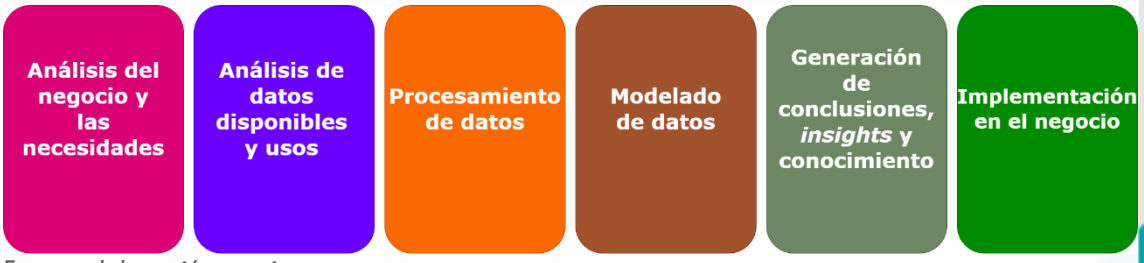
Módulo 2. Del dato a la información para la toma de decisiones

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 irrelevantes nos puede conducir por el camino equivocado. La minería de datos es una de las principales técnicas que podemos utilizar para generar información que nos permita tomar decisiones.

El *data mining* (o la minería de datos) es un conjunto de técnicas que se utiliza por parte de una compañía, una institución, 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, las empresas pueden lograr un mejor entendimiento de los datos y, así, tomar mejores decisiones" (Universidad ESAN, 2015, https://www.esan.edu.pe/apuntes-empresariales/2015/07/datamining-claves-procesos-mineria-datos/).

El *data mining* es un proceso que se realiza en etapas diferentes y que nos permite lograr el conocimiento posible de aplicar para mejorar el rendimiento de nuestro negocio, de nuestras campañas de *marketing*, etcétera.

Figura 1: El proceso de *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 compañía para resolver y una orientación respecto a qué es lo que se busca encontrar. En función de ese conocimiento del negocio (y del modelo del negocio, principalmente) se puede definir qué tipos de datos debemos recabar.

- **Análisis de datos disponibles y usos**: una vez que sabemos qué datos nos interesa conseguir, debemos encontrarlos. Es importante evaluar qué posibilidad real tenemos de captarlos y organizar el proceso para conseguirlos. Para esto

analizaremos las diferentes fuentes donde se consiguen los datos y los formatos en que los tomaremos.

- **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.

- **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 hemos verificado esa información, podemos arribar a conclusiones, obtener *insights* de la información y generar conocimiento que nos permita tomar decisiones informadas para nuestro negocio.

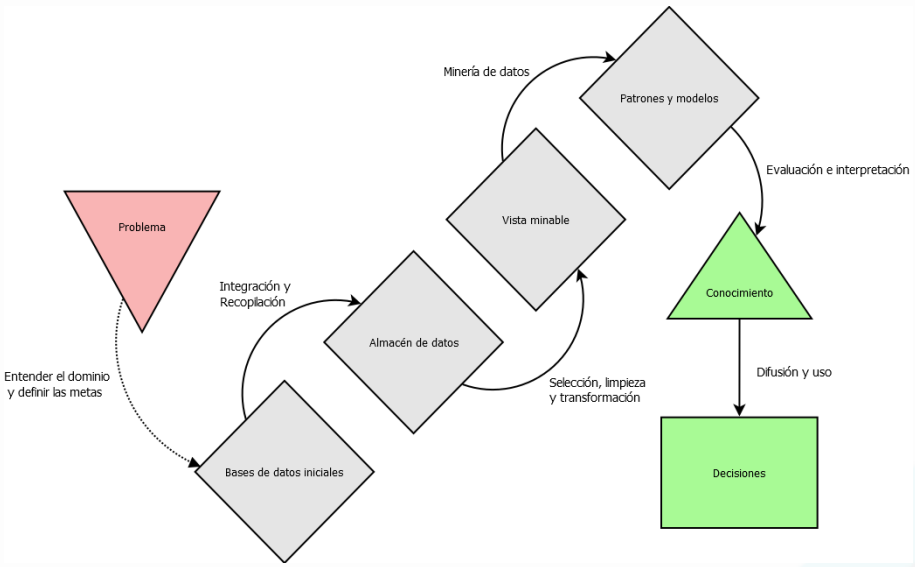
- **Implementación en el negocio**: cuando las decisiones están tomadas, solamente resta implementarlas y, por supuesto, medir el resultado de tal implementación.

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 analizar bloques de datos e información para extraer conocimiento aplicable a nuestro negocio. Desglosemos esta definición un poco más:

- **Descubrimiento**: nos permite determinar dos cuestiones centrales:
    - Es un proceso, por ende, hay una serie de pasos involucrados en el desarrollo de una determinada actividad. En este caso, extraer conclusiones sobre las relaciones y patrones que nos permitan tomar decisiones de negocio tendrá ciertos pasos a seguir para lograrlo.
    - 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 a entender que suceden y por qué.
- **Implica un conocimiento**: es el objetivo del proceso que realizamos. Nos interesa encontrar relaciones causales, correlaciones, elementos verificables que nos permitan saber qué decisiones tendrán mayores chances de éxito.
- **Se realiza en bases de datos**: este aspecto nos determina el universo de análisis donde trabajaremos con esta técnica.

Figura 2: Proceso de KDD



Fuente: Diagramas UML (2019). Fases del kdd. Recuperado de https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/.

2.1.1 *Knowledge discovery* y *data mining*

En una amplia variedad de disciplinas vemos cómo los datos creados a partir de la implementación de nuevas tecnologías se acumulan a un ritmo dramático. Existe una necesidad imperiosa de contar con herramientas y técnicas que 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. El *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, tendencias o reglas" (Muñoz de Frutos, 2017, https://computerhoy.com/noticias/internet/que-es-data-mining-70663) entre diferentes campos en bases de datos relacionales de tamaño considerable.

Entonces, ¿en qué se diferencian ambos conceptos? Básicamente, 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. Este último concepto se refiere, en 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*, para lograr cumplir el proceso de *knowledge discovery*: selección y preparación de los datos, limpieza de datos, incorporación de información previa, interpretación de los datos, etcétera. Los veremos en la figura 2.

> 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 tanto 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, 2019, https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/)

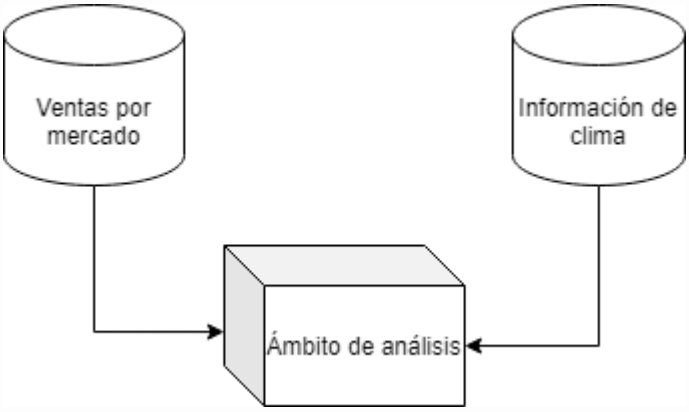
Los para realizar el *knowledge discovery son:*

1. **Identificar el problema:** el problema es la razón de ser del proceso, lo que nos determina la necesidad de realizar un análisis.
2. **Entender el dominio y definir las metas:** este problema debe ser definido de manera concreta y, para ello, debemos plantearnos objetivos.
3. **Conseguir las bases de datos iniciales:** comenzaremos nuestro proceso de uso de datos con la comprensión de aquella información que ya tenemos.

Debemos listar las fuentes internas y externas que poseemos, y reconocer cuáles requerimos conseguir.

4. **Integrar y recopilar datos:** en esta fase debemos vincular los datos con los que trabajaremos. Relacionaremos las fuentes (internas y externas), identificaremos la manera en que se pueden igualar y procesar. Tomemos un pequeño ejemplo: si quisiéramos entender si existe una relación entre las ventas de un producto y las condiciones climáticas, seguramente deberemos integrar y recopilar las bases de datos propias de la venta del producto y las bases de datos externas de la información climática en los diversos mercados donde estamos presentes.

Figura 3: Ejemplo de vinculación de datos



Fuente: elaboración propia.

5. **Crear el almacén de datos:** es un "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, 2019, https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/).

6. **Selección, limpieza y transformación:** como las fuentes originales que usamos de los datos no están necesariamente desarrolladas o diseñadas para vincularse entre sí, es posible que percibamos imperfecciones, desfasajes, incongruencias u otros inconvenientes que es importante subsanar antes de realizar la minería de datos.

- **Datos irrelevantes:** existen datos que caen fuera de nuestro universo de análisis y que, de utilizarlos, sumarán una complejidad innecesaria al estudio o desviarán los resultados. En el ejemplo anterior, si estamos analizando el comportamiento de las compras en función del clima, probablemente debamos definir qué datos de ventas nos resultan útiles (estudiar algunos productos en particular, por ejemplo, y no todas las ventas del negocio). Asimismo, elegiremos qué información del clima analizaremos (quizás la temperatura y las condiciones de lluvia sean factores clave, y no así la presión atmosférica).

- **Valores atípicos o *outliers***: son datos que, aun siendo válidos, pueden distorsionar la muestra porque se encuentran en puntos extremos. Podemos reunir estos valores y transformarlos en una categoría posible de análisis. En el caso del ejemplo sería, si un día hubo 14 grados bajo cero, categorizarlo junto a otros días de "mucho frío" o de "temperaturas bajo cero" y reunir allí a todos los *outliers*.

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

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

7. **Crear la 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. Es la creación de la base procesada una vez completo el proceso anterior.

8. **Realizar la minería de datos:** como hemos mencionado, la minería de datos busca explorar y analizar datos de gran volumen para intentar develar patrones y reglas de comportamiento relevantes. Se puede realizar, principalmente, de tres 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 y algoritmo de vecino más próximo.

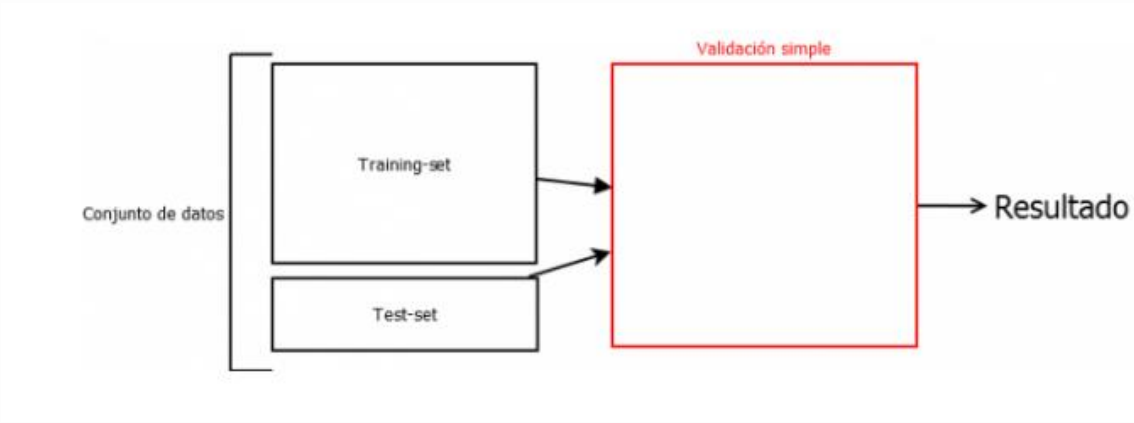
    - **Aprendizaje no supervisado**: busca entender y describir los datos con la intención de descubrir patrones de comportamiento subyacentes. Los algoritmos de recomendación son un claro ejemplo de este método. Algunos de los modelos que se utilizan en aprendizaje no supervisado son la agrupación, los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

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

9. **Patrones y modelos:** de acuerdo a 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, la categorización y la predicción de eventos futuros.

10. **Evaluación e interpretación:** en este paso debemos ver cómo los patrones detectados y los conocimientos generados 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*. Ambos *sets* se preparan previamente para realizar verificaciones no sesgadas.

Figura 4: *Data sets*



Fuente: Diagramas UML (2019). Validación simple. Recuperado de https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/.

11. **Conocimiento:** luego de validar 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, 2019, https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/)

12. **Difusión y uso:** se trata de lograr que el conocimiento generado llegue a la gente que lo necesita. Para ello, debemos transformar el conocimiento en un entregable fácil de comprender y de aplicar.

13. **Decisiones:** finalmente, debemos poder tomar decisiones (o ayudar a quien deba hacerlo) de acuerdo con el 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 muchos *softwares* específicos que se utilizan para realizar esfuerzos de *data mining*. Estos programas permiten analizar un mismo conjunto de datos desde

diferentes dimensiones o ángulos, categorizarlo y resumir las relaciones y patrones encontrados en base a búsquedas abiertas.

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

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

1. **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 restaurante utiliza los datos de los pedidos más frecuentes en determinado día y horario para definir sus menús especiales y, de esta manera, aumentar el tráfico en días y horas definidas.
2. **Clústeres:** relación que se verifica cuando los datos se agrupan de acuerdo a vinculaciones lógicas. Un ejemplo de esta actividad es el análisis de datos de ventas y perfiles demográficos a fin de identificar segmentos de mercado o preferencias de los consumidores.
3. **Asociaciones:** los datos también pueden analizarse para identificar asociaciones por correlación simple. La relación, conocida y verificada, entre la compra de cerveza y pañales es un ejemplo de esto.
4. **Patrones secuenciales:** la relación que existe cuando los datos se analizan para encontrar patrones y/o tendencias. Un ejemplo de esto es la capacidad de un comerciante de artículos de campamento para predecir la compra de una mochila, basado en la previa adquisición de una bolsa de dormir, una tienda de campaña, etc.

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

- Extracción, transformación y carga de datos en un sistema de almacenamiento de datos (*data warehouse system*).
- Acumulación y gestión de los datos en un sistema de bases de datos multidimensionales.
- Acceso a los datos por parte de analistas de datos y profesionales de IT (*information technology*).
- Análisis de los datos utilizando un *software* específico.
- Presentación de los datos en un formato adecuado.

Para realizar análisis descriptivos (aquellos limitados a las relaciones o patrones de comportamientos existentes y verificados) y estudios predictivos (aquellos que van más allá y definen potenciales relaciones o comportamientos futuros en base a los datos actuales) existen métodos de *data mining*. Los más comúnmente utilizados son:

- **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. La estimación de densidad probabilística se encuentra 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* implican la construcción de algoritmos específicos para implementar 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.

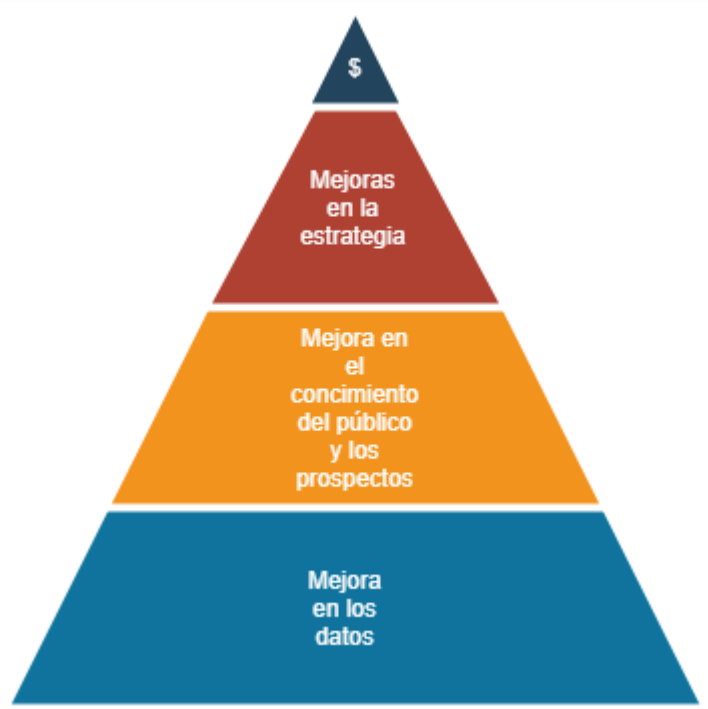
Figura 5: *Software* de *data mining*

Fuente: Microstrategy, 2020.

2.1.3 Usos y aplicaciones de *data mining* para *marketing*

Las actividades de *data mining* se desarrollan, prioritariamente, en compañías con un foco especial en sus consumidores (o usuarios) y que, a su vez, poseen los mecanismos y la infraestructura apropiada para obtener una considerable cantidad de datos acerca de los mismos. Estas organizaciones, generalmente, analizan sus datos internos (precios o posicionamiento de marca, por ejemplo) en búsqueda de relaciones y/o patrones de comportamiento vinculados con datos externos (características demográficas de sus consumidores, actividades de sus competidores o indicadores económicos). Este análisis permite determinar el impacto de las variaciones en dichas variables sobre el volumen de ventas, satisfacción de los clientes, etc.

Figura 6: Pirámide de *data mining* para *marketing*



Fuente: elaboración propia.

A medida que profundicemos en esta práctica nuestros datos nos permitirán trepar en la pirámide y lograr resultados en *marketing*. Vamos a analizar cada aspecto de la figura:

- **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 tienen mejor y peor desempeño, 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) gracias a la 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 de pedidos y los costos de producción y *servucción*, entre otros beneficios.

2.1.4 Herramientas e infraestructura requeridas

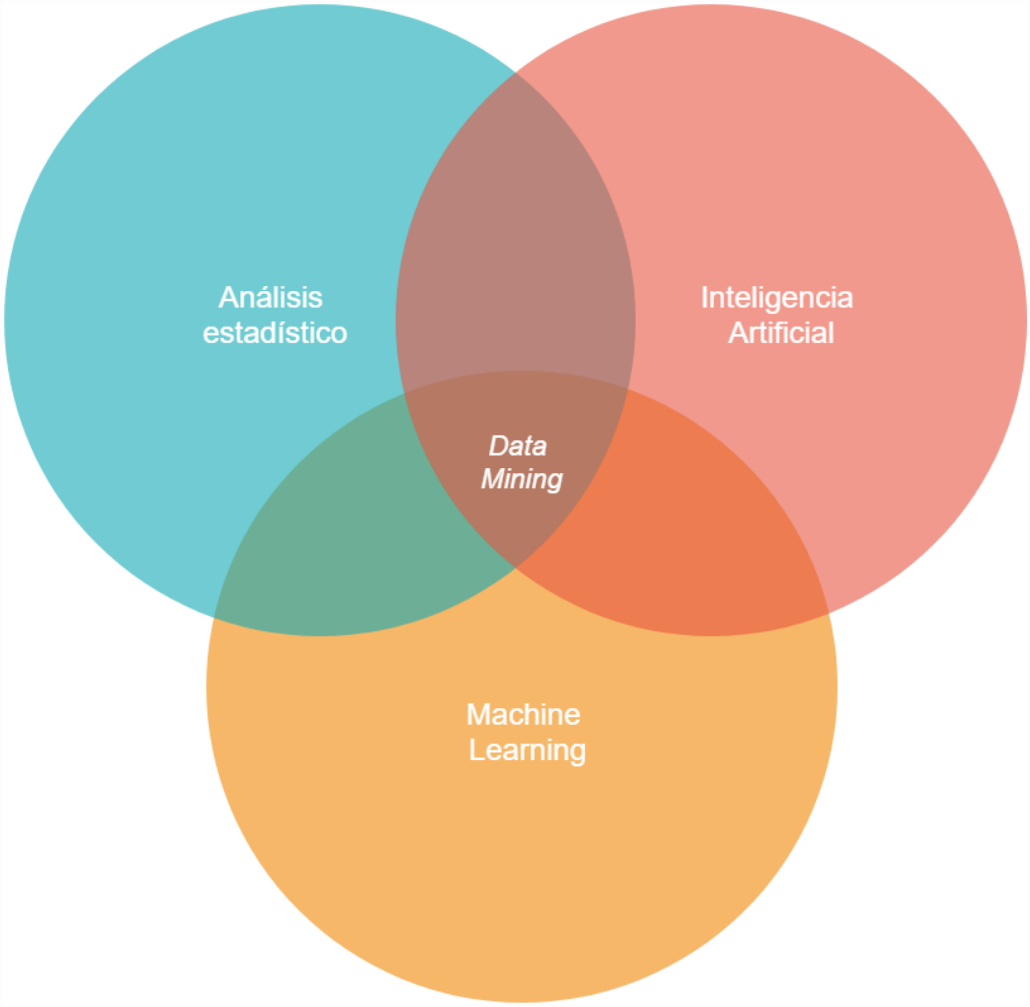
En la actualidad, hay aplicaciones de *data mining* disponibles para infraestructuras de todo tipo. Los precios van desde los cientos hasta varios millones de dólares. Las variables críticas son:

- **Tamaño de las bases de datos**. Mientras más cantidad de datos se almacenen, procesen o mantengan, serán necesarios sistemas e infraestructuras más potentes.

- **Complejidad y frecuencia del análisis**. Mientras más complejos sean los análisis requeridos y mayor la frecuencia de consulta, se requerirán sistemas e infraestructura más potentes.

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

Figura 7: Las herramientas e infraestructura en *data mining*



Fuente: elaboración propia.

El *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; es decir, el uso de *software* o máquinas para desarrollar procesos de análisis y de generación de conocimiento.
- El *machine learning*; es decir, 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 recolectarlos para responder las preguntas que nos ayudarán a resolver nuestros problemas. De acuerdo con las técnicas que elijamos utilizar, obtendremos información numérica en datos cuantitativos o información ilustrativa, como los datos cualitativos. También podemos adquirir una combinación de ambas opciones si decidimos 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.

Podemos trabajar con los datos 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 realizadas son nuestro *set* de datos (los valores de las variables obtenidas para una muestra de unidades) o matriz de datos (donde 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).

- Datos cuantitativos: requieren uso de análisis estadístico. Las variables pueden ser identificadas y sus relaciones, medidas. Se cuentan o expresan de manera numérica.

- Datos cualitativos: examinan datos no numéricos en busca de patrones y significados. Son recolectados y analizados con algún mayor grado de subjetividad.

- Datos mixtos: pueden explicar algunos resultados inesperados (los denominados *outliers* o excepciones) que, utilizando un solo enfoque, no se puede. (Paz, 2016, 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 para evitar que nos conduzcan a conclusiones erróneas. No obstante, no conviene eliminar ningún dato por ser meramente anormal, sino simplemente diferenciarlo del conjunto de los demás datos.

- **Normalizar o reducir los datos**. Significa eliminar todos aquellos datos que son irrelevantes o que, si bien tienen alguna influencia sobre las variables, no resultan de interés en el momento del análisis.

En el análisis propiamente dicho 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 busca 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 quiere analizar las diferentes variables medidas en forma inconexa o las relaciones entre dichas variables.

Otra dimensión importante hace referencia al propósito final del proyecto. Se trata de preguntarse: ¿el objetivo del análisis es describir cómo es el estado actual o, por el contrario, se busca predecir comportamientos futuros en función de la información obtenida acerca de las variables independientes y/o sus relaciones? De esto dependerá si se utilizan herramientas de análisis estadístico descriptivo o análisis estadístico predictivo.

Revisemos nuestros conocimientos básicos de estadística. Repasemos algunas definiciones claves:

- **Población**: la población es la colección de todos los individuos o ítems que están bajo consideración en un estudio estadístico.
- **Muestra**: la muestra es la parte de la población de la que recolectamos información.
- **Variable**: una variable es un valor que puede cambiar de acuerdo a las condiciones o por diferentes situaciones. Es un elemento o factor que puede variar, por no ser fijo o consistente. Por ejemplo: la altura, el peso, la cantidad de puntos por partido, la velocidad, etcétera. (Paz, 2016, p. 46)

Población y **muestra** son dos conceptos básicos que debemos comprender. Tienen que ver con el ámbito que se estudia o analiza en un momento en particular. Son dos aspectos íntimamente relacionados ya que la **población** es todo el *set* de personas o de objetos sobre los cuales queremos obtener conclusiones; sin embargo, al no poder acceder a la totalidad de estos datos, debemos recurrir a una **muestra**, que es una porción de la población.

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 surge de una investigación.

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, se conocen. Un parámetro es un resumen numérico desconocido de una población que nos permite hacer inferencias. Para definir los parámetros se utilizan técnicas estadísticas.

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 conjunto 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. La media de la variable en una muestra es la suma de todos los valores observados dividido sobre la cantidad de valores observados. (Paz, 2016, p. 51)

La fórmula de la media es:

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

> **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, p. 52)

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, haciendo 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 conjunto de datos.

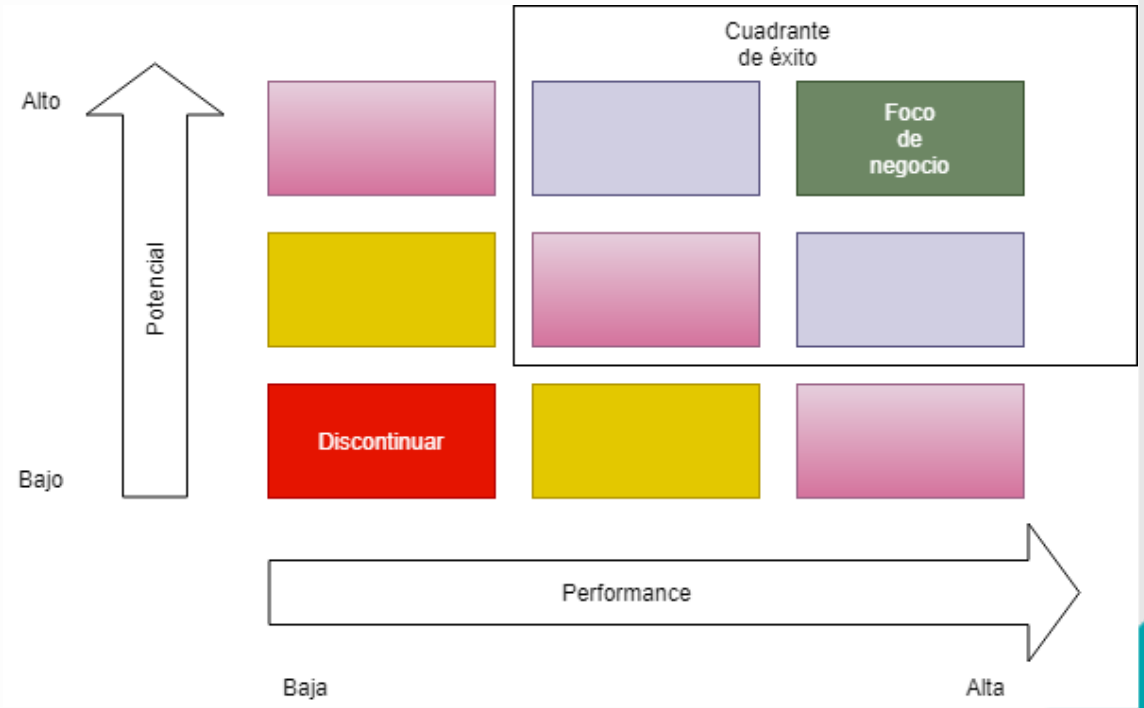
2.2.2 Análisis de variables individuales

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 retomar una categorización de las variables de los autores Kerlinger y Lee (en Barrionuevo, 2019) que las divide en:

- **Cualitativas.**
    - **Dicotómicas (o binarias)**: solo aceptan dos valores.
    - **Polinómicas:** aceptan diversos valores.

- **Independientes o 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.
- **Continua**: es capaz de asumir un conjunto ordenado de valores dentro de cierto rango (se asocia con escalas y puntuaciones).
- **Categórica**: está basada en mediciones nominales. Se organiza en base a características.
- **Variable latente**: es una entidad no observada que se presume subyace a las variables observadas.

Figura 9: Análisis de variables individuales



Fuente: elaboración propia.

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

Si identificamos algún aspecto que tiene mucho potencial de impacto y, además, es de alta *performance* para nuestra compañía, debemos aprovecharla como foco de negocio y 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 resultan una pérdida de tiempo y recursos.

2.2.3 Análisis de relación entre variables

De acuerdo con Bologna (en Barrionuevo, 2019) podemos analizar las relaciones entre las variables en distintos sentidos:

a) **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.
- **Simétricas**: se produce una covariación cuando no es posible señalar cuál variable es anterior. Se asocian a estudios explicativos.

b) **Desde el punto de vista de la dirección**:

- **Directa**: suceden cambios ascendentes en A y le siguen cambios ascendentes en B.
- **Inversa**: suceden cambios ascendentes en A y le siguen cambios descendentes en B.
- **Monótona**: cuando se espera que todos los resultados de una serie sean directos o inversos.

c) **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 evolucionan modo tal que, en alguna medida, se siguen entre ellas, podemos decir que existe una asociación o covarianza estadística entre ellas. Por ejemplo, la altura y peso de la gente están estadísticamente asociadas: aunque el peso de nadie esté causado por su altura ni la altura por el peso es; no obstante, 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 y hasta más variables. Los medios más simples son los medios de presentación gráfica y tabulación. La intensidad de la asociación entre variables puede también describirse como una estadística especial, como el coeficiente de contingencia y una correlación para lo que hay varios métodos de análisis disponibles.

(…)

- **Cociente de contingencia** puede aplicarse a todo tipo de variables, incluyendo aquellas que se han medido solo con una escala de clasificación. Una estadística alternativa es Chi cuadrado.

- **Correlación ordinal** es adecuada cuando al menos una de las variables se ha medido con una escala ordinal. La otra puede ser u ordinal o aritmética.

- Para variables sobre escalas aritméticas, el método usual es la **correlación** estándar, mejor dicho, la 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 es alto, en otras palabras, 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] se aproxima a la ecuación $y = ax + b$ [es decir, es fuerte]. (Routio, 2007, http://www.uiah.fi/projects/metodi/280.htm)

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 algoritmo del análisis de regresión construye una ecuación con una o más variables independientes. Además, esa ecuación posee parámetros (*a1*, *a2*, etc.) y valores (*b*) que se expresan de la siguiente forma:

$$y = a\,1x1 + a\,2x2 + a\,3x3 + \ldots + b$$

Donde:

y es la variable dependiente.

x1, *x2*, etc. son las variables independientes.

a1, *a2*, etc. son los parámetros.

b es la constante.

El análisis de regresión posee la limitación de que solo identifica 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 conocer por qué los usuarios y/o clientes se comportan de determinado modo. 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* a 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, tal como su oficina u hogar.

Según Kaushik (2009) estos son los pasos para preparar un testeo cualitativo:

1. Identificar las tareas críticas a testear.
2. Crear escenarios críticos a testear.
3. Identificar los criterios de éxito para cada escenario (es decir, preguntarse ¿cuándo se considera que la tarea ha sido cumplida?).
4. Definir quién participará del testeo.
5. Determinar la compensación para los participantes.
6. Contratar al reclutador, agencia de investigación, proveedor de encuestas, etc.
7. 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 *debriefing* lo antes posible con todos aquellos que hayan participado del esfuerzo de investigación a 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 a las tareas que se les encomendaron en cada escenario.
- Realizar análisis en profundidad a 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 podrían mejorar 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 es la posibilidad de explicar las relaciones existentes entre ellos. El mapa de conceptos es una técnica que se utiliza para esto. Se trata de una metodología que 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 que motiva a ese comportamiento. Para ello, generamos hipótesis basadas en diferentes técnicas cualitativas como:

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

Estos análisis nos ofrecen respuestas anecdóticas, a nivel individual, sobre la respuesta aplicable a uno de los casos. Debemos contrastar esas respuestas con el *set* general de datos para, a partir de esos resultados, analizar cuáles son las posibles explicaciones y generalizaciones que podremos validar.

Referencias

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

Diagramas UML (4 de julio de 2019). *¿Qué es el KDD o Proceso de descubrimiento de conocimiento?* Recuperado de https://diagramasuml.com/que-es-el-kdd-o-proceso-de-descubrimiento-de-conocimiento/.

Frand, J. (2 de marzo de 2014). Anderson School of Business. Obtenido de http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

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

Muñoz de Frutos, A. (2017). *¿Qué es Data Mining?* Recuperado de https://computerhoy.com/noticias/internet/que-es-data-mining-70663.

Microstrategy. (03 de 12 de 2020). Microstrategy.com. Recuperado de https://www3.microstrategy.com/es/resources/introductory-guides/data-mining-explained

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

Routio, P. (2007). *Análisis cuantitativo*. Recuperado de http://www.uiah.fi/projects/metodi/280.htm.

Universidad ESAN (31 de 07 de 2015). *Data mining: las claves de los procesos de minería de datos*. Recuperado de https://www.esan.edu.pe/apuntes-empresariales/2015/07/datamining-claves-procesos-mineria-datos/.