SPORT MARKETING INTELLIGENCE

MÓDULO 2. DEL DATO A LA INFORMACIÓN PARA LA TOMA DE DECISIONES

- CONMEBOL -
EVOLUCIÓN

Introducción

La minería de datos es una de las principales técnicas que podemos utilizar para generar información accionable que nos permita tomar decisiones.

El dato en sí mismo, y por sí solo, no nos sirve para tomar decisiones, pues no nos brinda información, e incluso, en algunos casos, la extracción de datos anecdóticos nos puede llevar a equivocaciones.

Según lo establece Conexión ESAN (2015):

> El *data mining*, o la minería de datos, es todo un conjunto de técnicas que se utiliza por parte de una organización o un equipo investigador para explorar y analizar grandes bases de datos, con el fin de establecer patrones o tendencias en la información. Mediante este proceso, se puede lograr un mejor entendimiento de los datos y así tomar mejores decisiones. (párr. 2).

El *data mining* es un proceso que se realiza en etapas diferentes, que nos permite, en última instancia, lograr conocimiento que podamos aplicar para mejorar el rendimiento de nuestras campañas de *marketing*.

Figura 1: El proceso del *data mining*

Fuente: elaboración propia.

- **Análisis del negocio y las necesidades**: el proceso debe comenzar con un entendimiento del negocio, un análisis de las necesidades que tiene la organización para resolver, y qué

es lo que se busca encontrar como respuesta. Según ese conocimiento del negocio, y del modelo del negocio principalmente, se puede definir qué tipos de datos debemos recabar. Si los datos que necesitamos pertenecen a una investigación sobre la cantidad de socios que visitan nuestro museo oficial del club, la selección será específica al comportamiento de nuestros aficionados en torno a esa pata de nuestro negocio (diferente a la que buscaríamos si, por el contrario, estuviésemos estudiando la necesidad de lanzar una plataforma OTT).

- Análisis de datos disponibles y usos: ya sabiendo qué datos nos interesa conseguir, debemos encontrarlos. Ver qué posibilidad real tenemos de captarlos, organizar el proceso para conseguirlos. Analizar las diferentes fuentes en las que conseguiremos los datos, y los formatos en que los tomaremos. ¿Necesitamos datos digitales? ¿Preferimos conseguirlos con encuestas físicas los días de partido? ¿Tal vez una combinación entre ambas opciones? Dependiendo del objeto de estudio, este paso es fundamental para mantenernos centrados en lo que necesitamos.

- Procesamiento de datos: debemos integrar las bases de datos de manera de poder procesarlos. Cotejar las bases de datos, purgar los datos erróneos, desechar duplicados, etcétera. Sobre esto, hemos profundizado ya en lecturas previas de este curso.

- Modelado de datos: aplicamos diferentes técnicas estadísticas, de analítica y modelos matemáticos de análisis de patrones dentro de la información. De este modo, obtenemos relaciones entre los datos.

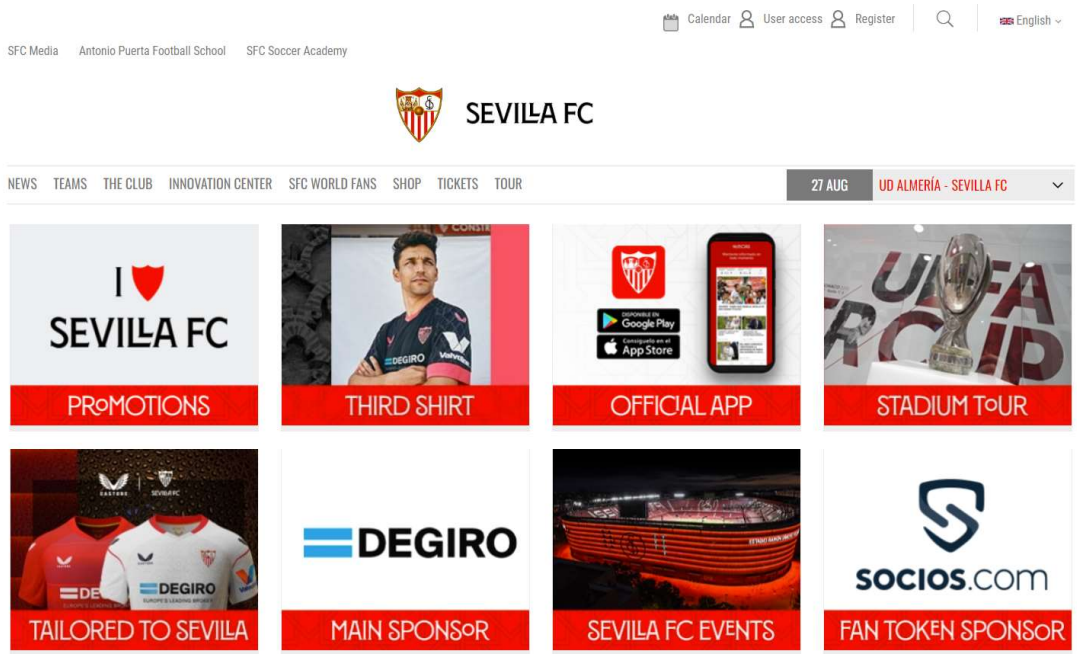
- Generación de conclusiones, *insights* y conocimiento: una vez que hemos detectado los patrones y las relaciones en los datos, y que hemos verificado esa información, podemos arribar a conclusiones, obtener *insights* de la información y conocimiento que nos permita tomar decisiones informadas para nuestro negocio. Qué cosas debemos mejorar en nuestros procesos, cuáles son los puntos a fortalecer y más: toda esa toma de decisiones debe estar alineada con datos concretos (o, como suele decirse popularmente, hechos, no opiniones).

- Implementación en el negocio: cuando las decisiones están tomadas, solamente resta ponerlas en marcha, y, por supuesto, medir el resultado de tal puesta en marcha.

En España, uno de los clubes modelo a la hora de transformar sus decisiones con base en la colección de datos es el Sevilla Fútbol Club. La institución, cuyo equipo es habitual en la Liga de Primera División y competiciones europeas, entendió en 2017 que debía concentrar su modelo de negocio en las necesidades de los aficionados, e inició entonces un proceso de

reestructuración interna que le permitiese trabajar pura y exclusivamente con base en datos duros.

Figura 2: Captura de pantalla del sitio web del Sevilla FC



Fuente: captura de pantalla de sitio web oficial Sevilla FC (https://sevillafc.es/en).

Para lograrlo, el club estableció una estrategia de contacto y captación **omnicanal**, en la que todos los puntos de contacto entre los aficionados y la institución se dan de manera directa y sencilla, independientemente de la plataforma (digital o física) en que se concrete.

Esta modalidad le ha permitido al club recolectar infinidad de datos con el objetivo de comprender mejor el comportamiento y la potencialidad de negocios con su afición a largo plazo, algo que vemos bien planteado en la página web (ver gráfico anterior): en este caso, la *landing page* cuenta con ocho servicios específicos orientados a generar ingresos.

Estos bloques, además, varían dependiendo de dos factores: ubicación geográfica del usuario (no mostrará lo mismo a alguien que viva en Sevilla o fuera de Europa) y rendimiento. Aquellos que no generen lo planificado son evaluados, modificados o retirados.

Unidad 2.1 Knowledge discovery en bases de datos (KDD)

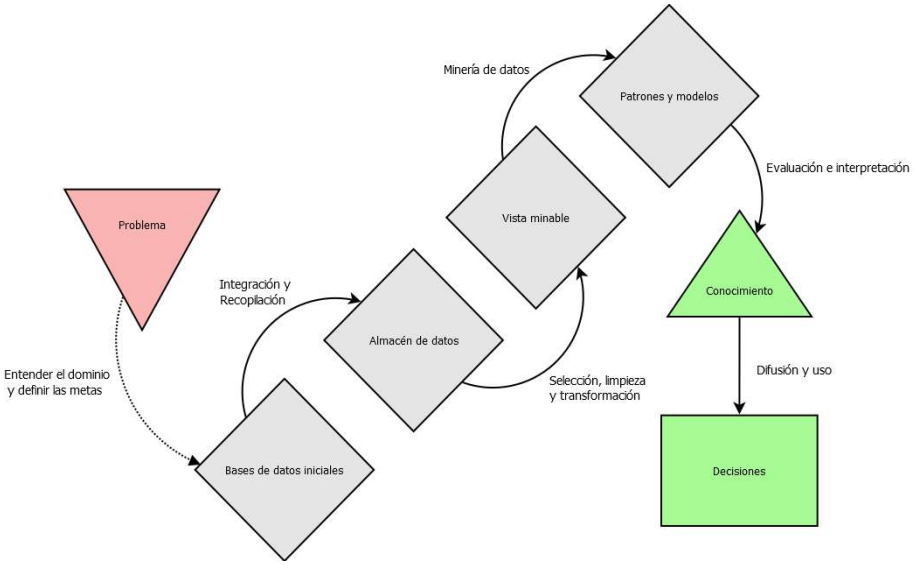
Como su nombre lo indica, el KDD o *knowledge discovery in databases* (descubrimiento de conocimiento en bases de datos) implica el proceso de análisis de bloques de datos e información, para extraer conocimiento aplicable a nuestro modelo de negocio en la organización.

Podemos desglosar la definición un poco más.

- Descubrimiento: nos permite determinar dos cuestiones centrales.
  - Es un proceso, por ende, una serie de pasos involucrados en el desarrollo de una determinada actividad. En este caso, el proceso implica una serie de pasos a realizar para extraer conclusiones sobre las relaciones y los patrones que nos permitan tomar decisiones estratégicas de negocio.
  - Implica un descubrimiento, y no una invención. Esto significa que los patrones y las relaciones existen en sí mismas, aun cuando nosotros las desconozcamos. No vamos a crear las relaciones, sino entender que suceden y por qué.
- De conocimiento: es el objetivo del proceso que realizamos. Nos interesa encontrar relaciones causales, correlaciones, elementos verificables que nos permitan que las decisiones que tomemos tengan mayores chances de éxito.
- En bases de datos: nos determina el universo de análisis en el que trabajaremos con esta técnica.

Para simplificar la tarea, pensemos en un cronograma ordenado con pasos específicos y necesarios para alcanzar un resultado exitoso. Empezaremos con el problema (existente, concreto), para luego pasar las diferentes instancias de datos: la base inicial, su integración y recopilación; la selección de los preponderantes con su correspondiente limpieza y transformación, su posterior minería, evaluación e interpretación, para arribar al conocimiento y una toma de decisiones más saludable.

Figura 3: Proceso de KDD



Fuente: [imagen sin título sobre proceso de KKD], (s. f.), https://bit.ly/3Nmpm8N.

2. 1. 1. *Knowledge discovery and data mining*

En una amplia variedad de disciplinas, y utilizando numerosas herramientas, muchas de ellas creadas a partir de las nuevas tecnologías, los datos se acumulan a un ritmo dramático. Y este fenómeno, que sin dudas nos brinda una gran oportunidad, puede generarnos también varios dolores de cabeza, a menos que utilicemos herramientas y técnicas que nos ayuden a hacer sentido y construir conocimiento (*knowledge*) a partir de esos datos recolectados. El conjunto de esas herramientas y técnicas es lo que se conoce como *knowledge discovery* en base de datos.

Existe una gran cantidad de actividades involucradas en el proceso de *knowledge discovery*. Entre ellas, las actividades de *data mining* (o minería de datos) son críticas para el esfuerzo de construcción de conocimiento a partir de datos aislados disponibles en base. *Data mining*, como hemos mencionado, es el proceso de analizar datos provenientes de diferentes perspectivas y resumirlos en información útil para la toma de decisiones. Técnicamente, *data mining* es el proceso de encontrar patrones o correlaciones entre diferentes campos en bases de datos relacionales de tamaño considerable.

Pero ¿en qué se diferencian estos dos conceptos? Básicamente, en que el *knowledge discovery* es el proceso de identificar patrones válidos, novedosos, y potencialmente útiles en los datos; mientras que *data mining* es uno de sus pasos. El concepto de *data mining* hace

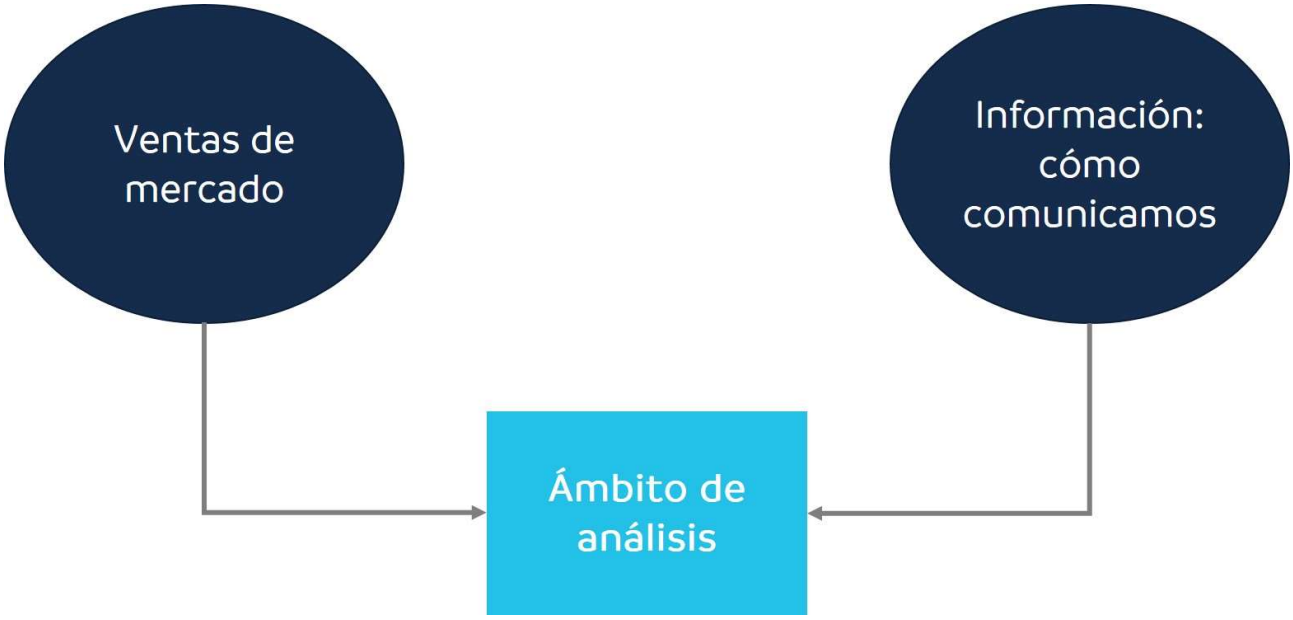
referencia, de forma específica, a las actividades de análisis involucradas en la identificación de esos patrones de comportamiento en los datos. Existen otros pasos, además de las tareas de *data mining*, involucrados en el proceso de *knowledge discovery*. Estos son selección y preparación de los datos, limpieza de datos, incorporación de información previa, interpretación de los datos, etcétera, como pudimos ver en la figura 3.

> La mayor parte del proceso de KDD, así como de cada una de las fases, es iterativo e interactivo. Se entiende por iterativo que la estructura temporal no sigue una progresión lineal, sino que el hecho de terminar una fase puede requerir avanzar a una fase posterior o regresar para repetir una fase anterior con mayor precisión. Por interactivo se entiende la necesidad del usuario, que además debe estar familiarizado con el proceso, debe apoyar cada una de las fases de forma activa. (Diagramas UML, 2018, párr. 5).

- Problema: el problema es la razón de ser del proceso, lo que nos determina la necesidad de realizar un análisis. Por ejemplo, entre tantos otros, podríamos citar un hipotético caso en el que un club, con numerosa afición los días de partido, no consigue transformar ese fenómeno en mayor número de ventas en la tienda oficial.
- Entender el dominio y definir las metas: este problema debemos definirlo en concreto, y, para ello, plantear los objetivos que perseguiremos. En esa misma línea hipotética planteada en el punto anterior, el objetivo de esta investigación pasaría por confirmar los motivos de desconexión entre aficionados y la tienda en primera instancia; e invertir esa relación a partir de la implementación de nuevas decisiones y estrategias.
- Bases de datos iniciales: comenzaremos nuestro proceso de uso de datos con la comprensión sobre la información con la que ya contamos. Debemos listar tanto las fuentes internas como externas que tenemos, y cuáles requerimos conseguir. Cuánta gente asiste al estadio, quiénes son, si son asistentes esporádicos o cuentan con un abono para toda la temporada, etc.

- **Integración y recopilación**: en esta fase, debemos vincular los datos con los que trabajaremos. Vinculamos tanto las fuentes internas como las externas, así como la manera en que se puedan igualar y procesar. En este caso, si quisiéramos entender si existe una relación entre las escasas ventas de un producto y la comunicación existente sobre este servicio entre nuestra afición, seguramente, deberemos integrar y recopilar las bases de datos propias de venta del producto, y las bases de datos externas de información y comunicaciones en las diversas plataformas en las que estamos presentes.

Figura 4: Ejemplo de necesidades de datos



Fuente: elaboración propia.

- **Almacén de datos:** "conjunto de datos históricos, internos o externos, y descriptivos de un contexto o área de estudio, que están integrados y organizados de tal forma que permiten aplicar eficientemente herramientas para resumir, describir y analizar los datos" (Diagramas UML, 2018, párr. 14).

- **Selección, limpieza y transformación**: dado que las fuentes originales que usamos de los datos no siempre están necesariamente desarrolladas o diseñadas para vincularse entre sí, es posible que se generen imperfecciones, desfasajes, incongruencias u otros inconvenientes que es importante subsanar previo al trabajo de minería de datos.

    - **Datos irrelevantes**: existen datos que caen fuera de nuestro universo de análisis y que, de utilizarlos, incorporan complejidad al estudio, o pueden desviar los resultados. En el ejemplo anterior, si estamos analizando el comportamiento de las compras en función

nuestra comunicación con la afición que asiste al estadio, probablemente, debamos definir qué datos de ventas nos resultan útiles (algunos productos en particular, por ejemplo, y no toda la venta del negocio), y, por otro lado, la información sobre nuestra comunicación (quizás la asiduidad con que publicitamos la tienda, o qué tan visibles somos en el estadio los días de partido).

  - Valores atípicos o *outliers*: son datos que, aun siendo válidos, pueden distorsionar la muestra por encontrarse en algún extremo. Estos valores se pueden discretizar para transformarlos en una categoría utilizable. En el caso del ejemplo, si un día hubo apenas un 30 % de aforo debido a que gran parte del estadio se encontraba en obras, se lo puede discretizar en una categoría de «poca asistencia» o «estadio en obras» que integren a los *outliers*.

  - Valores erróneos: son datos que no responden a la generalidad por algún error de captura, tipeado, tipificación, formato, etcétera.

  - Valores faltantes: son datos que no se capturaron o no se guardaron.

- Vista minable: una vez que la información está procesada y los datos curados, estamos en condiciones de someter el cuerpo de análisis a la minería de datos; por esta razón, las bases procesadas, una vez completo el proceso anterior, se denominan «vista minable».

- Minería de datos: como hemos mencionado, busca explorar y analizar datos de gran volumen intentando desvelar patrones y reglas de comportamiento relevantes. Se puede realizar, principalmente, de dos maneras:

  - Aprendizaje supervisado: su objetivo es predecir o clasificar. Se busca encontrar una variable de salida sencilla en función de los datos de entradas. Entre los modelos que se utilizan, encontramos las regresiones lineales, regresiones logísticas, series temporales, árboles de regresión o clasificación, redes neuronales, algoritmo de vecino más próximo.

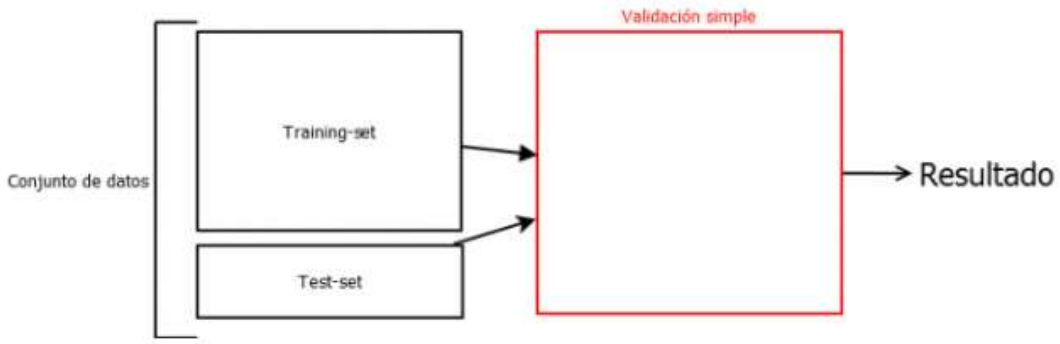
  - Aprendizaje no supervisado: busca entender y describir los datos, con la intención de descubrir patrones de comportamiento subyacentes. Los algoritmos de recomendación son un claro ejemplo de este método. Algunos de los modelos que se utilizan en aprendizaje no supervisado son la agrupación, los análisis de asociación y los análisis de componentes principales.

- Metodologías mixtas: se pueden utilizar técnicas de aprendizaje no supervisado para detectar patrones, y luego aprendizaje supervisado para trabajar sobre un número más pequeño y manejable de variables.

- **Patrones y modelos**: de acuerdo con las necesidades y lo que nos devuelva la realidad del análisis, develaremos patrones y modelos de comportamiento, que nos ayudarán en la descripción de la realidad, en la categorización y en la predicción de eventos futuros.

- **Evaluación e interpretación**: en este paso, debemos ver cómo los patrones detectados y ese conocimiento generado se aproximan a la realidad. Para ello, se debe trabajar con un *set* de datos de entrenamiento, y luego un *set* de datos de testeo. Se separa el **set** de entrenamiento y el *set* de testeo, para poder realizar verificaciones no sesgadas.

Figura 5: *Data sets*



Fuente: [imagen sin título sobre *data sets*], (s. f.), https://bit.ly/3Nmpm8N.

- **Conocimiento**: una vez que hemos validado las evaluaciones, podemos arribar a conclusiones aceptables, que se transforman en conocimiento:

    - En una clasificación, se medirá el número de entradas clasificadas correctamente entre el número de entradas de prueba.

    - En una regresión, se medirá la distancia (generalmente al cuadrado, que tendrá más en cuenta las distancias más grandes) entre el valor que se ha predicho y el valor real.

- En un agrupamiento, se medirá la distancia al punto medio del grupo y la distancia entre grupos.
- En una tarea de reglas de asociación, se evaluará de forma separada cada una de las reglas. (Diagramas UML, 2018, párr. 2).

- **Difusión y uso**: ya tenemos el conocimiento generado. Ahora, debemos hacer que el conocimiento llegue a la gente que requiere utilizarlo. Para ello, debemos transformar el conocimiento en un entregable fácil de comprender y aplicar, para quien lo requiere. Cabe recordar, una vez más, que esta es, llamativamente, una de las partes más complejas e importantes del proceso: los datos en sí mismos, sin una visualización clara, ordenada y transparente para aquellos que no están acostumbrados a trabajar con un proceso de minería de datos, no ayudarán con el objetivo final. De hecho, muy posiblemente compliquen el proceso.
- **Decisiones**: finalmente, debemos poder tomar decisiones, o ayudar a quien deba hacerlo, en función al conocimiento generado. Para ello, es importante mantener el modelo actualizado, y revisar periódicamente los cambios en el comportamiento, especialmente cuando se producen cambios notorios en las variables.

2. 1. 2. Elementos y métodos de *data mining*

Existen muchas opciones de *softwares* específicos que se utilizan para realizar esfuerzos de *data mining*. Este *software* permite a los usuarios analizar un mismo conjunto de datos desde diferentes dimensiones o ángulos, categorizarlos y resumir las relaciones y patrones encontrados con base en búsquedas abiertas.

Hay diferentes tipos de *software* disponibles para este fin, dependiendo de su base de análisis. Esta puede ser análisis estadístico, *machine learning*, *neural networks* (o redes neuronales).

De acuerdo con Frand (2014), las actividades de *data mining*, generalmente, buscan encontrar 4 tipos de relaciones fundamentales.

- **Clases**: la relación que existe cuando los datos acumulados se utilizan para encontrar otros datos en grupos predeterminados. Esta relación se establece cuando, por ejemplo, un club utiliza datos sobre los horarios y días de mayor tráfico en sus plataformas digitales para definir en qué momentos lanza los sorteos y concursos apuntados a premiar la lealtad de sus usuarios (o clientes).

- Clústeres: relación que se verifica cuando los datos se agrupan de acuerdo con relaciones lógicas. Un ejemplo de esta actividad es el análisis de datos de ventas y perfiles demográficos en nuestras plataformas, con el fin de identificar segmentos de mercado o preferencias de los consumidores.
- Asociaciones: los datos también pueden analizarse para identificar asociaciones (por correlación simple). La relación, conocida y verificada, entre la compra de entradas y vestimenta oficial, es un ejemplo de esto.
- Patrones secuenciales: la relación que existe cuando los datos se analizan para encontrar patrones o tendencias. Un ejemplo de esto es la capacidad de un club de predecir la compra de una camiseta en la tienda oficial, basada en la previa adquisición de un *short* y una campera del equipo, etc.

Existen 5 elementos (o pasos) fundamentales en el proceso de *data mining*:

1. extracción, transformación y carga de datos en un sistema de almacenamiento de datos (*data warehouse system*).
2. Acumulación y gestión de los datos en un sistema de bases de datos multidimensionales.
3. Acceso a los datos por parte de analistas de datos y profesionales de IT (*information technology*).
4. Análisis de los datos utilizando *software* específico para tal fin.
5. Presentación de los datos en un formato adecuado.

Tanto para realizar análisis descriptivos (aquellos limitados a las relaciones o patrones de comportamientos existentes y verificados) como predictivos (aquellos que van más allá y definen potenciales relaciones o comportamientos futuros según los datos actuales), los siguientes métodos de *data mining* son los más comúnmente utilizados. No ahondaremos en la teoría, pero siempre es importante comprender de qué se tratan para un mejor entendimiento del proceso:

- clasificación. Involucra catalogar los datos en uno de los tipos (o clases) predefinidos.
- Regresión: consiste en vincular un dato con una variable predicha en virtud de las relaciones funcionales existentes entre dichas variables.

- *Clustering*: consiste en identificar un conjunto de categorías (o clústeres) que comparten determinadas variables. El método de la estimación de densidad probabilística está muy relacionado con este método.
- **Sumarización**: involucra encontrar una descripción compacta de una serie de variables; esto es un resumen de reglas de asociación y el uso de técnicas de visualización multivariable.
- **Modelos de dependencia**: consiste en describir una cantidad significativa de dependencias entre variables.
- **Detección de cambio y desviación**: involucra descubrir los cambios más significativos en los datos.

Además de los métodos generales, las actividades de *data mining* involucran la construcción de algoritmos específicos para aplicar cada uno de esos métodos. En cada uno de esos algoritmos, es posible identificar 3 componentes básicos: representación del modelo, criterios para la evaluación del modelo y métodos de búsqueda.

Figura 6: *Software* de *data mining*

Fuente: Microstrategy, 2020, https://bit.ly/3swX5CL.

2. 1. 3. Usos y aplicaciones de *data mining* para *marketing*

Las actividades de *data mining* en el fútbol se utilizan mayormente para analizar la performance de los y las futbolistas, además de estudiar su carga de trabajo; pero en lo que a *marketing* se refiere, se desarrollan, prioritariamente, en organizaciones con hincapié en sus consumidores (o usuarios) y que, a su vez, poseen los mecanismos y la infraestructura apropiada para hacerse con una considerable cantidad de datos sobre ellos. Estas organizaciones, generalmente, analizan sus datos internos (precios, procesos o posicionamiento de marca, por ejemplo) en búsqueda de relaciones o patrones de comportamiento vinculados con datos externos (características demográficas de sus consumidores, actividades de sus competidores o indicadores económicos). Este análisis les permite determinar el impacto de las variaciones en dichas variables sobre el volumen de ventas, la satisfacción de los clientes, etc.

Figura 7: Pirámide de *data mining* para *marketing*

Fuente: elaboración propia.

A medida que trabajemos en nuestros esfuerzos de *marketing*, con nuestros datos, podremos ir trepando en la pirámide de la aplicación de *data mining*.

- **Mejora en los datos**: cuando contamos con mejoras en los datos, podremos estandarizar algunos procesos, comparar información, evitar duplicaciones e ineficiencias operativas, mejorar los reportes, etcétera.
- **Mejora en el conocimiento del público y los prospectos** : podremos conocer los perfiles de los clientes, detectar los grupos de públicos más rentables, los segmentos que mejor y peor nos rinden (en nuestro plan de negocio), etcétera.
- **Mejoras en la estrategia**: nuestras campañas de *marketing* y nuestro proceso de captación y retención de clientes puede mejorar, con impacto en las tasas de adquisición de clientes, con optimización del *mix* de *marketing* y el armado de nuestras campañas, con la detección de oportunidades de venta cruzada, etcétera.
- **Mejoras en el negocio**: en este nivel, podemos acceder a mejoras en los costos, maximización de las tasas de respuesta, precios promedio de compra, valor de vida del cliente, optimización de *upsell*, mejoras en los procesos de distribución y logística, optimización del pedido perfecto y más.

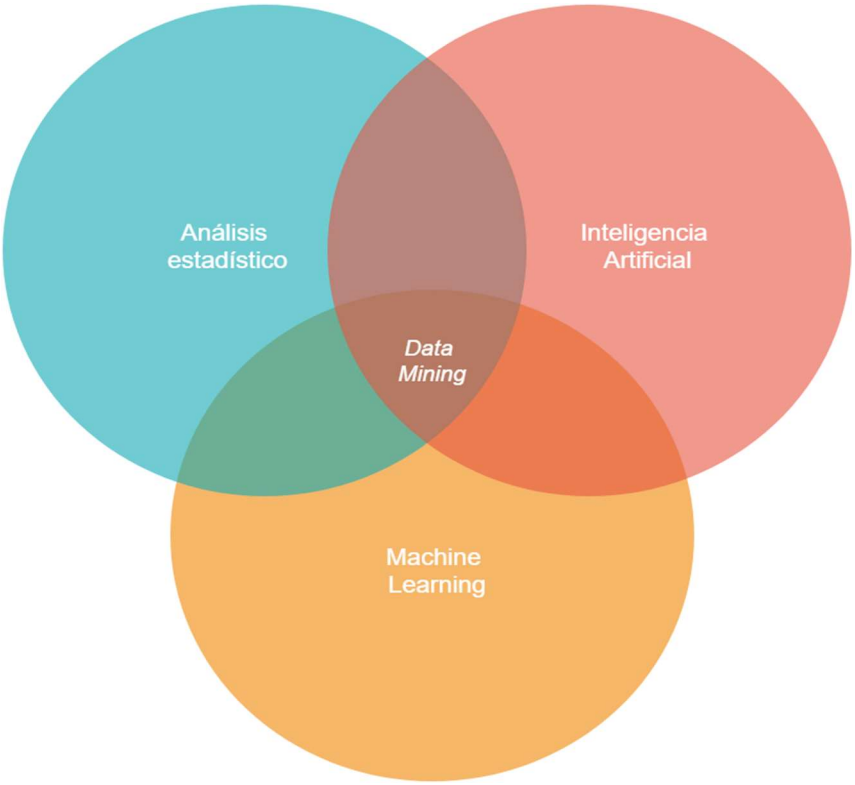
2. 1. 4. Herramientas e infraestructura requeridas

En la actualidad, aplicaciones de *data mining* están disponibles para infraestructuras de todos tipos. Los precios van desde los cientos hasta varios millones de dólares. Las variables críticas son las siguientes:

- **tamaño de las bases de datos**. Mientras más cantidad de datos se almacena, procesa y mantiene, más potentes son los sistemas e infraestructura que se necesitan.
- **Complejidad y frecuencia del análisis**. Mientras más complejos son los análisis y mayor la frecuencia de consulta, más potentes son los sistemas e infraestructura que se necesitan.

La principal de las herramientas a disposición es SAS, que se concentra en el proceso de encontrar anomalías, patrones y correlaciones en grandes bloques de datos, con la intención de predecir sucesos y comportamientos.

Figura 8: Las herramientas e infraestructura en *data mining*



Fuente: elaboración propia.

Data mining combina, fundamentalmente, tres herramientas y técnicas de trabajo:

- la analítica de estadística. Es decir, el estudio numérico de relaciones entre datos.
- La inteligencia artificial, que se refiere al uso de *software* o máquinas para desarrollar procesos de análisis y de generación de conocimiento.
- *Machine learning*: el uso de algoritmos que pueden aprender de los *sets* de datos para realizar predicciones.

Unidad 2.2 Análisis de datos

Existen muchos tipos de datos y múltiples maneras de recolectar datos para responder las preguntas que surgen de nuestros problemas; y las técnicas de recolección de datos que utilicemos pueden producir información numérica en datos cuantitativos, o pueden ser ilustrativos, como los datos cualitativos, o pueden incluir una combinación de ambas opciones y trabajar con datos mixtos. Por supuesto, determinar qué tipo de dato necesitaremos para responder las preguntas planteadas en nuestro problema nos dirá cuál es la técnica que necesitaremos usar.

Los datos los podemos trabajar de diferentes maneras: observar los valores de las variables para el fenómeno que estamos analizando es la manera de captar o recolectar datos. Cada dato individual se denomina «observación», y la colección de las observaciones que se realizan son nuestro *set* de datos (los valores de las variables obtenidas para una muestra de unidades) o matriz de datos (en la que los valores de cada variable particular se organizan dentro de una misma columna, y los valores de las variables forman las columnas de la matriz de datos).

De acuerdo con Paz (2016):

- Datos cuantitativos: requieren uso de análisis estadístico. Las variables pueden ser identificadas y sus relaciones medidas. Se cuentan o expresan de manera numérica. Cuántos aficionados asisten al estadio los días de partido, cuántos conocen la opción de comprar en la tienda oficial, cuántos lo hacen, etc.
- Datos cualitativos: examinan datos no-numéricos en busca de patrones y significados. Son recolectados y analizados con algún mayor grado de subjetividad. La tienda no luce lo suficientemente llamativa, la promoción de sus productos no es suficiente, etc.
- Datos mixtos: pueden explicar algunos resultados inesperados (los denominados *outliers* o excepciones) que utilizando un solo enfoque no se puede. (p. 42).

2. 2. 1. Técnicas de análisis cuantitativo

Una vez recolectados los datos, y antes de someterlos al análisis, es útil llevar a cabo algunas tareas preliminares. Estas, generalmente, incluyen:

- apartar los datos erróneos. Es importante diferenciar aquellos datos incorrectos sin eliminar ningún dato por ser meramente anormal (esto es diferente del conjunto de los demás datos). Nos puede llevar a conclusiones incorrectas.
- Normalizar o reducir los datos. Significa eliminar todos aquellos datos que son irrelevantes o que, si bien se conoce que posee una influencia sobre las variables, no es de interés en el momento del análisis.

En el análisis propiamente dicho de los datos, el objetivo es extraer una estructura que funcione como base para el desarrollo de información y la creación de conocimiento. Generalmente, al comienzo de un proyecto, el investigador posee un modelo matemático que aplicará a los datos. Este modelo se desarrolla a partir de la hipótesis de trabajo, aun cuando esta no sea exacta y deba definirse más claramente durante el análisis. Los datos empíricos se analizan de acuerdo con el modelo, y después se considera en qué grado el marco es adecuado a los datos o si debe buscarse un modelo que se adapte mejor.

El investigador suele decidir qué tipo de patrón de comportamiento está buscando en los datos, dado que esto determinará los métodos para realizar el análisis matemático. Así, una de las primeras cuestiones a resolver es si se quieren analizar las diferentes variables medidas de forma inconexa o las relaciones entre dichas variables.

Otra dimensión importante hace referencia al propósito final del proyecto. Esto puede ser lo siguiente: ¿el objetivo del análisis es meramente describir cómo es el estado actual (no suficientes aficionados compran en la tienda) o, por el contrario, predecir comportamientos futuros en función de la información obtenida acerca de las variables independientes o sus relaciones (mejoraremos la compra si…)? De esto dependerá la utilización de herramientas de análisis estadístico descriptivo o análisis estadístico predictivo.

Revisemos nuestros conocimientos básicos de estadística.

Población y muestra son dos de los conceptos básicos de la estadística que debemos comprender. Tienen que ver con el ámbito que estaremos estudiando o analizando en cada momento en particular. Son dos aspectos íntimamente relacionados, ya que la población podemos afirmar que es todo el *set* de personas o de objetos sobre los cuales queremos obtener conclusiones; sin embargo, muchas veces no podemos acceder a datos de la

totalidad de ellos, y es por ello que debemos recurrir a una muestra, que es una parte de la población. Vemos algunas definiciones clave:

- Población: la población es la colección de todos los individuos o ítems que están bajo consideración en un estudio estadístico. En este caso, los asistentes habituales al estadio.
- Muestra: la muestra es la parte de la población de la que recolectamos información. Como no podemos recolectar datos de todos los asistentes, reducimos el número a una muestra representativa.
- Variable: una variable es un valor que puede cambiar, de acuerdo con las condiciones, o diferentes situaciones. Es un elemento o factor que puede variar, por no ser fijo o consistente. Por ejemplo: edad, género, sector del estadio, poder adquisitivo, etc.

Las poblaciones, a su vez, pueden ser finitas o hipotéticas. Una población finita es aquella que puede ser listada físicamente: una población hipotética, en cambio, es una entidad más abstracta que puede surgir de lo que se está investigando.

La definición de parámetros en la estadística es otro de los elementos que se suele trabajar, ya que los parámetros numéricos, una vez conocidos, nos permiten resumir los hallazgos que hemos logrado. Por ello, muchas veces en las investigaciones estadísticas nos interesa definir estos parámetros que, anticipadamente, no son conocidos. Un parámetro es un resumen numérico desconocido de una población; para hacer inferencias sobre los parámetros, se utilizan estadísticas conocidas.

Superada esa parte teórica, y por demás compleja, analicemos ahora algunas de las principales medidas descriptivas de la estadística.

Existe todo un bloque de medidas llamadas «medidas de centro» que intentan indicar dónde se encuentra el centro o el valor más típico de la variable dentro de un *set* de medidas. Las que se utilizan son la media, la moda y la mediana.

La media: es una de las herramientas cuantitativas más usadas, y coloquialmente se la conoce también como promedio. De acuerdo con Paz (2016): “La media de la variable en una muestra es la suma de todos los valores observados divididos sobre la cantidad de valores observados” (p. 51).

La fórmula entonces de la media es la que se puede observar en la figura 9.

Figura 9: Fórmula de la media

$$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Fuente: elaboración propia.

La moda: la moda implica obtener la frecuencia en la que cada valor observado de una variable se encuentra y detectar la mayor frecuencia. Puede haber más de una moda en una muestra. La moda puede ser un valor numérico o cualitativo (Paz, 2016).

La mediana: es la frontera que divide la muestra en dos. La mediana de una muestra es el valor de la variable que divide el *set* de datos por la mitad, lo que hace que los valores observados en una de las mitades sean todos inferiores o iguales a la mediana, mientras que en la otra mitad serán todos los valores serán iguales o mayores que la mediana. Podríamos decir que es el valor central de un *set* de datos.

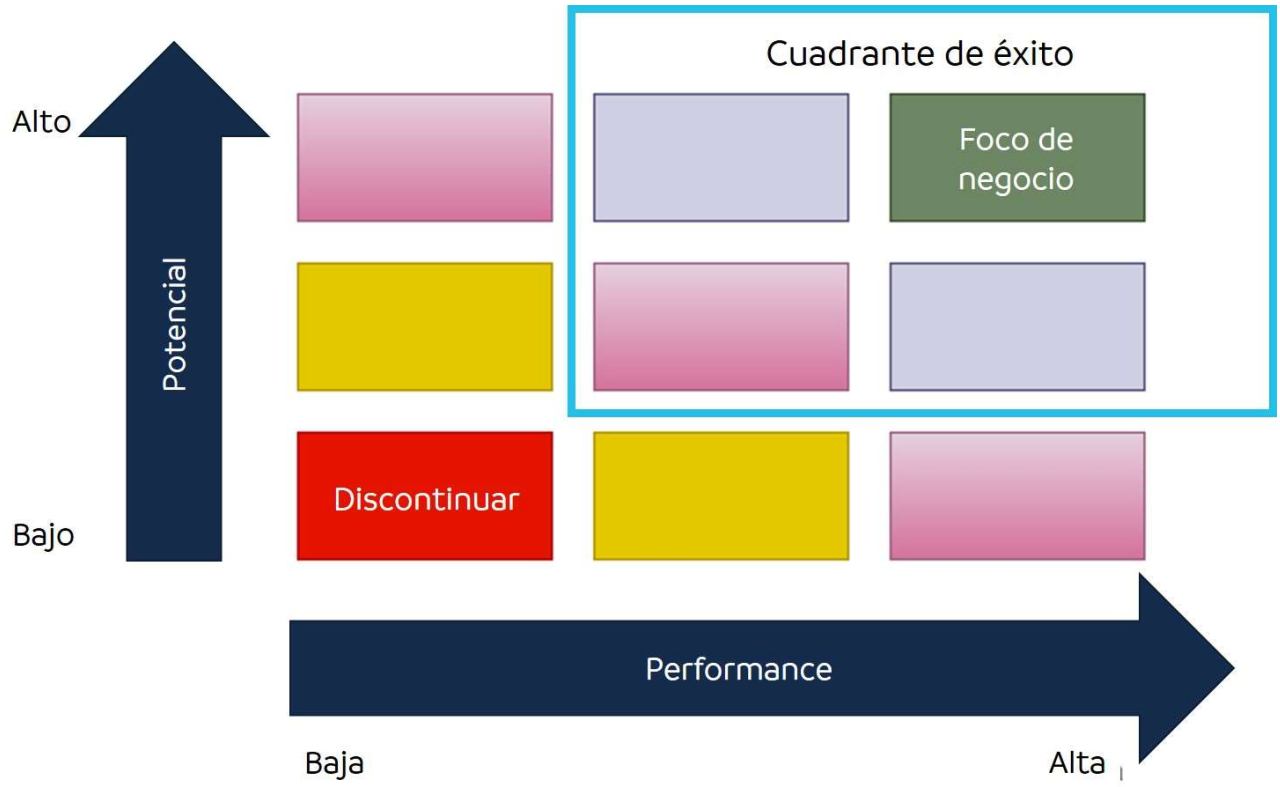
2. 2. 2. Análisis de variables individuales

Una variable es una propiedad que asume diversos valores y que varía en cada una de las observaciones. Es un símbolo al que se le asignan valores o números. Podemos ver una categorización de las variables siguiendo a Barrionuevo (2019), al analizar lo expuesto por Kerlinger y Lee (2002).

- Cualitativas
  - **Dicotómicas (o binarias)**: solo aceptan dos valores. Como ejemplo, podemos referirnos a respuestas «sí-no», o «socio-no socio» si nos mantenemos en el campo teórico que comenzamos al inicio de esta lectura.
  - **Polinómicas**: diversos valores. Podemos referirnos, para seguir en línea con lo mencionado previamente, a la edad del objeto de estudio como un valor de este grupo.
- **Independientes - dependientes**: la variable independiente es la causa supuesta de una variable dependiente.

- **Activa**: es cualquier variable que sea manipulada.
- **Variable atributo**: es aquella que ya viene con el sujeto de estudio. Cuando se trata de individuos o grupos humanos, se los llama atributos activos porque pueden cambiar en el tiempo. Puede referirse a cuestiones genéricas como el color de pelo, estado civil, etc.
- **Continua**: es capaz de asumir un conjunto ordenado de valores dentro de cierto rango (se asocia con escalas y puntuaciones). Podríamos preguntar, por ejemplo, qué tantas posibilidades hay de que un aficionado ingrese a comprar vestimenta a la tienda oficial si ofrecemos un programa de premio a la lealtad en la compra (en una escala del 1 al 10).
- **Categórica**: se basa en mediciones nominales. Se organiza con base en características. Ingreso económico (bajo, clase media, clase alta) es un buen ejemplo para graficarla.
- **Variable latente**: es una entidad no observada, que se presume subyace a las variables observadas. Normalmente, esta variable es inferida gracias a un método matemático, por lo que no haremos hincapié en ella.

Figura 10: Análisis de variables individuales



Fuente: elaboración propia.

Como vemos en la figura anterior, al analizar las diferentes variables de nuestro modelo de negocio, podemos detectar las que tienen mayor o menor potencial de impacto en él, así como analizar la *performance* actual de la variable.

Si algo tiene mucho potencial de impacto, y además es de alta *performance* en nuestra organización, debemos aprovecharla como foco de negocio, centralizar nuestra propuesta de valor alrededor de tal variable. Por otra parte, en el otro extremo, debemos discontinuar todos los esfuerzos que tengan bajo potencial y mala *performance*, pues son una pérdida de tiempo y de recursos.

2. 2. 3. Análisis de relación entre variables

Siguiendo a Bologna (2018), citado en Barrionuevo (2019), podemos analizar la clasificación de las relaciones entre variables.

- Desde el punto de vista del tiempo:
  - Asimétricas: una variable cambia a continuación de la otra en un sentido temporal o lógico (esto no implica que sea a causa de la otra). Se asocian a estudios descriptivos (por ejemplo: a mayor afición en el estadio, mayores ventas en la tienda oficial).
  - Simétricas: se produce una covariación cuando no es posible señalar cuál variable es anterior. Se asocian a estudios explicativos.
- Desde el punto de vista de la dirección:
  - directa. A cambios ascendentes en A le siguen cambios ascendentes en B.
  - Inversa: a cambios ascendentes en A le siguen cambios descendentes en B.
  - Monótona: cuando se espera que todos los resultados de una serie sean directos o inversos.
- Desde el punto de vista de la intensidad: medida de qué tan fuerte es la incidencia (asimétrica o simétrica).
  - Fuerte: es la variable que concentra la mayoría de los casos.
  - Débil: es la que tiene menos casos.

Si dos variables se comportan de tal manera que en alguna medida «se siguen» entre ellas, es posible establecer que existe una asociación o covarianza estadística entre ellas —lo cual no

implica causalidad—; por ejemplo, la venta de prendas en la tienda oficial está estadísticamente relacionada con la cantidad de personas que ingresan a la tienda; aun cuando se sabe que, en teoría, podría llegar a venderse más un día con menos visitantes. Pero se sabe que es habitual que se venda más cuando hay más gente. Si nos referimos a algo genérico, podemos hablar del peso y la altura de la gente: están estadísticamente relacionadas aun cuando se sabe que el peso de nadie está causado por su altura y viceversa, pero se sabe que es habitual que las personas altas pesen más que las personas bajas.

La ciencia de la estadística ofrece numerosos métodos para revelar y presentar las asociaciones entre dos o más variables. La intensidad o fuerza en la relación entre dos variables puede medirse a través del coeficiente de contingencia o la correlación, para lo que existen numerosos métodos de cálculo disponibles.

El coeficiente de contingencia puede aplicarse a todo tipo de variables incluyendo aquellas que se han medido solo con una escala de clasificación. El método de correlación ordinal es adecuado cuando al menos una de las variables se ha medido con una escala ordinal. Para variables sobre escalas aritméticas, el método más frecuentemente utilizado es la correlación estándar, también conocido como correlación del momento-producto o correlación de Pearson.

La correlación del momento-producto suele abreviarse con la letra r. Si el coeficiente de correlación (r) es bajo, por ejemplo, entre -0,3 y +0,3, las dos variables tienen una baja relación entre sí. Si el coeficiente de correlación (r) es alto, por ejemplo, si su valor se aproxima ya sea a +1 o a -1, esto significa que la relación entre las dos variables (ya sea directa o inversa) es fuerte.

Es posible que existan razones para creer que una variable es causalmente dependiente de otra u otras variables. Si existen suficientes datos en este sentido, el análisis de regresión es el método más apropiado para revelar el patrón exacto de esa asociación.

El análisis de regresión consiste en encontrar la ecuación lineal que explica la relación entre las variables y que se desvía lo menos posible de las observaciones individuales. El cálculo es extremadamente técnico, por lo que no le dedicaremos mucho tiempo en esta lectura.

No obstante, para mantenerlo como registro, es importante mencionar que el algoritmo del análisis de regresión construye una ecuación con una o más variables independientes. Además, esa ecuación posee parámetros a1, a2, etc. y b valores que se expresan de la siguiente forma:

$y = a1x1 + a2x2 + a3x3 + \ldots + b$

donde:

y = la variable dependiente;

x1, x2 etc. = variables independientes;

a1, a2 etc. = parámetros;

b = constante.

El análisis de regresión posee la limitación de solo identificar relaciones lineales entre las variables. Esto quiere decir que no puede manejar relaciones del tipo: y = ax2 + bx + c (entre otras), aun cuando esas asociaciones existan.

2. 2. 4. Técnicas de análisis cualitativo

El análisis cualitativo es crítico en la comprensión del gran volumen de datos que el medio *online* recolecta y almacena. Es la forma de responder al porqué del comportamiento de los usuarios y/o clientes. Estudios de usabilidad, testeos remotos y encuestas son algunos de los métodos más comúnmente utilizados.

User research (o investigación de usuarios) es la ciencia de observar y monitorear cómo los usuarios interactúan a diario con sitios web, *hardware* y *software*, con el fin de extraer conclusiones de cómo optimizarlos. Algunas veces, estos estudios se realizan en laboratorios y otras veces en el medio (natural) de los usuarios, como, por ejemplo, en su oficina u hogar.

Según Kaushik (2009), los pasos en la preparación de un testeo cualitativo son los siguientes:

- identificar las tareas críticas a testear.
- Crear escenarios críticos a testear.
- Identificar los criterios de éxito para cada escenario (¿cuándo se considera que la tarea ha sido cumplida?).
- Definir quién participará del testeo.
- Determinar la compensación para los participantes.
- Contratar al reclutador, agencia de investigación, proveedor de encuestas, etc.
- Realizar pruebas internas con el cuestionario, guion u otros materiales antes de exponer a los entrevistados.

Una vez que se ha recolectado la información a través de algunos de los métodos de análisis cualitativos mencionados, corresponde realizar las siguientes tareas:

- realizar una sesión de *debrief* lo antes posible con todos aquellos que hayan participado del esfuerzo de investigación, con el fin de compartir notas y percepciones.
- Intentar comprender patrones de comportamiento y tendencias.
- Sacar conclusiones respecto del éxito (o fracaso) de los participantes en relación con las tareas que se les encomendaron en cada escenario.
- Realizar análisis en profundidad con el fin de comprender relaciones entre las observaciones efectuadas.
- Hacer recomendaciones sobre cómo resolver los problemas identificados en cada tarea crítica. Esto incluye señalar puntos de falla, realizar recomendaciones concretas que mejorarán la experiencia, categorizar las recomendaciones como urgentes, posibles; positivas, pero no necesarias; importantes, pero no urgentes, etcétera.

Una de las principales funciones del análisis cualitativo de un *set* de datos está basada en la posibilidad de explicar las relaciones de los datos.

El mapa de conceptos se utiliza como técnica para intentar estas explicaciones.

Como metodología busca vincular las hipótesis de uso, a través de casos y sesiones individuales, en la replicación al incorporar más volumen.

Analicemos un ejemplo.

Hemos detectado diferentes categorías de público, y pudimos entender el comportamiento efectivo de ese público. Ahora, necesitamos explicar la razón del comportamiento. Para ello, generamos hipótesis, que basamos en diferentes técnicas cualitativas como las siguientes:

- grupos en enfoque.
- Entrevistas en profundidad.
- Análisis de sesiones individuales.
- Revisión de recorrido individual.

Estos análisis nos ofrecen respuestas anecdóticas, en lo individual, sobre la respuesta aplicable a uno de los casos. Debemos contrastar esas respuestas con el *set* general de datos

para, a partir de estas hipótesis, analizar cuáles son las posibles explicaciones y generalizaciones que podremos validar.

Referencias

Barrionuevo, D. (2019). *Apuntes de investigación*. Social Media Trends.

Bologna, E. (2018). *Métodos estadísticos de investigación*. Brujas.

Conexión ESAN. (2015). Datamining: las claves de los procesos de minería de datos. https://www.esan.edu.pe/apuntes-empresariales/2015/07/datamining-claves-procesos-mineria-datos/.

Diagramas UML. (2018). ¿Qué es el KDD o Proceso de descubrimiento de conocimiento? https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/.

Frand, J. (2014). Anderson School of Business. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[Imagen sin título sobre *data sets*]. (s. f.). https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/.

[Imagen sin título sobre proceso de KKD]. (s. f.). https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/.

Kaushik, A. (2009). *Digital Analytics 2.0*. Sybex.

Kerlinger, F., y Lee, H. (2002). *Investigación del comportamiento*. McGraw Hill.

Microstrategy. (2020). Data Mining Explained. https://www.microstrategy.com/en/resources/introductory-guides/data-mining-explained#:~:text=Data%20Mining%20allows%20organizations%20to,protect%20customers%20against%20identity%20theft.

Paz, G. (2016). *Analytics. Análisis y tratamiento de datos deportivos*. FC Barcelona Universitas.