Gradient Boosted Trees*. A continuación, se presentarán los resultados obtenidos en cada uno de estos modelos optimizados. Los resultados expuestos corresponden a la prueba realizada mediante 5 iteraciones de validación cruzada.

El modelo que utiliza el algoritmo *KNN*, arroja resultados con una exactitud del 77.92%, superando el *baseline* por alrededor de 7 puntos porcentuales. La varianza que muestra este modelo es de 0.36%. En la Figura 4 se puede observar su matriz de confusión. Resulta interesante remarcar que el parámetro k (cantidad de grupos) que optimiza el modelo es 2. Por lo tanto, se presupone que cada uno de estos dos grupos se asemeja a los posibles métodos de pago elegidos por el cliente. Por consiguiente, y en base a los datos de cada una de las instancias, estas observaciones se irán posicionando en el plano definido y finalmente

[9] Curva de característica operativa del receptor.
[10] Área bajo la curva *ROC*.

se clasificarán en base a su distancia Euclídea entre su posición y los dos diferentes grupos (elegirá el más cercano).

Figura 4

Matriz de confusión del modelo KNN

accuracy: 77.92% +/- 0.36% (micro average: 77.92%)

	true 1	true 0	class precision
pred. 1	21973	4038	84.48%
pred. 0	4006	6417	61.57%
class recall	84.58%	61.38%	

Figura 4: Resultados en RapidMiner Studio del modelo de KNN. Fuente: (Mierswa, I.; Klinkenberg, R.; RapidMiner 9.7, 2020).

En cuanto al modelo optimizado que utiliza Redes Neuronales, se definieron 3 capas ocultas con 50, 100 y 50 neuronas respectivamente. En la Figura 5 se observa que la exactitud de este modelo es de 76.38% y su varianza es de 0.41%. También se advierte que el modelo trabaja muy bien prediciendo "Otros Métodos de Pago" (logra una exactitud del 92.85%) y no así para la clasificación "Efectivo" (35.45%).

Figura 5

Matriz de confusión del modelo Redes Neuronales

accuracy: 76.38% +/- 0.41% (micro average: 76.38%)

	true 1	true 0	class precision
pred. 1	24121	6749	78.14%
pred. 0	1858	3706	66.61%
class recall	92.85%	35.45%	

Figura 5: Resultados en RapidMiner Studio del modelo de Redes Neuronales. Fuente: *(Mierswa, I.; Klinkenberg, R.; RapidMiner 9.7, 2020).*

Por último, se encuentra el modelo creado que utiliza la técnica de *Gradient Boosted Trees*. Este algoritmo de tipo supervisado trabaja de forma tal que genera múltiples árboles de decisión que pueden ser definidos como "débiles", los cuales van tomando los resultados del modelo generado anteriormente para crear uno más "fuerte" y que como consecuencia prediga mejor que el preliminar. Tal como se puede advertir en la Figura 6, este modelo presenta los mejores resultados. El modelo predice al 87.18% en términos de exactitud (*accuracy*) y su varianza es de 0.20%.

Figura 6

Matriz de confusión del modelo Gradient Boosted Trees

accuracy: 87.18% +/- 0.20% (micro average: 87.18%)

	true 1	true 0	class precision
pred. 1	24514	3205	88.44%
pred. 0	1465	7250	83.19%
class recall	94.36%	69.34%	

Figura 6: Resultados en RapidMiner Studio del modelo de *Gradient Boosted Trees*. Fuente: *(Mierswa, I.; Klinkenberg, R.; RapidMiner 9.7, 2020).*

Por lo tanto, al comparar los diferentes modelos se puede observar que el que mejor funciona es el que utiliza la técnica de *Gradient Boosted Trees* (como se exhibe en la Tabla 2).

Tabla 2

Comparación de los resultados de los modelos

	Modelos			
Cross Validation	5 Iteraciones		10 Iteraciones	
Métricas	Exactitud	Varianza	Exactitud	Varianza
KNN	77.92%	0.36%	78.64%	**0.43%**
Redes Neuronales	76.38%	0.41%	76.60%	0.61%
Gradient Boosted Trees	**87.18%**	**0.20%**	**87.94%**	0.59%

Tabla 2. Comparación entre resultados de los diferentes modelos predictivos optimizados. Fuente: *elaboración propia.*

Tal como se planteó en la sección de Métricas, y con el propósito de validar el mejor modelo creado, en la siguiente figura, se presenta la curva *ROC* y la métrica de *AUC* del modelo que presenta mejor exactitud.

El *AUC* alcanzado es de 92.5%. Estos resultados son muy convincentes e incluso superan los obtenidos a través del uso de la exactitud como métrica principal. Consecuentemente, se puede confirmar la validez de este modelo.

Figura 7

AUC y curva ROC del modelo Gradient Boosted Trees

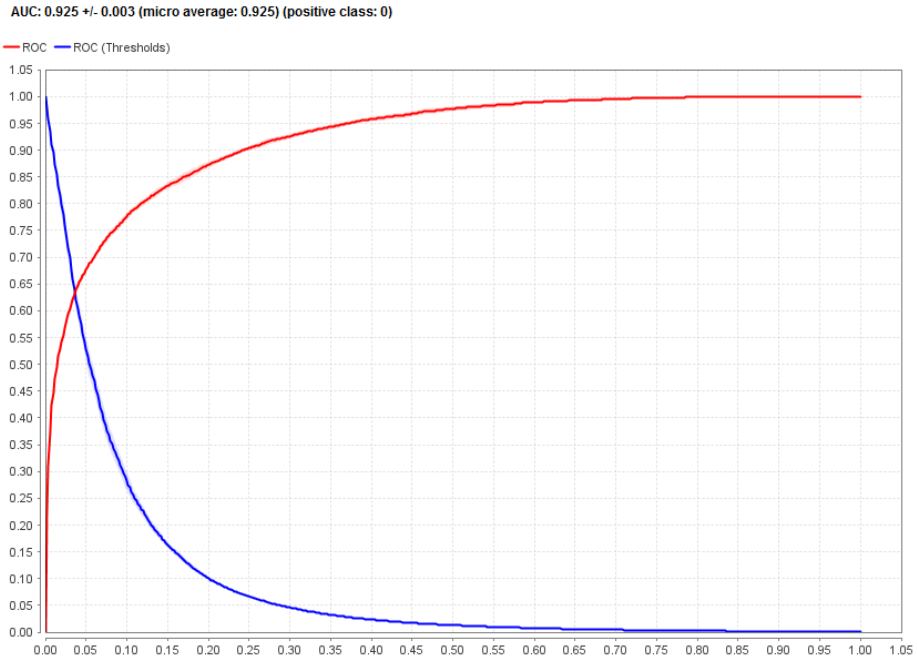


Figura 7: Resultados en RapidMiner Studio del modelo de *Gradient Boosted Trees*. Fuente: *(Mierswa, I.; Klinkenberg, R.; RapidMiner 9.7, 2020).*

Con el propósito de comprender e interpretar la importancia de los atributos en el modelo predictivo, se agruparon en diferentes categorías y se probó el impacto en su exactitud en el caso que se decidiera excluirlas. Las categorías creadas se observan en la siguiente tabla y se ubican en orden en base a su importancia:

Tabla 3

Atributos agrupados en categorías

Categoría	Precio	Fecha	Hora	Vendedor
Atributos	TOTAL_COBRANZA MONTO_COBRADO TK_PRECIO_UNIT TK_MONTO_CON_IVA	CPBTE_FECHA Cot_Dolar FECHA_DIA	CPBTE_TURNO CPBTE_HORA	VENDEDOR_COD
Categoría	**Cantidad Artículos**	**Sucursal**	**Compra**	**Detalle**
Atributos	Cant_Art_Comp	SUCURSAL Total_Ventas_Local Cant_Ventas_Local Nivel_Facturación	TIPO_CPBTE_NOMBRE Compra_Con_Cambio TK_SENIA	PRODUCCION ITEM_RUBRO

Tabla 3. Categoría de atributos ordenadas en base a su importancia en los modelos. Fuente: *elaboración propia.*

En resumen, podemos asegurar que la categoría “Precio” es la más importante en el modelo mientras que la categoría “Detalle” es la de menor importancia.

En el Anexo se presenta en detalle la descripción de cada uno de los modelos con los parámetros y métricas utilizadas.

Apartado 3. Implementación de Modelos de Aprendizaje Automático

Durante el transcurso del trabajo final, se remarcaron ciertas utilizaciones que se le pueden dar a los datos y el beneficio que traería aparejado su explotación.

El modelo desarrollado anteriormente muestra como ejemplo que, con los datos disponibles, se puede predecir el método de pago de los clientes en los locales físicos.

Con el fin de demostrar otro tipo de utilizaciones de los datos, y otras herramientas que se podrían aprovechar, en el siguiente apartado, se mostrarán los pasos necesarios para construir un modelo de aprendizaje automático que funcione como un sistema de recomendación de los productos en la página web de la organización utilizando el servicio provisto por Azure Machine Learning Studio (Microsoft, 2020). Luego, a través de los datos que se utilizaron para generar el sistema de recomendación, se mostrarán ciertos reportes creados con la herramienta Power BI (un software desarrollado por Microsoft de *Business Intelligence*[11]*)* que podrían ser útiles para la toma de ciertas decisiones. Finalmente, se reflexionará acerca de la metodología que resulta más apropiada para realizar la implementación de este proyecto.

3.1. Implementación de un Sistema de Recomendación

Al igual que en el modelo desarrollado para predecir el método de pago, los pasos realizados pueden enumerase de la siguiente forma:

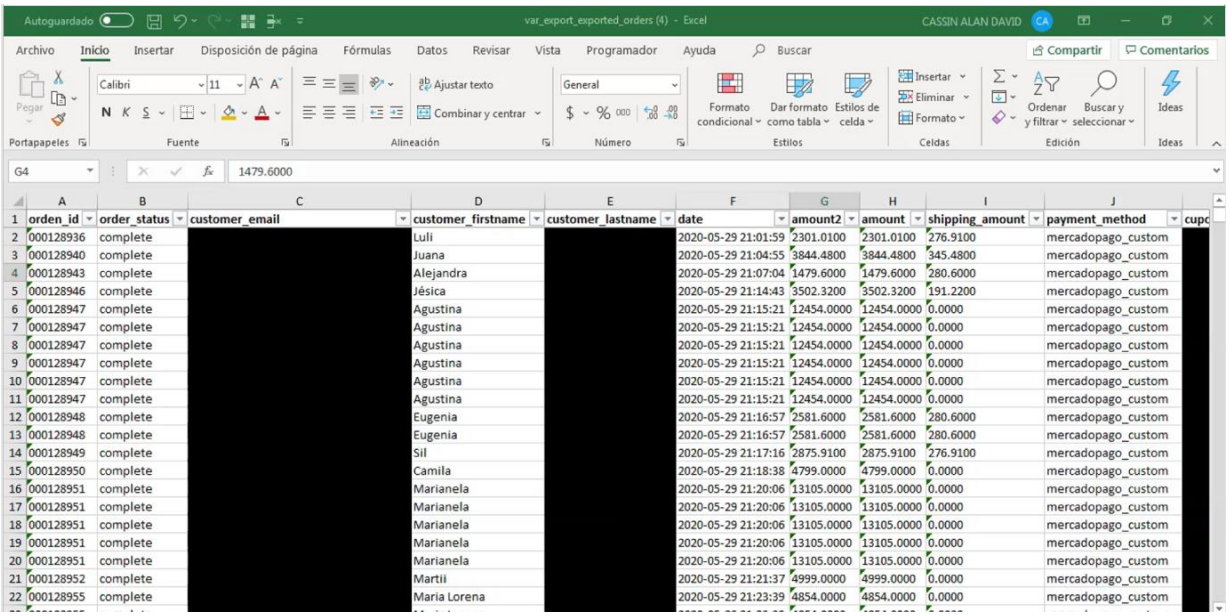
1. Entendimiento del problema: definición del problema y entendimiento de los datos.
2. Preparación de los datos.
3. Modelado: entrenamiento y selección del modelo.
4. Operacionalización: despliegue del modelo, monitoreo de predicciones e integración.

A continuación, se mostrarán ciertas capturas de pantalla realizadas en este proceso y en el apéndice se adjuntará un hipervínculo que contiene el video en el que se explica el paso a paso perpetrado:

[11] Es el conjunto de procesos requeridos para ofrecer una solución informática que nos permita analizar cómo está funcionando nuestra empresa. Este conocimiento hará que optimicemos dicho funcionamiento mediante la toma de decisiones pertinentes.

Figura 8

Limpieza y preprocesamiento de los datos



	A	B	C	D	E	F	G	H	I	J
1	orden_id	order_status	customer_email	customer_firstname	customer_lastname	date	amount2	amount	shipping_amount	payment_method
2	000128936	complete	[redacted]	Luli	[redacted]	2020-05-29 21:01:59	2301.0100	2301.0100	276.9100	mercadopago_custom
3	000128940	complete		Juana		2020-05-29 21:04:55	3844.4800	3844.4800	345.4800	mercadopago_custom
4	000128943	complete		Alejandra		2020-05-29 21:07:04	1479.6000	1479.6000	280.6000	mercadopago_custom
5	000128946	complete		Jésica		2020-05-29 21:14:43	3502.3200	3502.3200	191.2200	mercadopago_custom
6	000128947	complete		Agustina		2020-05-29 21:15:21	12454.0000	12454.0000	0.0000	mercadopago_custom
7	000128947	complete		Agustina		2020-05-29 21:15:21	12454.0000	12454.0000	0.0000	mercadopago_custom
8	000128947	complete		Agustina		2020-05-29 21:15:21	12454.0000	12454.0000	0.0000	mercadopago_custom
9	000128947	complete		Agustina		2020-05-29 21:15:21	12454.0000	12454.0000	0.0000	mercadopago_custom
10	000128947	complete		Agustina		2020-05-29 21:15:21	12454.0000	12454.0000	0.0000	mercadopago_custom
11	000128947	complete		Agustina		2020-05-29 21:15:21	12454.0000	12454.0000	0.0000	mercadopago_custom
12	000128948	complete		Eugenia		2020-05-29 21:16:57	2581.6000	2581.6000	280.6000	mercadopago_custom
13	000128948	complete		Eugenia		2020-05-29 21:16:57	2581.6000	2581.6000	280.6000	mercadopago_custom
14	000128949	complete		Sil		2020-05-29 21:17:16	2875.9100	2875.9100	276.9100	mercadopago_custom
15	000128950	complete		Camila		2020-05-29 21:18:38	4799.0000	4799.0000	0.0000	mercadopago_custom
16	000128951	complete		Marianela		2020-05-29 21:20:06	13105.0000	13105.0000	0.0000	mercadopago_custom
17	000128951	complete		Marianela		2020-05-29 21:20:06	13105.0000	13105.0000	0.0000	mercadopago_custom
18	000128951	complete		Marianela		2020-05-29 21:20:06	13105.0000	13105.0000	0.0000	mercadopago_custom
19	000128951	complete		Marianela		2020-05-29 21:20:06	13105.0000	13105.0000	0.0000	mercadopago_custom
20	000128951	complete		Marianela		2020-05-29 21:20:06	13105.0000	13105.0000	0.0000	mercadopago_custom
21	000128952	complete		Martii		2020-05-29 21:21:37	4999.0000	4999.0000	0.0000	mercadopago_custom
22	000128955	complete		Maria Lorena		2020-05-29 21:23:39	4854.0000	4854.0000	0.0000	mercadopago_custom

Figura 8: captura de pantalla de la hoja de cálculo utilizada para el preprocesamiento de los datos. Fuente: *elaboración propia.*

Figura 9

Análisis exploratorio de los datos

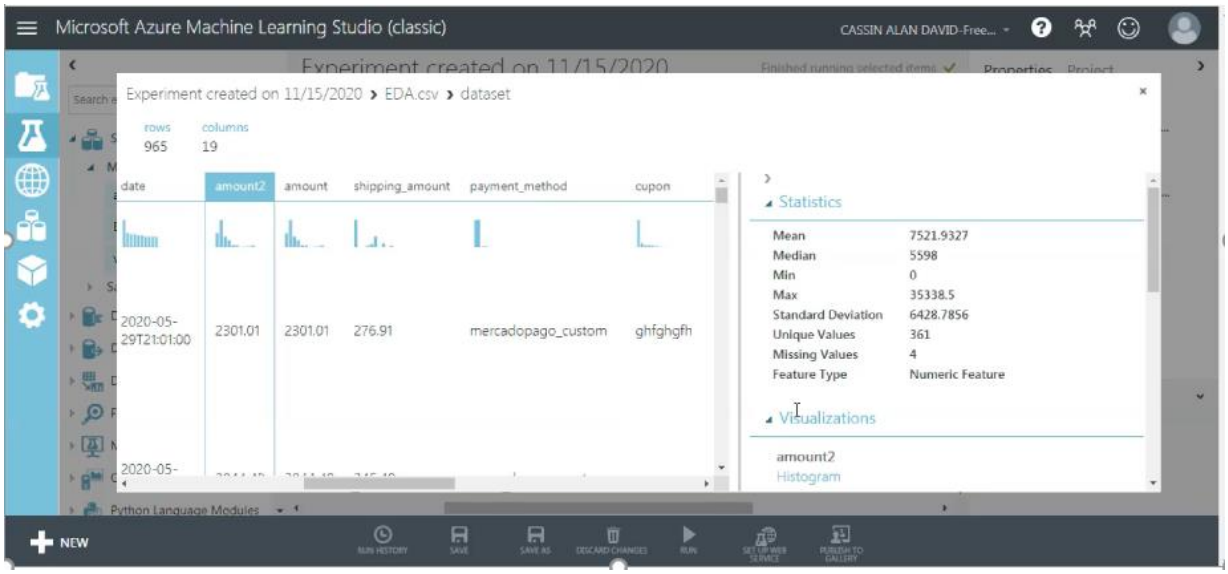


Figura 9: captura de pantalla de la herramienta utilizada para realizar el entendimiento de los datos. Fuente: *(Microsoft, 2020).*

Figura 10

Modelado del sistema de recomendación

Figura 10: captura de pantalla de la herramienta utilizada para desarrollar el modelo de aprendizaje automático. Fuente: *(Microsoft, 2020).*

3.2. Visualización de Reportes

En esta sección, se mostrarán algunos reportes que fueron creados con la herramienta Power BI utilizando las bases de datos provistas para el desarrollo del sistema de recomendación.

Este tipo de visualizaciones resultan muy amigables y pueden mostrar información en tiempo real, generando una herramienta adicional para los empleados de la organización a la hora de tomar decisiones.

Seguidamente, se expondrán algunos de estos reportes:

Figura 11

Compras por Rubro y Ubicación

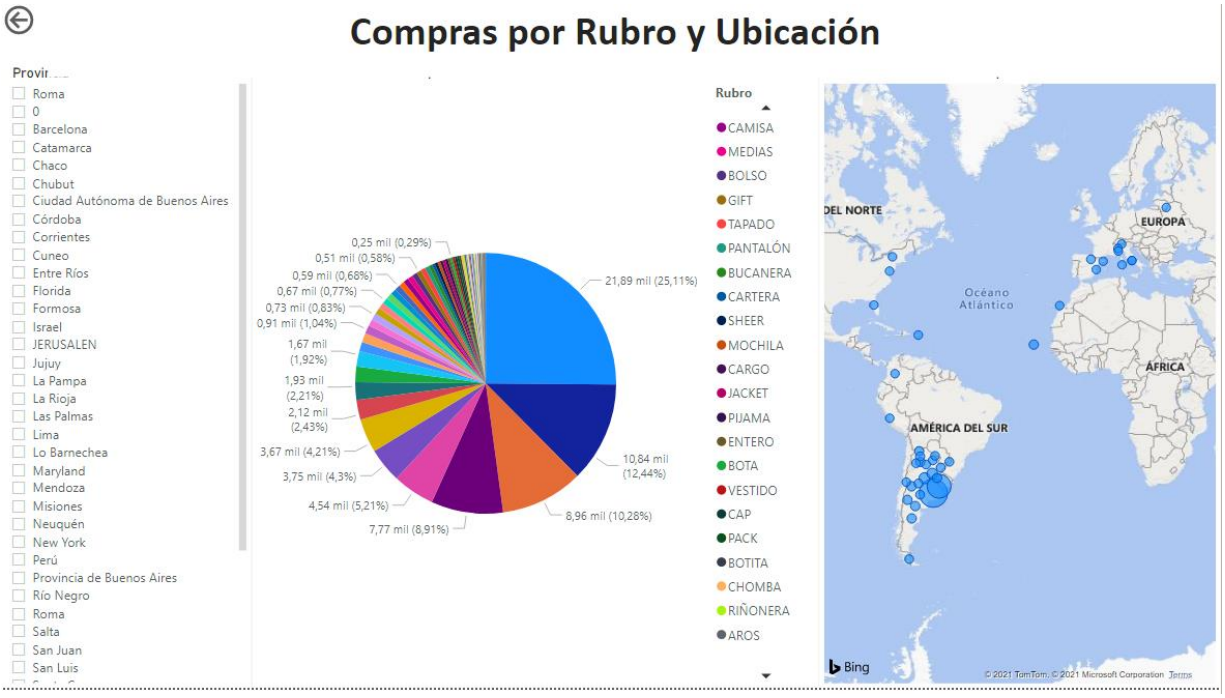


Figura 11: informe que muestra las ventas realizadas por rubro y locación. Fuente: *(Microsoft, 2021).*

Figura 12

Ventas de artículos y su rentabilidad

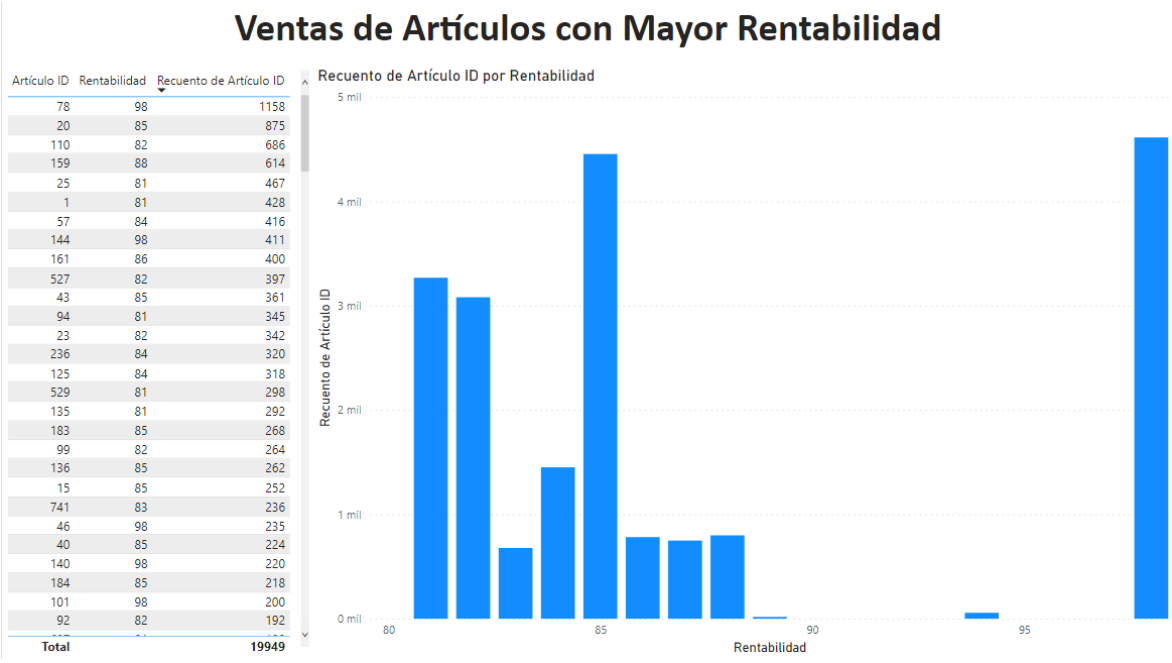
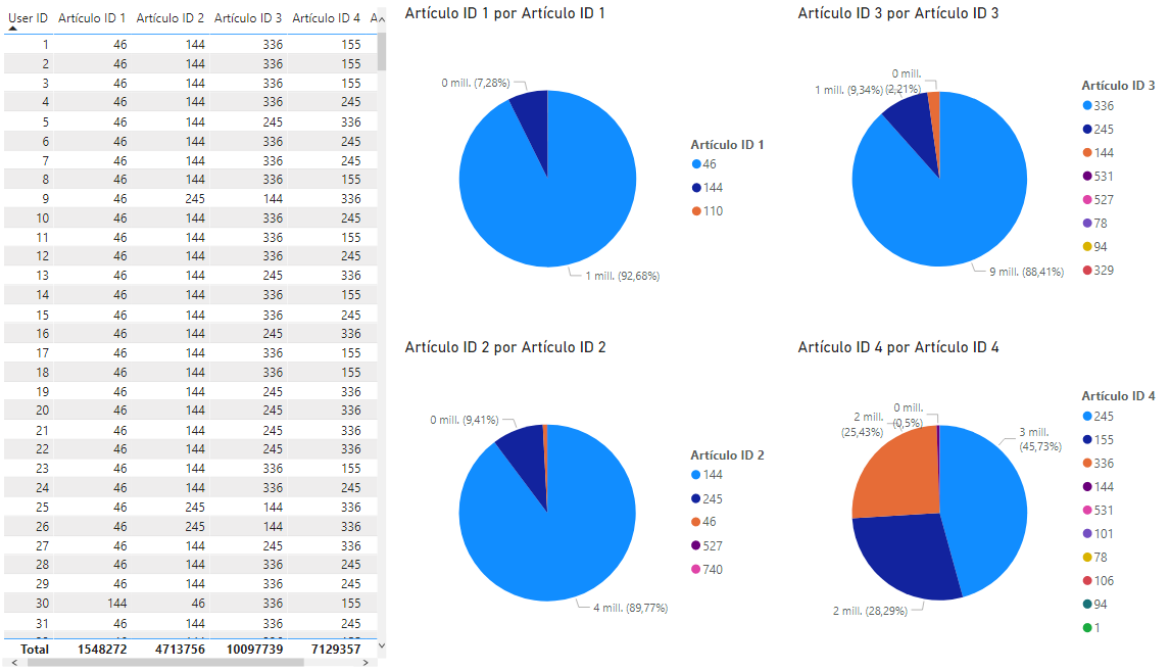


Figura 12: informe que muestra los artículos más vendidos y su rentabilidad. Fuente: *(Microsoft, 2021).*

Figura 13

Recomendaciones realizadas a los usuarios en base a la rentabilidad de los artículos



User ID	Artículo ID 1	Artículo ID 2	Artículo ID 3	Artículo ID 4
1	46	144	336	155
2	46	144	336	155
3	46	144	336	155
4	46	144	336	245
5	46	144	245	336
6	46	144	336	245
7	46	144	336	245
8	46	144	336	155
9	46	245	144	336
10	46	144	336	245
11	46	144	336	155
12	46	144	336	245
13	46	144	245	336
14	46	144	336	155
15	46	144	336	245
16	46	144	245	336
17	46	144	336	155
18	46	144	336	155
19	46	144	245	336
20	46	144	245	336
21	46	144	245	336
22	46	144	245	336
23	46	144	336	155
24	46	144	336	245
25	46	245	144	336
26	46	245	144	336
27	46	144	245	336
28	46	144	336	245
29	46	144	336	245
30	144	46	336	155
31	46	144	336	245
Total	1548272	4713756	10097739	7129357

Figura 13: informe que muestra los artículos recomendados. Fuente: *(Microsoft, 2021).*

3.3. Metodologías Ágiles

La gestión de este proyecto se enfocará en desarrollar e implementar un sistema de recomendación de filtrado colaborativo el cual utiliza un modelo de aprendizaje automático supervisado. Este sistema sería aplicado en su portal de *e-commerce*. El propósito de la creación de este sistema de recomendación es incrementar las ventas y proveer al usuario de una mejor experiencia en su compra al mostrarle los artículos que sean de su interés.

La metodología para llevar adelante este proyecto será *Lean*. Ésta resulta ser la más apropiada para aplicar en este proyecto y organización ya que es simple, y comparada a otras metodologías, resulta ser de bajo coste. También, por la reutilización y la iteración del modelo que irá evolucionando, resulta útil la adopción de esta metodología (Méndez Johegyi, 2020).

Primeramente, se preparará un set de requerimientos que actuarán como guía de problemas a resolver (comenzando por la extracción de los datos necesarios para desarrollar

el modelo y finalizando con la integración a los sistemas existentes). Luego, se construye el modelo y se testea de manera iterativa hasta que sea lo suficientemente bueno.

Como se mencionó anteriormente, la organización no posee un departamento propio de *IT*, y tanto la cultura de la empresa como la mentalidad de los que toman las decisiones, dificulta el desarrollo de nuevas tecnologías que podrían ser de gran importancia para lograr los objetivos propuestos por el directorio. Es por eso, que "verán con mejores ojos" la utilización de una metodología más barata como comienzo en la incursión de inversión en tecnología e innovación, y más precisamente en el desarrollo de un sistema de recomendación.

El plan del proyecto consistirá y tendrá que cumplir con el ciclo de vida definido en sus diferentes fases. A continuación, se describen las primeras tres fases:

1. Planificación: se define la necesidad de mejorar la experiencia del usuario en las compras online, para que consecuentemente se incrementen las ventas online. Se considera el desarrollo de un sistema de recomendación que le muestre al cliente los principales artículos que son de su interés. Este sistema deberá alimentarse de las compras históricas realizadas por cada usuario (de los cuales se posee su dirección y otros datos personales) y del interés mostrado por los usuarios mediante sus interacciones al explorar la página web de la marca.
2. Viabilidad: averiguar si es posible obtener y extraer los datos analíticos de la exploración del usuario en la web y conectarlos con el motor del sistema de recomendación. Además, se conocerá el costo estimado del desarrollo de esta herramienta. Por último, se realizarán estimaciones de las ventas adicionales que puede generar esta herramienta y se evaluarán contra su costo.
3. Resumen: se plasma lo que se espera lograr con este sistema de recomendación. Como, por ejemplo, mayor tiempo promedio de exploración por la página web y mayores ventas.

Para poder llevar adelante este proyecto, se contratará un proveedor que posea experiencia. El equipo del proveedor externo deberá estar compuesto por un *scrum master*[12] y un programador. Adicionalmente, se contratará un analista de datos (interno de la

[12] Facilitador de proyectos, es la figura que lidera los equipos en la gestión ágil de proyectos. Su misión es que los equipos de trabajo alcancen sus objetivos hasta llegar a la fase de sprint final, eliminando cualquier dificultad que puedan encontrar en el camino.

organización) que pueda trabajar junto al equipo externo de *IT* para poder proveer de las bases de datos necesarias para que funcione el modelo de aprendizaje automático. Además, proveerá el código final al equipo de *IT* para que lo incorporen a la página web. Este analista será el precursor de un nuevo departamento en la compañía y cumplirá este rol de gestión, análisis y explotación de datos para futuros proyectos. Las herramientas desarrolladas darán soporte al directorio en la toma de decisiones.

Con respecto al equipo del proveedor externo, el programador se encargará de desarrollar el código del sistema de recomendación, (en base a las bases de datos que el analista de datos le proveerá), realizará el testeo y por último la producción de este. El *scrum master* se asegurará que el programador cuente con todo lo necesario para poder llevar adelante su trabajo. También coordinará con el analista de datos y el equipo externo de *IT* reuniones diarias para alinear los requerimientos y las expectativas. Estas reuniones servirán para solucionar cualquier inconveniente que pueda surgir.

El éxito de la herramienta desarrollada se medirá en base a dos métricas:

1. Clics en los productos recomendados. Se creará un indicador que muestre los usuarios que dieron clic en alguno de los productos recomendados durante su visita en la página web sobre la cantidad total de usuarios que entraron a la página web.
2. Compras de los productos recomendados. Se creará un indicador que muestre si la compra del usuario fue alguno de los productos que se le recomendó (positivos sobre el total usuarios).

En cuanto al éxito del proyecto, el mismo se medirá en base al cumplimiento de los plazos preestablecidos y a las estimaciones de costos y ventas incrementales realizadas en el resumen (fase 3).

Conclusiones

En el inicio del trabajo final, se planteó el interrogante sobre la posibilidad de desarrollar un modelo de aprendizaje automático que pueda predecir el método de pago que utilizarán los clientes en las tiendas físicas al efectuar una compra. Observando los resultados obtenidos, se puede afirmar que el modelo de predicción creado que utiliza la técnica de *Gradient Boosted Trees*, predice el método de pago elegido por el cliente a la hora de realizar una compra en el comercio en estudio con un 87.18% de exactitud. También se puede aseverar que se trata de un modelo robusto ya que la precisión es alta y su volatilidad es baja. Por lo tanto, se puede aseverar que se ha cumplido el objetivo general propuesto.

Los parámetros más importantes que optimizaron el mejor modelo fueron: el uso de 800 árboles de decisión; 30 como medida máxima de profundidad; el uso de un mínimo de 10 instancias; y la utilización de 20 discretizaciones. El detalle de los demás parámetros se expone en la sección b del Anexo.

En cuanto a los demás modelos creados, se comprobó que la técnica que utiliza *KNN* como algoritmo arroja mejores resultados que el de Redes Neuronales. Estos superan el *baseline* del 71.30% por 6.62 y 5.08 puntos porcentuales respectivamente. Aun así, estos resultados se encuentran lejanos de los obtenidos con la técnica de *Gradient Boosted Trees* ya que superan el *baseline* por 15.88 puntos porcentuales.

La creación de este modelo significa una ventaja que tendrá el directorio de la compañía a la hora de tomar decisiones. Algunos ejemplos de las ventajas que acarrea esta herramienta son: estimar de manera más precisa el flujo de caja de la empresa y así lograr una mejor posición financiera, negociar mejores condiciones con las marcas asociadas a las tarjetas de crédito o incentivar una forma de pago específica.

Además, el modelo nos permite analizar y concluir que la categoría de atributos más importante para determinar el método de pago son las variables asociadas con el precio que pagará el cliente. Aunque cabe resaltar que esta categoría no es la única significativa y la exactitud del modelo se construye en base a todas. Por lo tanto, también se pueden realizar diferentes análisis con respecto a otras categorías que permitan responder ciertas consultas como, por ejemplo: ¿qué vendedores son los que generan más ventas en efectivo? o ¿en qué fecha del mes o en qué día de la semana existen más pagos realizados con otros métodos de

pago? Las respuestas a estas preguntas le brindarán nuevas herramientas al directorio de la compañía para tomar mejores decisiones e incentivar el uso del método de pago que a ellos les sea más conveniente en cada ocasión. Estos interrogantes podrán ser respondidos en futuros trabajos.

Finalmente, con el propósito de generar datos de mayor calidad que permitan mejorar el modelo creado y realizar otro tipo de análisis u otras predicciones sobre otras variables, se recomendaría la inclusión de atributos demográficos y más descriptivos sobre los clientes como pueden ser la edad y el domicilio. De esta manera, se podrían segmentar los clientes en diferentes *clusters* o grupos para identificar ciertos patrones en cada uno de estos conjuntos.

El desarrollo de este modelo permite comprobar empíricamente que el *Big Data* puede aplicarse a una PyME y que su correcta utilización puede ayudar a mitigar ciertas problemáticas de la organización. Tanto la generación como la explotación de los datos resulta muy importante para cualquier organización ya que es una herramienta que le brindará a los directivos información estadística para que puedan tomar mejores decisiones, y como resultado, obtener mejores rendimientos.

Por lo expuesto anteriormente, la metodología ágil óptima para llevar adelante este tipo de proyectos en esta organización resulta ser *lean*. Como consecuencia, se recomienda que esta primer experiencia y trabajo final sirva como el puntapié inicial para futuras aplicaciones que la compañía le pueda brindar a los datos que genera.

Referencias Bibliográficas

Akred, J. (6 de Septiembre de 2014). *Big Data Made Simple (BDMS)*. Obtenido de Big Data Made Simple: https://bigdata-madesimple.com/what-is-big-data-definitions-from-40-thought-leaders/

Banco de La Nación Argentina. (Septiembre de 2020). *Banco Nación*. Obtenido de https://www.bna.com.ar/Personas

Bosco Mendoza Vega, J. (Septiembre de 2020). *Medium.* Obtenido de https://medium.com/@jboscomendoza/xgboost-en-r-398e7c84998e

Ceurvels, M. (14 de Diciembre de 2020). Latin America will be the fastest-growing retail ecommerce market this year. *eMarketer*. Obtenido de https://www.emarketer.com/content/latin-america-will-fastest-growing-retail-ecommerce-market-this-year

Donza, E., Poy, S., & Salvia, A. (2019). *Heterogeneidad y fragmentación del mercado de trabajo (2010-2018).* Ciudad Autónoma de Buenos Aires: Educa.

Espino Timón, C. (2017). *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso.* Barcelona: Universitat Oberta de Catalunya. Obtenido de http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59565/6/caresptimTFG0117mem%C3%B2ria.pdf

Estado Argentino. (15 de Abril de 2019). *Estado Argentino*. Obtenido de Portal del Estado Argentino: https://www.argentina.gob.ar/noticias/nuevas-categorias-para-ser-pyme-3

EY. (2015). *Becoming an analytics-driven organization to create value.* Cleveland, Ohio: EYGM. Obtenido de https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/digital/ey-global-becoming-an-analytics-driven-organization.pdf

Instituto Nacional de Estadística y Censos (INDEC). (2019). *Encuesta Nacional de Gastos de los Hogares 2017-2018.* Ciudad Autónoma de Buenos Aires.

Joyanes Aguilar, L. (2013). *Big Data: Análisis de Grandes Volúmenes de Datos en Organizaciones.* México: Alfaomega.

Kotu, V., & Deshpande, B. (2014). *Predictive Analytics and Data Mining.* Waltham: Elsevier Inc.

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety.* META Group. Obtenido de https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an...

McAfee, A., & Brynjolfsson, E. (October de 2012). Big Data: The Management Revolution. *Harvard Business Review Home*. Obtenido de https://hbr.org/2012/10/big-data-the-management-revolution

Méndez Johegyi, C. A. (2020). *DEFINICIÓN DEL MÉTODO DE IMPLEMENTACIÓN DE UN SOFTWARE ERP EN LAS PYMES FUSIONANDO LOS MÉTODOS ÁGILES Y LEAN THINKING.* Facultad de Ciencias de la Administración, Escuela de Ingeniería de Sistemas y Telemática. Cuenca: Universidad del Azuay. Obtenido de http://201.159.222.99/bitstream/datos/10320/1/15949.pdf

Microsoft. (Noviembre de 2020). *Microsoft Azure Machine Learning Studio*. Obtenido de Microsoft Azure Machine Learning Studio: https://studio.azureml.net/

Microsoft. (Febrero de 2021). *Microsoft Power BI*. Obtenido de Microsoft Power BI: https://powerbi.microsoft.com/

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

Página 12. (4 de Enero de 2021). Las ventas minoristas cayeron 21,4 por ciento en 2020. *Página 12*. Obtenido de https://www.pagina12.com.ar/315077-las-ventas-minoristas-cayeron-21-4-por-ciento-en-2020

Reinsel, D., Gantz, J., & Rydning, J. (2018). *The Digitization of the World: From Edge to Core.* Framingham, MA: IDC. Obtenido de https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Sosa Escudero, W. (2020). *Big Data.* Ciudad Autónoma de Buenos Aires: Siglo XXI Editores Argentina.

Tablada, C. J., & Torres, G. A. (2009). Redes Neuronales Artificiales. *Revista de Educación Matemática*. Obtenido de https://www.famaf.unc.edu.ar/~revm/digital24-3/redes.pdf

Tafra, V. L. (2016). *EVOLUCIÓN DE LOS MÉTODOS DE PAGO: ANÁLISIS DE COSTOS Y BENEFICIOS DE UNA TRANSICIÓN HACIA UNA SOCIEDAD SIN DINERO EN EFECTIVO Y DIAGNÓSTICO DE LA CONDICIÓN ACTUAL DE LA REPÚBLICA DE CHILE.* Departamento de Industrias. Valparaíso: Universidad Técnica Federico Santa María. Obtenido de https://repositorio.usm.cl/bitstream/handle/11673/23265/3560900232543UTFSM.pdf?sequence=1&isAllowed=y

Team, RStudio. (2020). *RStudio: Integrated Development Environment for R.* Obtenido de RStudio, PBC: http://www.rstudio.com/

Tovar, C. (2017). *INVESTIGACIÓN SOBRE LA APLICACIÓN DE BUSINESS INTELLIGENCE EN LA GESTIÓN DE LAS PYMES DE ARGENTINA.* Universidad de Palermo. Ciudad Autónoma de Buenos Aires: Palermo Business Review. Obtenido de https://www.palermo.edu/economicas/cbrs/pdf/pbr15/PBR_15_05_Tovar.pdf

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining - Practical Machine Learning Tools and Techniques.* Cambridge: Elsevier Inc.

Anexo

a. Descripción de la Base de Datos

A continuación, en la Tabla 4, se detallan los atributos existentes en la base de datos original:

Tabla 4

Atributos del dataset original

Nombre del atributo	Tipo de dato	Descripción
SUCURSAL	Texto (string)	Sucursal de venta
ART_COD_MODELO	Texto (string)	Código del Artículo
GP_COMPROBANTE	Número entero (integer)	Número de Factura
CPBTE_FECHA	Fecha (date)	Fecha de Transacción
FECHA_DIA	Texto (string)	Día de la semana
CPBTE_TURNO	Texto (string)	Mañana o Tarde
CPBTE_HORA	Número entero (integer)	Hora de la Transacción
CPBTE_HORA_MINUTOS	Texto (string)	Hora y Minutos de la Transacción
VENDEDOR_COD	Texto (string)	Código del Vendedor
EFECTIVO	Número real (double)	Cantidad del importe abonado en efectivo
VUELTOS	Número real (double)	Cantidad de vuelto
CUPONES	Número real (double)	Cantidad del importe abonado con tarjeta
CUENTA_CORRIENTE	Número real (double)	Cantidad del importe que posee el cliente adeudado o a favor (por cambio)
TOTAL_COBRANZA	Número real (double)	Valor del artículo
CANT_CUPON	Número entero (integer)	Cantidad de tarjetas utilizadas
TIPO_CPBTE_NOMBRE	Texto (string)	Tipo de factura
TARJETA_MARCA_CODIGO	Texto (string)	Código de la forma de pago
TARJETA_MARCA_BANCO	Texto (string)	Descripción de la forma de pago
TARJETA_CANT_CUOTA	Número entero (integer)	Cantidad de cuotas o código de estas

Nombre del atributo	Tipo de dato	Descripción
TEMPO_DESC	Texto (string)	Descripción de la temporada
COLOR_COD	Número entero (integer)	Código del color
COLOR_DESC	Texto (string)	Descripción del color
TALLE_COD	Texto (string)	Código del talle
ITEM_RUBRO	Texto (string)	Rubro del artículo
ITEM_SRUBRO	Texto (string)	Subrubro del artículo
TK_CANTIDAD	Número entero (integer)	Cantidad de artículos
TK_PRECIO_UNIT	Número real (double)	Precio unitario del artículo
TK_MONTO_CON_IVA	Número real (double)	Precio total (precio unitario x cantidad)
TK_DESCUENTOS	Número real (double)	Monto de descuento
TK_SENIA	Número real (double)	Monto de seña
PRDUCCION	Texto (string)	Licencia

Tabla 4. Tipo de dato y descripción de los atributos originales. Fuente: *elaboración propia.*

Los atributos de la base de datos modificada que fueron utilizados en los modelos son los siguientes:

Tabla 5

Atributos de la base de datos modificada utilizados para los modelos

Nombre del atributo	Tipo de dato	Descripción
SUCURSAL	Número entero (integer)	Código de Sucursal de venta
Art_Comprobante	Número entero (integer)	ID de la transacción
CPBTE_FECHA	Número entero (integer)	Fecha de Transacción
Cot_Dolar	Número entero (integer)	Cotización del dólar
FECHA_DIA	Número entero (integer)	Código del día de la semana
CPBTE_TURNO	Número entero (binomial)	Código de Mañana o Tarde
CPBTE_HORA	Número entero (integer)	Hora de la Transacción

Nombre del atributo	Tipo de dato	Descripción
TOTAL_COBRANZA	Número real (double)	Valor del artículo
SUCURSAL	Número entero (integer)	Código de Sucursal de venta
Art_Comprobante	Número entero (integer)	ID de la transacción
CPBTE_FECHA	Número entero (integer)	Fecha de Transacción
Cot_Dolar	Número entero (integer)	Cotización del dólar
FECHA_DIA	Número entero (integer)	Código del día de la semana
CPBTE_TURNO	Número entero (binomial)	Código de Mañana o Tarde
CPBTE_HORA	Número entero (integer)	Hora de la Transacción
VENDEDOR_COD	Número entero (integer)	Código del Vendedor
TOTAL_COBRANZA	Número real (double)	Valor del artículo
Nombre del atributo	Tipo de dato	Descripción
SUCURSAL	Número entero (integer)	Código de Sucursal de venta
Art_Comprobante	Número entero (integer)	ID de la transacción
CPBTE_FECHA	Número entero (integer)	Fecha de Transacción

Tabla 5. Tipo de dato y descripción de los atributos de la base de datos modificada. Fuente: *elaboración propia*.

b. Modelos Utilizados y Parámetros

En la siguiente sección se presentarán 3 diferentes tablas (6, 7 y 8) con la descripción de los modelos, parámetros y métricas utilizadas:

Tabla 6

Descripción de los parámetros y métricas utilizados en el modelo KNN

Descripción del Modelo	
Modelo	KNN
K	2
Weighted Vote	True
Measure Types	MixedMeasures
Mixed Measure	MixedEuclideanDistance
Main Criterion	First
Accuracy	True
AUC	True

Tabla 6. Métricas y parámetros utilizados para el modelo *KNN*. Fuente: *elaboración propia.*

Tabla 7

Descripción de los parámetros y métricas utilizados en el modelo Redes Neuronales

Descripción del Modelo	
Modelo	Redes Neuronales
Hidden Layers	3 (tamaños: 50, 100 y 50)
Training Cycles	200
Learning Rate	0.01
Momentum	0.9
Decay	False
Shuffle	True
Normalize	True
Error epsilon	1.0E-4
Use local random seed	False
Main Criterion	First
Accuracy	True

Tabla 7. Métricas y parámetros utilizados para el modelo Redes Neuronales. Fuente: e*laboración propia.*

Tabla 8

Descripción de los parámetros y métricas utilizados en el modelo Gradient Boosted Trees

Descripción del Modelo	
Modelo	Gradient Boosted Trees
Number of Trees	800
Reproducible	False
Maximal Depth	30
Min Rows	10
Min Split Improvement	1.0E-5
Number of Bins	20
Learning Rate	0.01
Sample Rate	1
Distribution	AUTO
Early Stopping	False
Main Criterion	First
Accuracy	True
AUC	True

Tabla 8. Métricas y parámetros utilizados para el modelo *Gradient Boosted Trees.* Fuente: *elaboración propia.*

Apéndice

A continuación, se muestran las capturas de pantalla de los códigos utilizados y los procedimientos realizados:

Figura 14

Código utilizado en el software RStudio para realizar la estadística descriptiva

```
library (ggplot2)
library (psych)
library(tidyverse)

ventas <- read.csv(file.choose(), header=TRUE)
ventas <- data.frame(ventas)
ventas

ventas2 <- data.frame(ventas[,2:31])
ventas2

descriptivo <- describe(ventas2)
descriptivo

a <- ventas %>%
  pivot_longer (-ï..Art_Comprobante)

names(ventas)

ggplot(data = a, aes(x = name, y = value, fill = name)) +
  geom_boxplot() +
  facet_wrap(~ name, scales = "free") +
  theme_minimal() +
  theme(strip.text.x = element_blank(),
        legend.position = "none") +
  xlab("Atributos") +
  ylab("Valores") +
  ggtitle("Boxplots")
```

Figura 14: Captura de pantalla del código utilizado para realizar el análisis descriptivo de la base de datos y gráficos. Fuente: *(Team, RStudio, 2020).*

Figura 15

Procedimiento realizado para correr los modelos en RapidMiner Studio

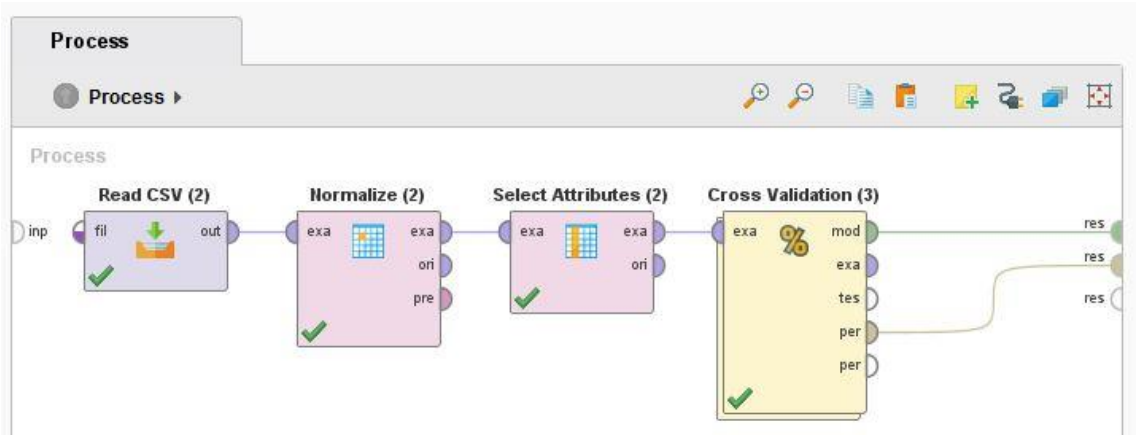


Figura 15: Captura de pantalla del procedimiento realizado para correr los modelos. Fuente: *(Mierswa, I.; Klinkenberg, R.; RapidMiner 9.7, 2020).*

Figura 16

Procedimiento realizado dentro de la Validación Cruzada de Rapid Miner Studio para el modelo KNN

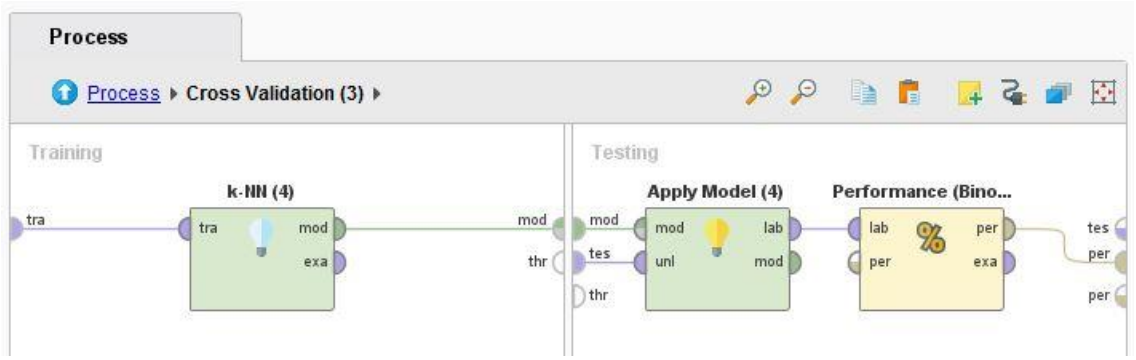


Figura 16: Captura de pantalla del procedimiento realizado para correr el modelo de *KNN*. Fuente: *(Mierswa, I.; Klinkenberg, R.; RapidMiner 9.7, 2020).*

Figura 17

Procedimiento realizado dentro de la Validación Cruzada de Rapid Miner Studio para el modelo Redes Neuronales

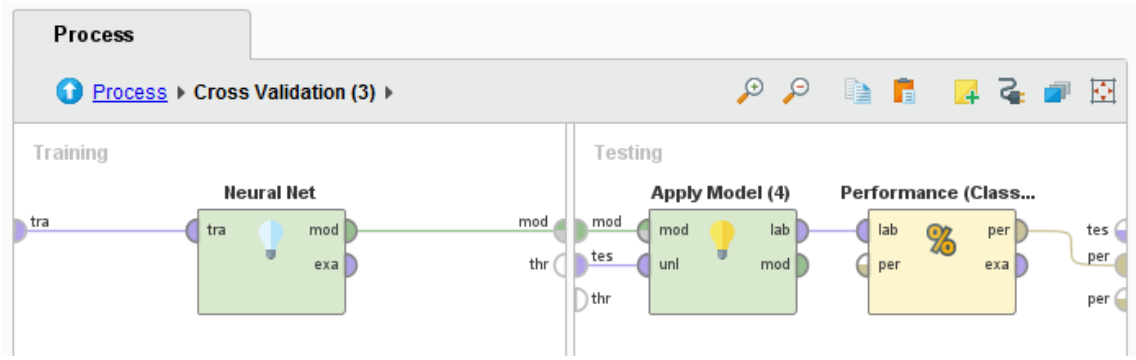


Figura 17: Captura de pantalla del procedimiento realizado para correr el modelo de Redes Neuronales. Fuente: *(Mierswa, I.; Klinkenberg, R.; RapidMiner 9.7, 2020).*

Figura 18

Procedimiento realizado dentro de la Validación Cruzada de Rapid Miner Studio para el modelo Gradient Boosted Trees

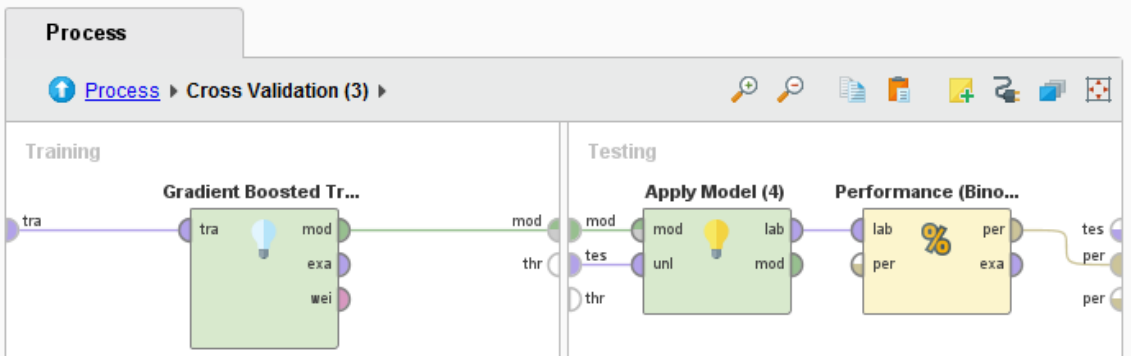


Figura 18: Captura de pantalla del procedimiento realizado para correr el modelo de *Gradient Boosted Trees*. Fuente: *(Mierswa, I.; Klinkenberg, R.; RapidMiner 9.7, 2020)*.

A continuación, se comparten los hipervínculos que contienen los videos con los diferentes pasos realizados para el desarrollo del sistema de recomendación:

- Video parte 1: https://youtu.be/rGY6XJIX37U
- Video parte 2: https://youtu.be/s1wBe7U9E_s
- Video parte 3: https://youtu.be/qUPIk54qOJE
- Video parte 4: https://youtu.be/Eon_W47KzDA