TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO SUPERIOR DE LOS RÍOS

Academia de Ingeniería en Sistemas Computacionales

Proyecto:
Análisis del uso de algoritmos de Minería de Datos y Machine Learning para Marketing Digital

Memoria de Residencia Profesional

Presenta:
Pedro Arcos Méndez

Asesor interno:
M.C.C. Edna Mariel Mil Chontal

Asesor externo:
Ing. Luis Alberto de la Cruz Díaz

Balancán, Tabasco, México. Enero del 2020

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1. Introducción

El área del marketing digital es la encargada de las estrategias volcadas en la promoción de una marca en internet usando canales, medios de comunicación y métodos que permite el análisis en tiempo real; es uno de los que se está adaptando al uso de la Inteligencia Artificial (AI), Minería de Datos (DM) y Machine Learning (ML), el marketing digital ayuda a potenciar la creatividad de los expertos en el área, maximizando sus resultados factibles en más campañas.

En el área de marketing digital, la minería de datos analiza grandes cantidades de información para conocer o predecir el comportamiento de los usuarios. Por lo que en este trabajo se explican sus principales herramientas y técnicas.

Cada día, las empresas producen una enorme cantidad de información obtenidas mediante la interacción de los clientes. Estos datos pueden ser de gran utilidad para los negocios. Sin embargo, pocos tienen el conocimiento y las herramientas necesarias para transformarlos en conclusiones relevantes. Esto se debe a que dicha información no es procesada.

La minería de datos utiliza esta información para personalizar y optimizar los productos y servicios. Se trata de un conjunto de tecnologías y técnicas computacionales para explorar las bases de datos de forma automática o semi automática. El objetivo de la minería de datos es extraer conclusiones de un conjunto de datos aplicando estructuras de análisis reutilizables. Para lograr esto, usa métodos de la inteligencia artificial, machine learning, estadísticas y sistemas de bases de datos.

El aprendizaje automático, común mente conocido como machine learning, mejora el rendimiento en las tareas comunes del marketing, como la segmentación de clientes, la generación de material publicitario de marca, la extracción y clasificación de contenido relevante, la comunicación con el cliente, la productividad y el rendimiento en general.

El éxito del marketing depende de muchos factores, necesita una investigación precisa del consumidor para construir una estrategia de marca, contenido atractivo para atraer al público, una comprensión razonable de la economía conductual y una habilidad casi mística para intuir cómo la gente recibirá el mensaje

enfrente a los de los competidores.

Por ello, en este proyecto investigación se analiza el uso de la minería de datos y machine learning en el área de marketing digital, y cómo algunas empresas están aplicando estas herramientas para optimizar sus estrategias de marketing llevándolos a ser más competitivos; de esta manera, este documento sea de utilidad para que usuarios (empresas, programadores, etc.) relacionados con el marketing digital puedan aplicar las herramientas sugeridas para dicho fin.

1.2. Descripción de problema

El uso de la minería de datos y machine learning en el marketing digital ha revolucionado la manera de realizar publicidad. En un mercado competitivo la intuición y la subjetividad no tienen lugar, por ello tienen que estar en constante actualización y manejar información en tiempo real de las tenencias y necesidades de los clientes. Con la minería de datos y machine learning las marcas crean estrategias con bases científicas para captar y fidelizar a los clientes, permitiendo realizar recomendaciones personalizadas, creación de contenido y personalización de experiencia.

Los beneficios de utilizar tecnologías de minado de datos y aprendizaje automático en marketing digital es la extracción de datos útiles en función de patrones específicos de cada cliente interpretando grandes volúmenes masivos de datos que se generan cada hora, lo que permite anticipar la demanda de un producto o servicio. Sin embargo, estos son utilizados en muchas áreas, por ejemplo: seguridad en la red, áreas médicas, etc., ayudando a tomar mejores decisiones para solucionar una necesidad utilizando algoritmos inteligentes.

Las áreas en donde se implementan estos algoritmos inteligentes aún están en su etapa inicial, no se cuenta con un análisis y estudio profundo de explotar al máximo esta tecnología.

1.3. Justificación

La minería de datos y machine learning es una herramienta que ayuda a realizar esas tareas para poder aprovechar la información que se haya almacenado y poder utilizarlo para generar nuevo material publicitario, sin embargo su uso no es algo que todas las organizaciones conozcan o realicen.

Las empresas ahora están aprovechando la minería de datos y el machine learning para mejorar todo, desde sus procesos de ventas hasta la interpretación de finanzas con fines de inversión.

El proyecto de investigación ayudará a las empresas que aún no implementan estas herramientas inteligentes, generando un panorama de información con los enfoques analizados y las estrategias que se está utilizando. De igual manera, aportara conocimientos a las personas que quieran aprender a utilizar estas herramientas, así como los expertos en tecnología de información y comunicación que aún no se han adentrado al estudio de estas, permitiendo explotar nuevas áreas de conocimiento y poder utilizarlos como nuevos campos de oportunidades en su área profesional.

1.4. Objetivos

1.4.1. Objetivo General

Analizar los algoritmos y enfoques de Minería de Datos y Machine Learning aplicados en Marketing Digital para el uso de estrategias en la micro, mediana y grandes empresas.

1.4.2. Objetivos particulares

- Realizar un estudio del estado del arte de minería de datos, machine learning y herramientas relacionadas
- Estudiar los diferentes métodos de minería de datos, machine learning y marketing digital.
- Analizar los algoritmos de minería de datos y machine learning que se utilizan en el marketing digital.
- Analizar diferentes enfoques que implementen minería de datos en sistemas de recomendación aplicado en marketing digital
- Analizar los métodos y técnicas por empresas que han implementado el uso de machine learning y minería de datos en sus estrategias de marketing
- Generar una lista de técnicas para demostrar la factibilidad del uso de minería de datos y machine learning que ayuden en las estrategias para el marketing digital

1.4.3. Alcances y limitaciones

Actualmente se genera una cantidad inmensa de datos cada minuto proveniente de distintas fuentes, conocido como big data debido a su volumen y complejidad, se utilizan ciencias de datos para su interpretación a través de algoritmos y métodos científicos. Las empresas pueden aprovechar esta información para optimizar sus estrategias de marketing digital basados en comportamiento de los usuarios de internet.

Las empresas en la actualidad están hablando mucho sobre el concepto de “organización exponencial”, entendido como aquellas organizaciones que son capaces de entender, asimilar, interiorizar y sacar el máximo provecho a las tecnologías exponenciales para crecer más rápido y ser más rentables que sus competidores. Si hablamos de “tecnologías exponenciales” nos referimos a todos aquellos campos de innovación que se ven afectados por la digitalización, sin duda alguna a las organizaciones les está causando un gran impacto por la automatización.

Los mayores beneficiados con la realización de esta investigación serán la micro, mediana y macro empresas locales y nacionales que aún no están haciendo usos de herramientas inteligentes para optimizar sus estrategias de marketing y poder generar mayores ingresos permitiéndoles la expansión de sus negocios. Ya que en esta investigación se explica de manera detallada los que se debe realizar al momento de utilizar estas herramientas, de igual manera se plantean enfoques de la aplicación de estas.

Sin embargo, la realización de esta investigación tiene sus limitantes, ya que para realizar el análisis se necesita de recolección de estados del arte, por lo que se necesita acceso a internet de manera estable. De igual manera existen recursos no gratuitos, por lo que se necesita una pequeña inversión para tener acceso a esta información y poder realizar el análisis.

La falta de administración del tiempo puede ser una limitante si no se tiene la ética profesional de realizar las actividades en tiempo y forma, por ello se debe realizar un cronograma de actividades para hacer más optimo el proceso de la investigación.

Capítulo 2

Marco Teorico

En el presente capítulo se hace una revisión del estado del arte de las temáticas relacionadas con el marketing digital: minería de datos y machine learning.

2.1. Minería de datos

El elemento fundamental de la minería de datos es tener el conocimiento de que son los datos. Los datos, elemento fundamental para la minería de datos son representaciones simbólicas de un determinado atributo o variable cualitativa o cuantitativa, es decir, es la materia principal para obtener la información. La información son los datos analizados de forma adecuada y de interés a fin, de ellos se obtiene el conocimiento. El conocimiento es la información ya procesada para poder emitir juicios que nos lleven a hechos, es decir, tener meta conocimiento que son las reglas que nos permiten obtener conocimiento.

Los datos se obtienen de bases de datos ya sean relacionales, temporales, documentales, multimedia, etc., e igual la podemos obtener de internet, en los registros e interacción por parte de los usuarios de esta. En la minería de datos existen dos tipos de modelos: 1) predictivo, que estiman valores de variables de interés a partir de otras variables y 2) descriptivo, que identifican los patrones que explican los datos creando reglas de asociación de datos.

Bajo lo anteriormente dicho, la minería de datos es la extracción no trivial de conocimiento procesable de una base de datos utilizados para la toma de decisiones haciendo uso de técnicas y algoritmos inteligentes.

Según Han Jiawei *et al.* [1] la minería de datos es el proceso de descubrir conocimiento interesante de grandes cantidades de datos almacenadas en bases de datos, data warehouses (colección de datos) u otro repositorio de información.

2.1.1. Métodos de minería de datos

El proceso de extraer conocimiento a partir de grandes volúmenes de datos ha sido reconocido por muchos investigadores como un tópico de investigación clave en los sistemas de bases de datos, y por muchas

compañías industriales como una importante área y una oportunidad para obtener mayores ganancias [2].

Para poder llevar a cabo el proceso de minería de datos se hace uso de metodologías, ésts metodologías integran el proceso que se debe llevar a cabo al momento de extraer información relevante de datasets de las compañías dedicadas al comercio y prestación de servicios.

La necesidad de estructurar, racionalizar y enriquecer el estudio de minería de datos, llevó a la comunidad de los años 90's a desarrollar una metodología de minería de datos. Como resultado de esta tarea se creó el Proceso estándar de la industria cruzada para minería de datos CRISP-DM (Cross Industry Standard Process for Data Mining), KDD (Knowledge Discovery in Databases) y SEMMA (Sample, Explore, Modify Model, Assess [3].

La metodología CRISP-DM se describe en términos de un modelo de proceso jerárquico, que consiste en conjuntos de tareas descritas en cuatro niveles de abstracción (de general a específico): fase, tarea genérica, tarea especializada e instancia de proceso [4]. La metodología KDD es un proceso centrado en el usuario, que tiene la propiedad de ser altamente interactivo, y que debe ser guiado por las decisiones que toma el usuario, o también por un agente inteligente [5].

Por su parte, SEMMA se considera una metodología general de minería de datos, sin embargo, se afirma que es "más bien una organización lógica del conjunto de herramientas funcionales" de uno de los productos de SAS Institute, para llevar a cabo las tareas centrales de procesamiento de datos. SEMMA se centra principalmente en las tareas de modelado de proyectos de minería de datos, dejando de lado los aspectos comerciales (a diferencia de, por ejemplo, CRISP-DM y su fase de comprensión comercial) [6].

Cabe mencionar que la minería de datos es solo una etapa dentro de la metodología KDD y CRISP-DM en la etapa 4 de ambas metodologías. En el presente trabajo se analiza la metodología KDD y CRISP-DM.

Metodología KDD

La metodología KDD es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente preprocesar los datos, hacer minería de datos (data mining) y presentar resultados [7].

El proceso KDD lo podemos ver en la Figura 2.1, involucra numerosos pasos interviniendo el usuario para la toma de decisiones. Esto se resume en las siguientes etapas [2].

A continuación, se explicará cada una de estas etapas [7]:

A)Etapa de selección

En la etapa de selección, una ves identificado el conocimiento y definido los objetivos del proceso, desde el punto de vista del usuario final, se crea datos objetivos seleccionando todo el conjunto de datos o una

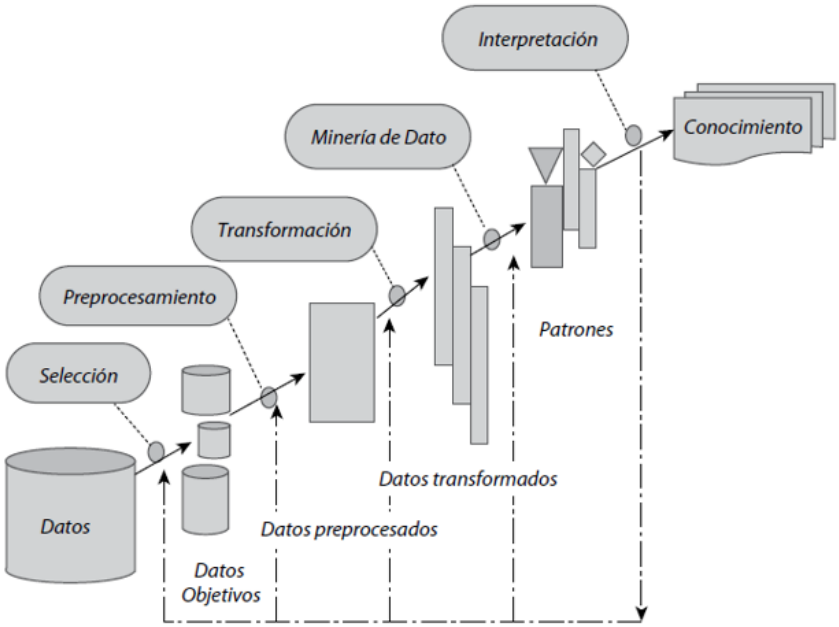

Figura 2.1: Etapas del procesamiento KDD.

muestra de esta, sobre el cual se realiza el descubrimiento de conocimiento. La selección de los datos varia de acuerdo a los objetivos de la organización.

B) Etapa de preprocesamiento/limpieza

En esta etapa se analizan la calidad de los datos, se aplican operaciones para eliminar datos atipicos y se reliza la sección de estrategias para manejar datos desconocidos, nulos, suplicados, esta etapa es de suma importanciala la interacción con el usuario final.

Los datos atípicos son valores que estan significativamente fuera de los rangos de valores esperados (ver Figura 2.2) y esto se debe principlamente por errores humanos, cambios en el sistema, información fuera de tiempo, entre otras. Los datos desconocidos son aquellos que no corresponden a un valor del mundo real. En el proceso de limpieza estos valores se ignoran y se reeemplazan por un valor temporal o por el valor más cercano, es decir, se usan medidas estadísticas como lo son la media, moda, mínimo y máximo para reemplazarlos.

C) Etapa de transformación/reducción

En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos.

D) Etapa de minería de datos

El objetivo de la etapa minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación, clustering, patrones secuenciales y

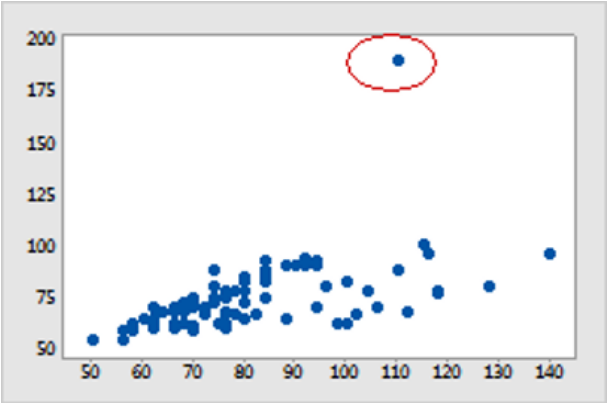

Figura 2.2: Grafica de datos para identificar datos atípicos.

asociaciones, entre otras.

Las técnicas de la minería de datos crean modelos para predecir o describir. Los modelos predictivos buscan encontrar valores futuros o desconocidos de varibales de interes, que son denominadas valores objetivos, dependientes o clases usando otras variales llamadas independiente o predictivas.

Por lo tanto, escoger un algoritmo de minería de datos incluye la selección de los métodos aplicaran en una búsqueda de patrones de datos, asi como la decisión de los modelos y parámetros más apropiados a utilizar dependiendo el tipo de datos.

E) Etapa de interpretación/evaluación de los datos

En esta etapa se interpretan los patrones que obtuvimos e igual se pueden retornar en las etapas anteriores para posteriores iteraciones. Se integrar la visualizacion de patrones extraídos, la remocion de los patrones irrelevantes y la interpretacion de los valores relevantes a modo que esta sea entendible para el usuario. De igual manera se asegura el conocimiento descubierto para anexarlo en otro sistema para siguientes acciones o solamente para documentarlo y reportarlo a las partes interesadas.

Metodología CRISP-DM

CRISP-DM, es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema [8].

El modelo de proceso actual para la minería de datos proporciona una visión general del ciclo de vida de un proyecto de minería de datos. Contiene las fases de un proyecto, sus tareas respectivas y las relaciones entre estas tareas. En este nivel de descripción, no es posible identificar todas las relaciones. Pueden existir relaciones entre cualquier tarea de minería de datos en función de los objetivos, los antecedentes y el interés del usuario, y lo más importante, en los datos.

El ciclo de vida de un proyecto de minería de datos consta de seis fases [5], que se muestran en la Figura 2.3. La secuencia de las fases no es rígida. Siempre se requiere avanzar y retroceder entre las diferentes fases. El resultado de cada fase determina qué fase, o tarea particular de una fase, debe realizarse

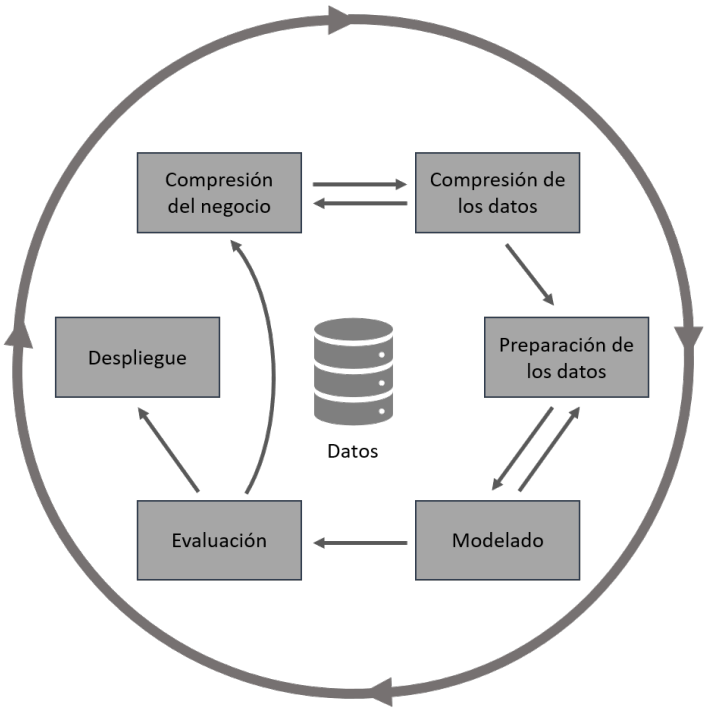

Figura 2.3: Etapas del modelo de referencia CRISP-DM.

a continuación. Las flechas indican las dependencias más importantes y frecuentes entre fases.

A continuación, se explicará cada una de estas faces [9]:

A) Comprensión del negocio

En esta fase trataremos de conseguir desde una clara perspectiva de negocio cuáles son los objetivos del mismo, tratando de evitar el gran error de dedicar el esfuerzo de todo el proyecto a proporcionar respuestas correctas a preguntas equivocadas.

Con los objetivos de negocio en mente, elaboraremos un estudio de la situación actual del negocio respecto de los objetivos planteados, en este punto, trataremos de clarificar recursos, requerimientos y limitaciones que se utilizarán para llevar a cabo la minería de datos en la etapa de modelado, contribuyendo claramente a la consecución de los objetivos primarios.

Finalmente, elaboraremos un plan de proyecto en el que detallaremos las fases, tareas y actividades que nos deberán llevar a alcanzar los objetivos planteados.

B) Comprensión de los datos

Comprensión se refiere a trabajar los datos con el objetivo de familiarizarse al máximo con ellos, saber de dónde provienen, en qué condiciones nos llegan, cuál es su estructura, qué propiedades tienen, qué inconvenientes presentan y cómo podemos mitigarlos o eliminarlos.

Se trata de una fase crítica puesto que es donde trabajamos de lleno con la calidad de los datos, que

por otro lado debemos ver como la materia prima para la minería de datos.

Tener una buena calidad de los datos será siempre una condición necesaria, aunque no suficiente para tener éxito en el proyecto.

C) Preparar los datos

El objetivo de esta fase es el de poder disponer del juego de datos final sobre el que se aplicarán los modelos. También se desarrollará la documentación descriptiva necesaria sobre el juego de datos.

Deberemos dar respuesta a la pregunta ¿qué datos son los más apropiados para alcanzar los objetivos marcados? Esto significa evaluar la relevancia de los datos, la calidad de los mismos y las limitaciones técnicas que se puedan derivar de aspectos como el volumen de datos.

Documentaremos los motivos tanto para incluir datos, como para excluir datos.

D) Modelado

El objetivo último de esta fase será el de disponer de un modelo que nos ayude a alcanzar los objetivos minería de datos y los objetivos de negocio establecidos en el proyecto.

Podemos entender el modelo como la habilidad de aplicar una técnica a un juego de datos con el objetivo de predecir una variable objetivo o encontrar un patrón desconocido.

El hecho de que esta fase entre en iteración tanto con su antecesora, la preparación de los datos, como con su sucesora, la evaluación del modelo, nos da una idea de la importancia de la misma en términos de la calidad del proyecto.

E) Evaluación del modelo

En fases anteriores nos hemos preocupado de asegurar la fiabilidad y plausibilidad del modelo, en cambio en esta fase nos centraremos en evaluar el grado de acercamiento a los objetivos de negocio y en la búsqueda, si las hay, de razones de negocio por las cuales el modelo es ineficiente.

Una forma esquemática y gráfica de visualizar el propósito de un proyecto data mining es pensar en la siguiente ecuación:

$$Resultados = Modelos + Descubrimiento$$

Es decir, el propósito de un proyecto data mining no son los modelos, que son por supuesto importantes, sino también los descubrimientos, que podríamos definir como cualquier cosa aparte del modelo que contribuye a alcanzar los objetivos de negocio o que contribuye a plantear nuevas preguntas, que a su vez son decisivas para alcanzar los objetivos de negocio.

F) Despliegue

En esta fase se organizarán y ejecutarán tanto las tareas propias del despliegue de los resultados como del

mantenimiento de las nuevas funcionalidades, una vez el despliegue haya finalizado.

Si el despliegue de los resultados del proyecto afecta a la actividad operativa de la organización, se hace imprescindible planificar y llevar a cabo tareas específicas de seguimiento y mantenimiento de las nuevas funcionalidades.

Evaluaremos las cosas que se han hecho bien y las que no se han hecho tan bien e identificaremos puntos y aspectos a mejorar.

Diferencias de la metodología KDD y CRISP-DM

Las metodologías presentadas comparten el mismo objetivo ya que están estructuradas en diversas etapas relacionadas entre sí con el objetivo del desarrollo de proyectos de minería de datos. De igual manera, cada una de están metodologías contemplan tareas específicas para en entendimiento, selección y preparación de los datos; así como la aplicación de algoritmos para descubrimiento patrones de interés, así mismo la etapa de evaluación forma parte importante en todas las metodologías, debido a la importancia de validar los resultados obtenidos, por ejemplo, en KDD la validación está en función de los objetivos del proyecto, para el caso de CRISP-DM los resultados se evalúan con base en el desempeño del modelo y el cumplimiento de los requerimientos iniciales del proyecto [10]. Además, KDD y CRISP-DM fueron diseñadas como metodologías neutras, de libre distribución, es decir, sin costo, lo que les permite adaptarse a cualquier herramienta ya sea libre o comercial.

En el Cuadro 2.1 se presenta un cuadro comparativo con las principales características de las cinco metodologías presentadas. Se incluye sus etapas, el tipo de herramientas utilizadas, el objetivo de su evaluación, el año de su creación, entre otros aspectos.

	KDD	**CRISP-DM**
Fases	- Integración y recopilación - Selección, limpieza y transformación - Minería de datos - Evaluación - Difusión y uso	- Comprensión del negocio - Comprensión del os datos - Preparación de los datos - Modelado - Evaluación - Despliegue
Etapas iterativas	Si	Si
Elección de herramientas	Libres y comerciales	Libres y comerciales
Tipo de evaluación del resultado	Basados en los objetivos del proyecto	Basado en el modelo y los objetivos del proyecto
Diseñada para minería de datos	Si	Si

Cuadro 2.1: Comparativa entre las metodologías tradicionales de minería de datos.

2.1.2. Herramientas de minería de datos

Existe una variedad de herramientas para realizar minería de datos para extraer conocimiento, en el Cuadro 2.2 se encuentra la lista de herramientas más usadas y aptas para el minado de datos.

Herramientas de minería de datos	Tipo de software	Plataforma	Algoritmos	Tipo de modelo
Weka	Libre	Todas las plataformas	Clustering y regresión	Predictivo
Clemetine	Libre	Windows, Linux	Red neuronal, logística	Predictivo
KNIME	Libre	Windows, Linux, Mac Os	Algoritmos de segmentación, árboles de decisión, redes neuronales	Predictivo
IBM SPSS	Comercial	Windows, Linux	Ecuaciones Estructurales	Predictivo
RapidMiner	Libre	Windows, Linux	Clustering, arboles de decisión, redes neuronales	Predictivo

Cuadro 2.2: Herramientas de minería de datos.

2.1.3. Aplicaciones de minería de datos

La minería de datos tiene una aplicación valiosa para las empresas, su importancia viene de la filtración y estudio de los datos internos que pueden ayudar a las empresas a plantear sus estrategias.

El DM es un campo de la estadística computacional dedicado a descubrir patrones en grandes volúmenes de conjuntos de datos. Su función general es estructurar dicha información y volverla comprensible, de modo que sirva para los intereses de las marcas o empresas [11].

Detección de fraudes

Falcon Fraud Manager es un sistema inteligente con el cual se puede examinar transacciones, propietarios de tarjetas y datos financieros. Se empleaba inicialmente para detectar y paliar el número acciones fraudulentas, las cuales hacían perder mucho dinero a las entidades financieras norteamericanas.

Su sofisticada combinación de modelos de redes neuronales, utilizada para analizar pagos mediante tarjeta y detectar los más remotos casos de fraude, permite ahorrar más de US$ 600 millones al año. Actualmente cuenta con funciones analíticas que procesan los datos de más de 500 millones de cuentas en el mundo.

Migración de clientes

Recientemente, una operadora española de telefonía móvil estudió la migración de sus clientes hacia otra operadora. Mediante minería de datos se analizó las diferencias entre los clientes que dejaron la operadora (12.6 %) y los que se mantuvieron (87.4 %). El resultado arrojó que la mayor parte de los migrantes

recibían pocas promociones y registraban un número de incidencias por encima del promedio.

A partir de dicho estudio, la operadora replanteó sus ofertas y analizó de forma exhaustiva las quejas reportadas por los clientes perdidos. Desde entonces, diseñó un trato más personalizado para los usuarios de perfil similar al de los que se fueron. El mismo método sirvió también para predecir el comportamiento de los nuevos clientes y trabajar mejor su fidelidad con la compañía.

Tamaño de las audiencias televisivas

Compañías televisivas, como la British Broadcasting Corporation (BBC) del Reino Unido, cuentan con un sistema de predicción del tamaño de las teleaudiencias, que les permite conocer la hora óptima de emisión para cada uno.

De esta manera, la BBC puede determinar qué criterios deben aplicarse para la transmisión de cada programa, ya que conoce qué series, películas y noticieros. Cada uno con sus géneros y bloques cuentan con mayor visualización a determinada hora. El sistema, por supuesto, se encuentra en constante actualización de datos.

En sector retail

Hace unos años, la compañía Walmart analizó cuáles eran los productos que se vendían con mayor frecuencia junto a los pañales. El estudio señaló que era la cerveza, descubriéndose también que ambos productos eran comprados mayormente los viernes por la tarde, por clientes varones de entre 25 y 35 años de edad.

Sucede que son los padres quienes suelen comprar los pañales, ya que éstos se venden en paquetes voluminosos que las madres procuran no cargar. Entonces, se detectó que los hombres los compraban más en ese día y a esa hora, para aprovechar en llevar las cervezas para el fin de semana.

La estrategia de Walmart fue ubicar la estantería de la bebida junto a la de los pañales, obteniendo como resultado un incremento en considerable en las ventas de ambos productos.

2.2. Machine Learning

Machine learning se originó en el campo de la inteligencia artificial, este lo forma un conjunto de métodos matemáticos y estadísticos, que su atarea involucra reconocimiento, diagnosticó, predicción, etc., existen múltiples definiciones, Hurwitz & Kirsch [12] lo definen como, una rama de AI que permite a un sistema aprender de los datos, es decir, través de programación explicita, sin embargo el aprendizaje automático no es un proceso fácil.

Se requiere técnicas de ML para mejorar los modelos predictivos. Esto depende del problema que estamos enfrentando, existen varias formas basados en el tipo y volumen de los datos. De acuerdo con la literatura, en machine learning existen categorías de aprendizaje, los cuales son: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo [13].

2.2.1. Tipos de conocimiento en machine learning

Dentro del aprendizaje automático hay dos tipos de tareas principales: supervisadas y no supervisadas, entra uno que esta fuera de ambos el cual es el aprendizaje por refuerzo. Por lo tanto, el objetivo del aprendizaje supervisado es aprender una función, dado una muestra deseada, se aproxima mejor a la relación de entrada y salida observables en los datos. El aprendizaje no supervisado, por otro lado, no tiene valores etiquetados porque su objetivo es la inferir la estructura natural presente dentro de un conjunto de datos. El aprendizaje por refuerzo su objetivo es determinar qué acciones debe escoger un agente de software en un entorno dado con el fin de maximizar alguna noción de "recompensa" o premio acumulado.

Aprendizaje supervisado

La información para construir un algoritmo contiene información sobre las características en estudio, que no está presente en los datos futuros. Por tanto, la información que se quiere predecir o por la que se quiere clasificar una población está disponible en los datos para construir el modelo. Más formalmente, el objetivo del aprendizaje supervisado es entrenar una aplicación de un conjunto de variables en una variable (factores) "x" en una variable output "y", a partir de un conjunto de datos (muestra de entrenamiento) de pares $\Delta = \{(x_i, y_i) i \in 1, ..., N\}$, donde "N" es el tamaño de la muestra [14].

Aprendizaje no supervisado

Como concepto opuesto al caso anterior, no se dispone en la muestra de construcción de la información anterior de una variable que se quiera predecir. Por tanto, en este caso no se dispone en la variable output, por lo que el conjunto de datos de la forma es de la forma $\Delta = \{x_i, i \in 1, ..., N\}$, donde "N" es el tamaño de la muestra. El objetivo de este tipo de problemas es encontrar relaciones o patrones en los datos [12].

Aprendizaje por refuerzo

Es un modelo de aprendizaje del comportamiento. El algoritmo recibe retroalimentación del análisis de los datos por lo que el usuario es guiado para el mejor resultado. El aprendizaje por refuerzo se diferencia de otros tipos de aprendizaje supervisado porque el sistema no se entrenó con el conjunto de datos de la muestra. Por el contrario, el sistema aprende por ensayo y error. Por lo tanto, una secuencia de decisiones exitosas resultará en ser el proceso de "armado", ya que mejor resuelve el problema en cuestión [13].

2.2.2. Arquitectura de machine learning

Machine learning tiene una arquitectura de proceso específico para realizar sus tareas. En la Figura 2.4 se puede ver los componentes de esta arquitectura [15].

A) Recopilación de datos

En este paso, los datos se recopilan de una variedad de fuentes de acuerdo a la declaración del problema. Este componente de la arquitectura es importante porque ML a menudo comienza con la recopilación de grandes volúmenes de datos de una variedad de fuentes potenciales. Esta parte de la arquitectura contiene los elementos necesarios para garantizar que la ingesta de datos de ML sea confiable, rápida y elástica.

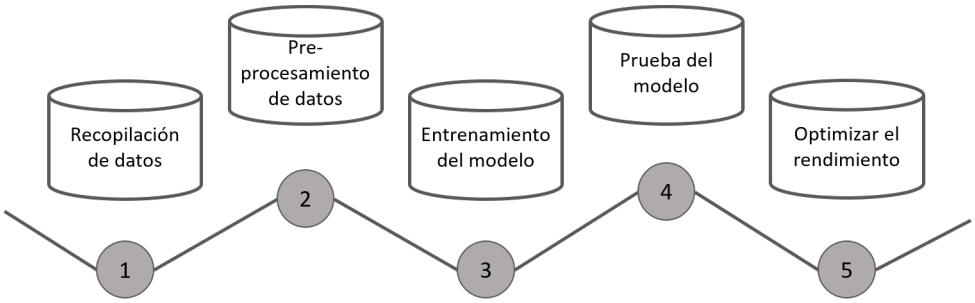

Figura 2.4: Arquitectura de proceso machine learning.

B) Pre-procesamiento de datos

En este paso, se envían los datos ingeridos para los pasos avanzados de integración y procesamiento necesarios para preparar los datos para la ejecución de ML. Esto puede incluir módulos para realizar cualquier transformación inicial de datos, normalización, limpieza y codificación que sean necesarios. Además, si se utiliza el aprendizaje supervisado, los datos deberán tener pasos de selección de muestra realizados para preparar conjuntos de datos para la capacitación.

C) Entrenamiento de datos

En este paso, se seleccionan los algoritmos y se adaptan para abordar el problema que se examinará en la fase de ejecución. Por ejemplo, si la aplicación de aprendizaje implicará análisis de conglomerados, los algoritmos de conglomeración de datos serán parte del modelo de datos ML utilizado aquí. Si se supervisa el aprendizaje a realizar, los algoritmos de entrenamiento de datos también estarán involucrados.

D) Prueba del modelo

Una vez que los datos están preparados y los algoritmos han sido modelados para resolver un problema comercial específico, se prepara el escenario para que las rutinas ML se ejecuten en la parte de ejecución de la arquitectura. La rutina de ML se ejecutará repetidamente, a medida que se realizan ciclos de experimentación, prueba y ajuste para optimizar el rendimiento de los algoritmos y refinar los resultados, en preparación para el despliegue de esos resultados para el consumo o la toma de decisiones.

E) Optimizar el rendimiento

Es importante evaluar el rendimiento del modelo y comparar diferentes algoritmos para estimar las propiedades del modelo y luego almacenada, archivada y aplicada, o puede regresar al componente de procesamiento para ser reprocesada.

En muchos casos, la salida de ML persiste en tableros que alertan al tomador de decisiones de un curso de acción recomendado. Al poner en funcionamiento los programas de ML, el sistema de aprendizaje automático se convierte en una consulta avanzada no determinista que depende de la potencia de cálculo para la ejecución.

El despliegue de la información resultante, las herramientas o la nueva funcionalidad generada por la rutina de aprendizaje automático variará según el tipo de ML que se esté utilizando y el valor que se pretende generar. Los resultados desplegados podrían tomar la forma de información reportada, nuevos

modelos para complementar las aplicaciones de análisis de datos o información para almacenar o alimentar en otros sistemas.

Para que cada etapa de la arquitectura funcione de manera adecuada es necesario tener las técnicas (algoritmos de aprendizaje). La comprensión de los diferentes algoritmos nos permite identificar los mejores, de acuerdo con la literatura los algoritmos más utilizados son: bayesiano, agrupación, Arboles de decisión, redes neuronales y aprendizaje profundo, regresión lineal, entre otras [16].

2.2.3. Herramientas de machine learning

Las herramientas de aprendizaje automático como herramienta para el análisis de datos podrían aumentar la precisión y la eficiencia de su investigación en el ámbito de la ciencia de datos sin requerir costos iniciales sustanciales de los equipos. Eso es porque las opciones existen en la nube [17].

Existe una variedad de herramientas para generar nuevo conocimiento utilizando machine learning. En el Cuadro 2.3 se encuentra la lista de herramientas más usadas y aptas para el minado de datos.

HerramientasML	**Tipos de software**	**Plataforma**
Azure Machine Learning Studio	Propietario	Windows
Amazon Machine Learning	Propietario	Windows, Linux, MacOS
Watson Machine Learning	Propietario	Windows, Linux, MacOs
Google Cloud Machine Learning Engine	Propietario	Windows, Linux, MacOs
BigML	Propietario	Windows, Linux, MacOs
Dataiku	Propietario	Windows, Linux, MacOs
Sciki-learn	Libre	Windows, Linux, MacOs
Accord.Net framework	Libre	Windows
Weka	Libre	Windows, Linux, MacOs
Deeplearn.js	Libre	Windows, Linux, MacOs
ConvNetJS	Libre	Windows, Linux, MacOs

Cuadro 2.3: Herramientas para crear aprendizaje automático.

2.2.4. Aplicaciones de machine learning

Las industrias que trabajan con grandes cantidades de datos han reconocido el valor de la tecnología del machine learning. Obteniendo datos relevantes en tiempo real, las organizaciones pueden trabajar de manera más eficiente o lograr una ventaja sobre sus competidores. A continuación se mencionan algunas áreas en donde se está aplicando el aprendizaje automático.

Servicios financieros

Los bancos y otras empresas de la industria financiera utilizan la tecnología del aprendizaje basado en máquina para dos fines principales: identificar ideas importantes en los datos y prevenir el fraude. Las ideas pueden identificar oportunidades de inversión o bien ayudar a los inversionistas a saber cuándo vender o comprar. La minería de datos también puede identificar clientes con perfiles de alto riesgo o bien utilizar cíber vigilancia para detectar signos de advertencia de fraude.

Gobierno

Dependencias de gobierno como seguridad pública y los servicios públicos tienen una necesidad particular del machine learning porque tienen múltiples fuentes de datos de las que se pueden extraer ideas. Por ejemplo, el análisis de datos de sensores identifica formas de incrementar la eficiencia y ahorrar dinero. Asimismo, el aprendizaje basado en máquina puede ayudar a detectar fraude y minimizar el robo de identidad.

Atención a la salud

El machine learning es una tendencia en rápido crecimiento en la industria de atención a la salud, gracias a la aparición de dispositivos y sensores de vestir que pueden usar datos para evaluar la salud de un paciente en tiempo real. Asimismo, la tecnología puede ayudar a expertos médicos a analizar datos para identificar tendencias o banderas rojas que puedan llevar a diagnósticos y tratamientos mejorado.

Marketing y ventas

Los sitios Web que le recomiendan artículos que podrían gustarle con base en compras anteriores, utilizan el machine learning para analizar su historial de compras y promocionar otros artículos que podrían interesarle. Esta capacidad de capturar datos, analizarlos y usarlos para personalizar una experiencia de compra (o implementar una campaña de marketing) es el futuro del comercio detallista.

Petróleo y gas

Cómo encontrar nuevas fuentes de energía. Análisis de minerales del suelo. Predicción de fallos de sensores de refinerías. Optimización de la distribución de petróleo para hacerla más eficiente y económica. El número de casos de uso del machine learning en esta industria es vasto y continúa creciendo.

Transporte

Analizar datos para identificar patrones y tendencias es clave para la industria del transporte, que se sustenta en hacer las rutas más eficientes y anticipar problemas potenciales para incrementar la rentabilidad. Los aspectos de análisis y modelado de datos del machine learning son herramientas importantes para las compañías de mensajería, transporte público y otras organizaciones de transporte.

2.3. Algoritmos de minería de datos y machine learning

Un algoritmo de minería de datos es un conjunto de heurísticas y cálculos que crea un modelo de minería de datos. Para crear un modelo, el algoritmo analiza primero los datos proporcionados, en busca de tipos específicos de patrones o tendencias. El algoritmo usa los resultados de este análisis para definir los parámetros óptimos para la creación del modelo de minería de datos [18].

Para poder descubrir conocimientos en base de datos se utilizan algoritmos inteligentes. Lo mismo sucede para el machine learning, hacen usos de algoritmos de aprendizaje.

En el Cuadro 2.4, se describen algunos de los algoritmos más populares que son utilizados en la minería de datos y machine Learning, lo cual algunos algoritmos pueden ser utilizados para ambos casos.

Algoritmos	Descripción	ML	DM
Clúster	El algoritmo de clouster o agrupamiento es una técnica utilizada en la minería de datos, consiste en dividir los datos en grupo de objetos similares, este algoritmo simplifica la información [19].	No	Si
Series de tiempo	El algoritmo de series de tiempo asume datos pasados para poder predecir datos futuros. Los datos de series de tiempo a menudo se presentan en un formato gráfico, con el intervalo de tiempo a lo largo del eje x de una tabla y los valores a lo largo del eje *y* [20].	Si	Si
Arboles de decisión	En minería de datos, un árbol de decisión describe datos, pero no las decisiones; más bien el árbol de clasificación resultante puede ser un usado como entrada para la toma de decisiones. El objetivo es crear un modelo que predice el valor de una variable de destino en función de diversas variables de entrada [21].	Si	Si
Regresión lineal	La regresión lineal es un algoritmo de aprendizaje supervisado que se utiliza en machine learning y en estadística. En su versión más sencilla, lo que haremos es «dibujar una recta» que nos indicará la tendencia de un conjunto de datos continuos [22].	Si	Si
Regresión logística	La regresión logística es una técnica de aprendizaje automático que proviene del campo de la estadística. A pesar de su nombre no es un algoritmo para aplicar en problemas de regresión, en los que se busca un valor continuo, sino que es un método para problemas de clasificación, en los que se obtienen un valor binario entre 0 y 1 [23].	Si	Si
Red neuronal	Consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. La información de entrada atraviesa la red neuronal (donde se somete a diversas operaciones) produciendo unos valores de salida [24].	Si	Si

Cuadro 2.4: Algoritmos utilizados en minería de datos y machine learning.

Capítulo 3

Minería de Datos y Machine Learning utilizados en Marketing Digital

En este capítulo se da a conocer de manera detallada las tareas que realiza cada una de las etapas de la arquitectura de minería de datos y machine learning enfocados al marketing digital, el análisis de los distintos enfoques que se han implementado para la extracción de información relevante usando minería de datos y la generación de nuevo material utilizando machine learning para optimizar la toma de decisiones y predecir datos relevantes para las estrategias de marketing digital.

3.1. Influencia del crecimiento tecnológico en el marketing digital

La revolución del uso del internet en las personas ha cambiado las maneras de comunicarse, realizar negocios y manejar a las empresas, la globalización y el crecimiento tecnológico exige una constante evolución. Así mismo el comercio electrónico a nivel internacional está evolucionando la manera de realizar sus estrategias de marketing llevándolos a utilizar tecnologías de la información y comunicación (TIC) para adaptarse a los constantes cambios tecnológicos. Sin embargo, muy pocas empresas están utilizando estas herramientas inteligentes, ya que desconocen cómo aplicarlas o incluso no tienen conocimiento de su existencia.

El uso de estas herramientas está siendo fundamental para desarrollar estrategias de marketing en los medios sociales, así como lo es la inversión en estas tecnologías para prosperar [25].

Los consumidores se están volviendo más exigentes por la aparición de las tecnologías digitales, ya que con esta poseen más información de los productos y servicios que se ofrecen por internet [26]. El cliente busca previamente información en la web, muchas veces antes de comprar, y puede utilizar las plataformas online más de una vez antes y después de tomar su decisión de compra [27].

3.2. Aplicación de la Minería de Datos en Marketing Digital

La implementación de minería de datos para la extracción de información es de gran relevancia, ya que es utilizada como herramientas para la toma de decisiones por empresas dedicadas al comercio elec-

trónico, permitiéndoles obtener datos precisos a través de datos futuros utilizando técnicas y herramientas inteligentes [28].

Actualmente la implementación de estas herramientas inteligentes aún está en desarrollo, de acuerdo a la literatura existe mucho conocimiento teórico, el número de empresas que implementan técnicas de minería de datos aplicadas a las ventas y el marketing continúa en aumento [29]. Algunas de las herramientas que nos facilita el manejo de minería de datos en marketing son [30]:

- **Segmentación de mercado:** facilitan la identificación de las características de un cliente al momento de comprar un producto.
- **Detección de riesgo de pérdidas de clientes:** con los datos históricos se predicen clientes están propenso a dejar la compañía para ir con un competidor.
- **Marketing:** identifica la perspectiva que debe contener una lista de correo electrónico para poder obtener una elevada probabilidad de respuesta.
- **Marketing interactivo:** realiza la comprensión de lo que el cliente le interesa, prediciendo material publicitario cuando éste navega en la web.
- **Análisis de la encuesta de la compra:** extraen información necesaria para extraer que productos y servicios se compran habitualmente juntos.
- **Análisis de tendencia:** revelan los diferentes hábitos del cliente de acuerdo a la temporada.

3.2.1. Principales etapas de la metodología CRISP-DM

La implementación de la minería de datos para la extracción de información cuenta con metodologías que ayudan para realizar un proceso ágil, sin embargo, no todas las que existen son aptas para aplicarlas en marketing digital, Una de las metodologías más utilizadas y enfocadas al negocio es CRISP-DM [8], , por lo que en este proyecto se pretende analizar esta metodología.

La metodología CRISP-DM este compuesto por seis etapas para llevar a cabo el proyecto de minería de datos, en la Figura 2.3 se muestra cada una de ellas. A continuación, se describe detalladamente de acuerdo a la literatura los procesos y tareas que se llevan a cabo en cada fase de la metodología [4].

Comprensión del negocio

A) Determinar objetivos comerciales

Principalmente el analista de datos debe realizar la compresión a fondo con una perspectiva de negocio lo que realmente se quiere lograr, a menudo los expertos de marketing tienen muchas ideas competitivas la cual tienen que equilibrarse. El objetivo del analista es descubrir factores importantes, una posible consecuencia es gastar tiempo y esfuerzo para brindar respuestas correctas a las preguntas incorrectas.

Organización

- Realizar diagramas para la identificación de las divisiones, departamentos y grupos del proyecto. El diagrama debe identificar las tareas de y responsabilidades de los gerentes
- Identificar las personas de importancia y sus roles que apoyaran en el proyecto
- Identificación del patrocinador (patrocinador financiero y experto en el dominio)
- Identificar al comité directivo y enumerar los miembros
- Identificar las unidades de negocio que se ven beneficiadas al implementar el proyecto de minería de datos

Área problematica

- Identificar el problema (por ejemplo, marketing, atención al cliente, desarrollo comercial, etc.)
- Describir el problema en general
- Verificar el estado actual del proyecto (verificar si está claro dentro de la unidad de negocios en el que se realizara un proyecto de minería de datos)
- Aclarar los requisitos previos del proyecto (por ejemplo, ¿Cuál es la motivación de realizar el proyecto?, ¿la empresa ya utiliza la minería de datos?, etc.)
- Si se necesita, se deben preparar presentaciones para explicar cómo se implementará y cuáles son los beneficios que tendrá al implementar minería de datos en la empresa
- Identificar los objetivos para el resultado del proyecto (por ejemplo, ¿se espera que se entregue un informe técnico de alta dirección para los usuarios finales?)
- Identificar las necesidades y expectativas de los usuarios finales.

Solución actual

- Describir cualquier solución actual para resolver el problema
- Describir las ventajas y desventajas de la situación actual y el nivel de aceptación que los usuarios tienen

B) Evaluar la situación

Esta tarea implica una búsqueda de hechos más detallada sobre los recursos, restricciones, suposiciones y otros factores que se deben tomar en cuenta al determinar los objetivos del análisis de datos y desarrollar el plan del proyecto.

Recursos de hardware

- Identificar el hardware base
- Establecer la disponibilidad de hardware para el proyecto de minería de datos

- Comprobar si el programa de mantenimiento de hardware entra en problema con la disponibilidad del hardware para el proyecto de minería de datos
- Identificar el hardware disponible para la herramienta de minería de datos que se utilizará

Fuente de datos y conocimiento

- Identificar la fuente de datos
- Identificar el tipo de fuente de datos
- Identificar fuentes de conocimiento
- Identificar el tipo de fuente de conocimiento
- Verificar las herramientas y técnicas disponibles
- Describir a fondo el conocimiento relevante

C) Determinar objetivos de minería de datos

Un objetivo empresarial establece objetivos en términos empresariales, de igual manera, un objetivo de minería de datos establece los objetivos del proyecto en términos técnicos. Por ejemplo, el objetivo de un negocio podría ser un aumento en sus ventas y clientes, mientras que el objetivo de la minería de datos podría ser, "cuantos artículos comprara un cliente, dado a los datos de compras en los últimos dos años, información demográfica relevante y el precio del artículo".

Actividades

- Interpretar las preguntas empresariales a los objetivos de minería de datos (por ejemplo, al realizar una campaña de marketing se requiere la segmentación de los clientes para poder decidir a quién va dirigida la campaña, y el tamaño de los segmentos debe ser especifico)
- Especificar los tipos de problemas de minería de datos (por ejemplo, clasificación, descripción, predicción y clustering)

Criterios de éxito

Definir los criterios en términos técnicos para obtener resultados de éxito en el proyecto, pueden ser de cierto nivel de predicción o de un grado determinado de la necesidad del cliente).

- Especificar criterios de evaluación para los modelos
- Definir prueba de rendimiento o comparativa para los criterios de evaluación
- Especificar los criterios que se abordaran (los criterios de evaluación subjetiva)

D) Producir el plan del proyecto

Describir el plan previsto para poder alcanzar los objetivos de minería de datos previstos y así poder lograr los objetivos del negocio. Enumerar las etapas, duración, recursos requeridos, entradas, salidas y dependencias que se ejecutaran en el proyecto.

Actividades

- Definir el proceso inicial del plan y discutir la viabilidad con el personal involucrado
- Combinar los objetivos identificados y las técnicas seleccionadas en un procedimiento coherente que resuelva las preguntas y los criterios de éxito del negocio
- Estimar el esfuerzo y recurso necesario para lograr e implementar la solución
- Identificar los pasos críticos
- Marcar puntos indecisos
- Identificar iteraciones principales

Comprensión de los datos

A) Producir el plan del proyecto

Acceso a los datos enumerados en los recursos del proyecto. Esta colección inicial incluye la carga de datos, si es necesario para comprender los datos. Por ejemplo, si tienen la intención de utilizar una herramienta específica para comprender los datos, es necesario cargar los datos en esa herramienta.

Planificación de los requerimientos de los datos

- Planear la información necesaria
- Comprobar si toda la información necesaria está disponible

Criterio de selección

- Especificar los criterios de selección (es decir, ¿Qué atributos son necesarios para la minería de datos especificadas?, ¿Qué atributos son irrelevantes?, ¿cuántos atributos se pueden manejar con la técnica elegida?)
- Seleccionar los archivos de interés
- Seleccionar los datos dentro de un archivo
- Pensar en cuánto tiempo se debería utilizar un historial (por ejemplo, si existe 24 meses de datos disponibles, solo se pueden utilizar 12 meses para el ejercicio)

B) Producir el plan del proyecto

Examinar las propiedades de los datos adquiridos y el informe de los resultados. Se deben describir los datos adquiridos, incluyendo el formato de los datos, la cantidad de los datos, las entidades de los campos y cualquier característica superficial que sea descubierto.

Análisis de volumen de datos

- Identificar los datos y métodos de captura
- Acceder al origen de los datos

- Utilizar análisis estadísticos si es conveniente
- Utilizar tablas de informes y sus relaciones
- Comprobar volumen de datos, complejidad y número de múltiplos
- Observar si los datos contienen entradas de textos gratuitos

Tipos de atributos y valores

- Comprobación de accesibilidad y disponibilidad de atributos
- Comprobar los tipos de atributos (numérico, simbólico, taxonomía, etc.)
- Comprobar el rango del valor del atributo
- Analizar las correlaciones del atributo
- Comprender el significado y valor del atributo en términos empresariales
- Para cada atributo, calcular las estadísticas básicas (por ejemplo, distribución de computación, media, máxima, mínima, varianza, etc.)
- Analizar las estadísticas básicas y relacionar los resultados con su significado en términos empresariales
- Determinar si el término del atributo se utiliza constantemente
- Entrevistar a expertos de dominio para obtener un punto de vista y opinión sobre la relevancia del atributo
- Decidir si es necesario equilibrar los datos

C) Exploración de los datos

Esta tarea aborda las preguntas de la minería de datos que se pueden abordar utilizando técnicas de consulta, visualización e informe. Estos análisis pueden abordar directamente los objetivos de la minería de datos. Sin embargo, también pueden ayudar a la descripción de los datos y los informes de calidad.

Exploración de datos

- Analizar las propiedades de los atributos de más interés detalladamente
- Identificar las características de las subpoblaciones

Exploración de datos

- Considerar y evaluar la información y resultados en el informe de descripción de datos
- Formas hipótesis e identificar las acciones
- Transformar la hipótesis con el objetivo de minería de datos
- Aclarar y hacer más preciso los objetivos de la minería de datos

- Realizar análisis básicos para verificar la hipótesis

D) Verificar calidad de los datos

Examinar la calidad de los datos utilizando preguntas como: ¿se cubren todos los casos requeridos de los datos?, ¿es viable o contiene errores? Si hay errores, ¿Qué tan comunes son?, ¿hay valores que faltan en los datos?, entre otras preguntas más que ayuden a resolver cualquier duda con la calidad.

Actividades

Enumerar los resultados de verificación de la calidad de los datos, si hay posibles problemas de calidad, listar las posibles soluciones.

- Identificar los valores especiales y catalogar su significado

Revisar claves y atributos

- Verificar la cobertura
- Verificar las claves
- Verificar los significados de los atributos y valores contenidos
- Identificar atributos faltantes en campos vacíos
- Establecer el significado de los datos faltantes
- Verificar los atributos con valores diferentes que tengan significados similares (por ejemplo, bajo contenido en grasa, dieta)
- Verificar la ortografía y el formato de los valores
- Revisar si existen desviaciones y tomar una decisión si esta desviación pueda indicar un fenómeno interesante
- Verificar la posibilidad de valores

Preparación de los datos

Esta es la descripción de los conjuntos de datos utilizados para el modelado o para el trabajo de análisis importante del proyecto.

A) Selección de los datos

Decidir los datos que se utilizaran para el análisis. La relevancia de los criterios de minería de datos, calidad y limitaciones técnicas como los límites de volumen o tipos de datos.

Actividades

- Recolectar datos adicionales apropiados
- Realizar prueba de recolección y significados para decidir si los campos deben ser incluidos

- Reconsiderar los criterios de sección de datos de acuerdo con la experiencia de calidad de datos y exploración de datos (es decir, si se pueden incluir o excluir otros conjuntos de datos)
- Reconsiderar los criterios de selección de datos de acuerdo con la experiencia del modelado (es decir, realizar una evaluación del modelo para mostrar que se necesitan otros conjuntos de datos)
- Seleccionar varios subconjuntos de datos (por ejemplo, atributos diferentes sin embargo solo serán datos que cumplan determinadas condiciones)
- Considerar el uso de técnicas de muestreo
- Documentar la justificación de inclusión o exclusión
- Consultar técnicas de muestreo de datos

B) limpieza de los datos

Aumentar la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas. Esto puede incluir la selección de subconjuntos de datos limpios, la inserción de defectos convenientes, o técnicas más ambiciosas, como la estimación de datos faltantes por el modelado.

Actividades

- Reconsiderar como tratar cualquier tipo de ruido observado
- Corregir, quitar o ignorar ruido
- Decir cómo lidiar con los valores especiales y su significado. El área de valor especiales puede dar lugar a múltiples resultados extraños y se debe examinar cuidadosamente
- Reconsiderar los criterios de selección de datos de acuerdo a la experiencia de limpieza de datos (es decir, poder incluir otros subconjuntos de datos)

C) Construcción de datos

Esta tarea incluye operaciones constructivas de preparación de datos, como la producción de atributos derivados, completos, nuevos registros, o valores transformados para atributos existentes.

Actividades

- Comprobar los mecanismos de construcción disponibles con la lista de herramientas sugeridas para el proyecto
- Decidir si es conveniente realizar la construcción dentro de la herramienta o fuera
- Considerar los criterios de selección de datos de acuerdo con la experiencia de construcción de datos

Atributos derivados

- Decidir si algún atributo debe normalizarse (por ejemplo, cuando se utiliza un algoritmo de agrupamiento)

- Considerar agregar nueva información sobre la importancia pertinente de los atributos agregando nuevos atributos
- ¿Cómo poder construir o imputar loa atributos faltantes? (decidir el tipo de construcción)
- Agregar nuevos atributos a los datos de acceso

D) Integrar datos

Esta tarea emplea métodos para combinar información de varias tablas u otras fuentes de información para crear nuevos registros o valores.

Actividades

- Comprobar si las instalaciones de integración son capaces de integrar las fuentes de entrada como sean necesarias
- Integrar fuentes y almacenar resultados
- Reconsidere los criterios de selección de datos de acuerdo a las experiencias de integración de datos (es decir, puede incluir o excluir otros conjuntos de datos)

E) Formato de datos

El formato de las transformaciones se refiere principalmente a las modificaciones sintácticas realizadas con los datos que no cambian su significado, pero que puede ser requerido por la herramienta de modelado.

Reordenar atributos

Algunas herramientas tienen requerimientos en el orden de los atributos, como el primer campo que es un identificador único para cada registro o el último campo que es el campo de resultado que el modelo es predecir.

Reformando valor interno

- Estos son cambios puramente sintácticos realizados para satisfacer los requerimientos de la herramienta de modelado específica
- Reconsiderar los criterios de selección de datos de acuerdo a las experiencias de limpieza de datos (es decir, puede incluir o excluir otros conjuntos de datos)

Modelado

A) Selección de la técnica de modelado

Como primer paso en el modelaje, se realiza la selección de la técnica de modelado inicial. Si se van a aplicar varias técnicas, se realiza esta tarea por separado para cada técnica.

Cabe mencionar que no todas las herramientas y las técnicas son aplicables a todas y cada tarea. Para ciertos problemas, sólo algunas técnicas son apropiadas. “Requisitos de política”, y otras limitaciones limitan aún más las opciones disponibles para la persona encargada de la minería de datos. Puede ser que

sólo una herramienta o técnica esté disponible para resolver el problema a mano, y que la herramienta puede no ser absolutamente la mejor, desde un punto de vista técnico.

Actividades

- Definir todas las suposiciones incorporadas hechas por la técnica sobre los datos (por ejemplo, calidad, formato, distribución)
- Comparar estos supuestos con los del informe de descripción de datos
- Asegurar de que estas suposiciones retener y volver a la fase de preparación de datos, si es necesario

B) Generar diseño de prueba

Antes de construir un modelo, es necesario definir un procedimiento para probar la calidad y validez de los modelos. Por ejemplo, en tareas supervisadas de minería de datos, como la clasificación, es común utilizar las tasas de error como medidas de calidad para los modelos de minería de datos. Por lo tanto, el diseño de prueba especifica que el conjunto de datos debe ser separado en la formación y los conjuntos de pruebas. El modelo se construye sobre el conjunto de entrenamiento y su calidad estimada en el conjunto de pruebas.

Actividades

- Revisar los diseños de pruebas existentes para cada objetivo de minería de datos
- Decidir los pasos necesarios (número de iteraciones, número de pliegues, etc.)
- Preparar los datos necesarios para las pruebas

C) Producir el plan del proyecto

Ejecutar la herramienta de modelado en el conjunto de datos preparado para crear uno o más modelos.

- Establecer parámetros iniciales
- Documentar los motivos para elegir esos valores

Modelos

Ejecutar la herramienta de modelado en el conjunto de datos preparado para crear uno o más modelos.

- Ejecutar la técnica seleccionada en el conjunto de datos de entrada para producir el modelo
- Postprocesar los resultados de la minería de datos (por ejemplo, reglas de edición, árboles de visualización)

Descripción del modelo

Describir el modelo resultante y evalúe su exactitud esperada, robustez y posibles deficiencias. Informe sobre la interpretación de los modelos y las dificultades encontradas.

D) Evaluación del modelo

El modelo debe evaluarse ahora para asegurar que cumple los criterios de éxito de la minería de datos y pasa los criterios de prueba deseados. Esta es una evaluación técnica puramente basada en el resultado de

las tareas de modelado.

Actividades

- Evaluar los resultados con respecto a los criterios de evaluación
- Probar los resultados según una estrategia de prueba (por ejemplo: validación cruzada, bootstrapping, etc.)
- Comparar resultados de evaluación e interpretación
- Crear la clasificación de los resultados con respecto a los criterios de éxito y evaluación
- Seleccionar los mejores modelos
- Interpretar los resultados en términos de negocios (en la medida de lo posible en esta etapa)
- Obtener comentarios sobre modelos por dominio o expertos de datos
- Comprobar la plausibilidad del modelo
- Comprobar el efecto sobre el objetivo de minería de datos
- Revisar el modelo de acuerdo con la base de conocimientos dada, para ver si la información descubierta es correcta
- Comprobar la confiabilidad del resultado
- Analizar el potencial de despliegue de cada resultado
- Si hay una descripción verbal del modelo generado (por ejemplo, a través de las reglas), evaluar las reglas: ¿Son lógicas, son factibles, hay demasiados o muy pocos, que ofenden el sentido común?
- Evaluar los resultados
- Obtener información sobre por qué una determinada técnica de modelado y determinados parámetros de configuración conducen a buenos o malos resultados

Evaluación

A) Evaluar los resultados

Este paso evalúa el grado en que el modelo cumple los objetivos empresariales, y busca determinar si hay alguna razón de negocio por la que este modelo es deficiente. Otra opción es probar los modelos en las aplicaciones de prueba en la aplicación real, si las limitaciones de tiempo y presupuesto permiten.

Además, la evaluación también evalúa otros resultados generados por minería de datos. Los resultados de la minería de datos cubren modelos relacionados con los objetivos empresariales originales y con todos los demás resultados. Algunos están relacionados con los objetivos empresariales originales, mientras que otros pueden revelar desafíos adicionales, información o sugerencias para futuras direcciones.

Actividades

- Comprender los resultados de la minería de datos
- Interpretación de los resultados en términos de la aplicación
- Comprobar el efecto sobre el objetivo de minería de datos
- Comprobar el resultado de la minería de datos contra la base de conocimiento dada para ver si la información es descubierta
- Evaluar los resultados con respecto a los criterios de éxito empresarial (es decir, el proyecto ha alcanzado los objetivos empresariales originales)
- Comparar resultados de evaluación e interpretación
- Comprobar los resultados con respecto a los criterios de éxito empresarial
- Comprobar el efecto del resultado en el objetivo de aplicación inicial
- Determinar si hay nuevos objetivos de negocio que se abordarán más adelante en el proyecto, o en nuevos proyectos
- Recomendaciones estatales para futuros proyectos de minería de datos

B) Revisión del proceso

En este punto, el modelo resultante parece ser satisfactorio y parece satisfacer las necesidades del negocio. Ahora es apropiado hacer una revisión más exhaustiva de la contratación de minería de datos a fin de determinar si hay algún factor o tarea importante que de alguna manera se haya pasado por alto. En esta etapa del ejercicio de la minería de datos, la Revisión del Proceso toma la forma de una Revisión de Garantía de Calidad.

Actividades

- Proporcionar una visión general del proceso de minería de datos utilizado
- Analizar el proceso de minería de datos. Para cada etapa del proceso pregunte:
    - ¿Era necesario?
    - ¿Fue ejecutado óptimamente?
    - ¿De qué maneras podría mejorarse?
- Identificar fallos
- Identificar pasos erróneos
- Identificar posibles acciones alternativas y/o rutas inesperadas en el proceso
- Revisar los resultados de los datos con respecto a los criterios de éxito empresarial

C) Determinar la siguiente etapa

En base a los resultados de la evaluación y a la revisión del proceso, el equipo del proyecto decide cómo proceder. Las decisiones que se deben hacer incluyendo si finalizar este proyecto y pasar a la implementación, iniciar iteraciones o crear nuevos proyectos de minería de datos.

Actividades

- Analizar el potencial de despliegue de cada resultado
- Estimación del potencial de mejora del proceso actual
- Comprobar los recursos restantes para determinar si permiten iteraciones adicionales de procesos (o si se pueden hacer disponibles recursos adicionales)
- Recomendaciones alternativas
- Refinar el proceso del plan

Despliegue

A) Plan de despliegue

Esta tarea comienza con los resultados de la evaluación y concluye con una estrategia para el despliegue del resultado de minera de datos en el negocio.

Actividades

- Resumir los resultados desplegables
- Desarrollar y evaluar planes alternativos para implementar
- Decidir por cada conocimiento distinto o resultado de información
- Determinar cómo se propagarán los conocimientos o la información a los usuarios
- Decidir cómo se controlará el uso del resultado y se medirán sus beneficios (cuando corresponda)
- Decidir por cada modelo desplegable o resultado del software
- Establecer cómo se desplegará el modelo o resultado del software dentro de los sistemas de organización
- Determinar cómo se controlará su uso y medir sus beneficios (cuando corresponda)
- Identificar posibles problemas durante despliegue (trampas a evitar)

B) Plan de monitoreo y mantenimiento

El monitoreo y el mantenimiento son asuntos importantes si los resultados de la minería de datos se convierten en parte del negocio cotidiano y su entorno. Una preparación cuidadosa de una estrategia de mantenimiento ayuda a evitar períodos innecesariamente largos de uso incorrecto de resultados de la minería de datos. Para supervisar el despliegue de los resultados de la minería de datos, el proyecto

necesita un plan detallado para el monitoreo y mantenimiento. Este plan tiene en cuenta el tipo específico de despliegue.

Actividades

- Comprobar si los aspectos dinámicos (es decir, ¿qué podrían cambiar las cosas en el entorno?)
- Decidir cómo se supervisará la precisión
- Determinar cuándo no se debe utilizar el resultado o modelo de la minería de datos. Identificar criterios (validez, umbral de exactitud, nuevos datos, cambio en el dominio de la aplicación, etc.), y lo que debe suceder si el modelo o resultado ya no podrían utilizarse. (modelo de actualización, configure proyecto de minería de datos, etc.)
- ¿Se modificarán los objetivos empresariales del uso del modelo con el tiempo?. Completamente documentar el problema inicial que el modelo estaba intentando resolver
- Desarrollar un plan de monitoreo y mantenimiento

C) Realizar un reporte final

Al final del proyecto, el equipo del proyecto escribe un informe final. Dependiendo del plan de despliegue, este informe puede ser sólo un resumen del proyecto y su experiencia, o una presentación final de los resultados de la minería de datos.

Actividades

- Identificar los informes necesarios (presentación de diapositivas, resumen de gestión, conclusiones detalladas, explicación de modelos, etc.)
- Analizar si se ha cumplido los objetivos iniciales de la minería de datos
- Identificar grupos de destino para informes
- Estructurar el contenido de los informes
- Seleccionar los resultados que se incluirán en los informes
- Escribir un informe

D) Revisión del proyecto

Evaluar lo que salió bien y lo que salió mal, lo que se hizo bien, y lo que hay que mejorar.

Actividades

- Entrevistar a todas las personas significativas involucradas en el proyecto y pregúnteles sobre su experiencia durante el proyecto
- Si los usuarios finales del negocio trabajan con los resultados de la minería de datos, entrevístalos: ¿Están satisfechos?, ¿Qué se podría haber hecho mejor?, ¿Necesitan soporte adicional?

- Resumir los comentarios y escriba la documentación de la experiencia
- Analizar el proceso (cosas que funcionaron bien, errores cometidos, lecciones aprendidas, etc.)
- Documentar el proceso específico de minería de datos (¿Cómo se pueden alimentar los resultados y la experiencia de la aplicación del modelo?)
- Generalizar de los detalles para que la experiencia útil para futuros proyectos

3.2.2. Minería de datos en marketing digital

Las empresas diariamente registran y acumulan una gran cantidad de información proporcionado por sus clientes potenciales, por ello se realizan grandes inversiones en tecnologías de la información utilizando los medios de comunicación para implementar estrategias de marketing con el objetivo de ganar popularidad en el mercado de competencias e incrementar la comercialización de sus productos o servicios.

La minería de datos es una de las herramientas inteligentes que las compañías e investigadores en el área de estrategias de marketing están utilizando, ya que les permite obtener datos exactos sobre tendencias y predicción de un del gusto de un cliente pasado a los datos registrados que estos tienen.

A continuación, describiremos algunos enfoques en donde aplicaron la minería de datos para la optimización de análisis de datos y predicción de tendencias.

Minería de datos para la predicción de tendencia y preferencia del cliente en el área de MODA

Contreras y Rosales [31], implementaron minería de datos para analizar el comportamiento de los clientes de una empresa dedicada a la moda, ellos implementaron técnicas de minería de datos en la cuenta de red social Instagram con la que la empresa cuenta.

Ellos optaron por utilizar la metodología CRISP-DM (ver Figura 3.1) para evaluar modelos descriptivos de técnicas de reglas de agrupación y asociación con el fin de obtener información relevante para diseños de estrategias de marketing de acuerdo a la preferencia de los usuarios.

A continuación, se explica de manera breve la estrategia que se utilizaron para la extracción de información relevante haciendo uso de la metodología mencionada anteriormente.

Comprensión del negocio: realizaron el estudio actual del negocio tomando en cuenta que la empresa se encarga de vestir figuras públicas (famosos), analizaron el estatus en el que la empresa se encuentra en el uso de redes sociales con el fin de obtener los medios para extraer información, así como las limitaciones de poder tener la interfaz publica de aplicaciones API (Application Programming Interface). Una vez identificado todos los elementos necesarios, se realizó un plan en la cual se detallaron las fases y tareas necesarias asignando actividades de esta y poder iniciar con la siguiente etapa.

Figura 3.1: Modelo CRIPS-DM análisis de gusto de clientes.

Selección de los datos: en esta fase de la metodología seleccionan la fuente de datos a utilizar, por lo que hicieron uso de la red social de Instagram para extraer el conjunto de datos de las publicaciones que han realizado entre agosto de 2014 hasta abril del 2015. Los datos fueron extraídos mediante la utilización de la API de Instagram Python-Instagram para desarrolladores.

Comprensión de los datos: el conjunto de datos obtenidos durante la extracción estaba compuesto por 1,435 registros y 8 atributos como se muestra en el Cuadro 3.1.

NOMBRE	DESCRIPCIÓN
FECHA_CREACION	Fecha de publicación
LINK	URL del archivo de Instagram
TIPO_MEDIA	Tipos de archivo (imagen o video)
TIPO_FILTRO	Filtro digital que tiene el archivo
LIKES	Número de “me gusta” de imagen o video
COMENTARIOS	Número de comentarios que tiene la imagen o video
LIKE_USUARIO	Valor booleano que describe si el usuario le dio “me gusta” a su publicación
TAGS	Número de usuarios etiquetados en la imagen o video

Cuadro 3.1: Conjunto de datos.

Se categorizaron algunos atributos para poder replicar el algoritmo de clustering, ya que este solo permite valores enteros, como se muestra en el Cuadro 3.2.

Modelado: utilizaron el modelo de clustering con los algoritmos K-Mens[1] y reglas de asociación con operadores FP-Growth[2]. Estos operadores se encuentran disponible en el software de minería de datos

[1]https://stanford.edu/c̃piech/cs221/handouts/kmeans.html
[2]https://docs.rapidminer.com/lates/studio/operators/modeling/associations/fp_growth.html

LIKES	DISCRETIZACIÓN	COMENTARIOS	DISCRETIZACIÓN
0-50	1	0-3C	1
50-100	2	3-6C	2
100-150	3	6-10C	3
150-200	4	>10	4
200-250	5		
>250	6		

TIPO_MODA	DISCRETIZACIÓN
50S	1
ACCESORIO	2
BATA	3
BLUSA	4
CAMISA HOMBRE	5
CAMISERO	6
CAMISETA HOMBRE	7
CASUAL	8
COCTEL CORTO	9
COCTEL LARGO	10
FALDAS	11
VESTIDOS DE MATRIMONIO	12

Cuadro 3.2: Categorización de atributos.

RapidMiner[3].

Evaluación: evaluaron los modelos de clustering visualizando los datos de manera gráfica del clúster para tomar los valores más relevantes y eliminar los datos atípicos. De igual manera con la regla de asociación analizaron los datos para formar ítems frecuentes.

Despliegue: según los resultados de la evaluación, llegaron a la conclusión que la categoría de tipo moda es la más preferida por los usuarios, basado en los datos de interacción que los usuarios de las redes sociales realizaban.

Par la empresa del caso de estudio, el uso de herramientas de DM los lleva a conocer la preferencia de sus clientes, de igual manera se convierte en un insumo importante para desarrollar nuevas estrategias de mercado. La información obtenida es para la empresa puesto que, puede diseñar e implementar estrategias que tengan en cuenta este tipo de preferencias y lograr la satisfacción y lealtad de los clientes.

De igual manera, se evidencio que el aprovechamiento de la información generada por las redes sociales, en conjunto con técnicas computacionales modernas como la minería de datos, permite a las empresas conocer las preferencias y el comportamiento de sus clientes, sin necesidad de realizar encuestas ni cualquier otro trabajo de campo.

[3]https://rapidminer.com/

Minería de datos para modelos predictivos de los factores de mercado en línea basados en la demanda del cliente

Truong Van Nguyen *et al.* [32] desarrollaron un enfoque comprensible de predicción de minería de datos, con el fin de lograr dos objetivos: (1) proporcionar un modelo de predicción de demanda altamente preciso y robusto de productos remanufacturados; y (2) arrojar luz sobre el efecto no lineal de los factores del mercado en línea como predictores de la demanda del cliente, basado en el conjunto de datos de Amazon del mundo real.

Optaron por utilizar el proceso de la metodología CRISP-DM (ver Figura 3.2), adaptándolo al proyecto. A continuación, se describe de manera breve los pasos que siguieron de la metodología para llevar a cabo el proyecto.

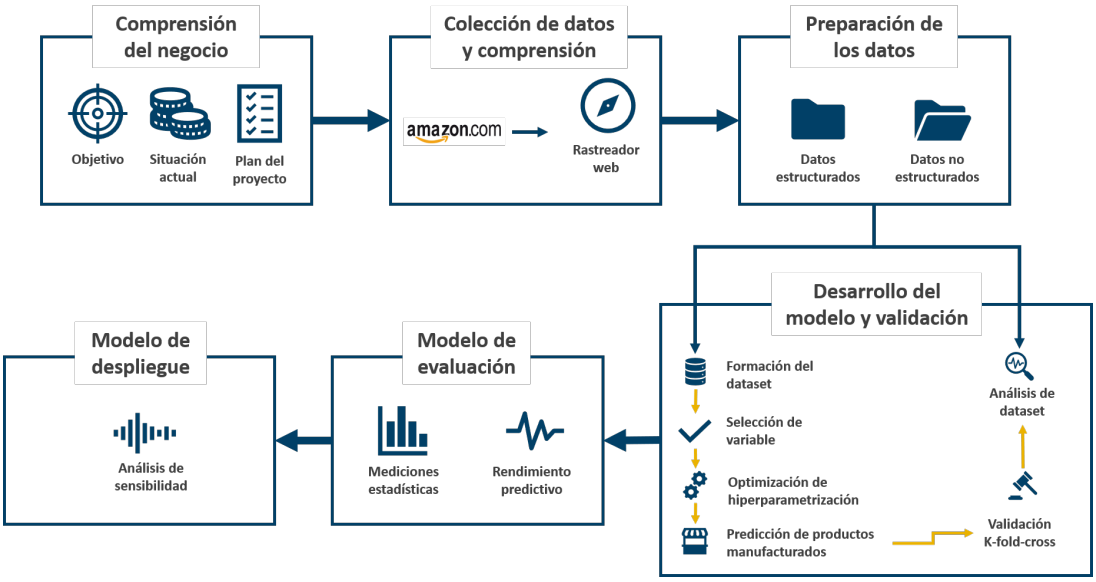

Figura 3.2: Modelo CRISP-DM basado en datos de Amazon.

Comprensión del negocio: plantearon su objetivo principal y procedieron al análisis de la situación actual en la que se encontraban para realizar un plan óptimo que cumplan las tareas en cada etapa.

Colección de datos y comprensión: la fuente de datos que utilizaron fueron los datasets de los registros de consulta que se realiza en el buscador de Amazon, ha sido elegido como fuente de datos para este estudio ya que Amazon clasifica las condiciones de los productos en tres categorías principales: nuevas, certificadas, renovadas (otro término para productos remanufacturados), y utilizadas. para lograr esto utilizaron rastreador web, permitiendo obtener el conjunto de datos de las búsquedas que se realizaban dentro de la tienda online.

Preparación de los datos: en esta etapa organizaron los datos de manera estructurada (datos numéricos) y no estructurado (datos textuales) con el fin de seleccionar las técnicas de preprocesamiento antes de analizarse. Con los datos estructurados realizaron un tratamiento externo de los datos para asegurar su consistencia e identificar valores fuera de lo ordinario. Con los datos no estructurados utilizaron un

tratamiento de transformación a datos utilizables haciendo uso de la tokenización, filtrado de palabras claves, etiquetado.

Desarrollo del modelo y validación: con los datos estructurados crearon los datasets para introducirlos el en el algoritmo, seleccionaron las variables para hacer la hiperparametrización[4] analizando de manera sensible cada modelo con respecto a sus propios parámetros de afinación y utilizando los conjuntos de entradas Boruta[5] y RFE[6], realizaron la validación cruzada con K-Fold[7] para comparar la exactitud de la predicción de varios modelos y evitar la edición del sesgo referente a muestreo de datos.

Modelo de evaluación: se calcularon tres mediciones estadísticas de uso común: medida de error absoluto (MAE), medida de error cuadrático (RMSE) y coeficientes de determinación (R2), para evaluar el rendimiento predictivo de diferentes modelos.

Modelo de despliegue: se realiza utilizando el método de análisis de la sensibilidad basada en la fusión de la información (IFSA) que ha utilizado comúnmente para el VIR en la minería de datos, con el fin de medir la importancia relativa de cada variable independiente en el modelo de predicción.

Como resultado, se revelan varias perspectivas de marketing impulsadas por datos, proporcionando directrices para ayudar a los gerentes a diseñar una estrategia efectiva de marketing específica para productos remanufacturados. Por lo tanto, puede ser beneficioso para los datos agregados de varias fuentes, los modelos de ML se consideran a menudo como cajas negras, que limita sus usos en industria.

Extracción de conocimiento de un canal de marketing de empresas

Por su parte Esra Kahya *et al.* [34] utilizaron la minería de datos para extraer el conocimiento de un canal de firma de marketing internacional para mejorar la eficiencia del sistema de marketing usado para predecir las firmas del canal de comercialización con las quejas muy altas. Optaron por el uso de la metodología CRISP-DM (ver Figura 3.3). Para analizar los datos de la encuesta que realizaron, aplicaron Mapa de Autoorganización Kohonen (SOM) para reducir los atributos.

A continuación, se describirá de manera breve cada una de las etapas de la metodología:

Comprensión del negocio: en esta etapa plantearon su objetivo el cual es predecir el canal de comercialización con quejas altas. Para llevar a cabo el proyecto de minería de datos necesitaron hacer uso de las Redes Kohonen y algoritmos como árboles de decisión lo cuales implementaran en la etapa de modelado.

Selección de los datos: en esta etapa se recolectaron 300 firmas que participan en un canal de marketing internacional con una encuesta de una sola aplicación. Las encuestas recogidas se convirtieron en una matriz de datos que incluye 20 características para cada empresa, incluyendo educación, sexo, edad, estado civil, acuerdos de pago, las preferencias de las instalaciones proporcionadas por la empresa, profesión,

[4]https://www.interactivechaus.com/manual/tutorial-de-machine-learning/hiperparametros
[5]https://www.rdocumentation.org/packages/Boruta/versions/6.0.0/topics/Boruta
[6]https://www.scikit-yb.org/en/latest/api/model_selection/rfcv.html
[7]https://www.interactivechaus.com/python/funtion/kfold

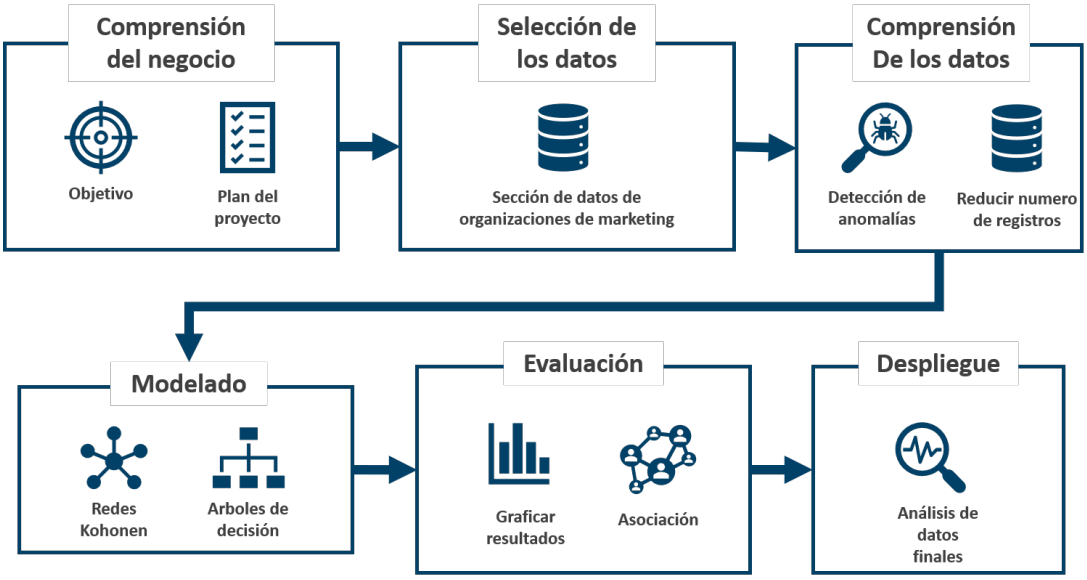

Figura 3.3: Modelo CRIPS-DM para predicción compañías con quejas.

número de marcas, etc. Se realizaron gráficas web para entender la relación entre el estado educativo y las expectativas de los propietarios de las empresas del canal de comercialización de la empresa fabricante (ver Figura 3.4).

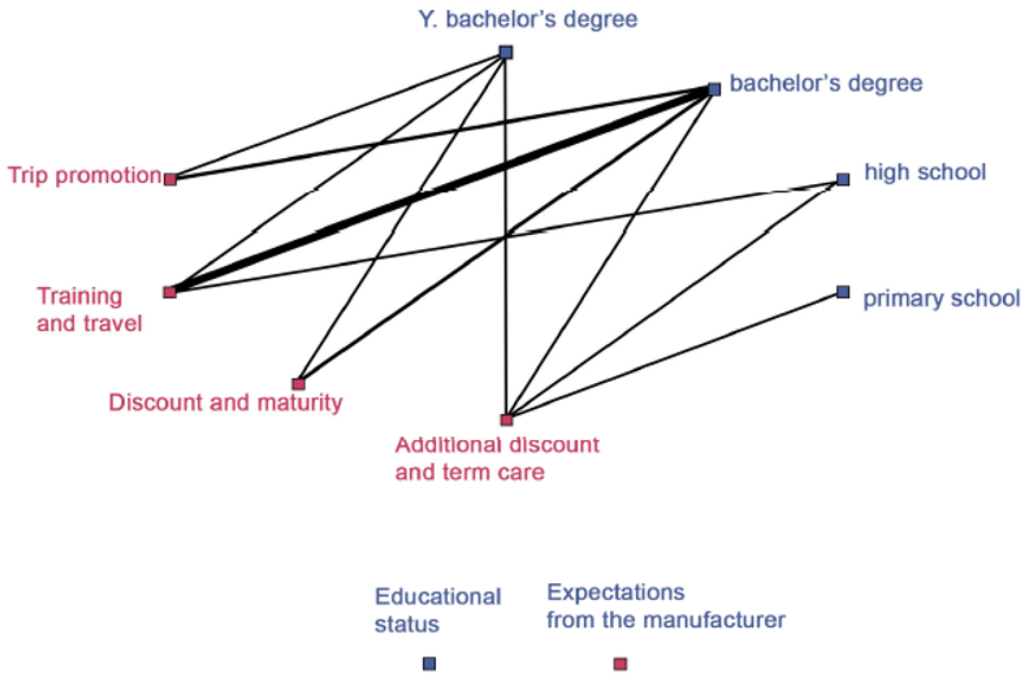

Figura 3.4: Gráfico web del estado educativo y las expectativas del fabricante.

Comprensión de los datos: en esta se detectan registros anómalos que muestren un comportamiento diferente de los valores medidos previamente en la matriz de datos. Después de aplicar análisis de anomalía, detectaron dos grupos de registros, uno con 236 y el otro con 64 registros con índice de anomalías.

Modelado: aplicaron la Redes Kohonen para elegir un subconjunto de variables de entrada eliminando características con poca información predictiva. Implementaron el algoritmo de Arboles de decisión para reducir los datos y almacenar los datos relevantes en un dataset.

Evaluación: Se utilizó una evaluación de precisión de validación cruzada estratificada diez veces para entrenar y probar la matriz de datos. La tasa de precisión del modelo es del 92,67 %, que se muestra en el Cuadro 3.3.

	Cantidad de registros	**Porcentaje de registros**
Registros correctos	278	92.67 %
Registros incorrectos	22	7.33 %
Total de registros	300	100 %

Cuadro 3.3: Precisión del modelo.

Despliegue: analizaron los datos finales que se muestra en el cuadro anterior con los expertos encargados en marketing de las compañías evaluadas, dieron sus opiniones respecto al resultado de la predicción de quejas frecuentes y vieron que la implementación de técnicas de análisis de datos de la red Kohonen y arboles de decisión.

Con el proyecto realizado llegaron a la conclusión, que las decisiones del canal de marketing son tan importantes como las decisiones que las empresas toman sobre las características y precios de los productos aplicando DM a los canales de marketing para mejorar la eficiencia del sistema de marketing.

3.3. Aplicación de Machine Learning en Marketing Digital

Vivimos en un mundo cambiante donde las tecnologías evolucionan a diario, y esto implica que se requieran perfiles profesionales que conozcan las últimas novedades del sector. En el ámbito de formación en marketing digital, se ha observado cómo disciplinas como la de Community Manager o el posicionamiento en buscadores SEO surgieron con la llegada de las redes sociales y de los motores de búsquedas respectivamente. Hoy en día, nos encontramos con muchos más avances y, por consiguiente, con la necesidad de formar a un amplio rango de perfiles digitales.

La IA y el machine learning, aplicados al Marketing, sin dudas son herramientas poderosas que transformarán la manera de crear estrategias de mercadeo digital. El machine learning diseña y desarrolla complejos algoritmos creados para aprender por sí mismos con base a la identificación de patrones de comportamiento (o aplicado al marketing, de gustos y preferencias) mediante el análisis de una gran cantidad de datos [35]. De acuerdo con Forbes, durante 2018. De acuerdo con Brenda Aranda [36], 84 % de las empresas de marketing comenzaron a implementar o expandir su uso de IA con machine learning y el 75 % de éstas lograron incrementar un 10 % la satisfacción de sus clientes.

3.3.1. Principales etapas de arquitectura de machine learning

Machine learning es la tendencia en la actualidad para predecir el gusto de los clientes mediante aprendizaje basado a datos introducidos y de esta manera predecir el gusto de un cliente, así como una tendencia, sin embargo, aún se encuentra en sus etapas iniciales y no se le ha dado un mayor beneficio en las distintas áreas que se pueden aplicar, una de ellas es el sector de marketing digital.

En la figura 2.4 se mostró la arquitectura general la cual cuenta con 5 etapas fundamentales para el desarrollo de un modelo de aprendizaje. En las siguientes subsecciones se describe detalladamente los procesos y tareas que se llevan a cabo en cada fase de la arquitectura aplicado al marketing digital [16].

Recopilacion de datos

En esta etapa se recolectan datos de varias fuentes de repositorios de datos para poder introducirlos en la plataforma de procesamiento. Esta parte de la arquitectura contiene los elementos necesarios para asegurar que la ingestión de datos ML sea fiable, rápida.

Actividades

Cómo se manejan estos datos antes de la ingestión depende de si los datos están viniendo en fragmentos discretos o un flujo continuo.

- Almacenar y reenviar los datos confidenciales a través de un almacén de datos por lotes
- Utilizar datos de transmisión continua (especialmente si las corrientes de datos grandes y erráticas están alimentando el proceso ML)
- Utilizar una plataforma de procesamiento de corriente
- Filtrar los datos que no se necesitan para procesar
- Almacenar algunos en el almacén de datos para futuros informes o pasar una porción a lo largo de si es necesario para el procesamiento inmediato

Sugerencia de Gartner: buscar herramientas que admitan estrategias de ingestión por lotes y datos en tiempo real para aprovechar los datos en movimiento para el procesamiento de ML.

Preprocesamineto de datos

En esta etapa se reenvían los datos ingeridos para la integración avanzada y los pasos de procesamiento necesarios para preparar los datos para la ejecución de ML. Esto puede incluir módulos para realizar cualquier transformación de datos, normalización, limpieza y codificación de los pasos necesarios. Además, si se está utilizando el aprendizaje supervisado, los datos necesitarán tener pasos de selección realizados para preparar conjuntos de datos para la formación.

Actividades

Hay varias consideraciones clave que pueden influir en las opciones de esta parte de la arquitectura.

- Procesar los datos en tránsito o en reposo
- Procesamiento continuo o alto rendimiento
- Implementar una arquitectura Lambda (opcional)
- Realizar procesamiento en memoria para procesamiento de alta velocidad
- Integrar datos (en esta capa pueden incluir si se utilizan una aplicación de integración de datos independiente o una plataforma de integración como una oferta de servicio

Gran parte de los datos ingeridos para el procesamiento pueden incluir características (variables de alias) que son redundantes o irrelevantes. Por lo tanto, los profesionales técnicos deben permitir la posibilidad de seleccionar y analizar un subconjunto de los datos para reducir el tiempo de entrenamiento, o simplificar el modelo. En muchos casos, el análisis de características es una parte de la selección de la muestra. Sin embargo, es importante destacar este subcomponente para filtrar datos que puedan violar las condiciones de privacidad o promover predicciones no éticas. Para combatir la privacidad y las preocupaciones éticas, los usuarios deben centrarse en la eliminación de características de ser utilizado en el modelo.

Tenga en cuenta que es un buen principio extraer la mayor cantidad de datos posible de las fuentes cuando estén disponibles. Esto se debe a que es difícil predecir qué campos de datos son útiles. La obtención de copias de las fuentes de datos de producción puede ser difícil y sujeta al control de cambios estricto. Por lo tanto, es mejor obtener un superconjunto de los datos disponibles y luego restringir los datos que se utilizan en el modelo mediante el uso de filtrado o de vistas de la base de datos. Si durante el desarrollo se hace evidente que se necesitan más campos de datos, entonces es posible simplemente relajar los criterios de filtrado o de vista, y los datos adicionales están inmediatamente disponibles. El almacenamiento es barato, y esto hace que el proceso sea mucho más ágil.

Las herramientas de preparación de datos de autoservicio se utilizan a menudo para realizar análisis y selección de funciones.

Recomendación de Gartner: buscar las herramientas que apoyan la preparación de datos del autoservicio para proporcionar a equipos de la ciencia de datos y los reveladores con la capacidad de manipular datos para apoyar algoritmos o modelos de ML. La gobernanza desempeñará un papel importante en el componente de su arquitectura. Además, considere asegurar las partes de aprendizaje y clasificación de su arquitectura para asegurar que las consideraciones de privacidad o éticas no se infringen a través del adversarial ML.

Entrenamiento del modelo

La parte de modelado de entrenamiento de la arquitectura es donde los algoritmos son seleccionados y adaptados para abordar el problema que se examinará en la fase de ejecución.

Actividades

- Identificar el tipo de aprendizaje (supervisado, no supervisado)

- Identificar las técnicas de análisis de datos
- Seleccionar algoritmos de formación de datos.

Note que los algoritmos no necesitan ser construidos desde cero por su equipo de ciencia de datos. Muchas bibliotecas útiles de algoritmos extensibles están disponibles en el mercado, y se pueden ampliar y adaptar para su propio uso. Al comenzar con ML, la experiencia se puede ganar obteniendo algunos algoritmos comunes - supervisados o sin supervisión - del mercado y desplegándolos en la nube con algunos datos para realizar experimentos. Éstos pueden revelar avenidas prometedoras para futuros valores empresariales y, eventualmente, expandirse en implementaciones formales de ML.

Recomendación de Gartner: Considerar los toolkits del ML en vez de desarrollar algoritmos a partir de cero. Además, considere bases de datos de valor clave para almacenar metadatos asociados con los modelos ML. Por ejemplo, se puede utilizar un almacén de valor clave para almacenar información semántica o dependiente del contexto sobre los modelos o algoritmos de ML.

Prueba del modelo

Una vez que los datos están preparados y los algoritmos han sido modelados para resolver un problema específico de negocio, la etapa se establece para las rutinas de ML que se ejecutarán en la parte de ejecución de la arquitectura.

Actividades

- Ejecutar repetidamente el modelo ML (como ciclos de experimentación, pruebas)
- Realizar la afinación para optimizar el rendimiento de los algoritmos
- Refinar los resultados (en preparación para el despliegue de esos resultados para consumo o toma de decisiones)

Una consideración clave en esta área es la cantidad de potencia de procesamiento que será necesaria para ejecutar eficazmente las rutinas de ML, ya sea que la infraestructura esté alojada local u obtenida como un servicio de un proveedor de nubes. Dependiendo de lo avanzado que sean las operaciones de ML, el desempeño necesario aquí puede ser significativo. Por ejemplo, para una red neural relativamente simple con sólo cuatro o cinco entradas (o "funciones") en ella, el procesamiento podría manejarse utilizando una CPU normal en un servidor de escritorio o un ordenador portátil. Sin embargo, una red que tiene numerosas características diseñadas para realizar rutinas avanzadas y de aprendizaje profundo, probablemente necesitará una potencia de cómputo de alto rendimiento en la plataforma de ejecución en forma de clusters de alto rendimiento (HPC), o calcular kernels ejecutándose en unidades de procesamiento de gráficos de alta potencia (GPUs).

Los algoritmos de ML pueden ser altamente no deterministas, lo que significa que los algoritmos pueden ejecutarse de forma diferente y producir resultados y resultados de performance diferentes dependiendo de comportamiento variable inesperado, secuencias de datos y evaluación de características. Una consideración importante aquí es examinar las infraestructuras que escalan automáticamente para evitar

interrupción o latencia excesiva debido a las técnicas de estrangulado. Un entorno adecuado para ejecutar ML está en la nube. Los entornos de nubes son altamente elásticos, lo que puede ahorrar en la sobrecarga de las soluciones de ingeniería en los locales.

Recomendación de Gartner: Considere las herramientas que ofrecen la supervisión y la ejecución de los experimentos de ML, la colaboración, y la reutilización del código. Es esencial ver el rendimiento de diferentes experimentos de ML hacia la optimización.

Optimización del rendimiento o despliegue

La salida de ML es similar a cualquier otra salida de aplicación de software, y puede persistir en almacenamiento, archivo, memoria o aplicación o hasta el componente de procesamiento para ser reprocesado. En muchos casos, la salida de ML es persistida a los paneles de instrumentos que alertan a un tomador de decisión de un curso de acción recomendado.

Comprender que el despliegue de la información resultante, las herramientas o la nueva funcionalidad generada por la rutina de machine learning varían dependiendo del tipo de ML que se esté utilizando y de qué valor se pretende generar. La salida desplegada podría tomar la forma de información reportada, nuevos modelos para complementar las aplicaciones de análisis de datos, o información que se almacenará o alimentará en otros sistemas.

Una consideración importante es si esta parte de la arquitectura necesitará ser operacionalizada (definir estrictamente variables en factores medibles). Esto es poco probable que sea el caso si la rutina de ML es exploratoria en la naturaleza porque la naturaleza de los resultados finales, y cómo ellos están desplegados es menos probable ser capaz de ser planeado o predicho por adelantado. Sin embargo, para las rutinas de ML no exploratorias, es posible que sea necesario planificar la operacionalización[8] de las fases de ejecución e implementación en la arquitectura.

Recomendación de Gartner: desarrollar un proceso que mueva sin problemas los experimentos de ML en producción. Esto se puede hacer a través de métodos de aplicación tradicionales o a través de productos de software COTS. Tradicionalmente, un reto en el despliegue ha sido que los lenguajes necesarios para definir estrictamente variables en factores medibles de los modelos que han sido diferentes de los que se han utilizado para desarrollarlos. Tenga eso en mente a medida que adquiera COTS para operacionalizar los programas de ML.

3.3.2. Machine Learning en Marketing Digital

Las empresas y compañías que han implementado minería de datos para el análisis de ventas o las tendencias que han tenido en temporadas, cuentan con datos relevantes y limpios para ser procesados utilizando machine learning, esta herramienta de aprendizaje optimiza la toma de decisiones y puede predecir material publicitario y estrategias de marketing.

[8] https://explorable.com/es/operacionalizacion

En la actualidad muy pocas empresas optan por el uso de estas herramientas ya que desconocen de los b beneficios que estas tienen o la manera de cómo aplicarlos dentro de sus negocios.

A continuación, describiremos algunos enfoques que han implementado y han propuesto modelos de arquitectura de machine learning para optimizar el trabajo del análisis de datos y de esta manera predecir tendencias y detectar anomalías en el marketing digital.

Adopción de herramientas de machine learning para uso en marketing digital de empresas

Andrej Miklosik *et al.* [37] se centraron en la selección y adopción de las herramientas analíticas impulsadas por ML por tres grupos distintos: agencias de marketing, empresas de medios y anunciantes. Los resultados destacan: 1) el importante papel de las herramientas analíticas inteligentes en la creación y despliegue de estrategias de marketing; 2) la falta de conocimiento sobre tecnologías emergentes, como ML e inteligencia artificial (IA); 3) la posible aplicación de las herramientas de ML en marketing, y; 4) el bajo nivel de adopción y utilización de las herramientas analíticas basadas en ML en la gestión de marketing.

Desarrollaron un marco compuesto de habilitadores y un mapa de proceso para ayudar a las organizaciones a identificar las oportunidades y ejecutar con éxito proyectos orientados hacia el despliegue y adopción de las herramientas analíticas de ML en el marketing digital.

Estructuraron un modelo de aprendizaje automático para ser utilizado en marketing digital (ver Figura 3.5), y lo pusieron a prueba en nueve compañías.

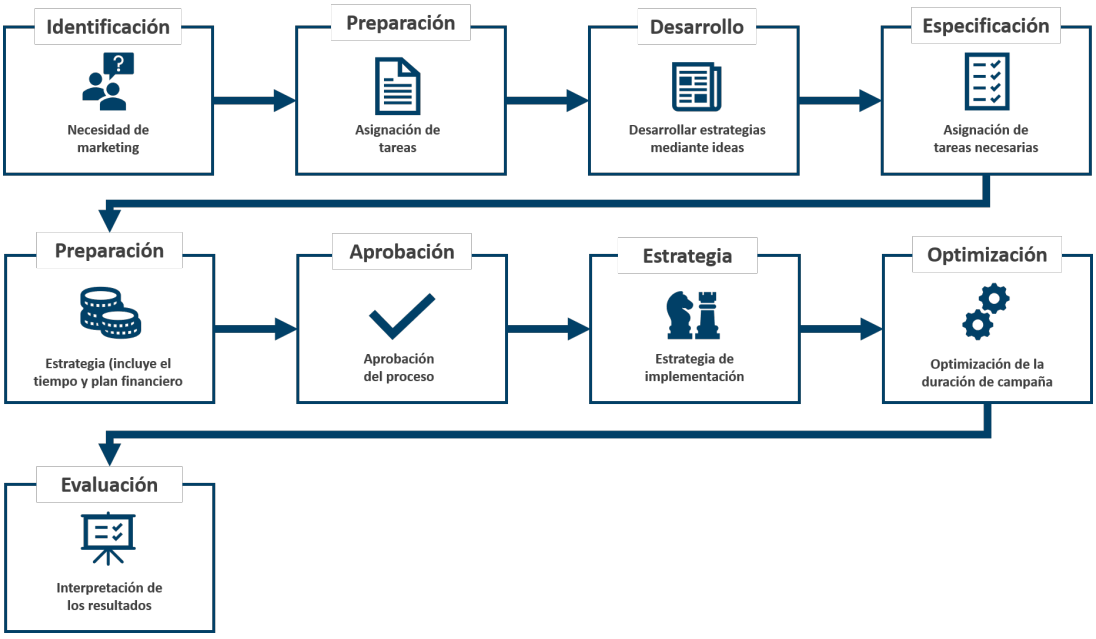

Figura 3.5: Proceso de Machine Learning implementado en marketing digital.

A continuación, se describe de manera breve en cada una de las etapas del modelo propuesto.

Identificación: en esta etapa se identifica la necesitad o el problema a resolver en el área de marketing, lo cual conlleva en crear un objetivo y un plan o estrategias para resolver el problema.

Preparación: se realiza la asignación de las tareas, es decir, lo que se realizara para resolver el problema sin salir del objetivo. Esto debe estar estructurado de manera técnica. Desarrollo: se crean estrategias mediante las ideas del equipo de trabajo que llevara a cabo el proyecto de machine learning, se documenta de manera técnica los que se realizara en las propuestas.

Especificaciones: de acuerdo con las estrategias propuestas se asignan las tareas necesarias para realizar el modelo de aprendizaje que es la encargada de generar conocimiento analítico y aplicarlos en las campañas de marketing.

Preparación: para llevar a cabo la realización de la estrategia, se realiza un cronograma de acuerdo al tiempo que se necesita para llevar a cabo el proyecto. De igual manera el plan financiero, en este caso es la inversión que se utilizara para el desarrollo del proyecto.

Aprobación: en esta etapa se realiza la validación de la estrategia, es decir, que cumpla con los puntos necesarios para resolver el problema, y que cumpla con los objetivos planteados.

Estrategia: en esta etapa se lleva a cabo la estrategia de implementación del modelo de aprendizaje automático, se aplican los algoritmos de aprendizaje (redes neuronales, arboles de decisión, etc.) que el experto en ML seleccione para resolver el problema. Si este no tiene los resultados favorables se regresa nuevamente a la etapa de desarrollo para plantear una nueva estrategia.

Optimización: si el modelo de aprendizaje es viable y cumple con la solución del problema, en esta etapa se optimiza la duración del aprendizaje, se afinan y se hacen pruebas para hacer más eficiente el modelo.

Evaluación: en esta etapa se evalúan los resultados de las pruebas del modelo de aprendizaje, y se selecciona el más óptimo para su uso.

Como resultado llegaron a la conclusión que las nueve compañías confirmaron que trabajan con datos y usan análisis de datos casi diariamente. Todos los que implementaron, acordaron que no tomarían más pasos sin la información obtenida del análisis de la comercialización. Esto también se encuentra en el modelo (Figura 3.5) donde el análisis de marketing y las herramientas analíticas que sirven como fuente de información de entrada son una parte esencial de la preparación y ejecución de los pasos 2, 3, 5, 7, 8 y 9.

Las compañías confirmaron la dependencia de los comerciantes sobre análisis de marketing y herramientas analíticas, ya que la información apoya la toma de decisiones en el desarrollo de estrategias de marketing digital.

Propuesta de modelo de machine learning para la práctica en marketing digital

Kensa Bayoude *et al.* [15] presentaron propuestas referentes a machine learning para mejorar las estrategias de marketing. Propusieron un modelo ML (ver Figura 3.6) para la predicción de productos, así como otras herramientas, con el fin de revolucionar las estrategias de marketing.

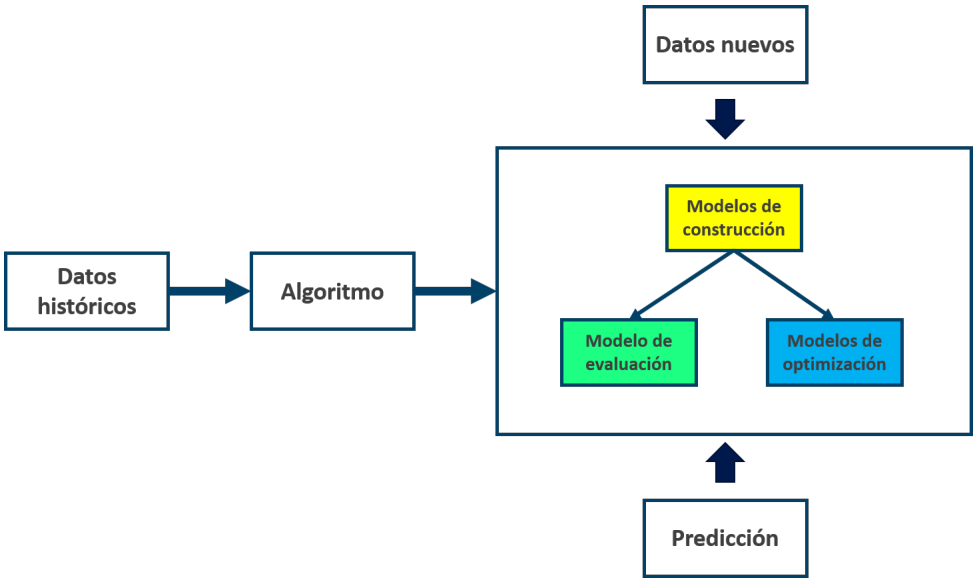

Figura 3.6: Modelo de Machine Learning propuesto para la predicción de material publicitario y productos en marketing digital.

A continuación, se describe cada etapa del modelo propuesto.

Datos históricos: en esta etapa se introducen los datos relevantes de años anteriores o temporadas pasadas de acuerdo al problema que se quiere resolver, estos deben estar en conjunto, para ser procesados.

Algoritmo: en esta etapa se selecciona el algoritmo que nos ayudara en nuestro modelo de aprendizaje, puede ser de red neuronal, clúster, etc., esto de acuerdo al tipo de aprendizaje que se utilizara (supervisado, no supervisado).

Modelo: en esta etapa se lleva a cabo el desarrollo del modelo, es decir todo el proceso que se realizara incluyendo los algoritmos de aprendizaje seleccionado, se evalúan los modelos, y el que mejor se apegue a la solución del problema se optimiza, refinando el proceso de predicción.

Es decir, cada vez que se introduce un nuevo conjunto de datos, este arrojara la predicción de una tendencia de temporada, el gusto del cliente y nuevas opciones de realizar contenido publicitario en marketing digital.

Con el estudio de las herramientas de machine learning en marketing digital, llegaron a la conclusión de que los avances tecnológicos siempre han ayudado a las empresas creando nuevas oportunidades para llegar a los clientes. Una de las mejores tecnologías de nuestro tiempo es el aprendizaje automático. Crea

nuevas oportunidades de narración y marketing que es el cambio de cómo interactúan las personas con información, tecnología, marcas y servicios.

Detección de ventas de sustancias ilícitas en Twitter utilizando machine learning

Tim Mackey *et al.* [38] desarrollaron e implementaron una metodología de machine learning para detectar con precisión la comercialización de productos ilegales. (ver Figura 3.7) Utilizaron la API de la red social Twitter para filtrar palabras claves comunes, utilizaron el aprendizaje automático no supervisado para aislar los clústeres de datos asociados con marketing. Revisaron los tweets para evaluar la característica de los vendedores ilegales.

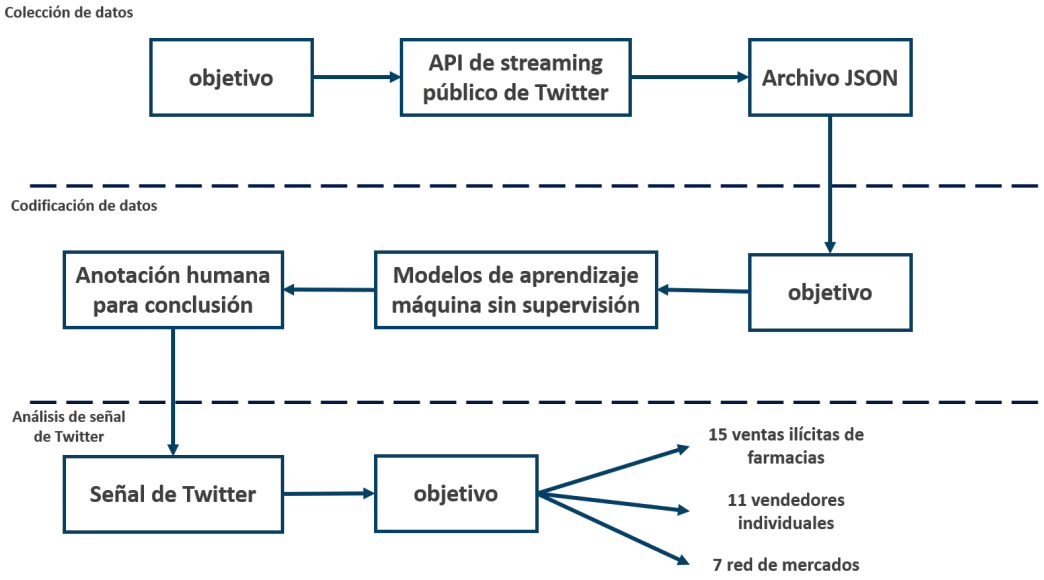

Figura 3.7: Modelo de metodología de estudio.

Colección de datos: procedieron a recopilar mensajes (es decir, tweets) publicados en Twitter durante un período de aproximadamente 20 días a partir del 15 de noviembre, 2017 al 5 de diciembre de 2017 (el día antes del inicio del Código-a-Thon). La interfaz pública de aplicaciones de streaming (API) disponible en Twitter fue usado con ciertas palabras clave preseleccionadas que eran una combinación de Nombres no propietarios internacionales y nombres de marcas de opiáceos comúnmente abusados.

Identificaron los tweets relevantes relacionados con el problema de desafíos (es decir, datos de la señalización y datos) en grandes volúmenes de datos de Twitter (en los cientos de miles).

Codificación de datos: especificaron, métodos no supervisados para obtener un resumen de los temas subyacentes presentes en un gran cuerpo de texto, eligieron métodos basados en el modelado de temas sin supervisión sobre otros enfoques, como los basados principalmente en analizar co-ocurrencias de hashtag. Utilizaron modelo llamado Biterm topic model (BTM) diseñado para detectar temas y patrones en corpus de textos cortos (como tweets), que anteriormente hemos utilizado para examinar el comportamiento de abuso de medicamentos recetados, marketing y acceso en línea (vea la Figura 3.7) BTM fue elegido porque está específicamente diseñado para trabajar en escenarios donde la longitud de los documentos o

mensajes son cortos.

Sin embargo, para asegurar que BTM fuera el método más adecuado realizaron experimentos con otros tres modelos temáticos: (1) asignación Latente de Dirichlet, (2) Factorización de Matriz Nonnegativa, (3) y Kernel k-medios. Sobre la base de estas pruebas, BTM[9] anotó más alto a través de ambas métricas para números de racimo diferentes. Por lo tanto, optaron por el uso de BTM para esta fuente de datos, en particular tenía un alto rendimiento en comparación con otros modelos de temas.

Análisis de señal de Twitter: para todos los tweets de señalización que se categorizaron específicamente como farmacias en línea ilegales, también cruzaron referencias a las URL de estos sitios web con la base de datos externa de LegitScript[10] que incluye una clasificación legal. La clasificación legal de LegitScript se basa en su propia evaluación de si el sitio web es (1) el fabricante de los pícaros: vendedor contratado en actividades ilegales, inseguras o engaña, (2) no aprobado por el tejido: proveedor con un problema de cumplimiento regulatorio o riesgo en una o más jurisdicciones, (3) no verificado por la persona: no sujeto a revisión o monitoreo de LegitScript, o (4) la norma de la certificación LegitScript. Las consultas de clasificación de LegitScript ofrecen otra capa de verificación con respecto a un estado legal de la farmacia en línea y pueden ayudar a confirmar que los sitios presentan alto riesgo para los consumidores. También revisaron los datos de WHOIS para determinar la dirección del protocolo de Internet (IP) y la ubicación del propietario registrado para enlaces clasificados como farmacias en línea.

[9] https://rdrr.io/cran/BTM/man/BTM.html
[10] https://www.legitscript.com/

Capítulo 4

Resultados

En este capítulo se dan a conocer las estrategias de marketing digital tomados de diferentes enfoques relacionados con minería de datos y machine learning.

4.1. Análisis de la metodología CRISP-DM en gestión de proyectos de minería de datos

La metodología KDD y CRISP-DM tiene ciertas etapas y una estructura de modelo diferente a las otras metodologías, utilizan nodos para realizar minería de datos, sin embargo, existen diferencias y áreas en donde aplicarlo.

4.1.1. Escenarios y puntos de partida considerados para el proyecto

Entre los dos modelos analizados, CRISP-DM y KDD inician con un análisis del negocio y del problema organizacional, con el fin de planear una estrategia y asignar las actividades que se realizaran en las etapas.

4.1.2. Estructura de fase del proceso de minería de datos

La metodología KDD y CRISP-DM contemplan el análisis y comprensión del problema antes de comenzar el proceso de minería.

En ambos modelos se contempla la selección y preparación de los datos (ver Cuadro 4.1). Esta situación se repite para la fase de modelado, donde se aplican las técnicas de minería para obtener los nuevos patrones.

La fase de evaluación de los patrones obtenidos está presente también en todas las metodologías.

En CRISP-DM, se propone además una planificación para el control futuro y un análisis de cierre del proyecto (análisis postmortem). El análisis postmortem consiste en encontrar información objetiva acerca de la trayectoria de un proyecto, con la finalidad de poder hacer una evaluación abierta del equipo de trabajo, de las decisiones tomadas a lo largo del mismo, de las tecnologías empleadas y sus consecuencias, con el objetivo de incorporar lo aprendido en proyectos futuros [39].

Fases	KDD	CRISP-DM
Análisis y comprensión del negocio	- Compresión del dominio de aplicación	- Comprensión del negocio
Selección y preparación de los datos	- Crear el conjunto de datos - Limpieza y preprocesamiento de los datos - Reducción y proyección de los datos	- Entendimiento de los datos - Preparación de los datos
Modelado	- Determinar la tarea de minería - Determinar el algoritmo de minería - Minería de datos	- Modelado
Evaluación	- Interpretación	- Evaluación
Implementación	- Utilización del nuevo conocimiento	- Despliegue

Cuadro 4.1: Fases del proceso de minería de datos en la metodología KDD y CRISP-DM.

4.1.3. Nivel de detalle en las tareas de cada fase

El modelo KDD propone sólo los pasos generales del proyecto de minería de datos, sin especificar puntualmente las tareas que deben llevarse a cabo en cada una de sus fases. En cambio, el modelo CRISP-DM, especifican con mayor detalle las actividades del proceso.

KDD se acerca más a un modelo de proceso que a una metodología, ya que sólo definen las fases generales. En proyectos donde se desee aplicar los mismos, cada organización deberá establecer las tareas y las actividades que implementará en cada etapa.

Si bien, CRISP-DM no llegan a especificar con un alto nivel de detalle cómo realizar todas las tareas, podrían ser considerados una metodología ya que describen y puntualizan las actividades específicas a realizar en cada fase del proceso.

4.1.4. Actividades para la gestión de proyectos

En el Cuadro 4.2, se puede observar que la metodología CRISP-DM propone actividades de planificación para las distintas áreas de la gestión del proyecto, pero no explican tareas de control y monitoreo. KDD no incluye actividades de gestión del proyecto.

En CRISP-DM las actividades de planificación se ven reflejadas en las tareas “Evaluación de la situación” y “Crear un plan para el proyecto de minería de datos”. Si bien no se explicitan tareas de seguimiento y control, en el modelo se aclara que el plan del proyecto debe ser revisado (y de ser necesario modificado), antes de comenzar con cada fase del proceso.

Fases	KDD	CRISP-DM
Gestión del alcance	—	Planificación del alcance de la tarea
Gestión del tiempo	—	Planificación del tiempo de la tarea
Gestión del costo	—	Planificación del costo de la tarea
Gestión del riesgo	—	Gestión del riesgo en la tarea
Gestión de los recursos humanos	—	Planificación de los recursos humanos en la tarea

Cuadro 4.2: Actividades de la gestión de proyectos en cada modelo.

4.2. Técnicas que las empresas utilizan en la minería de datos y machine learning.

4.2.1. Minería de datos aplicado en una empresa de moda

La minería de datos está revolucionando las técnicas de marketing digital, actualmente las empresas dedicadas al negocio de vestir figuras públicas (famosos) utilizan estas herramientas. Un ejemplo de ello fue la empresa dedicada a vestir figuras públicas, implemento técnicas de minería de datos para pronosticar el tipo de moda que es preferido para los usuarios.

En la Figura 4.1 se muestra el modelo, técnicas y herramientas que se utilizaron para realizar el proyecto de minería de datos, utilizaron la metodología CRISP-DM, como se mencionó anteriormente está enfocada a negocios por ser un modelo descriptivo.

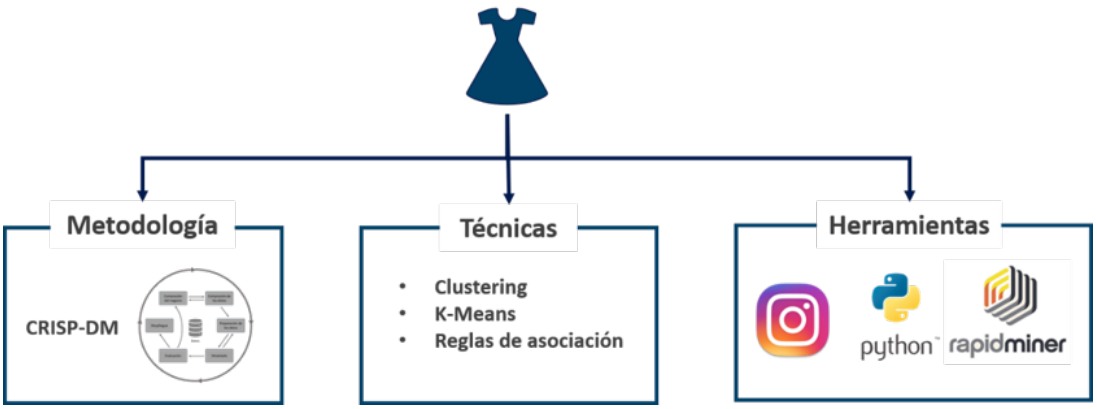

Figura 4.1: Metodología, técnicas y herramientas utilizadas en la empresa de moda.

Como herramientas utilizadas para realizar el proyecto de minería, utilizaron la API de Instagram para poder tener acceso a los datos con ayuda del lenguaje de programación Python siendo este un lenguaje de alto nivel y óptimo para la extracción de datos. Para realizar la minería de datos utilizaron el software Rapidminer ya que es un software muy amigable para el usuario y trabaja mediante bloques.

4.2.2. Minería de datos en el SEO de Amazon.com

El comercio electrónico es uno de los sectores principales en la aplicación de minería de datos, en la sección "*Minería de datos para modelos predictivos de los factores de mercado en línea basados en*

la demanda del cliente“ del capítulo 3 Truong Van Nguyen *et al.* [34] utilizan el SEO de Amazon para estrategias de mercadeo con el fin de identificar la preferencia de los clientes. En la Figura 4.2 se observa la metodología, técnicas y herramientas utilizadas en este proyecto.

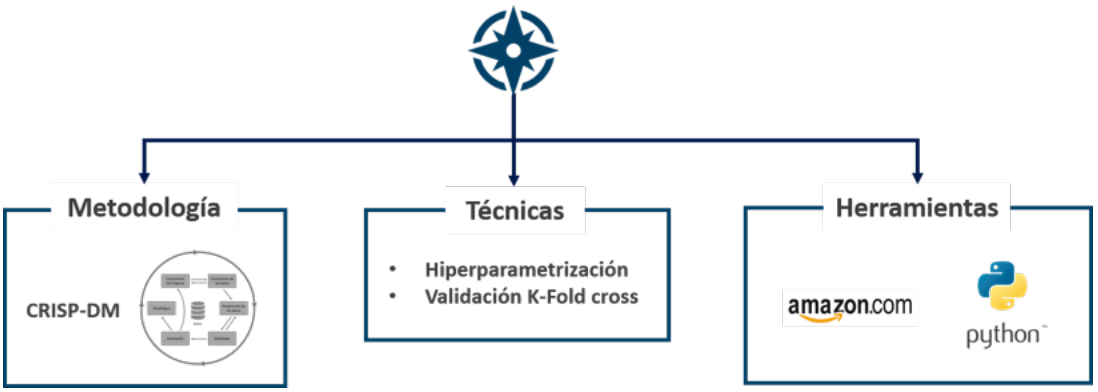

Figura 4.2: Metodología, técnicas y herramientas utilizadas por Amazon.com.

La metodología utilizada en este proyecto de minería fue CRISP-DM, como el caso anterior, esta metodología es óptima para el área de negocios y adaptativa, las técnicas utilizadas fue la hiperparametrización ya que aproxima las funciones de datos, de igual manera utilizaron la validación cruzada K-Fold, toma los datos originales y crea dos conjuntos separados.

Las herramientas utilizadas para la minería de datos fue SEO de Amazon.com, ya que este realiza el registro de los que los clientes buscan, el lenguaje de programación Python se utilizó para extraer los datos almacenados.

4.2.3. Machine learning en empresa de seguridad de comercio electrónico

No solo las empresas utilizan minería de datos, existen empresas que hacen uso de machine learning para generar conocimiento, estas generan nuevos contenidos y material publicitario, así como la detención de comercio de productos ilícitos.

En la sección ”*Detección de ventas de sustancias ilícitas en Twitter utilizando machine learning*“ del capítulo 3, una empresa de seguridad aplico minería de datos y machine learning para la detección de ventas de productos ilícitos en la red social de Twitter, en la Figura 4.3 se muestran las técnicas y herramientas utilizadas en el proyecto de minería de datos.

Las técnicas fueron, Biter Topic Model utilizado para la clasificación de textos cortos que se realizan en cada tweet, factorización de matriz para realizar análisis multivariado, K-means para agrupar los tweets en diferentes grupos, LegitScript para monitorear los pagos que se realizan.

Las herramientas que utilizaron para este proyecto fueron: API de Twitter para realizar la extracción de los tweets que se realizan, y el lenguaje de programación Python para utilizar la API de la red social ya mencionada.

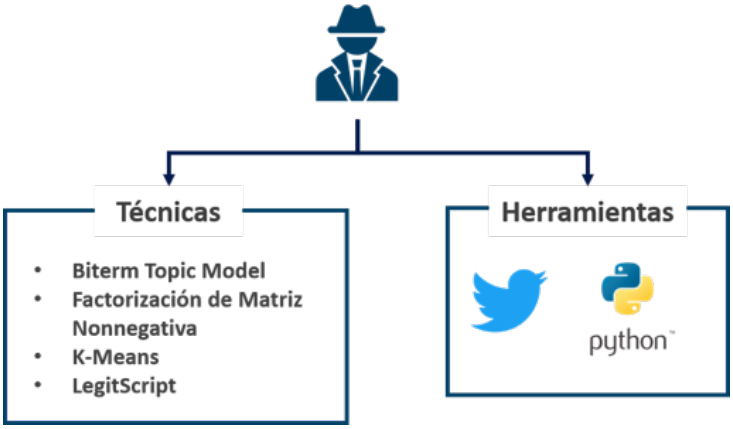

Figura 4.3: Técnicas y herramientas para la detección de comercio ilícito.

4.3. Arquitectura de la propuesta de un modelo de Machine Learning

Actual mente existe pocos trabajos reportados en la literatura que hayan abordado machine learning aplicados al marketing digital, la mayoría son propuestas de modelos que hacen uno se la ML en marketing digital. Sin embargo, son modelos básicos sin la explicación adecuada de las actividades que se deben realizar.

En la Figura 4.4 se muestra una propuesta de un modelo de ML aplicado al marketing digital, este modelo consta de 5 etapas que son: Adquisición de datos, procesamiento, modelo de aprendizaje, ejecución e implementación.

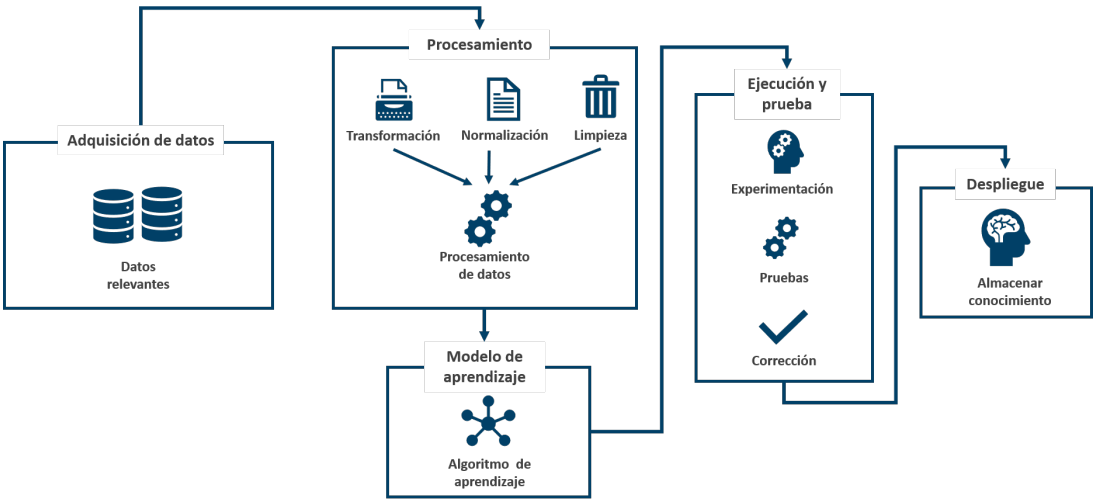

Figura 4.4: Propuesta de modelo de ML en marketing digital.

De acuerdo a las etapas de la arquitectura general del modelo de machine learning descritos en la *sección 3.3.1* del *capítulo 3*, las etapas del modelo propuesto serán las que se describirán a continuación:

Adquisición de datos: en esta etapa se ingresan en conjunto de datos seleccionado que ayudara a

resolver el problema, los datos pueden provenir de un proyecto de minería de datos, sin embargo, estos datos deben de tener ciertas características como:

- Relevancia: los datasets deben contener datos importantes, lo cual al transformarlos en información ayudaran a generar un conocimiento adecuado al aplicar los algoritmos de aprendizaje
- Limpieza: estos datos no deben contener datos irrelevantes, los datasets deben contener solo datos referentes al problema que se enfrentara
- Estructurado: los datasets deben contener datos estructurados, es decir estar agrupado de manera adecuada

Preprocesamiento: se transforman los datos a lenguaje máquina, ya que los algoritmos de aprendizaje traban de manera binomial, se normalizan los datos y se realiza una limpieza para prevenir que se introduzcan datos irrelevantes al ejecutar el algoritmo de aprendizaje. Lo siguiente es procesar los datos para proseguir con la etapa de aprendizaje.

Modelo de aprendizaje: se selecciona el algoritmo de aprendizaje que se utilizara (pueden ser arboles de decisión, redes neuronales, series de tiempo, regresión, etc.). El algoritmo de aprendizaje debe cubrir la necesidad del objetivo.

Ejecución y prueba: una vez generado el conocimiento se procede a experimentar el producto, se le realizan pruebas para encontrar buscar errores y se evalúa para verificar si es óptimo para realizar las correcciones adecuadas.

En caso de que el producto obtenido no cumple con las necesidades para la solución del problema, se regresa a la etapa de modelo de aprendizaje, procediendo a la selección de un nuevo algoritmo de aprendizaje que cumpla con la solución del problema.

Despliegue: en esta etapa final se almacena el conocimiento obtenido que logra resolver el problema planteado. Lo que ayudara predecir, generar y agilizar contenido.

Si se aplica en estrategias de marketing digital optimizara el tiempo de los expertos al momento de aprender de los datos históricos, lo que llevara a realizar campañas de marketing.

4.3.1. Propuesta de una arquitectura de minería de datos y machine learning en marketing digital

De acuerdo en la literatura de los enfoques analizados, existe una gran variedad de propuesta de arquitectura de modelo de machine learning y minería de datos en marketing digital, sin embargo, en los enfoques no implementan una arquitectura en utilice DM y ML, es decir, no se plantea una arquitectura general que se implemente DM y ML en conjunto en el área de marketing digital.

Por ello, se plantea la propuesta de una arquitectura de minería de datos y machine learning para ser aplicados en marketing digital, con el fin de extraer datos y procesarlos de manera automática para generar nuevos contenidos publicitarios (ver Figura 4.5). La metodología de minería de datos que se estaría usando en esta propuesta es CRISP-DM junto con el modelo de ML propuesto.

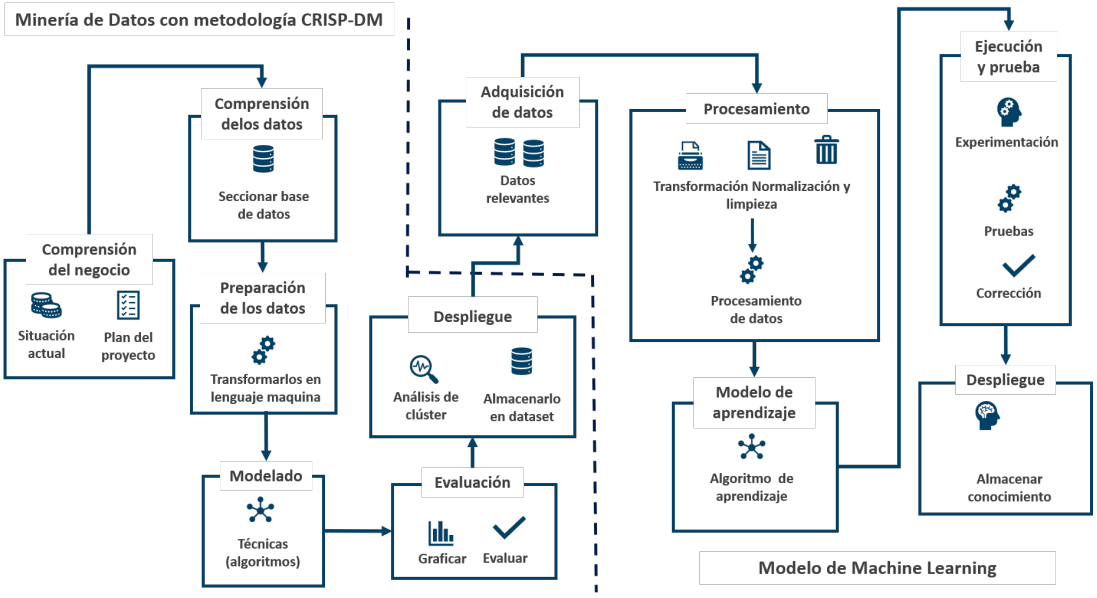

Figura 4.5: Propuesta de modelo de proceso de DM y ML para predicción de productos y material publicitario en Marketing Digital.

En este caso, las actividades y tareas que se realizarían en cada una de las etapas de minería de datos son los descritos en la *sección 3.2.1* del *capítulo 3*, ya que en ella se encuentra descrita de manera detallada las tareas a realiza. De igual manera con el modelo de machine learning se utilizaría lo descrito en el modelo propuesto.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

El uso de minería de datos y machine learning aplicados al marketing digital en las empresas de comercio está ganando relevancia con el paso del tiempo, estas herramientas inteligentes están superando el rendimiento humano en número creciente de campos de acción. Las empresas de retail que utilizan estas herramientas tienen un crecimiento en comercio global, obteniendo ingresos millonarios.

De acuerdo con la literatura, ML especialmente en sistemas de capacitación monitoreados mediante redes neuronales han despertado un gran interés en el ámbito de marketing digital, ya que los sistemas de reconocimientos pasaron a tener un margen de error muy bajo durante los últimos años equiparándose con el margen de error humano. De igual manera, la minería de datos está causando impacto en las empresas retail, ya que esta ayuda a la identificación de patrones, intereses y hábitos de los usuarios siendo clave en las estrategias de marketing para la creación de contenidos personalizados.

De acuerdo al análisis realizado, la metodología CRISP-DM podría ser considerado la mejor para realizar proyectos de minería de datos en áreas de comercio, por el nivel de detalle con el que describen las tareas en cada fase del proceso, y porque incorporan actividades para la gestión del proyecto (como gestión del tiempo, costo, riesgo). En este aspecto, la metodología KDD no incorpora actividades para el control y monitoreo del plan de trabajo. Si hablamos de algoritmos en minería de datos, de acuerdo con la literatura los mejores para ser aplicados con la metodología CRISP-DM son: reglas de asociación, arboles de decisión y redes neuronales, ya que estos algoritmos buscan relacionar los datos utilizando parte de algoritmos de aprendizaje para encontrar similitud en los datos y conservar únicamente los datos relevantes.

Por su parte, machine learning a pesar de no ser utilizado en su máximo potencial se está utilizando en la actualidad en muchas compañías de comercio electrónico, ya que les permite crear su propio modelo de aprendizaje basado en la arquitectura general para ser utilizadas en estrategias de marketing. La importancia del machine learning aparece como una solución, ya que permite analizar en tiempo real la información procesada. Los algoritmos más utilizados para esta son; redes neuronales y arboles de decisión ya que permite extraer conocimiento a través de patrones específicos, ya que analiza cada nodo respecto a las relaciones que estas tienen con otros nodos para aprender y mejorar el rendimiento.

5.2. Trabajo futuro

Debido a que en este trabajo únicamente se analizaron técnicas y herramientas de minería de datos y machine learning enfocados al marketing digital, se propone que en un futuro se implemente un proyecto de minería de datos para el pronóstico de ventas basado en datos históricos. De igual manera poder implementar la propuesta del modelo de proceso de DM y ML para predicción de productos y material publicitario en Marketing Digital, y seguir con la investigación de las nuevas herramientas inteligentes que se están utilizando para mejorar las estrategias de marketing digital.

Bibliografía

[1] J. HAN, M. KAMBER y J. PEI, Data Mining Concept and Techniques, 3dr ed., Morgan Kaufmann Publisheres, 2001.

[2] R. TIMARÁN, «Una mirada al descubrimiento de conocimiento en bases de datos,» Ventana Informática, pp. 39-58, 2009.

[3] J. Romero, «Jorge Romero,» Meotodologías de Minería de Datos, [En línea]. Available: https://jorgeromero.net/metodologias-de-mineria-de-datos/. [Último acceso: 2020 01 22].

[4] P. CHAPMAN, J. CLINTON, R. KERBER, T. KHABAZA, T. REINARTZ, C. SHEARER y R. WIRTH, CRISP-DM 1.0, 2000.

[5] H. O. NIGRO, D. XODO, G. CORTI y D. TERREN, KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario, Campus Universitario - Paraje Arroyo.

[6] A. I. AZEVEDO, I. ROJÁO y M. FILIPE, KDD, SEMMA and CRISP-DM: a parallel overview, Instituto Politécnico do Porto. 2008

[7] S. R. TIMARÁN PEREIRA, I. HERNÁNDEZ ARTEAGA, S. J. CAICEDO ZAMBRANO, A. HIDALGO TROYA y J. C. ALVARADO PÉREZ, «El proceso de descubrimiento de conocimiento en bases de datos,» de *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas*, Bogotá, Ediciones Universidad Cooperativa de Colombia, 2016, pp. 63-86.

[8] V. GALÁN CORTINA, Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario, Universidad Carlos III de Madrid, 2015.

[9] J. GIRONÉS ROIG, «Metodologías y estándares,» de *Business Analytics*, Catalunya, España, Universitat Oberta de Catalunya, 2013, pp. 1-55.

[10] A. A. AQUINO, G. MOLERO y R. ROJANO, Hacia un nuevo proceso de minería de datos centrado en el usuario, Instituto Tecnológico de Celaya, 2015.

[11] «CONEXIONESAN,» Cuatro interesantes aplicaciones empresariales de data mining, 08 agosto 2017. [En línea]. Available: https://www.esan.edu.pe/apuntes-empresariales/2017/08/cuatro-interesantes-aplicaciones-empresariales-de-data-mining/. [Último acceso: 28 enero 2020].

[12] J. HURWITZ Y D. KIRSCH, Machine Learning: for Dummies, IBM Limited Edition, John Wiley & Sons, Inc, 2018.

[13] M. MOHAMMED, M. BADRUDDIN Y E. B. MOHAMMED, Machine Learning: Algorithms and Applications, Taylor & Francis Group, LLC, 2017.

[14] M. SOLUTIONS, «Machine Learning: una pieza clave en la transformación de los modelos de negocio,» Management Solutions, pp. 18-29, 2018.

[15] K. BAYOUDE, Y. OUASSIT, S. ARDCHIR Y M. AZOUAZI, «How Machine Learning Potentials are transforming the Practice of Digital Marketing: State of the Art,» *Periodicals of Engineering and Natural Sciences*, vol. 6, nº 2, pp. 373-379, 2018.

[16] C. E. SAPP, «Preparing and Architecting for Machine,» de *Preparing and Architecting for Machine*, 2017.

[17] «SITIO BIG DATA,» 6 herramientas de servicio Machine Learning para el análisis de datos, 27 agosto 2018. [En línea]. Available: http://sitiobigdata.com/2018/08/27/6-herramientas-servicio-machine-learning-analisis-de-datos/. [Último acceso: 11 noviembre 2019].

[18] «MICROSOFT,» Algoritmos de minería de datos (Analysis Services: Minería de datos), 05 marzo 2017. [En línea]. Available: https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2014. [Último acceso: 11 noviembre 2019].

[19] «ECURED,» Clustering, [En línea]. Available: https://www.ecured.cu/Clustering. [Último acceso: 11 noviembre 2019].

[20] «DBA BUSINESS ANALYTICS,» El algoritmo de series temporales, 31 diciembre 2013. [En línea]. Available: https://francescsanchezbi.webnode.es/news/el-algoritmo-de-series-temporales/. [Último acceso: 11 noviembre 2019].

[21] L. ROKACH Y O. MAIMON, Data Mining with Decision Trees: Theory and Aplications, New Jersey: World Scientific, 2008.

[22] «APRENDE MACHINE LEARNING,» Regresión Lineal en español con Python, 13 mayo 2018. [En línea]. Available: https://www.aprendemachinelearning.com/tag/regresion-lineal/. [Último acceso: 11 noviembre 2019].

[23] D. RODRÍGUEZ, «Analytics Lane,» La regresión logística, 23 julio 2018. [En línea]. Available: https://www.analyticslane.com/2018/07/23/la-regresion-logistica/. [Último acceso: 11 noviembre 2019].

[24] «R. SALAS,» Redes neuronales artificiales. Universidad de Valparaıso, Departamento de Computación, 1, 2004.

[25] J. M. VEGA, S. A. ROMERO Y G. GUZMÁN, «Digital Marketing and the Finances of SMES,» de *Revista de Investigación en Tecnologpias de la Información*, 2018.

[26] A. NOVOA, M. SABOGAL Y C. VARGAS, «Estimación de las relaciones entre la inversión en medios digitales y las variables financieras de la empresa: una aproximación para Colombia,» de *Revista Escuela Administración de Negocios*, Colombia, 2016.

[27] . STRAUSS Y R. FROST, E-Marketing, New Jersey: Pearson Education, Inc, 2014.

[28] J. GUTIÉRREZ Y B. MOLINA, «Identification of data mining techniques to support decision-making in solving business problems,» *Revista Ontare*, pp. 33-52, 2016.

[29] D. ARTEAGA, R. REMIGIO Y D. M. CALDERÓN, «Minería de Datos Aplicado al Marketing.,» *Número Especial de la Revista Aristas: Investigación Básica y Aplicada*, vol. 6, nº 12, pp. 23-29, 2018.

[30] A. CRAVERO Y S. SEPÚLVEDA, Aplicación de Minería de Datos para la Detección de Anomalías: Un Caso de Estudio, *D. d. I. d. Sistemas*, Chile: Universidad de La Frontera, 2009.

[31] L. CONTRERAS Y K. ROSALES, «Análisis del comportamiento de los clientes en las redes sociales mediante técnicas de Minería de Datos,» de *VIII Congreso Internacional de Computación y Telecomunicaciones*, Colombia, 2016.

[32] T. VAN NGUYEN, L. ZHOU, A. Y. LOONG CHONG, L. BOYING Y P. XIAODIE, «Predicting Customer Demand for Remanufactured Products: A Data-Mining Approach,» de E*uropean Journal of Operational Research*, 2019.

[33] . CHEN, Y. CHEN Y A. OZTEKIN, «A hybrid data envelopment analysis approach to analyse college graduation rate at higher education institutions,» *INFOR: Information Systems and Operational Research*, vol. 55, nº 3, pp. 188-210, 2017.

[34] E. KAHYA OZYIRMIDOKUZ, K. UYAR Y M. HAKAN OZYIRMIDOKUZ, «A Data Mining Based Approach to a Firm's Marketing Channel,» *22nd International Economic Conference – IECS 2015*, vol. 27, pp. 77-84, 2015.

[35] «DIGITAL 57,» Machine Learning aplicado al Marketing digital: un paso más allá de la Inteligencia Artificial, 19 enero 2018. [En línea]. Available: https://www.digital57.co/machine-learning-aplicado-al-marketing-digital-paso-mas-alla-la-inteligencia-artificial/. [Último acceso: 29 enero 2020].

[36] B. ARANDA, «El Financiero,» El uso de "Machine Learning" en la estrategia de Marketing Digital de su empresa, 30 enero 2019. [En línea]. Available: https://elfinanciero.com.mx/monterrey/el-uso-de-machine-learning-en-la-estrategia-de-marketing-digital-de-su-empresa. [Último acceso: 29 enero 2020].

[37] A. MIKLOSIK, M. KUCHTA, N. EVANS Y S. ZAK, Towards the Adoption of Machine Learning-Based Analytical Tools in Digital Marketing, *Research Project VEGA (S.G.A.)*, 2019.

[38] T. MACKEY, J. KALYANAM, J. KLUGMAN, E. KUZMENKO Y R. GUPTA, «Solution to Detect, Classify, and Report Illicit Online Marketing and Sales of Controlled Substances via Twitter: Using Machine Learning and Web Forensics to Combat Digital Opioid Access,» *Journal of Medical Internet Research*, vol. 20, nº 4, pp. 1-14, 2018.

[39] J. M. MOINE, S. GORDILLO Y A. S. HAEDO, «Análisis comparativo de metodologías para la gestión de proyectos de minería de datos,» *XVII Congreso Argentino de Ciencias de la Computación*, pp. 931-938, 2011.