Analista de Big data

Competencia 1: Analista de Big data

Índice:

El impacto de los datos masivos en la sociedad actual.

Big Data es la consecuencia de que estamos inmersos en un mundo tremendamente digital, que esto genera cantidad de interacciones de datos, nos deja un rastro y hoy en día existe la tecnología para poder capturar toda esta información, analizarla, procesarla y utilizarla para tomar decisiones, entre otras cosas. Por lo tanto, ¿qué impacto tiene Big Data para nosotros como individuos, para sociedad, para las empresas? En primer lugar, Big Data tiene un impacto para nosotros como individuos, en el sentido de tener mucho más acceso a la información por múltiples canales, nos impacta también como ciudadanos, por el hecho de poder acceder a servicios digitales de mucho más valor, nos impacta como consumidores porque debemos interaccionar con nuestros proveedores de servicios de manera más digital, pero también nosotros, gracias a Big Data, podemos tener muchísima más información para tomar también nuestras propias decisiones. Tenemos una interacción multicanal con los proveedores de productos, con los proveedores de servicios, y en esta experiencia digital podemos tener acceso a productos y servicios mucho más personalizados, a ofertas mejores para nosotros, gracias al Big Data. Por lo tanto, Big Data supone para nosotros como ciudadanos, como consumidores, un mundo de oportunidades, de generación nuevas fuentes de ingresos, de nuevos modelos profesionales, incluso de nuevos modelos de negocio, y una mejor relación también como ciudadanos con estas administraciones y como consumidores con las empresas que nos proporcionan productos y servicios.

Impacto del Big Data en la sociedad

Big Data y Analítica Avanzada para:

- Personalizar los servicios
- Mejorar eficiencias de procesos
- Disminuir riesgos

Desde el punto de vista de la sociedad, Big Data ha permitido y está permitiendo grandes avances, primer lugar, en el mundo científico, en el mundo médico, por ejemplo, lo que es la medicina personalizada, la mejora de los tratamientos de, y diagnóstico, enfermedades tan serias como cáncer están viendo grandes avances gracias al Big Data, a la medicina preventiva, a la medicina personalizada.

Impacto del Big Data en la sociedad

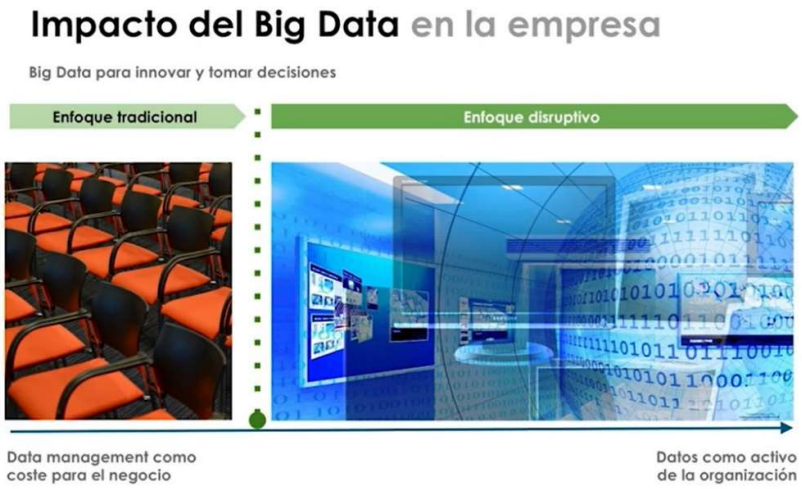
- → Desarrollo de medicina preventiva
- → Tratamientos personalizados
- → Detección de patrones en enfermos

(...)

Gracias al Big Data tenemos acceso a herramientas de formación y educación online como las diversas plataformas que hoy en día existen, en la que podemos personalizar los contenidos y adaptarlos cada vez más a las necesidades individuales de formación.

Gracias a Big Data tenemos redes sociales como LinkedIn, como Facebook, como Twitter, que también nos permiten conectarnos permanentemente a una red ingente de contactos a nivel mundial, y la sociedad también gracias a eso puede avanzar en la mejora de las relaciones entre los ciudadanos y las administraciones. También Big Data se aplica a nivel social en la predicción del terrorismo, en la predicción del crimen y también se está aplicando cada vez más para poder resolver los desafíos grandes que tenemos de la humanidad, la predicción de la pobreza, la predicción de plagas, la predicción de terremotos, todo eso detrás están tecnologías Big Data hoy en día.

Y por último, Big Data está dando muchas oportunidades en el entorno empresarial, está cambiando, en primer lugar, la manera en que las empresas deben entender la información para tomar decisiones; en segundo lugar, permite mejorar la manera en que las empresas trabajan, permite ayudarles a innovar y a desarrollar nuevos productos y servicios, y cómo mejorar también la eficiencia de sus procesos, pero también Big Data nos permite desarrollar nuevos modelos de negocio basados en datos que están transformando todas las industrias y que probablemente están disrumpiendo, generando las industrias del futuro.



Introducción al Big data

En la analítica tradicional, una persona, el analista, el científico, toma información de bases de datos o de ficheros y mediante técnicas analíticas obtiene modelos que pueda utilizar en negocio, en investigación o en el ámbito en el que trabaje. Para ello, lo normal es utilizar su ordenador, pero también puede utilizar **servidores o supercomputadores** si necesita capacidad de cómputo elevada. Con el **Big Data** una de las soluciones que se obtiene es que ahora podemos almacenar muchísima más información.

El coste de almacenamiento de información se ha reducido muchísimo por lo tanto podemos almacenar mucho más **volumen**. En 1992, con 500 dólares se podía almacenar sólo 1 giga de información mientras ahora podemos almacenar 26 mil gigas de información con ese monto. A esto hay que sumarle que cada 24 meses se está duplicando la capacidad de los microprocesadores gracias a lo que establece la **Ley de Moore** que lleva vigente los últimos 50 años. Esto permite que podamos procesar muchísima información en mucho menos tiempo y con algoritmos mucho más potentes.

También, aparte de utilizar **CPU's** se están evolucionando a utilizar, en el tratamiento información, otro tipo de procesadores como las **GPU** o las **TPU,** de forma que podamos procesar más información en menos tiempo con determinados tipos de algoritmos.

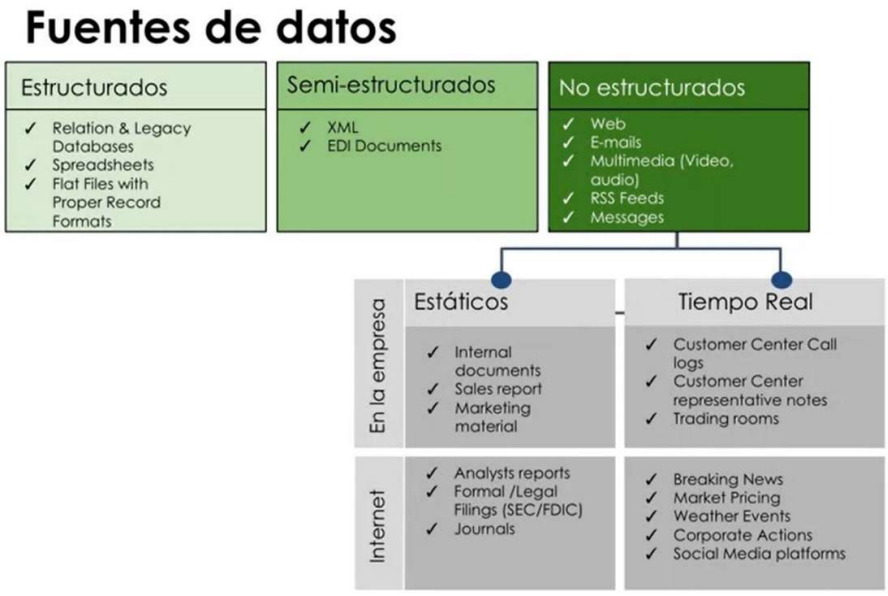
También, nos afecta la **velocidad**, que las comunicaciones a nivel mundial están mejorando muchísimo.

De forma que ya no tenemos que tener el ordenador cerca, sino que podemos considerar otro tipo de soluciones como trabajar en la nube o tener soluciones híbridas. Entonces, por un lado, tenemos que podemos almacenar muchísima cantidad información, y por otro lado tenemos, que podemos procesarla una velocidad muy elevada. Históricamente, como era muy costoso almacenar información, lo que hacíamos era seleccionar cuál era la información más importante de la que disponíamos y almacenábamos solamente esa información. Ahora, como es mucho más barato, lo que estamos haciendo es almacenar toda la información posible para no perder conocimiento. A la información estructurada que hemos tenido históricamente ahora se suman información no estructurada, como por ejemplo las imágenes, los textos y los sonidos. Esto se ha debido gracias a la transformación digital que nos está permitiendo transformar fotos en datos, sonidos en datos, y texto en datos de forma que podemos procesarlo.

A esto hay que sumarle que gracias a que la velocidad de las comunicaciones ha aumentado, estamos cambiando el paradigma de la computación pasando de una computación centrada en servidores y supercomputadores a una computación centrada en sistemas distribuidos, gracias a que ahora los ordenadores pueden comunicarse muy rápidamente entre sí. Las ventajas de estos sistemas distribuidos es la alta disponibilidad, la tolerancia a fallos, la escalabilidad, y el bajo coste que tiene. Mas allá de estas ventajas, posee pequeños inconvenientes a la hora de entrenamiento de información.

Primero, la seguridad:

- **Hay más puntos de entrada y por lo tanto hay problemas de seguridad.**
- **Hay más puntos de falla al tener muchas más máquinas que hay que coordinar.**



Por último, para la algoritmia avanzada, necesitamos **algoritmos específicos y software específico de computación distribuida**.

Esto ha permitido que muchos algoritmos, principalmente de **Machine Learning** de inteligencia artificial, que antes no era posible utilizar porque no convergían por la falta de información o la falta de capacidad de cómputo, ahora si lo hacen, de forma que podemos utilizar técnicas mucho más sofisticadas.

Por eso actualmente se están desarrollando tan rápidamente la inteligencia artificial y en particular el **Deep Learning**, el **Speech Recognition** y el **Natural Language Processing**. Si volvemos en el tiempo, vemos que la analítica tradicional basada en datos estructurados, técnicas analíticas y trabajando sobre servidores y supercomputadores para generar modelos analíticos cambia el mismo usuario. Ahora utiliza datos estructurados, no estructurados, provenientes de imágenes, de texto, de sensores. Estos datos pueden estar alojados en un ordenador o incluso pueden estar alojados en una nube, en un sistema distribuido y podemos utilizar ahora ya no solo las técnicas tradicionales clásicas sino

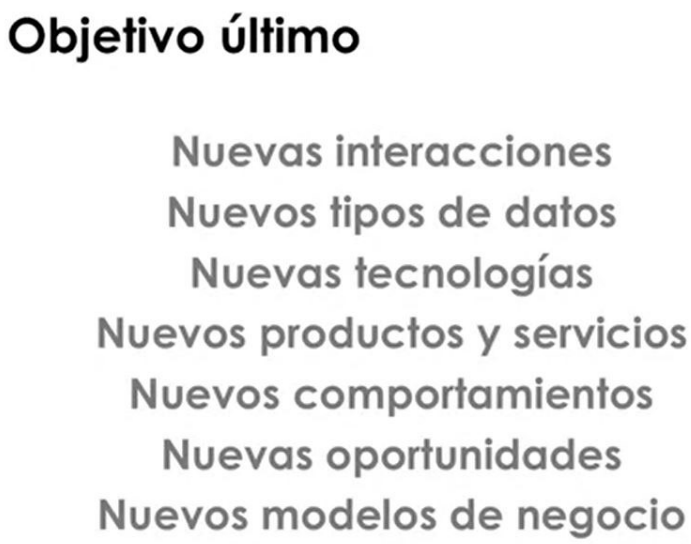
además técnicas mucho más potentes de **Machine Learning** o de inteligencia artificial, que lo que nos permiten obtener modelos analíticos más potentes.

Como resumen de las V's del Big Data hemos hablado que el big data nos aporta:

- **Volumen**: Podemos almacenar muchísima información.
- **Velocidad**: Los datos se generan a una gran velocidad. Por ese motivo, muchos de ellos quedarán obsoletos en cuestión de poco tiempo y perderán su valor cuando aparezcan otros más recientes.
- **Variabilidad**: Los datos provienen de diferentes canales: redes sociales, dispositivos propios
- **Valor**: Hace referencia a la información útil que se puede extraer de los datos. Esa misma que te ayudará a generar un valor agregado para tu negocio
- **Veracidad**: Es probable que debido al gran volumen de datos que recibimos, algunos lleguen incompleto. Y es que todo lo que recibimos de Internet y sobre todo de las redes sociales no es fiable.

Estos son los grandes retos que tiene el Big data desde el punto de vista del tratamiento de información. Los principales retos que vamos a tener son primero, la complejidad tecnológica y segundo la complejidad algorítmica. Después, necesitamos **equipos multidisciplinarios**,

porque en esta complejidad, no solo una persona es capaz tener todo el conocimiento para poder desarrollarlo y, por último, necesitamos **conocimiento experto**, porque toda esa información necesitamos alguien que la interprete, que le saque valor y a los resultados también, porque después hay que comunicarlos. Para poder realizar todo esto, lo que es necesario es establecer una metodología del procesamiento de datos con una metodología de procesamiento Big Data.

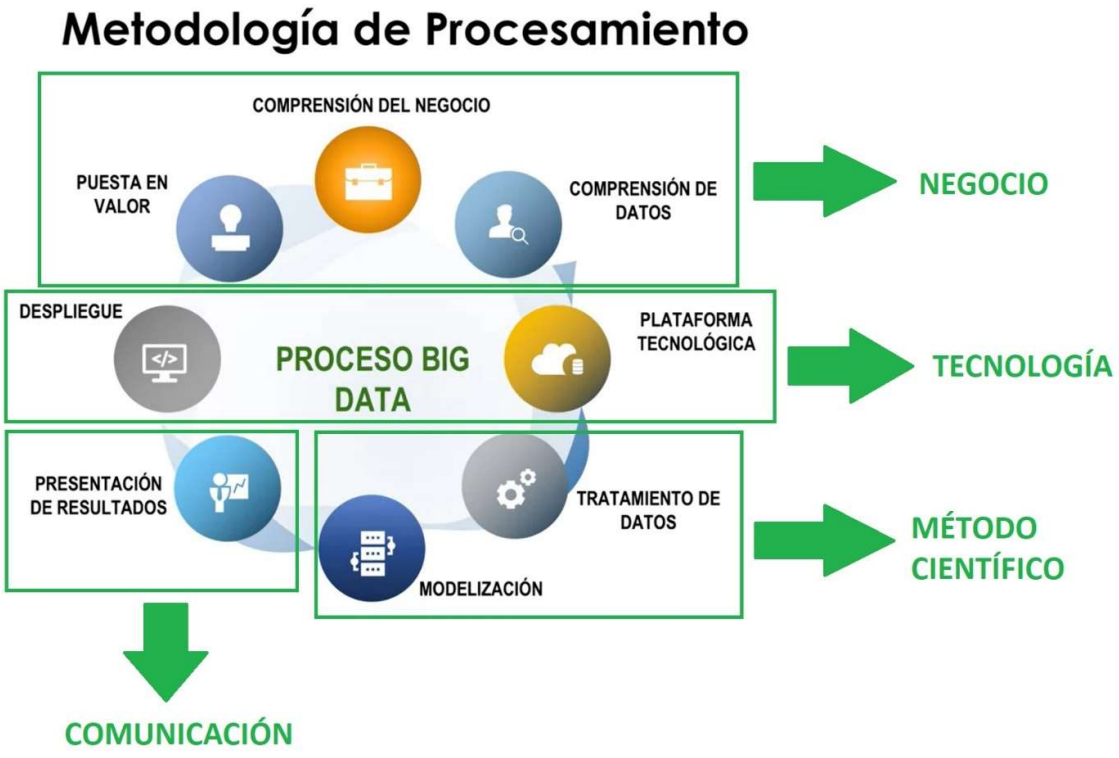


Introducción al procesamiento de datos masivos.

La metodología de procesamiento de grandes volúmenes de datos permite transformar los datos en conocimiento. Para ello, vamos a ver tres grandes cosas. Primero, la propia metodología de procesamiento de datos; segundo, cuáles son las componentes de esa metodología en alto nivel; y por último, los factores de éxito para que esta metodología sea una realidad en las compañías.

Comenzamos con la metodología de procesamiento. Está compuesta de ocho fases, si comenzamos arriba y vamos recorriendo en el sentido de las agujas del reloj, la primera fase es la comprensión del negocio, en donde lo que se pretende es saber cuál es el problema y cuál es el objetivo. La segunda fase es la comprensión de datos, en la que lo que queremos es conocer cuáles son los datos necesarios para poder resolver el problema. La tercera fase es la plataforma tecnológica, ¿dónde vamos a trabajar? ¿Qué tecnología necesitamos? ¿Qué componentes? La cuarta fase es el tratamiento de datos. Una vez que ya tenemos los datos, ¿cómo los tenemos que procesar? ¿Cómo los vamos a integrar? ¿Qué vamos a hacer

con ellos? La siguiente fase es la modelización, donde lo que hacemos es, con técnicas estadísticas avanzadas, crear modelos que nos permitan extraer el conocimiento de los datos. La siguiente es la presentación de los resultados. Todo aquello que hemos obtenido, tenemos que contarlo a nuestros stakeholders, los resultados que hemos obtenido. La siguiente es el despliegue. Una vez que tenemos la aprobación, tenemos que hacer el despliegue del modelo. Y, por último, está la puesta en valor. Una vez que ya tenemos el modelo, ¿cómo lo utilizamos? ¿Para qué nos sirve? Esta metodología se divide en varias componentes. La principal es la componente de negocio, porque realmente un modelo lo importante es que sea accionable, un modelo tiene que resolver un problema real. Entonces, es muy importante que tengamos ese conocimiento experto. Por tanto, es fundamental que desde el principio tengamos esa visión de la componente de negocio. La segunda componente es la tecnología. Sin esta tecnología, no vamos a ser capaces de procesar grandes volúmenes de información. Las fases en la que afecta más la tecnología son la de plataforma tecnológica y la de despliegue. La tercera componente es la componente científica. Estos modelos construyen aplicaciones utilizando el método científico, utilizando técnicas analíticas. Por lo tanto, es importante que tengamos claro que tenemos que utilizar esa componente científica. Y, por último, tenemos que tener una componente de comunicación. Es fundamental tener claro que, si no somos capaces de comunicar resultados, si no somos capaces de comunicar ese conocimiento que hemos adquirido, perderemos valor.



Para poder implementar esta metodología en cualquier empresa o en cualquier ámbito, lo importante es que tengamos en cuenta lo siguiente. Para empezar, tenemos que tener datos, si no tenemos datos no vamos a poder hacer nada. Lo siguiente es, talento, fundamental. Es necesario que contemos con las personas con los conocimientos adecuados y las capacidades adecuadas para poder tratar toda esta información. Después, obviamente, herramientas analíticas y tecnología. Si tenemos datos, pero no somos capaces de tratarlos, no somos capaces de tener una plataforma tecnológica suficientemente potente para poder desarrollar esos modelos analíticos, no vamos a poder terminar el trabajo. Y por último, para que todo esto sea una realidad, hace falta una cultura organizacional en la que se premie toda esta visión de negocio y, sobre todo, se entienda que la ciencia viene a aportar valor.

Ya vimos de manera general como se componen las diferentes fases de la metodología de procesamiento en Big Data, pero vamos a profundizar un poco mas en cada una de ellas a continuación.

- **Comprensión de negocio.**

El objetivo de esta fase es identificar, analizar y comprender el problema y traducirlo a un problema analítico. Las etapas de esta fase son la identificación del problema, la fijación de los objetivos, la identificación de los implicados y por último la fijación de la tipología de análisis.

Comprensión de Negocio

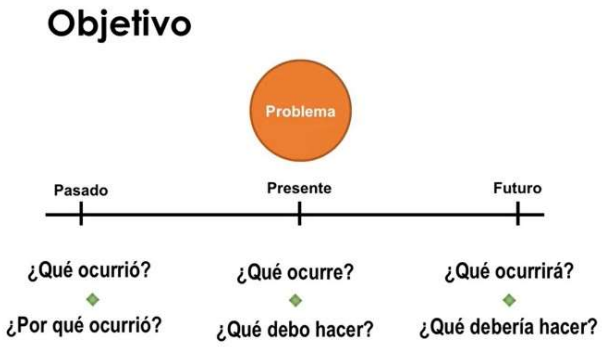
Objetivo

Identificar, analizar y comprender el problema y traducirlo a un problema analítico

Etapas

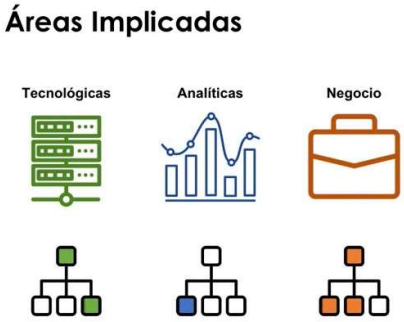
- Identificación del problema
- Fijación de los objetivos
- Identificación de los implicados
- Fijación de la tipología de análisis

Inicialmente necesitamos identificar y obviamente motivar el problema a resolver. ¿Cuál es el problema de negocio? ¿Qué es lo que se pretende resolver? ¿Por qué se tiene que abordar? ¿Qué valor esperamos obtener al resolverlo? Esto es fundamental empezar porque nos va a permitir establecer cuál va a ser el foco en nuestro trabajo y cuál va a ser el objetivo final. Cuando definimos el problema de negocio lo que nos tenemos que plantear es realmente que queremos saber, normalmente nos quedamos a un nivel muy alto. Por ejemplo, quiero saber porque mis clientes abandonan. El abandono es un problema muy habitual, pero tenemos distintos enfoques que podemos hacer desde un punto de vista analítico, queremos saber cosas del pasado, ¿qué ocurrió? ¿Por qué abandonaron? ¿O por qué ocurrió? Queremos saber cosas del presente, ¿qué ocurre? ¿qué debo hacer para evitar abandono? O queremos saber cosas del futuro, ¿qué va a ocurrir? ¿Quién va a abandonar? ¿O qué debería hacer para mitigarlo? Es muy importante que tengamos claro el objetivo porque cada objetivo tiene una técnica analítica distinta y el tratamiento de información es diferente.



Una vez que tenemos claro el objetivo, tenemos que pasar a identificar a las personas que van a estar involucradas, al personal del ámbito de tecnología que tendrán que estar dentro del problema a resolver, habrá personas desde el plano analítico que tendrán que plantear todo el modelo y el tratamiento de información, y después habrá personas del plano de

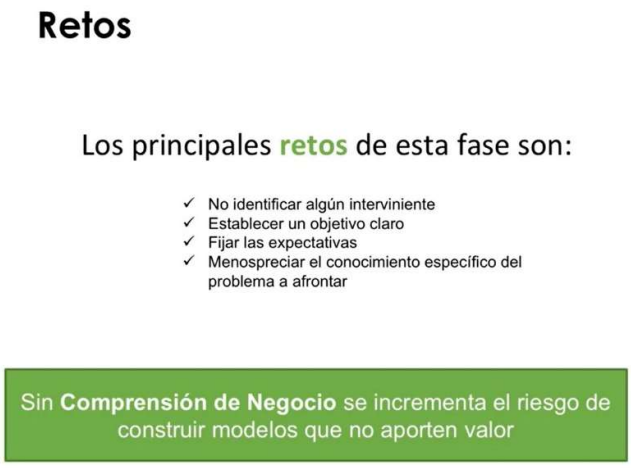
negocio. Tendremos que entender a quién impacta este problema, quién nos puede aportar conocimiento y quién lo va a utilizar finalmente.



Posteriormente tenemos que traducir ese problema de negocio a un problema analítico subyacente. Esto puede ser un análisis descriptivo, vamos a mostrar mediante estadísticas cuál es la realidad que está pasando, podemos hacer un análisis inferencial generalizando conclusiones muestrales a poblaciones completas o incluso haciendo relaciones de correlaciones entre variables o incluso de causa-efecto También podemos hacer análisis predictivos determinando datos futuros en función a la información pasada o podemos hacer análisis prescriptivos que nos permitan elegir la acción óptima a realizar en función del futuro incierto. Es fundamental y clave de toda la metodología del procesamiento de dato comprender el problema de negocio.

Los retos de esta fase son variados, los principales es no identificar algún interviniente inicialmente y que conforme vayamos desarrollando la metodología nos demos cuenta que necesitamos un apoyo o alguien que tenía que haber aportado algo anteriormente o no establecer un objetivo claro, no fijar las expectativas y que algún stakeholder considere que va a recibir más o algo distinto a lo que realmente se está construyendo, o menospreciar el conocimiento de negocio específico en pro del conocimiento analítico. Por lo tanto, antes de

empezar a trabajar con datos y algoritmos es fundamental conocer y comprender el problema de negocio.



- **Comprensión de los datos.**

Una vez que ya tenemos claro el problema de negocio y su objetivo la siguiente fase es la comprensión de datos. En esta fase vamos a identificar las fuentes de información y analizar su conveniencia para su posterior captura y almacenamiento. Las etapas de esta fase son: el inventario de información, la identificación de fuentes, la disponibilidad de las fuentes, la relación de la información y por último la representación funcional de los datos.

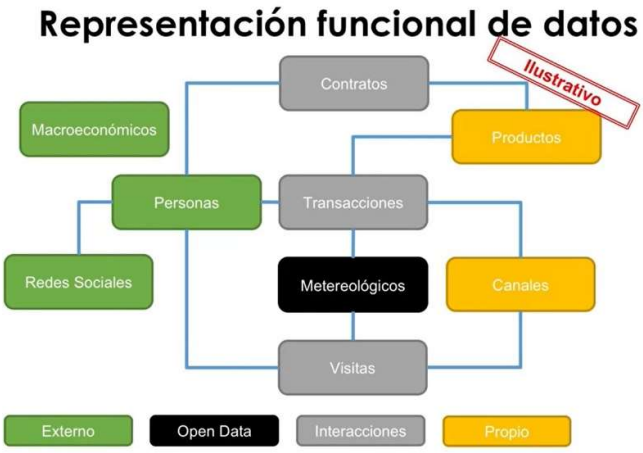
En esta fase todavía no tenemos datos, todavía estamos analizando los datos desde un punto de vista conceptual y el primer paso es hacer inventario de información. ¿cuál sería o cuál es la información que nos gustaría tener? ¿Cuál sería la información adecuada para el problema de negocio? Entonces lo que tenemos que hacer es hacer un listado de la información que sería necesaria tener, para esto lo mejor es considerar con conocimientos

expertos del problema de negocio, gente de negocio que nos puedan ayudar. Una vez que tenemos esa lista de la información que nos gustaría tener lo siguiente es identificar las fuentes de información asociadas. ¿Dónde se encuentra esa información? ¿O dónde podrí estar esa información? ¿Se trata de fuentes internas? ¿Se trata de fuentes externas? Por ejemplo, ¿también podrían ser redes sociales? ¿Podría ser open data? Lo importante es identificar dónde se encuentran esos conceptos de información para poder analizar la dificultad de capturar y almacenar esa información.

Una vez que tengamos identificada la información lo que nos queda es plantearnos un doble check, ¿esa información que queremos identificada la fuente podemos capturarla y almacenarla? ¿O no? Si podemos capturarla y almacenarla o si ya la tenemos capturada y almacenada perfecto, si no podemos actualmente por problemas técnicos, tecnológicos o por alguna otra causa lo que tendremos que hacer es un plan de adquisición de fuentes para que a futuro tengamos disponible toda esa información para enriquecer nuestros análisis. Por lo tanto, lo importante es tener claro en este punto si es o no es accesible la información antes de empezar a trabajar con ella.

Una vez que tengamos la información entonces ya bajamos al dato, ya bajamos a la información que tenemos capturada y almacenada y lo que nos planteamos es, ¿estos datos cómo se generan? Inicialmente los datos son puntos en el espacio, son datos independientes y lo que tenemos que plantearnos es, ¿cómo podemos relacionar todos estos conceptos? Claro si yo pienso en los datos que ya tengo en la red puedo tener redes sociales, puedo hacer transacciones con tarjeta de crédito, puedo enviar un Whatsapp a un amigo, toda esa información que es mía no está tan claro que sea mía porque no hay un identificador único común que me permita integrarlo todo. Entonces es muy importante comenzar a encontrar esos identificadores que nos van a permitir agrupar esa información y ver si somos capaces de tener un identificador único. Entonces nos podemos encontrar que vamos dejando información por nuestro correo electrónico, vamos dejando información de pagos que pagamos en efectivo que pagamos con tarjeta, los mensajes que hacemos con el móvil. El problema fundamental siempre de la relación de información es cómo puedo yo relacionarla y esto es crítico porque si no tendríamos mucha información inconexa, independiente y no tendríamos la visión global. Así que es muy importante hacer un esfuerzo de entender cómo se relacionan los datos o sino crear una regla que nos permita relacionar los datos.

Por último, todos esos datos una vez que los tengamos relacionados tenemos que hacer una relación funcional desde los datos en donde veamos a nivel conceptual como se relacionan. Como por ejemplo en los datos que vemos en el siguiente gráfico, podemos ver que hay personas, donde tenemos datos de redes sociales de esas personas, después hay productos y la relación entre las personas y los productos pueden ser a través de contratos, a través de transacciones, etcétera. Después nos podemos encontrar conceptos de información como los conceptos macroeconómicos que a lo mejor no somos capaces de almacenarlos, pero es importante tener claro y conceptualmente cuál es la información de la que dispongo, cuáles son las fuentes de información y cómo se relacionan.



En esta fase los retos que nos encontramos son los siguientes. Primero, tener claro la identificación de las fuentes de información asociadas al problema de negocio, puede ser que no identifiquemos alguna fuente de información o porque no sepamos que existen o porque no tengamos claro que esa información nos puede aportar valor. Segundo, comprender la información contenida en los datos, una vez que tengamos la información que queremos, la fuente y ahondemos en el dato a lo mejor el nombre del campo no me ayuda y necesito comprender realmente que es esa información y de dónde ha surgido. Después relacionar los conceptos es crítico, si no somos capaz de relacionar los conceptos, si no soy capaz de relacionar quién hace que, dónde y de qué forma, no voy a ser capaz de extraer el conocimiento que hay dentro de esos datos. Y por último muy importante no focalizarse en los datos disponibles, uno de los mayores errores de esta fase es centrarse en qué datos tengo voy a trabajar con los datos que tengo, es importante hacer ese ejercicio previo de analizar qué información me gustaría tener y dónde puedo encontrarla. Por lo tanto, antes de ponerse a trabajar con datos y algoritmos lo importante es pensar desde un punto de vista de negocio cuál es la información que realmente me gustaría tener, si está accesible y disponible y cómo puedo trabajar con ella.

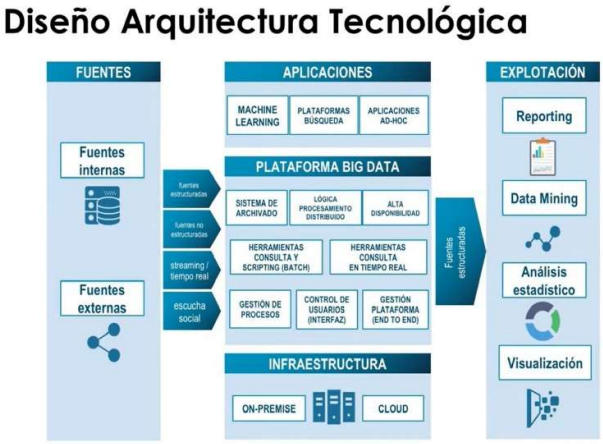
- **Plataforma tecnológica.**

El objetivo de esta fase es disponer de una plataforma tecnológica para la construcción del modelo analítico. Esta fase se divide en tres etapas: primero, el diseño de la arquitectura tecnológica; segundo, la selección de las componentes Big Data adecuadas; y tercero, establecer la estrategia de implantación de esta plataforma tecnológica.

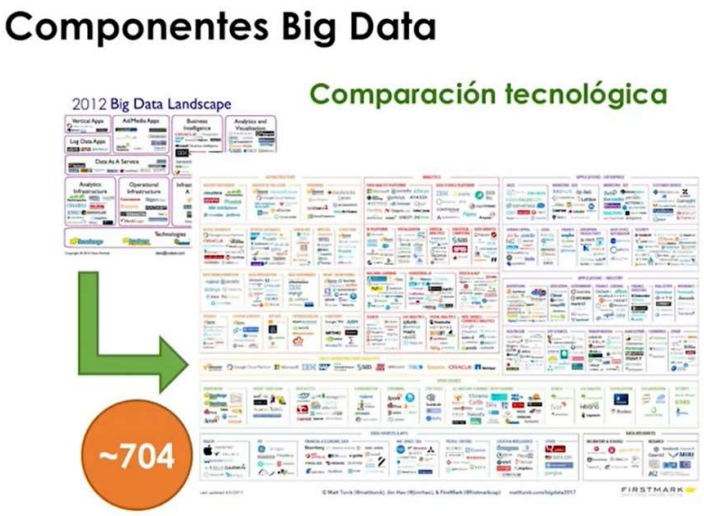
Para empezar, cuando vayamos a hacer el diseño de la arquitectura tecnológica, lo importante es establecer la composición de las componentes estructurales necesarias para soportar las acciones asociadas a la construcción y explotación de modelos analíticos. Esto es, tenemos que pensar en la captura de información, tenemos que pensar en el almacenamiento de datos, tenemos que pensar en el procesamiento de los datos y tenemos que pensar en la explotación del modelo. Tenemos que tener toda la visión global y cómo se relacionan las componentes.

Por lo tanto, el diseño tiene que contar con una parte de entrada de información, de captura de información de las fuentes internas o externas, tenemos que contar con bases de datos donde se aloje toda esa información, tenemos que contar con motores de procesamiento, y todo eso estará soportado sobre una infraestructura que puede ser On-premise, Cloud, etcétera. Además, tendremos que tener componentes para la explotación de esos modelos analíticos, estadísticos o de "machine learning" que queramos hacer, y, por supuesto

herramientas o aplicaciones de visualización. Entonces, es muy importante que tengamos claro que tenemos que cubrir todas esas necesidades.



Ahora, ¿con qué?, ¿con qué componentes Big Data? En el 2012, el número de componentes Big Data que había importantes, más o menos, eran unas 84, que se dividían entre herramientas de visualización, distintas tecnologías, bases de datos estructuradas, etcétera. Actualmente, tenemos más de 704 componentes y hay un crecimiento exponencial. Realmente, es muy difícil conocer todas las componentes que hay actualmente en el mundo de Big Data y, sobre todo, cómo se relacionan entre sí. Por eso, es fundamental que haya un perfil específico, dedicado a comprender las componentes tecnológicas.



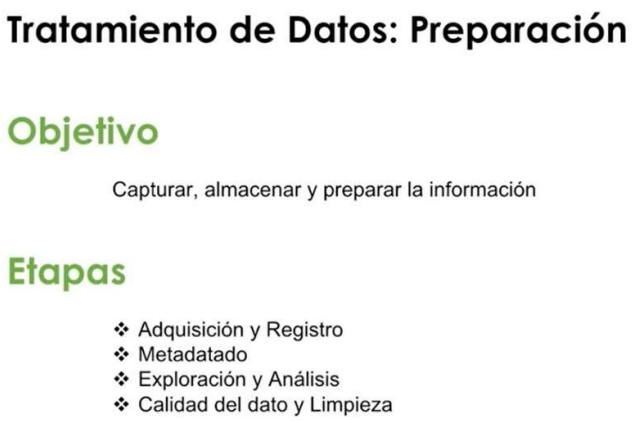
Además, es importante establecer la estrategia de implantación. ¿Qué es lo que queremos? Queremos tener todo el control de toda la infraestructura, de todas las aplicaciones, de todos los datos al lado nuestro, o queremos utilizar soluciones Cloud que nos permitan externalizar y que un proveedor nos puede dar prácticamente todos los servicios. Entonces, tenemos un amplio rango de posibilidades, desde On-premise hasta SaaS ("software as a service"),

pasando por infraestructura como servicio o plataforma como servicio, para que nosotros podamos elegir cuál es la estrategia implantación que más se adecúa a nuestros datos, a nuestros procedimientos y al modelo que vamos a construir.

Los retos en esta fase son los siguientes. Para empezar, considerar todas las implicaciones de la arquitectura diseñada, es decir, esa arquitectura tiene que dar servicio a toda la creación del modelo y a toda la puesta en producción posterior y la explotación, por lo tanto, hay que pensar todos los detalles. Segundo, hay que estar muy al día de toda la evolución de todas las componentes Big Data, porque realmente el crecimiento está siendo exponencial, y es muy importante conocer las distintas versiones de las tecnologías para ver que son compatibles. Después, hay que dimensionar de forma adecuada los recursos, para que se pueda procesar en el tiempo adecuado y tenga suficiente potencia para que los algoritmos analíticos funcionen. Por último, es muy importante, desde el principio, tener clara cuál es la estrategia adecuada de implantación. Por tanto, es vital que pensemos que la solución que queramos construir, tenemos que construirla sobre una plataforma tecnológica y ésta tiene que estar diseñada. Por tanto, es muy importante dedicarle un buen esfuerzo.

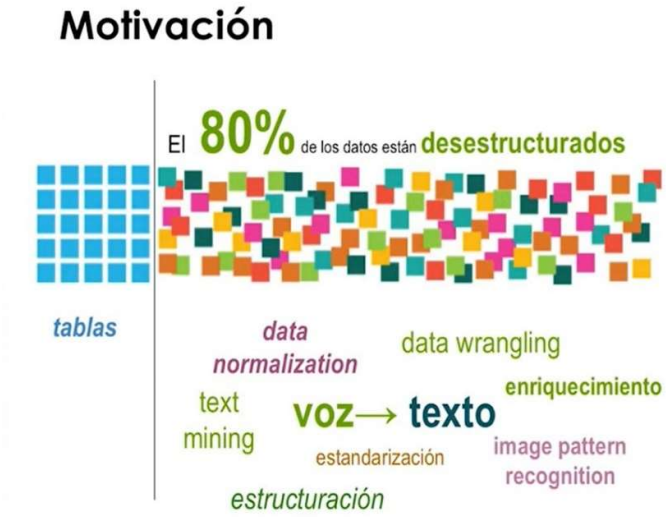
- **Tratamiento de datos: Preparación**

En este momento de la metodología ya tenemos el problema de negocio, ya comprendemos los datos y ya tenemos una plataforma tecnológica preparada para trabajar. El objetivo ahora es capturar, almacenar y preparar la información para que nos sirva para el modelado de datos. Las etapas que vamos a ver son las siguientes, si tiene registro, metadato, exploración y análisis y calidad del dato y limpieza. Esta fase es una de las más críticas porque es la que requiere más trabajo. Normalmente se estima que el 80% del tiempo que emplea un científico de datos trabajando con información es realizando el tratamiento de datos. Así que es muy importante que tengamos claro las distintas etapas.

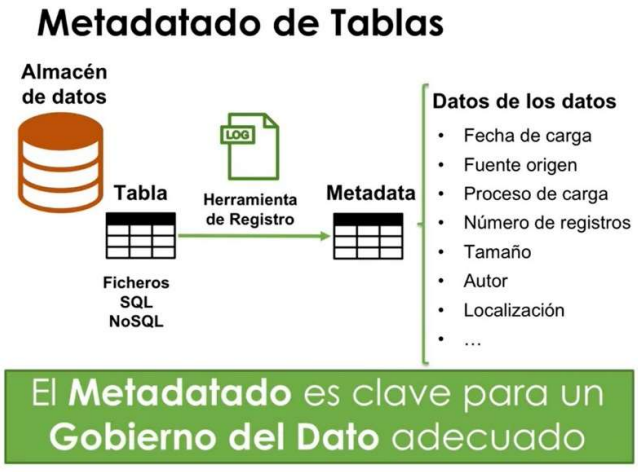


Comenzamos indicando el por qué, ¿por qué se dedica tanto tiempo al tratamiento de información? La información estructurada está en formato de tablas y es la que normalmente solemos trabajar, pero actualmente el 80% de la información viene desestructurada como hemos comentado, puede ser vídeo, puede ser texto, puede ser voz y esa información además no está relacionada y normalmente la calidad no suele ser máxima por lo que requerimos un trabajo y un esfuerzo muy grande en ordenar toda esa información, darle un formato adecuado para que esté lista para los algoritmos para conseguir modelos. La primera etapa es la de adquisición y registro. Una vez que tenemos identificada las distintas fuentes de información, fuentes internas, fuentes externas, open data, lo que nos tenemos que plantear es cómo vamos a capturar esa información. Hay muchas herramientas en función de la naturaleza de los datos que son estructurados o no estructurados y en función de cómo queramos hacer esa planificación de captura. Podemos hacer un proceso Batch donde vayamos capturando todos los datos en un momento determinado del día o de la semana o podemos hacer un proceso en real time, near real time, que cada poco tiempo vaya ingestando la información o directamente como se genere. En función de esa planificación y esa naturaleza tenemos que utilizar una o varias herramientas de ingesta o APIs, conectores, por scrapping, etcétera y además tendremos que incluir una herramienta

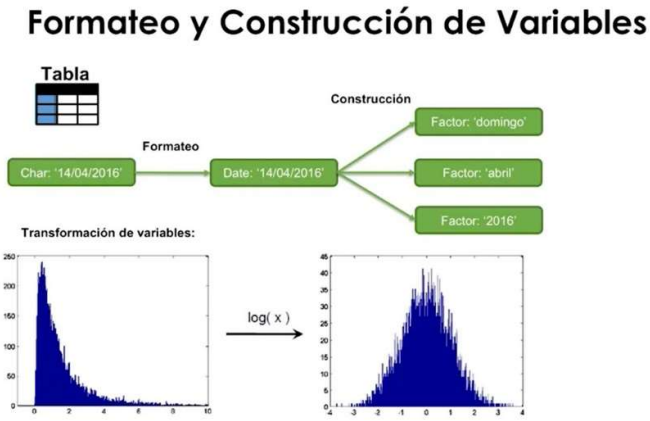
de registro. Es importante que quede siempre claro la información, cuándo se ha capturado, donde se ha capturado, cómo se ha capturado, qué se ha capturado, antes de incluirla en nuestro almacén de datos.



La siguiente etapa es el metadato de los datos. Una vez que tenemos la información en el almacén, toda esa información que tenemos sobre esos datos, cuál es la fuente de información, cuando se ha procesado, donde se guarda, todo eso es el metadato. El metadato son los datos de los datos, que hace referencia a la fecha de carga, a la fuente de origen del proceso de carga, al tamaño, al autor, a la organización y toda aquella información que queremos o queramos que esté asociada a la tabla. Esto es muy importante y es fundamental para tener un gobierno del dato adecuado. Para eso pues necesitamos una herramienta de registro que nos permita guardar toda esta información y asociarla a cada una de las tablas, ficheros o base de datos que tengamos para que esto si se tiene que consultar en el futuro, esté disponible.

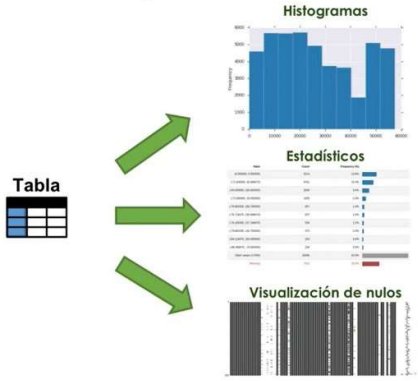


La siguiente etapa es el formateo y construcción de variables. Una vez que la información ya la tenemos alojada en nuestro almacén de datos, lo primero que hay que plantearse es, ¿tiene el formato adecuado? Es decir, ¿los números están guardados como números? Cuando tenemos algo guardado como número, ¿realmente es un número? Por ejemplo, si yo tengo el número de hijos, 1, 2, 3, 4, podemos tener claro que es un número, pero si yo estoy pensando en los códigos postales o en las provincias y que son las provincias y veo el número 30, a lo mejor no hace referencia al número que se tenga que sumar sino realmente hace referencia a una provincia. Entonces es importante que tengamos un formato. Esto es especialmente crítico con las fechas. Normalmente las fechas cada una puede tener un formato distinto, hay que trabajar muy bien para que la herramienta con la que estemos trabajando o el software, interprete estas fechas en forma adecuada. También en esta fase se hace construcción de variables sencillas. Por ejemplo, podemos crear, a partir de una fecha, que día de la semana es o en qué mes estamos o cual es el año e incluso podemos hacer transformaciones más complejas utilizando funciones matemáticas.



Posteriormente lo que se hace es una exploración y análisis de variables. Para cada una de las variables y las tablas podemos hacer una serie de agregaciones, una serie de resúmenes utilizando gráficos o estadísticos, visualizando la información que tenemos. ¿Para qué? Para poder comprender lo que realmente hay en los datos, intentar identificar problemas que puedan surgir, para los nulos, atípicos, etcétera.

Exploración y Análisis de Variables



El siguiente paso una vez identificadas las cosas que puedan tener los datos es proceder a su tratamiento y limpieza, esta es la fase de calidad del dato y limpieza. Tenemos que identificar todos esos valores vacíos, esos valores incoherentes, los valores atípicos, los outliers, los errores que puedan tener los datos y ver cómo los vamos a tratar, ¿qué queremos hacer? Aquí hay que decidir qué vamos a hacer con cada uno de esos problemas que surgen. Cuando encontremos campos vacíos, los vamos a imputar, vamos a borrar los registros, ¿qué vamos a hacer? Cuando tengamos valores incoherentes, ¿los vamos a poner en cuarentena? ¿Vamos a crear variables adicionales para controlar estos valores incoherentes? y así con todas las problemáticas que vayan surgiendo. Es muy importante tener claro las decisiones que se tomen aquí, porque a la hora de poner el modelo en producción tendremos que tener en cuenta todas las decisiones tomadas en este punto.

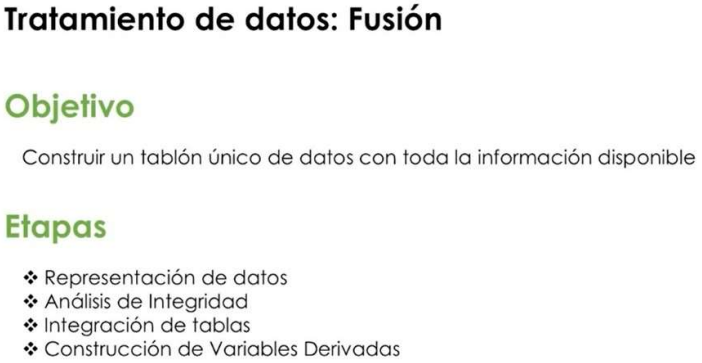
Calidad del dato y Limpieza

Los retos de esta fase son primero, evaluar la calidad de los datos. Sin ninguna duda es uno de los grandes problemas que tenemos siempre, ¿cuál es la calidad del dato que tengo? Por un lado, hablamos de los nulos, hablamos de problemas de errores atípicos, pero también podemos hablar de, si la calidad. Es importante tener claro que podemos tener incluso problemas de calidad de información. Después tenemos que plantearnos cómo vamos a tratar toda esa información estructurada. No solo en mismos procesos. Si se trata de vídeos

o si se trata de textos o si se trata de imágenes. También es importante fijar los criterios de tratamiento y tienen que ser homogéneos a lo largo de toda la metodología. Si el tratamiento de los vacíos se hace de una determinada forma tenemos que hacerlo con todos los vacíos de la misma manera o es conveniente hacerlo de la misma manera y después lo tenemos que tener también en cuenta cuando vayamos a poner el modelo en producción en la fase de despliegue. Y también tenemos que crear una política de metadato que además tendrá que ser compatible a la política de metadato que tenga toda la compañía, para que toda la información, toda la fuente de información estén igualmente identificada. Todos los procesos estén claros y organizados.

- **Tratamiento de datos: Fusión.**

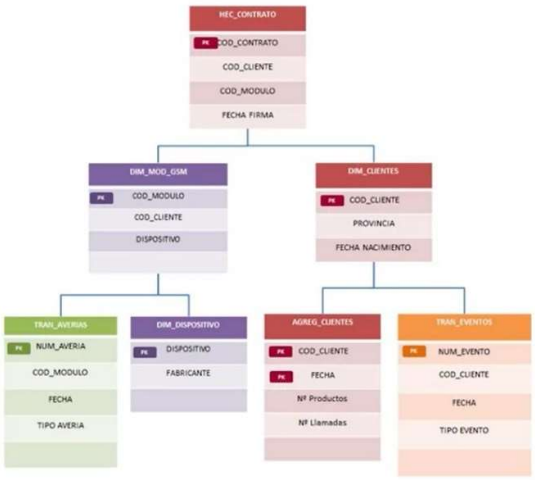
El objetivo es construir un tablón único de datos con toda la información disponible que esté preparado para el modelado. Las etapas de esta parte del tratamiento de datos es uno, la representación de los datos, dos, el análisis de integridad, tres, la integración de las tablas y por último construcción de variables derivadas.



Empezamos con la representación de datos. En la comprensión de datos estuvimos comentando que era importante relacionar los conceptos, una vez que tenemos los

conceptos relacionados, ya tenemos la fuente, ya tenemos los datos lo importante es bajar al nivel adecuado para entender realmente cómo se relacionan los datos. Es decir, cuáles son los campos que se basan unos con otros y para eso es importante llegar a tabla a tabla e identificar cuáles son los campos y cómo se relacionan entre sí. Para tener un esquema de la base de datos con la que estamos trabajando que integre todas las fuentes.

Representación de los datos



Esquema de Bases de Datos

Lo siguiente es hacer el análisis de integridad. Para cada una de las tablas puede tener claves primarias y claves foráneas. Las claves primarias son aquellos datos que son únicos que representan el registro, por ejemplo, en el siguiente grafico tenemos tres tablas, una de alumnos, otra de cursos y otra de matrícula, entonces los alumnos tienen un identificador único que es el código de alumno. Ese código tiene que ser único, no puede haber dos alumnos con el mismo código, sin embargo, el nombre sí puede hacer referencia a dos personas que se llamen igual o la edad puede ser igual, así que no tiene que ser una clave primaria, solamente el código de alumno es una clave primaria. Sin embargo, después cuando relacionamos conceptos podemos crear las matrículas que tienen su propio código, cada matrícula tiene que ser única, pero sí que es cierto que un alumno y un curso, una combinación de alumno y curso sí que pueden estar matriculados en distintas asignaturas y de distintas formas. Por eso es muy importante tener claro cuáles son las claves primarias y cuáles son las claves foráneas. Las foráneas son aquellas claves primarias de otras tablas que se encuentran en una tercera tabla. Para analizar la integridad lo que tenemos que ver es primero la integridad de entidad. ¿Alguna clave primaria tiene algún problema en cuanto, puede ser que no aparezca en algún registro o que esté duplicada?, puede ser que alguna clave foránea no exista. Por ejemplo, en este caso nos podemos encontrar con alumnos sin código de alumno, eso sería un problema de integridad de entidad. Nos podríamos encontrar

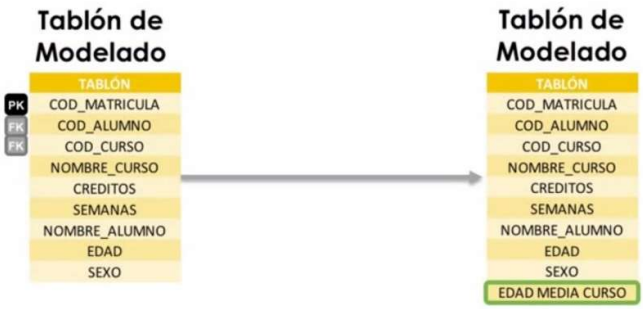
una matrícula de un alumno de un curso que cuando vayamos a buscar los datos del alumno no exista, eso sería un problema de integridad referencial. Entonces es muy importante establecer cuáles son las reglas que vamos a seguir para integrar la información. Una vez que hemos hecho ese análisis, toca el momento integrar las tablas.

Entonces lo que queremos es juntar todos los conceptos en una única tabla que esté preparada para el modelado, y es donde entran en juego las reglas de integración, por ejemplo, qué hacemos con aquellas claves foráneas que no existen. Si yo tengo una matrícula de un alumno en un curso y no existe el alumno, ¿cómo corrijo ese problema? esas reglas de integración tenemos que definirlas muy bien porque van a influir después de la parte de modelado y sobre todo hay que tenerlas muy en cuenta en la parte de despliegue del modelo. Una vez que tengamos claras las reglas e integremos las tablas, nos vamos a encontrar con el tablón de modelado. El tablón de modelado es una única tabla que tiene todos los datos necesarios preparados para poder modelar, de forma que tenemos toda la información concentrada en un único punto.

Integración de tablas

Por último, una vez que tenemos el tablón de modelado, lo que podemos hacer es crear nuevas variables que incluyan conceptos de las diferentes tablas independientes que había entre sí. Por ejemplo, en este caso podríamos crear una variable que se llamase edad media del curso, estamos mezclando un concepto de los alumnos con un concepto de los cursos, de forma que lo podemos crear una vez que tengamos este tablón de modelado y no antes.

Construcción de Variables Derivadas

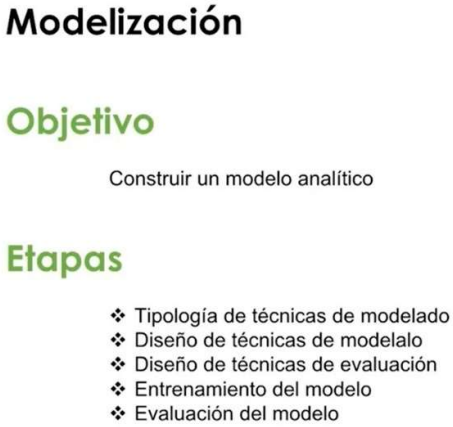


Los retos de esta fase son los siguientes, lo primero hay que diseñar el modelo de datos con el que queremos trabajar. Si ya está construido hay que comprenderlo, pero si no está construido, ese diseño nos va a ayudar a enfocar mejor el problema. Lo segundo es ver cómo evolucionamos este modelo de datos, claro en el primer problema analítico que hagamos tendremos un modelo de datos, pero conforme queramos crear más modelos y queramos introducir más información, este modelo de datos va a ir creciendo en complejidad y hay que tener muy claro cuál es la estructura, el orden y sobre todo el gobierno de este modelo de datos. Y por último la gestión de las incidencias en la integración de tablas, porque cuando hagamos un primer análisis con los datos que tengamos encontraremos unos

problemas, pero tenemos que definir unas reglas, ¿por qué? Porque conforme pase el tiempo tendremos nuevos alumnos, nuevos cursos, nuevas matrículas y tendremos que ir aplicando todas esas reglas de forma continua en el tiempo, por tanto, tienen que estar muy claras y muy especificadas. Una vez que tengamos el tablón de modelado, ya estamos listo para pasar a la siguiente fase, la fase de modelado.

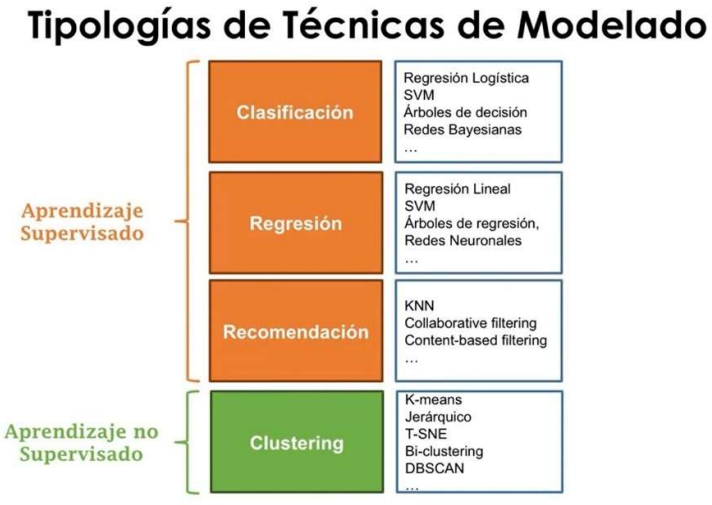
- **Modelización.**

El objetivo de esta fase es construir un modelo analítico a partir del tablón de datos que ya tenemos generado. Vamos a ver una introducción a todas las etapas de esta fase de modelización. Las etapas son las siguientes. Lo primero, conocer la tipología técnica de modelado. Lo segundo, es el diseño de esa técnica de modelado. Lo siguiente es diseñar la evaluación que vamos a utilizar seguido del entrenamiento de construcción de modelo para acabar con la evolución del modelo.



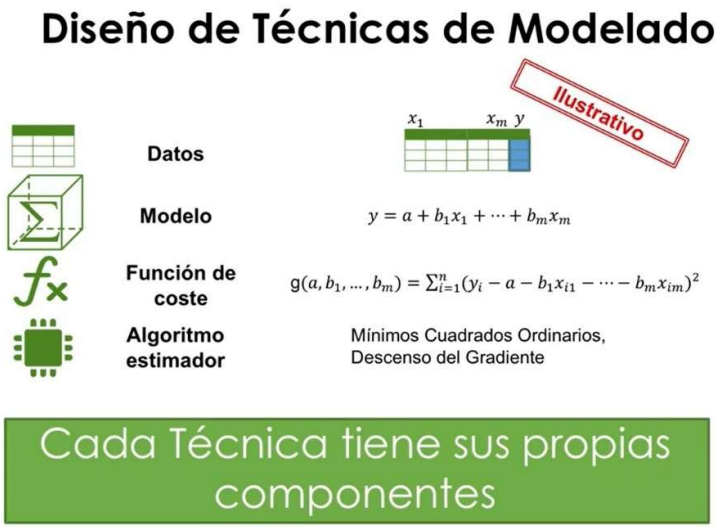
Inicialmente tenemos que elegir cuál es la técnica o familia de técnicas que vamos a utilizar para construir el modelo. Para eso tenemos que tener claro si se trata de un problema de aprendizaje supervisado o un problema de aprendizaje no supervisado. Los problemas de

aprendizaje no supervisados son aquellos donde nosotros tenemos muchísimos datos y le pedimos al algoritmo que encuentre o identifique patrones sin que le demos ninguna pista. Los problemas de aprendizaje supervisado son aquellos en los que sí vamos a tener esas pistas porque tenemos ocurrencias pasadas que nos van a permitir decirle al algoritmo qué es lo que debe o no aprender. Por ejemplo, nosotros podemos querer clasificar a todos los alumnos de un curso, podemos decirle, “estos son todos los alumnos, ¡clasifícalos!”, sin decirle nada más. Eso ha sido un aprendizaje no supervisado y lo que hará será hacer grupos de personas parecidas. Que es lo que se llaman clusters, mediante técnicas de clustering y las más habituales son K-means, el clustering jerárquico, el T-SNE, el Bi-clustering, el DBSCAN, etcétera. Sin embargo, si le decimos, queremos saber cuál es la prioridad que un alumno apruebe el curso o no, si yo tengo datos pasados de alumnos y además también si aprobaron o no aprobaron el curso, yo puedo introducir esa pista y decirle, lo que quiero es que busques patrones enfocados a aprobar o no el curso. Entonces, esto es aprendizaje supervisado y normalmente se divide en tres, a groso modo. Clasificación, cuando le pedimos que clasifique en base a dos o más categorías, por ejemplo, aprobar o no aprobar, o podemos tener tres, si es alto, medio o bajo el resultado de algo, o más categorías. Después puede ser regresión cuando lo que queremos es que nos estime una cantidad. Por ejemplo, podríamos querer estimar cuál es la nota que va a sacar un alumno o cuál es el precio de una vivienda, etcétera, entonces es un problema de regresión. Y, por último, problema de recomendación si lo que queremos es que nos cree recomendaciones personalizadas o bien de contenidos o de cualquier otro tipo de recomendación.



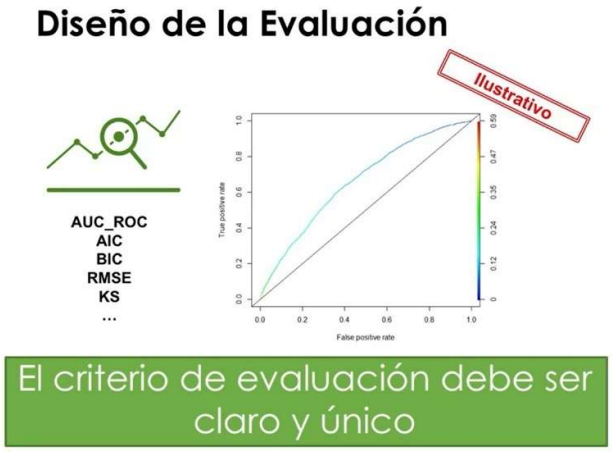
Ahora tenemos que diseñar cómo vamos a realizar el modelado. Para eso partimos del tablón de datos que, si es una clasificación clasificada, además tendremos una columna que

será el target, que está representada en color azul y tendremos un modelo de una familia de modelos que lo hemos representado a través de una fórmula matemática. El objetivo es que ese modelo tenemos que calcular los parámetros que hacen que sea óptimo y que se adecúe a los datos y que nos aporte conocimiento. ¿Esto cómo se hace? Se hace a través de algoritmos estimadores que son normalmente algoritmos recursivos que tratan de encontrar cuáles son los mejores valores de los parámetros, los valores que hacen que el modelo sea el mejor. ¿Pero qué es esto que el modelo sea el mejor? ¿El mejor con respecto a qué? El mejor con respecto a una función de costes que decidamos. El objetivo es encontrar los mejores parámetros para función de costes. Por ejemplo, si yo considero que mi modelo es una recta y = a + b x, yo lo que quiero encontrar es cuáles son los valores de a y de b que mejor se adecúan a los datos para relacionar esas variables x e y. Para eso yo tengo que identificar una función de costes. ¿Cuál es la función de costes?, puede ser que quiero minimizar el error de la estimación, puede ser que quiero maximizar los beneficios, también podrían ser funciones de negocio, puede ser cualquier función que se nos ocurra con función de costes. Al final lo que hacemos es que el algoritmo busque cuáles son los valores de a y de b que hacen esa función de costes sea mínima o máxima en función de lo que estemos definiendo. Pero es muy importante que establezcamos cuál es el criterio porque con distintos criterios llegaremos a distintos modelos. Y hay que buscar siempre el criterio adecuado para el problema que queramos resolver.



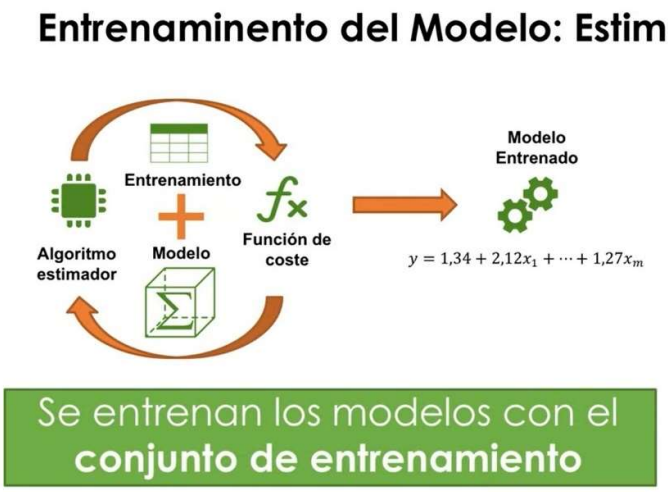
Después tenemos que diseñar la evaluación. Estos modelos que hemos construido con esta técnica de modelado, ¿cómo los vamos a evaluar? Métricas de evaluación hay muchísimas. Hay veces que se utilizan varias métricas de evaluación para ver si hay algún modelo que puede ser mejor que los demás en todas o en la mayoría de las métricas de evaluación. Pero lo importante es definir claramente el criterio al principio, ¿para qué? Para que siempre se utilice la misma metodología y se elija el modelo en base a esa métrica de evaluación. Es

muy importante no elegir cualquier métrica porque las métricas se suelen adecuar a la técnica de modelo y al problema a resolver, por lo tanto, hay que hacer un ejercicio previo de analizar y elegir las métricas adecuadas. Pero lo importante es que el criterio sea claro y único para que tengamos siempre una forma de elegir el modelo finalista.



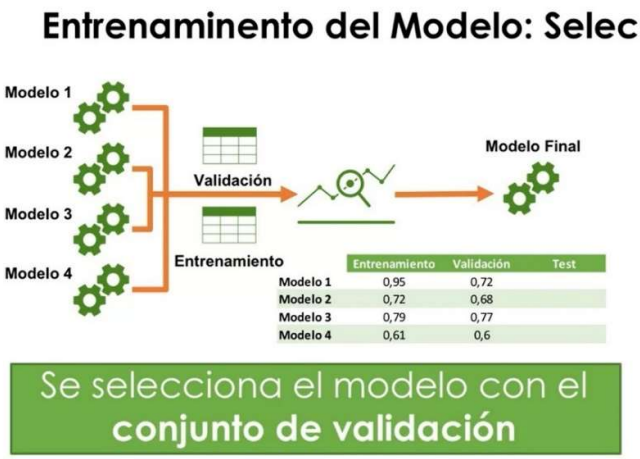
Una vez que tenemos ya el diseño de cómo vamos a modelar y tenemos diseñado cómo vamos a evaluar, empezamos con el entrenamiento. Para eso hay que tener una metodología de división de los datos para poder realizar esta metodología. Por ejemplo, podemos dividir el conjunto en tres, conjunto de entrenamiento, validación y test. El conjunto de entrenamiento suele ser el más grande y es el que se va a utilizar para construir los modelos. El conjunto de validación es el que se va a utilizar para elegir cuál es el modelo finalista o también para si hay que hacer algún tipo de cálculo de estimación de hiper parámetros o fine tuning, lo haremos con el conjunto de validación. Y por último reservaremos un conjunto de datos de test para realmente evaluar la capacidad del modelo. Este conjunto de test no se va a utilizar nunca a la hora de construir el modelo. Esta metodología lo que asegura es conocer la capacidad predictiva real del modelo y evitar problemas que pueden surgir como el overfitting.

Una vez que tenemos la división de los conjuntos, pasamos a estimar esos parámetros, para eso tenemos por un lado el modelo, que hemos dicho que es esa fórmula donde no conocemos los parámetros y los datos, cogemos los datos de entrenamiento. Y utilizando el algoritmo estimador y la función de costes, de forma recursiva, vamos calculando cuáles son los mejores parámetros para obtener ese modelo. Una vez que tengamos los mejores parámetros, tendremos un modelo entrenado que será una fórmula donde ya los parámetros tendrán un valor concreto. Lo importante es destacar que este entrenamiento de este modelo, de todos los modelos que elijamos, lo haremos utilizando el conjunto de entrenamiento, ¿para qué? Para cuando los evaluemos, los evaluemos en un conjunto que no se haya utilizado para entrenar. Es muy importante que el examen que le hagamos al modelo sea con información diferente a la que he utilizado para aprender. Así evitaremos que, en lugar de aprender, memorice.



Los siguiente es seleccionar. Hemos entrenado varios modelos con distintos algoritmos y distintas funciones de coste. ¿Cómo podemos ahora elegir cuál es el modelo mejor o el modelo final? Bueno, podemos utilizar esa métrica de evaluación que hemos elegido y vamos a calcularla tanto en el conjunto de entrenamiento como en el conjunto de validación. Entonces una vez que tengamos todos los datos, lo que buscaremos es aquel modelo que tiene un valor mayor en el conjunto de validación. Después podremos ver que hay algunos casos en los que se utilizan otras técnicas, pero en principio, en general lo que se quiere encontrar es el modelo que mejor actúa en unos datos que no he utilizado de entrenamiento y eso nos permite conocer si los modelos son robustos o no. Además, en este momento si vemos que las familias de modelos que hemos utilizado no son adecuadas o no aprenden, podremos cambiar las técnicas y volver a empezar a entrenar modelos y volver a pasar por la validación. Esto es un proceso cíclico en donde hasta que no tengamos un modelo que pensemos o que consideremos que es el modelo finalista, no terminaremos nunca. Pero

importante, la decisión se toma con los resultados en el conjunto de validación que no se utilizaron en el entrenamiento.

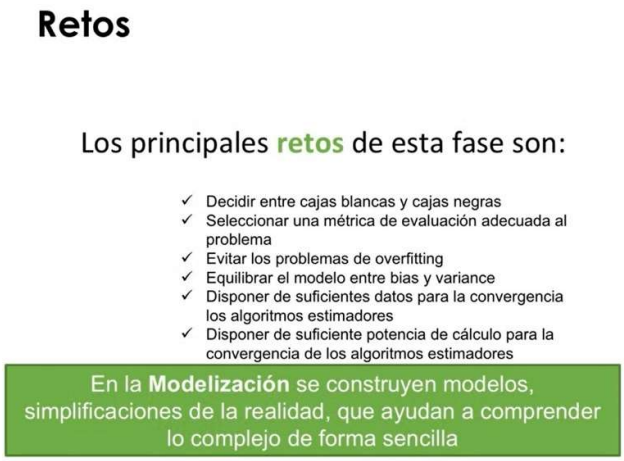


Por último, tenemos la evaluación final del modelo, es decir, ¿cuál es la capacidad real de un modelo? Cuando para esto se utiliza ese conjunto de test que hemos obtenido del conjunto inicial de forma aleatoria extrayendo una muestra aleatoria y lo que hacemos es evaluar el modelo en ese conjunto de test que no se ha utilizado ni en el entrenamiento ni en la validación. Porque esto nos hace referencia y nos demuestra cómo funcionaría un modelo cuando lo utilicemos con datos nuevos que nunca ha utilizado. Entonces, este valor es el que nos medirá la capacidad real del modelo. Realmente la fase de modelado es la fase más crítica. Aunque el objetivo es obtener un modelo realmente en esta fase también el objetivo es conocer mejor el problema, identificar cuáles son las variables que están relacionadas, entender cómo se relacionan entre sí y cómo se hacen la proyección para también hacer propuestas de construcción de nuevas variables, tratamiento diferente de variables que habíamos hecho, nuevas técnicas de filtrado, etcétera, de forma que es probable que una vez que lleguemos a la parte de modelado tengamos que volver a empezar otra vez por la comprensión de negocio, porque hemos detectado en la fase de modelización que hemos encontrado nuevo conocimiento, nueva información que sería accesible con un tratamiento distinto. Por eso es muy importante que no solo nos focalicemos en la construcción del modelo, que hay que construirlo, sino también en ir afinando las preguntas, ¿hemos hecho las preguntas correctas? ¿Hemos creado el tablón de forma correcta? ¿El target que hemos puesto realmente representa el problema de negocio que queremos? ¿Los resultados que obtenemos los contrastamos con negocio para saber que tienen sentido y que son coherentes? Bueno, es muy importante, así que tiene que quedar claro que el modelo no es el objetivo, es el vehículo para aprender el conocimiento y aprender la experiencia que tienen los datos.

Evaluación del Modelo

La capacidad analítica se calcula en el **conjunto de test**

En esta fase hay muchos retos que asumir, el primero es decidir entre cajas blancas y cajas negras. Las cajas blancas son aquellos algoritmos que son interpretables, que tienen una fórmula sencilla que podemos comprender porque obtenemos la proyección que tenemos. Mientras que las cajas negras son algoritmos más complejos, normalmente que predicen mejor pero que es muy difícil interpretarlos. Según quién sea el interlocutor y cómo se vaya a utilizar pues tendremos que decidir si queremos elegir caja blanca, más interpretable, o caja negra, más precisa menos interpretable. Ese equilibrio hay que buscarlo. Además, tenemos que intentar buscar una métrica de evaluación adecuada al problema, aunque las métricas de evaluación están muy enfocadas a la parte analítica, sí que es cierto que podemos utilizar métricas más enfocadas a la parte de negocio. Tenemos que analizar problemas que pueden surgir como el problema del overfitting de la multicolinealidad y buscar el equilibrio entre el sesgo y la varianza que es muy difícil buscar el punto intermedio. También es importante tener claro si disponemos de suficientes datos para la convergencia de los algoritmos más complejos. Puede ser que necesitemos más datos para que ciertos algoritmos puedan converger. Eso es importante para verlo lo antes posible. Y también por supuesto disponer de la suficiente potencia adecuada. Claro lo que no podemos hacer es considerar algoritmos muy complejos con muchísimos datos y no tener la potencia adecuada y tener que esperar, cinco, seis, siente hora para obtener el resultado de un modelo. ya que la tecnología lo permite y podemos escalar, lo importante es contar con la tecnología adecuada. Esta fase de modelado es una de las fases más críticas, es una de las fases más difíciles donde realmente se tienen que tomar muchísimas decisiones, y todas esas decisiones tienen que estar totalmente documentadas y totalmente validadas para que después a la hora de poner el modelo en funcionamiento, las tengamos todas claras.



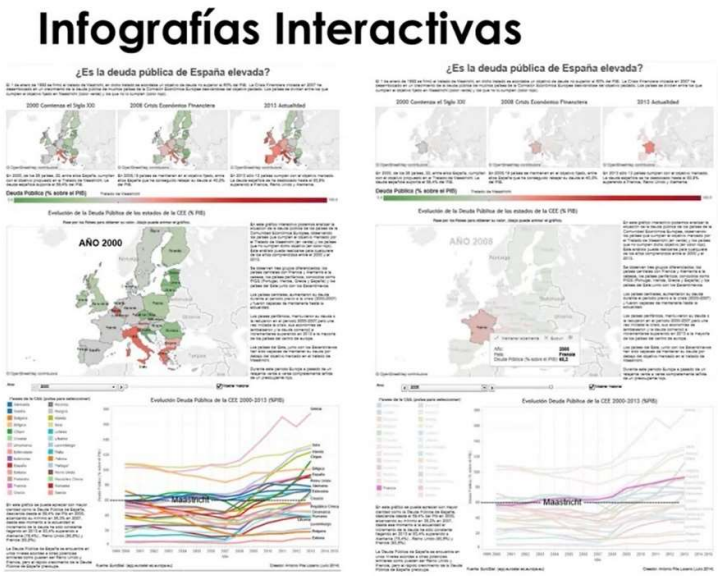
- **Presentación de resultados.**

En esta fase el objetivo es trasladar el conocimiento al resto de intervinientes implicados. Entonces es muy importante tener claro que no todos los intervinientes tendrán el conocimiento adecuado para comprender todo lo que se ha hecho hasta ahora, por eso es muy importante que elijamos cuál es la modalidad de presentación de resultados, donde predominan principalmente las modalidades gráficas. Vamos a ver cuatro modalidades, informes y reportes, visualizaciones, infografías y cuadros de mando. Pero hay muchísimas formas de presentar resultados. Lo importante es trasladar el conocimiento a los intervinientes.

Entonces podemos empezar haciendo informes y reportes donde haya un complemento entre texto escrito, estadísticos y gráficos de forma que se muestre todo el conocimiento que se ha extraído del modelo, cómo se ha construido el modelo, qué información aporta el modelo, pero sobre todo es muy importante que nos focalicemos en la parte de qué conocimiento se ha extraído del modelo, para qué nos sirve el modelo, cómo vamos a utilizar

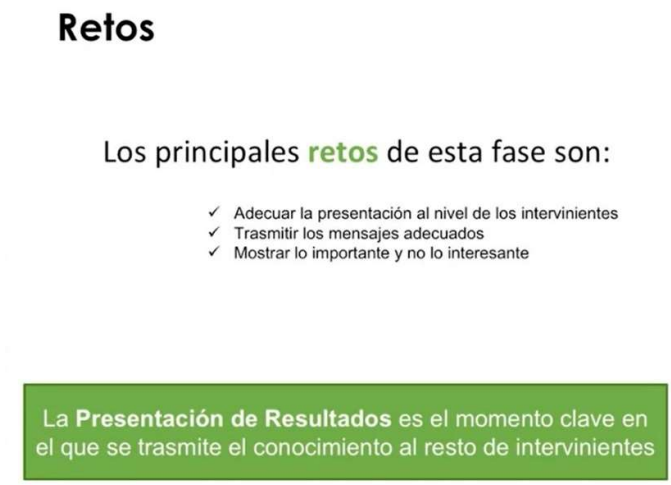
el modelo, qué valor aporta el modelo, eso es lo que realmente tendríamos que mostrar en ese informe reporte.

Lo siguiente que podemos utilizar son visualizaciones, siempre se ha dicho que una imagen vale más que mil palabras por eso es muy importante que trabajemos muy bien las visualizaciones para que muestren la información que realmente queremos mostrar. Hay muchísimas herramientas y muchísimas técnicas de visualización. También podemos utilizar infografías interactivas donde tenemos visualizaciones en donde el usuario puede navegar, profundizar, analizar, porque le va a servir para poder hacer sus propias preguntas y obtener respuestas con la información o el conocimiento que ponemos disposición de los usuarios.



Por último, podemos utilizar cuadros de mando en donde se vaya haciendo el seguimiento de distintos KPI y distintas métricas que puede ser en tiempo real o casi en tiempo real o en procesos batch de forma que podamos tener una foto de evolución de esas métricas y esos valores que nos ayuden a seguir ese conocimiento.

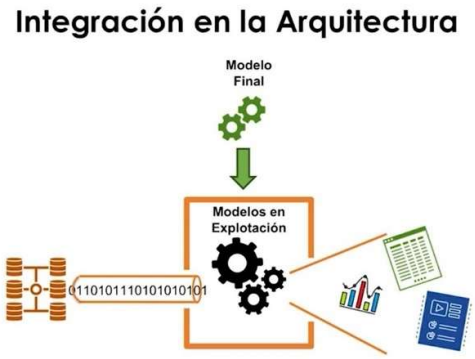
Los retos de esta fase principales son los siguientes, lo primero es adecuar la presentación al nivel de los intervinientes. No sirve de nada hablar a unos intervinientes de aspectos técnicos si nos son capaces de entender esos aspectos. Por eso es muy importante tener claro cuál es el interlocutor y adecuar los mensajes. Esto es fundamental para no perder la oportunidad de trasladar todo ese conocimiento y todo el trabajo realizado al resto de intervinientes en el proyecto, en el proceso o en la investigación. Además, hay que elegir los mensajes adecuados, estamos en el mundo de Big data donde tenemos muchísima información y podemos hacer muchísimos análisis, de forma que podemos llegar a muchísimas conclusiones. Lo importante es cuáles son los mensajes que queremos trasladar, simplificar toda esa información en un recurso visual, en un entregable sencillo de forma que el mensaje quede claro, y por último y muy importante es focalizarse en lo importante y no en lo interesante. Realmente hay que volver hacia atrás a la conclusión de negocio y plantearnos: qué es lo que queríamos resolver, cuál era el problema de negocio, ¿lo estamos resolviendo con esta presentación? ¿Estamos dando las pautas, el conocimiento y la información adecuada? Esta fase es fundamental porque es criticada porque muchos de los intervinientes no habrán conocido nada del proyecto hasta que lleguemos a esta fase y que se le traslade el conocimiento adecuado es el éxito del problema analítico.



- **Despliegue.**

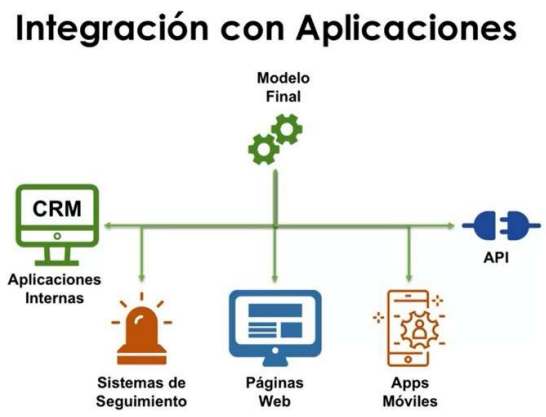
El objetivo es desplegar en la plataforma tecnológica de explotación el modelo construido. Las etapas son tres. Primero integración en la plataforma, en la arquitectura. Segundo planificación temporal y tercero integración con las aplicaciones.

Lo primero es integrar en la arquitectura la tecnología de la compañía o la empresa o la universidad donde estemos trabajando para eso nosotros tendremos ya un proceso donde tendremos los datos aquellos se van ingestando, habrá modelos en explotación que nos van dando los resultados de los modelos que pueden ser informes, gráficos, predicciones, etcétera. Lo que tenemos que hacer es tomar este algoritmo nuevo, este modelo que hemos entrenado e introducirlo dentro de esa familia de modelos en explotación, tendremos que ver si están las librerías adecuadas, las versiones, siempre es cuesta un poco más pasar de un entorno en desavío a un entorno de producción, pero tendremos que hacerlo de la forma más cómoda posible.

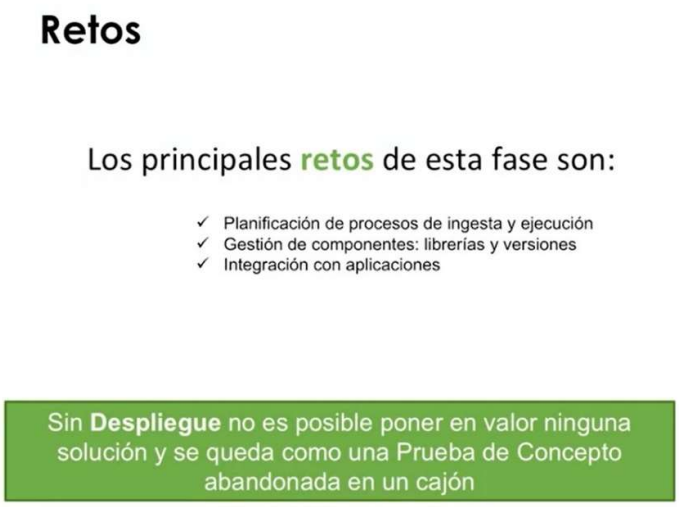


Lo siguiente es establecer la planificación temporal, para explotar este modelo necesitamos datos y esos datos tienen una cadencia de producción o una cadencia de captura y almacenamiento probablemente distinta, tendremos que ver cuándo se capturan los datos, cómo se capturan los datos para establecer una planificación para ejecutar el modelo en el momento que tengamos toda la información disponible.

Y por último tendremos que plantearnos cómo vamos a utilizar este modelo con aplicaciones, es decir, podemos tomar los outputs del modelo e integrarlos en herramientas internas cómo un CRM, podemos disponibilizar los outputs del modelo a través de algún tipo de conector como un API o podemos plantearnos incluirlo dentro de cualquier tipo de aplicación como por ejemplo en las páginas web para que directamente se aplique el modelo cuando se haga algún tipo de interacción, en apps móviles, etcétera. Es muy importante que tengamos claro cuál va a ser el uso del modelo y cómo vamos a utilizar el modelo.



Los principales retos de la fase de despliegue son, primero planificar los procesos de ingesta y ejecución porque si no tenemos los datos en el momento adecuado de tomar la decisión no podremos explotar el modelo. Segundo, gestionar todos los componentes porque las librerías pueden cambiar, pueden cambiar las versiones, pueden cambiar las componentes y entonces tendremos que ir siempre viendo que los modelos que hemos entrenado se adecuan bien a todas esas nuevas componentes y nuevas librerías que vamos introduciendo. Y tercero la interacción de las aplicaciones, hay pensar muy bien cómo vamos a integrar estos modelos en las aplicaciones y eso normalmente requerirá desarrollos en otros lenguajes de programación o en otras plataformas. Lo importante es que el modelo que se quede en explotación resuelva el problema de negocio como se había planteado al principio de la comprensión del lenguaje.

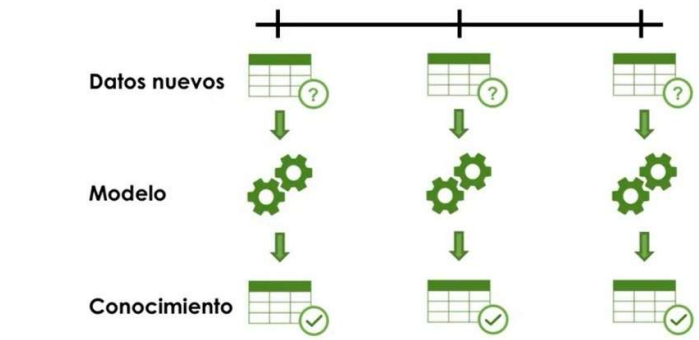


- **Puesta en valor.**

Una vez que ya tenemos el código construido y está desplegado sobre la plataforma tecnológica, lo que tenemos que hacer ahora es integrarlo dentro de las operaciones. Para

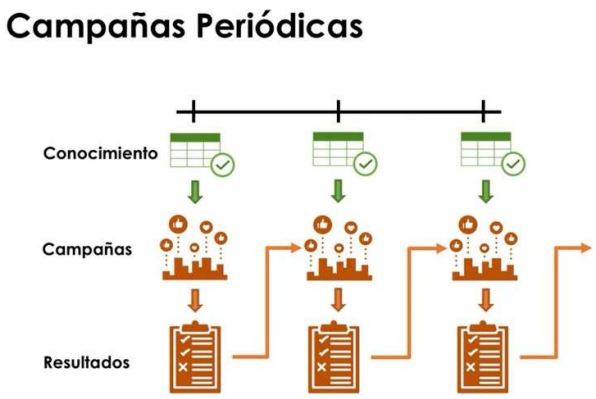
ello hay varias modalidades. Vamos a presentar algunas de ellas para que tengamos la introducción. Las modalidades que vamos a ver son la toma de decisiones, cómo nos apoya en el modelo a la toma de decisiones, campañas periódicas, cómo podemos utilizar el modelo en campañas periódicas. Y, por último, que el modelo se utilice para tomar decisiones desde un punto de vista autónomo.

Es importante tener claro que los modelos no aciertan siempre. Los modelos tienen incertidumbre, porque son predicciones a futuro y el futuro es incierto. Pero lo importante es que son útiles, que nos pueden ayudar a entender el futuro y a poder tomar decisiones. Un modelo en el fondo es una simplificación de la realidad y como tal no puede integrar toda la información disponible por lo que, lo que hay que hacer, es considerar estos modelos como aproximaciones de la realidad que nos pueden ayudar y nos pueden guiar a tomar mejores decisiones. Y lo importante, si nosotros construimos un modelo y el modelo no se acciona, si el modelo no se utiliza, si el conocimiento que tiene el modelo no lo utilizamos, al final el modelo no sirve para nada. Si yo sé qué clientes van a abandonar y no hago nada, no sirve para nada. Si yo sé qué clientes van a dejar de pagar y no haga nada, no sirve para nada. Si yo sé qué día va a llover y no hago nada, no sirve para nada. Entonces es muy importante que tengamos claro que los modelos, su principal valor es que sean accionables. Una vez que tengamos el modelo integrado y desplegado en la plataforma tecnológica, lo normal es que vayan viniendo datos nuevos de forma periódica, o puede ser que sea en real time, pero lo normal es en batch y que vayamos explotando con el modelo y obtengamos conocimiento. Ese conocimiento hay que ponerlo en valor. ¿Cómo podemos ponerlo en valor? Hay distintas modalidades, la primera es la toma de decisiones.

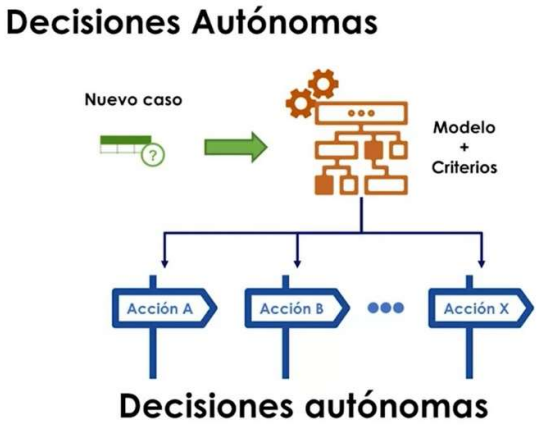


Una vez que yo tengo ese conocimiento, lo podemos presentar como hemos dicho, en un informe o reporte, en una infografía o en un cuadro de mandos interactivo. Y para que esté disponible, para que las personas que tienen que tomar decisiones puedan tomar las decisiones en base a ese conocimiento.

Otra modalidad es utilizarlo en campañas periódicas, con acciones periódicas. Es decir, ese conocimiento que vamos obteniendo de forma periódica conforme tenemos más datos, podemos generar una serie de acciones proactivas por nuestra parte, basadas en ese conocimiento, y después medir cuáles han sido sus resultados, cuáles han funcionado mejor, cuáles han funcionado peor, para que nos sirva de input para ir optimizando esas acciones o esas campañas que nos permitan mejorar u optimizar nuestros procesos.



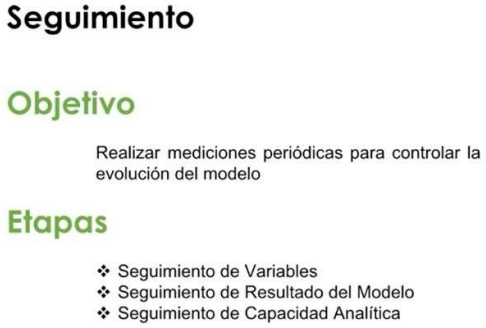
También podemos utilizar el modelo para que tome decisiones autónomas. Nosotros podemos tomar el modelo y en base a ese modelo y unos criterios que establezcamos, podemos hacer que tome esas decisiones autónomas. Por ejemplo, podemos poner un modelo en una web para que en función al visitante y las características del visitante se le vayan mostrando diferentes contenidos y, en función a su consumo, que vaya cambiando ese modelo, esos contenidos que va ofreciendo, para que sea totalmente personalizado.



Los principales retos de esta fase son, primero, acceder al modelo analítico. La mayoría de los proyectos analíticos no llegan a coger valor simplemente porque se acaban en el modelo y nadie es capaz de accionarlos. Es muy importante que el modelo se active, que el modelo se utilice, que ese conocimiento sirva para algo. También es importante la coordinación de todos los departamentos implicados, porque aquí ya hay que introducir este modelo en la operativa normal de las compañías y eso requiere interacción con muchísimos departamentos, tanto tecnológicos como operativos, como de negocio. Y, por último, hay que establecer una sistemática de explotación que tenga en cuenta la medición. Es decir, es importante actuar, medir, aprender, actuar, medir, aprender. Con este objetivo conseguiremos que el conocimiento sea retroactivo y cada vez sea mejor nuestro modelo.

- **Seguimiento, retroalimentación y reentrenamiento.**

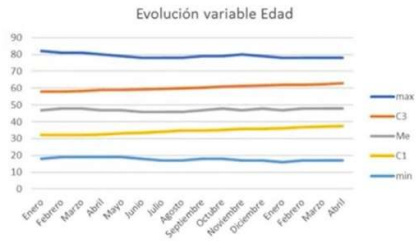
Durante la fase de puesta en valor hemos comentado que hay que hacer el seguimiento de todas las acciones que se realicen para que ese conocimiento nuevo, eso que se aprenda, se pueda introducir después en las nuevas acciones a realizar de forma que vayamos optimizando la operación, pero también es importante hacer el seguimiento del modelo analítico. ¿Por qué? El modelo analítico ha aprendido con unos datos basados, unos patrones de comportamiento, si esos datos cambian o cambian los patrones de comportamiento el modelo puede perder capacidad. Entonces es importante que vayamos midiendo esta capacidad en el tiempo para detectar esas pérdidas por si tenemos que plantearnos volver al anterior modelo o hacer alguna otra acción. Las etapas que vamos a ver en el seguimiento del modelo son, primero vamos a hacer el seguimiento de las variables, segundo vamos a hacer el seguimiento de los resultados del modelo y tercero vamos a hacer el seguimiento de la capacidad analítica del modelo.



Para empezar, es importante conocer las variables, si siguen siendo las mismas o se comportan de forma similar al momento de la extracción que nos sirvió para entrenar el modelo. Para eso es importante utilizar técnicas como la estabilidad de las variables que lo que nos van haciendo o lo que nos van mostrando es cómo evolucionan algún estadístico o

varios estadísticos de esas variables para conocer si en el tiempo se comportan igual y tienen la misma distribución. Si las variables cambiasen de comportamiento, si sus distribuciones cambiasen, serían una señal de alerta que deberíamos analizar.

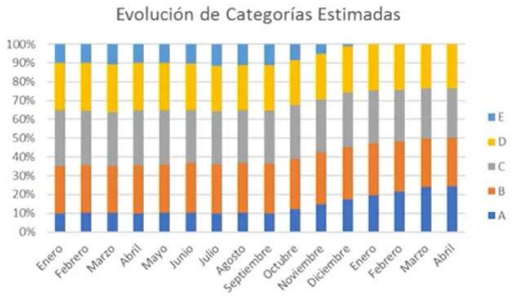
Estabilidad de Variables



Si cambian las variables hay que evaluar su impacto en el modelo

Otro análisis que se puede realizar es la estabilidad del modelo, el output del modelo, podemos analizar cuál es su distribución, cuáles son ciertos estadísticos y ver si en el tiempo se mantienen constantes. Si cambiase la estabilidad del modelo por ejemplo en un problema de clasificación que una de las categorías desapareciese y no se clasificase ningún elemento en esa categoría sería algo atípico, sería algo anormal que tendríamos que analizarse, sería conveniente hacer un análisis más profundo para ver qué ha cambiado.

Estabilidad del Modelo



Si el modelo no es estable hay que revisarlo

También tenemos que analizar la capacidad del modelo, una vez que elijamos la métrica de evaluación que utilizamos en modelización, tendríamos que continuar siguiendo esa métrica y lo habitual y normal es que con el tiempo el modelo vaya perdiendo capacidad. Si el modelo

es robusto y no hay cambios en los patrones de comportamiento los modelos mantendrán la capacidad predictiva más tiempo, si el modelo no es tan robusto la construcción, la metodología, no se ha seguido muy bien o hay cambios en los patrones de comportamiento puede ser que la capacidad del modelo vaya menguando. Es muy importante hacer el seguimiento y fijar un umbral a partir del cual actuaremos y veremos que hacemos en esas actuaciones.

Capacidad del Modelo

Si baja la capacidad analítica hay que revisar el modelo

Las actuaciones que podemos realizar básicamente son dos, primero será re-estimar, es decir manteniendo las mismas variables que tiene el modelo vamos a ver si podemos cambiar los coeficientes o ajustar los coeficientes para que pueda volver a tener la misma capacidad o tendremos que hacer un re-entrenamiento que es volver a empezar toda la metodología, volver a partir del folio en blanco con todo el conocimiento obviamente que ya se tiene pero volver a hacer todas las pautas y todos los pasos para poder tener otra vez un modelo con una capacidad predictiva adecuada.

Acciones a Realizar

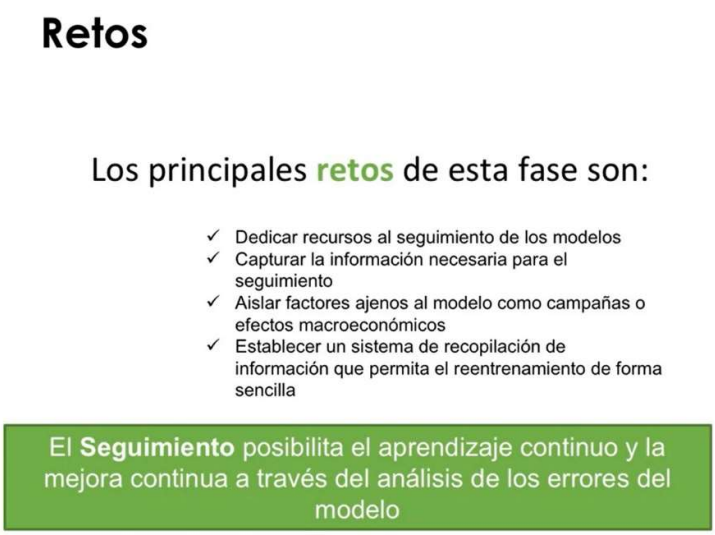
❖Re-estimación

Estimar el modelo utilizando únicamente las variables del modelo y datos más recientes

❖Re-entrenamiento

Realizar el proceso completo de entrenamiento de modelo desde la comprensión de negocio hasta la puesta en valor.

Los principales retos de esta fase son los siguientes, primero, para hacer este seguimiento vamos a necesitar recursos por lo tanto tenemos que tener claro que este es un coste adicional a la construcción del modelo. Segundo, aparte de capturar la información necesaria para explotar el modelo vamos a tener que capturar información adicional para hacer todo este seguimiento, es importante tenerlo en cuenta en la fase de despliegue. Tercero, cuando analicemos los factores que influyen en la capacidad del modelo o en la evolución de las variables o en la evolución del resultado del modelo tenemos que ser capaces de aislar efectos internos o factores internos como son las campañas que hagamos y efectos externos como pueden ser los factores macroeconómicos que influyan. Además, habrá que establecer un sistema de recopilación de información que permita hacer un reentrenamiento de forma sencilla. Como conclusión tenemos que tener claro que los modelos son objetos vivos, las variables pueden cambiar, los resultados pueden cambiar y todo debido a cambios de comportamiento o cambios tecnológicos. Por lo tanto, es muy importante que hagamos el seguimiento y para detectar esas pérdidas de capacidad predictiva y actuar para recuperar esa capacidad predictiva del modelo.



Análisis y desarrollo de herramientas de Big Data.

Ya vimos cómo es el camino que nos permite realizar un correcto análisis, investigación, prueba y puesta en producción de nuestro modelo de Big data, pero ¿qué pasa con las herramientas?, si bien ya habíamos hablado un poco que existen diferentes modelos y metodologías para llegar a nuestro objetivo, no dijimos mucho sobre lo que el mercado puede ofrecernos como usuario. En esta sección veremos solo algunas de las herramientas que nos permiten trabajar con datos. Poder aprovechar los datos y transformarlos en

conocimiento para ser usados en las organizaciones se ha vuelto el objetivo principal del Big Data. Las características de las herramientas de Big Data sirven para tomar decisiones en torno a esto para poder comprender los grandes volúmenes que se generan. Por esto, el Big Data tiene un papel protagonista y es algo imprescindible para cualquier empresa. El análisis de datos se vuelve de vital importancia para captar nuevos clientes como para incrementar ventas y generar estrategias comerciales. En cualquiera de los casos, contar con herramientas de Big Data es tan necesario como la recolección misma. Muchos datos se obtienen en estos procesos y a veces resultan difíciles de analizar.

Tendremos en cuenta las siguientes 10 herramientas de Big Data:

1. Apache Hadoop
2. Elasticsearch
3. Apache Storm
4. MongoDB
5. Apache Spark
6. Python
7. Apache Cassandra
8. Lenguaje R
9. Apache Drill
10. Oozie

- **Apache Hadoop.**

Hadoop es tal vez la herramienta más utilizada para realizar el análisis de datos. Compañías muy grandes como The New York Times y hasta Facebook la emplean para tomar los datos que recolectan y poder hacer cosas con ellos. Al mismo tiempo, ha servido como modelo para otras herramientas de Big Data. La característica principal de Hadoop es que es un framework que permite procesar volúmenes de datos muy grandes en lotes. Además, se organizan en lotes que usan modelos de programación simple por lo que resulta amigable y muy sencilla. Otra de las ventajas es que es escalable. Esto quiere decir que puede operar ya sea con uno o con muchos servidores.

- **ElasticSearch.**

Una herramienta conocida dentro del mundo del Big Data es Elasticsearch. Algunas de las empresas que trabajan con ella son Mozilla y Etsy. En este software para Big Data podrás procesar grandes cantidades de datos e ir viendo la evolución que tengan en tiempo real. Además, cuenta con elementos para el análisis de Big Data tales como gráficos que permiten comprender con más facilidad la información que vayas obteniendo. Una de las ventajas de esta herramienta de Big Data es que permite aplicarle una expansión. ¿Qué quiere decir? Significa que se puede complementar con un paquete de productos extra que sirven para aumentar sus prestaciones. Este conjunto de productos para Elasticsearch se llama Elastic Stack y lo puedes descargar en su sitio web gratis. Algo a destacar de esta herramienta de Big Data es que es un motor de búsqueda y analítica de código abierto y gratuito.

- **Apache Storm.**

Otra de las herramientas de Big Data que es de código abierto y que puede ser usada con cualquier lenguaje de programación es Storm. Este software de Big Data funciona procesando en tiempo real y de forma sencilla mucha cantidad de datos. El sistema de Storm va creando topologías con los macrodatos (aquellos más amplios y menos específicos) y los transforma para analizarlos. Este análisis de Big Data se realiza de forma continua a medida que los flujos de información van alimentando el sistema constantemente. Apache Storm es un sistema para machine learning que puedes descargar en su sitio oficial.

- **MongoDB.**

Esta herramienta de Big Data es una base de datos optimizada para trabajar con grupos que resultan variables frecuentemente. Además, sirve para datos que no son estructurados o que son semiestructurados. Su función principal es almacenar los datos de aplicaciones móviles y sistemas de gestión de contenidos. Las empresas que la usan son Bosch y Telefónica.

- **Python.**

Tal vez te has preguntado qué es Python y para qué se usa ya que es muy popular hoy en día. Esta herramienta de Big Data cuenta con una ventaja fundamental a comparación de otras de esta lista: los conocimientos que son necesarios para usarla son básicos y mínimos. Para saber usar Python basta con tener una mínima idea de programación e informática y no tendrás mayores problemas. Esto hace que tenga una gran comunidad de usuarios y que sea una de las herramientas de Big Data más conocidas y más difundida no solo para Big Data. Se consolida como uno de los lenguajes más sencillos para programar y resulta fácil de aprender. Python tiene una gran comunidad que crean sus propias librerías y las comparten en muchas plataformas. El inconveniente que presenta esta herramienta para manejar Big Data es que es bastante más lenta que el resto de las existentes en el mercado.

- **Apache Cassandra.**

Cassandra es una herramienta de Big Data que se desarrolló en un principio por Facebook. Es una base de datos y resulta tu mejor opción si necesitas escalabilidad y disponibilidad alta, pero sin afectar el rendimiento. Algunos de los usuarios de Cassandra son Netflix y Reddit. Puedes descargarla desde su sitio oficial en el que también encontrarás documentación interesante y una comunidad para resolver tus dudas.

- **Lenguaje R.**

Esta herramienta de Big Data es un lenguaje de programación y un entorno que se enfoca en el análisis de datos estadístico ya que se parece mucho al lenguaje matemático. Se emplea para el análisis de Big Data y cuenta con una comunidad de usuarios que generan una serie de librerías y bibliotecas extensas. En su sitio web puedes encontrar información actualizada y herramientas. El lenguaje R está muy usado en la minería de datos también.

- **Apache Drill.**

Esta herramienta de Big Data es un framework de código abierto que permite un trabajo en el análisis de datos interactivo. Esto lo realiza en grupos y a gran escala. Su diseño fue pensado para alcanzar y procesar petabytes de datos y miles y miles de registros en pocos segundos. Soporta mucha variedad de sistemas y bases de datos y se puede descargar en su sitio web oficial.

- **Oozie.**

La última herramienta de Big Data de la lista es Oozie. Es un sistema que permite definir un rango de trabajos en diferentes lenguajes de programación. Permite a los usuarios que realizan en ella su análisis de Big Data establecer relaciones con estos trabajos. Además, sirve como programador para trabajar en conjunto con Hadoop.

Metadatos en el entorno del Big Data

Pongamos un ejemplo. Cuando escuchamos una canción o vemos una película, tanto el sonido, como el vídeo se puede considerar dato, sin embargo, cuestiones como título, autor, género que definen el contenido de ambos son metadatos.
Los metadatos se han convertido hoy en una fuente de información de gran valor para las estrategias de Big Data. Éstos son los encargados en cierta forma de facilitar el flujo de trabajo y, sobre todo, la comprensión de los datos y la información en general, ya que son pieza clave para la mejora de la eficiencia a la hora de gestionar la información. Para hacernos una idea los metadatos son algo así como etiquetas que ayudan a gestionar y localizar la información.
La definición más correcta indica que los metadatos son “datos que describen otros datos” o dicho de otra forma: “Son el conjunto de datos que proporciona información de un recurso, es decir de otros datos como un archivo de imagen o un documento de texto, siendo algunos de estos metadatos la fecha de creación, la de la última modificación o la resolución en el caso de una imagen”.

Por medio de los metadatos podemos, por ejemplo, saber el nombre, el número de teléfono, la localización, duración de la llamada, etc. En realidad, los metadatos no son más que información que describe el contenido, calidad, condiciones, historia, disponibilidad y otras características de los datos.

METADATOS

Normalmente sirven para facilitar las búsquedas de tal manera que proporcionan suficiente información entre la colección de datos para seleccionar aquellos que más nos interesen o simplemente para saber que existen.
La información de una organización proviene de múltiples fuentes de datos y la comprensión de la misma es una necesidad absoluta en todas las compañías para poder tomar decisiones más acertadas, evaluar las acciones de futuro o fijar sencillamente otros objetivos.
Tener una visión clara de los datos que se manejan y una óptima comprensión de los mismos supone un paso importante dentro de las compañías Data-Driven. En este contexto, los metadatos son el gran aliado del Big Data, pero sobre todo, del Data Analytics, ya que hacen posible desde localizar rápidamente la información, descartar aquella que es irrelevante o tener resultados de confianza para cada consulta.
Es, por tanto, una información valiosísima que resulta imprescindible para un óptimo gobierno de datos, es decir: garantizar de forma satisfactoria la administración eficiente y eficaz del dato.

Reglas de asociación.

Las reglas de asociación tienen como misión encontrar los elementos (*itemsets*), y generar las normas adecuadas, atendiendo a una clasificación para predecir cualquier atributo o combinaciones de los atributos. Una salida profesional y un sector en auge está relacionado con las soluciones que aportan dichas reglas en análisis de datos. Identificarlas, saber qué tipo de aplicaciones tienen en diferentes sectores y empresas, así como determinar los datos que incluyen se hace fundamental en la tecnología. Tanto en la minería de datos como el aprendizaje automático, las reglas de asociación se utilizan para analizar la información dentro de un determinado conjunto de datos.

Las reglas de asociación se definen como un conjunto de técnicas que permiten establecer relaciones de interés con la finalidad de descubrir hechos que aporten valor dentro de las variables que facilitan los datos que son enormes.

Estas técnicas utilizan diferentes algoritmos que generan y testean distintas pautas. La aplicación más práctica se puede realizar en el análisis de la cesta de una compra online.

Aquí podrás observar qué tipo de productos se compran con mayor frecuencia, si quieres realizar una estrategia de *marketing* que sea realmente eficaz.

- **Objetivo y aplicaciones**

El objetivo principal es encontrar correlaciones entre los diferentes elementos u objetos de las bases de datos relacionales, transaccionales o *data-warehouses*. Asimismo, es de vital importancia describir el algoritmo, explicar sus fases y definir las medidas alternativas para el proceso de descubrimiento de estas asociaciones.

Estas reglas cuentan con diversas aplicaciones:

- El soporte para poder proceder a la toma de decisiones.
- Hay un diagnóstico en las telecomunicaciones.
- Se analiza la información en las ventas.
- Se distribuye la mercancía en las tiendas.
- Se segmentan los clientes en función a su patrón de compra.

- **Técnicas para una clasificación exitosa**

A través de un conjunto de técnicas se pueden identificar las categorías a las que pertenecen los datos, si se tiene en cuenta un conjunto de pruebas que contengan los datos categorizados. Es una forma de predecir el comportamiento de un grupo segmentado de clientes. Se podrá analizar el comportamiento del cliente en función de su decisión de compra, ratio de abandono, tasa de consumo o cualquier otra variable que aporte información valiosa.

- Análisis *cluster*

A través del análisis *cluster* o de conglomerados se pueden clasificar los objetos en grupos más pequeños con características similares que se han conocido previamente. Con este método se segmenta a los consumidores en grupos muy parecidos, para posteriormente realizar acciones de marketing.

- Técnica de data mining

En este apartado, también queremos destacar la técnica de data mining para extraer patrones de *datasets* mediante la combinación de métodos estadísticos. Se incluyen técnicas de aprendizaje de reglas de asociación, el análisis de agrupamiento y la clasificación. Si aplicamos la minería de datos a los clientes, podremos llegar a modelar el comportamiento realizado en la cesta de la compra.

- Algoritmos genéticos

Por otra parte, también existen los algoritmos genéticos como una técnica que ayuda a optimizar los datos y tiene en cuenta la supervivencia de los que mejor se adaptan. Estos algoritmos evolutivos son muy funcionales para resolver problemas no lineales. Se puede mejorar la planificación de las tareas en la industria manufacturera u optimizar el rendimiento de una cartera de inversión.

- Aprendizaje automático

Otra técnica interesante es el aprendizaje automático. Una especialidad de la inteligencia artificial que se ocupa del diseño y desarrollo de algoritmos que te ofrecen datos empíricos. El objetivo principal es aprender de forma autónoma a reconocer patrones complejos y tomar decisiones según los datos obtenidos.

- **Características principales de las reglas de asociación**

Según la anterior explicación, existen diversas técnicas para analizar los datos, que aplicadas con eficacia te sirven para extraer los resultados que ofrecen los grupos grandes de datos. A continuación, enumeramos algunas de sus características principales:

- Cumplir con los niveles mínimos de soporte y confianza y que sus subconjuntos también lo cumplan.
- Si algún *item* no se rige por el nivel mínimo, no debe considerarse como superconjunto.
- Generar reglas con un solo consecuente para construir dos o más, de forma sucesiva.
- El esfuerzo que se realiza depende de la cobertura mínima que se requiere.
- Si un conjunto de *items* no supera la prueba de soporte, ninguno de sus superconjuntos la pasará.

- **Extensiones de las reglas de asociación**

Posiblemente, la extensión más exitosa es FP-Growth al ser capaz de calcular eficientemente los conjuntos de elementos frecuentes en ejemplos, utilizando la estructura de datos del árbol FP. Utiliza un método que es capaz de descomponer una tarea en subtareas más reducidas. Por ejemplo, cuando se hace una compra online, se tiene en cuenta una sugerencia donde se explica que los clientes que compraron un determinado artículo también realizaron la compra de otro. Esto es un ejemplo de regla de asociación. Cuando se sabe qué elementos coexisten con mayor frecuencia, el algoritmo *FP-Growth* entra en juego y tiene un papel importante que desempeñar.

El papel de las diferentes abstracciones es otro aspecto a tener en cuenta. Se comienza con una clase superior con el propósito de filtrar clases inferiores. Cuando se encuentran las reglas con diferentes niveles de abstracción se podría generar una redundancia, por ese motivo es importante incorporar mecanismos de filtrado. Las reglas de asociación al igual que las de clasificación funcionan con atributos discretos. Por esa razón, hay que tener enfoques comunes y discretos antes de minar jerarquías predefinidas. Asimismo, hay que conseguir aumentar la confianza y reducir la longitud de las reglas. No siempre que te encuentras una regla de asociación quiere decir que sea útil.