“MINERÍA DE DATOS PARA EMPRENDEDORES”

Silvana Pompeya, Martinez Campos

Ingeniería en Informática
Facultad de Ingeniería
Año 2021

Título:

Ingeniera en Informática

Profesor Guía:

Nombre: Gustavo Ramiro, Rivadera Firma: _____

Tribunal Evaluador:

Nombre: _____ Firma: _____

Nombre: _____ Firma: _____

Nombre: _____ Firma: _____

Alumna:

Nombre: Silvana Pompeya, Martinez Campos Firma: _____

Fecha de Exposición:

____/____/____

Dedicatoria

A mis padres Viviana Campos y Carlos Martinez,
por guiarme y darme esta oportunidad.
A mi padrino Ricardo Martinez,
por ser mi inspiración en la vida.

Agradecimientos

Agradezco a Dios por guiar mis pasos y darme la fortaleza necesaria para lograr este objetivo, a mis padres por brindarme la posibilidad de estudiar, por apostar y confiar en mí; a toda mi familia por el apoyo incondicional.

Agradezco a los profesores y administrativos de la facultad de Ingeniería de la Universidad Católica de Salta, por brindarme con amor el conocimiento y herramientas necesarias para formarme como persona y futura profesional; y sobre todo gracias a mis compañeros Santiago Alegre, Occhipinti Ignacio, Iván Maccio y Nicolás Yened por el compañerismo, el trabajo, esfuerzo y motivación de cada día para sacar esta cerrera adelante, en cada éxito y cada derrota.

Índice del trabajo

Índice de imágenes

Índice de tablas

Abstract

Minería de datos para emprendedores tiene como finalidad aplicar la minería de datos de reglas de asociación, para mejorar el proceso de análisis de la cesta de compras de una indumentaria, en búsqueda de patrones de productos que se adquieren juntos y usarlos como recomendaciones para ayudar a incrementar las ventas e iniciarse en e-commerce.

El proceso de análisis y generación de reglas de asociación se realiza mediante la herramienta de minería de datos llamada RapidMiner y se lleva a cabo con la metodología CRISP-DM.

El producto final del trabajo es: un informe de recomendaciones en base al modelo de reglas de asociación obtenido y un prototipo de tienda online que refleja el uso de las recomendaciones.

1 Introducción

Con el paso de los años los avances en la tecnología provocó que el comercio electrónico y la cantidad de datos crezcan a pasos agigantados y se conviertan en materia prima para las empresas, de grande, mediano y pequeño porte; que buscan un conocimiento inteligente de sus datos históricos y actuales para tomar decisiones exitosas. Esto los encamina a ser responsables directo del registro y análisis adecuado de los datos para obtener de ellos información de buena calidad y oportuna de la que se pueda extraer el conocimiento que ayude a tomar decisiones gerenciales.

Pero entre tanta cantidad de información ¿Cómo podemos encontrar la que estamos necesitando?, y más aún si pensamos que no solo se trata de encontrar la información adecuada, sino de obtener resultados pensantes hacia la toma de decisión, es decir ¿Cómo podemos adquirir conocimiento entre tanta información existente?, como punto de partida hacia las respuestas es importante saber que el conocimiento es una acción efectiva resultante de aplicar inteligencia sobre la información, en un contexto y con un propósito determinado.

Para poder extraer un conocimiento, en este proyecto se propone el uso de la minería de datos, que se enmarca en el Proceso de Descubrimiento del Conocimiento a partir de Base de Datos (KDD). Fayyad, Piatestky-Shapiro, & Smyth (1996) afirman que la minería de datos es: "El proceso no trivial de identificar patrones válidos, novedosos, potencialmente útil y en última instancia comprensibles a partir de datos" (p.30).

1.1 Definición del problema

Un comercio como FOER Deportes genera una gran cantidad de registros de ventas que son resguardados para poder ser consultados y utilizados para un posterior análisis. Los registros de ventas son en formato papel y consumen una gran cantidad de tiempo y recurso a la hora asociar los artículos que se compran con más frecuencia. Esta tarea intuitiva, manual o a ojo de analizar la cesta de compras para encontrar tendencias entre los productos, no es sencilla por lo que muchas veces resulta ser errónea e influye en las ventas y en el diseño de una página web. ¿Cómo posicionar los productos en el local? ¿Cómo pronosticar ventas y optimizar campañas de marketing? ¿Cómo remover e incrementar las ventas de un stock determinado? son cuestiones a resolver; además la falta de participación en la web le imposibilita abrirse a otros clientes.

1.2 Pasos a realizar

Ante la problemática de asociar de forma manual todas las ventas se considera a la minería de reglas de asociación como una solución informática, cuya tarea es el análisis automático o semiautomático de grandes cantidades de datos para extraer patrones interesantes.

Para contar con recomendaciones que indiquen el producto y sus tendencias es necesario realizar un proceso para generar las reglas de asociación. Este proceso posee varios desafíos; del conjunto de ventas en formato papel se realiza la selección de los datos, el reconocimiento de la presencia de valores atípicos y ausencia de datos para transformarlos en una cesta de compra digital comprensible. A la cesta de compra se le aplica un algoritmo de minería de

asociación para encontrar conjuntos de elementos frecuentes que se usa para encontrar el modelo de reglas de asociación. Luego este modelo se evalúa para comprobar su calidad y obtener un conocimiento potencialmente útil para el negocio. Estos resultados serán entregados a los interesados mediante un informe final y un prototipo de tienda online que refleja la implementación de los patrones encontrados.

Para analizar esta cesta de compra en búsqueda del conocimiento, empleamos la herramienta RapidMiner que nos proporciona la minería de datos. El objetivo es la obtención de un modelo que permita determinar cuáles son los artículos más demandados, determinar aquellos que se compran juntos o con qué medida la compra de un producto provoca la compra de un segundo.

El conocimiento extraído sobre el comportamiento de los clientes servirá de base a la gerencia para tomar decisiones sobre el diseño de una página web y sobre marketing, por ejemplo para realizar mejor los pedidos, sugerir productos dada la compra de otros, para aplicar la técnica de cross selling o ventas cruzadas y para diseñar espacios de compra en el local y en la web. Es esta manera se reduciría el tiempo de compra del cliente y se aumentaría la productividad de las ventas y análisis de los datos.

1.3 Motivación

La posibilidad de agilizar y mejorar el proceso manual de encontrar tendencias en las ventas del negocio, me motivaron a realizar este trabajo aplicando una solución de minería de datos para contribuir con su productividad. También me motivó el interés del usuario en considerar los avances de la tecnología y el uso del comercio electrónico como una herramienta fundamental para el crecimiento del negocio.

1.4 Metodología a usar

Para desarrollar este trabajo usaremos la metodología CRISP-DM, que se enfoca en estudiar el negocio y sus objetivos, de esta manera podemos conocer en profundidad qué aspectos son importantes para la empresa, cuáles son sus objetivos y alinearlos con un objetivo de minería de datos. La metodología describe el ciclo de vida del proceso de minería de datos en seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.

1.5 Criterios de éxito del trabajo

Los criterios de éxitos del trabajo:

- El modelo propuesto debe generar las reglas de asociación de los productos con al menos un 80% de confianza.
- El informe debe expresar recomendaciones basadas en cómo implementar las reglas de asociación obtenidas de la cesta de compra.
- El prototipo de la tienda online debe demostrar la implementación práctica de las recomendaciones expresadas en informe obtenido.

1.6 Organización del trabajo

El trabajo está organizado en seis capítulos:

Capítulo 1 - Introducción: a modo de resumen refleja la definición del problema, motivación, pasos a realizar, la metodología, los criterios de éxitos y organización del trabajo.

Capítulo 2 - Estado de la cuestión: refleja los antecedentes relacionados con el problema, fundamentos teóricos y prácticos que contribuyen a la solución escogida de minería de datos.

Capítulo 3 - Definición del problema: se definen le problema, los objetivos, alcance y alternativas para el desarrollo del proyecto.

Capítulo 4 - Solución propuesta: se establece y justifica la resolución del problema siguiendo las fases de la metodología CRISP-DM, desde la comprensión del negocio hasta la implementación. Se presenta el plan del proyecto y se realizan los análisis de: riesgos y contingencias, costos y beneficios, FODA y un análisis de factibilidad técnica, operativa y legal.

Capítulo 5 - Evaluación y resultados experimentales: realiza una evaluación de la propuesta y muestran los resultados obtenidos.

Capítulo 6 - Conclusión y futuras líneas de investigación: hace hincapié en la experiencia y contribución que se logra mediante el desarrollo del proyecto y establecen futuras líneas de investigación.

Capítulo 7 - Bibliografía

Capítulo 8 - Anexo

2 Estado de la cuestión

En este capítulo se desarrollan y describen los conocimientos recogidos sobre e-commerce y minería de datos, usados para establecer la situación actual en la que se plantea y relaciona el problema

2.1 Antecedentes relacionados

A continuación se muestran artículos y noticias recogidas sobre los casos y aplicaciones relacionadas el problema.

- "Economía digital: una oportunidad para las pymes argentinas":

Los cambios tecnológicos si son buenos y mejoran la productividad en una empresa. "La digitalización es una realidad en muchas empresas argentinas que están aprovechando los cambios tecnológicos para mejorar su productividad" (Mayer & Szenkman, 2018, p. 1).

- "Amazon y el servicio de atención al cliente":

Amazon aplicó minería de datos para analizar, estudiar y trazar un historial del comportamiento de sus usuarios para adelantarse a sus posibles crisis con el fin de resolverlas y asegurar su fidelidad. Con esta información logró aplicar campañas de marketing y email marketing, como estrategia para la atención al cliente (Carrion, 2019).

- "Data Mining en la campaña de Obama":

Obama para su campaña política de marketing se basó en un análisis de datos de la población y de la competencia, implementando minería de datos para mejorar su posición. Logró identificar la población, el canal y el momento indicado en el que su mensaje tendría mayor repercusión en los votantes (Cultura CRM, 2017).

- "E-commerce: análisis de la cesta de la compra con Data Mining":

El análisis de una cesta de compra de por sí brinda información útil para las tiendas tanto físicas como online, pero con la minería de datos se logra un análisis y extracción del conocimiento de forma automática o semiautomática.

> La cesta de la compra es muy útil para determinar las tendencias del mercado en un momento concreto, además gracias a data mining podemos actuar casi en tiempo real, mejorando las ofertas. Del mismo modo, esta estrategia de análisis de datos nos ayudará a mejorar los procesos de producción y ajustar las necesidades de stock. (Cultura CRM, 2017)

De los artículos anteriores se concluye que junto al crecimiento de la tecnología se encuentran los avances e implementación de la minería de datos como solución para predecir, de un conjunto de datos, patrones que ayudan a la tomar decisiones gerenciales. Este campo se puede aplicar en diversos ámbitos de la vida en donde exista volumen de datos, como ser en medicina y farmacia, comercios, bancos, e-commerce.

2.2 E-commerce

E-commerce, también es conocido como comercio electrónico y se refiere al intercambio comercial de productos o servicios a través de medios digitales, con el fin de crecer en el mercado:

> El comercio electrónico se entiende como la actividad de compra de productos y servicios mediante el uso de medios electrónicos como el internet, aplicaciones móviles y otras redes informáticas (…) e-commerce tiene como objetivo: compartir información comercial, conseguir la fidelidad de los clientes y ampliar el mercado. (Olivier Peralta, 2019)

2.2.1 Cifras del E-commerce en Argentina

En Argentina existe una Asociación Civil sin fines de lucro conocida como Cámara Argentina de Comercio electrónico (CACE). CACE tiene como propósito promover el uso y desarrollo de tecnologías en todos los ámbitos, su objetivo es lograr que el comercio electrónico sea una herramienta para el desarrollo socioeconómico; y tiene por iniciativa: realizar estudios anuales de comercio electrónico, realizar eventos masivos de ventas (Hot Sale y el Cyber Monday) y brinda capacitaciones a emprendedores y pymes través del E-commerce Day (CACE, 2020).

Los E-commerce Day son una iniciativa de E-commerce Institute que se realizan en diferentes países de América Latina en forma conjunta con sus respectivas Cámaras de Comercio Electrónico para unir y sumar a las empresas del mundo de los negocios digitales (Instituto Latinoamericano de Comercio Electronico, 2020).

E-commerce Institute lleva 14 años fomentando el desarrollo de la industria del comercio electrónico en América Latina a través del Tour E-commerce Day con más de 118 eventos realizados en 18 países de la Región

El evento eCommerce DAY está enfocado en la importancia de Internet y las nuevas tecnologías.

> El Tour de Evento eCommerce DAY crea un espacio para la difusión, promoción y reflexión sobre la importancia del impacto que ha producido Internet y las nuevas tecnologías en la vida, trabajo y negocios de las personas y empresas, permitiendo una mejora en la competitividad de nuestra economía y la reducción de la brecha que separa a nuestros emprendedores, empresas y profesionales en diferentes regiones de América Latina. (Instituto Latinoamericano de Comercio Electronico, 2020)

En Argentina se realizó de forma virtual el evento eCommerce Day, el 27 de agosto del año 2020, durante el evento se difundieron varios temas de los cuales se rescatan los tres más importantes.

Primero, el comercio electrónico creció de forma abrupta en los últimos meses, advierte Pueyrredon (2020), citado en (CACE, 2020):

> El e-commerce sufrió una aceleración abrupta, creciendo a tasas anuales en pocos meses y la proyección es que esto siga creciendo (...) Es por eso que debemos seguir trabajando en profesionalizar la oferta (...) y es allí donde eventos como el eCommerce Day ofrecen

> una agenda de capacitación especializada y adecuada a las demandas actuales para profesionales de la industria.

Segundo, el director de la Cámara Argentina de Comercio electrónico Sambucetti (2019), citado en (CACE, 2020) informó, según los resultados del Estudio Anual eCommerce, que las ventas online aumentaron y provocaron una reactivación de la economía en Argentina:

> En los primeros 6 meses del 2020 se facturó un 106% más que en el mismo periodo 2019. Las órdenes de compra, las unidades vendidas y la cantidad de sesiones aumentaron significativamente en relación al año pasado, marcando un nuevo hito en el comercio electrónico y contribuyendo a la reactivación de la economía Argentina.

En base al dicho comentario y resultado estadístico, considero que un punto clave para la sociedad emprendedora es la digitalización, remarcando a e-commerce como la mejor elección para los negocios ya que hoy en día se ha posicionado como una elección para mitigar los riesgos, como los del contagio de covid.

Tercero, e-commerce AWARD's Argentina 2020 (evento que premia a las empresas de mayor reconocimiento en América Latina) destacó a Mercado Libre en la categoría de Mejor iniciativa e-commerce (CACE, 2020). Este reconocimiento, a nivel Latinoamérica, me permite tomar a Mercado Libre como un ejemplo de plataforma online a seguir para la venta de productos.

Los avances de la tecnología, internet y el creciente consumo en los productos o servicios ofrecidos por la web implican el almacenamiento de un gran volumen de datos, que se tornan inmanejable a la hora de analizarlos; lo mismo ocurre si los datos se encuentran almacenados en formato papel. Por esto es necesario una herramienta potente que descubra, de forma automática, información valiosa sobre los datos y su conocimiento organizado que aporten a la toma de decisión. En este sentido Jiawei Han, Micheline Kamber, & Jian Pei (2012) afirman que:

> Este crecimiento explosivo del volumen de datos disponibles es el resultado de la informatización de nuestra sociedad y el rápido desarrollo de poderosas herramientas de recopilación y almacenamiento de datos. Las empresas de todo el mundo generan conjuntos de datos gigantes, que incluyen transacciones de ventas, registros de negociación de acciones, descripciones de productos, promociones de ventas, perfiles y rendimiento de la empresa y comentarios de los clientes. Por lo que se necesitan con urgencia herramientas poderosas y versátiles para descubrir automáticamente información valiosa de las enormes cantidades de datos y transformar dichos datos en conocimiento organizado. Esta necesidad ha llevado al nacimiento de la minería de datos (p.2).

En conclusión, el crecimiento del comercio electrónico en Argentina y su contribución en la reactivación de la economía del país, nos asegura que es una muy buena herramienta a considerar e implementar en las empresas, pymes o emprendedores argentinos.

En Argentina aún existen negocios como FOER Deportes que registran sus datos en papel, por lo que es ahí en donde la informática entra en juego, en esta oportunidad con la minería de datos y el comercio electrónico, para aportar un valor agregado al automatizar procesos del negocio que permiten mejorar su productividad e incrementar sus ventas. La minería de datos

como una herramienta para analizar y descubrir automáticamente tendencias en la cesta de compra le permite realizar marketing selectivo, planificar el espacio en los estantes y realizar ventas cruzadas lo que conducirá a mejorar las ventas. Y la herramienta e-commerce como otro medio de ventas que refleja el conocimiento obtenido de la minería de datos mediante recomendación de productos y cross selling. El cross selling "consiste en ofrecer al cliente o posible cliente un producto complementario a su intención de compra principal, haya finalizado su compra o no" (Quintana, 2021).

2.3 Minería de datos

Define de modo sencillo Marqués (2014) a la minería de datos como: "Un conjunto de técnicas encaminadas al descubrimiento de la información contenida en grandes conjuntos de datos. Se trata de analizar comportamientos, patrones, tendencias, asociaciones y otras características del conocimiento inmerso en los datos" (p.3).

También, Orallo (2004) sostiene a la minería de datos como: "El proceso de extraer conocimiento no trivial, comprensible, previamente desconocidos y potencialmente útil desde grandes cantidades de datos almacenados en distintos formatos" (p.82). Los patrones descubiertos deberán ayudar a tomar decisiones más seguras que reporten beneficios a la organización.

Para entender el ambiente de la minería de datos se tomará dos enfoques:

1. Micro minería de datos
2. Macro de datos

2.3.1 Micro minería de datos

Usaremos la palabra micro minería de datos para referimos a una mirada interna en la definición de minería de datos, por lo que se detectan tres conceptos importantes: dato, información y conocimiento.

En este sentido Vallejos (2006) afirma que:

> Los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación de la información y ese modelo representan un valor agregado, entonces nos referimos al conocimiento (p.7).

Estos conceptos se encuentran organizados jerárquicamente de tal forma que su orden en la pirámide representa su volumen y el valor que los responsables de las decisiones le dan en esa jerarquía (Imagen 2-1).

Definimos:

- Datos: conjunto de valores discretos. Por ejemplo, lo ítems comprados.
- Información: datos procesados y que tienen significado. Por ejemplo, los clientes compran gaseosa y golosinas.

- Conocimiento: capacidad de convertir datos e información en acciones efectivas. Por ejemplo, de los clientes que compras gaseosa, el 66% compran golosinas.

Imagen 2-1: Jerarquía de dato, información y conocimiento.

2.3.2 Macro minería de datos

Se usará la palabra macro minería de datos para referimos al estudio de la minería de datos en su entorno.

La minería de datos es el núcleo de todo un proceso llamado Descubrimiento del Conocimiento a partir Base de Datos (KDD - Knowledge Discovery in Databases) este proceso es definido por Bramer (2016) como: "Extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos" (p. 2).

Si bien los términos minería de datos y descubrimiento de conocimiento en bases de datos son usados como sinónimos, KDD describe el proceso completo de extracción de conocimiento a partir de los datos, que además de la obtención de los modelos o patrones como objetivo de minería de datos, incluye su evaluación e interpretación.

Las metas de KDD según Vallejos (2006) son:

1. Procesar automáticamente grandes cantidades de datos crudos.
2. Identificar los patrones más significativos y relevantes.
3. Presentarlos como conocimiento apropiado para satisfacer metas del usuario.

KDD consta de una secuencia ordenadas de tareas predefinidas (Imagen 2-2), de las que Bramer (2016) explica:

> Los datos llegan, posiblemente de muchas fuentes. Están integrado y colocado en algún almacén de datos común. Luego, se toma una parte y se procesa previamente en un formato estándar. Estos datos preparados se pasan luego a un algoritmo de minería de datos que produce una salida en forma de reglas o algún otro tipo de patrones. Luego, se interpretan para dar conocimientos nuevos y potencialmente útiles. (p. 3)

Lo principales pasos de este proceso iterativo KDD son:

- Etapa de recopilación: en la mayoría de los casos la información se encuentran en diferentes fuentes o bases de datos, tanto internas como externas. Su recopilación consiste en seleccionar los datos requeridos para cumplir con objetivo del negocio. Las primeras etapas son las que determinan que las siguientes sean capaces de extraer conocimiento valido y útil de la información original.

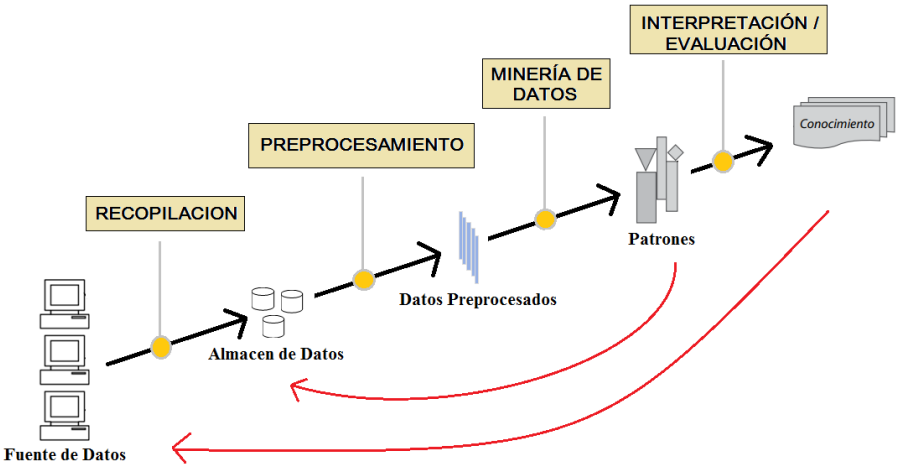


Imagen 2-2: Proceso de descubrimiento del conocimiento en Base de Datos (KDD).

- Etapa de pre-procesamiento: esta etapa se realiza por que en la mayoría de los casos los datos están sucios, por el simple hecho de que provienen del mundo real y de distintas fuentes. Que estén sucios quiere decir que se encuentran:
  - Incompletos: faltan valores en los atributos, faltan atributos interesantes o solo contienen datos agregados (ocupación= “”).
    Los datos incompletos pueden provenir de: errores (humanos, hardware o software), al cargar los datos no se consideraron que eran importantes, criterios diferentes entre el momento de la recogida de datos y el de su análisis.
  - Con ruido (valores incorrectos): contienen errores u outliers (salario = “-10”).
    Los datos con errores pueden provenir de: instrumentos defectuosos de recogida de datos, errores humanos o de programas de carga de datos, errores en la transmisión de datos. Los outliers son valores que están fuera de rango.
  - Inconsistentes: contienen discrepancia en códigos y nombres. (Valoración en un sitio es “1, 2,3”, en otro “A, B, C”).
    Los datos inconsistentes pueden provenir de: fuentes de datos diferentes, violación de dependencias funcionales (ejemplo: modificación de datos vinculados)

  Las tareas principales del pre-procesamiento son:
  - Limpieza de los datos:
    - Resolver datos faltantes: estos valores faltantes se ignoran, se reemplazan por un valor por omisión, o por el valor más cercano, es decir, se usan métricas de tipo estadístico como media, moda, mínimo y máximo para reemplazarlos. Se puede completar manualmente, automáticamente o ignorar todo el registro. También se puede eliminar el atributo columna cuando la proporción de nulos en ella es muy grande.

        - Resolver datos con ruido: el ruido es un error aleatorio en una variable medida que se soluciona eliminando el atributo, la instancia o suavizar ajustando los datos a las funciones de regresión.
        - Identificar o eliminar outliers.
        - Corregir datos inconsistentes.
        - Resolver redundancias causadas por la integración de los datos.
    - Integración de los datos:
        - Combinar datos de múltiples fuentes en un solo lugar, hacia la vista minable. La vista minable recoge toda la información necesaria (y sólo esa) para realizar la minería de datos. Es una tabla, la mayoría de los algoritmos solo pueden procesar una tabla.
        - Resolver problemas de representación y codificación.
        - Integrar los datos cuidadosamente desde diferentes tablas para crear información homogénea, lo que puede ayudar a evitar redundancia e inconsistente, mejorar la calidad y rapidez de la minería de datos.
    - Transformación de los datos: se buscan características útiles para representar los datos de manera tal de hacerlos usables y navegables, dependiendo del objetivo del trabajo.
        - Combinar dos o más atributos en un solo atributo
        - Conversión de atributos: algunos métodos (redes neuronales, regresión, vecino más próximo) precisan que los atributos sean numéricos
        - Normalización de los datos: "su objetivo principal es asociar formas similares a los mismos datos en una única forma de datos" (PowerData, 2017).
        - Desratización: lograr que los valores seas discretos ya que algunos algoritmos lo requieren así. Es muy útil para resumir los datos, y se puede lograr mediante particiones con el mismo ancho: divide el rango en N intervalos del mismo ancho, si A y B son el mínimo y máximo valor del atributo, ancho es $(B - A)/N$. También mediante particiones con la misma altura: divide el rango en N intervalos con aproximadamente el mismo número de elementos.
    - Reducción de los datos: obtiene una representación reducida del volumen pero que produce iguales o similares resultados analíticos. Es importante ya que algunos métodos de minería de datos pueden demorar mucho tiempo al aplicar el conjunto de datos. Existe estrategias para la reducción de datos, como ser:
        - Reducción de la dimensionalidad: puede ser horizontal o vertical:
            - Reducción horizontal: implica la eliminación de tuplas idénticas
            - Reducción vertical: implica la eliminación de atributos que son insignificantes o redundantes con respecto al problema, como la eliminación de columnas que dependen funcionalmente (por ejemplo, edad y fecha de nacimiento).
        - Reducción mediante muestreo de datos.

- Etapa de minería de datos: una vez obtenida la vista minable en la etapa anterior, se realiza el proceso de exploración y análisis por medios automáticos o semiautomáticos, de esa vista

obtenida; con el fin de descubrir patrones significativos aplicando técnicas de minería de datos.

- Etapa de interpretación y evaluación: se evalúan los patrones obtenidos para comprobar que tienen la calidad suficiente para poder realizar la interpretación y obtener conocimiento. Dicho conocimiento descubierto se consolida para poder incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto.

2.3.3 Modelos de minería de datos

- Modelo predictivo: "Pretenden estimar valores futuros o desconocidos de variables de interés (variables dependientes) usando otras variables (variables independientes)" (Orallo, 2004, p. 21). Por ejemplo: estimar la demanda de un nuevo producto en función de los gastos de publicidad

- Modelo descriptivo: "Exploran las propiedades de los datos para explicar o resumirlos" (Orallo, 2004, p. 21). Por ejemplo: determinar grupos de viajeros con intereses similares

2.3.4 Aprendizajes en minería de datos

Bramer (2016) afirma que:

> En general tenemos un conjunto de datos de ejemplos (llamado instancias), cada uno de los cuales comprende los valores de una serie de variables, que en la minería de datos a menudo se denominan atributos. Hay dos tipos de datos que se tratan radicalmente diferentes. (p.4)

- Aprendizaje supervisado, Bramer (2016) explica que:

  > Hay un atributo especialmente designado y el objetivo es utilizar los datos proporcionados para predecir el valor de ese atributo para instancias que aún no se han visto. Los datos de este tipo se denominan etiquetado. Si usamos los datos de esta forma, los datos se conocen como aprendizaje supervisado en minería de datos
  >
  > Si el atributo designado es categórico, es decir, debe tomar uno de varios valores distintos como 'muy bueno', 'bueno' o 'malo', o 'coche', 'bicicleta', 'persona', 'autobús' o 'taxi' la tarea se llama clasificación. Si el atributo designado es numérico, por ejemplo, el precio de venta esperado de una casa o el precio de apertura de una acción en el mercado de valores de mañana, la tarea se llama regresión. (p. 5)

- Aprendizaje no supervisado, Bramer (2016) explica que:

  > Los datos que no tienen ningún atributo especialmente designado se denominan sin etiquetar. La extracción de datos no etiquetados se conoce como aprendizaje sin supervisión. Aquí el objetivo es simplemente extraer la mayor cantidad de información que podamos de los datos disponibles. (p. 5)

2.3.5 Tareas de minería de datos

Existen diversas tareas dependiendo del modelo del que se trate.

- Modelo predictivo, tareas de minería de datos:

- Clasificación: la clasificación genera una organización entre los registros o instancias de una base una base de datos, considerando que cada registro tiene una etiqueta asociada al valor de un atributo que llamamos la clase de una instancia. El objetivo de este modelo es asignar una clase a registros no vistos previamente con tanta precisión como sea posible. De esta forma se clasifican los elementos en diferentes grupos. Es importante destacar que el algoritmo no es capaz de determinar a qué grupo pertenece un valor sino más bien logra relacionar características con etiquetas y así obtener un resultado.
- Regresión: la clasificación es una forma de predicción donde el valor a predecir es una etiqueta; otro caso es la predicción numérica, más conocida como regresión. En este caso deseamos predecir un valor numérico, como los beneficios de una empresa o el precio de una acción (Bramer, 2016).

- Modelo descriptivo, tareas de minería de datos:
  - Agrupamiento: Dado un conjunto de objetos, cada uno con un conjunto de atributos, y una medida de similaridad entre ellos, encontrar clústeres (agrupamientos); tales que los objetos del mismo grupo son similares entre sí y muy diferentes a los objetos de otros grupos.
    Se diferencia de la clasificación al hablar de grupos y no de clases, que analiza los datos para generarles una etiqueta; la clasificación analiza los datos etiquetados con una clase.
    Se usa, por ejemplo para la segmentación del mercado: recoger atributos de clientes basados en información geográfica o estilo de vida; encontrar grupos de clientes similares, etc.
  - Asociación: dado un conjunto de registros, cada uno con elementos de alguna colección dada, produce reglas de dependencia que predigan la ocurrencia de un elemento basado en la ocurrencia de otros elementos. Tiene como objetivo identificar relaciones entre atributos categóricos. Hernández Orallo (2004) afirma que: "Este tipo de tarea se utiliza frecuentemente en el análisis de la cesta de la compra, para identificar productos que son comprados juntos por un numero suficientemente grande de clientes, gestionar estante, promociones de marketing y ventas en un supermercado". (p.45)

A las técnicas, tareas, tipos de modelos y aprendizajes de minera de datos, se las puede organizar de forma jerárquica en un diagrama (Imagen 2-3) para comprender su distribución y que permita seleccionarlos adecuadamente a la hora de solucionar un problema con minería de datos.

2.3.6 Herramientas de minería de datos

Las herramientas software de minería de datos más usadas entre el año 2012 y 2013, según los resultados de una encuesta realizada por el prestigioso portal de internacional de minería de datos KDnuggets (Imagen 2-4), RapidMiner en su versión gratuita obtuvo el primer lugar, seguido de R, Excel, Weka y Python; por lo que continuación se detallan las principales características, ventajas y desventajas de cada una.

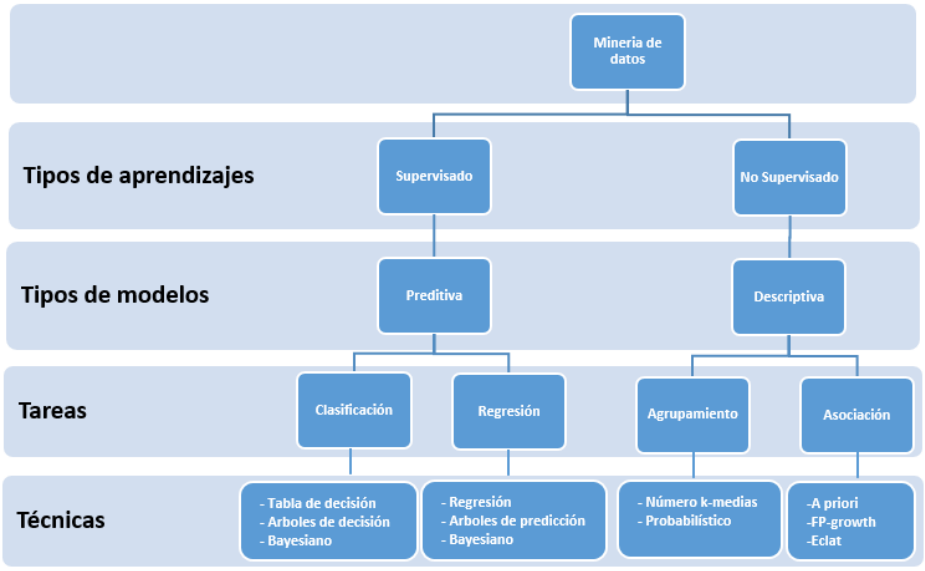


Imagen 2-3: Técnicas, tareas, tipos de modelos y aprendizajes en minería de datos.

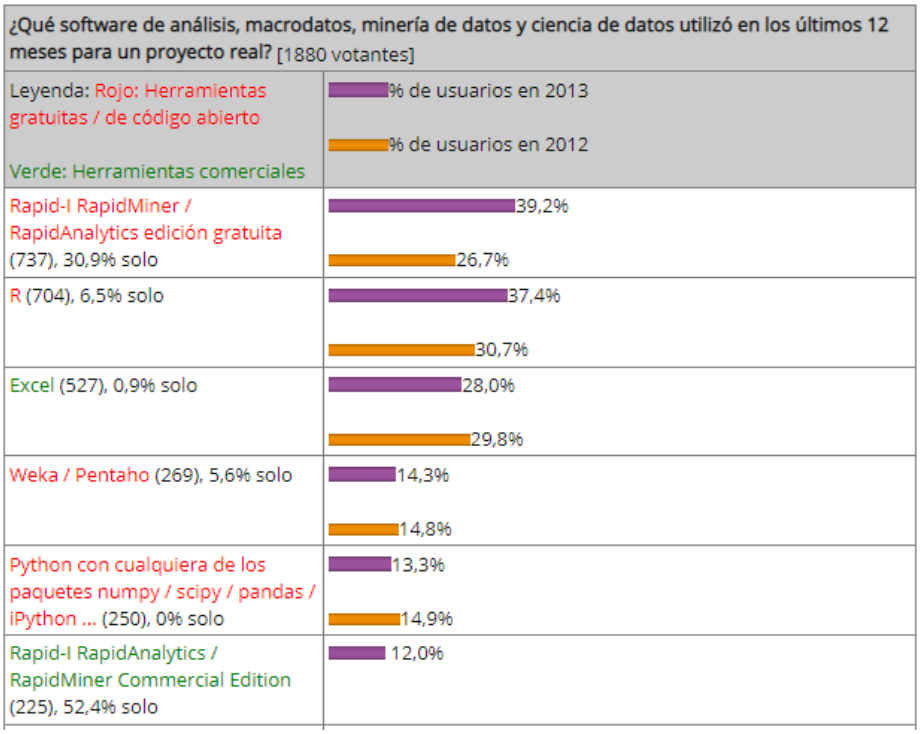


Imagen 2-4: Encuesta sobre el uso de las herramientas de minería de datos. Fuente: https://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html

- RapidMiner Studio:

RapidMiner es una plataforma unificada para la preparación de datos, el aprendizaje automático y la implementación de modelos. Permite el desarrollo de los modelos predictivos, modelos descriptivos, transformación de datos, series de tiempo, etc. Tiene un entorno de programación visual, fácil de usar, mediante el encadenamiento de operadores a través de un entorno gráfico e integrado para el aprendizaje automático, minería de datos, minería de texto, análisis predictivos y análisis de negocio, incorporando extracción, transformación y carga de datos, como así también presentación de informes de predicción. Es multiplataforma, se puede

ejecutar tanto en Windows como en Linux y maneja fuentes de datos como Excel, Access, Oracle, Microsoft SQL Server, MySQL, archivos de texto y otros (RapidMiner, 2020).

RapidMiner se utiliza para realizar análisis de minería de datos en aplicaciones empresariales, gobierno y academias, reconociendo todos los pasos del proceso de minería de datos incluyendo visualización, validación y optimización de los resultados.

Cuenta con la versión RapidMiner Studio 9.2 de código abierto que proporciona licencias de productos gratuitos para estudiante, profesores e investigadores.

- R:

Es un entorno de software libre bajo licencia GNU y un lenguaje de programación con un enfoque hacia el análisis estadístico. Uno de los más usados en investigación científica y popular en los campos de aprendizaje automático , minería de datos, investigación biomédica, bioinformática y matemáticas financieras. Incluye modelos lineales y no lineales, óptimo para la clasificación y el agrupamiento de los datos, facilitando así su posterior interpretación (R-project, 2020).

Es un lenguaje bastante adecuado para la estadística, ya que permite manipular los datos rápidamente y de forma precisa. Se puede automatizar fácilmente, gracias a la creación de scripts. Es compatible con lenguajes de programación como Fortran, C o C++ y funciona con sistemas UNIX, Windows y MacOS.

R no soporta gráficos en tres dimensiones o dinámicos, por lo que el resultado de algún informe puede ser algo pobre visualmente y bastante anticuado. Los algoritmos no están unificados; cada uno de ellos se almacena en un paquete distinto, por lo que hay que ir cambiando de opción para leer los datos obtenidos.

Aprender a manejarlo lleva bastante tiempo y no siempre es fácil de alcanzar este objetivo, más aún si se trata de personas sin conocimiento previo. Su actualización constante conlleva tener que aprender las características de cada apartado continuamente.

Sus restricciones y su incompatibilidad con otros lenguajes de programación impiden que se pueda usar para crear aplicaciones web.

Los datos ocupan demasiado espacio en la memoria física al ser almacenado en una única carpeta por defecto. Por lo que es necesario volcar los datos de forma periódica para evitar el colapso de los dispositivos de almacenamiento. R no tiene medidas de seguridad.

- Python:

Se trata de un lenguaje de programación multi-paradigma, ya que soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma (Python, 2020).

Python tiene numerosas librerías, se integra con otras aplicaciones como MongoDB, es un lenguaje de programación de uso general muy popular, general, fácil de usar y con una rápida

curva de aprendizaje junto con su versatilidad. Cuenta con una amplia gama de módulos específicos y soporte comunitario.

Para los fines específicos de análisis estadístico y de datos, la amplia gama de paquetes de R le da una ligera ventaja sobre Python. Es un lenguaje de tipo dinámico, lo que significa que debemos ser muy cuidadosos.

- Weka:

Weka es un conjunto de librerías JAVA para la extracción del conocimiento desde bases de datos y resolver problemas de minería de datos. Es una herramienta libre distribuida bajo licencia GNU-GPL, escrita en Java y se ejecuta en casi cualquier plataforma.

Posee una interfaz de usuario gráfica, así como también puede ser accedida mediante una interfaz de línea de comandos o una interfaz de programación de aplicaciones de Java.

Con Weka se puede resolver una amplia variedad de tareas de minería de datos especialmente pre procesamiento de datos, regresión, clasificación, clustering, asociación y visualización, sin necesidad de codificar.

2.4 Minería de reglas de asociación

La minería de reglas de asociación es un tipo de minería de datos que busca, entre un gran conjunto de datos transaccionales, relaciones frecuentes de conjuntos de elementos que conducen a descubrir asociaciones y correlaciones interesantes.

En este trabajo describiremos este tipo de tarea de la minería de datos por ser la adecuada al tipo de problema que se plantea resolver.

Para Jiawei, el descubrimiento de asociaciones interesantes ayuda a la toma de decisión comercial, al marketing y al análisis del comportamiento de clientes: “El descubrimiento de relaciones de correlación interesantes puede ayudar en muchos procesos de toma de decisiones comerciales, como el diseño de catálogos, el marketing cruzado y el análisis del comportamiento de compra del cliente” (Jiawei Han, Micheline Kamber, & Jian Pei, 2012, p. 244).

- Análisis de la cesta de mercado:

El análisis de la canasta de mercado es un ejemplo típico de este tipo de minería que analiza los hábitos de compra de los clientes al encontrar asociaciones entre los artículos que el cliente coloca en su canasta de compra, esta información sobre los artículos que los clientes compran juntos con frecuencia puede conducir a un aumento de las ventas al ayudar a los minoristas a realizar un marketing selectivo y planificar su espacio en los estantes.

Pensemos en que cada artículo tiene una variable booleana que representa la presencia o ausencia de ese artículo, por lo que cada canasta se representa como un vector booleano. De esta forma los vectores booleanos pueden analizarse en busca de patrones de compra que reflejan artículos que se compran juntos. Estos patrones se pueden representar en forma de reglas de asociación. Por ejemplo, la información de que los clientes que compran zapatillas

también tienden a comprar medias al mismo tiempo se representa en la siguiente regla de asociación:

Zapatilla ⇒ Media [soporte = 2%, confianza = 60%]

- Un soporte del 2% significa que el 2% de todas las transacciones bajo análisis muestran que la zapatilla y la media se compran juntas.
- Una confianza del 60% significa que el 60% de los clientes que compraron zapatilla también compraron media

El soporte y la confianza (medidas de interés de la regla) reflejan la utilidad y la certeza de la regla descubierta.

- Conjuntos de elementos frecuentes y reglas de asociación:

A un conjunto de elementos se lo denomina itemsets, un itemset que contiene k artículos es un k-itemset. Por ejemplo el conjunto {zapatilla, media} es un conjunto de 2 elementos. La frecuencia de ocurrencia de un itemset es el número de transacciones que contienen el conjunto de elementos, esto se conoce como la frecuencia o conteo de soporte. El conjunto k-itemsets frecuentes se denota comúnmente por L_K.

Jiawei Han, Micheline Kamber y Jian Pei (2012) advierten que la minería de reglas de asociación puede verse como un proceso de dos pasos, primero encontrar conjuntos de elementos frecuentes y segundos generar reglas de asociación:

> Primero, encuentra todos los conjuntos de elementos frecuentes: por definición, cada uno de estos conjuntos de elementos ocurrirá al menos con la misma frecuencia que un recuento mínimo de soporte predeterminado, min_sup. Segundo, genere reglas de asociación sólidas a partir de conjuntos de elementos frecuentes: por definición, estas reglas deben satisfacer un apoyo mínimo y una confianza mínima. (p. 247)

Sea:

- I = $\{I_1, I_2, \ldots, I_m\}$ un conjunto de elementos.
- D, los datos relevantes para la tarea, es decir un conjunto de transacciones de base de datos donde cada transacción T (asociada con un identificador TID) es un conjunto de elementos no vacío tal que T ⊆ I.
- A = $\{A_1, A_2, \ldots, A_m\}$ un conjunto de elementos. Una transacción T se dice que contiene A si A ⊆ T.

Una regla de asociación es una implicación de la forma A ⇒ B, dónde A ⊂ I, B ⊂ I, A≠∅, B≠∅, y A ∩ B = φ.

- La regla A ⇒ B se mantiene en el conjunto de transacciones D con soporte s, dónde s es el porcentaje de transacciones en D que contienen A ∪ B. Esta se toma como la probabilidad, P(A ∪ B).
- La regla A ⇒ B tiene confianza C en el conjunto de transacciones D, dónde C es el porcentaje de transacciones en D conteniendo A que también contienen B. Este es tomado como la probabilidad condicional, P (B | A).

Es decir:

$$Soporte\ (A \Rightarrow B) = P(A\ U\ B)$$

$$Confianza\ (A \Rightarrow\ B) = P(B \mid A) = \frac{Soporte\ (\text{A} \cup \text{B})}{Soporte\ (A)}$$

Normalmente las reglas se consideran interesantes si satisfacen un umbral mínimo de soporte y confianza, establecido por el usuario o experto en el dominio, por convención estos valores se escriben para que ocurran entre 0% y 100%, en lugar de 0 a 1.0.

Se pueden aplicar medidas de interés adicionales para el descubrimiento de relaciones de correlación entre elementos asociados, debido a que el segundo paso es menos costoso; el primero determina el desempeño general de las reglas de asociación minera.

Un desafío importante en la extracción de conjuntos de elementos frecuentes de un gran conjunto de datos es el hecho de que dicha extracción a menudo genera una gran cantidad de conjuntos de elementos que satisfacen el soporte mínimo, especialmente cuando se establece bajo. Esto se debe a que si un conjunto de elementos es frecuencia, cada uno de sus subconjuntos es frecuente también. Un conjunto de elementos largo contendrá un número combinatorio de subconjuntos de elementos más cortos y frecuentes.

2.5 Técnicas de reglas de asociación

Dependiendo del tipo de tarea de minería de datos se elige la técnica que la resuelva, para el proyecto se requiere de la tarea de asociación.

En este sentido en minería de reglas de asociación primero se deben encontrar los conjuntos de elementos frecuentes, para ello existen algoritmos a priori, algoritmos basados en patrones frecuentes de crecimiento y algoritmos que utilizan el formato de datos vertical:

> Encontrar primero conjuntos de elementos frecuentes (conjuntos de elementos como A y B, satisfaciendo un umbral mínimo de soporte), de las cuales se generan fuerte reglas de asociación de la forma de A ⇒ B. Estas reglas también satisfacen un umbral mínimo de confianza (una probabilidad pre especificada de satisfacer B bajo la condición de que A se satisfaga). Las asociaciones se pueden analizar más a fondo para descubrir reglas de correlación, que transmiten correlaciones estadísticas entre conjuntos de elementos A y B. Se han desarrollado muchos algoritmos eficientes y escalables para minería frecuente de conjuntos de elementos, de las cuales se pueden derivar reglas de asociación y correlación. Estos algoritmos se pueden clasificar en tres categorías: algoritmos a priori, algoritmos basados en patrones frecuentes de crecimiento, como FP-Growth, y algoritmos que utilizan el formato de datos vertical como Eclat. *(Jiawei Han, Micheline Kamber, & Jian Pei, 2012, p. 272)*

- Algoritmo a priori:

A priori es un algoritmo fundamental para extraer conjuntos de elementos frecuentes para las reglas de asociación booleanas y fue propuesto por R. Agrawal y R. Srikant en 1994. Su nombre se debe a que utiliza conocimiento previo de las propiedades frecuentes del conjunto de elementos.

Del conjunto de datos se toma el conjunto de un elemento y se le realiza la prueba de los subconjuntos, podando los elementos que no satisfacen el soporte mínimo. En cada etapa se suma una longitud de elemento al conjunto anterior y el proceso se repite hasta que no se puedan encontrar conjuntos de elementos frecuentes:

> Este algoritmo emplea un enfoque iterativo conocido como búsqueda por niveles, en el que se utilizan conjuntos de elementos k para explorar conjuntos de elementos (k + 1). Primero, el conjunto de conjuntos de 1 elemento frecuentes se encuentra escaneando la base de datos para acumular el recuento de cada elemento y recopilando aquellos elementos que satisfacen el soporte mínimo. El conjunto resultante se denota por L_1. A continuación, L_1 se usa para encontrar L_2 (el conjunto de conjuntos de 2 elementos frecuentes) que se usa para encontrar L_3, y así sucesivamente, hasta que no se puedan encontrar conjuntos de *k* elementos más frecuentes. El hallazgo de cada L_k requiere un escaneo completo de la base de datos. *(Jiawei Han, Micheline Kamber, & Jian Pei, 2012, p. 249)*

Para mejorar la generación de conjuntos de elementos frecuentes por niveles, se utiliza la llamada Propiedad A priori para reducir el espacio de búsqueda: todos los subconjuntos no vacíos de un conjunto de elementos frecuentes también deben ser frecuentes.

Por definición la propiedad se basa en lo siguiente: si un conjunto de elementos *I* no satisface el umbral de soporte mínimo (*min_sup*) entonces I no es frecuente, es decir, $P(I) < min\_sup$.Si se agrega un elemento *A* al conjunto de elementos *I*, entonces el conjunto de elementos resultante (es decir, $I \cup A$) no puede ocurrir con más frecuencia que *I*. Por lo tanto, $I \cup A$ tampoco es frecuente, es decir $P(I \cup A) < min\_sup$.

Esta propiedad pertenece a una categoría especial de propiedades llamada antimonotonicidad en el sentido de que si un conjunto no puede pasar una prueba, todos sus súper conjuntos también fallarán la misma prueba. Se llama antimonotonicidad porque la propiedad es monótona en el contexto de fallar una prueba.

Existen variaciones del algoritmo a priori que pueden reducir el número de escaneos de los datos:

> Se pueden usar variaciones que involucran hash y reducción de transacciones para hacer que el procedimiento sea más eficiente. Otras variaciones incluyen particionar los datos (minar en cada partición y luego combinar los resultados) y muestrear los datos (minar en un subconjunto de datos). Estas variaciones pueden reducir el número de escaneos de datos necesarios a tan solo dos o incluso uno. (Jiawei Han, Micheline Kamber, & Jian Pei, 2012, p. 272)

- FP-Growth:

Crecimiento de patrones frecuentes, FP-Growth o FP-crecimiento es un algoritmo propuesto por los investigadores Han et. al en el 2000.

Si bien el método de generación y prueba de candidatos a priori posee un buen rendimiento al reducir significativamente el tamaño de los conjuntos de candidatos, puede sufrir el costo de

generar una gran cantidad de conjuntos candidatos y el costo de escanear la base de datos varias veces:

> Es posible que aún deba generar una gran cantidad de conjuntos de candidatos y escanear repetidamente toda la base de datos y verificar un gran conjunto de candidatos mediante la coincidencia de patrones. Es costoso repasar cada transacción en la base de datos para determinar el soporte de los conjuntos de elementos candidatos. (Jiawei Han, Micheline Kamber, & Jian Pei, 2012, p. 257)

Por lo que el posterior diseño del método FP- Growth extrae el conjunto completo de conjuntos de elementos frecuentes sin un costoso proceso de generación de candidatos, comprimiendo la base de datos original en un compacto árbol FP y adoptando la estrategia divide y vencerás:

> Primero, comprime la base de datos que representa elementos frecuentes en un árbol de patrones frecuentes, o FP-árbol, que retiene la información de asociación del conjunto de elementos. Luego divide la base de datos comprimida en un conjunto de bases de datos condicionales (un tipo especial de base de datos proyectada), cada uno asociado con un elemento frecuente o "fragmento de patrón", y extrae cada base de datos por separado. Para cada "fragmento de patrón", solo es necesario examinar sus conjuntos de datos asociados. Por lo tanto, este enfoque puede reducir sustancialmente el tamaño de los conjuntos de datos que se buscarán, junto con el "crecimiento" de los patrones que se examinan. *(Jiawei Han, Micheline Kamber, & Jian Pei, 2012, p. 257)*

Según los investigadores la generación de bases de datos condicionales pequeñas permite la búsqueda de patrones frecuentes más cortos de forma recursiva lo que reduce los costos de búsqueda, considerando bases de datos proyectadas cuando la base de datos original es muy grande:

> El método de crecimiento FP transforma el problema de encontrar patrones frecuentes largos en buscar patrones más cortos en bases de datos condicionales mucho más pequeñas de forma recursiva y luego concatenando el sufijo. Utiliza como sufijo los elementos menos frecuentes, ofreciendo una buena selectividad. El método reduce sustancialmente los costes de búsqueda. Cuando la base de datos es grande, a veces no es realista construir un árbol FP principal basado en memoria. Una alternativa interesante es dividir primero la base de datos en un conjunto de bases de datos proyectadas y luego construir un árbol FP y extraerlo en cada base de datos proyectada. Este proceso se puede aplicar de forma recursiva a cualquier base de datos proyectada si su árbol FP todavía no puede caber en la memoria principal. *(Jiawei Han, Micheline Kamber, & Jian Pei, 2012, p. 259)*

- Eclat:

Los conjuntos de elementos frecuentes extraídos por los algoritmos anteriores provienen de un conjunto de transacciones de una base de datos en formato horizontal, es decir {TID: itemset} dónde TID es un ID de transacción y itemset es el conjunto de artículos comprados en la transacción TID. Los datos y también se pueden presenten en una base de datos con formato vertical por lo que la extracción de los conjuntos de elementos frecuentes se obtiene de manera eficiente con el uso del algoritmo Eclat. La idea de este algoritmo es usar la intersección de los

ID de transacciones de los k ítems para calcular los TID_set de los correspondientes (k+1) itemsets, este proceso se realiza de forma recursiva hasta que no se pueden encontrar conjunto de elementos frecuentes o candidatos:

> Primero, transformamos los datos formateados horizontalmente al formato vertical escaneando el conjunto de datos una vez. El recuento de soporte de un conjunto de elementos es simplemente la longitud del TID_set del conjunto de elementos. Comenzando con k = 1, los k-itemsets frecuentes se pueden usar para construir el candidato (k + 1)-itemsets basado en la propiedad A priori. El cálculo se realiza mediante la intersección de los TID_sets de los k-itemsets frecuentes para calcular los TID_sets de los correspondientes (k + 1) -itemsets. Este proceso se repite, con k incrementado en 1 cada vez, hasta que no se pueden encontrar conjuntos de elementos frecuentes o conjuntos de elementos candidatos. (Jiawei Han, Micheline Kamber, & Jian Pei, 2012, p. 261)

Este algoritmo además de aprovechar la propiedad a priori, reduce costos al no considerar necesario el escaneo de la base de datos para encontrar el soporte de los (k+1) itemsets para $k \geq 1$y al usar una técnica llamada diffset para reducir los costos de registrar los conjuntos TID largos:

> Además de aprovechar la propiedad Apriori en la generación de candidatos (k + 1) - itemsets frecuentes k- itemsets, otro mérito de este método es que no es necesario escanear la base de datos para encontrar el soporte de (k + 1) - itemsets (para k ≥ 1). Esto se debe a que el conjunto TID de cada k- itemset contiene la información completa requerida para contar dicho soporte. Sin embargo, los conjuntos de TID pueden ser bastante largos, ocupando un espacio de memoria sustancial, así como tiempo de cálculo para la intersección de los conjuntos largos. Para reducir aún más el costo de registrar conjuntos TID largos, así como los costos posteriores de las intersecciones, podemos usar una técnica llamada diffset, que realiza un seguimiento de solo las diferencias de los conjuntos TID de un (k + 1) -itemset y un correspondiente k- itemset. (Jiawei Han, Micheline Kamber, & Jian Pei, 2012, p. 262)

2.6 Generación de reglas de asociación

Una vez extraído, de una base de datos de transacciones, el conjunto de elementos frecuentes que satisfacen el soporte mínimo, generar las reglas de asociación fuertes a partir de ellos. Recordando que las reglas de asociación fuerte son las que satisfacen los valores mínimos de soporte y confianza, y que se pueden calcular usando la ecuación para la confianza:

$$Confianza\ (A \Rightarrow\ B) = P(B \mid A) = \frac{Soporte\ (\mathrm{A} \cup \mathrm{B})}{Soporte\ (A)}$$

Las reglas de asociación se pueden generar de la siguiente manera:

- Para cada conjunto de elementos frecuentes *l*, genere todos los subconjuntos no vacíos de *l*.
- Para cada subconjunto no vacío *s* de *l*, genere la regla:

$$s \Rightarrow (l - s) \quad if \quad \frac{soporte\ (l)}{soporte\ (s)} \geq \min\_conf$$

A continuación proponemos un ejemplo de generación de reglas de asociación, basado en una lista de ID de transaccionales de artículos:

TID	Lista de ID de articulo
T100	*l1, l2, l5*
T200	*l2, l4*
T300	*l2, l3*
T400	*l1, l2, l4*
T500	*l1, l3*
T600	*l2, l3*
T700	*l1, l3*
7800	*l1, l2, l3, l5*
T900	*l1, l2, l3*

Tabla 2-1: Ejemplo de lista de ID de artículos.

Mediante la aplicación de un algoritmo, este conjunto de datos contiene elementos frecuentes $X = \{l_1, l_2, l_5\}$.

Los subconjuntos no vacíos (s) de *X* son $\{l_1, l_2\}$, $\{l_1, l_5\}$, $\{l_2, l_5\}$, $\{l_1\}$, $\{l_2\}$ y $\{l_5\}$

Las reglas de asociación que se pueden generar a partir de X son las siguientes:

$\{l_1, l_2\} \Rightarrow l_5$, *confianza = 2/4 = 50%*
$\{l_1, l_5\} \Rightarrow l_2$, *confianza = 2/2 = 100%*
$\{l_2, l_5\} \Rightarrow l_1$, *confianza = 2/2 = 100%*
$l_1 \Rightarrow \{l_2, l_5\}$, *confianza = 2/6 = 33%*
$l_2 \Rightarrow \{l_1, l_5\}$, *confianza = 2/7 = 29%*
$l_5 \Rightarrow \{l_1, l_2\}$, *confianza = 2/2 =100%*

Si proponemos 70% como umbral mínimo de confianza, las reglas fuertes admitidas son: la segunda, tercera y última.

El soporte y la confianza son medidas de interés objetivas que utilizan la mayoría de los algoritmos como un paso para eliminar las reglas poco interesantes, aun así entra en juego el interés del usuario sobre las reglas presentadas. Para saber qué reglas son realmente interesantes existen medidas alternativas de correlación entre los conjuntos de elementos A y B. que ayudan a extraer relaciones de datos interesantes. Estas medidas de correlación llevan a reglas de correlación de la forma:

$$A \Rightarrow \text{B [} \textit{soporte, confianza, correlación} \text{]}$$

La medida de correlación más simple es el Lift, quien nos indica que la ocurrencia de itemset A es independiente de la ocurrencia de itemset B si *P(A ∪ B) = P (A) P (B)*; de lo contrario, los itemset A y B están dependientes y correlacionados como eventos. El lift entre la ocurrencia de *A* y *B* puede ser medido por computación:

$$lift\ (A, B) = \frac{P\ (A \cup B)}{P(A)P(B)}$$

El valor resultante de la ecuación puede clasificarse:

- Si lift $(A \rightarrow B) > 1$, A y B están correlacionadas positivamente, lo que significa que la ocurrencia de uno implica la ocurrencia del otro.
- Si lift $(A \rightarrow B) < 1$, la ocurrencia de A esta correlacionado negativamente con la ocurrencia de B, lo que significa que la ocurrencia de uno probablemente conduce a la ausencia del otro.
- Si lift $(A \rightarrow B) = 1$, A y B están independientes y no existe correlación entre ellos.

El lift de la regla de asociación (o correlación) A ⇒ B y es equivalente a:

$$Lift\ (A \rightarrow B) = \frac{Confianza(A \rightarrow B)}{Soporte(B)}$$

2.7 Metodologías de minería de datos

En la imagen 2-5 se observan los resultados de una encuesta realizada en Octubre de 2014 por KDnuggets (portal internacional de minería de datos), sobre las metodologías usadas para proyectos de minería de datos.

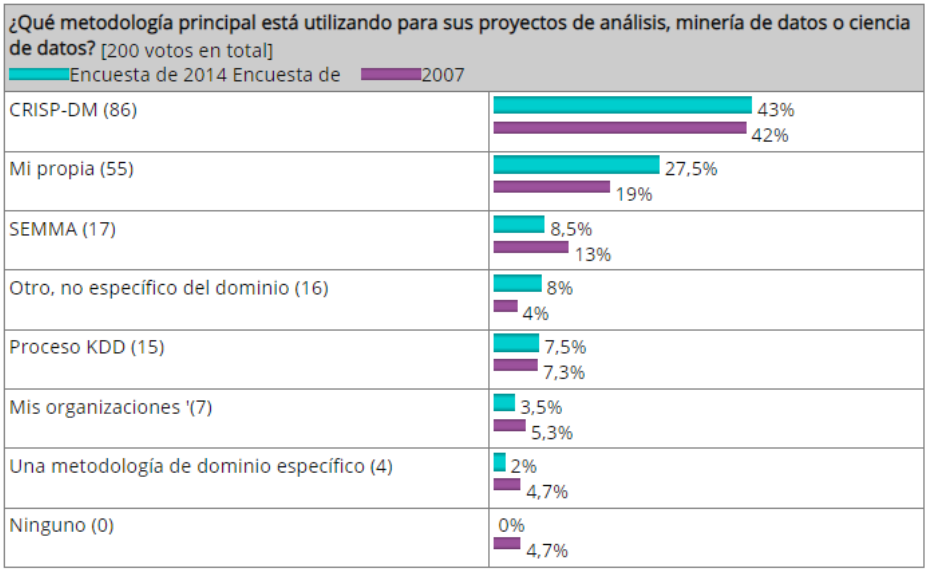


Imagen 2-5: Encuesta sobre el uso de las metodologías para minería de datos. Fuente: https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Según los resultados de la encuesta las metodologías CRISP-DM, metodología propia y SEMMA, son las tres más usadas en proyecto de minería de datos; siendo la metodología propia (en inglés my own) una metodología no estandarizada.

2.7.1 CRISP-DM

Afirma Sheare (2003) que: "CRISP-DM es una metodología integral de minería de datos y un modelo de proceso que proporciona a cualquier persona, desde principiantes hasta expertos en minería de datos, un plan completo para llevar a cabo un proyecto de minería de datos" (p.7).

CRISP-DM divide el ciclo de vida de un proyecto de minería de datos en seis fases: comprensión empresarial, comprensión de datos, preparación de datos, modelado, evaluación e implementación. En la imagen 2-6 se observan las fases de la metodología, las flechas indican

las dependencias más importantes y frecuentes entre ellas, mientras que el círculo exterior simboliza la naturaleza cíclica de la minería de datos en sí e ilustra que las lecciones aprendidas durante el proceso de minería de datos y de la solución implementada pueden desencadenar nuevas preguntas comerciales, a menudo más centradas.

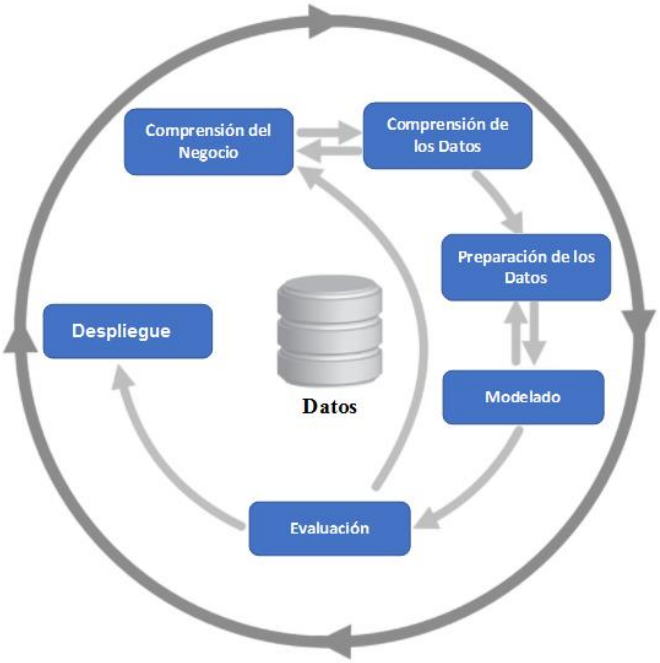


Imagen 2-6: Fases de la metodología CRISP-DM. Fuente https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview

Fase 1: Comprensión del negocio

Esta fase se la considera importante en un proyecto de minería de datos, se centra en comprender los objetivos del proyecto desde una perspectiva empresarial, convertir este conocimiento en una definición de problema de minería de datos y luego desarrollar un plan preliminar diseñado para lograr los objetivos. Es vital conocer el negocio del cual está buscando una solución, para comprender qué datos deben analizarse y cómo.

La fase de comprensión empresarial implica:

- Determinación de los objetivos del negocio
- Valoración de la situación.
- Determinación de los objetivos de la minería de datos.
- Elaboración del plan del proyecto.

Fase 2: Comprensión de los datos

Esta fase comienza con una recopilación de datos iniciales, luego el analista de datos procede a familiarizarse con ellos, identificar problemas de calidad, descubrir conocimientos iniciales y/o a detectar subconjuntos interesantes para formar hipótesis sobre información oculta.

La fase de comprensión de datos comprende:

- Recopilación de datos iniciales.

- Descripción de datos.
- Exploración de datos.
- Verificación de la calidad de los datos.

Fase 3: Preparación de los datos

Esta fase se la considera la más importante, requiere mayor tiempo y esfuerzo. En este sentido (Sheare, 2003) advierte que para un proyecto de minería de datos, los estándares de cronograma de la industria generalmente aceptados son:

> Del 50% al 70% implica la fase de preparación de datos, del 20% al 30% implica la fase de comprensión de datos, entre el 10% y el 20% se gasta en cada una de las fases de modelado, evaluación y comprensión empresarial y del 5% al 10% implica la fase de planificación de la implementación. (p15)

La preparación de datos cubre todas las actividades para construir el conjunto de datos final hacia la vista minable, para luego introducirlo en el modelado.

Los cinco pasos en la preparación de datos son:

- Selección de datos.
- Limpieza de datos.
- Construcción de datos.
- Integración de datos.
- Formateo de datos.

Fase 4: Modelado

En esta fase, se seleccionan y aplican diversas técnicas de modelado y sus parámetros se calibran a valores óptimos. Algunas técnicas tienen requisitos específicos sobre la forma de los datos. Por lo tanto, puede que sea necesario volver a la fase de preparación de datos.

Los pasos de modelado incluyen:

- Selección de la técnica de modelado.
- Generación del diseño de prueba.
- Creación de modelos.
- Evaluación de modelos.

Fase 5: Evaluación

En este capítulo se realizara la evaluación sobre los resultados del modelo de minería de datos y evaluaciones generales de recursos, restricciones y facilidades; para corroborar si las tareas realizadas responden a los objetivos del proyecto. Finalmente el líder del proyecto mostrara los resultados del modelo de minería de reglas de asociación y se describirá como utilizarlos.

Los pasos clave aquí son:

- Evaluación de resultados.
- Revisión del proyecto.
- Determinación de los próximos pasos.

En esta fase se prueban los modelos con nuevos datos reales y también se busca revelar información o sugerencias para futuras direcciones.

Fase 6: Despliegue

Generalmente la creación del modelo no es la etapa final de un proyecto de minería de datos, el conocimiento obtenido de tal debe organizarse y presentarse al usuario final para que pueda utilizarlo. Puede ser mediante un informe final o mediante la implementación de un proceso de minería de datos repetible en toda la empresa. En ambos casos es importante que el cliente, quien lleva adelante la implementación como usuario final, comprenda de antemano qué acciones deben tomarse para poder hacer uso real de los modelos creados

Los pasos clave aquí son:

- Implementación del plan.
- Monitoreo y el mantenimiento del plan.
- Producción del informe final.
- Revisión del proyecto.

2.7.2 SEMMA

La metodología SEMMA fue propuesta por SAS[1], empresa multinacional conocida como la principal fabricante de inteligencia empresarial de software, incluyendo su propio software, lo cual ayuda a dar acceso, gestionar, analizar y reportar sobre data para apoyar a la toma de decisiones.

SEMMA es el acrónimo a las cinco fases (sample, explore, modify, model, assess; en español muestreo, exploración, modificación, modelado y evaluación) es definida por SAS (2017) como: “El proceso de selección, exploración y modelamiento de grandes cantidades de datos para descubrir patrones de negocios desconocidos”. SAS también propone un software integrado y fácil de usar para el proceso de minería de datos, este proceso consiste en las fases que se muestran en la imagen 2-7.

[1] https://es.wikipedia.org/wiki/SAS_Institute

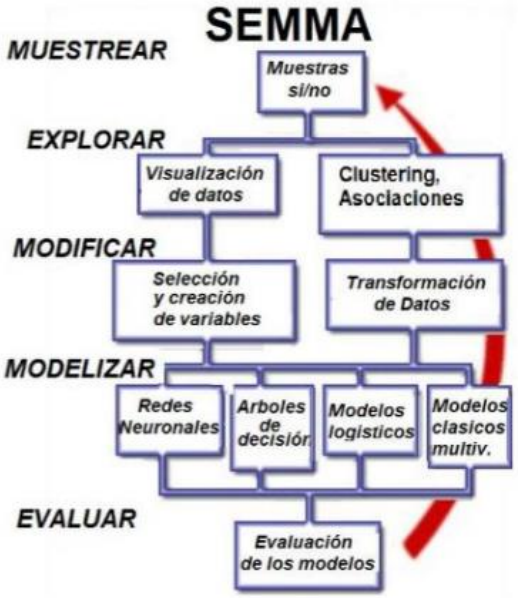


Imagen 2-7: Fases de la metodología SEMMA

La primera fase consiste en mostrar los datos creando una o más tablas de datos, considerando que el tamaño de la muestra debe ser lo suficientemente grande para contener información significativa y lo suficientemente pequeña para procesarlas. En la fase de exploración se buscan relaciones, tendencias y anomalías en los datos, para obtener una comprensión e idea sobre los mismos. Luego se en la fase de modificación se crean, seleccionan y transforman las variables necesarias para guiarlas hacia la fase de modelado. En el modelado se usan las herramientas analíticas para buscar combinaciones entre los datos que predigan de forma confiable un resultado deseado. Finalmente la fase de evaluación, evalúa la utilidad y confiabilidad de los resultados encontrados en el proceso de minería de datos.

3 Definición del problema

3.1 Definición del problema

La PyMe FOER Deportes surge por la pasión por el deporte de su fundador. Esta se dedica a la venta de indumentaria deportiva desde su local ubicado en Salta Capital y no cuenta con una estructura organizacional. Sus ventas son solo en efectivo y se registran manualmente en papel.

FOER Deportes realiza el análisis de sus ventas de forma intuitiva, por lo que el error de la intuición en las ventas es mayor, se requiere demasiado tiempo y atención para realizar dicha tarea. La interpretación sobre el comportamiento de sus clientes no es acertada a la hora de recomendarle productos u ofrecerle otro en combinación, y además la falta de participación en la web le quita la posibilidad de incrementar sus ventas.

En base a lo manifestado se propone este trabajo informático bajo el lema "Minería de datos para emprendedores", que tiene como objetivo aplicar un modelo de predicción de ventas que ayuda al negocio a tomar mejores decisiones gerenciales para mejorar la productividad e incursionarse en el comercio electrónico.

Se plantea este problema y no otro porque surge de la base de las ventas, siendo ésta la que genera los ingresos del negocio. La importancia de este problema es el análisis automático de las ventas (cesta de compra) mediante la minería de datos para extraer información e inclusive conocimiento oculto en los datos a partir de patrones interesantes hasta ahora desconocidos.

3.2 Objetivo general del proyecto

- Obtener un modelo descriptivo de asociación capaz de identificar las relaciones entre dos o más productos correspondientes a la cesta de compra de los clientes.

3.3 Objetivos específicos del proyecto

- Recolectar el conjunto de datos de ventas de la indumentaria.
- Generar y comprender un repositorio digital de los datos.
- Aplicar tareas de pre procesamientos al repositorio de datos.
- Seleccionar una técnica de minería de datos que permita descubrir asociaciones entre productos.
- Aplicar la técnica seleccionada de minería de datos que permita descubrir asociaciones entre productos que se venden juntos.
- Validar la técnica comprobando que esta se ajuste adecuadamente a los requerimientos del problema.
- Evaluar e interpretar los resultados del modelo de minería de datos.
- Tomar los resultados del modelo para dar una breve recomendación en base a las necesidades del negocio.
- Tomar los resultados del modelo para aplicarlos en el prototipo de una tienda online.

3.4 Alcance del proyecto

El desarrollo del proyecto tiene como alcance obtener un modelo descriptivo de ventas que permita identificar reglas de asociación entre las categorías de los productos de una cesta de

compra, que predigan la ocurrencia de un elemento basado en la ocurrencia de otro; para elaborar un informe final con recomendaciones y combinaciones de productos que ayuden incrementar las ventas y aplicarlas en una tienda online.

Para la etapa de minería de datos se usa RapidMiner como programa informático y CRISP-DM como metodología de proyecto de minería de datos. Para el desarrollo del proyecto también se tomarán ciertos puntos de la guía del PMBOK[2], que contiene una descripción general de los fundamentos de la Gestión de Proyectos, reconocidos como buenas prácticas y desarrollado por el PMI, para lograr un gerenciamiento eficaz y eficiente del proyecto.

El entregable final del trabajo está basado en el modelo obtenido y en las necesidades del negocio. Se hace entrega de un informe final que describe la interpretación del modelo encontrado en una serie de recomendaciones para que el dueño del negocio lo comprenda en un lenguaje natural y de esta forma lo aplique e integre tanto en su local físico como en la tienda online.

Con el objetivo de mostrar cómo se haría uso del informe de recomendaciones en la web se desarrolla el prototipo de una tienda online. El prototipo refleja los resultados del modelo de reglas de asociación, resumidos en el informe de recomendaciones, para mostrar las dependencias de los productos, ofertas e implementar la tarea de cross selling en la tienda online. Como segundo entregable de proyecto, se presenta un video sobre la tienda online, este formato se debe a que la contratación de un hosting incurre ciertos costos de los cuales no se podrán cubrir.

3.5 Alternativas para el desarrollo del proyecto

Para el desarrollo del proyecto se consideraron las alternativas en cuanto a herramientas de minería de datos (descriptas en el Capítulo 2 - Sección 2.3.6), metodologías de minería de datos (descriptas en el Capítulo 2 - Sección 2.7) y las siguientes alternativas tecnológicas en cuanto a sistema operativo y hardware.

3.5.1 Alternativas de sistemas operativos

- Windows: Es un sistema operativo desarrollado por Microsoft para ser utilizado en computadoras personales, es popular a nivel mundial y reconocido por contar con una interfaz de ventanas que permiten realizar diversas tareas. (Microsoft , 2020)

  Ventajas:
  - Las aplicaciones asociadas se pueden trabajar en cualquier equipo con Windows
  - Es simple e intuitivo para los usuarios.
  - Dispone de un gran soporte técnico a nivel global.
  - Ofrece licencias para estudiantes.

[2] https://www.pmi.org/pmbok-guide-standards/foundational/pmbok?sc_camp=D750AAC10C2F4378CE6D51F8D987F49D

- Cuenta con actualizaciones constantes utilizando descargas para mantener el proceso del sistema
- Reutiliza el código al momento de ejecuta los diversos programa y procesos para aumentar su rendimiento, transparencia y eficiencia.
- Soporta la mayoría del hardware disponible.

Desventajas:

- Su popularidad y distribución hacen que su seguridad no sea muy buena por lo que es objeto de la piratería y de malware.
- Consume muchos recursos de la computadora y memoria RAM.
- Su costo en licencias y mantenimiento es elevado.
- Con cada actualización de las versiones de Windows los equipos requieren de mayor cantidad de recursos para así tener un buen rendimiento.
- Es de código cerrado, es decir, que no comparte su código de modo que la comunidad no tiene permitido de modificar y cambiar cualquier aspecto del sistema.
- Los problemas que surgen en este sistema operativo generalmente no son fáciles de solucionar.
- Es multitarea y multiusuario

- Linux:

  Tiene como filosofía mantener una estructura de sistema operativo lo más sencilla posible; es por ello que todos los componentes, incluso los equipos y los procesos, son un archivo y los ajustes en el núcleo del sistema se pueden llevar a cabo en cualquier momento. (Mantenimiento Informático , 2020)

  Ventajas:

  - Linux es muy robusto, estable, seguro y rápido
  - Linux es libre, nos da la posibilidad de manipular en código fuente y cuenta con una gran cantidad de aplicaciones libres en Internet.
  - Es gratuito.
  - Mayor estabilidad y uso en los servidores de alto rendimiento
  - Las vulnerabilidades son detectadas y corregidas más rápidamente que cualquier otro sistema operativo.
  - Pocos requisitos de hardware.

- Linux, al poder funcionar exclusivamente en modo texto sin la necesidad de cargar un entorno gráfico
- Multitarea real: Es posible ejecutar varias aplicaciones y procesos simultáneamente.

Desventajas:

- En la mayoría de distribuciones hay que conocer nuestro Hardware a la hora de instalar.
- Se debe contar con un previo conocimiento en informática.
- En cuanto al soporte, muchas distribuciones de Linux no tienen una empresa que los respalde.
- No existen muchos software comerciales.

- MacOS: es un sistema operativo diseñado por Apple y para computadoras fabricadas específicamente por esta compañía, conocidas generalmente como Mac. Por lo cual tanto el hardware como el software son totalmente compatibles, lo que hace que el ordenador tenga un mejor rendimiento y procese la información de manera rápida.

  Ventajas
  - Mejor interfaz gráfica del mercado
  - Ideal para diseño gráfico.
  - Es muy estable
  - Generalmente son inmune a ataques de virus

  Desventajas
  - Costo elevado en el mercado
  - Existe poco software para este sistema
  - Su soporte es escaso.
  - Específico para equipos de Apple
  - Los recursos del equipo no son ampliables

3.5.2 Alternativas de hardware

Para el desarrollo del proyecto se puede usar cualquier modelo de computadora (Lenovo, Samsung, Mac) ya sean de escritorio o portátil, basta con que cumplan los siguientes requisitos mínimos:

- Dos núcleos
- Procesador de 2 GHz
- 4 GB de RAM
- > 1 GB de espacio libre en disco

- Resolución: 1280x1024

Recomendado:

- Cuatro núcleos
- Procesador de 3GHz o más rápido
- 16 GB de RAM
- > 100 GB de espacio libre en disco

En el Capítulo 4, apartado selección de las herramientas utilizadas se da a conocer las alternativas seleccionadas y su justificación.

4 Solución propuesta

Como solución se propone realizar un análisis semiautomático de la cesta de compra aplicando la tarea de asociación de la minería de datos, su resultado es un modelo de predicción de ventas en un dominio de reglas de asociación de la forma X → Y, que permiten extraer el conocimiento adecuado que ayuden a tomar decisiones aplicables en una tienda tanto física como online.

Se plantea esta solución con minería de reglas de asociación por la gran cantidad de datos con la que se cuenta, el tamaño de la base de datos puede llegar a interpretarse como un volumen pequeño, pero si la considero suficiente para el fin del estudio que pretendemos realizar. Esta solución de sebe porque la minería de datos ayuda a mejorar y automatizar procesos de análisis de grandes cantidades de datos que tienden a ser engorrosos a la hora de analizarlos.

El desarrollo del prototipo de una tienda online como solución para ingresar a e-commerce, que refleja cómo se haría uso del modelo de reglas de asociación obtenido, se debe a su consideración estadística como una de las herramientas potentes que ayudan a un negocio a incrementar las ventas y llegar a más clientes.

Selección de las herramientas para el desarrollo:

Para llevar a cabo el proyecto es fundamental contar con la cesta de compra a procesar, la herramienta de minería de datos, el sistema operativo que permite su procesamiento, el hardware sobre el cual se trabajar y la metodología para el desarrollo del proceso de minería de datos.

De estas alternativas descriptas (Capítulo 3 - Sección 3.5) se optaron las siguientes para llevar a cabo el desarrollo de proyecto:

- Cesta de compra:

Se debe contar con la cesta de compra para realizar el modelo de minería de datos, teniendo en cuenta que se encuentra en formato papel, deberá ser digitalizada en formato .xlsx para su análisis y procesamiento.

- Hardware:

Se optó por utilizar una notebook disponible ya que la misma cumple con los requerimientos mínimos necesarios (Capitulo 3 - Sección 3.5.2) y además por ser compatible con el software y herramienta de minería de datos necesaria para el desarrollo de proyecto.

- Software:

En el equipo se encuentran instalados y en funcionamiento el siguiente software:

  - Sistema Operativo: se optó por Windows 10 detallado en Capítulo 3 - Sección 3.5.1. Teniendo en cuenta el hardware con el que se cuenta para el desarrollo del proyecto, se descartó el sistema operativo MacOS, ya que solo se puede correr en un hardware específico. Esta elección se debe a que se ha comprobado que las distintas herramientas de minería de datos si son compatible con Windows 10, además este sistema es

intuitivo, conocido, no requiere un equipo específico para ser ejecutado y dispone de un gran soporte a nivel global.

- Herramienta de minería de datos: se utiliza RapidMiner detallado en Capítulo 2 – sección 2.3.6. Se eligió esta herramienta al ser un software dedicado exclusivamente a la tarea de minería de datos, siendo que para cubrir todo el proceso, desde la preparación de los datos hasta la obtención del modelo y su evaluación, proporciona los elementos necesarios. También soporta diferentes fuentes de datos acordes a la problemática. Es una herramienta fácil e intuitiva, basada en el uso de módulos, dónde cada uno de ellos realiza una función particular, permitiendo su parametrización. No requiere de conocimientos sobre programación, por lo que es de utilidad para el público en general.

- Metodología de minería de datos:

Se utiliza la metodología CRISP-DM, detallada en Capítulo 2 - Sección 2.7.1, como guía para el desarrollo del proyecto.

Se descarta la metodología SEMMA, detallada en Capítulo 2 - Sección 2.7.2, porque se necesita de su herramienta software asociada para su implementación, a diferencia de CRISP-DM que no depende de la herramienta software para llevar a cabo el proceso de obtención de conocimiento.

SEMMA excluye las etapas importantes de análisis del negocio y detección del problema organizacional, por lo que se evidencia que es una metodología orientada a aspectos técnicos; a diferencia de CRISP-DM que nos permite comenzar dichas etapas e incorporar actividades para la gestión del proyecto (como gestión del tiempo, costo y riesgos, gestión de los recursos humanos).

Es importante destacar que las etapas de selección, preparación y modelado se llevan a cabo en ambas metodologías; pero la etapa de evaluación e interpretación de los patrones en SEMMA se realiza sobre el desempeño del modelo, mientras que en CRISP-DM se realiza en función de la utilidad que aportan los mismos al dominio de aplicación o al problema organizacional.

Por ultimo CRISP-DM dispone de una etapa de implementación de los resultados y además propone una planificación para el control futuro y análisis de cierre del proyecto, cuestiones que SEMMA las excluye.

Por estas razones se considera y selecciona la metodología CRISP-DM como la más adecuada para cumplir con los objetivos y requisitos de un trabajo final de grado.

4.1 Fase1: Comprensión del negocio

La PyME FOER Deportes se dedica a la venta de indumentaria deportiva desde su local ubicado en Salta Capital y no cuenta con una estructura organizacional. Sus ventas son solo en efectivo y produce grandes volúmenes de registros de ventas en formato papel. Para hacer el análisis de sus ventas tomó la decisión de realizarlo de manera intuitiva. Actualmente este análisis no brinda información certera a la gerencia, sobre la relación entre los productos que se compran juntos, para tomar decisiones que ayuden a aumentar las ventas de otro stock; esto genera como consecuencia mayor tiempo, atención y dificultad en el análisis de los datos y una

disminución en las ventas. Además la falta de participación en la web le imposibilita abrirse a más clientes, a diferencia de otros mercados que el negocio detectó.

A esta fase se la considera la más importante, se conoce en detalle el negocio del cual se está buscando una solución para comprender qué datos, cómo se analizarlos, cómo lograr los objetivos planteados y sus necesidades a través de un plan preliminar.

Con la ejecución de esta fase se establecen:

- Objetivos del negocio
- Evaluación de la situación.
- Objetivos de la minería de datos.
- Plan del proyecto.

4.1.1 Objetivo del negocio

- Encontrar tendencias y recomendaciones que impliquen relaciones entre los productos que se compran juntos con el fin de ayudar a la toma de decisión gerencial y usarlas como criterios para iniciarse en el comercio electrónico.

4.1.2 Evaluación de la situación

4.1.2.1 Recursos disponibles

- Software: RapidMiner.
- Hardware: Notebook y disco externo.
- Personal disponible: Director del proyecto y Analista de datos.

4.1.2.2 Fuentes de datos disponibles

El negocio tiene disponible como fuente de datos las anotaciones de las ventas en formato papel, registradas entre los meses de Abril y Julio del año 2020; los datos serán prestados y se emplearan adecuadamente, poniendo énfasis en la importancia de mantener la confidencialidad de los mismos y que fueran utilizados para fines académicos únicamente. Cabe destacar que la cantidad de datos en pocos meses puede considerarse poca, pero es suficiente para el lograr el análisis de minería que se quiere llevar a cabo

4.1.2.3 Requerimientos, supuestos y restricciones

Requerimientos del proyecto:

- Los requerimientos del proyecto en cuanto a software, hardware y cesta de compra se encuentran detallados al inicio del respectivo capitulo en “Selección de las herramientas y técnicas utilizadas”.

Supuestos del proyecto:

- Se da por hecho que todas las ventas registradas en formato papel son reales y que de las mismas sólo se tomará como dato la categoría de los productos.

Restricciones del proyecto:

- El resultado esperado del desarrollo del proyecto es la obtención de un modelo que permita la asociación de un conjunto de ventas de productos y el desarrollo de un prototipo de tienda online que reflejen dichas asociaciones para mejorar las ventas en el negocio. El propósito del proyecto en así, está centrado en la obtención del modelo.

4.1.2.4 Análisis de riesgos y contingencia

A continuación se analizan y evalúan los posibles riesgos que tienen un impacto negativo durante el desarrollo del proyecto:

- Deserción de algún miembro del equipo: esto provocará una relentización de las tareas planificadas, afectará el estado de ánimo del resto de los miembros del equipo y provocará gastos económicos.
- Desperfectos en equipamiento informático: en el desarrollo del proyecto se usa un equipo informático disponible; una falla, robo o desperfecto en el hardware o software perjudicará el proyecto provocando la pérdida de datos críticos y atraso en la realización de las tareas en el tiempo estipulado.
- Cambios en la economía del país: estos cambios provocarán una desvalorización o sobrevaloración en el costo del proyecto, tanto de recursos humanos, hardware y software.

Matriz de probabilidad e impacto:

La matriz de probabilidad e impacto de riesgos (tabla 4-1) nos permite determinar la prioridad de los riesgos, su ocurrencia e impacto que provocarán en el proyecto si se producen. El producto entre la ocurrencia y el impacto dan como resultado la valoración. La prioridad toma valores entre uno y cinco, donde cinco representa una mayor prioridad.

Riesgo	Ocurrencia	Impacto	Valoración	Prioridad
Deserción de algún miembro del equipo	2	2	4	2
Desperfectos en equipamiento informático	2	4	8	1
Cambios en la economía de país	5	3	15	3

Tabla 4-1: Matriz de probabilidad e impacto de riesgos

La ocurrencia y el impacto toman valores entre uno y cinco, donde cinco cuantifica una ocurrencia segura o impacto catastrófico (tabla 4-2).

Ocurrencia / Impacto	Raro (1)	Poco probable (2)	Probable (3)	Muy probable (4)	Casi seguro (5)
Despreciable (1)	-	-	-	-	-
Menores (2)	-	Deserción de algún miembro del equipo	-	-	-
Moderadas (3)	-	-	-	-	Cambios en la economía de país
Mayores (4)	-	Desperfectos en equipamiento informático	-	-	-
Catastróficas (5)	-	-		-	-

Tabla 4-2: Identificación de los riesgos en la matriz

Plan de contingencia:

El plan de contingencia conforma los siguientes procedimientos que se deben seguir para disminuir el impacto de los riesgos y garantizar la continuidad del funcionamiento del proyecto:

- Deserción de algún miembro del equipo: tener varios currículum analizados por si es necesario reemplazar algún miembro.
- Desperfectos en equipamiento informático: realizar backup semanales de los diferentes documentos del proyecto almacenándolos en dispositivos externos y se contará con un equipo informático de respaldo.
- Cambios en la economía del país: para disminuir el impacto de este riesgo se motiva al equipo con premios por objetivos cumplidos en el menor tiempo. También se considera por contrato la financiación del proyecto en dos cuotas, pagadas del 1 al 10 de cada mes, de modo de minimizar el impacto de los cambios económicos. El incumplimiento del cliente con lo pactado, generará un incremento del 5% en cada cuota.

4.1.2.5 Análisis de costos y beneficios

Gestión de los costos:

- Costo de recurso humano: a continuación se presenta la descripción de los cargos requeridos en el proyecto (tabla 4-3) y la estimación de los costos por recurso humano (tabla 4-4) que corresponden a los honorarios que brinda el Copaipa a la fecha 10/11/2020

Rol	Descripción del cargo	Nombre	Autoridad
Director de proyecto	Analizar, planificar, coordinar, gestionar, ejecutar y controlar el proyecto	Pompeya Martínez	
Analista de datos	Llevar a cabo las tareas de minería de datos	Ian Martínez	Pompeya Martínez

Tabla 4-3: Descripción de los cargos

RRHH	Cantidad	Costo por Hs	Hs por día	Días	Precio Total
Líder de Proyecto	1	$ 720	8	13	$ 74880
Analista de Datos	1	$ 600	8	28	$ 134400
TOTAL					**$ 209280**

Tabla 4-4: Costo de RRHH

- Costos de Hardware: la estimación del costo de hardware se presenta en la tabla 4-5 y corresponde al monto establecido por un proveedor a la fecha 10/11/2020.
  La adquisición de disco externo es similar al de la marca Seagate:
  Capacidad: 1TB
  USB 3.0

Hardware	Cantidad	Costo unitario	Precio Total
Disco Externo	1	$ 5400	$ 5400
TOTAL			**$ 5400**

Tabla 4-5: Costo de Hardware

El proyecto se desarrolla con una computadora disponible por el Director del Proyecto y no se incluye en el presupuesto de hardware.

- Costos de Software:

Software	Cantidad	Costo	Precio Total
Licencia Office 2013 Profesional Plus	1	$ 4400	$ 4400
Licencia Project 2013 Profesional	1	$ 3300	$ 3300
TOTAL			**$ 7700**

Tabla 4-6: Costo de Software

- Gastos varios:

Concepto	Meses	Costo por mes	Precio Total
Papelería y útiles de oficina	2	$ 200	$ 400
TOTAL			**$ 400**

Tabla 4-7: Costo varios

- Costo del proyecto:

Concepto	Costo
RRHH	$ 209280
Hardware	$ 5400
Software	$ 7700
Costos varios	$ 400
TOTAL	**$ 222780**

Tabla 4-8: Costo del Proyecto

Análisis de beneficios:

El beneficio inmediato del proyecto es la optimización de la tarea intuitiva de análisis de la cesta de compra, logrando que se realice de forma automática mediante la aplicación de la minera de datos. Con este análisis automático se obtienen patrones más certeros sobre el comportamiento de los clientes y de los productos que se compran juntos, se percibe de mejor manera los patrones que a simple vista (por apreciación) son difíciles de identificar y ayudan a la gerencia a tomar mejores decisiones en base al conocimiento obtenido para remover a incrementar las ventas de los productos.

Otro benéfico es la mayor visualización de los productos en el local y en el desarrollo de un prototipo de tienda online, posicionándolos de acuerdo a los hábitos de los clientes, permitiendo mayor agilidad en la búsqueda, influencia a la hora de la compra y un aumento en las ventas y lealtad de los consumidores.

4.1.2.6 Análisis de FODA

Para el análisis de FODA, primero se realiza un análisis interno y externo, a continuación una breve descripción de sus componentes:

Análisis Interno, se analizan fortalezas y debilidades:

- Fortalezas, son las capacidades y recursos con los que se cuenta para alcanzar el objetivo del proyecto de minería de datos.

- Debilidades, son factores que provocan una situación desfavorable para la ejecución de objetivo del proyecto de minería de datos, por lo que deben ser atendidas y mejorarse.

Análisis Externo, se analizan oportunidades y amenazas.

- Oportunidades, factores que resultan positivos y favorables en el entorno del proyecto de minería de datos.
- Amenazas, situaciones que proviene del entorno atentando contra la estabilidad del proyecto.

Una vez que se definen las principales fortalezas, oportunidades, debilidades y amenazas del proyecto, se plantean estrategias alternativas que alinean los factores internos y externos (tabla 4-9):

- Estrategias FO, también conocidas como estrategias ofensivas, son estrategias que utilizan las fortalezas para maximizar las oportunidades.
- Estrategias FA, también conocidas como estrategia reactiva, son estrategias que utilizan las fortalezas para disminuir las amenazas.
- Estrategias DO, también conocidas como estrategia adaptativas, son estrategias para minimizar las debilidades aprovechando las oportunidades.
- Estrategias DA, también conocidas como estrategia defensivas, superan debilidades y evitan amenazas.

4.1.2.7 Análisis de Factibilidad

Factibilidad Técnica: El proyecto es factible técnicamente, ya que tanto los requerimientos de Hardware como de Software, se encuentran disponibles, son fáciles de conseguir y económicos.

Factibilidad operativa: El proyecto es factible operativamente porque cuenta con el personal necesario y capacitado para un mismo fin, atendiendo los lineamientos del mismo. También se cuenta con una buena comunicación durante las distintas etapas del desarrollo del proyecto y no existe resistencia al cambio.

Factibilidad Legal: Legalmente la solución propuesta y los datos utilizados no infringen ninguna ley; los mismos son resguardados a pesar de no referirse a datos personales, solo la persona responsable de la extracción y procesamiento de datos tiene acceso a ellos y firmará un contrato de confidencialidad de los datos.

4.1.3 Objetivo de minería de datos

4.1.3.1 Objetivo general de minería de datos

- Aplicar la técnica de minería de datos por regla de asociación para obtener un modelo que permita relaciones entre los productos de una cesta de compra, con el fin de brindar recomendaciones y mostrarlas en un prototipo de tienda online.

4.1.3.2 Criterios de éxito de minería de datos

- El modelo propuesto debe generar las reglas de asociación de los productos con al menos un 80% de confianza.

INTERNAS / **EXTERNAS**	**FOTALEZAS**	**DEBILIDADES**
	-Experiencia y conocimientos sobre el uso de la herramienta RapidMiner de minería de datos. -Conocimientos adquiridos sobre minería de datos e informática durante la formación académica. -Conocimiento sobre las metodologías de proyecto CRISP-DM y PMI, para el desarrollo del proyecto. -Asesoramiento y disposición de profesores idóneos en temas de minería de datos.	-No poder contar con una gran cantidad de datos en la cesta de compra. -Limitación en la distribución de proyecto como producto final.
OPORTUNIDADES	**ESTRATEGIAS OFENSIVAS (FO)**	**ESTRATEGIAS ADAPTATIVAS (DO)**
-Demanda creciente del mercado en búsqueda de análisis de predicción de ventas que conlleven a un análisis automático. -Proyecto con un precio final a bajo costo. -Ampliar el espectro de venta del negocio. -Expandir el nicho del mercado tradicional al virtual.	-Crear un modelo de reglas de asociación que me permita predecir los productos que se compran juntos. -Aprovechar el conocimiento de la minería de datos, CRISP-DM y RapidMiner, para poder atender a una demanda creciente de predicción de ventas. -Aprovechar el conocimiento de la metodología PMI para gestionar un proyecto final eficiente y de bajo costo.	-Aprovechar la expansión al mercado virtual para ampliar los datos de la cesta de compra
AMENAZAS	**ESTRATEGIAS REACTIVAS (FA)**	**ESTRATEGIAS DEFENSIVAS (DA)**
-Ingreso de proyectos finales de grados con el mismo enfoque. -Falsificación de datos en la cesta de compra. -Políticas socioeconómicas inestables. -Competencia significativa en el mercado con servicios similares de predicción de ventas. -Continua actualización de las tecnologías y herramientas de minería de datos.	-Mejorar la calidad de los resultados del producto final haciendo uso eficiente de los conocimientos de minería de datos para diferenciarse de otros proyectos. -Hacer un uso apropiado de la informática para disminuir errores humanos en la carga de datos para la cesta de compra.	-Agilizar el tiempo de desarrollo del proyecto para evitar que actualizaciones tecnológicas puedan dejarlo obsoleto y los cambios socioeconómicos, inestable. -Permanecer actualizados e informados sobre nuevas tendencias tecnológicas del mercado.

Tabla 4-9: Matriz FODA

4.1.4 Plan del Proyecto

El siguiente plan para la dirección de este proyecto de minería de datos se basa en la guía PMBOOK y contempla:

4.1.4.1 Gestión del Alcance

La descripción del alcance se encuentra en Capitulo 3 - sección 3.4.

4.1.4.2 Estructura de desglose del trabajo

La EDT (Estructura de Desglose de Trabajo) "es una descomposición jerárquica del alcance total del trabajo a realizar por el equipo del proyecto para cumplir con los objetivos y crear los entregables requeridos" (Project Management Institute, Inc., 2013). Las tareas para cumplir con el objetivo del proyecto se muestran en el siguiente EDT basada en la metodología CRISP-DM (Imagen 4-1).

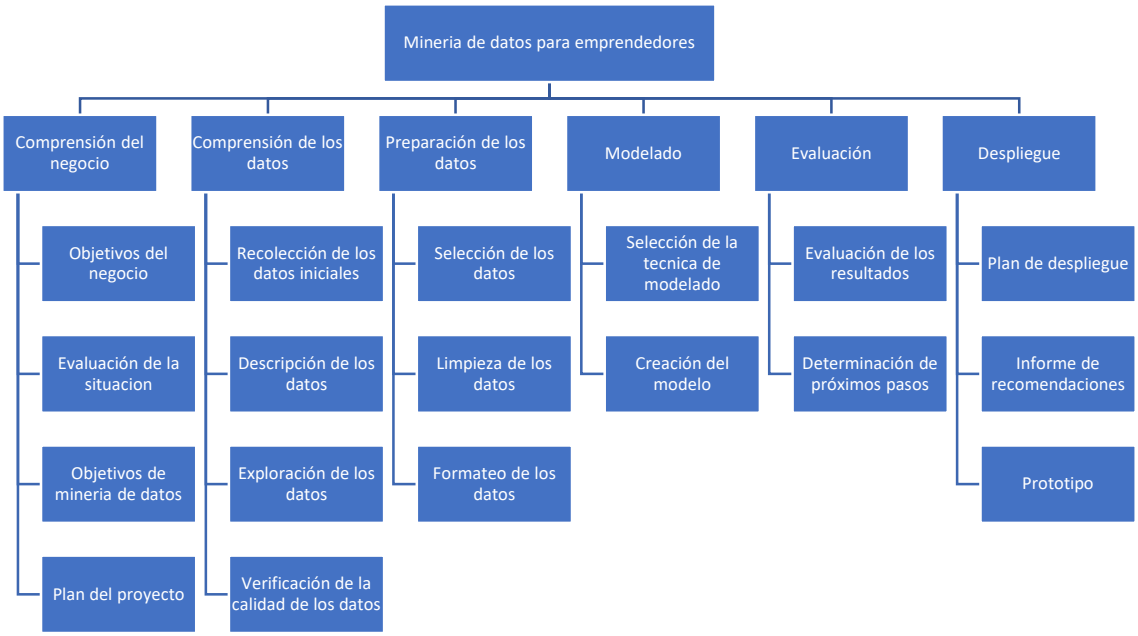


Imagen 4-1: EDT del proyecto minería de datos para emprendedores

Diccionario de la EDT:

1. Comprensión del Negocio: se considera la más importante del proyecto minería de datos, centrándose en comprender los objetivos del proyecto desde una perspectiva empresarial luego convertir este conocimiento en una definición de problema de minería de datos y desarrollar un plan preliminar diseñado para lograr los objetivos. Es vital conocer el negocio del cual está buscando una solución, para comprender qué datos deben analizarse y cómo.

1.1 Objetivos del negocio: consiste entender la verdadera necesidad del cliente para descubrir los factores importantes involucrados en el proyecto.

1.2 Evaluación de la situación: consiste en comprender el estado actual del proyecto de minería de datos para abordarlo, desde los recursos disponibles, personal, software, datos disponibles, riesgos del proyecto, soluciones a estos riesgos y elaborar un análisis de costos beneficio del proyecto.

1.3 Objetivos de minería de datos: establece el objetivo de la minería de datos y se definen criterios de éxitos.

1.4 Plan del proyecto: consiste en elaborar y comunicar a los interesados el plan de acción del proyecto de minería de datos para lograr los objetivos del negocio; gestionando el alcance, tiempo, costo, riesgos y analizar las factibilidades del proyecto.

2. Comprensión de los datos: esta fase comienza con una recopilación de datos iniciales, luego el analista de datos procede a familiarizarse con ellos, identificar problemas de calidad y descubrir conocimientos iniciales.

2.1 Recopilación de los datos iniciales: se recolectan los datos necesarios, incluyendo la carga a través de un formulario de Google para su posterior procesamiento.

2.2 Descripción de los datos: se examinan, describen y comprenden las propiedades de los datos adquiridos y se informa sobre el número de registros, atributos y campos.

2.3 Exploración de los datos: consiste en ejecutar una visualización de los datos y exponer los primeros hallazgos a través de la herramienta RapidMiner.

2.4 Verificación de la calidad de los datos: se examina y se verifica la calidad de los datos, teniendo en cuenta atributos faltantes, campos en blanco y la ortografía de los valores.

3. Preparación de los datos: cubre todas las actividades para construir el conjunto de datos final hacia la vista minable, para luego introducirlo en el modelado.

3.1. Selección de los datos: se justifican los registros y atributos seleccionados por inclusión o exclusión; teniendo en cuenta su relevancia en base a los objetivos, sus limitaciones técnicas y de calidad, limite en el volumen de datos o tipo de datos e importancia en atributos.

3.2. Limpieza de los datos: se aborda el problema del valor faltante detectado en el paso de "Verificación de la calidad de los datos" para optimizar su calidad y consistencia hacia la vista minable y necesaria para la fase de modelación.

3.3. Formateo de los datos: consiste en transformar sintácticamente los datos sin modificar su significado, se trata de cambios en el formato o diseño de los datos para adecuarlos a la herramienta de modelado especifica. Finalmente se realiza un reporte final de la calidad de los datos.

4. Modelado: existen diversas técnicas de modelado para un problema de minería de datos por lo que se selecciona y se aplica una, calibrando sus parámetros a valores óptimos.

4.1. Selección de la técnica de modelado: se selecciona la técnica más adecuada de modelado que cumple con los objetivos del proyecto.

4.2. Creación del modelo: consiste en ejecutar la técnica seleccionada sobre los datos preparados, para crear el modelo y probar la variación de parámetros.

5. Evaluación

5.1. Evaluación de resultados: consiste en interpretar y determinar si los resultados del modelo y su eficiencia cumplen con los objetivos del negocio. Se resumen los resultados de la evaluación en términos de criterios de éxito comercial, incluida una declaración final sobre si el proyecto ya cumple con los objetivos comerciales iniciales.

5.2. Determinación de próximos pasos: el líder del proyecto debe decidir si finalizar este proyecto y pasar a la implementación o si iniciar más iteraciones o configurar nuevos proyectos de minería de datos.

6. Despliegue

6.1. Plan de implementación: consiste en tomar reglas obtenidas del modelo y llevarlas a la práctica mediante una estrategia de implementación.

6.2. Informe de recomendaciones: consiste en detallar recomendaciones para la implementación del modelo en el negocio y en la web, estas recomendaciones surgen de los resultados del modelo encontrado.

6.3. Prototipo: se desarrolla un prototipo de tienda online con el objetivo de mostrar cómo se usarían las recomendaciones basadas en el modelo final encontrado.

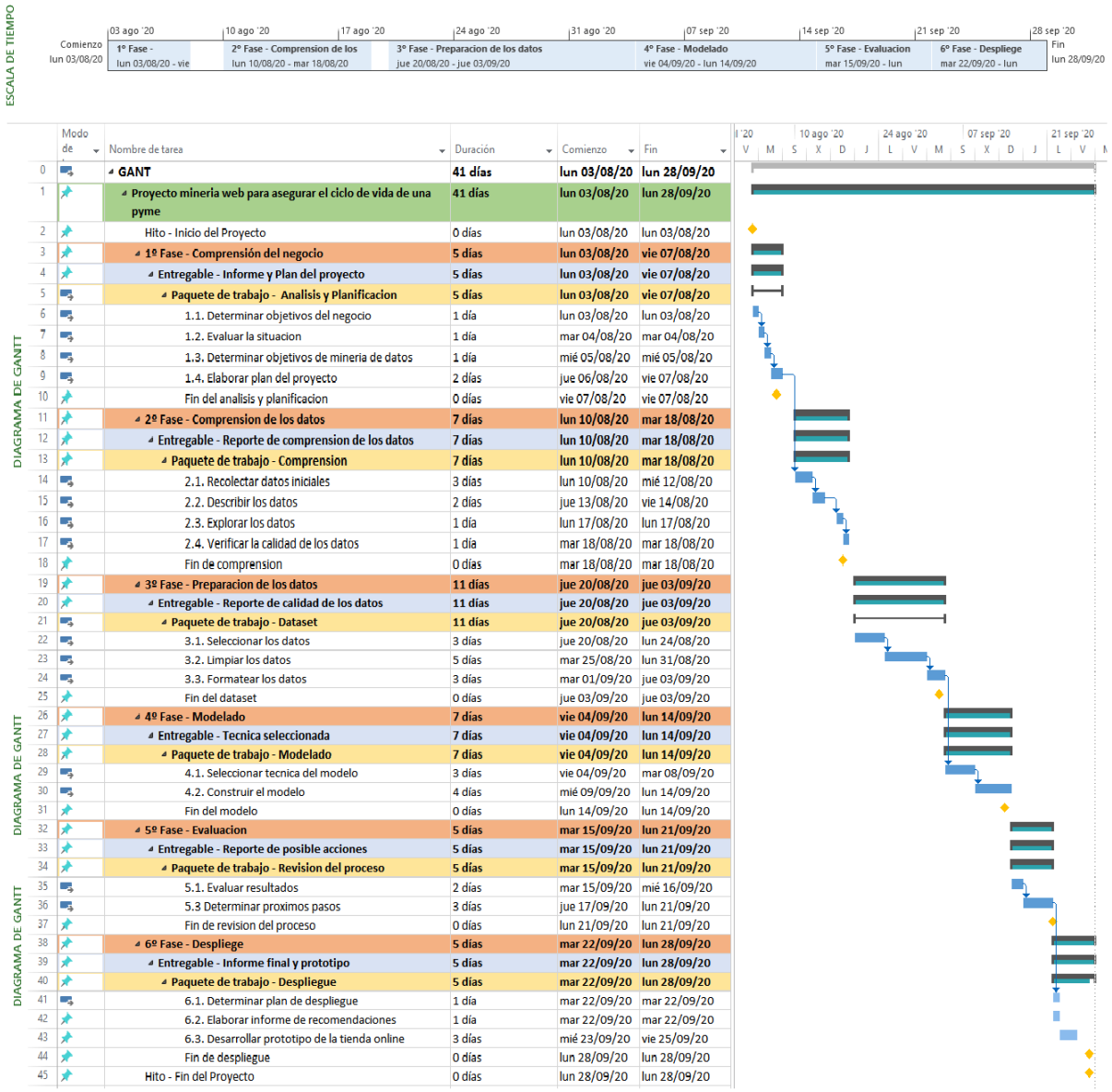
4.1.4.3 Gestión del tiempo

En el diagrama de Gantt (Imagen 4-2) se representa cada tarea en forma de barra, sobre una escala gráfica, manteniendo una relación de proporcionalidad entre sus duraciones y su posición respecto del punto de origen y fin dentro del proyecto.

4.2 Fase 2: Comprensión de los datos

Para entender mejor la fase de comprensión de los datos se propone un diagrama de guía con los elementos que se usan para llevarla a cabo (Imagen 4-3).

1. Cuaderno con el registro manual de las ventas, para extraer los datos.
2. Formulario de Google: permite la carga de los datos de ventas, es un formulario con extensión *.form* hecho a medida que permite recopilar diversos tipos de información de una manera simple y eficiente, a través de internet. Los resultados del formulario se registran automáticamente en una hoja de cálculo del tipo *.sheet* que se genera por defecto en la nube de Google, por lo que podemos consultarlos en cualquier momento y desde cualquier dispositivo.
3. Google Sheet: es una hoja de cálculo con extensión *.sheet* que se genera por defecto en la nube de Google por su respectivo formulario, almacenando en cada fila el resultado de un formulario, permite visualizar la carga de los datos en tiempo real.
4. Excel: la hoja de cálculo *.sheet* también puede visualizarse en la pc como una hoja de cálculo de Excel con extensión *.xlsx* este tipo de archivo es el ideal para el desarrollo del trabajo ya que es uno de los formatos que admite Rapid Miner.
5. Rapid Miner, para realizar el proceso de minería de datos.



	Modo de	Nombre de tarea	Duración	Comienzo	Fin
0		GANT	41 días	lun 03/08/20	lun 28/09/20
1		Proyecto mineria web para asegurar el ciclo de vida de una pyme	41 días	lun 03/08/20	lun 28/09/20
2		Hito - Inicio del Proyecto	0 días	lun 03/08/20	lun 03/08/20
3		1º Fase - Comprensión del negocio	5 días	lun 03/08/20	vie 07/08/20
4		Entregable - Informe y Plan del proyecto	5 días	lun 03/08/20	vie 07/08/20
5		Paquete de trabajo - Analisis y Planificacion	5 días	lun 03/08/20	vie 07/08/20
6		1.1. Determinar objetivos del negocio	1 día	lun 03/08/20	lun 03/08/20
7		1.2. Evaluar la situacion	1 día	mar 04/08/20	mar 04/08/20
8		1.3. Determinar objetivos de mineria de datos	1 día	mié 05/08/20	mié 05/08/20
9		1.4. Elaborar plan del proyecto	2 días	jue 06/08/20	vie 07/08/20
10		Fin del analisis y planificacion	0 días	vie 07/08/20	vie 07/08/20
11		2º Fase - Comprension de los datos	7 días	lun 10/08/20	mar 18/08/20
12		Entregable - Reporte de comprension de los datos	7 días	lun 10/08/20	mar 18/08/20
13		Paquete de trabajo - Comprension	7 días	lun 10/08/20	mar 18/08/20
14		2.1. Recolectar datos iniciales	3 días	lun 10/08/20	mié 12/08/20
15		2.2. Describir los datos	2 días	jue 13/08/20	vie 14/08/20
16		2.3. Explorar los datos	1 día	lun 17/08/20	lun 17/08/20
17		2.4. Verificar la calidad de los datos	1 día	mar 18/08/20	mar 18/08/20
18		Fin de comprension	0 días	mar 18/08/20	mar 18/08/20
19		3º Fase - Preparacion de los datos	11 días	jue 20/08/20	jue 03/09/20
20		Entregable - Reporte de calidad de los datos	11 días	jue 20/08/20	jue 03/09/20
21		Paquete de trabajo - Dataset	11 días	jue 20/08/20	jue 03/09/20
22		3.1. Seleccionar los datos	3 días	jue 20/08/20	lun 24/08/20
23		3.2. Limpiar los datos	5 días	mar 25/08/20	lun 31/08/20
24		3.3. Formatear los datos	3 días	mar 01/09/20	jue 03/09/20
25		Fin del dataset	0 días	jue 03/09/20	jue 03/09/20
26		4º Fase - Modelado	7 días	vie 04/09/20	lun 14/09/20
27		Entregable - Tecnica seleccionada	7 días	vie 04/09/20	lun 14/09/20
28		Paquete de trabajo - Modelado	7 días	vie 04/09/20	lun 14/09/20
29		4.1. Seleccionar tecnica del modelo	3 días	vie 04/09/20	mar 08/09/20
30		4.2. Construir el modelo	4 días	mié 09/09/20	lun 14/09/20
31		Fin del modelo	0 días	lun 14/09/20	lun 14/09/20
32		5º Fase - Evaluacion	5 días	mar 15/09/20	lun 21/09/20
33		Entregable - Reporte de posible acciones	5 días	mar 15/09/20	lun 21/09/20
34		Paquete de trabajo - Revision del proceso	5 días	mar 15/09/20	lun 21/09/20
35		5.1. Evaluar resultados	2 días	mar 15/09/20	mié 16/09/20
36		5.3 Determinar proximos pasos	3 días	jue 17/09/20	lun 21/09/20
37		Fin de revision del proceso	0 días	lun 21/09/20	lun 21/09/20
38		6º Fase - Despliege	5 días	mar 22/09/20	lun 28/09/20
39		Entregable - Informe final y prototipo	5 días	mar 22/09/20	lun 28/09/20
40		Paquete de trabajo - Despliegue	5 días	mar 22/09/20	lun 28/09/20
41		6.1. Determinar plan de despliegue	1 día	mar 22/09/20	mar 22/09/20
42		6.2. Elaborar informe de recomendaciones	1 día	mar 22/09/20	mar 22/09/20
43		6.3. Desarrollar prototipo de la tienda online	3 días	mié 23/09/20	vie 25/09/20
44		Fin de despliegue	0 días	lun 28/09/20	lun 28/09/20
45		Hito - Fin del Proyecto	0 días	lun 28/09/20	lun 28/09/20

Imagen 4-2: Diagrama Gantt de la planificación del proyecto

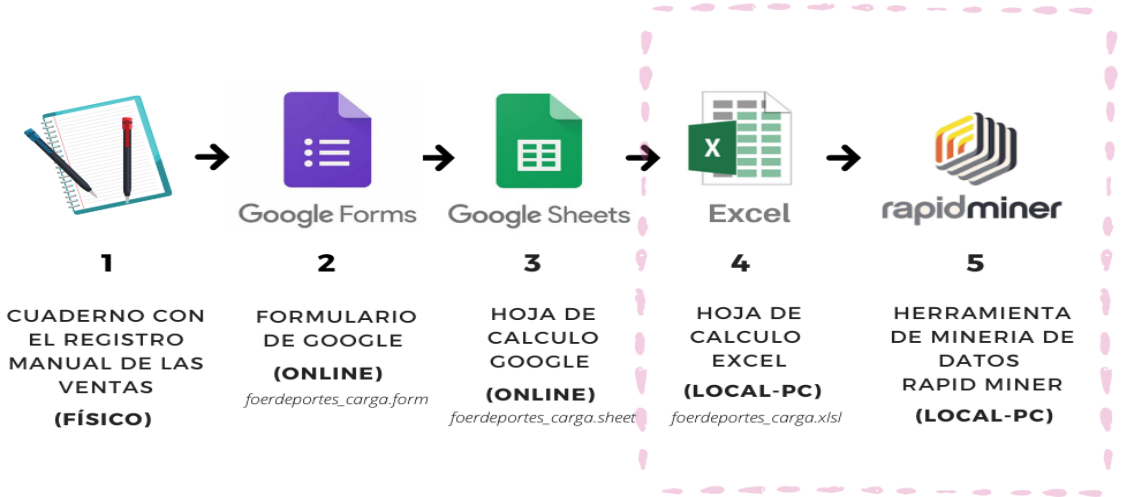


Imagen 4-3: Diagrama con los elementos necesarios para la comprensión de los datos

4.2.1 Recopilación de datos iniciales

Los datos sobre las ventas de la indumentaria deportiva se recopilan de un cuaderno, registradas en forma escrita. Mediante la técnica de observación se detectó que los registros contienen información completa de la transacción de los clientes con respecto a la categoría de los productos adquiridos y no presentan valores atípicos.

Para representar las transacciones de ventas en una base de datos existen tres formas:

1. Lista: cada fila representa una transacción con la lista de artículos comprados por el consumidor y pueden tener un número indefinido de columnas; como se muestra en la tabla 4-10.

Zapatilla	Media	Remera	
Buzo	Media		
Gorra	Short	Media	Remera

Tabla 4-10: Representación de los datos como una lista de artículos comprados

2. Representación vertical: conformada por dos columnas, en donde la primera es el número o ID de la transacción (TID) y la segunda indica qué artículo está presente en dicha transacción, (se repiten tantas filas TID como artículos tenga la transacción del cliente); como se muestra en la tabla 4-11.

TID	Ítem
1	Zapatilla
1	Media
1	Remera
2	Buzo
2	Media
3	Gorra
3	Short
3	Media
3	Remera

Tabla 4-11: Representación de los datos de forma vertical

3. Representación horizontal (una matriz de unos-ceros o true-false): cada fila de la matriz representa una transacción y cada columna representa un artículo, salvo la primera que representa el ID de la transacción (TID). Si el articulo está presente en la transacción lo representamos con true y si está ausente con false; como se muestra en la tabla 4-12.

TID	Zapatilla	Remera	Musculosa	Short	Buzo	Jogging	Gorra	Media
1	true	true	false	false	false	false	false	true
2	false	false	false	false	true	false	false	true
3	false	true	false	true	false	false	true	true

Tabla 4-12: Representación de los datos de forma horizontal

Si bien la representación vertical es la más recomendada y usada para grandes volúmenes de datos porque ocupa menos espacio en memoria, se considera el formato horizontal ya que se

admite en la mayoría de los algoritmos como A priori y FP-Growth. Habiendo seleccionado la representación horizontal, el formulario de Google se diseña como se muestra en la imagen 4-4.

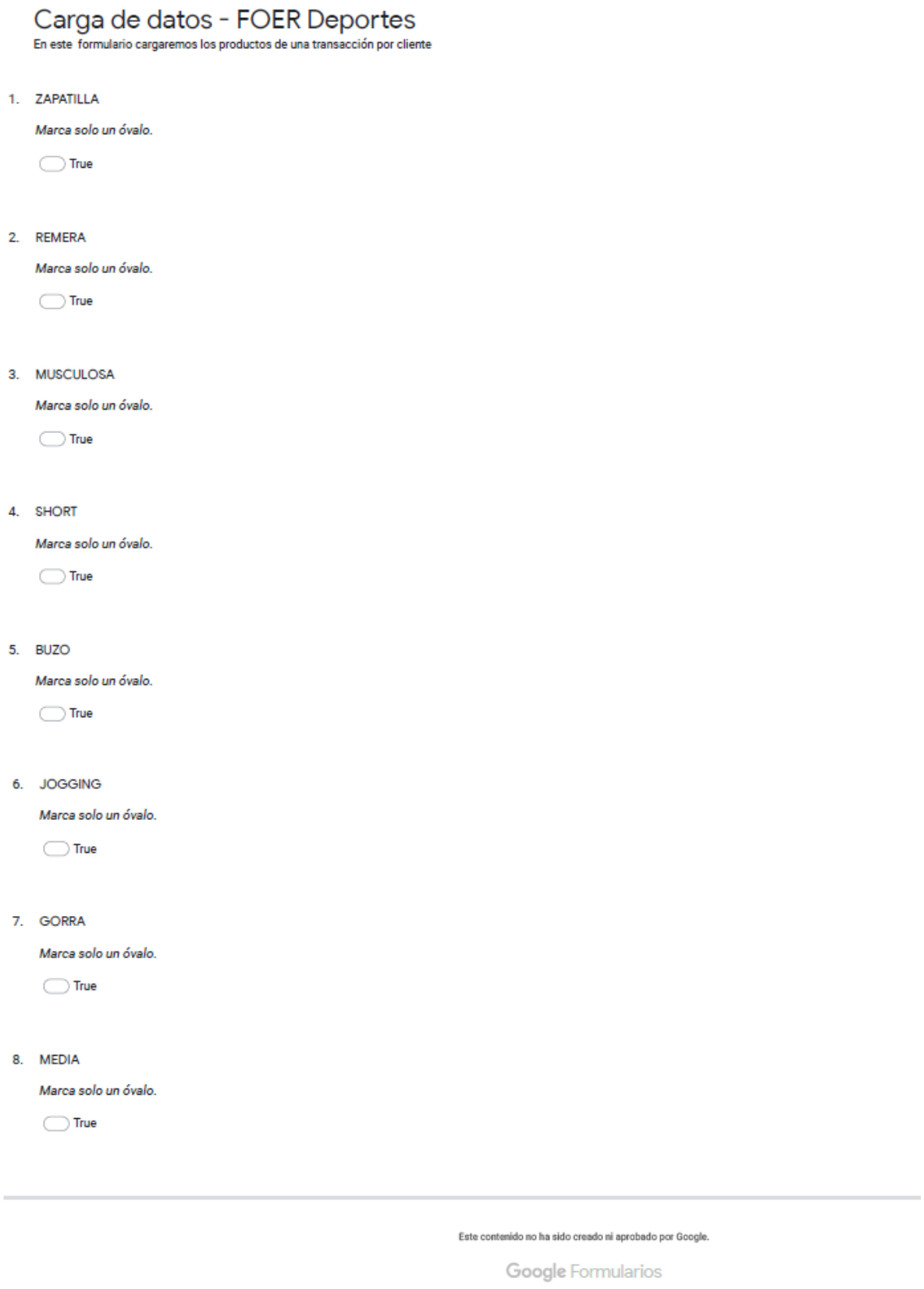


Imagen 4-4: Formulario de Google para carga de datos carga de datos

El dueño del negocio o interesado es quien procede a la carga de los datos a través del link del formulario de Google (*foerdeportes_carga.form):*
https://docs.google.com/forms/d/e/1FAIpQLSferIWS0ytkAkMxoC3tDb-7tVQ5sGzSW7GycQ3sFXv6XTUHMw/viewform

Este formulario automáticamente genera el almacenamiento de las respuestas en la hoja de cálculo (*foerdeportes_carga.gsheet*):
https://docs.google.com/spreadsheets/d/1cxyOhVyc3Xc8GoAjZ6IiYERuy6sSg6V98fkccESnaBI/edit?usp=sharing

Esta hoja de cálculo también se visualiza de forma local en el directorio de la PC del Director del proyecto y Analista de datos, como *foerdeportes_carga.xlsx*.

Cabe aclarar que en el desarrollo del proyecto se trabaja con los datos históricos de las ventas registradas entre los meses de Abril y Julio del año 2020.

En el caso de querer agregar, eliminar o modificar productos en el formulario, en el anexo A se detallan las instrucciones relacionadas (es necesario contar con un correo de Google) y a continuación se presenta el link del formulario para su edición:
https://docs.google.com/forms/d/17CPwyf2qfTNPAqW6-0v8BmpWy7yQYQBrzXQpQMEPEi8/edit

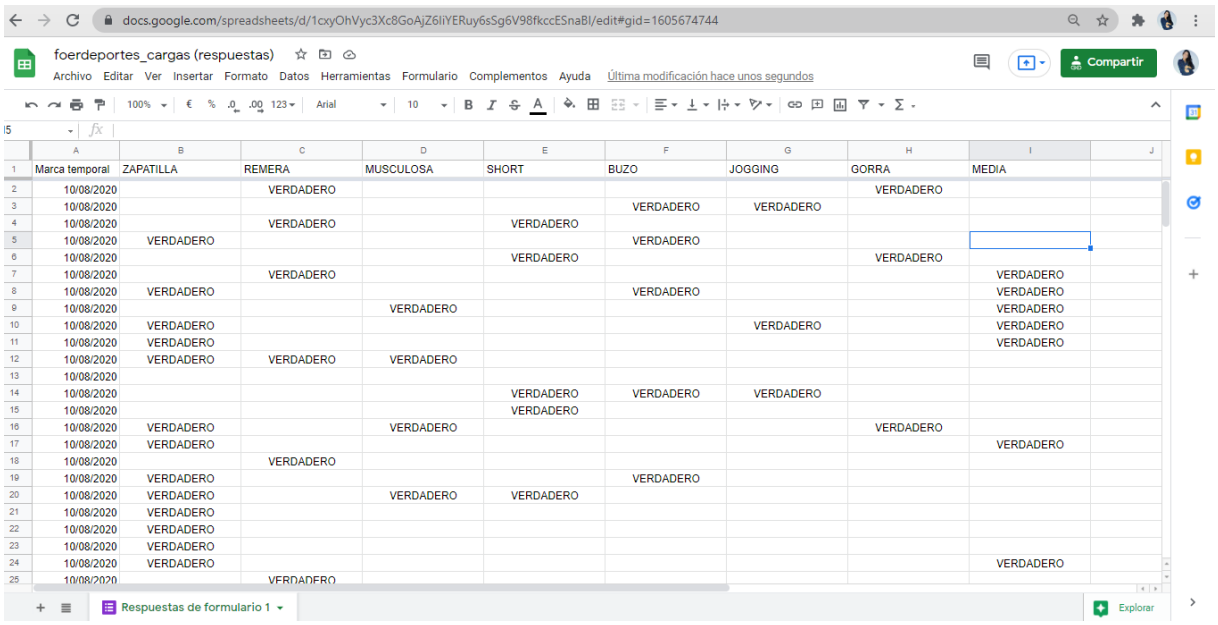
4.2.2 Descripción de los datos

El conjunto de datos obtenido cuenta con 186 transacciones de ventas y los siguientes ítems son una descripción de los campos usados para representar la base de datos, siendo true (compra) o false (no compra):

1. TID: es un número que identifica de forma única a la transacción o compra realizada por el cliente.
2. Zapatilla: representa el artículo zapatilla mediante un valor true o false.
3. Remera: representa el artículo remera mediante un valor true o false.
4. Musculosa: representa el artículo musculosa mediante un valor true o false.
5. Short: representa el artículo short mediante un valor true o false.
6. Buzo: representa el artículo buzo mediante un valor true o false.
7. Jogging: representa el artículo jogging mediante un valor true o false.
8. Gorra: representa el artículo gorra mediante un valor true o false.
9. Media: representa el artículo media mediante un valor true o false.

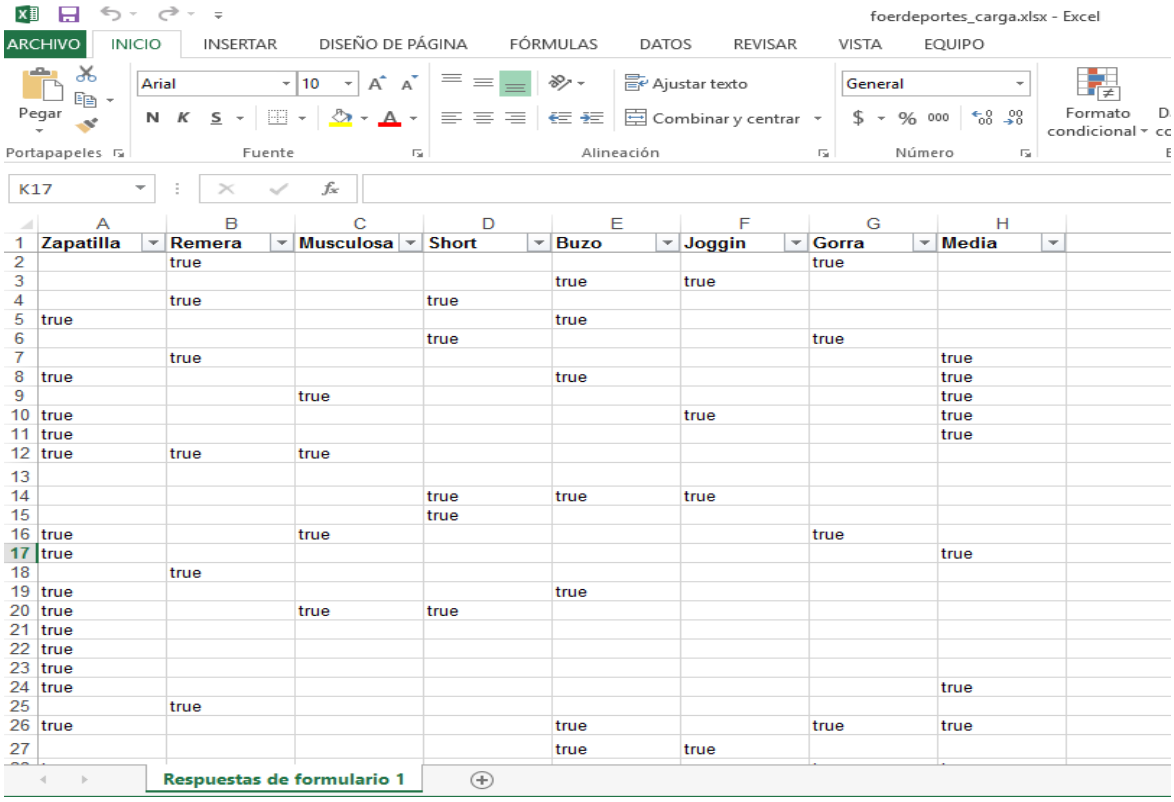
4.2.3 Exploración de los datos

Archivo *foerdeportes_carga.gsheet* que se obtuvo de la carga de los productos con el formulario *foerdeportes_carga.form* (Imagen 4-5)
Archivo *foerdeportes_carga.xlsx* que se obtuvo de la carga de los productos con el formulario *foerdeportes_carga.form* (Imagen 4-6)



Marca temporal	ZAPATILLA	REMERA	MUSCULOSA	SHORT	BUZO	JOGGING	GORRA	MEDIA
10/08/2020		VERDADERO					VERDADERO	
10/08/2020					VERDADERO	VERDADERO		
10/08/2020		VERDADERO		VERDADERO				
10/08/2020	VERDADERO				VERDADERO			
10/08/2020				VERDADERO			VERDADERO	
10/08/2020		VERDADERO						VERDADERO
10/08/2020	VERDADERO				VERDADERO			VERDADERO
10/08/2020			VERDADERO					VERDADERO
10/08/2020	VERDADERO					VERDADERO		VERDADERO
10/08/2020	VERDADERO							VERDADERO
10/08/2020	VERDADERO	VERDADERO	VERDADERO					
10/08/2020								
10/08/2020				VERDADERO	VERDADERO	VERDADERO		
10/08/2020				VERDADERO				
10/08/2020	VERDADERO		VERDADERO				VERDADERO	
10/08/2020	VERDADERO							VERDADERO
10/08/2020		VERDADERO						
10/08/2020	VERDADERO				VERDADERO			
10/08/2020	VERDADERO		VERDADERO	VERDADERO				
10/08/2020	VERDADERO							
10/08/2020	VERDADERO							
10/08/2020	VERDADERO							
10/08/2020	VERDADERO							VERDADERO
10/08/2020		VERDADERO						

Imagen 4-5: Archivo foerdeportes_carga.gsheet



	Zapatilla	Remera	Musculosa	Short	Buzo	Joggin	Gorra	Media
2		true					true	
3					true	true		
4		true		true				
5	true				true			
6				true			true	
7		true						true
8	true				true			true
9			true					true
10	true					true		true
11	true							true
12	true	true	true					
13								
14				true	true	true		
15				true				
16	true		true				true	
17	true							true
18		true						
19	true				true			
20	true		true	true				
21	true							
22	true							
23	true							
24	true							true
25		true						
26	true				true		true	true
27					true	true		

Imagen 4-6: Archivo foerdeportes_carga.xlsx

Se ejecuta el proceso (Imagen 4-7) en RapidMiner para iniciar el proceso de exploración y análisis del repositorio *foerdeportes_carga.xlsx*.

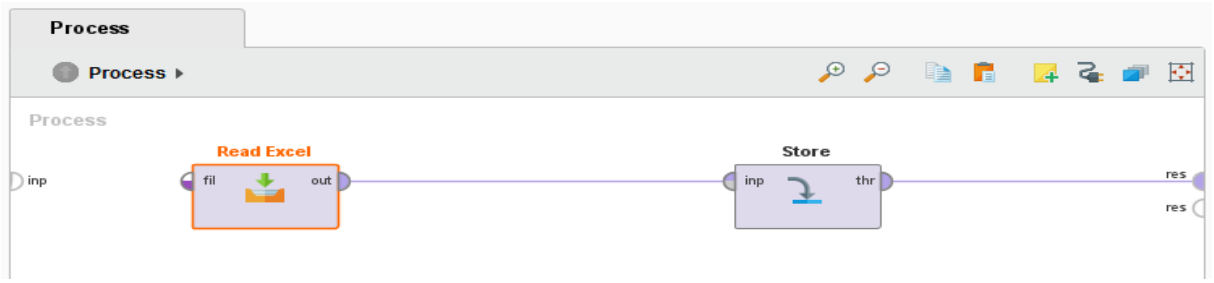


Imagen 4-7: Exploración de los datos en RapidMiner

El uso del operador "Read Excel" permite cargar los datos almacenados en la hoja de cálculo *foerdeportes_carga.xlsx*, técnicamente convierte este archivo en un objeto "example set" (conjunto de muestras). En este operador se configura el formato de los atributos como polinomial (Imagen 4-8).

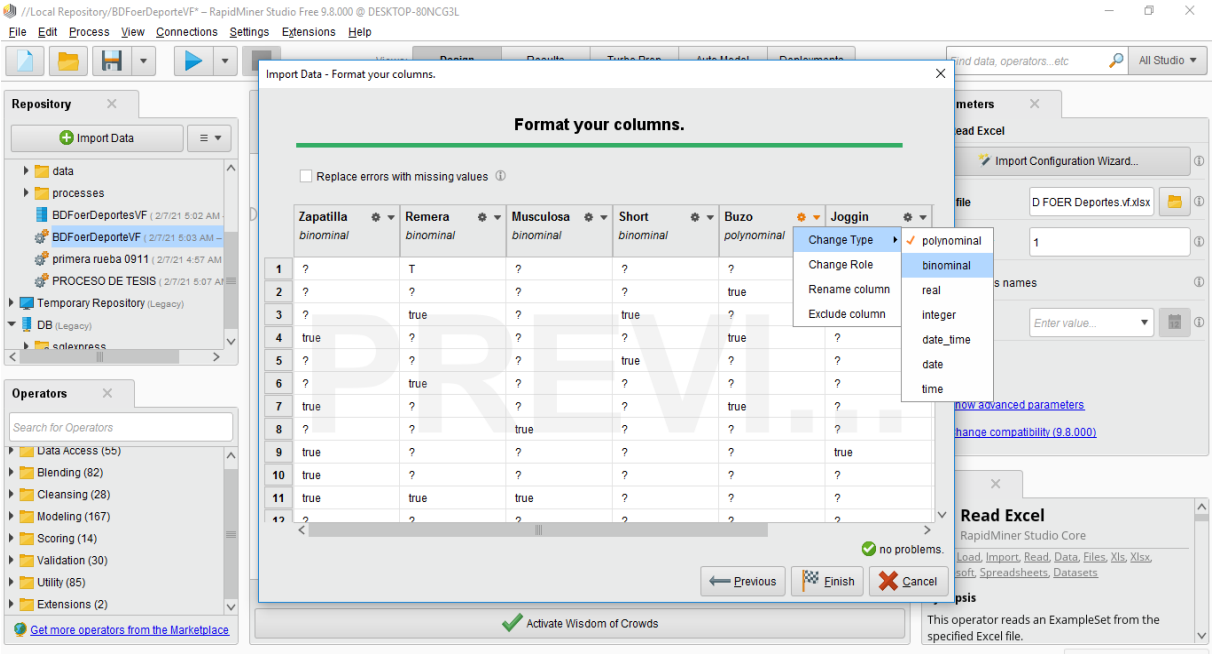


Imagen 4-8: Formato de las columnas en RapidMiner

El uso del operador "Store" permite guardar el objeto "example set" en el repositorio local de RapidMiner con el fin de utilizarlo de forma práctica en otros procesos, al repositorio lo guardamos con el nombre *BDFoerDeportesVF*.

Luego se crea y ejecuta un nuevo proceso con el operador "Retrive", quien permite hacer la lectura de los datos del repositorio *BDFoerDeportesVF*, como se muestra en la imagen 4-9:

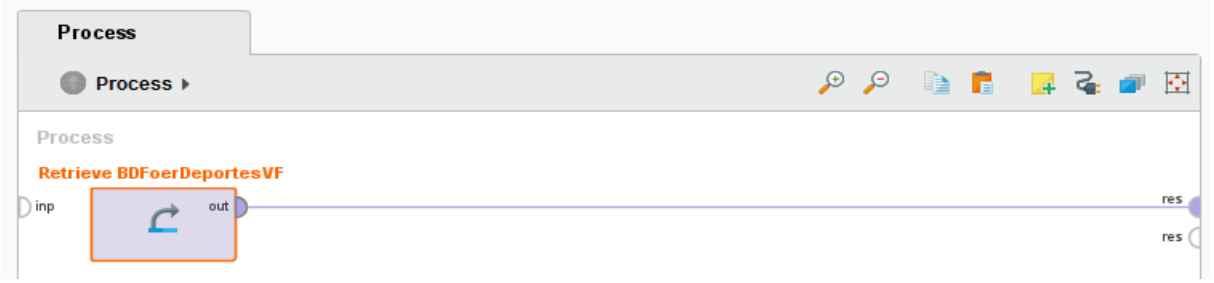
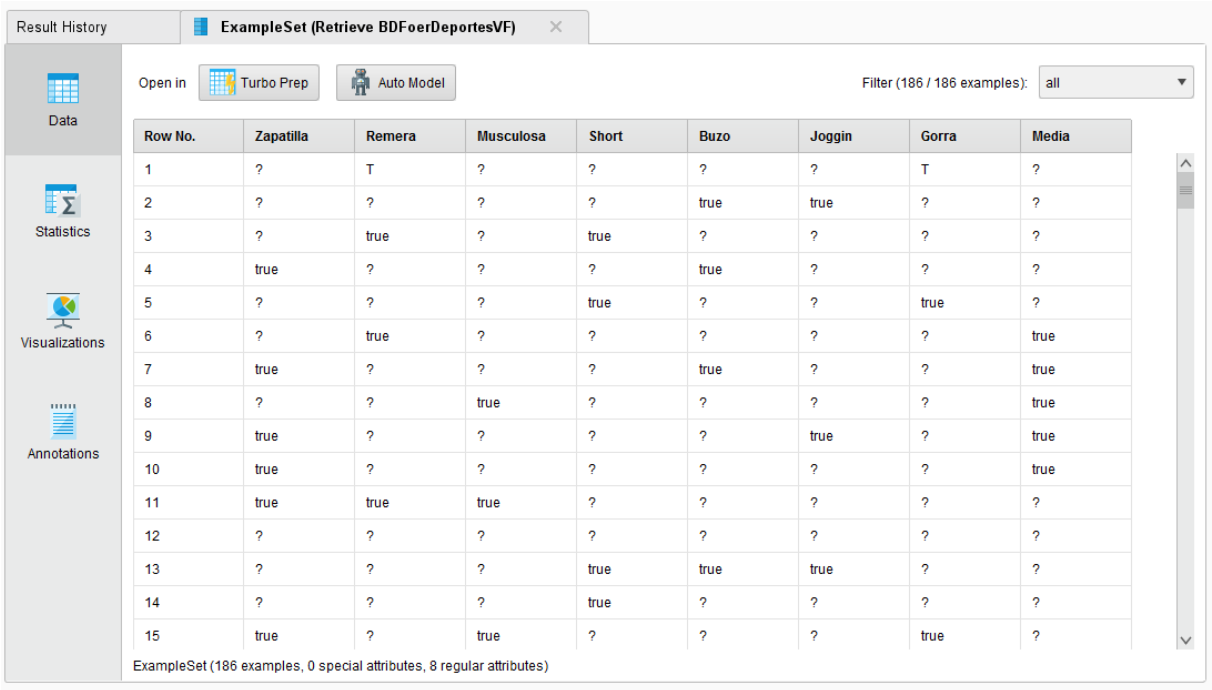


Imagen 4-9: Operador Retrive de RapidMiner

Como resultado de la ejecución de este proceso se obtiene una vista de los datos de la hoja de cálculo *foerdeportes_carga.xlsx* como se muestra en la imagen 4-10:



Row No.	Zapatilla	Remera	Musculosa	Short	Buzo	Joggin	Gorra	Media
1	?	T	?	?	?	?	T	?
2	?	?	?	?	true	true	?	?
3	?	true	?	true	?	?	?	?
4	true	?	?	?	true	?	?	?
5	?	?	?	true	?	?	true	?
6	?	true	?	?	?	?	?	true
7	true	?	?	?	true	?	?	true
8	?	?	true	?	?	?	?	true
9	true	?	?	?	?	true	?	true
10	true	?	?	?	?	?	?	true
11	true	true	true	?	?	?	?	?
12	?	?	?	?	?	?	?	?
13	?	?	?	true	true	true	?	?
14	?	?	?	true	?	?	?	?
15	true	?	true	?	?	?	true	?

Imagen 4-10: Lectura de datos con el operador Retrive en RapidMiner

Como primer análisis para entender la distribución de los datos, en la opción statistics de RapidMiner se observa el total de vendidos por cada artículos (Imagen 4-11), este resultado se resume en la tabla 4-13.

Producto	Vendidos (true)
Zapatilla	81
Remera	42
Musculosa	23
Short	36
Buzo	24
Jogging	22
Gorra	41
Media	71

Tabla 4-13: Cantidad de Productos Vendidos

4.2.4 Verificación de la calidad de los datos

En cuanto a la calidad de los datos se reconoce que la variable true es una representación significativa, adecuada para el proceso de reglas de asociación, el formato de datos es polinomial y se observa que existen valores faltantes identificados con el signo (?) en la imagen 4-10.

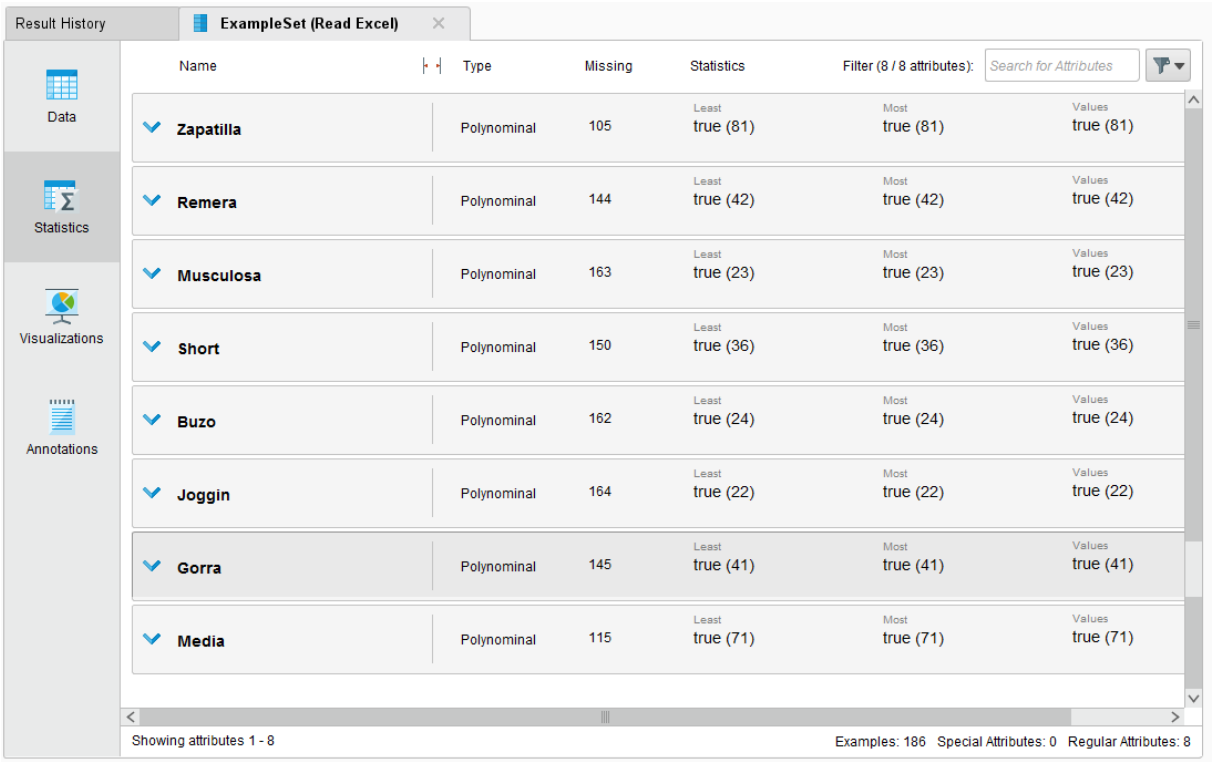


Imagen 4-11: Distribución de los datos recolectados en RapidMiner

4.3 Fase 3: Preparación de los datos

En esta fase se desarrollan las siguientes tareas para lograr la vista minable:

4.3.1 Selección de los datos

Se seleccionan todos los datos cargados ya que estos son relevantes para cumplir con el objetivo de minería de datos, tanto en calidad, cantidad, tipo de dato y atributos. Si bien el volumen de datos puede plantearse como pequeño, si lo consideramos suficiente para el estudio que pretendemos realizar.

4.3.2 Limpieza de los datos

Sin datos limpios, los resultados de un análisis de minería de datos están en duda, por lo que se abordaran los problemas de calidad de los datos (informados en la Fase 2 - Verificación de la calidad de los datos) de la siguiente manera:

- Para los valores faltantes identificados con el signo de pregunta, se completan con la palara false, ya que solo se cargaron los productos vendidos con true.

Esto se soluciona con el operador *"Map"* (Imagen 4-12), que mediante el parámetro value mappings se ingresa a la lista *qué* palabras reemplazar por *otra* (Imagen 4-13).

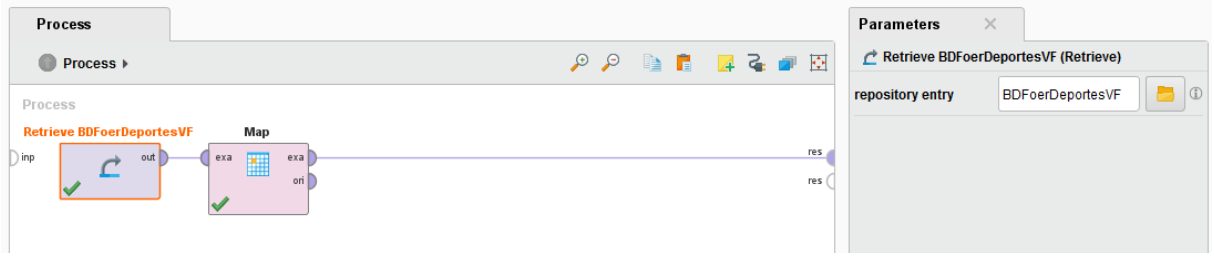


Imagen 4-12: Uso del operador Map para corregir datos en RapidMiner

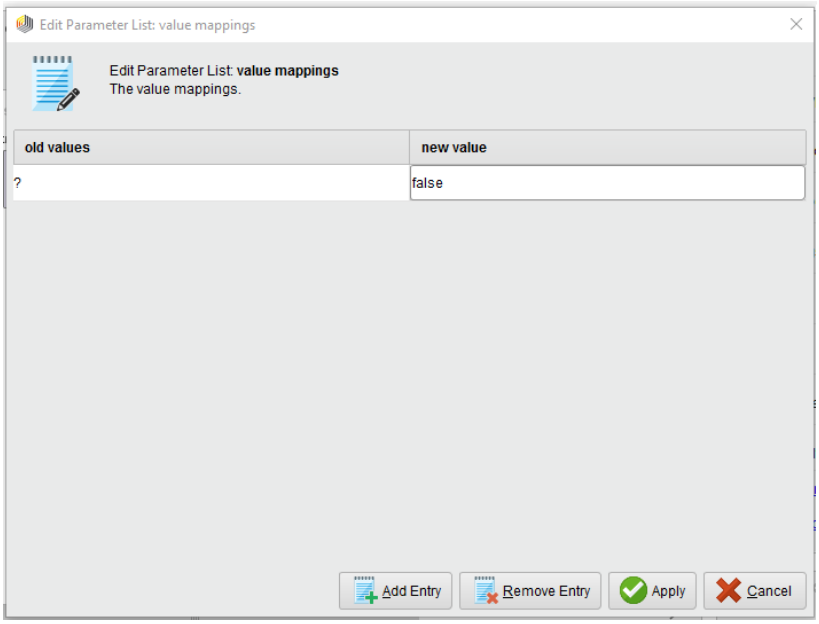


Imagen 4-13: Edición de valores con en el operador Map

El resultado final de ejecutar el proceso queda de la siguiente manera (Imagen 4-14):

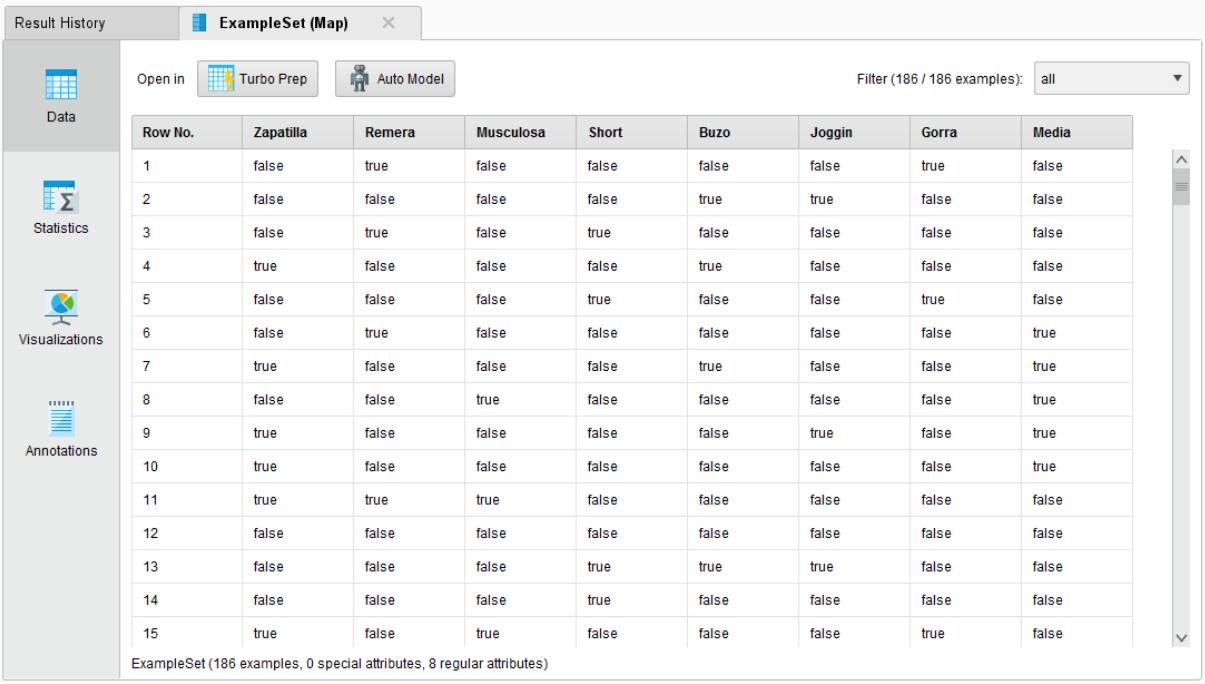


Imagen 4-14: Resultados de los datos al ejecutar el proceso con el operador Map

En general los algoritmos de minería de reglas de asociación requieren como entrada un conjunto de datos del tipo binomial, por lo que es necesario unificar el repositorio *BDFoerDeportesVF* en un espacio de variables booleanas. Para lograrlo se agrega el operador "*Nominal to Binomial"* que retorna una notación unificada booleana (Imagen 4-15).

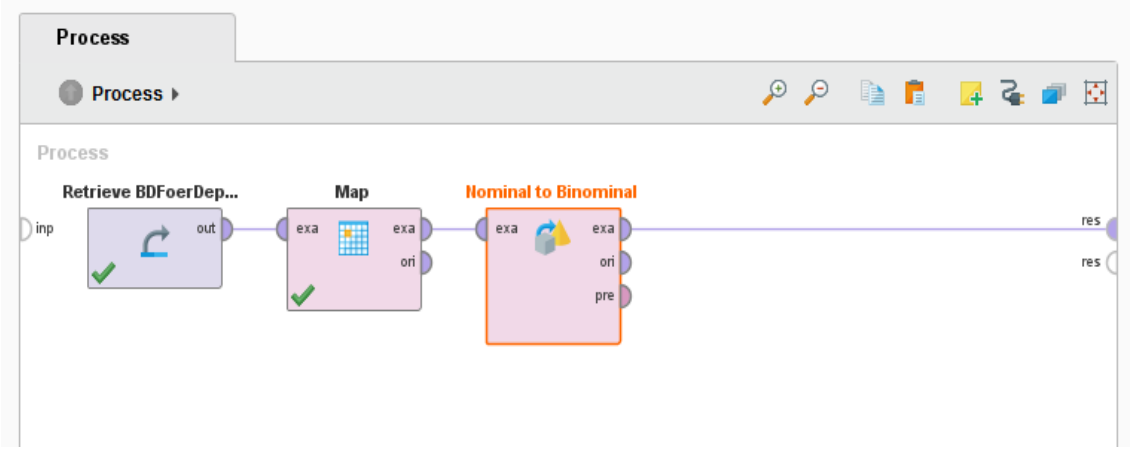


Imagen 4-15: Operador Nominal to Binomial en RapidMiner

Los resultados de la preparación de los datos quedan como se muestra en la imagen 4-16

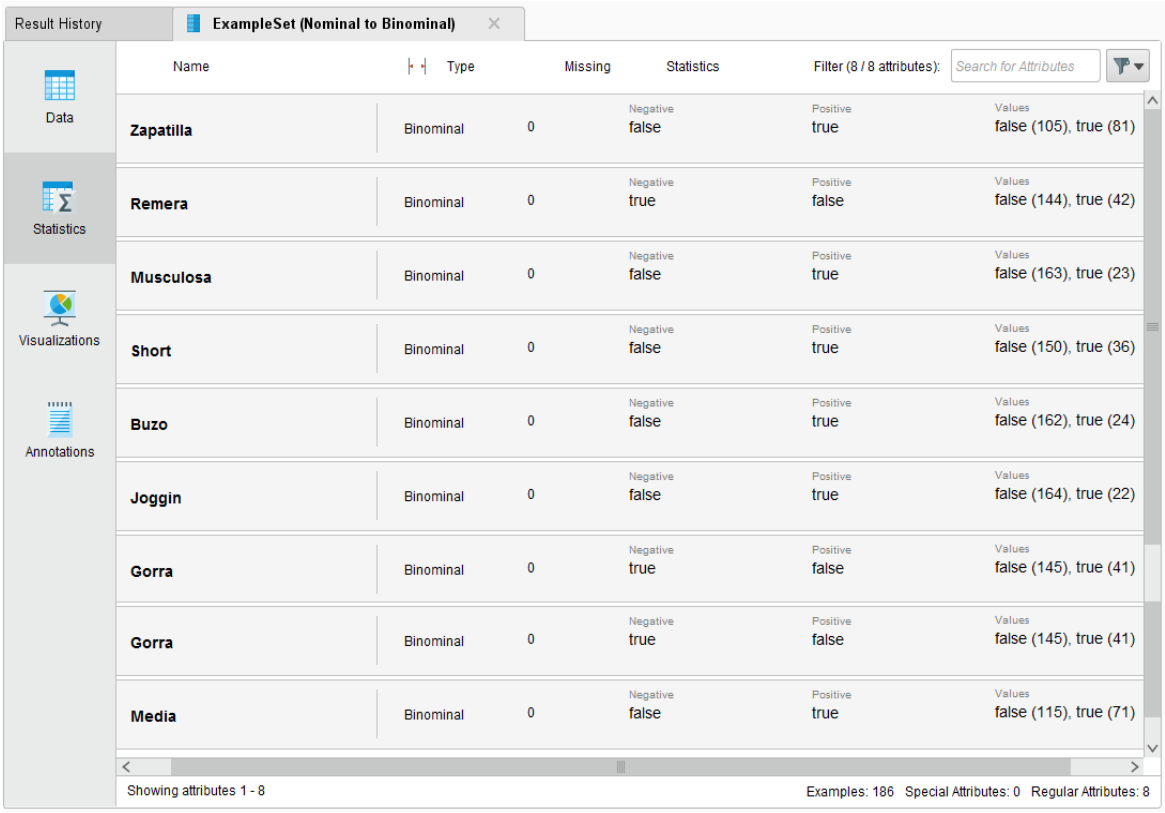


Imagen 4-16: Resultados de la preparación de los datos

Finalmente se logra definir en esta instancia los atributos, el tipo de datos y el valor con el que se creó la vista minable para la fase de modelado, la misma se resume en la tabla 4-14.

4.3.3 Formateo de los datos

En esta instancia se realiza un reporte final de la calidad de los datos y se concluye que el formato y el diseño de los datos es el adecuado, los mismos están listos y preparados para ser usados en la fase de modelado

Atributo	Descripción	Tipo	Valor
Zapatilla	Representa el artículo zapatilla mediante un valor true o false.	Binomial	True o False
Remera	Representa el artículo remera mediante un valor true o false.	Binomial	True o False
Musculosa	Representa el artículo musculosa mediante un valor true o false.	Binomial	True o False
Short	Representa el artículo short mediante un valor true o false.	Binomial	True o False
Buzo	Representa el artículo buzo mediante un valor true o false.	Binomial	True o False
Jogging	Representa el artículo jogging mediante un valor true o false.	Binomial	True o False
Gorra	Representa el artículo gorra mediante un valor true o false.	Binomial	True o False
Media	Representa el artículo media mediante un valor true o false.	Binomial	True o False

Tabla 4-14: Vista minable

4.4 Fase 4: Modelado

En esta instancia a la vista minable obtenida se le aplica técnicas para la construcción del modelo que responda como solución al problema.

4.4.1 Selección de la técnica de modelado

De acuerdo al tipo de tarea a realizar, se selecciona la técnica de minería de datos. La tarea de encontrar productos que se compran juntos en las transacciones corresponde al tipo de tarea descriptiva y para resolverla se emplea la técnica de extracción de reglas de asociación haciendo uso del algoritmo FP-Growth.

Se selecciona este algoritmo por ser de aprendizaje no supervisado y basado en reglas de asociación, fácil de implementar, entender y útil en grandes cantidades de datos. Cabe destacar que FP-Growth funciona exactamente igual que el algoritmo A priori, pero A priori se descarta solo porque RapidMiner no lo admite entre sus operadores. Además, teniendo en cuenta que la representación de nuestra base de datos es horizontal, se descarta el algoritmo Eclat porque éste necesita de una representación vertical de los datos, seleccionando así el algoritmo FP-Growth como el más adecuado para el modelado. En el Capítulo 2 - Sección 2.5 se complementa la información sobre estos algoritmos.

4.4.2 Creación del modelo

En esta instancia el analista de datos utiliza una serie de operadores y el conjunto de datos preparado para crear el modelo de asociación. La imagen 4-17 muestra el resultado final del modelado para la generación de reglas de asociación en RapidMiner.

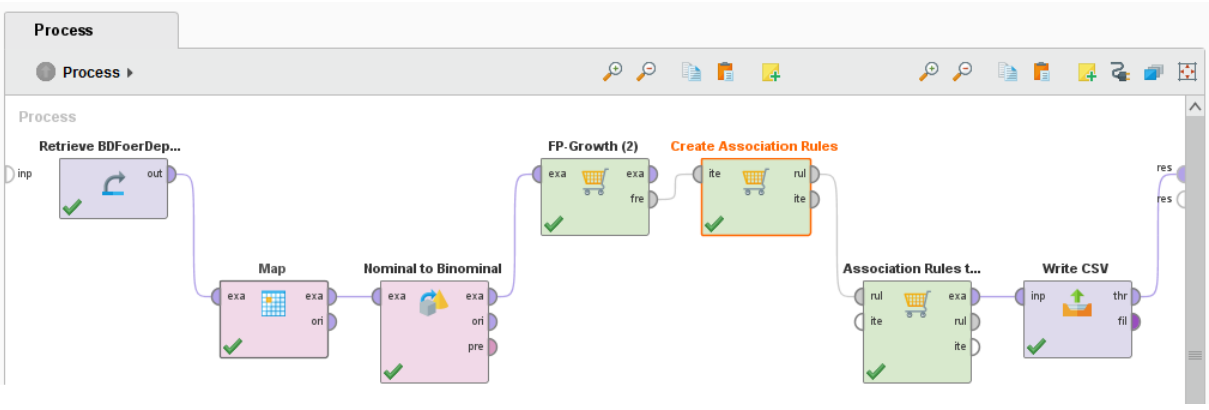


Imagen 4-17: Modelado para la generación de reglas de asociación en RapidMiner

Se recuerda que para generar las reglas de asociación, primero se necesita encontrar los conjuntos de elementos frecuentes por eso se usa el operador:

- FP-Growth (Imagen 4-18).

  Calcula y muestra todos los conjuntos de elementos frecuentes de un conjunto de datos, de manera eficiente y utilizando la estructura de datos de árbol FP, para ello es obligatorio que todos sus atributos de entrada sean binominales. (RapidMiner Documentation, 2021)

Imagen 4-18: Operador FP-Growth en RapidMiner. Fuente: https://docs.rapidminer.com/8.0/studio/operators/modeling/associations/fp_growth.html

El operador FP-Growth trabaja de dos maneras:

1. Encontrar al menos el número especificado de conjuntos de elementos con el soporte más alto sin tener en cuenta el 'soporte mínimo'. Este modo está disponible cuando el parámetro de número mínimo de búsqueda de conjuntos de elementos se establece en verdadero. Luego, este operador encuentra el número de conjuntos de elementos especificado en el parámetro de número mínimo de conjuntos de elementos. En este caso, se ignora el parámetro de soporte mínimo.

2. Encontrar todos los conjuntos de elementos con un soporte mayor que el soporte mínimo especificado. El soporte mínimo se especifica mediante el parámetro de soporte mínimo. Este modo está disponible cuando el parámetro de búsqueda de número mínimo de conjuntos de elementos se establece en falso.

Para el modelo se usa el segundo modo de trabajo por lo que se prueban diferentes valores para el parámetro min_support en RapidMiner, los resultados obtenidos son los siguientes:

- Con un min_support de 0.9. Buscamos que el soporte de los items que aparezcan juntos con frecuencia sean mayor o igual al min_support. Se observa en la imagen 4-19 que para la cesta de compra en estudio no existen itemset frecuentes que cumplan con el umbral min_support (0.9), esto sucede porque el soporte es muy alto; por lo que el valor de esta medida se descarta.

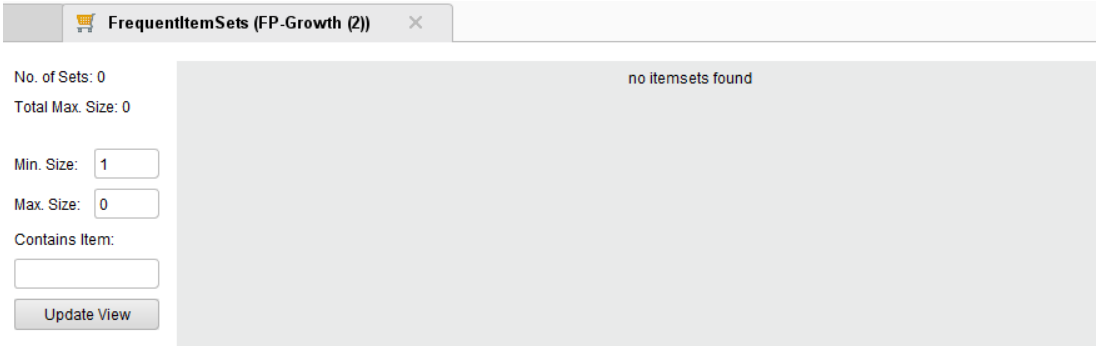
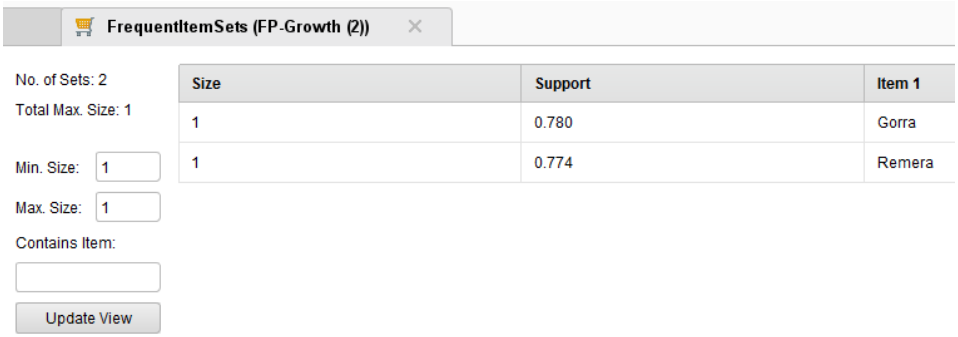


Imagen 4-19: Itemset frecuentes con min_support de 0.9

- Con un min_support de 0.6. Buscamos que el soporte de los items que aparezcan juntos con frecuencia sean mayor o igual al min_support.

Se observa en la imagen 4-20 que para la cesta de compra en estudio existen 2 itemset frecuentes que superan el umbral min_support (0.6), pero su cantidad no es suficiente para generar las reglas de asociación, por lo que el valor de esta medida se descarta.

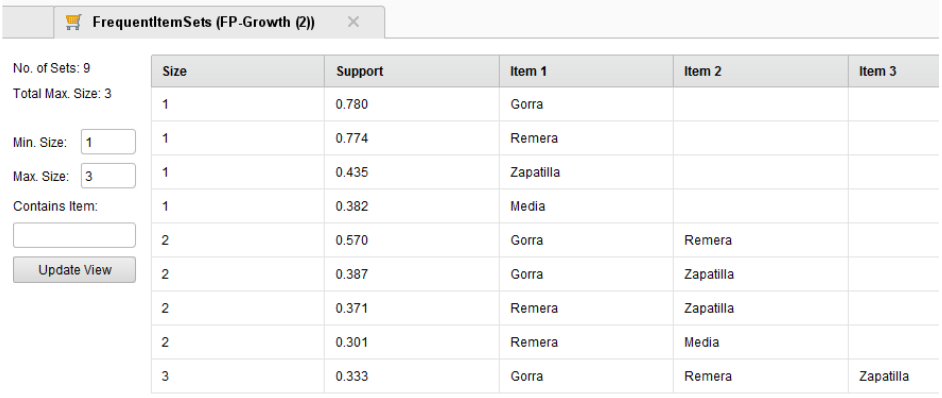


Size	Support	Item 1
1	0.780	Gorra
1	0.774	Remera

Imagen 4-20: Itemset frecuentes con min_support de 0.6

- Con un min_support de 0.3. Buscamos que el soporte de los items que aparezcan juntos con frecuencia sean mayor o igual al min _soporte (0.3).

Se observa en la imagen 4-21 que para la cesta de compra en estudio existen 9 itemset frecuentes que superan el umbral min_support (0.3), este valor no se descarta y se lo toma como un candidato a elegir.



Size	Support	Item 1	Item 2	Item 3
1	0.780	Gorra		
1	0.774	Remera		
1	0.435	Zapatilla		
1	0.382	Media		
2	0.570	Gorra	Remera	
2	0.387	Gorra	Zapatilla	
2	0.371	Remera	Zapatilla	
2	0.301	Remera	Media	
3	0.333	Gorra	Remera	Zapatilla

Imagen 4-21: Itemset frecuentes con min_support de 0.3

- Con un min_support de 0.1. Buscamos que el soporte de los items que aparezcan juntos con frecuencia sean mayor o igual al min _soporte (0.31).

Se observa en la imagen 4-22 que para la cesta de compra en estudio existen 28 itemset frecuente que superan el umbral min_support (0.1), este valor no se descarta y se lo toma como un segundo candidato a elegir.

Al variar el parámetro min_support entre 0 y 1, se deduce que a medida que el umbral de min_support se acerca a cero se encuentra una mayor cantidad de itemset frecuentes, esto puede incrementar el número de candidatos y la longitud de los itemset frecuentes.

FrequentItemSets (FP-Growth (2))

No. of Sets: 28
Total Max. Size: 4

Min. Size: 1
Max. Size: 4
Contains Item:

Update View

Size	Support	Item 1	Item 2	Item 3	Item 4
1	0.780	Gorra			
1	0.774	Remera			
1	0.435	Zapatilla			
1	0.382	Media			
1	0.194	Short			
1	0.129	Buzo			
1	0.124	Musculosa			
1	0.118	Joggin			
2	0.570	Gorra	Remera		
2	0.387	Gorra	Zapatilla		
2	0.296	Gorra	Media		
2	0.140	Gorra	Short		
2	0.124	Gorra	Buzo		
2	0.102	Gorra	Musculosa		
2	0.118	Gorra	Joggin		
2	0.371	Remera	Zapatilla		
2	0.301	Remera	Media		
2	0.172	Remera	Short		
2	0.118	Remera	Buzo		
2	0.108	Remera	Musculosa		
2	0.194	Zapatilla	Media		
3	0.333	Gorra	Remera	Zapatilla	
3	0.226	Gorra	Remera	Media	
3	0.124	Gorra	Remera	Short	
3	0.113	Gorra	Remera	Buzo	
3	0.151	Gorra	Zapatilla	Media	
3	0.161	Remera	Zapatilla	Media	
4	0.129	Gorra	Remera	Zapatilla	Media

Imagen 4-22: Itemsets frecuentes con min_support de 0.1

Neves (2003) recomienda que, como parámetros de entrada del algoritmo, se defina un valor bajo para el soporte y un valor elevado para la confianza:

> De esta forma, en primer lugar se genera una mayor cantidad de conjuntos frecuentes y una gran cantidad de reglas, posteriormente se verifica la cohesión de las mismas a través de la medida de confianza. Una regla de asociación con un valor de confianza bajo no expresará un patrón de comportamiento en los datos y, por otra parte, un valor de soporte muy elevado probablemente llevaría a la perdida de patrones. (p.10)

Bajo las recomendaciones de Neves, se toman los conjuntos de elementos frecuentes que superan el umbral mínimo del 0.3 y 0.1, y se procede a generar las reglas de asociación fuertes mediante el operador Create Association Rules.

- Create Association Rules (Imagen 4-23): “Genera un conjunto de reglas de asociación, por lo que necesita como entrada el conjunto de elementos frecuentes dados por el operador FP-Growth” (RapidMiner Documentation, 2021)

Imagen 4-23: Operador Create Association Rules en RapidMiner. Fuente: https://docs.rapidminer.com/latest/studio/operators/modeling/associations/create_association_rules.html

Las reglas de asociación son declaraciones si / entonces que ayudan a descubrir relaciones entre datos aparentemente no relacionados. Un ejemplo de una regla de asociación sería "Si un cliente compra huevos, es 80% probable que también compre leche". Una regla de asociación tiene dos partes, un antecedente (si) y un consecuente (entonces). Un antecedente es un elemento (o conjunto de elementos) que se encuentra en los datos. Un consecuente es un elemento (o conjunto de elementos) que se encuentra en combinación con el antecedente.

Las reglas de asociación se crean analizando los datos en busca de patrones frecuentes de si / entonces y utilizando el *soporte de* criterios y la *confianza* para identificar las relaciones más importantes. El soporte es una indicación de la frecuencia con la que aparecen los elementos en la base de datos. La confianza indica el número de veces que se ha determinado que las afirmaciones si / entonces son verdaderas.

Por el criterio de éxito de éxito de minería de datos (Capitulo4 - Sección 4.1.3.2), el modelo propuesto debe generar las reglas de asociación de los productos con al menos un 80% de confianza, por lo que se configura el parámetro (min_confidence) del operador Create Association Rules a un 80% y se prueban los conjuntos de elementos frecuentes correspondientes al min_support de 0.3 y 0.1:

- Modelo de reglas de asociación con un mínimo soporte de 0.3 y una mínima confianza de 80% (tabla 4-15).

Antecedente	Consecuente	Soporte	Confianza	Lift
Zapatilla	Remera	0.37	0.85	1.10
Gorra, Zapatilla	Remera	0.33	0.86	1.11
Zapatilla	Gorra	0.39	0.89	1.14
Remera, Zapatilla	Gorra	0.33	0.90	1.15

Tabla 4-15: Reglas de asociación con min_support 0.3 y min_confidence 80%

- Modelo de reglas de asociación con un mínimo soporte de 0.1 y una mínima confianza de 80% (tabla 4-16).

Antecedente	Consecuente	Soporte	Confianza	Lift
Remera, Zapatilla, Media	Gorra	0.13	0.8	1.03
Musculosa	Gorra	0.1	0.83	1.06
Zapatilla, Media	Remera	0.16	0.83	1.08
Zapatilla	Remera	0.37	0.85	1.1
Gorra, Zapatilla, Media	Remera	0.13	0.86	1.11
Gorra, Zapatilla	Remera	0.33	0.86	1.11
Musculosa	Remera	0.11	0.87	1.12
Gorra, Short	Remera	0.12	0.88	1.14
Zapatilla	Gorra	0.39	0.89	1.14
Short	Remera	0.17	0.89	1.15
Remera, Zapatilla	Gorra	0.33	0.9	1.15
Gorra, Buzo	Remera	0.11	0.91	1.18
Buzo	Remera	0.12	0.92	1.18
Remera, Buzo	Gorra	0.11	0.95	1.22
Buzo	Gorra	0.12	0.96	1.23
Jogging	Gorra	0.12	1	1.28
Buzo	Gorra, Remera	0.11	0.88	1.54

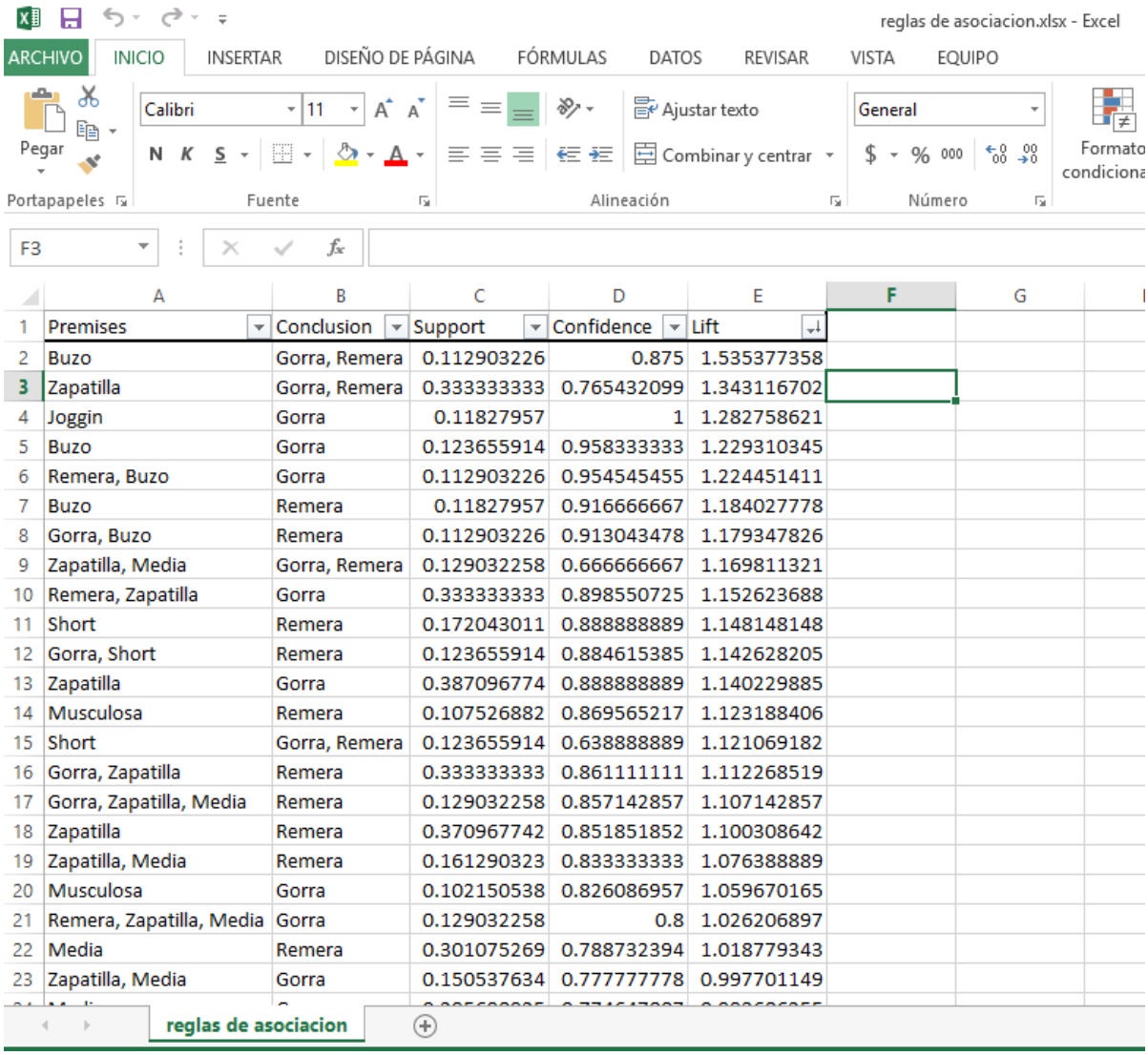
Tabla 4-16: Reglas de asociación con min_support 0.1 y min_confidence 80%

- Association Rules to ExampleSet (Imagen 4-24): "Convierte las reglas de asociación en un set de datos u hojas de datos, esto es necesario para la entrada del operador WriteCSV. Para encontrar este operador en RapidMiner es necesario instalar la extensión Converters" (RapidMiner Documentation, 2020).

Imagen 4-24: Operador Association Rules to Example Set de RapidMiner

- WriteCSV (Imagen 4-25): "Permite escribir un archivos csv, indicando el destino de almacenamiento local de la pc, el operador proporciona como objeto de salida el archivo *reglasdeasoc.csv* con las reglas obtenidas (Imagen 4-26)" (RapidMiner Documentation, 2020).

Imagen 4-25: Operador Write CSV de RapidMiner



	A	B	C	D	E	F	G
1	Premises	Conclusion	Support	Confidence	Lift		
2	Buzo	Gorra, Remera	0.112903226	0.875	1.535377358		
3	Zapatilla	Gorra, Remera	0.333333333	0.765432099	1.343116702		
4	Joggin	Gorra	0.11827957	1	1.282758621		
5	Buzo	Gorra	0.123655914	0.958333333	1.229310345		
6	Remera, Buzo	Gorra	0.112903226	0.954545455	1.224451411		
7	Buzo	Remera	0.11827957	0.916666667	1.184027778		
8	Gorra, Buzo	Remera	0.112903226	0.913043478	1.179347826		
9	Zapatilla, Media	Gorra, Remera	0.129032258	0.666666667	1.169811321		
10	Remera, Zapatilla	Gorra	0.333333333	0.898550725	1.152623688		
11	Short	Remera	0.172043011	0.888888889	1.148148148		
12	Gorra, Short	Remera	0.123655914	0.884615385	1.142628205		
13	Zapatilla	Gorra	0.387096774	0.888888889	1.140229885		
14	Musculosa	Remera	0.107526882	0.869565217	1.123188406		
15	Short	Gorra, Remera	0.123655914	0.638888889	1.121069182		
16	Gorra, Zapatilla	Remera	0.333333333	0.861111111	1.112268519		
17	Gorra, Zapatilla, Media	Remera	0.129032258	0.857142857	1.107142857		
18	Zapatilla	Remera	0.370967742	0.851851852	1.100308642		
19	Zapatilla, Media	Remera	0.161290323	0.833333333	1.076388889		
20	Musculosa	Gorra	0.102150538	0.826086957	1.059670165		
21	Remera, Zapatilla, Media	Gorra	0.129032258	0.8	1.026206897		
22	Media	Remera	0.301075269	0.788732394	1.018779343		
23	Zapatilla, Media	Gorra	0.150537634	0.777777778	0.997701149		

Imagen 4-26: Archivo *reglasdeasoc.csv* con las reglas del modelo de asociación

4.5 Fase 5: Evaluación

4.5.1 Evaluación de los resultados de la minería de datos

A continuación se realiza la evaluación de los resultados de la minería de datos en base su criterio de éxito (Capitulo 4 - Sección 4.1.3.2).

Para cada uno de los modelos de reglas de asociación obtenidos del conjunto de elemento frecuente con min_support de 0.3 y 0.1 (Capitulo 4 sección 4.4.2) se obtienen diferentes valores para sus parámetros que permiten determinar su nivel de precisión, considerando el lift como tercer medida. Estos valores se observan como resumen en la tabla 4-17.

Modelo	min soporte	min confianza	min lift	máx. lift
FP-Growth	0.3	80%	1.1	1.15
FP-Growth	0.1	80%	1.03	1.54

Tabla 4-17: Resumen de los parámetros para los modelos con min_support de 0.3 y 0.1

El primer modelo parte de un conjunto de elementos frecuentes que supera el mínimo soporte de 0.3 y la mínima confianza del 80%. Sin embargo obtener reglas con un valor de confianza

alto pueden ser un tanto engañoso o pueden haber sido generadas por casualidad, esto se debe muchas veces por la presencia de algunos items estadísticamente independientes, por lo que la confianza no llega a ser suficiente para medir la utilidad de la regla. Para detectar estas reglas espurias se usa la métrica de correlación llamada lift, que nos permite medir que tan lejos de la independencia está el antecedente y el consecuente.

En la tabla 4-15 (Capitulo 4 - Seccion4.4.2) se muestran las reglas con sus respectivos valores de lift (en forma ascendente) y se observan que varían desde 1,1 hasta 1,15. Para una visión práctica, en la tabla 4-16 solo se muestra el valor mínimo y máximo que tomó el lift para dicho modelo. Recordando que:

- Si el lift $(A \rightarrow B) = 1$, se asume que A y B están independientes y que no existe correlación entre ellos. (Capitulo 2 - Sección 2.6)

Por lo tanto, el valor del lift cercano a 1 indica que la calidad de la regla no es buena, por lo que las reglas del modelo uno quedan descartadas, remitiéndonos al siguiente modelo.

El segundo modelo parte de un conjunto de elementos frecuentes que supera el mínimo soporte de 0.1 y la mínima confianza del 80%. Si bien las reglas obtenidas varían entre una confianza del 80% al 100%, no es suficiente para medir la utilidad de la misma; por lo que también se usa el lift para encontrar las reglas que realmente son interesantes.

En la tabla 4-16 (capitulo 4 seccion4.4.2) se muestran las reglas con sus respectivos valores de lift (en forma ascendente) y se observan que varían desde 1,1 hasta 1,54. Para una visión práctica, en la tabla 4-16 sólo se muestra el valor mínimo y máximo que tomó el lift, recordando que:

- Si el lift $(A \rightarrow B) > 1$, A y B están correlacionadas positivamente, lo que significa que la ocurrencia de uno implica la ocurrencia del otro.
- Si el lift $(A \rightarrow B) < 1$, la ocurrencia de A esta correlacionado negativamente con la ocurrencia de B, lo que significa que la ocurrencia de uno probablemente conduce a la ausencia del otro.
- Si el lift $(A \rightarrow B) = 1$, A y B están independientes y no existe correlación entre ellos.

Bajo estos criterios sólo se toman la reglas que esta correlacionadas positivamente, es decir que la ocurrencia de un elemento implica la ocurrencia de otro, con un lift mayor o igual a 1.2 (tabla 4-18)

Antecedente	Consecuente	Soporte	Confianza	Lift
Buzo	Gorra, Remera	0.11	0.88	1.54
Jogging	Gorra	0.12	1	1.28
Buzo	Gorra	0.12	0.96	1.23
Remera, Buzo	Gorra	0.11	0.95	1.22

Tabla 4-18: Reglas interesante con un lift mayor a 1.2

Se interpreta que los clientes que han comprado buzo también compraron gorra y remera, con un soporte de 0,11 y una confianza de 0.88. Esto indica que en el 11% de las transacciones contienen ambas referencias juntas. Además, la confianza indica que en el 88% de los casos

que se compra buzo también se compra gorra, remera. Un lift de 1.54 mejora la confianza e indica que cuando aparece buzo aparece gorra y remera, más de una vez de lo que se podría esperar por azar. De forma análoga se interpretan las reglas de la tabla 4-18.

Se selecciona como mejor modelo el resultante de aplicar la técnica FP-Growth con un soporte de 0,1. Para elegir este modelo se tuvo en cuenta que cumple con el objetivo y con el criterio de éxito de minería de datos: obtener un modelo que permita las relaciones entre los productos de una cesta de compra, a través de reglas de asociación, que tengan una confianza mayor al 80%.

4.5.2 Determinación de los próximos pasos

Obtenido el modelo que cumple con los objetivos de minería de datos y luego de realizar la evaluación del mismo y visualización de sus resultados, se toma la decisión de implementar la

El siguiente paso es la planificación de la implementación del "informe final de recomendaciones" que describe la interpretación del modelo encontrado en una serie de recomendaciones para que el dueño del negocio lo comprenda en un lenguaje natural y de esta forma lo implemente en su negocio físico y lo refleje en una tienda online. Quedando bajo la responsabilidad del interesado la implementación de informe.

4.6 Fase 6: Despliegue

Es la fase final del proyecto de minería de datos y tiene por objetivo presentar al cliente el conocimiento obtenido de forma organizada para que pueda usarlo en los procesos de toma de decisión de la indumentaria, como así también en la personalización de una página web. El cliente es quien lleva a cabo los pasos de implementación por lo que es importante que comprenda de antemano qué acciones deben tomarse para hacer uso del informe del modelo creado.

4.6.1 Plan de despliegue

El analista de datos entrega el modelo de asociación obtenido al líder del proyecto como un archivo *reglasdeasoc.csv* que contiene el conjunto total de reglas identificadas de la cesta de compra.

El líder del proyecto será el responsable de aplicar las medidas necesarias para filtrar las reglas interesantes que aporten valor al conocimiento. Y será el responsable de describir la interpretación de estas reglas en una serie de recomendaciones y en un lenguaje natural para que de esta forma el dueño del negocio pueda comprenderlas.

El dueño del negocio será el responsable de recibir el informe de recomendaciones para usarlo en la toma de decisiones de ventas en local y en el desarrollo de una de tienda online que refleje las reglas y recomendaciones expresadas en el informe.

Para una mejor compresión se detalla en un diagrama de flujo el proceso de implementación o despliegue del modelo de reglas de asociación (Imagen 4-27)

4.6.2 Entregables

4.6.2.1 Informe final

El informe final de recomendaciones tendrá el formato de la imagen 4-28. El mismo expresa una serie de recomendaciones basado en el modelo encontrado, para ser integrados y aplicado tanto al negocio como a una tienda online.

4.6.2.1 Prototipo

Con el objetivo de mostrar cómo se haría uso del informe de recomendaciones basado en el modelo de regla de asociación, para iniciarse en e-commerce, se desarrolla el prototipo de una tienda online. En el anexo B se detalla toda la información relacionada con el prototipo, desde su interfaz, forma de implementación y funcionamiento.

El prototipo refleja las reglas obtenidas de asociación entre los productos, a través de cross selling, recomendaciones de productos y ofertas. Demostrando así la posibilidad de iniciarse en e-commerce mediante el uso de una tienda online como una de las herramientas potentes que ayudan a un negocio a incrementar las ventas y llegar a más clientes.

La forma en la que se integra el modelo con la tienda es detectando que en el informe de recomendaciones se encuentran las reglas interesantes expresadas como X → Y, técnicamente en la programación de la tienda online, a la configuración del producto X se le agrega el producto Y como recomendación o un producto distinto para venta cruzada (cross selling).

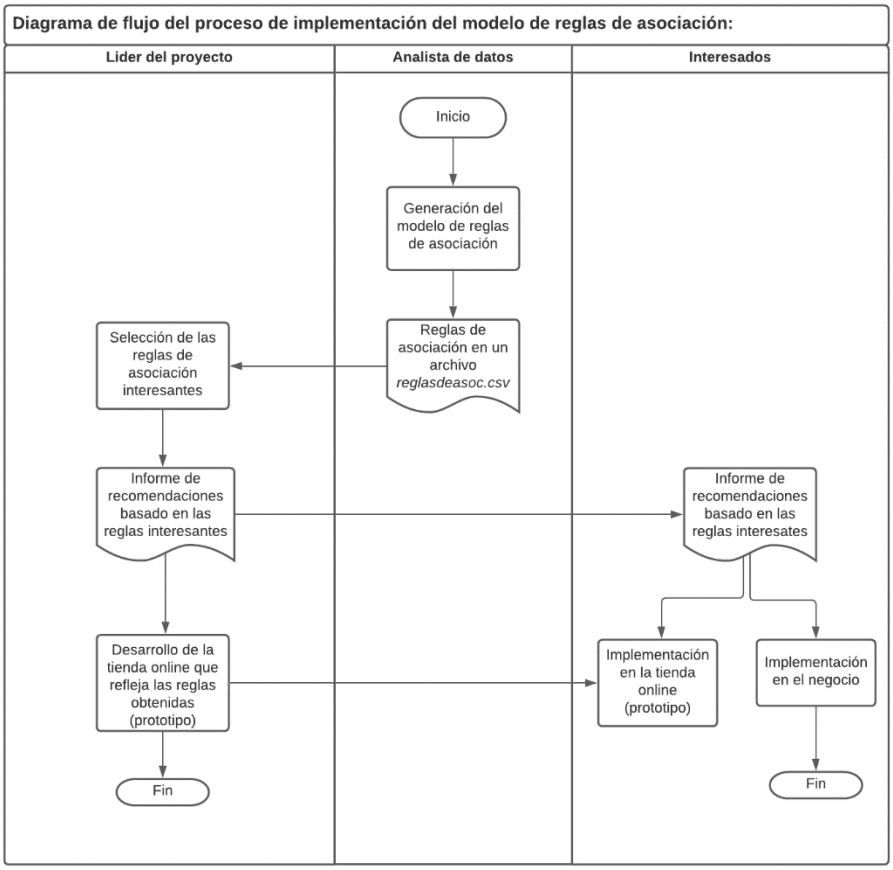


Imagen 4-27: Diagrama de flujo del proceso de implementación del modelo

INFORME DE RECOMENDACIONES

Reglas obtenidas de la cesta de compra:

Con una confianza mínima del 80%:

[Buzo] → [Gorra, Remera]
[Jogging] → [Gorra]
[Buzo] → [Gorra]
[Remera, Buzo] → [Gorra]

Estas reglas indican que:

Un consumidor que compra buzo es probable que compre gorra y remera.
Un consumidor que compra jogging es probable que compre gorra.
Un consumidor que compra buzo es probable que compre gorra.
Un consumidor que compra remera y buzo es probable que compre gorra.

Tomando como base este conocimiento sobre el comportamiento de los clientes se puede tomar decisiones gerenciales para mejorar las ventas, con respecto:

- En el negocio:
  - Reordenamiento de los estantes para levantar las ventas de otro stock tomando al antecedente de la regla. Los productos con menor venta (los que no aparecen en la regla) colocarlos cerca de los más vendidos. Esta cercanía provoca que el cliente que compra con mayor frecuencia un determinado producto visualice y tenga en cuenta otro producto determinado. Si a este comportamiento se le aplica una oferta o descuento al producto menos vendido se genera una mayor atención del cliente logrando así influir en su compra.
  - Reordenamiento de los estantes para levantar las ventas de otro stock tomando al antecedente y el consecuente de la regla. Los productos con menor venta (los que no aparecen en la regla) colocarlos entre medio de los estantes del antecedente y el consecuente. Es decir, que si un cliente compra con mayor frecuencia un determinado producto y a sabiendas que con mayor probabilidad se encamina hacia la búsqueda otro producto determinado, colocar en ese trayecto el o los productos con menor venta. Si a este comportamiento se le aplica una oferta o descuento al producto menos vendido hará que el cliente se detenga en el trayecto y lo observe detenidamente con la posibilidad de adquirirlo, logrando así influir en su carrito de compra.
  - Reordenamiento de la vidriera. Vestir los maniquíes con los productos que se compran juntos (como indican las reglas) en combinación con el resto de los artículos para motivar visualmente al cliente.
  - Aplicación de la técnica cross selling o venta cruzada. Esta técnica puede ayudar a mejorar las ventas mediante el uso del modelo de reglas de asociación obtenido. Al producto que ya está comprando o acaba de comprar el cliente, promocionarle artículos complementarios, considerando los que no tienen tanta demanda de compra con el fin de levantar las ventas de dicho stock.
- En e-commerce
  - Diseño de la tienda online:
    - Saber el comportamiento y gusto del cliente, permite ofrecerle de forma rápida lo que está buscando, por lo que mostrar en la pantalla principal las categorías de los productos que el cliente adquiere con mayor frecuencia permite agilizar su búsqueda.
    - Orden descendente de las categorías de productos en el menú principal, de tal forma que se encuentren los productos más vendidos primeros, ya que estos son los que ayudan a influir en la compra de otros productos.
  - Productos recomendados. Mostrar al usuario productos relacionados al que está interesado en agregar al carrito, es decir seleccionado un producto mostrar otro producto que tiene una mayor probabilidad de ser comprado también. Esta comportamiento representa la regla "si X entonces Y"
  - Aplicación de la técnica cross selling o venta cruzada. Esta técnica puede ayudar a mejorar las ventas online mediante el uso del modelo de reglas de asociación obtenido. Consiste en promocionarle al usuario artículos complementarios al que ya está agregado en el carrito de compras; considerando los que no tienen tanta demanda con el fin de levantar las ventas de dicho stock.

Imagen 4-28: Formato del informe de recomendaciones en base al modelo

5 Conclusión y futuras líneas de investigación

La aplicación de la minería de datos sobre una cesta de compra para lograr patrones de asociación, permite realizar este proceso de manera más eficiente de lo que a simple vista es difícil de identificar. Contar con estos patrones de asociación sobre el comportamiento de los clientes permite mejorar la toma de decisión gerencial sobre marketing y emplearlo de igual manera en la web.

El objetivo propuesto fue alcanzado al haber obtenido mediante el algoritmo *FP Growth* un modelo capaz de realizar la asociación de los artículos con al menos un 80% de confianza. El modelo encontrado supera el 80% de confianza y permite realizar la asociación de los artículos con un alto nivel de eficiencia e interés para el negocio (habiendo encontrado reglas con un lift mayor a 1.2 que indican que son positivas) y en un tiempo significativamente menor que el que emplearía una persona.

Como futuras líneas de investigación se podría considerar un sistema de gestión web para agilizar la carga de datos; y como mejor practica automatizar la integración del archivo de reglas de asociación .csv para que la tienda online entienda las reglas y las muestre en su interfaz con el usuario.

También se podría considerar otros atributos de los artículos como el talle, color, marca y otras características de la prenda que puedan enriquecer el análisis de minería de datos.

En cuanto a la metodología usada CRISP-DM se confirmó que los procesos de extracción, selección y limpieza de los datos son los que mayor tiempo conllevan, por lo que es importante brindarle atención y dedicación, como base para iniciar el proceso de minería de datos.

Teniendo en consideración lo mencionado y actuando en consecuencia, se podría lograr una mejor utilidad de los recursos del negocio, con la aplicación de la minería de datos de asociación, y obtener un resultado de mayor calidad en la gerencia.

6 Bibliografía

E.J. Gómez Aguilera, P. Chausa Fernández, C. Cáceres Taladriz, F García Alcaide, & J.M. Gatell Artigas. (2006). Extracción de reglas de asociación en una base de datos clínicos.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Reglas de asociación de minería entre conjuntos de elementos en grandes bases de datos. *Actas de la conferencia internacional ACM SIGMOD de 1993 sobre Gestión de datos.* Washington: DC.

Bramer, M. (2016). *Principles of Data Mining.* Portsmouth, Hampshire, UK: Springer.

CACE. (12 de 08 de 2020). *Camara de Comercio Argentino*. Obtenido de https://www.cace.org.ar/sobre-cace

Camara de Comercio Argentino. (27 de Agosto de 2020). *CACE*. Obtenido de https://www.cace.org.ar/noticias-record-de-audiencia-mas-de-13600-personas-de-todo-el-pais-se-conectaron-el-ecommerce-day-argentina-online-live-experience

Carrion, L. (9 de 10 de 2019). *Mailelay*. Obtenido de https://blog.mailrelay.com/es/2019/10/09/data-mining#Amazon_y_el_servicio_de_atencion_al_cliente

Clara, B. L. (2001). *Manual de derecho informático.* Santa Fe, Rosario, Argentina: Ed. jur. Nova Tesis. Recuperado el 20 de Mayo de 2020

Cultura CRM. (04 de 01 de 2017). *Cultura CRM*. Obtenido de https://culturacrm.com/data-mining/data-mining-casos-exito

Cultura CRM. (09 de 02 de 2017). *Cultura CRM*. Obtenido de https://culturacrm.com/data-mining/ecommerce-cesta-compra-data-mining

Domingues, M. (2004). *Generalização de regras de associação.* Instituto de Ciências Matemáticas e de Computação,. São Carlos, Brasil: Instituto de Matemáticas e Informática.

Fayyad, U., Piatestky-Shapiro, G., & Smyth, P. (1996). The kdd Process for Extracting Useful Knowledge from Volumes of Data. *COMMUNICATIONS OF THE ACM*.

Guillermo Varela. (11 de 08 de 2020). eCommerce Day. *eCommerce Day Uruguay ONLINE [LIVE] EXPERIENCE*. Uruguay. Recuperado el 13 de 09 de 2020, de https://www.youtube.com/watch?v=DcmZ0yjzWcI

Instituto Latinoamericano de Comercio Electronico. (30 de 09 de 2020). *Instituto Latinoamericano de Comercio Electronico*. Obtenido de eCommerce Institute: https://www.einstituto.org/site/iniciativas/ecommerce-day/

Jiawei Han, Micheline Kamber, & Jian Pei. (2012). *Data Mining Concepts and Techniques.* 225 Wyman Street, Waltham, MA 02451, USA: MK.

Mantenimiento Informático . (2020). *Mantenimiento informático*. Obtenido de https://mantenimientoinformaticoeconomico.com/2020/07/14/sistema-operativo-linux-ventajas-y-desventajas/

Marqués, M. P. (2014). *Minería de datos a través de ejemplos* . Madrid: RC LIBROS.

Mayer, M., & Szenkman, P. (25 de 08 de 2018). *Economía digital: una oportunidad para las pymes argentinas*. Obtenido de INFOBAE: https://www.infobae.com/opinion/2018/08/25/economia-digital-una-oportunidad-para-las-pymes-argentinas/

Microsoft . (2020). *Windows*. Obtenido de https://www.microsoft.com/es-ar/windows

Neves, J. (2003). *Ambiente de pós-processamento para regras de associação.* Universidad de Porto, Facultad de Economía. Porto, Brasil: Illinois.

Olivier Peralta, E. (2019). *Genwords*. Recuperado el noviembre de 2020, de https://www.genwords.com/blog/tipos-de-ecommerce

Orallo, H. (2004). *Introduccion a la mineria de datos.* Alhambra.

Pinho, J. L. (2010). *Métodos de clasificación basados en asociación aplicados a sistemas de recomendación.* Universidad de Salamanca.

PowerData. (30 de 01 de 2017). *PowerData*. Obtenido de https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/preprocesar-y-normalizar-datos-4-pasos-para-limpiar-y-mejorar-datos

Pueyrredon, M. (2020). *CACE*. Obtenido de https://www.cace.org.ar/noticias-record-de-audiencia-mas-de-13600-personas-de-todo-el-pais-se-conectaron-el-ecommerce-day-argentina-online-live-experience

Python. (2020). *Python*. Obtenido de https://www.python.org/

Quintana, P. (03 de 2021). *Maketing Ecommerce* . Obtenido de https://marketing4ecommerce.net/que-es-la-venta-cruzada-o-cross-selling-y-su-primo-el-up-selling/

RapidMiner. (2020). *Rapid Miner*. Obtenido de https://rapidminer.com/

RapidMiner Documentation. (2020). *Associate Rules to Examplesset*. Obtenido de https://docs.rapidminer.com/latest/studio/operators/modeling/associations/association_rules_ to_ exampleset.html

RapidMiner Documentation. (2020). *Write CSV*. Obtenido de https://docs.rapidminer.com/latest/studio/operators/data_access/files/write/write_csv.html

RapidMiner Documentation. (2021). *Create Association Rules*. Obtenido de https://docs.rapidminer.com/latest/studio/operators/modeling/associations/create_association_rules.html

RapidMiner Documentation. (2021). *FP-Growth*. Obtenido de https://docs.rapidminer.com/8.0/studio/operators/modeling/associations/fp_growth.html

R-project. (2020). *The Comprehensive R Archive Network*. Obtenido de https://cran.r-project.org/

Sambucetti, G. (2019). Obtenido de https://www.cace.org.ar/noticias-record-de-audiencia-mas-de-13600-personas-de-todo-el-pais-se-conectaron-el-ecommerce-day-argentina-online-live-experience

Sheare, C. (2003). *The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Wraehousing.* Elsevier Science.

™, S. E. (30 de 08 de 2017). *Centro de ayuda de SAS®*. Obtenido de https://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm

Vallejos, S. J. (2006). *Mineria de datos.* Corrientes, Argentina.

Anexo

A. Google Forms

Requisitos para agregar, editar, eliminar y cargar lista de productos vendidos es tener una cuenta de Google y contar con internet.

a. Agregar productos

1. Ingresar al link: https://docs.google.com/forms/d/17CPwyf2qfTNPAqW6-0v8BmpWy7yQYQBrzXQpQMEPEi8/edit
2. Agregar cada producto haciendo click en el icono de agregar pregunta (Imagen A-1)

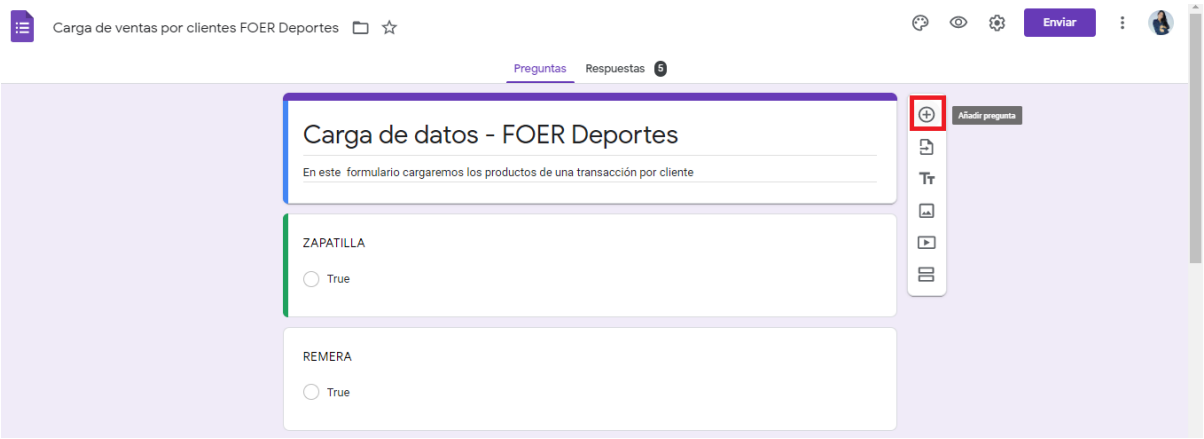


Imagen A-1: Agregar pregunta

3. Agregamos el producto en el casillero pregunta y como respuesta añadimos TRUE, recordar que true representa el producto añadido a la compra (Imagen A-2).

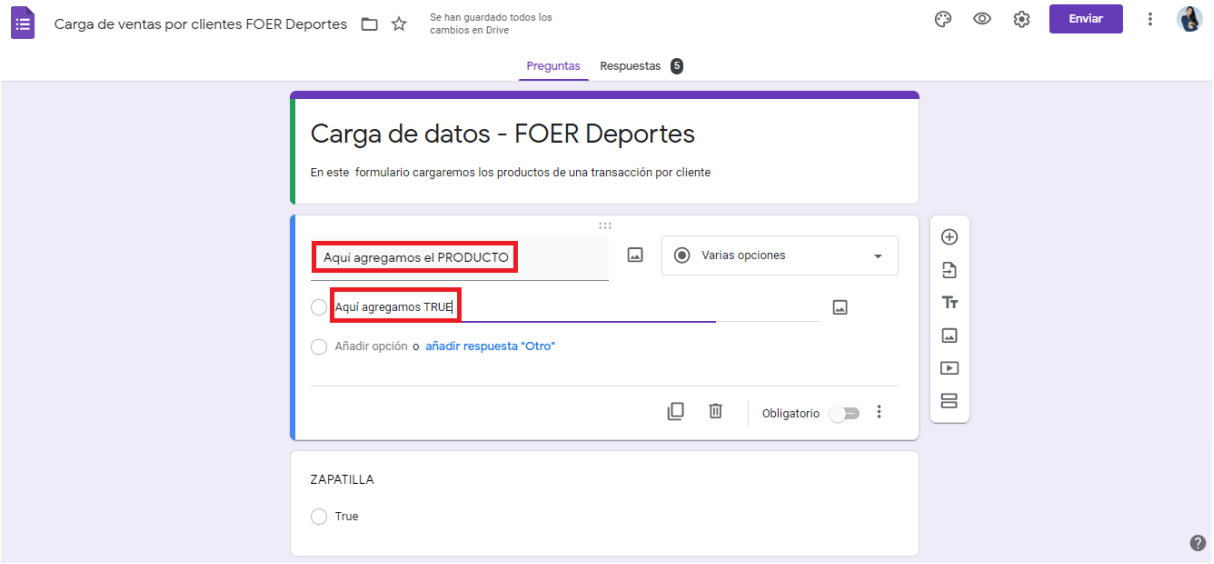


Imagen A-2: Agregar producto y tipo de respuesta

4. Una vez agregado todos los productos necesarios, el formulario queda guardado automáticamente en la cuenta de Google.

b. Editar producto

1. Ingresar al link, con una cuenta de Gmail: https://docs.google.com/forms/d/17CPwyf2qfTNPAqW6-0v8BmpWy7yQYQBrzXQpQMEPEi8/edit
2. Hacer clic en el apartado del producto para editar su nombre (Imagen A-3).

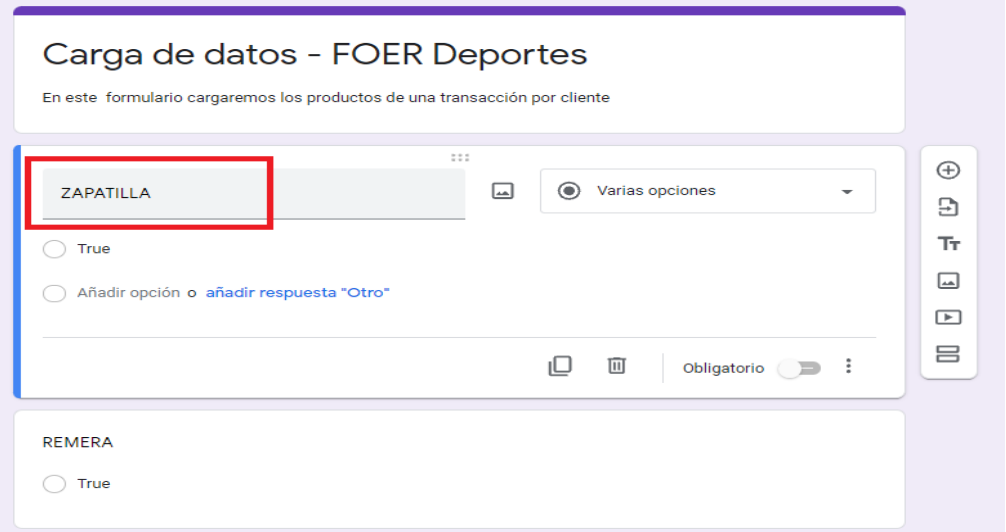


Imagen A-3: Editar producto

c. Eliminar producto

1. Ingresar al link, con una cuenta de Gmail: https://docs.google.com/forms/d/17CPwyf2qfTNPAqW6-0v8BmpWy7yQYQBrzXQpQMEPEi8/edit

2. Hacer click en el símbolo del cesto de basura para eliminar producto

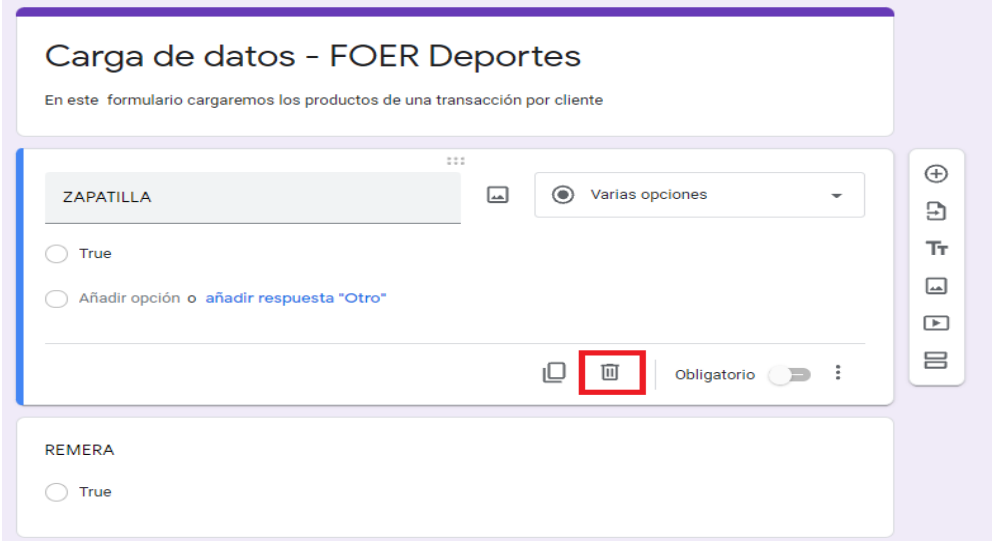


Imagen A-4: Eliminar producto

d. Cargar lista de productos vendidos

1. Ingresar al link: https://docs.google.com/forms/d/e/1FAIpQLSferIWS0ytkAkMxoC3tDb-7tVQ5sGzSW7GycQ3sFXv6XTUHMw/viewform?pli=1

2. Seleccionar los productos que se compraron en la transacción de la venta, por ejemplo ZAPATILLA y MUSCULOSA (Imagen A-5*).*

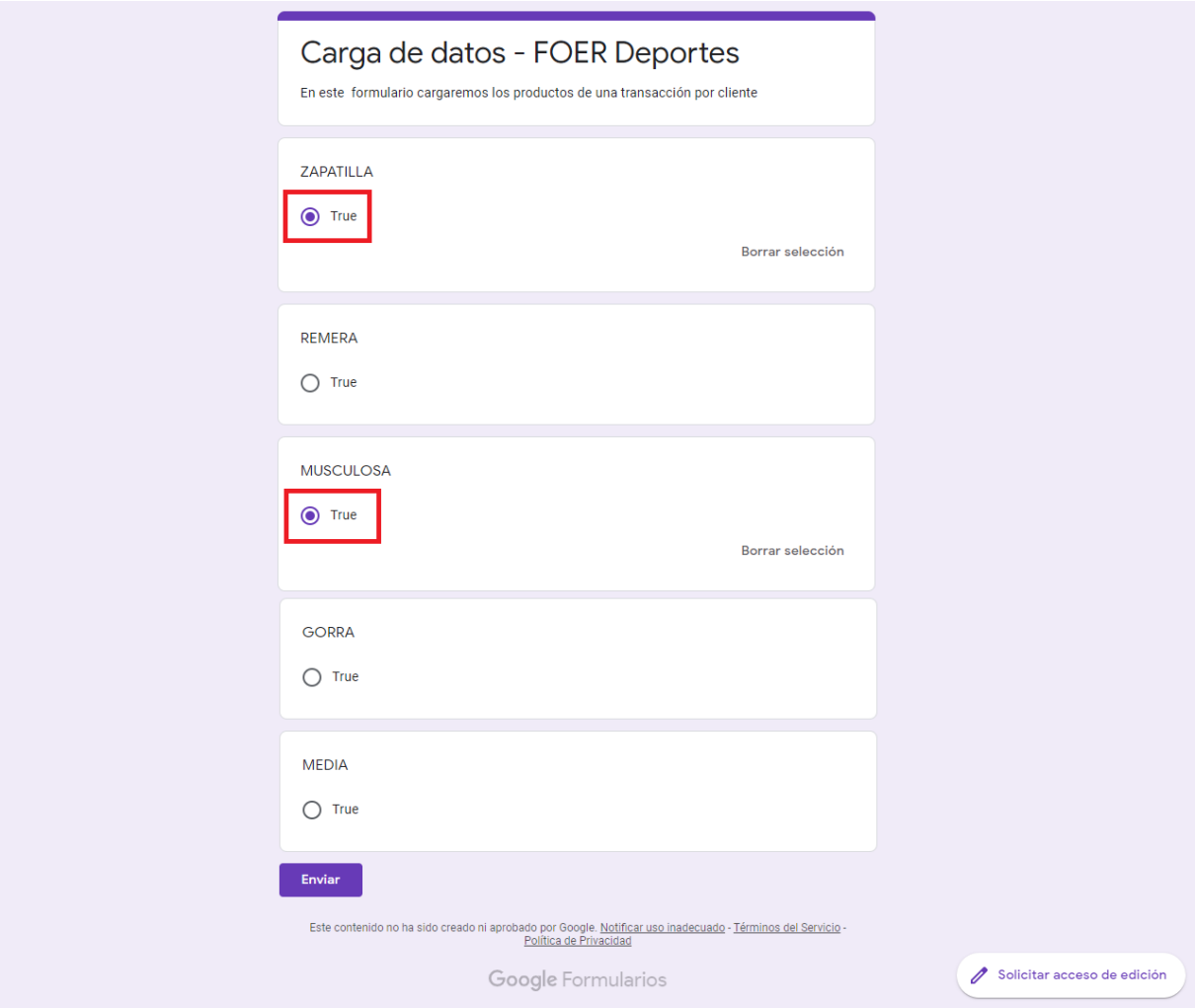


Imagen A-5: Cargar lista de productos vendidos

3. Una vez finalizada la selección de los productos de una transacción por cliente, hacemos click en Enviar (Imagen A-6).

B. Prototipo

El objetivo del desarrollo del prototipo de una tienda online es mostrar la implementación del informe de recomendaciones que está basado en el conocimiento extraído del modelo de asociación, como resultado del proceso de minería de datos.

Cabe aclarar que el archivo plano de reglas de asociación .csv no se integra (no se carga) a la página web, por lo tanto la página web entiende las relaciones porque se configuran manualmente las reglas de asociación (expresadas en el informe de recomendaciones). Como se muestra en la imagen B-1.

El prototipo es una tienda online creada en WordPress con el plugin woocommerce de comercio electrónico y de forma local en la PC. XAMPP como sistema de gestión de base de datos MySQL y servidor web Apache. WordPress es un sistema de gestión de contenidos

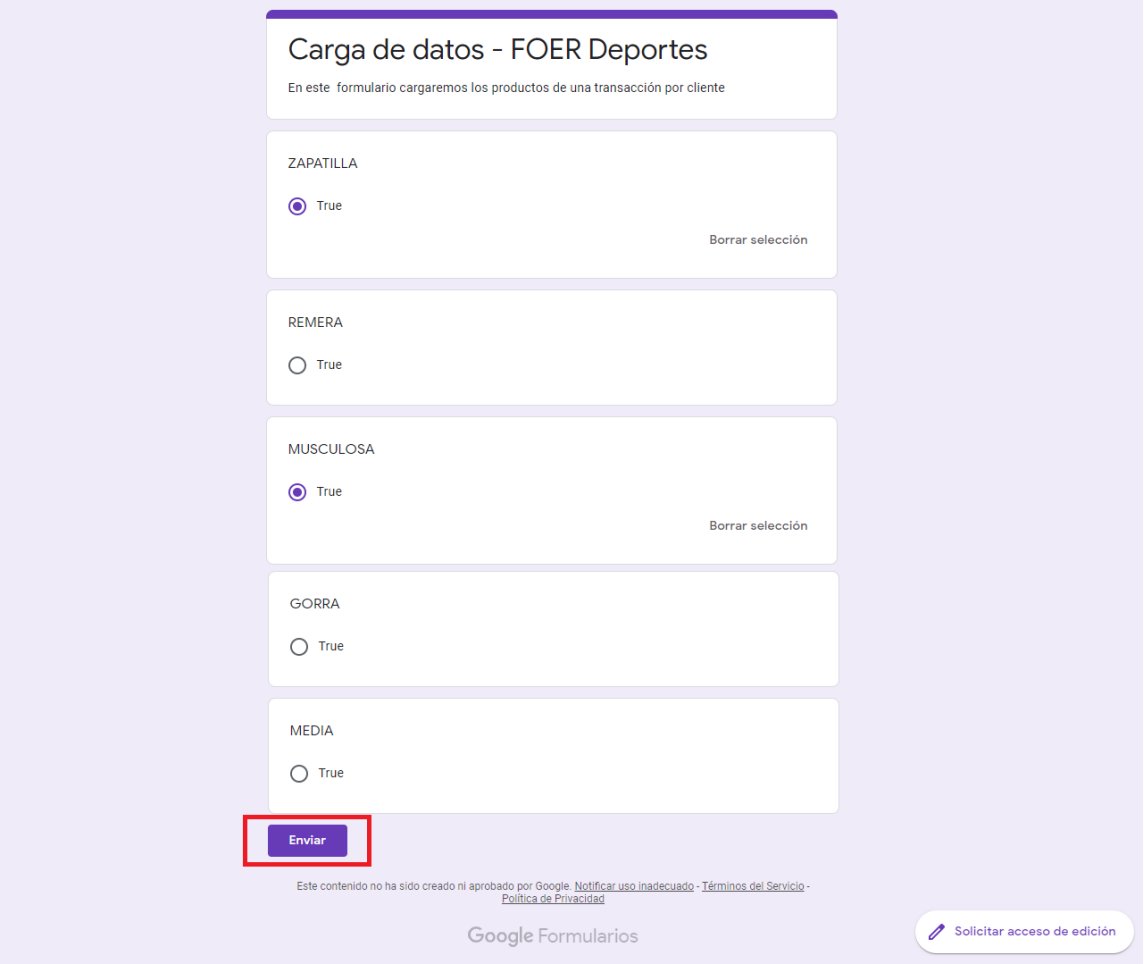


Imagen A-6: Finalizar carga de la lista de productos vendidos

enfocado en la creación de cualquier tipo de página web, está desarrollado en el lenguaje PHP para entornos que ejecuten MySQL y Apache, bajo la licencia GPL[3] y es software libre. XAMPP es un paquete de software libre, que consiste en el sistema de gestión de base de datos MySQL, el servidor Apache y los intérpretes para el lenguaje PHP.

El prototipo estará publicado en formato video en internet: https://drive.google.com/file/d/1x_UL2SHnJ2AY6fHFr1n4KcA0sCh970Xl/view?usp=sharing

La tienda online las siguientes páginas: inicio, tienda, producto, carrito y finalizar compra.

- Inicio: es la portada de la tienda online, se muestras las categorías destacadas o más vendidas.
- Tienda: se muestran todos los productos disponibles y permite filtrar por categorías.
- Producto: permite elegir talles el producto de una categoría.
- Carrito: muestra los productos agregados al carrito y los productos relacionados (venta cruzada, aplica técnica cross selling).

[3] GPL: Licencia Pública General, es una licencia de derecho de autor usando en el mundo del software libre y código abierto

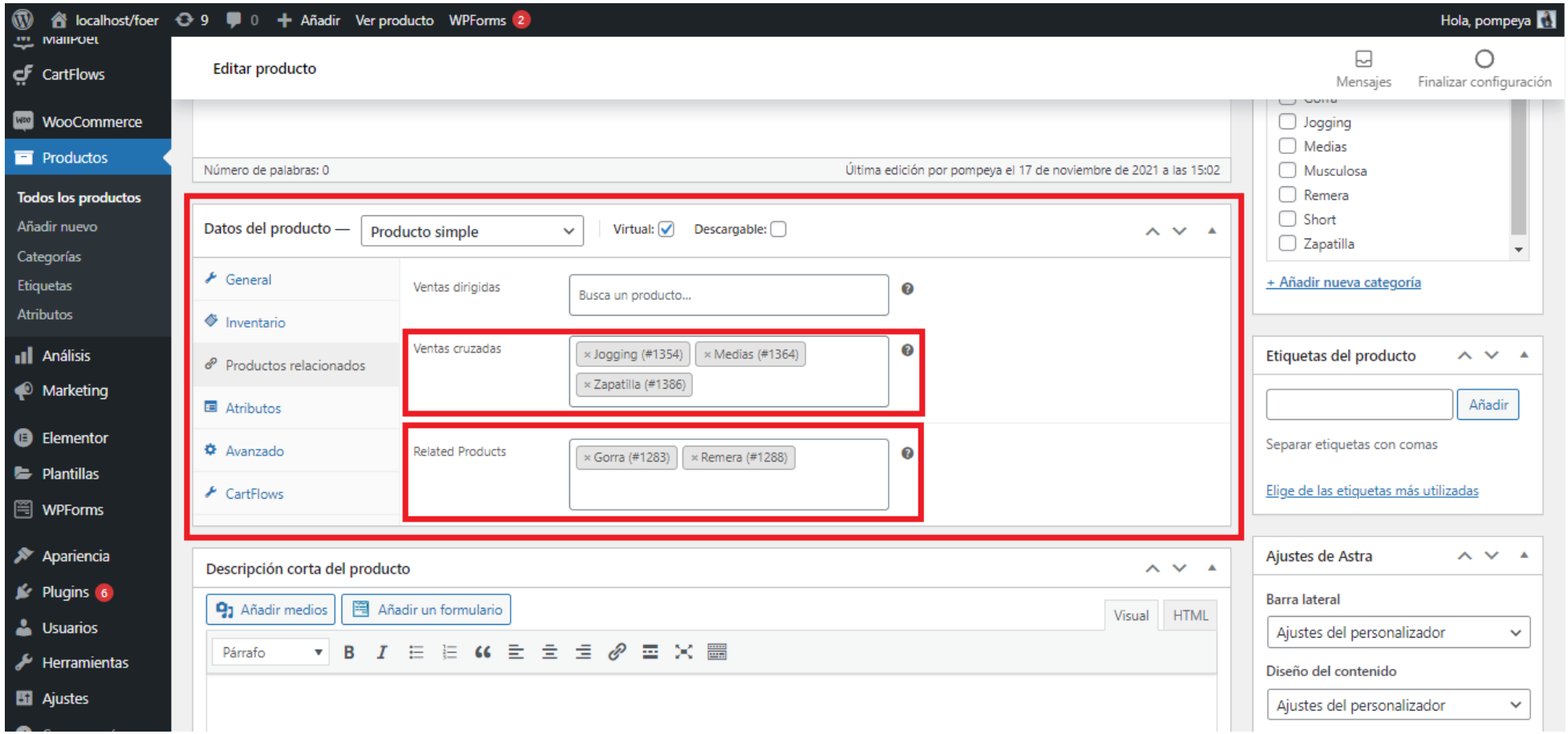
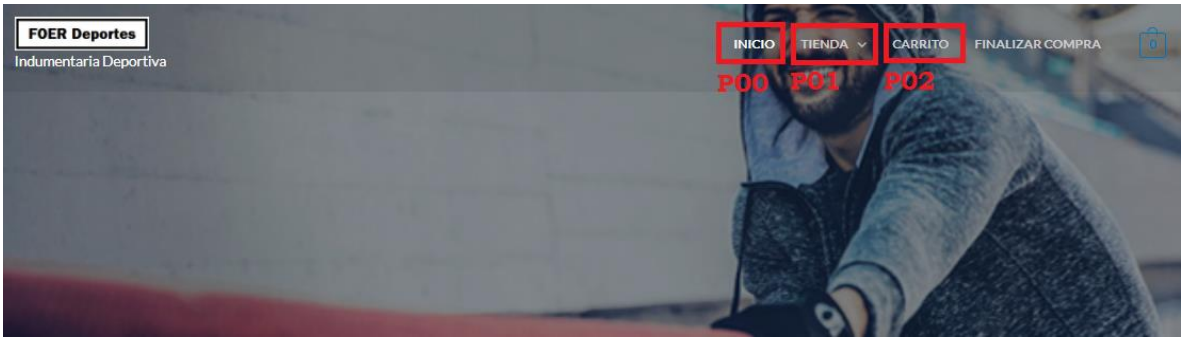


Imagen B-1: Configuración manual de las reglas de asociación en la página web.

Tomando como ejemplo la regla **[Buzo] → [Gorra, Remera],** en la imagen B-1 se muestra la configuración del producto buzo. En la casilla related products se agregan los productos que también tienden a comprase (gorra y remera) y en la casilla ventas cruzadas se agregan los producto que el negocio quiere remover stock.

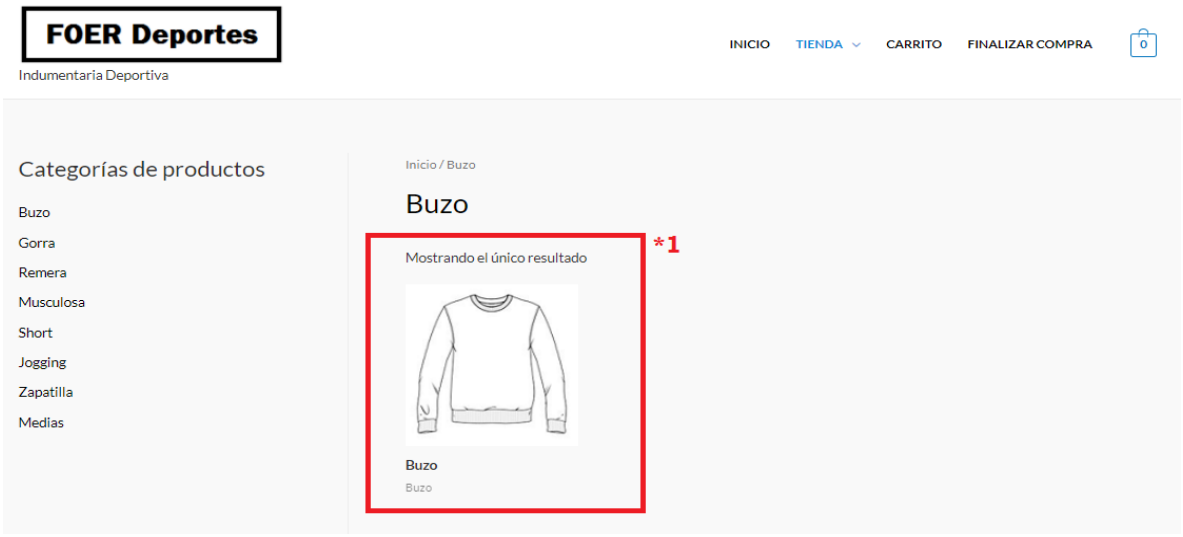
A continuación se muestra un diseño de interfaz estático para la implementación del modelo a través de la navegación en la tienda online, tomando como base general la regla: **[Buzo] → [Gorra, Remera]**

Tienda online Pantalla: Pantalla de inicio	
	

Tienda online Proceso: Usuario visualiza los productos destacados Pantalla: Pantalla de inicio	P00

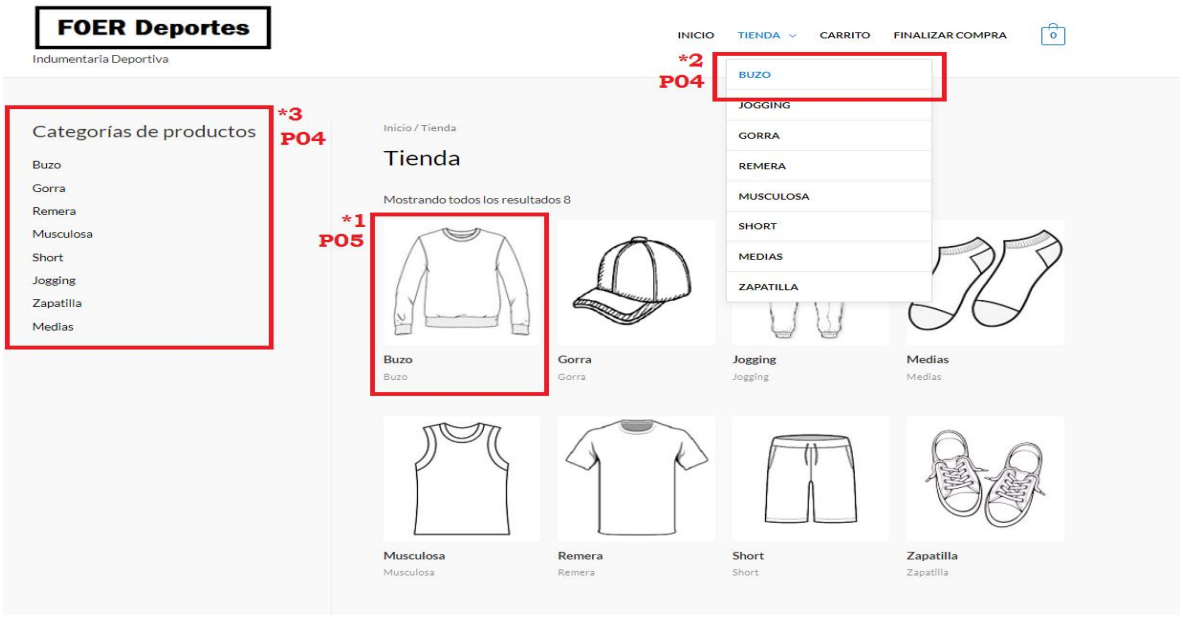
*1 Redirecciona y muestra en la pantalla P06 los productos disponible de la categoria buzo.
*2 Redirecciona y muestra en la pantalla P06 los productos disponible de la categoria remera.
*3 Redirecciona y muestra en la pantalla P06 los productos disponible de la categoria joggins.
*4 Redirecciona y muestra en la pantalla P06 los productos disponible de la categoria gorra.
*5 Se permite visualizar y seleccionar los productos destacados, redireccionando a la pantalla P05 correspondiente al producto seleccionado.
Estos productos destacados son los que corresponden las reglas de asociación.

Tienda online Proceso: Usuario navega entre los productos disponible por categoría Pantalla: Pantalla productos por categoría	P06



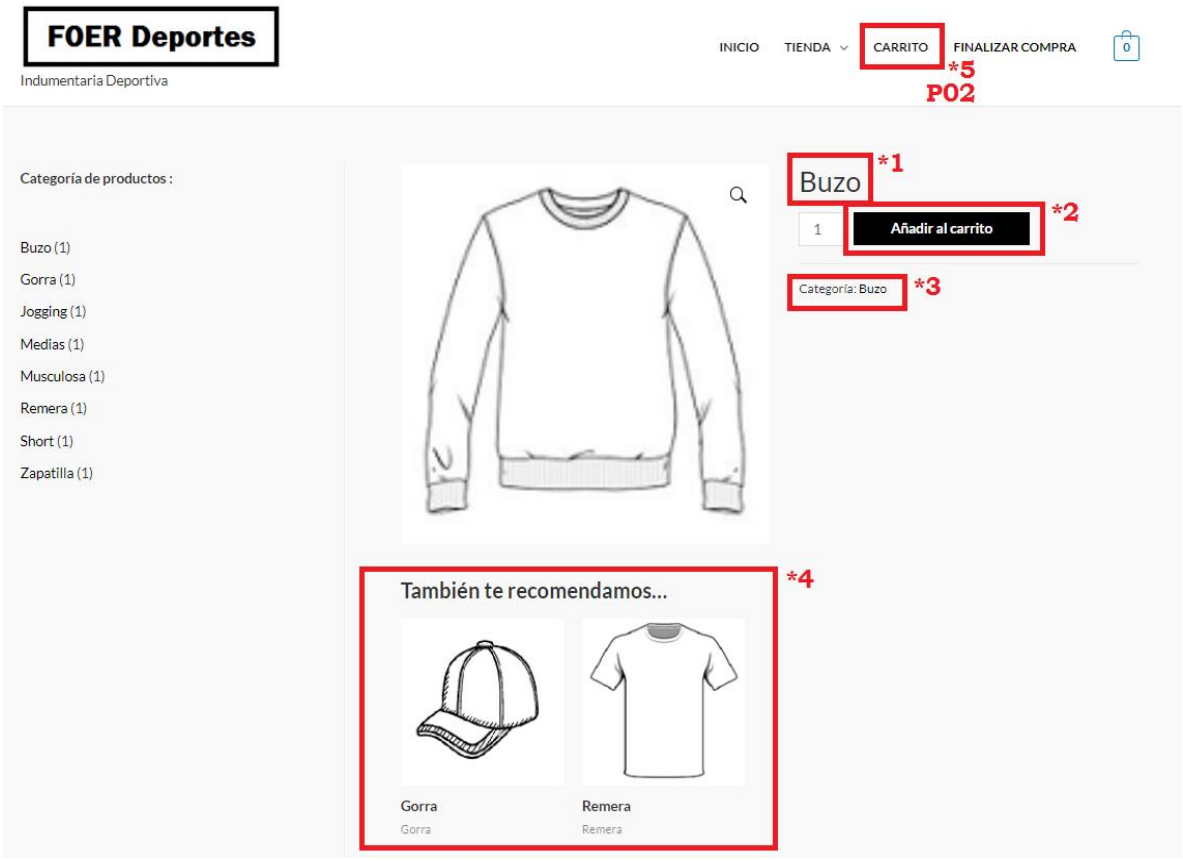
*1 Se permite navegar y seleccionar entre los productos disponibles de la categoría especificada, redireccionando a la pantalla P05 al seleccionar el producto.

Tienda online Proceso: Usuario navega entre los productos disponibles Pantalla: Pantalla general de la tienda	P01



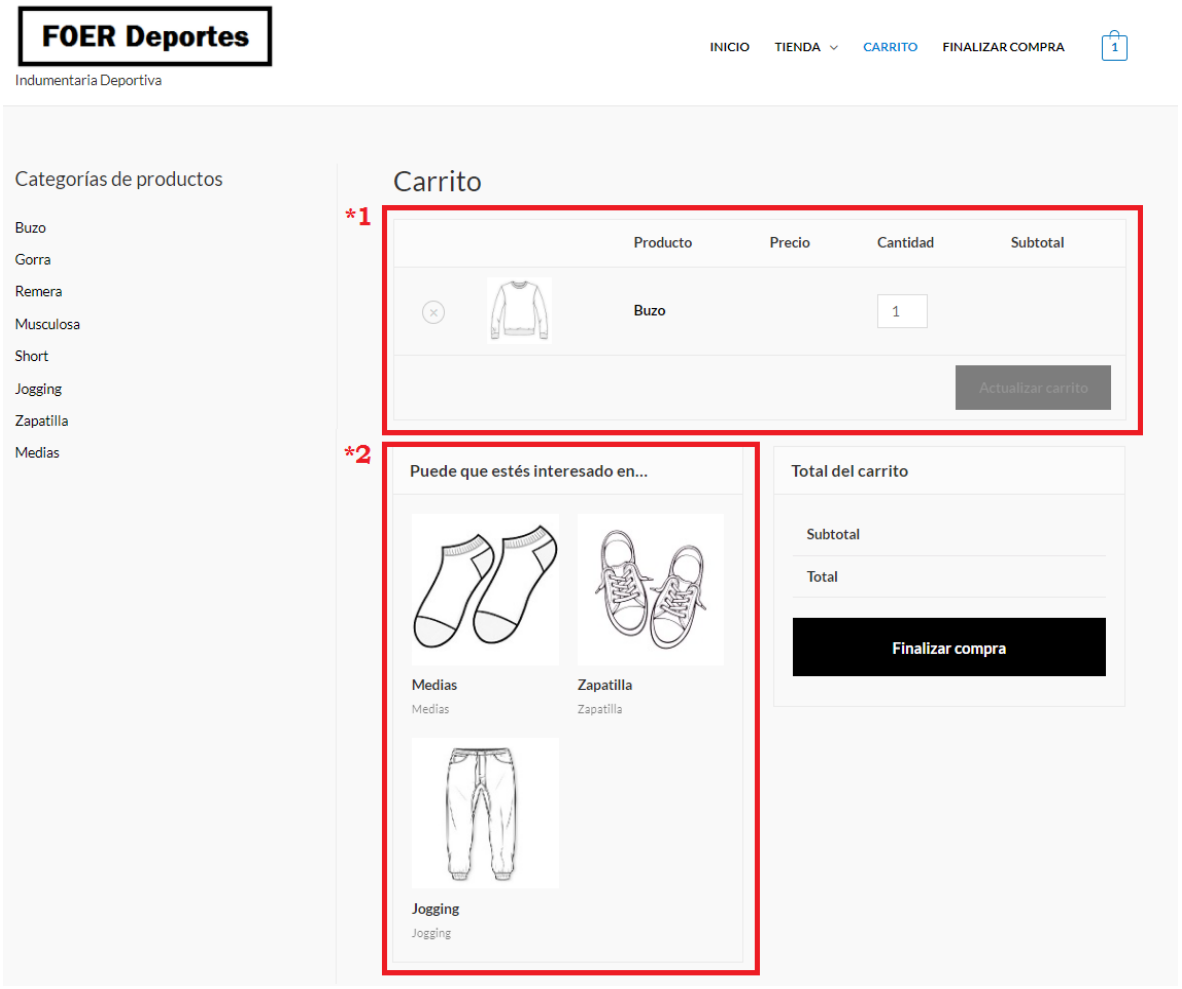
*1 Se permite navegar y seleccionar entre los productos disponibles de en stock, redireccionando a la pantalla P05 al seleccionar el producto.
*2 Se permite seleccionar una categoría de prodcuto del menú desplegable “Tienda”, redireccionando a la pantala P04. Las categorias estan ordenadas de acuerdo a su demanda (mayor demanda arriba).
*3 Se permite seleccionar una categoría de producto del listado “Categoría de productos”, redireccionando a la pantala P04. Las categorias estan ordenadas de acuerdo a su demanda (mayor demanda arriba).

Tienda online Proceso: Usuario selecciona el producto Pantalla: Pantalla de un producto	P05



*1 Indica el nombre del producto
*2 Añade el producto al carrito de compras.
*3 Muestra la categoría del producto.
*4 Muestra y permite seleccionar los productos recomendados según el modelo de asociación obtenido. Para el ejemplo, si el usuario compra **[Buzo]** probablemente compre **[Gorra, Remera]** por lo que se le recomienda gorras y remeras.
*5 Permite visualizar el carrito de compras redireccionando a la pantala "carrito" P02.

Tienda online Proceso: Usuario visualiza el carrito de compras Pantalla: Pantalla del carrito de compras	P02



*1 Permite mostrar el prodcuto que se agregó al carrito de compras.
*2 Muestra y permite seleccionar productos que combinan o se relacionan con el producto agregado al carrito (técnica cross selling).
Para el ejemplo, si el usuario compra **buzo** puede que esté interesado en combinarlo con **jogging** y/o **zapatilla,** que se encuentran en oferta para estimular al cliente en la compra y remover el stock de los productos que menos se venden.