Conviértete en Data Scientist

Apuntes de www.nataliejulian.com (asistiendo)

https://www.linkedin.com/learning/paths/conviertete-en-data-scientist

- **Aproximación al curso Aprende data science**
- A principios de los 90, el trabajo del futuro era el de desarrollador web. Solo los superhéroes con destreza tecnológica podrían guiarnos por la red enmarañada. Iban a liderar el comercio electrónico. Pero no ha sido así. Con herramientas como Wordpress, CRM y Salesforce, todos podemos construir una página web sin necesidad de ningún superhéroe. La mayoría de las organizaciones tiene equipos de desarrollo web. El gestor de proyectos, el analista comercial y los diseñadores gráficos colaboran para crear soluciones complejas basadas en web. Los desarrolladores son parte de un equipo mayor. Con la ciencia de datos, pasará algo parecido. En lugar de un salvador, la organización tendrá equipos de ciencia de datos, que se dedican a colaborar para poner en valor los datos que se recogen. Saben hacer las preguntas adecuadas y obtienen los mejores resultados a través de la investigación y el desarrollo. El científico de datos es solo uno de los miembros del equipo de ciencia de datos. Soy Doug Rose, y este curso está pensado para quien sienta interés por la ciencia de datos, aunque no desee convertirse en un científico de datos a tiempo completo. Descubriremos los desafíos básicos de esta ciencia. Veremos las bases de datos más comunes y los tipos de datos que almacenan. Repasaremos algunas nociones de estadística y análisis predictivo. Por último, veremos algunos inconvenientes típicos de trabajar en equipo. Tanto si eres gestor de proyectos como un desarrollador novel, este curso te permitirá destacarte en esta área. Aprendamos la ciencia de datos.

- **Define una práctica multidisciplinar con muchos significados**
- ¿Qué hace un científico de datos? No es una pregunta fácil. Definir al científico de datos no es tan fácil como con otras ciencias. Si eres científico político o medioambiental, tienes un título que responde a un programa específico. El término «científico de datos» ya se utilizaba antes de que la ciencia de datos se considerara una disciplina. Hoy, los que se autodenominan científicos de datos provienen de varias áreas de investigación. Como disciplina, la ciencia de datos todavía no está definida. Es como la primera arqueología: cualquiera que con una pala buscara artefactos del pasado podía decir que era arqueólogo. Hoy, para serlo, hay que pasar por una universidad reconocida e investigar durante varios años. Como con la arqueología temprana, la ciencia de datos es más una práctica que una disciplina: eres científico de datos si trabajas con datos de forma científica. Así que queda al criterio de cada persona si se quiere presentar como científico de datos. Aun así, algunas profesiones están más preparadas que otras. Si eres estadístico, o analista de datos o trabajas en ciencias biológicas, podrías afirmar que siempre has sido científico de datos, con el énfasis en los datos o en lo científico. Algunos de los primerosque se autodenominaron científicos de datos eran matemáticos, otros venían de ingenierías de sistemas e información; e incluso algunos provenían de los negocios y las finanzas. Si habías trabajado con números y sabías algo sobre datos, podías autodenominarte científico de datos. Como la demanda de científicos de datos no para de crecer, se intentará que las habilidades se estandaricen para que las empresas sepan a quién están contratando. Pero eso aún no ha ocurrido. De hecho, existe el riesgo de que se piense que cualquiera que trabaja con datos y tiene dos dedos de frente es científico de datos. La mejor forma de pensar en la ciencia de datos es centrándose en la ciencia y no en los datos. En este contexto, hablamos del empirismo, que reacciona ante los datos, mediante experimentos y preguntas. Es probable que uses el método empírico todo el tiempo sin pensarlo como un método. Un científico de datos usa este método cada día. Una aproximación empírica combina el conocimiento y la práctica. Veamos un ejemplo de cómo uso yo una aproximación empírica. Como asesor y formador, viajo bastante a menudo y paso las noches en habitaciones de hotel con diferentes estilos de cuarto de baño. No dejo de sorprenderme ante la variedad de canillas, llaves y dispositivos que hay en el mundo. Siempre me cuesta descubrir cómo se enciende la ducha en cada hotel. Comienzo formulando la pregunta empírica: «¿Cómo enciendo la ducha?» Luego pruebo con un experimento, aprieto un botón y el agua llena la bañera. Si aprieto otro, la lluvia comienza a caer. Con el agua ya fluyendo, debo controlar la temperatura. Uso los distintos tiradores y botones para conseguirlo. Si giro demasiado de uno, el agua sale demasiado caliente; si giro el otro, sale demasiado fría. Con preguntas y reevaluaciones consigo que el agua salga a temperatura confortable. Es verdad que podría adoptar una teoría sobre cómo conseguir la temperatura confortable, meterme de un salto, girar una perilla y cruzar los dedos. El problema es que tengo la misma probabilidad de congelarme que de quemarme. Los científicos de datos usan esta aproximación empírica todo el tiempo. Hacen preguntas a los datos y las ajustan para ganar entendimiento, ajustan las perillas e intentan mejorar la calidad de las preguntas. La ciencia de datos es una disciplina fluida, con muchos profesionales de formación variada, que se autodesignan científicos de datos. Céntrate en el método científico y utiliza la aproximación empírica para obtener conocimiento a partir de los datos y refuerza el lado científico, no el de los datos. Te convertirás en un mejor científico de datos mientras no se sistematice el tema.

- **Los paquetes estadísticos y las herramientas informáticas**
- Como la ciencia de datos todavía se está definiendo, es importante estandarizar el software y las herramientas. Ocurre como con los primeros arqueólogos. El software sería como los pinceles y piquetas para excavar. No concentres demasiados esfuerzos en aprender todas las herramientas, ya que no te transformarán en científico de datos. Es el método científico y no las herramientas lo que hace al científico de datos. Las herramientas caben en tres categorías básicas: almacenamiento, depuración y análisis. Para almacenar los datos, puedes usar hojas de cálculo, bases de datos y bases de datos clave-valor. Las más utilizadas son Hadoop, Cassandra y POST REST SQL. La depuración se usa sobre todo para que sea más fácil trabajar con los datos. Se utilizan editores de texto, herramientas de secuencia de comandos y lenguajes de programación tipo Python y SCALLOP. Por último, hay paquetes estadísticos que ayudan a analizar los datos. Los más usados son el paquete de código abierto R, SBSS, y las bibliotecas de datos de Python. Con estas herramientas, también se pueden visualizar los datos y crear gráficos y diagramas. Veamos qué herramientas para almacenar datos debes conocer. Se escucha mucho la idea de los «desafíos del big data». El volumen de información es tal, que los sistemas de gestión en base de datos no pueden gestionarlo La conexión entre la ciencia de datos y el big data es tan íntima, que muchos creen que son lo mismo. Pero la ciencia de datos aplica el método científico para estudiar los datos. No hace falta que estén dentro de la categoría de big data. Hay un libro fantástico titulado Data Smart, que nos presenta la estadística para ciencia de datos mediante simples hojas de cálculo. Sin embargo, las áreas más activas de esta ciencia trabajan con big data. Actualmente, el programa Hadoop de código abierto es el más utilizado. Hadoop usa un sistema de archivos distribuido para almacenar los datos en servidores estándar. Esta red de servidores se denomina clúster Hadoop. El clúster divide las tareas para que puedas ejecutar las aplicaciones. Puedes tener petabytes de datos distribuidos en cientos o miles de servidores. Puedes ejecutar procesos sobre los datos en el clúster. Los procesos más comunes que verás son MapReduce y Apache Spark. MapReduce trabaja con los datos por lotes. Y Spark puede procesar los datos en tiempo real. Una vez que hayas recogido los datos, es hora de un poco de limpieza. A menudo, la información que recoges no se puede utilizar. Imagina que reúnes millones de tuits de tus clientes. Si tienes una cuenta de Twitter, ya sabes que a veces hay texto y otras veces hay imágenes. Cuando recoges estos datos, podrías crear un script o comando que divida todos los tuits en texto o imagen. Luego, vuelves a almacenarlo en tu clúster y así podrás analizar cada uno de estos grupos de diferente forma. Si lo haces con frecuencia, te conviene crear una pequeña aplicación con Python para que repita la tarea una y otra vez. La mayor parte del tiempo, un científico de datos se la pasa depurando los datos. Algunos dicen que consumen hasta el 90 por ciento del tiempo depurando la información para que se pueda usar. Ahora se puede usar R o Python para analizar los datos. R es un lenguaje de programación estadística. Te permite crear conexiones y establecer correlaciones entre los datos. Luego se pueden presentar con la función de visualización de datos incluida en R. Consigues un informe eficiente con un diagrama visual. Imaginemos un tipo de informe que quieres crear. Tu empresa quiere saber si hay alguna conexión entre las opiniones positivas y el momento del día. Capturas los datos de Twitter en tu clúster de Hadoop, y depuras los datos para categorizar los tuits como positivos o negativos. Por último, con un paquete estadístico tipo R creas una correlación e imprimes un informe con su correspondiente diagrama. Estas son solo algunas de las herramientas más populares. Si estás en un equipo de data science, oirás hablar de al menos una de ellas. Existen muchas más que automatizan la recolección, la depuración y el análisis, y diferentes organizaciones invierten en satisfacer a una base de clientes que no para de crecer. Se trata de priorizar el análisis. Las herramientas y los datos solo son el vehículo para aumentar el conocimiento. Así que limita el gasto en nuevo software.

- **Crea conocimiento sobre el negocio**
- En los últimos 20 años, muchas organizaciones han optado por aumentar su eficiencia operativa. Se han esforzado por optimizar los procesos corporativos. La idea es ganar en eficiencia y flexibilidad. Hacen preguntas operativas como ¿se puede trabajar más rápido y de forma más eficaz? La ciencia de datos es diferente. No se organiza por objetivos que cumplir. Es exploratoria y utiliza el método científico. Intenta obtener conocimiento útil para los negocios. Con la ciencia de datos se hacen varios tipos de preguntas. ¿Qué sabemos de nuestros clientes? ¿Cómo podemos ofrecerles un producto mejor? ¿Qué hacemos mejor que la competencia? ¿Qué pasaría si desapareciéramos del mercado? Son preguntas que exigen un nivel elevado de pensamiento organizativo, y muchas entidades no están preparadas para hacerse estas preguntas. Están repletas de empleados que trabajan en dar respuestas y marcar hitos con sus objetivos. No se les paga por demostrar escepticismo o un espíritu de prospección. Recuerda la última reunión a la que hayas asistido. Piensa qué pasaría si alguien tuviese preguntas tipo: ¿por qué lo hacemos así? ¿Por qué creen que funcionará de esta forma? ¿Por qué creen que es buena idea? Es probable que todo el mundo se sintiera incómodo. Alguien le reclamará ¿no te has leído el memorándum? Pero son habilidades que necesitas para afianzar el conocimiento organizacional. Son el tipo de preguntas que tu equipo de data science debería hacerte. Pero la mayoría de las empresas se centra en la producción de resultados. Explorar otras posibilidades sería un cambio. Las preguntas se perciben como un obstáculo al progreso, pero como empresa, solo ganarás conocimiento haciendo preguntas interesantes. Una vez trabajé para una web que conectaba compradores de autos con las concesionarias. En la web, había cientos de etiquetas de información. Con las etiquetas se podía ver si el cliente solo curioseaba o hacía clic en algún enlace. Esta información se almacenaba en un clúster Hadoop. Conseguían terabytes de datos cada semana. La empresa archivaba datos recogidos hacía años. Habían destinado mucho presupuesto al departamento que se encargaba de almacenar y mantener los datos. Recoger esos datos era fácil, los programas que usaban eran simples y fáciles de crear. Lo complicado era decidir qué se hacía con esa información. Comenzaban a pensar en los tipos de preguntas que podría hacerse. Este es un problema típico de muchas organizaciones cuando se introducen en la ciencia de datos. Lo perciben como un cambio operativo, sobre todo. Al principio se trata de recoger los datos. Ponen todo el esfuerzo en recoger tantos datos como sea posible y almacenarlos en su clúster Hadoop. Un proyecto así da importancia al aspecto técnico de los datos, inevitablemente. No entienden la ciencia de datos como ciencia. Se explica fácilmente porque recoger datos es bastante barato y muy fácil de explicar. Todo el mundo puede sumar sus esfuerzos, y crear múltiples clústeres para incluir en ellos datos de toda la organización. Pero casi todas las empresas lo tienen más difícil con la parte científica. No están acostumbradas a hacer buenas preguntas. ¿Qué preguntas podría hacerse la web de mi ejemplo? ¿Y si experimentaran con el color de los autos? Podrían establecer la probabilidad de que un cliente haga clic en una imagen roja, azul o amarilla. El informe podría demostrar que los clientes hacen clic un dos por ciento más de veces si el auto es rojo. Podrían compartir esa información con las concesionarias de autos, y obtener más beneficios. ¿Qué tal si experimentan para ver si la página está mostrando demasiados autos a la vez? Podrían probar qué sucede con los clics si se limita el número de autos disponibles, y crear un informe con los resultados. Este tipo de investigación empírica es lo que un científico de datos debería pensar siempre. Se trata de exprimir los datos mediante preguntas interesantes, experimentar con los resultados y producir informes bien diseñados. La ciencia de datos comienza con una pregunta, pero debe incluir experimentos. Cuando los hayas ejecutado, puedes utilizar las hojas de cálculo y los programas para crear los informes. Analízalos, y verás si obtienes alguna revelación respecto del negocio.

- **Haz conexiones con bases de datos relacionales**
- Los científicos de datos trabajan de varias formas. Extraen los datos de bases de datos antiguas y de hojas de cálculo. También trabajan con imágenes y videos. Deberías familiarizarte con las formas más comunes de almacenar datos que tienen las organizaciones. La mayoría tiene un amplio abanico de opciones. Algunas son muy modernas y otras, no tanto. La mejor forma de entender estas tecnologías es desde el principio. Hasta las bases de datos más modernas se fundamentan en tecnologías con 50 años de antigüedad. Las bases de datos modernas comenzaron la misión espacial Apolo a finales de los 60. La NASA trabajó con IBM para crear un sistema de gestión de la información, o IMS. Los cohetes que lanzaban a la luna necesitaban millones de piezas, y la NASA trabajaba con estos prototipos que se parecen mucho a una hoja de cálculo moderna. Era un archivo informático con varias columnas y largas listas de filas. Una tabla con un millón de filas es difícil de gestionar. Piensa en una hoja de cálculo de un millón de filas en una pantalla en blanco y negro. IBM comercializó más tarde el IMS creado para la NASA. A mediados de los 70, desarrollaron un lenguaje de consulta estructurada, el SQL, para asistir en la búsqueda de datos de sus clientes. Al mismo tiempo se estaban creando las primeras bases de datos relacionales, que separaban los datos en grupos de tablas. cada una de estas tablas parece una hoja de cálculo, pero con menos información. Después crearon relaciones entre las tablas. En lugar de una sola lista de millones de partes, podían crear 50 tablas con 20 000 piezas cada una. Por eso se llaman bases de datos relacionales, porque se fundamentan en un grupo de tablas que se relacionan entre sí. Los primeros ingenieros se esforzaron por aprender cuál era la forma más eficiente de agrupar las tablas. Crearon mapas que mostraban las relaciones entre tablas y los llamaron esquemas. En este trabajo pionero se percibe la dificultad de crear estos esquemas. ¿Creamos una tabla para las piezas más grandes? Tal vez sea mejor una para los propulsores y otra para los tanques de combustible. El problema está en que, si cambias el diseño del cohete, tienes que cambiar el diseño de la base de datos. Las tablas se podrían crear para cada fabricante de cada pieza; el problema en este caso sería que tal vez un fabricante produce miles de piezas, mientras que otro solo produce un par de docenas. A día de hoy es un problema sin resolver. Las bases de datos relacionales necesitan su buena cantidad de diseño previo. Hace falta tener muy clara la apariencia de la información antes de empezar a recogerla. Si te equivocas, el esfuerzo para corregir el diseño es muy costoso. SQL es un lenguaje elegante que puede extraer datos de muchas tablas relacionales diferentes. Es capaz de reconectar varias tablas y presentar los datos en una tabla virtual que se llama Vista. SQL tuvo tanto éxito que todavía hoy es uno de los lenguajes más usados. Si buscas en LinkedIn, verás que es una de las habilidades más populares. Con los años, se añadieron muchas funciones a las bases de datos relacionales y se creó un sistema de gestión de bases de datos relacionales llamado RDBMS. Algunas de las empresas de software, como IBM, Microsoft y Oracle aún dan asistencia y desarrollan este tipo de sistemas de gestión.

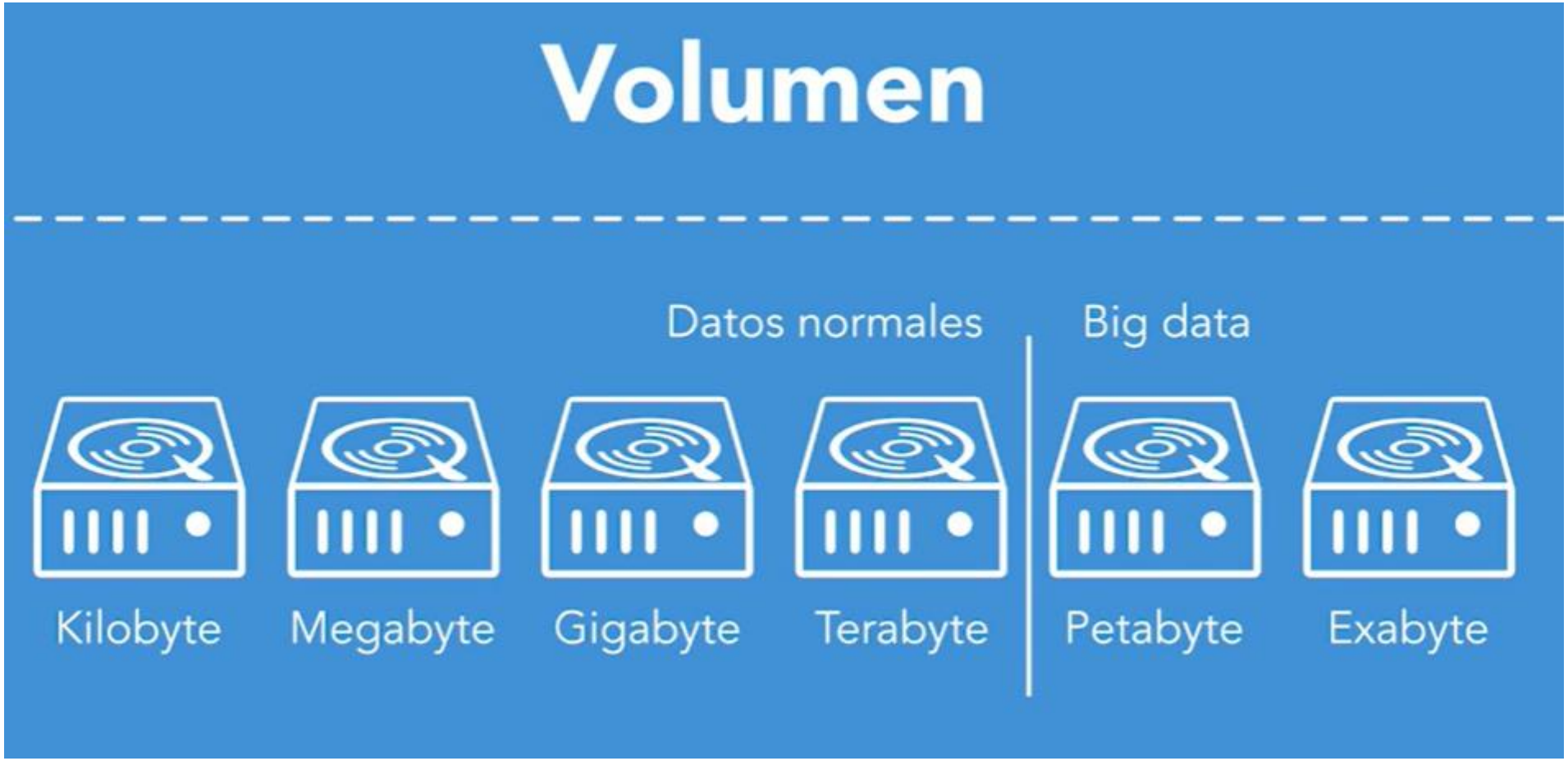
- **Guarda la información en almacenes de datos mediante ETL**
- Muchos conceptos de la ciencia de datos se basan en el trabajo previo con bases de datos. Las empresas hace décadas que capturan y analizan los datos. El sistema de gestión de bases de datos relacionales es la piedra angular de los datos corporativos. Tienes que entender los términos de nuestro DBMS en tus proyectos de data science. Es probable que te los encuentres cuando te centres en un almacén de datos. Un almacén de datos o EDW es un tipo especial de base de datos relacional que se dedica a analizar la información. Las bases de datos tradicionales están optimizadas para el procesamiento de transacciones en línea, o OLTP. Un almacén de datos se usa para el procesamiento analítico, conocido como OLAP. para que lo entiendas: la base de datos típica se dedica a trabajar con datos en tiempo real. Un almacén de datos se dedica a analizar lo que ya ha sucedido. Imagina que tienes una página que vende zapatillas de correr. Contratas a un ingeniero para que cree tu base de datos y crea decenas de tablas y relaciones diferentes entre ellas. Hay una tabla de direcciones de clientes, una de zapatillas, una de opciones de envío… Tu servidor web tendría que usar una instrucción SQL para buscar entre los datos. Cuando un cliente compra su calzado, esa instrucción une su dirección con el zapato y les asigna la opción de envío. Es deseable que esta base de datos sea rápida y eficiente. Es una base de datos transaccional OLTP. Se optimiza para que, cuando el cliente encuentre el calzado, lo una a su dirección, haga clic en enviar a su domicilio y la base de datos reaccione velozmente. El cliente espera que sea en tiempo real. Puedes pedirle al ingeniero de sistemas que cree una secuencia de comandos que suba los datos cada día al almacén. Tu almacén de datos está optimizado para procesamiento analítico. Es una base de datos OLAP dedicada a crear informes. Puedes pedirle a un analista de datos que cree un informe para comprobar si existe alguna conexión entre la dirección de un cliente y el tipo de calzado que compra. Si descubres que los clientes que viven en zonas más templadas compran zapatos más coloridos, puedes usar esta información para actualizar la presentación de las zapatillas en la web. Podrías cambiar la web para que los clientes de zonas templadas vean primero las zapatillas de colores claros. Ahora imagina que tu página es muy exitosa y la compra una empresa de ropa deportiva que ya tiene un almacén para todas sus páginas web. Quieren extraer los datos de tu página web y combinarlos con los de sus otras páginas. La empresa tiene llevar a cabo un proceso llamado ETL, que quiere decir: extraer, transformar y cargar. Extraen los datos de todas sus páginas y los cargan en su propio almacén de datos corporativo. La empresa extrae los datos de tu página web en un formato estándar. Luego los transforma en el tipo de información que necesita para su almacén de datos, que podría tener un esquema diferente al tuyo. El analista de datos pasa mucho tiempo depurando y uniendo datos para que se ajusten. Por último, cargan los datos transformados en su almacén. Cualquier equipo de data science tiene estos retos. Es probable que usen el mismo lenguaje. Puede que escuches por ahí: «Tenemos que ETL los datos en el almacén para moverlos al clúster Hadoop». Esto quiere decir que el analista de datos debe transformar tus datos antes de poder moverlos a tu clúster. Otro punto importante es que muchas organizaciones ven Hadoop como un reemplazo de sus almacenes de datos. Muchas empresas reescriben sus archivos ETL para poder cargar los datos en un clúster Hadoop a medida que eliminan sus almacenes de datos por etapas o los cierran directamente. Las organizaciones esperan ahorrar gastos con el almacenamiento de datos en hardware de precio accesible. Los almacenes de datos son un servicio muy costoso. Te cruzarás con estos conceptos cuando trabajes en un equipo de data science. Intenta que este lenguaje no te frustre demasiado. Una gran parte de la ciencia de datos todavía tiene que ver con la recogida de datos. Pasarás horas en reuniones de ETL antes de poder hacer preguntas interesantes. Si comprendes los términos y desafíos, es más probable que consigas los datos que necesitas.

- **Olvídate del pasado con NoSQL**
- Para muchas empresas, las bases de datos relacionales son la columna vertebral de las transacciones digitales y creen que los almacenes de datos son la piedra angular de las analíticas corporativas. Las bases de datos relaciones han cumplido su cometido, pero las aplicaciones contemporáneas nos presentan desafíos que superan este modelo. Suele pasar que el equipo de data science necesite formas más flexibles de almacenar datos. Recuerda que estas bases de datos se basan en esquemas. Tienes que conocer los datos a fondo antes de poder ingresarlos en la base de datos. Hace falta mucha planificación: primero, conocer el formato de la información, si es audio, texto o video. Luego, organizas los campos dentro de la tabla. Y por último, estableces las relaciones entre las tablas. Tenemos una para vender zapatillas deportivas. Eres un cliente que ha encontrado el calzado que quiere. La página enlaza tu zapatilla con una dirección de envío. Ya puedes pasar por la página de pedidos. Esa página sola tiene que acceder a cuatro bases de datos diferentes. La de zapatillas, la de clientes, la de direcciones y por último la de envíos. Es demasiado trabajo para una base de datos transaccional. Cuanto más trabajo tenga la base de datos, más lenta será tu página web. Aumentar la velocidad también es complicado. ¿Compras un servidor con más capacidad? ¿O separas las tablas en varios servidores? ¿Puedes sincronizar unos cuantos servidores a través de tu red? Cuando hablamos de páginas web muy extensas, estas alternativas no son prácticas. Piensa en una base de datos que almacene todos los datos de la página de pedidos en una sola transacción. Crea un registro del calzado, el cliente, su dirección y el envío, todo en una transacción. No hace falta que dividas los datos en varias tablas. No tienes que preocuparte de crear las relaciones. Alcanza con volcar los datos. Es la idea que sostiene al NoSQL. NoSQL nació como etiqueta de Twitter; la usaban los desarrolladores que querían superar las bases de datos relacionales. No es una negación del SQL. De hecho, el NoSQL no tiene casi nada que ver con el SQL. Solo señala las limitaciones del modelo relacional. En general, una base de datos NoSQL es no-relacional, sin esquemas, apta para los clústeres e, idealmente, de código abierto. Estas cualidades son atractivas para un equipo de data science. Si una base de datos no es relacional, es más fácil cambiarla y usarla. No existe una gran diferencia entre el funcionamiento de tu aplicación web y la forma en la que almacenas los datos en la base. No hace falta que pases por procesos agotadores de creación y división de las tablas existentes, para crear diferentes vistas. Esto se denomina «normalizar tu base de datos». Sin esquemas, no hace falta saber cómo será todo antes de comenzar. Imagina que una empresa más grande compra tu página web de calzado. Esta empresa quiere sumar a tus clientes a su programa de compradores frecuentes. Con una base de datos relacional, estás ante un problema de arquitectura. ¿Deberías tener a los compradores frecuentes marcados en la tabla de clientes? Tal vez haga falta crear una tabla entera con números de identificación de los clientes frecuentes. ¿Se puede asignar más de un número de cliente a cada cliente? ¿Pueden compartir número de cliente dos clientes? Todo esto hay que pensarlo antes. Debes rediseñar la base de datos y buscar la solución para los datos que falten. Sin un esquema, es muy simple añadir un campo nuevo. Se almacena como una nueva transacción. Si el cliente tiene un número de cliente frecuente, se carga como parte de la transacción. Si no lo tiene, no existe ese campo para ese cliente. Una base de datos NoSQL es apta para clústeres. Se pueden almacenar los datos en varios cientos o incluso miles de servidores de bases de datos. En una base de datos NoSQL, los registros se guardan en una transacción que se llama agregado, que contiene toda la información. El calzado, la dirección del cliente y la información del envío. Es fácil sincronizar estos agregados a través de varios servidores de bases de datos. Muchos servidores trabajan en clústeres. Así pueden sincronizarse y enviar sus actualizaciones a otros clústeres. La palabra «clúster» ya te suena. Hadoop trabaja así con sus conjuntos de datos. Está construido sobre Hbase, una base de datos NoSQL de código abierto. Cuando trabajas en un equipo de data science, es muy probable que te cruces con NoSQL. Es una buena forma de lidiar con conjuntos masivos de datos. Gracias a su diseño simple, también los desarrolladores pueden crear aplicaciones web. Estas aplicaciones crecen enseguida a escala corporativa.

- **Aborda los problemas de los datos masivos**
- Para algunas empresas, big data y la ciencia de datos son conceptos mezclados. La ciencia de datos usa el método científico con los datos que tienes. No hace falta que sean muchos datos para que puedas hacer las preguntas. Los datos masivos ofrecen una fuente de datos contundente. Esta nueva fuente te permite preguntar lo que conjuntos pequeños de datos no podrían responder. Por lo general, a mayor cantidad de puntos de información, mayor capacidad de análisis estadístico. «Big data» suena a título de película de terror de los 60: una mujer gritando tras sus gafas retro mientras la sepulta una montaña de información. En realidad, «big data» no era una frase nominal. En el documento original de la NASA, se describía como un «problema de datos masivo». Se podía leer en inglés como una de dos alternativas. Es un problema de «datos masivos» o un «problema masivo» de datos. Cuando lees el artículo entero, da la sensación de que el énfasis estaba en el «problema». No son «datos masivos», sino un problema de exceso de información que almacenar. También lo encontramos en el informe McKinsey, donde se refieren a los datos masivos para hablar de la información que excede la capacidad del hardware. ¿Por qué pensar en el big data como un problema y no como una frase nominal? Porque la mayoría de las empresas con proyectos de big data no tiene datos masivos. Puede que creas que son masivos porque son muchos. También representa un problema porque no sabes dónde almacenarlos, pero no es un problema de datos masivos. Una forma de definir si tienes un problema de big data es averiguar si tus datos entran en una de cuatro categorías. Son las cuatro V. Puedes preguntar: ¿Tengo un volumen elevado de datos? ¿Hay una gran variedad de datos? ¿Los datos llegan a gran velocidad? ¿Los datos que recojo son veraces? ¿Me darán conocimiento o revelaciones? para que sean datos masivos, deben cumplir con los cuatro atributos. Podrías dudar si el volumen es suficientemente elevado o no, pero esta pregunta suele ser fácil. Si recoges petabytes de información cada día, tienes un volumen suficiente. Por supuesto, podría no ser un problema. En el futuro, tal vez haga falta un exabyte para que el volumen se considere un problema. También piensa en la variedad de los datos. Esta es más difícil de responder. Piensa en la bolsa de Nueva York. Gestionan millones de transacciones cada día. Seguro que tienen un volumen elevado de datos. También entran a gran velocidad. Los precios de las acciones entran y fluctúan en milisegundos, pero si lo piensas un poco, verás que es el mismo tipo de información. Es el símbolo de la acción y el precio. Básicamente, es texto, no hay imágenes ni sonidos ni clips de noticias. No tienen un problema masivo. Recogen muchos datos, sin duda, pero la tecnología que usan es más que suficiente para gestionar ese problema. Por último, ¿los datos son veraces? Imagina que quieres crear una base de datos que recoja todos los tuits y posts de Facebook que hablen de tu página web. Juntas los videos, las imágenes y el texto, varios petabytes de datos dirigidos a tu clúster cada día. Creas informes para descubrir si tus clientes se sienten a gusto con tus productos. Cuando recorres los datos, te das cuenta de que no hay información que defina el estado de ánimo del cliente. El esfuerzo se desperdició en recoger datos inútiles. Cuando pienses en los datos masivos, recuerda las cuatro V. Así podrás decidir si tienes un problema de big data, como, por ejemplo, el de los autos sin conductor. Tienes que recoger todo tipo de datos. Son cantidades masivas de videos, sonidos, informes de tránsito y datos de GPS. Todo irá entrando en la base de datos a tiempo real y alta velocidad. Luego, el auto tiene que decidir cuál de los datos es más veraz. ¿La persona que grita a un lado de la carretera está celebrando un evento deportivo? ¿O está gritando porque hay alguien en medio de la carretera? Un conductor real tiene segundos para darse cuenta. Un auto basado en big data tiene que procesar videos, audios y coordenadas de tránsito, para decidir si se detiene o hace caso omiso de ese ruido. Es un verdadero problema del big data. Recuerda además la diferencia con la ciencia de datos. Big data te permite hacer preguntas más interesantes, pero no significa que todas dependan de tener datos masivos. Céntrate en la ciencia, y podrás hacer las mejores preguntas, sin importar la cantidad de datos.

Las cuatro V

1. Volumen
2. Variedad
3. Velocidad


Volumen
Datos normales
Big data
Kilobyte
Megabyte
Gigabyte
Terabyte
Petabyte
Exabyte

- Test de capítulo

- Uno de los problemas de las bases de datos relacionales es que antes de diseñarlas tienes que tener mucha información sobre los datos.
- ¿Por qué es importante comprender los almacenes de datos? Porque muchas organizaciones utilizan el lenguaje y los conceptos de los almacenes de datos en la ciencia de datos.
- No es una característica de una base de datos NoSQL: es relacional.
- ¿Cuál de las siguientes no es una característica de big data? Los datos tienen verborrea.

- **Organiza la información con los datos estructurados**
- Los equipos de ciencia de datos tienen que tratar con varios tipos de datos, que son un factor clave a la hora de determinar cómo se almacenan los datos. Una tecnología tipo NoSQL aporta mucha flexibilidad para almacenar todo tipo de datos. Las bases de datos relacionales pierden flexibilidad a cambio de facilidad de uso. Para decidir cómo almacenar los datos, tienes que conocer los distintos tipos que existen. Es lo mismo con cualquier almacenamiento. Hay una base de datos optimizada para cada tipo de datos; así como un sándwich no se almacena en una jarra, no pondrás en una base de datos relacional el tipo incorrecto de datos. Existen tres tipos de datos que el equipo debe tener en cuenta. Hay datos estructurados, semiestructurados y no estructurados. El primer tipo es el más simple. Reciben el nombre de datos estructurados. Son los datos que tienen un formato específico y siguen cierto orden. Los datos estructurados son como los ladrillos y el cemento del mundo de las bases de datos. Son baratos, inflexibles y hace falta mucho diseño previo. Un buen ejemplo es la típica hoja de cálculo de oficina. Cuando llenas las filas de información, tienes que ser fiel a una estructura bastante rígida. Por ejemplo, creas una columna con Fecha de compra. Cada entrada de esa columna debe seguir unas directrices estrictas. No puedes poner «martes» en una fila y «marzo» en la siguiente, el formato es importante. Cada fila tiene que tener el mismo formato. Puedes adoptar un formato estándar, como el mes seguido de barra y el año. Esta estructura se llama «modelo de datos». Los datos estructurados se basan en el modelo de datos. Un modelo de datos es parecido a los esquemas de una base de datos relacional, salvo que el esquema acaba definiendo toda la estructura de la base de datos. Un esquema te muestra cómo organizar tu base de datos relacional. Incluye la tabla, las relaciones y las interconexiones que existen. Un modelo de datos define la estructura de los campos individuales. Es la forma de definir lo que cabe en cada campo de datos. Allí decides si el campo tendrá texto, números o fechas. En el caso de la hoja de cálculo, es fácil ver el problema de hacer caso omiso del modelo de datos. Si pones «martes» en el campo de fecha, la mayoría de las hojas de cálculo no te lo impedirá. Pero en la fila de abajo pones «marzo». Parece fácil, seguro que te cuesta percibir el error. El problema viene después. Imagina que quieres un informe con todas las compras hechas en marzo. ¿Cómo lo haces? ¿Usarías el número tres o la palabra «marzo»? Seguro que no usarás la palabra «martes». Si lo haces, la hoja de cálculo estaría llena de datos basura. Cada vez que quisieras ordenar los datos o emitir un informe, tendrías varias filas con datos inválidos. Tendrías que perder tiempo limpiándolos o eliminarlos del informe. Por eso las hojas de cálculos tienen reglas de formato. Son las reglas que te obligan a respetar el modelo cuando ingresas los datos. Lo mismo sucede con las bases de datos: la mayoría va a rechazar cualquier dato que no siga el modelo. Las páginas web o la fuente que uses para recoger datos suelen estar programadas para tipos y formatos específicos. Las bases de datos relacionales son excelentes para recoger datos estructurados, que abundan por todas partes. Mucha de la información que obtienes de páginas web o aplicaciones móviles proviene de datos estructurados. Los extractos bancarios, la información de vuelos, los horarios del bus, incluso tu agenda de direcciones, son formas de datos estructurados. Eso no quiere decir que la mayoría de la información sea estructurada. De hecho, la mayoría de la información no sigue un formato determinado. Algunos de los datos más interesantes no tienen estructura alguna. Los datos tipo videos, imágenes y audios no tienen estructura definida. Piensa en las imágenes que subes desde el teléfono. Puede ser una foto de cualquier cosa tomada en cualquier lugar. Podría tener altísima calidad o ser un desastre pixelado. Un archivo pesado o uno pequeño. No hay una estructura intrínseca a ese dato que ayude a la base de datos a almacenarlo. El equipo de data science debe combinar este tipo de datos con el método de recolección. Si usas una base de datos relacional, vas a estar limitado a los datos estructurados. Si usas un clúster NoSQL, cualquier tipo de datos es materia de trabajo, aunque será más complicado crear informes. Son decisiones que debe tomar el equipo. La ciencia de datos es tal porque aplica el método científico a tus datos. Los datos son la materia prima que te permiten hacer las preguntas. Como equipo, deben decidir qué material les permitirá obtener la información más interesante.

Tipos de datos

1. Estructurados
2. Semiestructurados
3. No estructurados

- **Comparte datos semiestructurados**
- Los equipos de data science trabajan con muchos tipos de datos. Las bases de datos relacionales son la mejor opción para datos estructurados. Con un modelo de datos estricto, los datos estructurados caben en el esquema. Es como una hoja de cálculos con filas y columnas fijas. Con los datos estructurados, los informes se crean con facilidad. Puedes usar un lenguaje de consulta estructurada como SQL para extraer los datos de tu base de datos y mostrarlos en un formato estándar. Cuando los datos estructurados se anidan en la base de datos relacional, parece que el mundo entero estuviera organizado. Es como las especias en sus especieros. Sabes dónde está cada cosa y dónde ir a buscarla. El problema es que muy pocas aplicaciones son tan sencillas. Tenemos una web de zapatillas deportivas, e imagina que utilizas una base de datos relacional. Tienes cuatro tablas, una para los zapatos, otra para los clientes, otra para sus direcciones y opciones de envío. Todos los datos estructurados caben en un modelo de datos. Las fechas son estándar, los códigos postales también. Sale todo bien, el mundo parece un lugar maravilloso. Hasta que recibes un mensaje de la compañía de reparto. El repartidor dice que podrías bajar los costos de forma sustancial añadiendo la información directamente en su base de datos. Debes consultar su base de datos, descargar uno de los códigos de envío regionales y añadirlo al pedido para crear un nuevo registro. Debería ser fácil, porque las bases de datos son similares. Son datos estructurados en bases de datos relacionales. El problema es que su esquema no es igual al tuyo. Tu código postal se llama ZIPCode. Ellos lo han llamado código PostalCode. A ti no te afecta si el envío va a una empresa o a una residencia. A ellos, sí. Tú no necesitas especificar si es una casa o un apartamento. Para ellos son tarifas diferentes. Tienes que encontrar la forma de intercambiar tus datos estructurados con los suyos, aunque tengan esquemas diferentes. Para resolverlo, tienes que descargar los datos y su base de datos. Cuando te envían una dirección, tiene que incluir los nombres de los campos y el modelo de datos. Cuando un cliente pide unas zapatillas, tu base de datos enviará el código postal a su base de datos. Te devolverá unos cuantos datos que incluyan su versión de la dirección con los nombres de campo que usan. Recuerda que usan la etiqueta «codigopostal» para el CP. Se incluirá en el dato de nueva creación. Los datos del repartidor tienen características de datos estructurados. Están bien organizados y tienen formato estándar. El campo de texto siempre será de texto. Los campos de fechas siempre serán de fechas. Pero los datos incluyen el esquema del repartidor, que puede utilizar los nombres que quiera. Por eso, este tipo de datos se llaman semiestructurados. Estos datos son más populares que los estructurados. Tienen cierta estructura, pero esta depende de la fuente. Trabajas con datos semiestructurados todo el tiempo. El correo electrónico es de datos semiestructurados, con una estructura bastante coherente. Siempre tienes un emisor y un destinatario, pero los nombres y el contenido del campo pueden variar. Los equipos de ciencia de datos trabajan más con datos semiestructurados que con los de tipo estructurado. Hablamos de cantidades de correos, blogs y redes sociales que se pueden analizar. Hay algunas formas establecidas de trabajar con datos semiestructurados. Uno de ellos es XML, el tipo antiguo de datos semiestructurados que se usa para intercambiar información. También existe JSON o JavaScript Object Notation, que es una forma actualizada de intercambiar datos de este tipo. Es el tipo de datos preferido para los servicios web. Es muy probable que tu página web de zapatillas reciba datos JSON del repartidor de envíos. Con datos semiestructurados, puedes hacerte preguntas más interesantes. Supón que queremos conocer la opinión de los clientes. ¿Están satisfechos con sus pedidos? Puedes descargar datos semiestructurados de redes sociales. Luego, combinas esos datos con los datos estructurados que tienes acerca del cliente. Si no está satisfecho, puedes enviarle un cupón de resarcimiento. Este tipo de preguntas son las que surgen del uso combinado de datos estructurados y semiestructurados.

- **Recopila datos no estructurados**
- Vamos a hacer un repaso de lo explicado hasta ahora. En general, los equipos de data science trabajan con tres tipos de datos. Tenemos los datos estructurados, como los de las hojas de cálculo. Tienen un orden establecido y un formato coherente. A menudo se almacenan en bases de datos relacionales. Luego están los datos semiestructurados. Tienen una parte de estructura, pero existe flexibilidad para cambiar los nombres de los campos. Por último, el tipo más popular de datos son los no estructurados, que abarcan todo lo demás. Según algunos analistas, el 80 por ciento de los datos son no estructurados. Si lo piensas, tiene sentido. Piensa en la información que creas cada día. Cada vez que dejas un mensaje de voz, o con cada foto que subes al Facebook. El memorándum escrito en Word para el trabajo o la presentación de PowerPoint. Incluso cuando buscas en internet, es de forma no estructurada. Si buscas «gatos» aparecen videos, canciones, libros e incluso música. ¿Qué tienen en común estos datos? Ese es uno de los retos clave. La respuesta corta es que no tienen casi nada en común. Carecen de esquema. El esquema es como el mapa que muestra los campos de datos, las tablas y las relaciones. Con los datos no estructurados no existe un esquema. Para estos datos el formato depende del archivo. Un documento de Word tiene un formato establecido, pero solamente lo puede utilizar ese programa. No es el formato de todos los documentos de texto. Por eso no se puede editar un documento de Word con otro programa. Eso también significa que no hay un modelo de datos establecido. No hay un lugar coherente para buscar nombres de campo y datos. ¿Cómo descubres el título y los contenidos de cientos de tipos diferentes de archivos? ¿Cómo se hace si tienes pdf, documentos de Word y presentaciones de PowerPoint? cada uno tiene su propio formato. No hay un campo que indique el título del documento. Es un problema que los buscadores como Google y Bing hace años que investigan. ¿Cómo se trabaja con datos que no tienen un formato establecido ni usan un modelo de datos coherente? Cada vez que buscas con estos motores, verás el fruto de sus esfuerzos. Si buscas un término como «gato», verás que han encontrado la forma de buscar texto, videos, imágenes y audios. Trabajar con datos no estructurados es la parte más interesante de la ciencia de datos. Las bases de datos más nuevas tipo NoSQL te permiten capturar y almacenar archivos grandes. Es mucho más fácil almacenarlos en una sola ubicación. Todos los archivos de audio, video, imagen y texto pueden ir a un clúster NoSQL. Puedes escalar los servidores de forma horizontal y usar herramientas y programas similares. Para capturar la mayor cantidad posible de datos también hay nuevas herramientas. Puedes usar tecnología de big data como Hadoop, para procesar los datos de un clúster. Luego puedes analizar esos datos con MapReduce o Apache Spark. Volvamos a nuestro ejemplo. El negocio ha crecido y ahora eres parte de un equipo de data science nuevo. Trabajas con los directivos y con marketing para formular las primeras preguntas. ¿Quién es el mejor cliente para zapatillas deportivas? Buscas la información biográfica básica. Es fácil buscarlo en la base de datos de clientes. Tienes su dirección de email y conoces la ciudad y región donde vive. Con esa información, puedes escanear la actividad del cliente en las redes sociales. Comienzas a recoger todos los datos no estructurados. Tal vez el cliente colgó un video corriendo al final de una maratón. Puedes enviarle un tuit de felicitaciones. También podrías curiosear los posts de los amigos de tu cliente, si lo ves en alguna foto saliendo a correr con amigos. Los datos no estructurados te permiten identificar a esas personas y enviarles promociones especiales. Estos proyectos se llaman de visión 360 grados de la relación con el cliente. Intentas descubrir todo lo que puedas sobre sus motivaciones. Luego, usas esa información para definir los mejores clientes y enviarles promociones. Podrías descubrir que algún cliente te recomienda a muchos de sus amigos. Podrías añadir algún incentivo o descuento. Con el tiempo, puedes recoger cada vez más datos no estructurados. Y las preguntas serán más sofisticadas. ¿Tus clientes suelen viajar? ¿Son más competitivos? ¿Con qué frecuencia salen a comer o a cenar? cada una de estas preguntas te ayuda a conectar con el cliente y a venderle más productos. Los datos no estructurados son un recurso en crecimiento constante. Piensa en todo lo que hiciste hoy que podría interesarle a alguna empresa. ¿Tuiteaste sobre el paseo que diste? Tal vez necesites mejor calzado nuevo. ¿Te quejaste porque era un día lluvioso? Podrías comprar un paraguas. Los datos no estructurados les permiten a las empresas ese nivel de interacción.

Visión 360° del cliente

Investiga las motivaciones

Envíale promociones

Ofrécele incentivos

Formula preguntas adicionales

- **Criba los datos que no utilices**
- Los datos no estructurados presentan desafíos. Una de las primeras preguntas que surgen es si deseas eliminar algunos datos. Debes utilizar el método científico con tus datos, para poder formular preguntas interesantes como equipo de data science. Debes decidir si pones un límite a las preguntas que vas a poder hacer alguna vez. Hay argumentos a favor de mantener y deshacerse de parte de los datos. Algunos analistas de datos argumentan que siempre puedes encontrar una pregunta para hacer que no conocías antes. También resulta barato mantener cantidades masivas de datos. Son unos pocos céntimos por cada gigabyte. No cuesta nada mantenerlos, y evitarse la difícil decisión de qué habría que eliminar. Saldría más barato comprar nuevos discos duros que las horas de reunión para decidir qué se mantiene. Por otro lado, algunos analistas afirman que es recomendable eliminar los datos. Se acumula demasiada basura en esos clústeres de big data. Cuanta más basura haya, más difícil será encontrar resultados interesantes. Algunos analistas llaman a este problema «ruido en los datos». Muchos científicos de datos intentan dar con una solución. ¿Cómo se gestiona tanta basura? Trabajé para una empresa que tenía este problema. Su página web conectaba clientes potenciales con concesionarios de coches. Crearon un sistema de etiquetas que registraba todo lo que miraba cada cliente en la web. Si pasaba por encima de una imagen, la base de datos creaba un registro. Este sistema de etiquetas registraba todos los clics y movimientos por la página. El sistema fue incorporando miles de etiquetas, cada una de ellas con millones de transacciones. En la empresa había unas pocas personas que entendían qué datos capturaba cada etiqueta, así que era difícil redactar informes productivos. Podían registrar cuántas personas entraban en cada etiqueta, pero unos pocos entendían que quería decir esa etiqueta. Usaban el mismo sistema de etiquetas con los datos no estructurados. Comenzaron a recolectar las publicidades y videos breves. Querían conectar cada etiqueta con la imagen y la transacción. Así, podían ver la imagen que el cliente había hecho clic. Existía una etiqueta que indicaba dónde se ubicaba en la página. Todo se almacenaba en el clúster que crecía. Algunos miembros del equipo decían que muchos de los datos eran obsoletos, y solo unos pocos entendían el sistema de etiquetas lo suficiente como para comprender la información. Las publicidades cambiaban constantemente, así que comenzaron a renombrar las etiquetas y muchos de los datos se quedaban obsoletos enseguida. Para otros, la cantidad de datos recogida era pequeña comparada con lo que podría almacenarse en un clúster Hadoop. ¿A quién podrían molestarle un par de gigabytes de datos obsoletos? No valía la pena el esfuerzo de limpiarlos. Este es el tipo de problemas a los que también te tendrás que enfrentar. Recuerda un par de detalles. Lo primero es que no hay una respuesta correcta. El equipo de data science tiene que definir qué le funciona mejor. Si decides mantener todo lo que recojas, tendrán que trabajar más arduo a la hora de crear informes de interés. Hará falta más filtros y tus datos tendrán más ruido. Si decides eliminar toda la basura, el clúster estará limpio. Podrías eliminar sin querer parte de la información que un día echarás en falta. Se parece a limpiar los armarios. No puedes saber si esa chaqueta de cuello ancho volverá a estar a la moda. Si te quedas con todas las chaquetas, no recordarás todo lo que ya tienes. Lo más importante es que el equipo tome una u otra decisión. La política de datos no puede cambiar cada pocos meses. Decide al principio si mantendrán todos los datos o van a ir eliminando diferentes grupos de datos. Trabaja con el equipo para que todos estén de acuerdo en qué información eliminar. Si no estableces una política de retención de datos, corres el riesgo de corromper los datos. Si no sabes lo que has eliminado, será más difícil que los informes tengan sentido. Decide lo antes posible qué prefieres para tu empresa.

Argumentos a favor de mantenerlos

Es difícil saber qué podrías necesitar más adelante

Almacenamiento barato

La eliminación consume tiempo

Argumentos a favor de eliminarlos

Ayuda a gestionar la basura y el ruido

Cada vez es más difícil encontrar resultados valiosos

- Test de capítulo

- ¿Cuál es el mejor ejemplo de dato estructurado? la factura del teléfono

- Los datos semiestructurados no incluyen ningún esquema. FALSO

- ¿Cuál de los siguientes es el mejor ejemplo de dato no estructurado? una foto que guardas en el móvil

- **Empieza con la estadística descriptiva**
- Un equipo de data science se dedica a recoger datos, depurarlos y almacenarlos. Luego formulan preguntas a partir de ellos. Crean informes mediante matemáticas para entender mejor esa información. La estadística es una disciplina muy valiosa. Para formar parte de un equipo de ciencia de datos, hacen falta algunas nociones básicas. Es útil recordar que las estadísticas son una herramienta para contar una historia. Pero no son la finalidad de la historia en sí. La mejor forma de darte cuenta de cuánto te falta entender de una historia es poner distancia cuando algo parece incorrecto. Mi hijo me contó un chiste al respecto. Demuestra cómo usar la estadística para contar una historia. Me preguntó: «¿Sabes por qué nunca ves un elefante escondido entre las ramas de un árbol?» Le respondí que no. Y dice: «Porque sabe esconderse». Recuerda este chiste cuando leas los informes. Creemos que las estadísticas son matemáticas puras. Nadie cuestiona que dos más dos son cuatro. En realidad, la estadística se parece más a la narración. Como con cualquier historia, se puede llenar de hechos, ficciones y fantasías. Se pueden esconder varios elefantes si no sabes dónde mirar. Es fácil verlo con ejemplos en la política. Un candidato afirma que en los últimos cuatro años el salario promedio aumentó 5000 dólares. El público aplaude y celebra. Luego su oponente afirma que no deberían alegrarse porque, de hecho, en los últimos cuatro años, cada familia de clase media ha perdido 10 000 dólares de ingresos. ¿Quién dice la verdad? Ambos candidatos. Utilizan las estadísticas a su favor con una historia diferente. Una habla de prosperidad y la otra de fracaso. Ambas son ciertas, pero ninguno de los políticos cuenta toda la historia. En cada historia hay que buscar el elefante. En este caso, cada candidato usa estadísticas descriptivas. Intentan describir la situación de todos los votantes sin tener que hablar de cada familia. Hablan de la familia típica. Un candidato usa la media, que es básicamente un promedio. Suma todos los ingresos de cada familia, lo divide por el número total de familias. Es una de las estrategias estadísticas más útiles y más usadas. Puedes calcular el promedio general de tus calificaciones estadísticas deportivas, tiempo estimado de viaje e inversiones. En este ejemplo, el político sumó los ingresos de todas las familias, y dividió el resultado por el número total de familias. Claro que cada familia ganó 5000 dólares más. Pero si te fijas, la media no es la única forma de describir el ingreso familiar. Su contrincante tiene otra estrategia. Usa el ingreso familiar medio. La mediana describe los ingresos de la familia que se encuentra en el centro de la distribución. Para calcularlo, se enumeran de menor a mayor ingreso todas las familias. Las numeras del principio al final. Divides el número total por dos y encuentras qué cantidad de ingresos corresponden a la mediana. La familia de la mitad de la lista tiene los ingresos medianos. Recuerda buscar lo que no te cuentan. Si existe una diferencia muy grande entre la media y la mediana, quiere decir que tus datos están sesgados. En ese caso, imagina que unas pocas familias son muy acaudaladas. Sus ingresos han aumentado muchísimo en los últimos años. Podríamos hablar de millones de dólares de patrimonio. Estas familias tergiversan los datos porque en un extremo encontramos un grueso de dinero. Eso aumenta la media, pero no tiene efecto sobre la mediana. En la media, sus ingresos se suman como los de los demás y se incluyen en el promedio. En la mediana, solo estarían ubicados en el extremo superior. Pero como el número de familias no cambia, tampoco se modifican los ingresos de la familia del punto medio. Este problema de la media y la mediana aparece por todas partes. Si hay dos personas en una habitación, su altura media podría ser de 1,75 metros. Si entra un jugador de básquet a la habitación, la media podría aumentar 20 centímetros. La altura mediana seguiría siendo más o menos la misma, pero el grupo estaría sesgado en altura. En el equipo de data science, siempre debes cuestionar las historias que se cuenten con estadísticas. Repasa siempre las justificaciones de cada afirmación. Intenta que los informes usen formas diferentes de describir los datos. Busca el elefante. Las estadísticas cuentan varias historias.

- **Entiende la probabilidad**
- La probabilidad es una parte de la estadística que te permite contar una buena historia. Es el cálculo de las posibilidades que existen de que una cosa se cumpla. Es una medición de los resultados posibles. Si arrojas una moneda, la probabilidad predice de qué lado caerá. La parte estadística de la probabilidad está en la distribución probabilística. Si lanzas un dado de seis caras, hay seis resultados posibles. La posibilidad de que salga un número determinado es de uno en seis. Cada vez que arrojas el dado, tienes un 17 por ciento de probabilidades de acertar el número que saldrá. La probabilidad también demuestra una secuencia de hechos. Puedes calcular la probabilidad de sacar el mismo número dos veces seguidas. Eso corresponde al 17 por ciento del 17 por ciento, lo que equivale al tres por ciento. Si estás jugando a los dados, es una probabilidad muy baja. La probabilidad es una herramienta útil para tu equipo de data science. Es una parte esencial de la analítica predictiva. Te ayuda a calcular la probabilidad de que un cliente tenga determinado comportamiento. Cuando trabajaba en una empresa de biotecnología, intentábamos predecir la probabilidad de que un paciente aceptara formar parte del ensayo clínico. Conseguir participantes para los ensayos clínicos es una tarea complicada. Existen varios ensayos y es muy costoso mantenerlos en funcionamiento, incluso si están vacíos. Si no se llenan, la empresa pierde beneficios. Por eso usan ciencia de datos para preguntar. ¿Qué hace que un paciente se abstenga de participar en un ensayo clínico? Hay varios aspectos que disminuyen la probabilidad de que un paciente participe. Si no está permitido cenar la noche anterior, hay un 30 por ciento menos de probabilidad de que participe. Habrá otro 20 por ciento menos de probabilidad si el ensayo incluye análisis de sangre y pinchazos. La empresa debe equilibrar la probabilidad de que los pacientes participen con la precisión que necesita de los datos. Pongamos que están probando la eficacia de un nuevo medicamento, que se puede medir con un análisis de sangre o uno de saliva. El análisis de sangre tiene un 10 por ciento más de probabilidad de ser preciso. Es fácil: que usen un análisis de sangre. Pero espera: si el ensayo exige un análisis de sangre, tendrá un 20 por ciento menos de participantes, lo que disminuye la cantidad de sujetos evaluados en el estudio. Se pierden los pacientes que rechazan los pinchazos como parte del estudio. Si necesitan 1000 participantes, son 200 personas menos. Con esto tenemos otra pregunta importante. Si el ensayo tiene 200 pacientes menos, ¿los resultados son menos precisos? Nuestro equipo de ciencia de datos creó otra distribución probabilística. ¿Qué pasa si el medicamento puede causar algún tipo de reacción? Tienes más mediciones con 1000 pacientes que con 800. El equipo debía tener eso en cuenta. ¿Convenía tener más pacientes en el estudio sin usar agujas, aunque bajara la precisión? Esto llevó a preguntas más interesantes. ¿Debería repetirse el análisis de saliva para mejorar la probabilidad de que sea preciso? Al final, esa era la decisión que el equipo de data science ayudó a tomar. Tal vez lo mejor fuera tener el mayor número de pacientes, para aumentar la probabilidad de que una reacción adversa se detectase; también se debía repetir más veces el análisis menos preciso, para mejorar la probabilidad de que los resultados fueran precisos. Así la empresa consiguió la máxima participación y al mismo tiempo un aumento de la posibilidad de su estudio. Todo gracias al poder de la probabilidad. Cuando trabajes con probabilidad, debes tener presentes un par de ideas. Primero, la probabilidad te llevará por caminos inesperados. ¿Quién hubiera imaginado que un estudio médico obtendría mejores resultados mediante un análisis menos preciso? Segundo, la probabilidad es una herramienta que permite hacer preguntas cada vez más interesantes. Que no te desanime descubrir que cada pregunta te lleva a más preguntas. Recuerda que la ciencia de datos aplica el método científico a tus datos. A veces, este recorrido lleva a destinos insospechados. Lo importante es no darse por vencido cuando el camino da un giro inesperado. Puede pasar, al trabajar con probabilidades, pero los giros inesperados suelen llevarte a las revelaciones más valiosas.

- **Busca correlaciones para mejorar resultados**
- La estadística descriptiva es una herramienta de análisis. Otra idea a considerar es la correlación. Muchas empresas la usan para adivinar qué tipo de productos comprarás. También se usa para conectarte con amigos y conocidos. Si tienes Netflix, seguro que te ha sorprendido la precisión con la que te recomienda películas que te acaban gustando. Amazon utiliza la correlación para recomendarte más productos. La correlación es una serie de relaciones estadísticas que miden el grado de relación entre dos variables. Se suele medir entre uno y cero. Si hay una correlación de uno, las dos variables se correlacionan con fuerza. Si la correlación es de cero, las dos variables no tienen relación. El uno también se puede expresar en positivo o negativo. Un uno negativo es una inversa, o una anticorrelación. Una correlación positiva es la que existe entre altura y peso. Cuanto más alta sea una persona, es mucho más probable que pese más. Si aumenta la altura, aumenta el peso. También hay ejemplos más sencillos. A mayor temperatura exterior, más probabilidad de que la gente compre helado. Si sube la temperatura, aumentan las ventas de helado. Un ejemplo de correlación negativa lo tenemos con los autos y la gasolina. Cuanto más pese el auto, menos va a rendir el kilometraje de combustible. Si el peso del auto aumenta, la distancia por litro disminuye. Es una relación inversa. Si te gusta correr, también sabes que vas más lento si vas cuesta arriba. También es una correlación negativa. A mayor pendiente, menor velocidad conseguirás. Si la pendiente aumenta, la velocidad disminuye. Las correlaciones son un instrumento muy interesante para analizar las relaciones entre dos variables. Una correlación inversa no es dañina, solo expresa otra forma de relación. Un equipo de data science busca correlaciones entre sus datos, para calibrar cualquiera de las relaciones. Por suerte, hay programas informáticos que hacen todos los cálculos necesarios para obtener una correlación. Una fórmula que se utiliza a menudo es la del coeficiente de correlación. Lo normal es que el número no sea redondo, es probable que veas una correlación de 0,5 o un 0,75 negativo. Así se demuestra si la correlación es fuerte o débil. Cuanto estés más cerca del uno o del menos uno, más fuerte será la correlación. Un ejemplo del trabajo de la ciencia de datos está en la función de Gente que podrías conocer de LinkedIn La empresa quería encontrar una forma de saber qué profesionales ya se conocían. Varios equipos trabajaron con datos de LinkedIn buscando correlaciones entre las conexiones. Luego buscaron la explicación de esas conexiones. Podría ser la escuela a la que asistieron, los trabajos que compartieron o que compartan grupos o intereses. El equipo de data science buscó correlaciones positivas y también negativas. Los datos pueden demostrar que te interesa un trabajo. Otra persona puede estar interesada y ambos trabajan en la misma empresa. El equipo de data science sabe qué trabajos buscas y dónde has trabajado antes. Es suficiente para establecer una correlación entre dos personas. Puede existir una correlación positiva muy fuerte entre la gente que trabaja en la misma oficina y está interesada en los mismos trabajos. La red social te acaba sugiriendo si quieres conectar con esa persona. El equipo de data science también establece una correlación entre tus conexiones y las de los demás. Si estás conectado con una persona, y ella tiene conexión con otros profesionales que tienen tus mismas habilidades, podría ser una buena sugerencia para ti. Tiene toda la lógica, si lo piensas. Es mucho más probable que conozcas a alguien que trabaja en la misma oficina. También es más probable que te conectes con los que tienen tus mismos intereses. A medida que el número de intereses en común aumenta, mayor es la probabilidad de que conozcas a esa persona. La correlación también permite poner a prueba las suposiciones. Podrías suponer que los clientes que pasan más tiempo en tu página web son los más satisfechos. Pero puede que no sea así. De hecho, podría existir una correlación negativa, y que quienes más gastan tengan las expectativas más imposibles de cumplir. Es fácil que los desilusiones y que acaben dejando una opinión negativa. El equipo de data science pone a prueba estas suposiciones usando correlaciones. Busca estrategias que consigan que los clientes más felices gasten más dinero. Aprende a gestionar las expectativas de los que más gastan en tu empresa. Si buscas estas correlaciones, detectarás relaciones que hubiesen pasado inadvertidas.

- **La correlación no implica causalidad**
- La correlación te ayuda a ver relaciones donde a primera vista no te dabas cuenta. Pero debes definir si la correlación es la causa de lo que estás analizando. ¿Es la correlación la causa del cambio? Como regla general, la correlación no implica causalidad. Eso quiere decir que una relación entre dos variables podría estar afectada por una tercera que no forma parte de tu análisis. No es fácil establecer la causalidad desde el equipo de data science. No quieres crear relaciones que no existen. Un ejemplo: nací en una de las zonas más frías de Estados Unidos. Cuando mis padres se jubilaron, se mudaron a Florida. Hoy viven felices en una soleada comunidad de jubilados en Las Vistas de Boca Lago. Los visitamos en familia cada pocos meses. Según las estadísticas, su comunidad es uno de los lugares más peligrosos del planeta. Cada vez que visitamos hay personas que están siendo hospitalizadas o algo peor. Existe una correlación muy alta entre su comunidad y la muerte o las lesiones graves. Podrías creer que por este motivo nunca visito a mis padres. Ese lugar es la descripción inicial de cualquier videojuego de francotiradores. Pero vamos de visita a menudo y nos sentimos a salvo, porque la correlación no implica causalidad. La causa real es que la edad mediana es mucho más alta. Los ancianos que viven en una comunidad de jubilados tienen una probabilidad más alta de lesionarse o morir. Si solo miras la correlación, parece que vivieran en una zona de guerra. No te creerías que se pasan el día jugando a las cartas junto a la piscina. Piensa en formas de aplicar estos conceptos desde el equipo de data science. Volvamos a la página web de calzado deportivo. El equipo identifica un aumento de ventas en enero. Hay una correlación entre el mes de enero y la cantidad de gente que compra calzado nuevo. El equipo se reúne para descubrir la causa. Se hacen preguntas interesantes. ¿Los clientes tienen más ingresos en enero? ¿Por qué salen más personas a correr durante los meses más fríos? ¿Son corredores noveles que están motivados por los propósitos de año nuevo? ¿Son clientes nuevos? ¿Qué tipo de zapatillas compran? El equipo debate sobre las preguntas y decide crear informes. El informe muestra que la mayoría de los clientes son nuevos y compran calzado caro. Con estos informes, el equipo decide que la causa del aumento de ventas es que los clientes nuevos tienen más dinero en enero. Puede que hayan recibido tarjetas de regalo. Al año siguiente, el equipo decide aprovechar esta causalidad. En diciembre, ofrecen tarjetas de regalo para navidades. También envían promociones a los clientes nuevos del año anterior. Unos meses después, vuelven a analizar los datos. Descubren que sus promociones y descuentos no han tenido efecto alguno. Resulta que hay el mismo número de compradores para la misma cantidad de zapatillas. Parece que la causa del aumento no era la disponibilidad de dinero. El equipo de data science vuelve a las preguntas que había hecho y redacta más informes. Descubren que todas las compras nuevas, ambos años, son de clientes nuevos y corredores novatos. ¿Por qué ese aumento abrupto de clientes nuevos que compran zapatillas costosas durante los meses más fríos? El equipo reflexiona al respecto y decide que la razón puede ser de conducta. Se hacen otra pregunta: ¿todos los clientes nuevos están pensando en ponerse en forma como propósito de año nuevo? Al año siguiente, deciden ofrecer otra promoción, basada en los propósitos de año nuevo. Envían un correo que anuncia: «¿Quieres cumplir con tu propósito de año nuevo?» Ofrece guías gratuitas para corredores y pulseras de actividad para mantenerlos interesados durante todo el año. La correlación y la causalidad son una dificultad para todos los equipos de data science. Es muy fácil establecer relaciones falsas. En estadística, se llaman relaciones espurias. Si encuentras la causa real, el valor añadido será mucho mayor. La mejor forma de evitar una relación espuria es seguir el método científico. Las preguntas deben estar bien formuladas y tus prejuicios no deben afectar a los resultados.

- **Combina las técnicas de la analítica predictiva**
- Todo lo que has visto hace referencia al pasado. Cómo recolectar distintos tipos de datos, y someterlos a análisis estadísticos, que son el punto de partida del verdadero conocimiento. El equipo de data science puede establecer correlaciones entre algunos hechos. Ahora podemos darles la vuelta a estas ideas. Usaremos el conocimiento para darle la vuelta y predecir el futuro. Esto se llama analítica predictiva. El término va muy asociado a la ciencia de datos. A veces se usan de forma intercambiable, pero no son exactamente lo mismo. La analítica predictiva es un tipo de ciencia de datos. La ciencia de datos se basa en aplicar el método científico a los datos. La analítica predictiva usa esos resultados y los transforma en acciones concretas. Piénsalo así: la meteorología es un tipo de ciencia. Estudia la física, la velocidad del viento y la atmósfera. Si sales con un meteorólogo, te enseñará por qué las nubes tienen formas diferentes y cómo la presión determina los movimientos que siguen. Es la ciencia de la meteorología. Intenta comprender el tiempo y las tendencias históricas. Aun así, la mayoría de la gente no habla de la meteorología como de una ciencia. Piensan en ella como la predicción del tiempo. La predicción del tiempo es lo que hacen los meteorólogos cuando usan la analítica predictiva. Con la probabilidad y la correlación predicen patrones del tiempo atmosférico. Los meteorólogos se valen de los datos históricos para asignar probabilidades. Pueden encontrar correlaciones entre los sistemas de baja presión y grandes tormentas. A medida que disminuye la presión, la gravedad de la tormenta aumenta. Es una correlación positiva entre la presión y las tormentas. Todos los análisis se reúnen en la respuesta del meteorólogo cuando le preguntan: «¿Cómo estará el clima mañana?» Lo que ha servido para entender el pasado se vuelve una predicción del futuro. El interés por el análisis predictivo se ha renovado. Las nuevas tecnologías e instrumentos permiten que el conocimiento sea más certero. Piensa en la predicción del tiempo. Ahora mismo, la agencia meteorológica tiene acceso a datos históricos de miles de estaciones. Pero imagina que distribuyeran millones de balizas. Los ciudadanos podrían instalarlas en sus casas y conectarlas a sus redes wifi. Con estos dispositivos económicos se puede registrar la presión del aire y la temperatura. También pueden grabar video y audio. Se podría subir toda esa información a un clúster Hadoop. Los científicos tendrían acceso a unos niveles de información sin precedente. Por eso la analítica predictiva se asocia directamente con los datos masivos de la ciencia de datos. A mayor volumen de datos, mejores serán las preguntas que el equipo pueda plantear. Así el equipo puede hacer análisis complejos. En este caso, podría observar los patrones del tiempo casa por casa y manzana por manzana, y crear modelos predictivos complejos basados en millones de hogares. Lo mismo vale para tu equipo. Tenemos una página web de zapatillas. Imagina que el equipo colecciona millones de tuits sobre la actividad de correr. Identifica a unos cuantos corredores influyentes. Podrías enviarles regalos o promociones, para que hablen bien sobre tu empresa. También se pueden usar estos datos para identificar eventos clave. Con estas nuevas herramientas se puede ganar en perspectiva sobre los datos. El equipo puede analizar millones de tuits, como la agencia meteorológica analiza petabytes de información. Puedes observar oleadas de información a tiempo real. Una regla general es que a mayor cantidad de datos, los resultados de la analítica predictiva será más robusta y precisa. Las organizaciones suelen entusiasmarse con la analítica predictiva, al punto de que no siempre dedican tiempo y recursos suficientes al equipo de data science. Quieren saltar directamente a las predicciones sin haber comprendido los datos. Cuando formas parte de un equipo de data science, tienes que explicar que la calidad de las predicciones depende del análisis previo que tu equipo haga de los datos. Tu equipo tiene que entender el pasado para poder predecir el futuro. No menosprecies el análisis. Haz buenas preguntas sobre tus datos y usa las herramientas estadísticas para crear informes. Así, tus predicciones serán muchísimo más precisas.

- Test de capítulo

- Las mismas estadísticas pueden contar historias distintas.

- **Céntrate en el conocimiento**
- Uno de los desafíos de la ciencia de datos es lo que llamo el clúster de los sueños. Se basa en una película con Kevin Costner llamada Campo de sueños. La película trata de un granjero que dedica todos sus ahorros a construir un campo de béisbol en su plantación. Lo visitan fantasmas de antiguos jugadores, para pedirle que acabe de construirlo. Le prometen que vendrán a jugar a su campo de béisbol. Muchas organizaciones caen en la misma trampa. Dedican su esfuerzo al hardware y a recoger cantidades masivas de datos. Luego invierten considerablemente en programas para analizar esos clústeres enormes. Su sueño es que con grandes inversiones en hardware y software el conocimiento que obtengan será mayor. Si lo construyen, vendrán. Tiene sentido que se sientan así, porque muchas organizaciones tienen experiencia en concretar proyectos de hardware exitosos. Saben hacerlo. La mayoría de las organizaciones tiene claro cómo hacerlo porque hace décadas que se dedican a ello. En cambio, la ciencia de datos es nueva. No es tan fácil decidir que se invierta dinero en investigar y preguntar. No estás aumentando la capacidad operativa. En su lugar debes cambiar completamente la perspectiva. El hardware es concreto, es tangible, se ve lo que compras. La exploración es más difícil de cuantificar. No tiene un ROI que encaje perfectamente en una cartera de proyectos. Puede ser volátil. Solo demuestra su valor una vez que ya lo has hecho. La Biblioteca del Congreso de los Estados Unidos lanzó un proyecto para recolectar 170 000 millones de tuits. Querían demostrar que podían trabajar en ciencia de datos. Compraron servidores y las computadoras para almacenar los tuits, pero no tenían ningún plan para toda esa información. Tampoco podían conceder permisos para acceder a esos datos. Habían pensado que si lo creaban, ya vendrían. Es una lástima, pero ahí está, ociosa, entre cientos de servidores. Es un monumento a la recogida de datos. Podría ser un caso excepcional, pero es bastante normal. Las organizaciones se centran en cimentar su capacidad. Se ponen unos objetivos con cierta cantidad de nodos en sus clústeres Hadoop. También dan importancia al uso de un conjunto de herramientas de visualización. El presupuesto se dedica a la estructura y los programas. No queda casi nada disponible para el equipo de ciencia de datos. Una vez trabajé para una organización que intentaba reemplazar su almacén de datos con un clúster de big data. Ya dedicaban millones a los servidores del almacén de datos, así que contrataban expertos para mantener esa inversión. Cuando se pasaron a Hadoop, lo hicieron con la misma perspectiva. Desarrollaron un proyecto multimillonario para crear tres clústeres separados. Todo el presupuesto se destinó a los servidores y los programas. Al cabo de dos años, tenían los tres clústeres, pero casi nadie había accedido a esos datos. Para empeorarlo, esas pocas personas trabajaban a distancia, cada una en un área funcional distinta. Tenían millones en sistemas y programas, pero ningún equipo de data science. Tras un par de años de proyecto, el clúster almacenaba unos pocos terabytes de información, lo mismo que se puede guardar en un disco rígido de precio económico. Un par de personas creaba informes para uno o dos departamentos. Vale la pena recordar un par de detalles para evitar esta equivocación. Lo primero es que los equipos de data science son exploratorios. Observan los datos para extraer algún tipo de conocimiento. Los datos no son el producto, es lo que se pueden entender con ellos. No hay ningún premio por tener un clúster enorme. Aunque el equipo pase la mayor parte del tiempo recogiendo datos, no significa que todo el valor provenga del material en crudo. Tener un cuchillo profesional no te convierte en chef profesional. Una colección enorme de datos no hace a un equipo de data science. Lo que lo constituye son las preguntas que sepa hacer. El método científico es lo que lo distingue. También debes recordad que la mayoría de los equipos de data science usan varios programas. A veces prefiero usar R en lugar de Python. Es más fácil para almacenar pocos datos en una base de datos relacional. También pueden cambiar de herramienta de visualización. Dales la oportunidad de que investiguen. Un equipo de data science puede conseguir más con varias herramientas gratuitas que con una sola que requiera una gran inversión. El equipo debería construir sus propias herramientas, también. Un buen equipo de data science siempre será caótico. Pondrán en uso varias herramientas y técnicas para discutir con sus datos y limpiarlos. En lugar de invertir en sistemas y programas, hazlo en formación y experiencia. Recuerda que la parte más importante de tu equipo de data science son las personas que hacen las preguntas.

Los equipos de data science buenos son caóticos.

- **Próximos pasos para aprender data science**

- Hemos puesto en perspectiva el término «data science». Seguro que no te cuesta encontrar quien esté de acuerdo con la importancia de la ciencia de datos. Es mucho más difícil encontrar a alguien que comparta la definición de científico de datos. A lo largo de este curso, he incluido varios enlaces a ejemplos de visualización de datos. Fueron creados con R y Python. Tienes más ejemplos e información en mi libro Data Science: Create Teams That Ask the Right Questions and Deliver Real Value. Está disponible en todas las librerías digitales, publicado por Apress. Espero que hayas disfrutado de este curso para entender la ciencia de datos. Puedes seguirme en LinkedIn, donde comparto más ejemplos de visualizaciones y estrategias que puedes probar con tu propio equipo. Gracias por tu compañía y buena suerte en el futuro.

Minería de datos

- **Minería de datos para científicos de datos**
- La minería de datos combina estadísticas, probabilidad, inteligencia artificial, aprendizaje automático y tecnologías de base de datos en un sistema experto que tiene como materia prima los conjuntos de datos. Los analistas de datos trabajan con cualquiera de estos procesos o con todos ellos, apoyándose en herramientas muy diversas. Emplear estas tecnologías resulta muy beneficioso para las empresas de hoy en día, que tratan con gran cantidad de información de calidad que proviene de fuentes muy variadas, incluidos sitios web, aplicaciones empresariales, redes sociales y dispositivos móviles de todo tipo. Aprende las bases de estos procesos de minería, eleva tu nivel de conocimiento para que puedas asumir el rol de científico de datos que tanto necesita tu empresa y tu entorno de negocio. Trabajando con estas tecnologías, podrás crear modelos que predigan las ventas y el comportamiento futuro a partir de valores existentes o modelos que ayuden a eliminar actividades fraudulentas que puedan perjudicar a tu empresa. Soy Ana María Bisbé, consultora y formadora Business Intelligence, MVP y Microsoft Partner en Power BI, y me gustaría que me acompañaras en este curso de LinkedIn Learning que te ayudará a convertirte en un buen analista de datos capaz de interpretar los datos que ves y, por tanto, serás capaz de identificar en ellos patrones, anomalías, así como las necesidades de limpieza y transformación. ¿Empezamos?

- **Necesidad de explorar datos**
- Ante un proceso de creación de informes analíticos predictivos, el primer paso siempre es la obtención de datos externos. Los datos pueden venir de orígenes muy diversos, y su calidad y estructura puede coincidir o no con las necesidades del escenario que se analiza. Por eso es muy importante profundizar en los datos nada más empezar el proceso. Es posible utilizar técnicas de filtrado, agrupación, ordenado, comprobación de valores nulos y extremos. Según el lenguaje de consulta o herramienta, habrá unas funcionalidades u otras, pero el objetivo será siempre el mismo: conocer el dato antes de tratarlo. La minería de datos es un sistema experto que aprende de la experiencia de datos almacenados. No es posible avanzar en las tareas de limpieza y tratamiento de datos si no se conocen previamente. Por eso los datos son el medio o la base para llegar a conclusiones y transformar los datos originales para convertirlos en información relevante para que las empresas puedan abarcar mejores soluciones que les ayuden a conseguir sus objetivos.

- **Mecanismos de exploración de datos**
- La preparación de datos incluye la exploración, limpieza y configuración de los datos para el proceso de minería de datos. Existen herramientas para investigar las anomalías de los datos. Al explorar datos, necesitamos obtener una vista previa de los mismos y recopilar información estadística que nos resultará útil para la limpieza o el diseño de la fase de modelado de datos. Para explorar valores discretos, es decir, conocer el tipo y la cantidad de datos, es posible utilizar un gráfico de barras que agrupa los valores y muestra el número de casos para cada valor. Para valores numéricos continuos, puede ser mediante un recuento de puntos de datos discretos o un gráfico con la distribución de valores numéricos. La primera acción importante relativa a la limpieza de datos es quitar valores atípicos. Para detectar valores atípicos se suelen aplicar sofisticados análisis de patrones a los datos para determinar si los distintos valores se corresponden con el comportamiento del resto de la muestra o no. Un valor atípico hace referencia a un valor de datos que es problemático por alguno de los motivos siguientes. El valor está fuera de intervalo. Es posible que los datos se hayan especificado de forma incorrecta. El valor es un valor ausente, es un espacio u otra cadena de tipo "null". El valor puede sesgar significativamente la distribución del modelo. Si los valores son erróneos o se salen demasiado del intervalo esperado, podrían indicar la presencia de errores que deberían corregirse, o podría tratarse de valores verdaderos que no obstante podrían afectar a los resultados del análisis. Para descartar valores atípicos, se pueden combinar gráficos de líneas o barras con segmentadores numéricos que van mostrando el efecto que se produce en los datos a medida que se eliminan los extremos. Una variante puede ser rellenar desde el ejemplo, es decir, detectar los valores ausentes, eliminar las filas o reemplazar estos ausentes por un valor promedio, extremo u otro valor. Otro mecanismo es quitar valores atípicos. Puede ser mediante una representación gráfica de la distribución de los valores para quitar los valores extremos. La segunda tarea relevante a destacar en la limpieza de datos es ajustar etiquetas de datos. Cambiar etiquetas de datos nos permite trabajar con los valores de datos y hacer que sean más fáciles de leer y comprender, así como asignar una etiqueta de grupo a intervalos de datos continuos. Se puede tratar además de sustituir códigos numéricos o abreviaturas por textos descriptivos.

- **Qué es la minería de datos**
- La minería de datos combina estadísticas, probabilidad, inteligencia artificial, aprendizaje automático y tecnologías de bases de datos en un sistema experto que tiene como materia prima los conjuntos de datos. Las organizaciones actuales están reuniendo volúmenes cada vez mayores de información de todo tipo de fuentes, incluidos sitios webs, aplicaciones empresariales, redes sociales, dispositivos móviles y en los últimos tiempos datos que provienen del Internet de las cosas. En este campo de la estadística se han recuperado algunas técnicas matemáticas y estadísticas clásicas que han tenido que ser actualizadas para adaptarse al tratamiento de grandes muestras de datos y a los requisitos del procesamiento automático de la información. La minería de datos deriva patrones y tendencias que existen en los datos. Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos que se aplica luego al conjunto definitivo de datos a evaluar. Permite descubrir información que no esperábamos obtener. Esto se debe a su funcionamiento con algoritmos, que admiten la validación de muchas combinaciones distintas de datos. Las personas que se dedican al análisis de datos a través de este sistema son conocidos como mineros o exploradores de datos. Estos intentan descubrir patrones en medio de enormes cantidades de datos. Su intención es la de aportar información valiosa a las empresas para así ayudarlas en la toma de decisiones futuras. Los análisis de datos mediante la minería de datos pueden aportar numerosas ventajas a las empresas para la optimización de su gestión y tiempo y también para la captación y fidelización de clientes, que les permitirá aumentar sus ventas.

- **Objetivos de la minería de datos**
- La minería de datos es un proceso para detectar información de grandes conjuntos de datos de la manera más automática posible. Su principal finalidad es explorar, mediante la utilización de distintas técnicas y tecnologías, bases de datos enormes de manera automática con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos que se han ido recopilando con el tiempo en un contexto específico. La minería de datos surgió con el objetivo de ayudar a analizar y permitir la comprensión de una enorme cantidad de datos y que estos pudieran ser utilizados para extraer conclusiones y así contribuir a la mejora y crecimiento de las empresas. Con este fin, hace uso de prácticas estadísticas y en algunos casos de algoritmos de búsquedas próximos a la inteligencia artificial y a las redes neuronales. Se encarga de explorar los datos, encontrar patrones o reglas que expliquen el comportamiento de los datos en un determinado contexto y realizar predicciones a partir de los datos originales y su combinación con los algoritmos. Desde el punto de vista de la empresa, la minería de datos tiene como objetivo ayudar a mejorar la atención al cliente a partir de la información obtenida y ofrecer a los clientes los productos o los servicios que se necesitan. antes de usar los modelos, estos son comprobados mediante estadísticas para verificar que las predicciones obtenidas son válidas. El correcto uso de estos modelos permite ahorrar costes a la empresa y abrir nuevas oportunidades de negocio. Hay que tener en cuenta que para alcanzar los objetivos pueden existir dificultades que dependen entre otras cosas de la calidad, cantidad o tipo de datos que se quieran recopilar. Esta tarea puede ser lenta y costosa. La inversión inicial para obtener las tecnologías necesarias para la recopilación de datos puede tener un coste elevado, por eso es tan importante centrarse desde el mismo inicio en la calidad y la estructura de los datos.

- **Escenarios de la minería de datos**
- La minería de datos se encuentra en muchas de las esferas de nuestra vida cotidiana, dentro y fuera de la actividad empresarial. Un ejemplo es el escenario meteorológico. Existen predicciones sobre el estado del tiempo desde hace muchísimos años, y desde entonces se podía predecir con cierta exactitud y alguna incertidumbre también el tiempo que iba a hacer a futuro. A día de hoy, es posible predecir situaciones complejas relativas a tormentas, movimientos de masas de aire o evoluciones de las temperaturas. Otro ejemplo es el mundo "online". Una de las mayores aplicaciones de la minería de datos es precisamente este entorno. Se trata de un entorno global, sin fronteras, en el que participan y, por tanto, aportan datos una gran cantidad de personas con atributos muy diferentes. La evaluación mediante procesos de minería de datos de las acciones, características, preferencias y patrones de compra de estas personas permiten ser más eficientes a la hora de ofrecer en décimas de segundos un anuncio, una promoción o un producto en concreto en base a lo que se está consultando o comprando. Este aspecto se vincula con la publicidad contextual de Google, que incluye análisis de idiomas y contenido de cada página en concreto para realizar la publicidad en el mismo idioma y en dependencia de las palabras que considera claves en el contenido. Se apoya en estos y otros elementos de minería de datos para hacer un análisis de la compra y proponer venta cruzada. Es decir, si estás viendo la ficha de un producto, te recomienda otro, y combina los precios, tiene en cuenta el histórico de productos, por ejemplo, libros que otros usuarios han comprado juntos en el pasado. Los analizadores de Google te ofrecen además ofertas personalizadas teniendo en cuenta las búsquedas que has realizado. Los anuncios que te ofrece la página están personalizados para tus gustos, siendo así inmensamente más efectivos. Conmigo suelen acertar los algoritmos. Las grandes superficies también son escenarios de minería de datos. Aplican análisis de cesta de la compra para definir la correcta ubicación de los productos en la tienda y para proponer, muchas veces mediante tiques de descuento, productos en los que podemos estar interesados. Si nos vamos al mundo empresarial, vemos que empresas de todo tipo pueden utilizar la minería de datos para analizar los patrones de ventas y así determinar de forma más eficiente cuándo se venderá un producto en particular y cuál es la expectativa de los beneficios. Además, utilizan estos procesos para crear campañas de marketing y publicidad dirigidas a identificar los factores que mejor predicen a los clientes que tienen más probabilidades de comprar un producto determinado y orientar la campaña a esas personas. Las empresas manufactureras también resultan un escenario interesante, ya que pueden utilizar la minería de datos para buscar patrones en el proceso de producción, de modo que puedan identificar con precisión los cuellos de botella y los métodos defectuosos y encontrar formas de aumentar la eficiencia. También pueden aplicar conocimientos de minería de datos al diseño de productos y hacer ajustes basados en la retroalimentación de las experiencias del cliente. Hay sectores más específicos que presentan escenarios más concretos, por ejemplo, el sector bancario. Las aplicaciones en banca son innumerables, desde la fidelización de clientes hasta la prevención del fraude, pasando por la creación de nuevos productos. Sin lugar a dudas, es el sector que más rendimiento le ha sacado a este tipo de técnicas, debido entre otras cosas a la calidad de los datos que manejan sobre los atributos y preferencias de los clientes. Esto les permite lanzar productos al mercado para un destinatario perfectamente definido y acotado, lo que hace que la inversión sea menor en coste y mayor en certeza. Los sectores bancarios y de seguros son escenarios claros para la evaluación de riesgos. El uso de técnicas de minería de datos a partir de valores resultantes de reclamaciones anteriores permite evaluar la probabilidad de que una reclamación sea fraudulenta. Las agencias de calificación crediticia pueden usar datos financieros y datos del historial del cliente para predecir qué clientes tienen más probabilidades de incumplir con un préstamo. El sector bursátil se puede beneficiar de la aplicación de redes neuronales para la predicción de series temporales, que resulta muy útil para los mercados financieros. La evolución de los mercados depende de multitud de variables, muchas de las cuales son difíciles de cuantificar, por lo que es difícil encontrar situaciones en las que únicamente un análisis numérico de los datos históricos permita realizar buenas predicciones.

- **Mitos acerca de la minería de datos**
- Uno de los principales obstáculos de los proyectos de minería de datos es el desconocimiento que existe sobre esta temática y el no reconocimiento de sus beneficios. Uno de los mitos más frecuentes está relacionado con la cantidad de datos necesaria para un proceso de minería de datos. No se trata de la cantidad en sí, se trata de la calidad del conjunto que permita definir un patrón que se pueda procesar luego con el modelo de minería de datos. Si el conjunto de datos tiene pocos miles de filas y no se destaca un patrón, serán necesarios más datos, pero si se tiene muchos millones y tampoco se deriva un patrón, será inútil. En cada escenario hay que hacer lo que se pueda con lo que se tenga y no hacer cumplir el mito de que se necesita una estructura de CRM o ERP para poder iniciar un proyecto de ese tipo. A mejor organización, mejor calidad, sin dudas. Hay un mito relativo a la calidad de datos que hay que tratar con cautela. Se trata de no iniciar un proceso porque los datos no son exactos o hay datos ausentes. Hay técnicas para rellenar los datos ausentes. Hay forma de limpiar los datos lo más posible. Los modelos estadísticos trabajan con grandes números tratando de identificar patrones genéricos. Si la cantidad de datos sucios es relativamente pequeña, podemos desarrollar tareas de limpieza y minería de datos. Seguramente estos datos no van a influir en el resultado del modelo. Uno de los mitos más frecuentes al que nos enfrentamos antes de iniciar un proyecto de minería de datos es que el cliente piensa que desarrollar un modelo lleva mucho tiempo, es muy complejo y es muy caro. Los algoritmos son complejos. La necesidad de contar con un experto es absolutamente real, pero a día de hoy se han democratizado tanto estos procesos que se puede avanzar mucho con ellos sin que represente un alto coste de tiempo y recursos. A día de hoy, estos procesos son accesibles a empresas de todo tipo y tamaño. Hay otro mito muy extendido relativo al objetivo principal del proyecto. cada uno en su escenario y cada uno con sus objetivos particulares. No se trata de hacer magia, se trata de utilizar la minería de datos para lo que ha sido diseñada, con los datos para identificar los patrones, con los algoritmos adecuados para cada caso, bien entrenados los modelos, dándoles el tiempo que requieren para su aprendizaje y luego poder interpretar los resultados y aplicar las acciones necesarias en cada negocio o escenario.

- **Implicaciones tecnológicas en proyectos de minería de datos**
- Hay un grupo de aspectos tecnológicos a tener en cuenta antes y durante el proyecto de creación de estructuras y modelos de minería de datos. Va a depender también de las necesidades de cada caso. Elementos como requerir inmediatez en la respuesta del modelo, el volumen y la complejidad de los datos van a impulsar la necesaria inversión en tecnología que garantice la mayor velocidad y exactitud posible de las predicciones a realizar. Habrá que hacer un estudio de rendimiento de consultas a datos, tiempos de procesado y recursos necesarios para los procesos de minería que afecten a otros procesos que se realicen en los servidores de la entidad. De este estudio se derivará la inversión necesaria para conseguir dar respuesta en los plazos adecuados. El proyecto de implementación de estructuras y modelos de minería de datos debe tener en cuenta el día después de la obtención de los resultados del modelo. ¿Qué hacemos ahora con estos resultados? ¿Cómo implementamos las mejoras en el proceso y validamos su efectividad? en este caso se trata además de pensar en el equipamiento periférico para el resultado final, por ejemplo impresoras, lectores de código y ese tipo de accesorios. Será necesario adaptar los sistemas de puntos de venta para que la venta contemple la consulta al sistema de previsión y reciba las órdenes de impresión de dichos tiques. Una vez puesto en marcha el sistema, hay que volver a controlar los recursos, especialmente el almacenamiento y tiempos de procesado. Aun cuando hayamos previsto el crecimiento, hay que comprobarlo día a día. A veces los problemas ocurren porque se dimensionó para un conjunto de usuarios para los que funciona bien y luego se implantó para muchos más sin que se evaluara este redimensionamiento. Desde el punto de vista de las personas involucradas en las acciones posproceso de minería de datos, hay que formarlas, garantizar que la información resultante se presente de forma clara y eficiente y prever la curva de aprendizaje y el rechazo que pueda existir al aplicar nuevos procedimientos. Otro aspecto relativo a las personas tiene que ver con los datos que se consumen de las propias personas y la seguridad de los mismos. Hay que garantizar todos los procesos de seguridad de la información. Hay que garantizar todos los medios técnicos para la entrada y salida de información. Las formas modernas de datos también requieren nuevos tipos de tecnologías, como reunir conjuntos de datos de una variedad de entornos informáticos distribuidos, y para datos más complejos como imágenes y vídeos, datos temporales y datos espaciales. Los datos que se extraen deben ser completos, precisos y confiables, y para garantizarlo hay que vencer todos y cada uno de los retos tecnológicos que se van presentando.

- Test de capítulo

- ¿A qué se dedican los especialistas en minería de datos? A descubrir patrones en medio de enormes cantidades de datos.

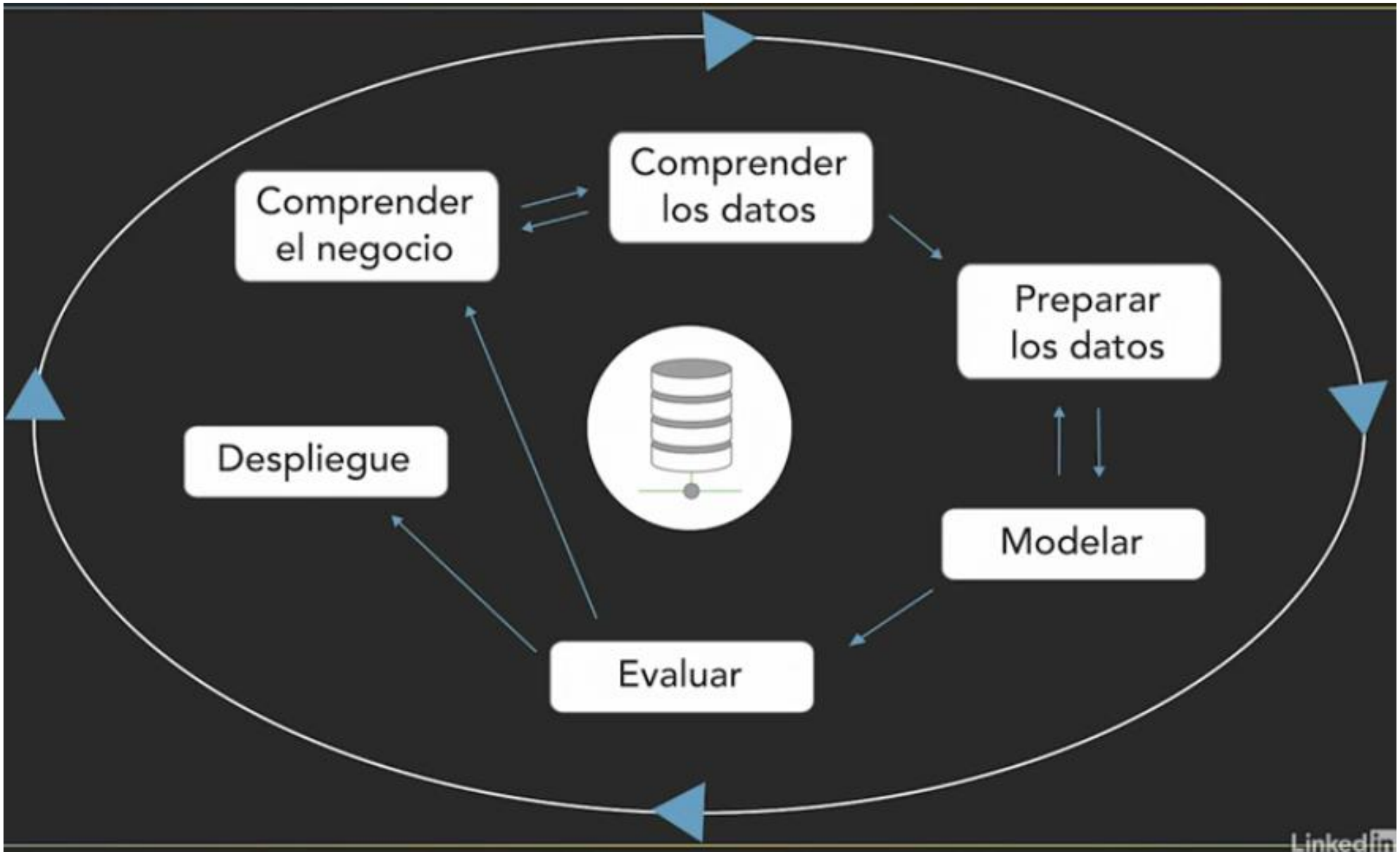
- ¿Cuál es el primer paso en el proceso de creación de análisis predictivos? la obtención de datos externos

- ¿Cuál de los siguientes no es un ejemplo de un valor atípico? Valor dentro de intervalo.

- ¿Cuál de las siguientes opciones no es un objetivo de la minería de datos? Optimizar el funcionamiento de las bases de datos. ¡Correcto! La minería de datos se enfoca en explorar datos para encontrar patrones repetitivos y realizar predicciones que ayuden a una empresa. Para optimizar el funcionamiento de una base de datos se usan otras herramientas.

- Para garantizar la mayor velocidad y exactitud posible de las predicciones a realizar hay que invetir en tecnología

- **Modelo CRISP en minería de datos**
- El modelo CRISP, que significa en inglés "Cross Industry Standards Process", refleja el ciclo interno del proceso de minería de datos. La fase de transformación comienza con el conocimiento y entendimiento tanto del negocio como de los datos. La fase de comprensión del negocio se centra en la comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición del problema de minería de datos y un plan preliminar diseñado para alcanzar los objetivos. Por su parte, la comprensión de datos comienza con una colección inicial de datos y procesos con actividades, con el objetivo de familiarizarse con los datos, identificar la calidad de los problemas para descubrir las primeras señales dentro de los datos y detectar temas interesantes para poder formular hipótesis de información oculta. Luego se preparan los datos y los modelos. Durante la preparación de datos, se realizan las tareas de selección y transformación de tablas, registros y atributos y, además, la limpieza de datos para las herramientas de modelado. Mientras que en la fase de modelado se seleccionan y aplican varias técnicas de modelado y se calibran los parámetros para obtener óptimos resultados. Se utilizan diferentes modelos con diferentes parámetros. Luego se evalúan y, en dependencia de los resultados, van a producción o se vuelve a comenzar el ciclo. En la etapa de evaluación se comparan los modelos y se adopta el de mayor calidad y fiabilidad desde la perspectiva de análisis de datos. Y, para finalizar, la fase de despliegue depende de los requerimientos, pudiendo ser simple como la generación de un informe o compleja como la implementación de un proceso de explotación de información que atraviese a toda la organización. Un proceso de minería de datos continúa después del despliegue de la solución. Las lecciones aprendidas durante el proceso pueden provocar nuevas preguntas de negocio. Unos procesos de minería de datos se beneficiarán de la experiencia de los anteriores. Existe el diagrama de proceso que muestra la relación entre las diferentes fases de CRISP- DM. Las flechas en el diagrama indican las dependencias más importantes y frecuentes entre las fases. El círculo exterior en el diagrama simboliza la naturaleza cíclica de la minería de datos en sí.


Comprender el negocio
Comprender los datos
Preparar los datos
Modelar
Evaluar
Despliegue
LinkedIn

- **Tipos de algoritmos de minería de datos**
- Existen distintos tipos de algoritmos que se pueden utilizar en procesos de minería de datos. Hay dos grandes grupos: directos o supervisados e indirectos. Los algoritmos directos o supervisados cuentan con una variable a predecir que define el funcionamiento del algoritmo. Se pueden dividir en clasificación, estimación y previsión. El conocimiento se utiliza para clasificar los casos y la estimación de valores continuos. Para definir qué tipo de algoritmo se necesita hay que encontrar la pregunta a responder en cada modelo. Es un algoritmo de clasificación si nos estamos preguntando ¿es un ejemplo de definición de tipo A o B?, ¿o quizás A, B o C? Si, en cambio, nos estamos preguntando por cuánto crecerá un determinado valor en dependencia de la combinación de otras variables, hablamos de una regresión. Por su parte, en los algoritmos indirectos o no supervisados se buscan patrones entre los datos. Entre ellos destacan los de agrupación por afinidad, cesta de la compra, "clustering" y descripción. Hablamos en este caso de estadística descriptiva. Para este grupo de algoritmos también se cumple que para definir qué tipo de algoritmo se necesita hay que encontrar la pregunta a responder en cada modelo. Si nos preguntamos ¿cuál es la estructura de los datos?, ¿cómo se pueden agrupar?, vamos a necesitar algoritmos tipo "clustering" o "K-Means". Si nos interesa obtener respuesta a ¿es este un caso inusual?, entonces se trata de un algoritmo de detección de anomalías. Y si lo que queremos es responder a la pregunta ¿alguna recomendación?, estaremos necesitando algoritmos de reglas de asociación. La elección del algoritmo es un paso muy importante en el proceso de minería de datos.

- **Pautas ante un proyecto minería de datos**
- Veamos algunas de las pautas a tener en cuenta cuando nos disponemos a realizar un proyecto de minería de datos. Analizar nuestras necesidades de información significa tener en cuenta el volumen de datos con el que contamos, e ir valorando la posibilidad de crecimiento en el futuro. El propio proyecto creará más datos y más consumo en servidores y repositorios de datos. Hay que evaluar a su vez el posible consumo de los modelos para conseguir dimensionar el proyecto de la forma adecuada en la actualidad y con vistas a futuro, ajustarse a lo estrictamente necesario, no invertir tiempo y esfuerzos en implantar algo que no se necesite de verdad. Una vez asumida su necesidad e implementado el proceso, hay que reevaluar su necesidad manteniendo una visión a medio y largo plazo. No hay que implantar sistemas que no se necesiten. Además de caros, serán un obstáculo y una hipoteca para futuros crecimientos. Garantizar la calidad del proceso. No es posible informatizar el caos. En su lugar, si hay un caos en los datos hay que revisar procedimientos, rutinas y la forma de trabajar. Podremos crear nuevas estructuras para poner todo en orden y buscar en todo esto poder desarrollar un proyecto fiable y que nos ayude además a recortar costes. Aprovechar al máximo los recursos de los que se disponen. Contratar un proyecto o un sistema más caro no va a hacer que aumente la rentabilidad de la gestión. A veces ya tenemos las herramientas necesarias, solo que no sabemos utilizarlas. Vencer las barreras del desconocimiento de la materia, sin complejos técnicos por no ser un experto. El conocimiento del negocio, el conocimiento de los datos y las necesidades del entorno son un excelente recurso para animarse a adentrarse en el apasionante mundo de la minería de datos. A esto se une la necesidad de pedir asesoramiento externo y contrastar opiniones, quizás evaluando distintos fabricantes. La visión de alguien experto en procesos de minería de datos, aunque sea ajeno a la compañía, introducirá un soplo de aire fresco y una visión más rica del problema. No temer al error en el proceso. La extracción de datos se realiza muchas veces mediante prueba y error. Por eso para los mineros de datos cometer errores es algo natural. Los errores pueden ser valiosos bajo ciertas condiciones porque se puede aprender de los intentos fallidos. Por otra parte, no debemos cometer errores como omitir controles de calidad de los datos ni poner en producción un modelo predictivo que no se haya probado. El control de calidad en la minería de datos es tan importante como en cualquier otro proceso de la vida. No debemos creer que un patrón en los datos prueba una relación de causa y efecto, podría ser apenas una coincidencia de una sola vez. Tampoco hay que asumir más allá de lo que los propios datos revelan y tener en cuenta las condiciones de obtención de la muestra. Y es muy importante saber que no tenemos que trabajar con algoritmos preferidos, no existe un tipo único de modelo de minería de datos que se adapte a cada situación. Hay preguntas que nos debemos hacer antes de empezar el proyecto de minería de datos, sobre el acceso a los datos externos, el tipo, contenido y función a desempeñar dentro del modelo, de las columnas a emplear, los algoritmos más adecuados y la necesidad de conjuntos de pruebas para la validación de los modelos.

- **Etapas de la minería de datos**
- El proceso de minería de datos incluye varias etapas. Primero hay que definir el problema: analizar los requisitos empresariales y definir el ámbito del problema, las métricas por las que se evaluará el modelo y el objetivo final del proyecto de minería de datos. Puede que necesitemos un estudio de disponibilidad de datos para investigar las necesidades de los usuarios de la empresa con respecto a los datos disponibles. Si los datos no son suficientemente buenos para satisfacer los objetivos, puede que debamos volver a definir el proyecto. En segundo lugar, necesitamos seleccionar y preprocesar los datos. Se trata de preparar los datos, filtrarlos, solucionar los problemas de incongruencia y mala calidad en los datos. Lo mejor es utilizar algún método de automatización para estas tareas. A continuación, viene el paso de selección de variables. Se trata de escoger, seleccionar entre los datos ya limpios y depurados aquellas columnas que sean relevantes para el escenario concreto. La selección de características reduce el tamaño de los datos sin sacrificar la calidad del modelo obtenido del proceso de minería. En esta misma fase hay que volver a explorar los datos, realizar cálculos estadísticos sobre los mismos y luego se puede decidir si el conjunto de datos contiene datos con errores y crear una estrategia para solucionar los problemas. Como cuarto paso, debemos extraer conocimiento y generar modelos. Mediante una técnica de minería de datos se obtiene un modelo de conocimiento que representa patrones de comportamiento observados en los valores de las variables del problema o las relaciones de asociación entre dichas variables. Es muy importante en este paso separar aleatoriamente los datos preparados en el conjunto de datos de entrenamiento, que se utilizan para generar el modelo y el conjunto de datos de comprobación, que sirve para comprobar la precisión del modelo mediante la creación de consultas de predicción. Una vez definida la estructura del modelo de minería de datos, es necesario procesarla rellenando la estructura vacía con patrones que describen el modelo. Esto se conoce como "entrenar el modelo". El modelo de minería de datos se define mediante un objeto de estructura de minería de datos u objeto de modelo de minería de datos y un algoritmo de minería de datos. Como quinto paso, tenemos la interpretación y evaluación, es decir, explorar y validar los modelos. Una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. La exploración de un modelo de minería de datos permite comprender el comportamiento del modelo antes de implementarlo. Cada algoritmo devuelve un tipo de resultados diferentes. Es importante comprobar si los modelos crean predicciones correctas mediante herramientas de validación, como el gráfico de elevación y la matriz de clasificación. Estas herramientas requieren los datos de comprobación que antes separamos del conjunto de datos original en el paso de generación del modelo. Y para finalizar, en sexto lugar hay que aplicar los algoritmos, es decir, implementar y actualizar los modelos. La elección del algoritmo apropiado para una tarea empresarial específica puede ser un trabajo difícil, debiendo implementar el modelo o modelos que funcionan mejor en un entorno de producción. Los algoritmos se pueden utilizar de forma combinada. Por ejemplo, utilizar algunos algoritmos para examinar los datos y después usar otros algoritmos para predecir un resultado específico basándose en estos datos. Por ejemplo, utilizar un algoritmo de clústeres que reconoce patrones para dividir los datos en grupos que sean más o menos homogéneos, y luego utilizar los resultados para crear un mejor modelo de árbol de decisión. Asimismo, se pueden utilizar varios algoritmos dentro de una solución para realizar tareas independientes. Por ejemplo, usar un algoritmo de árbol de regresión para obtener información de previsiones financieras y un algoritmo basado en reglas para llevar a cabo un análisis de la cesta de la compra. Los modelos de minería de datos pueden predecir valores, generar resúmenes de datos y buscar correlaciones ocultas. Una vez que los modelos de minería de datos se encuentren en el entorno de producción, es posible llevar a cabo otras tareas dependiendo de sus necesidades, por ejemplo la actualización del modelo, ya que a medida que la organización recibe más datos, debe volver a procesar los modelos para mejorar así su eficacia.

- Test de capítulo
- En esta etapa de la minería de datos se vuelven a explorar los datos, se realizan cálculos estadísticos sobre los mismos y luego se decide si el conjunto de datos contiene errores. ¿Podrías identificarla? : Selección de variables.

- ¿En qué fase del modelo CRISP se realizan las tareas de selección y transformación de tablas, registros y atributos? preparación de datos

- Los algoritmos directos cuentan con una variable a predecir que define el funcionamiento del algoritmo.

- ¿Qué debemos hacer al realizar un proyecto de minería de datos? Ajustarnos a lo estrictamente necesario.

- **De qué está compuesta la minería de datos**
- En los proyectos de minería de datos tenemos que distinguir varios componentes. El primero de todos, la vista de origen de datos, que es el punto de partida, el momento en el que se crea la conexión con el origen de datos externo. La vista es un objeto compartido que permite combinar varios orígenes de datos y usarlos como un origen único. Las propiedades de la vista del origen de datos se pueden utilizar para modificar tipos de datos, crear agregaciones o asignar alias a las columnas. A partir de estos datos, se crean para cada estructura los conjuntos de entrenamientos y de pruebas. Esto significa que una parte de los datos se emplea para entrenar los modelos para que detecten los patrones, y otro grupo se utiliza para comprobar con datos reales la eficiencia de los modelos una vez entrenados. La estructura de minería de datos, como su nombre lo indica, es la colección de columnas y modelos que participan en el proyecto. Es la estructura. Uno de los elementos a distinguir, además, es la tabla de casos que almacena los datos de origen para los modelos de minería de datos. Para cada columna es importante definir el tipo de dato y el tipo de contenido. Los tipos de contenido incluyen discreto, continuo y cíclico. Los tipos de contenido discreto son valores que no forman parte de una secuencia, por ejemplo, número de hijos, números de teléfonos o género. Como vemos, estos valores discretos pueden ser numéricos o no numéricos. Por su parte, los tipos de contenido continuo representan secuencias de datos numéricos en una escala, por ejemplo, de temperatura o peso. Y, por último, los tipos de contenido cíclico representan datos que se organizan en conjuntos limitados y ordenados que se repiten, como pueden ser los días de la semana o los meses del año numerados. Los modelos de minería de datos son las distintas combinaciones de columnas de la estructura. No tienen que ser todas, pero sí tienen que pertenecer a la estructura. Si una estructura tiene varios modelos, pueden utilizar columnas diferentes y pueden asignar valores de uso a las columnas en modelos diferentes. Además, se especifica el valor de uso para cada columna, que es como el tipo de función a realizar para cada columna. Si no se especifica un valor de uso para una columna, significa que no está incluida en el modelo. Los valores de uso disponibles son: llave o clave, que indica que la columna es una columna clave que contiene valores únicos que identifican individualmente a cada fila. Predecible, que indica la columna para la que se desea predecir valores en la extracción. Y, por último, entrada. Se asigna al resto de columnas y le indica al modelo que debe usar esta columna para ayudar a predecir los valores para la columna para la que estamos creando una predicción.

- **Estructuras y modelos de minería de datos**
- La estructura de minería de datos define los datos a partir de los cuales se generan los modelos de minería de datos. Especifica la vista de datos de origen, el número y el tipo de columnas y una partición opcional en conjuntos de entrenamiento y de pruebas. Una misma estructura de minería de datos puede admitir varios modelos de minería que comparten el mismo dominio. La configuración de una estructura de minería de datos consta de los pasos siguientes. Definir un origen de datos, seleccionar las columnas de datos que se van a incluir en la estructura y definir una clave, especificar si los datos de origen se deben separar en un conjunto de entrenamiento y en un conjunto de pruebas (este paso es opcional) y luego procesar la estructura. Una estructura de minería también puede contener tablas anidadas. Una tabla anidada representa una relación de uno a varios entre la entidad de un caso y sus atributos relacionados. Se trata de combinar, por ejemplo, un cliente y sus compras o un producto y sus ventas. Cuando se definen los datos para las estructuras de minería de datos, también es posible especificar que algunos de los datos se usan para entrenamiento y otros para pruebas. La información sobre los conjuntos de datos de pruebas y de entrenamiento se almacena en la memoria caché con la estructura de minería de datos. Por consiguiente, el mismo conjunto de pruebas puede usarse con todos los modelos que están basados en la misma estructura. En las estructuras de minería es posible habilitar la obtención de detalles, lo que significa que se pueden ver los detalles de los casos que componen luego los modelos. Lo que hacemos es incorporar columnas a la estructura que no se van a utilizar en los modelos. Y, por último, hay que procesar la estructura de minería de datos, ya que solo es un contenedor de metadatos hasta que se procesa. No es necesario volver a procesar la estructura cada vez que se agrega un nuevo modelo. Se puede procesar solamente el modelo. La estructura y el modelo de minería de datos son objetos independientes. La estructura de minería de datos almacena la información que define el origen de datos. Un modelo de minería almacena la información derivada del procesamiento estadístico de los datos, como los patrones encontrados como resultado del análisis. El modelo de minería de datos recibe los datos de una estructura y los analiza utilizando un algoritmo de minería de datos. El modelo está vacío hasta que los datos que proporciona la estructura de minería se procesan y analizan. Una vez procesado el modelo, contiene los metadatos, resultados y enlaces de la estructura de minería de datos.

- **Estructuras de minería de datos: tipos de datos**
- Las columnas de la estructura de minería de datos se agregan a la misma y definen el modo en el que un modelo de minería utiliza las columnas de un origen de datos. Las columnas de la estructura se diseñan para ser flexibles y extensibles, porque cada algoritmo que se utilice para generar un modelo de minería puede utilizar diferentes columnas de la estructura para interpretar los datos. Las columnas tienen un conjunto de propiedades que ayudan en su definición y son: identificador, nombre, tipo, contenido, distribución y marcadores de modelado. Cuando se agregan columnas a una estructura de minería de datos, también se debe definir el tipo de datos para dichas columnas, como texto, entero o numérico largo. Si los datos se pueden administrar como texto o como datos numéricos, el tipo de datos especificado afectará de forma importante a la hora de procesar los datos por parte del algoritmo. Los tipos de datos y los tipos de contenido que definen las columnas de la estructura se derivan del origen de datos que se utiliza para crear la estructura. Es posible cambiar esta configuración en la estructura de minería de datos y establecer indicadores de modelado y la distribución de columnas continuas. Las propiedades que se configuran para la estructura de minería se propagan a todos los modelos de minería asociados con esa estructura. Los tipos de datos indican al motor de minería de datos si los datos del origen son numéricos o de texto y cómo deben procesarse los datos, por ejemplo especificar si los números deben tratarse como enteros o utilizando posiciones decimales. Los tipos existentes son: text, long, boolean, double y date. Cada tipo de datos admite uno o varios tipos de contenido, que es la forma en que se procesan o se calculan los datos de la columna. En el modelo de minería, los datos numéricos se pueden tratar como números o como texto. Y si es un número, puede tener asociado un tipo de contenido discreto o continuo. Si cambia el tipo de datos de una columna, es necesario volver a procesar siempre la estructura de minería y todos los modelos de minería basados en esa estructura. En algunas ocasiones, al cambiar el tipo de datos la columna ya no se utilizará en el modelo en cuestión, por eso hay que reprocesar. También se pueden convertir los datos a un formato que el algoritmo pueda usar. "Discretizar" es convertir números continuos en un conjunto de valores discretos. En principio, todos los números son infinitos y, por tanto, continuos. Cuando se modela la información, puede resultar más eficaz discreti zar los valores disponibles, ya sea agrupándolos en un número de intervalos finito o aproximándolos a una media central o valor representativo. La discretización ofrece ventajas para el análisis, ya que reduce el espacio del problema. Sin embargo, no siempre es conveniente cambiar los números continuos a discretos, y algunos algoritmos funcionan solamente con datos numéricos continuos. Es bueno definir los tipos de datos de las columnas de una estructura de minería de datos para influir en el modo en que los algoritmos procesan sus datos al crear modelos de minería. No obstante, definir los tipos de datos de columnas proporciona a los algoritmos información únicamente sobre el tipo de datos de la columna, pero no acerca del comportamiento de estos datos. Por ese motivo, cada tipo de datos de minería admite uno o más tipos de contenido que puede utilizar para describir el comportamiento de los datos que contiene la columna. Al crear el modelo de minería de datos, es importante asegurarse de que las columnas de los datos contienen el tipo de datos correcto. Algunas columnas pueden incluir tantos valores que quizás el algoritmo no pueda identificar fácilmente patrones en los datos. Además, si los datos son numéricos, debes asegurarte de que son discretos o continuos según corresponda.

- **Contenido en las estructuras de minería de datos**
- Existen varios tipos de contenido, que es la forma en que se procesan o se calculan los datos de la columna en el modelo de minería. Una columna discreta contiene un número finito de valores, ya sea porque la entrada numérica se ha limitado a un conjunto determinado de valores o porque no hay ninguna serie numérica en los datos. Por ejemplo, el texto siempre se trata como valor discreto. Los valores de una columna discreta no implican que los datos estén ordenados. Aunque los valores sean numéricos, los valores discretos están claramente separados y no existe ninguna posibilidad de que se den valores fraccionados. Los códigos telefónicos de una zona son un buen ejemplo de datos numéricos discretos. Este tipo de contenido es compatible con todos los tipos de datos de minería de datos. Los datos numéricos continuos pueden incluir un número de valores fraccionados infinito. Una columna de precios o importes de ventas son ejemplos de columnas de atributos continuos. Una columna continua siempre contiene datos numéricos, representan datos de medidas, incluidas marcas de tiempo. Este tipo de contenido es compatible con los tipos de datos date, double y long. Una columna ha sido discretizada cuando la columna contiene valores que representan grupos o depósitos de valores que se derivan de una columna continua. Los depósitos se tratan como si fueran valores ordenados y discretos. Este tipo de contenido es compatible con los siguientes tipos de datos: date, double, long y text. Las columnas de tipo llave o clave identifican una fila de forma única, normalmente en una tabla de casos. La columna clave es un identificador numérico o de texto. Indica que la columna no debe utilizarse para el análisis, sino para realizar el seguimiento de los registros. Se ajusta a los tipos de datos date, doble, long y text. Las tablas anidadas también tienen claves, pero el uso de la clave de tabla anidada es diferente en una tabla anidada. El tipo de contenido clave indica que la columna es el atributo a analizar. Los valores de la clave de tabla anidada deben ser únicos para cada caso, pero puede haber duplicados en todo el conjunto de casos. En ocasiones, existe el tipo de contenido de clave de secuencia. La columna es un tipo específico de clave donde los valores representan una secuencia de eventos, están ordenados y no tienen que estar separados por una distancia equivalente. Puede ocurrir en tipos de datos double, long text y date. Existe un tipo de contenido de clave de tiempo parecido al anterior. La columna es un tipo específico de clave donde los valores representan valores que están ordenados y que ocurren en una escala de tiempo. Es compatible con los tipos de datos double, long y date. Se identifica un tipo de contenido ordenado si la columna contiene valores que definen un conjunto ordenado. Estas columnas se consideran discretas en términos de tipo de contenido y son compatibles con todos los tipos de datos de minería. Una columna se identifica como cíclica si contiene valores que representan un conjunto ordenado cíclico. Por ejemplo, los días enumerados de la semana son un conjunto ordenado cíclico, ya que el día número uno sigue al día número siete. Las columnas cíclicas se consideran ordenadas y discretas en términos de tipo de contenido. Este tipo de columnas es compatible con todos los tipos de datos de minería. Y, por último, podemos encontrar contenido de tabla, que es cuando una columna contiene una tabla de datos anidada con una o más columnas y una o más filas. Pueden existir varias columnas de tabla anidada en una tabla de casos. Para cada fila de la tabla de casos, la tabla anidada puede contener varios valores, todos ellos relacionados con el registro del caso primario. El tipo de datos de esta columna siempre es table. Algunos de los algoritmos que se utilizan para crear modelos de minería de datos requieren tipos de contenido específicos para funcionar correctamente. Por ejemplo, el algoritmo 'Naive Bayes' de Microsoft no puede utilizar columnas continuas como entrada y no puede predecir valores continuos. Otros algoritmos solo pueden utilizar valores continuos como entrada o no pueden predecir valores discretos. Cuando se crea un modelo de minería de datos, la ambigüedad en los datos puede generar errores o impedir que se utilicen algunos algoritmos. Por tanto, al crear el modelo de minería de datos, se recomienda definir explícitamente si los datos de una columna son discretos, continuos o discretizados, no es posible crear un modelo con datos continuos y después tratar la columna como discreta.

- **Modelos de minería de datos**
- Un modelo de minería de datos aplica un algoritmo de minería que se representa en una estructura de minería. Es un objeto que pertenece a una determinada estructura y hereda todos los valores de las propiedades que están definidas en la misma. El modelo puede utilizar todas las columnas o un subconjunto de las contenidas en la estructura. Es importante realizar un análisis detallado de cada columna de la estructura y agregar únicamente aquellas que tengan sentido en el análisis del modelo. Un modelo de minería de datos se crea mediante la aplicación de un algoritmo a los datos. Es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones. Tras procesar el modelo, almacena metadatos. Los metadatos especifican el nombre del modelo y del servidor donde están almacenados, la definición de las columnas de la estructura de minería que se usaron para generarlo, las definiciones de los filtros que se aplicaron al procesar y el algoritmo empleado para analizar los datos. Todos estos elementos tienen una influencia muy eficaz en los resultados del análisis. Cada tipo de modelo crea un conjunto diferente de patrones, conjuntos de elementos, reglas o fórmulas que se utilizan para crear predicciones. Por lo general, cada algoritmo analiza los datos de una forma diferente, por lo que el contenido del modelo resultante también se organiza en estructuras diferentes. En un tipo de modelo, los datos y los modelos pueden estar agrupados en clústeres, y en otro los datos pueden estar organizados en árboles, ramas y las reglas que los dividen y definen. El modelo también se ve afectado por los datos usados en el entrenamiento, en dependencia de los filtros aplicados. Los datos reales no se almacenan en el modelo, solo se almacenan las estadísticas de resumen con los datos reales que residen en la estructura de minería de datos. Los pasos para crear un modelo de minería son: crear la estructura con las columnas de datos que sean necesarias, seleccionar el algoritmo más adecuado para la tarea analítica, elegir las columnas de la estructura para usar en el modelo y especificar sus propiedades. Opcionalmente, se pueden establecer parámetros para ajustar el procesamiento del algoritmo y luego procesar la estructura y el modelo. Las propiedades de los modelos de minería son nombre, descripción, fecha en que se procesó por última vez, los permisos del modelo y los filtros que se usan en los datos para el entrenamiento. Además, tiene propiedades que se derivan de la estructura de minería de datos y que describen las columnas de datos que usa. Existen además dos propiedades especiales. La propiedad algoritmo sirve para identificar al algoritmo que va a depender de la tarea y del proveedor que se utilice en el proyecto y que es única para el modelo. La propiedad uso define cómo usa el modelo cada columna de forma individual. Los usos disponibles para las columnas que participan en el modelo son columna de entrada, de predicción, solo de predicción o clave. Se indica ignorar si la estructura contiene una columna que no se usa en el modelo. Si se modifica cualquier propiedad, hay que revisar el resto de propiedades para que se adapten bien al modelo, y hay que procesar nuevamente el modelo. Una vez procesado el modelo de minería, contiene una gran cantidad de información sobre los datos y los patrones encontrados mediante el análisis, incluyendo estadísticas, reglas y fórmulas de regresión. Se pueden utilizar los visores personalizados para examinar esta información o se pueden crear consultas de minería de datos para recuperarlas y usarlas para el análisis y la presentación.

Pasos para crear un modelo

- Crear la estructura de minería de datos
- Seleccionar el algoritmo más adecuado
- Elegir las que se van a usar en el modelo
- Definir las propiedades de las columnas
- Establecer parámetros
- Procesar la estructura y el modelo

- Test de capitulo
- ¿Qué tipo de columna contiene valores fraccionados infinitos? Continua
- La propiedad de uso define cómo usa el modelo cada columna de forma individual.
- ¿Cuáles son los tipos de contenido que representan datos numéricos en secuencia? Contenido continuo
- ¿Qué puede admitir una misma estructura de datos? Admitir varios modelos de minería que comparten el mismo dominio.
- Algunos tipos de datos existentes en minería de datos son: text, long, boolean, double, date

- **Introducción a algoritmos de minería de datos con Microsoft**
- El algoritmo de minería de datos es el mecanismo que crea modelos de minería de datos. Para crear un modelo, un algoritmo analiza primero un conjunto de datos buscando tendencias y patrones específicos. Después, el algoritmo utiliza los resultados de este análisis para definir los parámetros del modelo de minería. A continuación, estos parámetros se aplican en todo el conjunto de datos para extraer patrones procesables y estadísticas detalladas. La elección del mejor algoritmo para una tarea analítica puede ser un desafío. Se pueden utilizar diferentes algoritmos para realizar la misma tarea. cada uno de ellos genera un resultado diferente, y algunos pueden generar más de un resultado. En la minería de datos se incluyen los siguientes tipos de algoritmos. Algoritmos de clasificación, que predicen una o más variables discretas basándose en los demás atributos del conjunto de datos, por ejemplo árbol de decisión, red neuronal y Naive Bayes. Algoritmos de regresión, que predicen una o más variables numéricas continuas como pérdidas o ganancias, basándose en otros atributos del conjunto de datos. Son los casos de los algoritmos de series temporales, regresión lineal y regresión logística. Algoritmos de segmentación o "clustering", que dividen los datos en grupos de elementos que tienen propiedades similares. Algoritmos de asociación, que buscan correlaciones entre diferentes atributos de un conjunto de datos para la creación de reglas de asociación. Se trata del algoritmo del mismo nombre: asociación. Algoritmos de análisis de secuencias, que crean un resumen de las secuencias frecuentes o episodios en los datos, como una serie de clics en un sitio web. Esto se realiza mediante el algoritmo de clústeres de secuencia. Al seleccionar el tipo de algoritmo, ayuda a enfocarse en el tipo de tarea a realizar. Por ejemplo, para predecir un atributo discreto en una tarea como es clasificar la evolución de los pacientes y explorar los factores relacionados, los algoritmos que se pueden utilizar son árboles de decisión, Bayes Naive, clústeres y redes neuronales. Si lo que necesitamos predecir es una secuencia, algo como realizar un análisis de flujos de clics, lo más adecuado es utilizar algoritmos de clústeres de secuencia. Si el objetivo es buscar grupos de elementos comunes en las transacciones para sugerir a un cliente la compra de productos adicionales, se pueden utilizar algoritmos de asociación o de árboles de decisión. Y, por último, si queremos buscar grupos de elementos similares, por ejemplo, crear grupos de pacientes con perfiles de riesgo, en función de atributos como datos geográficos, demográficos y distintos comportamientos, utilizamos algoritmos de clústeres o clústeres de secuencia.

- **Algoritmo clústering para segmentación de datos**
- El algoritmo de "clustering" es del tipo segmentación. Estos algoritmos dividen los datos en grupos o clústeres de elementos que tienen propiedades similares que no se podrían derivar lógicamente a través de la observación casual. Se utiliza para identificar agrupamientos naturales de casos basados en un conjunto de atributos. Los casos dentro del mismo grupo tienen valores de atributos más o menos similares. La agrupación es una tarea de minería de datos no supervisada. Ningún atributo único se utiliza para guiar el proceso de capacitación. Todos los atributos de entrada son tratados por igual. La mayoría de los algoritmos de "clustering" construyen el modelo a través de una serie de iteraciones y se detienen cuando el modelo converge, es decir, cuando los límites de estos segmentos están estabilizados. El algoritmo de clústeres entrena el modelo de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo. Mediante la observación de las columnas que forman un clúster, podemos ver con mayor claridad la forma en que los registros de un conjunto de datos se relacionan entre sí. Algunos ejemplos son predecir un atributo discreto, marcar los clientes de una lista de posibles compradores como clientes con buenas o malas perspectivas. Calcular la probabilidad de que un servidor genere un error en los próximos, digamos, seis meses. Clasificar la evolución de los pacientes y explorar los factores relacionados. Buscar grupos de elementos similares. Crear grupos de pacientes con perfiles de riesgo en función de atributos como datos demográficos y comportamientos. Analizar usuarios mediante patrones de búsqueda y compra de productos. Identificar servidores con características de uso similares. Los requisitos para un modelo de agrupación en clústeres son los siguientes. Una columna de clave única. Única columna numérica o de texto. No se admiten columnas compuestas, que identifica cada registro de manera única. Columnas de entrada, al menos una columna de entrada y tantas como sea necesario con los valores que se utilizan para generar los clústeres. Y, además, una columna de predicción que es opcional. El algoritmo no necesita una columna de predicción para generar el modelo, es opcional agregar esta columna de casi cualquier tipo de datos. Los valores de la columna de predicción se pueden tratar como entradas al modelo, o se puede especificar que solo se utilicen para las predicciones. Por ejemplo, si deseas predecir los ingresos del cliente agrupando en clústeres de acuerdo con datos demográficos como la región o la edad, debes especificar los ingresos como "solo de predicción", no sería una columna de entrada, y luego agregar todas las demás columnas como columnas de entrada. El algoritmo de clústeres de Microsoft identifica primero las relaciones de un conjunto de datos y genera una serie de clústeres basándose en ellas. Un gráfico de dispersión es una forma útil de representar visualmente el modo en el que el algoritmo agrupa los datos. El gráfico de dispersión representa todos los casos del conjunto de datos. Cada caso es un punto del gráfico. Los clústeres agrupan los puntos del gráfico e ilustran las relaciones que identifica el algoritmo. después de definir los clústeres, el algoritmo calcula el grado de perfección con que los clústeres representan las agrupaciones de puntos. Y a continuación intenta volver a definir las agrupaciones para crear clústeres que representen mejor los datos. El algoritmo establece una iteración en este proceso hasta que ya no es posible mejorar los resultados mediante la redefinición de los grupos.

- **Algoritmo de reglas de asociación**
- Los algoritmos de asociación buscan correlaciones entre diferentes atributos de un conjunto de datos. Devuelven reglas que describen cómo se agrupan los productos en una transacción y las probabilidades de que dichos productos se adquieran juntos. En términos de asociación, cada producto o, mejor dicho, cada par de atributo-valor se considera un artículo. La tarea de asociación tiene dos objetivos, encontrar conjuntos de elementos frecuentes y encontrar reglas de asociación. El parámetro umbral de frecuencia o soporte sirve para que el modelo analice solo los artículos que aparecen en él, al menos en ese valor de los carritos de la compra. Por ejemplo, si tenemos X e Y y representan dos elementos que pueden formar parte de la cesta de la compra, el parámetro de soporte es el número de casos del conjunto de datos que contienen la combinación de ambos elementos. De esta forma, el algoritmo controla el número de conjuntos de elementos que se generan. El parámetro de probabilidad, también denominado parámetro de confianza, representa la fracción de casos del conjunto de datos que contiene X y que también contiene Y. De esta forma, el algoritmo controla el número de reglas que se generan. La aplicación más común de esta clase de algoritmos es la creación de reglas de asociación que pueden utilizarse en un análisis de cesta de la compra para identificar aquellos productos que se venden a menudo en la misma cesta y las reglas, con el propósito de realizar ventas cruzadas. Se puede utilizar el algoritmo de asociación para identificar los conjuntos de productos que suelen adquirirse juntos. Así, se pueden predecir los elementos adicionales en los que un cliente puede estar interesado basándose en aquellos que ya se encuentran en su cesta. Además de identificar los conjuntos de elementos frecuentes basados en el soporte, los algoritmos de tipo asociación encuentran reglas. Una regla de asociación tiene la forma "Si tengo A y B, entonces con toda probabilidad querré C". Los requisitos para un modelo de reglas de asociación son los siguientes. Una columna de clave única, única columna numérica o de texto, y no se admiten columnas compuestas. Esta columna identifica cada registro de manera única. Columnas de entrada, que deben ser discretas. Los datos de entrada de un modelo de asociación suelen encontrarse en dos tablas. Por ejemplo, una tabla puede contener la información del cliente y la otra, las compras de ese cliente. Es posible incluir estos datos en el modelo mediante el uso de una tabla anidada. Una única columna de predicción, un modelo de asociación solo puede tener una columna de predicción. Normalmente, se trata de la columna de clave de tabla anidada, como el campo que contiene los productos que se han comprado. Los valores deben ser discretos o discretizados. El algoritmo de asociación suele usarse para los motores de recomendación. Un motor de recomendación indica los elementos a los clientes basándose en aquellos que ya han adquirido. Los modelos de asociación se generan basándose en conjuntos de datos que contienen identificadores para casos individuales y para los elementos que contienen los casos. Un grupo de elementos de un caso se denomina un conjunto de elementos. Un modelo de asociación se compone de una serie de conjuntos de elementos y de las reglas que describen cómo estos elementos se agrupan dentro de los casos. Las reglas que el algoritmo identifica pueden utilizarse para predecir las probables compras de un cliente en el futuro basándose en los elementos existentes en la cesta de la compra actual del cliente.

- **Algoritmo para el análisis de secuencias**
- El análisis de secuencia se utiliza para encontrar patrones en una serie discreta. Una secuencia se compone de una serie de valores o estados discretos. Por ejemplo, una secuencia de clics en la web. Las compras de los clientes también se pueden modelar como datos de secuencia. El cliente compra un producto y luego otro, en ese orden. Un caso individual contiene un conjunto de elementos o estados. Los modelos de secuencia analizan las transiciones de estado. Con el modelo de secuencia "comprar el producto A antes que el producto B", indica claramente una secuencia diferente a "comprar el producto B antes que el producto A". El análisis de secuencia es una tarea de minería de datos relativamente nueva. Es cada vez más importante debido principalmente a dos tipos de aplicaciones: análisis de registro web y análisis de ADN. Hay varias técnicas de secuencias disponibles en la actualidad, como las cadenas de Markov. Los investigadores están explorando activamente nuevos algoritmos en ese campo. El algoritmo de clústeres de secuencia combina el análisis secuencial con la agrupación en clústeres. Se utiliza para explorar datos que contienen eventos que pueden vincularse con rutas o secuencias. Encuentra las secuencias más comunes y realiza una agrupación en clústeres para buscar secuencias que sean similares. Encuentra clústeres de casos que contienen rutas similares en una secuencia. El algoritmo examina todas las probabilidades de transacción y mide las diferencias o las distancias entre todas las posibles secuencias del conjunto de datos con el fin de determinar qué secuencia es mejor utilizar como entradas para la agrupación en clústeres. Los requisitos de un modelo de agrupación en clústeres de secuencias son los siguientes. Una columna de clave única, es decir, una clave que identifique los registros. Una columna de secuencia para los datos de la secuencia, el modelo debe tener una tabla anidada que contenga una columna de identificador de secuencia. El identificador de secuencia puede ser cualquier tipo de dato ordenable. Solo se admite un identificador de secuencia por cada secuencia y un tipo de secuencia en cada modelo. Un modelo de agrupación de clústeres de secuencia podría incluir atributos opcionales no relacionados con la secuencia. El algoritmo admite la incorporación de otros atributos que no tengan que ver con las secuencias. Estos atributos pueden incluir las columnas anidadas. Un modelo de agrupación en clústeres de secuencia podría incluir información de los pedidos como tabla de casos, datos demográficos sobre el cliente concreto de cada pedido, como atributos no relacionados con la secuencia y una tabla anidada que contenga la secuencia que siguió al cliente al examinar el sitio o colocar los artículos en el carro de la compra, como información de la secuencia. Mediante el uso de algoritmos de clústeres de secuencia, una empresa puede encontrar grupos de clientes que tienen patrones o secuencias de clics similares. La empresa puede usar estos grupos para analizar la forma en que los clientes se mueven por el sitio web, identificar qué páginas se relacionan más estrechamente con la venta de un producto en particular y predecir las páginas que tienen mayores probabilidades de ser visitadas a continuación. Otros ejemplos de uso son analizar los registros que enumeran eventos que preceden a un incidente, como errores de disco duro o inter bloqueos del servidor, o los registros que siguen las interacciones de un cliente o un paciente a lo largo del tiempo para predecir cancelaciones de servicio u otros resultados satisfactorios.

- Test de capítulo
- ¿Cuáles son los algoritmos que predicen una o más variables numéricas continuas como pérdidas o ganancias, con base en otros atributos del conjunto de datos? Correcto! Por ejemplo, los algoritmos de series temporales, regresión lineal y regresión logística.
- ¿Cuál de las siguientes opciones no es un requisito para un modelo de agrupación en clústeres? admite columnas compuestas

- ¿Qué algoritmo combina el análisis secuencial con la agrupación en clústeres? análisis de secuencias
- ¿Qué algoritmo se suele usar para los motores de recomendación? Algoritmo de asociación. ¡Correcto! Estos algoritmos buscan correlaciones entre diferentes atributos de un conjunto de datos.

- **Algoritmo de árbol de decisión**
- El algoritmo de árboles de decisión es un algoritmo de clasificación y regresión para el modelado de predicción de atributos discretos y continuos. Para atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores conocidos como estados de estas columnas para predecir los estados de aquella columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. Por esa razón, se puede utilizar el algoritmo de árbol de decisión no solo para la predicción, sino también como una forma de reducir el número de columnas de un conjunto de datos, ya que el árbol de decisión puede identificar las columnas que no afectan al modelo de minería de datos final. Puede mostrarse mediante un histograma. Algunos ejemplos de uso pueden ser marcar los clientes de una lista de posibles compradores como clientes con buenas o malas perspectivas, calcular la probabilidad de que un servidor genere un error en el futuro, clasificar la evolución de los pacientes y explorar los factores relacionados. Para atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión. en este caso cada nodo contiene una fórmula de regresión. En estos casos algunos ejemplos son pronosticar las ventas del año próximo, predecir los visitantes de un sitio a partir de las tendencias históricas y estacionales proporcionadas, generar una puntuación de riesgo a partir de datos demográficos. Si se define más de una columna como elemento de predicción, o si los datos de entrada contienen una tabla anidada que se haya establecido como elemento de predicción, el algoritmo genera un árbol de decisión independiente para cada columna de predicción. Ejemplos de uso. Usar el análisis de la cesta de la compra para determinar la posición del grupo. Sugerir a un cliente la compra de productos adicionales. También sirve para analizar los datos de una encuesta a los visitantes de un evento, y también para descubrir qué actividades estaban correlacionadas, con el fin de programar entonces actividades futuras. El algoritmo de árboles de decisión genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta. A medida que el algoritmo agrega nuevos nodos a un modelo, se forma una estructura en árbol. El nodo superior del árbol describe el desglose de la columna de predicción para toda la muestra, que puede ser la población global de clientes. A medida que el modelo crece, el algoritmo considera todas las columnas. Este algoritmo utiliza la selección de características para guiar la selección de los atributos más útiles. Esta selección de características es importante para evitar que los atributos irrelevantes utilicen tiempo del procesador. Los requisitos para un modelo de árbol de decisión son los siguientes. Una columna de una única clave. Cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas. Una columna de predicción. Se requiere al menos una columna de predicción, puede incluir varios atributos de predicción en un modelo y pueden ser de tipos diferentes: numéricos o discretos. Sin embargo, el incremento de números de atributos de predicción puede aumentar el tiempo de procesamiento. También columnas de entrada. Se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entradas también afecta el tiempo de procesamiento. Una vez procesado el modelo, los resultados se almacenan como un conjunto de patrones y estadísticas que se pueden usar para explorar las relaciones o para realizar predicciones.

- **Algoritmo de redes neuronales**
- El algoritmo de red neuronal prueba cada posible estado del atributo de entrada con cada posible estado del atributo de predicción y calcula las probabilidades de cada combinación según los datos de aprendizaje. Es posible utilizar estas probabilidades para tareas de clasificación o regresión, así como para predecir un resultado en función de algunos atributos de entrada. También se puede utilizar una red neuronal para el análisis de asociación. El número de redes creadas por el algoritmo en un modelo de minería de datos depende del número de estados o valores del atributo de las columnas de entrada, así como del número de columnas de predicción que utiliza el modelo de minería y el número de estados de dichas columnas. Este algoritmo es útil para analizar datos de entrada complejos, como los datos de un proceso comercial o de producción, o problemas empresariales para los que hay una cantidad importante de datos de entrenamiento disponibles pero en los que no es fácil derivar reglas mediante otros algoritmos. El modelo de red neuronal debe contener una columna de clave, una o más columnas de entrada y una o más columnas de predicción. Los casos sugeridos para utilizar el algoritmo de red neuronal son análisis de comercialización y promoción, como por ejemplo medir el éxito de una promoción por correo directo o una campaña publicitaria en la radio. Predecir los movimientos de las acciones, la fluctuación de la moneda u otra información financiera con gran número de cambios a partir de los datos históricos. Analizar los procesos industriales y de producción. Y, además, trabajar la minería de texto. Cualquier modelo de predicción que analice relaciones complejas entre muchas entradas y relativamente pocas salidas será un buen caso para trabajar estos algoritmos de redes neuronales.

- **Algoritmo Naïve Bayes**
- El algoritmo Bayes Naïve es un algoritmo de clasificación basado en los teoremas de Bayes y se puede utilizar para el modelado de predicción y exploración. La palabra "naïve", ingenua, en inglés, del término Bayes Naïve, proviene del hecho que el algoritmo utiliza técnicas bayesianas, pero no tiene en cuenta las dependencias que puedan existir. Calcula la probabilidad de cada estado de cada columna de entrada, dado cada posible estado de la columna de predicción. Es muy útil para generar rápidamente modelos de minería de datos que detecten las relaciones entre las columnas de entrada y las columnas de predicción. Puedes utilizar este algoritmo para realizar la exploración inicial de los datos y más adelante aplicar los resultados en otros modelos de minería. Mediante este algoritmo, un departamento de comercialización puede predecir un resultado para el perfil de un cliente concreto, y, por tanto, puede determinar qué clientes responderán con más probabilidad según sea el producto a comprar. Los requisitos para este tipo de modelo son los siguientes. Una columna de una sola clave, una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas. Necesita columnas de entrada, todas las columnas deben ser discretas o se deben haber discretizado los valores. Por otra parte, las variables deben ser independientes. Los atributos de entrada deben ser independientes unos de otros. Esto es particularmente importante al utilizar el modelo para la predicción. Para dos columnas de datos que ya están estrechamente relacionadas, el efecto sería multiplicar la influencia de estas dos columnas, lo que pudiese ocultar otros factores que influyen en el resultado. Al contrario, la capacidad del algoritmo de identificar las correlaciones entre las variables es útil cuando está explorando un modelo o conjunto de datos para identificar las relaciones entre las entradas. Necesita al menos una columna de predicción. El atributo de predicción debe contener valores discretos o discretizados. Los valores de la columna predecible se pueden tratar como entradas. Es interesante utilizar este algoritmo si se explora un nuevo conjunto de datos para encontrar posibles relaciones entre las columnas.

- Test de capítulo
- ¿Cuál es el algoritmo que identifica las columnas de entrada que se correlacionan con la columna de predicción? árbol de decision

- ¿Cuál es el algoritmo que prueba cada posible estado del atributo de entrada con cada posible estado del atributo de predicción? redes neuronales

- **Algoritmo de serie de tiempos**
- Un algoritmo de serie temporal proporciona varios algoritmos optimizados para la previsión en el tiempo de valores continuos. Estos modelos no necesitan columnas adicionales como información de entrada para predecir una tendencia. Un modelo de serie temporal puede predecir tendencias basadas únicamente en el conjunto de datos original utilizado para crear el modelo. También es posible agregar nuevos datos al modelo al realizar una predicción e incorporar automáticamente los nuevos datos en el análisis de tendencias. A la combinación de datos de origen y datos de la predicción se le denomina serie. El algoritmo de serie temporal es capaz de realizar predicciones cruzadas. Al entrenar al algoritmo con dos series independientes pero relacionadas, se puede utilizar el modelo generado para predecir el resultado de una serie basándose en el comportamiento de la otra. Por ejemplo, las ventas observadas de un producto pueden influir en las ventas previstas de otro producto. De forma predeterminada, el algoritmo de serie temporal utiliza una mezcla de dos algoritmos al analizar patrones y realizar predicciones. El algoritmo entrena dos modelos independientes sobre los mismos datos. Uno de los modelos usa el algoritmo ARTXP y obtienen mejores resultados en las predicciones a corto plazo. Y el otro modelo usa el algoritmo ARIMA optimizado para la predicción a largo plazo. Es posible también controlar la mezcla de algoritmos para favorecer la predicción a corto o a largo plazo en las series temporales. Ambos algoritmos pueden detectar estacionalidad en los datos en varios niveles. Por ejemplo, tus datos podrían contener ciclos mensuales anidados en ciclos anuales. Para detectar estos ciclos estacionales, es posible proporcionar una sugerencia de periodicidad o bien especificar que el algoritmo deberá detectar automáticamente la periodicidad. Además de la periodicidad, hay otros parámetros que controlan el comportamiento del algoritmo de serie temporal. Cuando este detecta la periodicidad, realiza predicciones o analiza casos. Cada modelo de predicción debe contener una serie de casos, que es la columna que especifica los intervalos de tiempo u otras series sobre las que se produce el cambio. Un modelo de serie temporal debe utilizar siempre una fecha, una hora o algún otro valor numérico único para su serie de escenarios. A continuación, el algoritmo combina los resultados de dos modelos para obtener la mejor predicción sobre un número variable de intervalos de tiempo. La predicción cruzada también es útil para crear un modelo general que se puede aplicar a múltiples series. Por ejemplo, digamos que las predicciones para una serie determinada (una región, digamos) son inestables debido a que la serie no dispone de datos de buena calidad. Es posible entonces entrenar un modelo general sobre la medida de todas las series, una para cada una de las regiones que haya y, a continuación, aplicar el mejor modelo a las series individuales para crear predicciones más estables para cada una de ellas, es decir, para cada región. Al utilizar el algoritmo de serie temporal en los datos históricos de los últimos años, una empresa puede crear un modelo de minería de datos que prevea la venta futura de determinados productos. Además, la organización puede llevar a cabo predicciones cruzadas para ver si las tendencias de ventas de tipos específicos del producto están relacionadas. Algunos casos de uso para este algoritmo cuya función es predecir un atributo continuo pueden ser pronosticar las ventas para el año próximo, predecir los visitantes a un sitio a partir de tendencias históricas o generar una puntuación de riesgo a partir de datos demográficos. Los requisitos para un modelo de serie temporal son los siguientes. Una única columna de clave de tiempo. Columna numérica o de fecha que se utilizará como serie de casos y que define los intervalos de tiempo que utilizará el modelo. La columna debe contener valores continuos y estos deben ser únicos para cada serie. La serie no puede estar almacenada en dos columnas, por ejemplo una columna Año y una columna Mes. Una columna predecible que debe contener valores continuos alrededor de la que el algoritmo generará el modelo de serie temporal. Y una columna clave de serie opcional con valores únicos que identifiquen a una serie.

- **Algoritmo de regresión lineal**
- El algoritmo de regresión lineal ayuda a calcular una relación lineal entre una variable independiente y otra dependiente y, a continuación, utilizar esta relación para la predicción. La relación toma la forma de una ecuación para la línea que mejor represente una serie de datos. Hay otros tipos de regresión que utilizan varias variables, y también hay métodos no lineales de regresión. Sin embargo, la regresión lineal es un método útil y conocido para modelar una respuesta a un cambio de algún factor subyacente. Se puede utilizar la regresión lineal para determinar una relación entre dos columnas continuas, por ejemplo, para calcular una línea de tendencias en los datos de fabricación o ventas o para el desarrollo de modelos de minería más complejos, con el fin de evaluar las relaciones entre las columnas de datos. Con el algoritmo de regresión lineal se calculan y se prueban automáticamente todas las posibles relaciones entre las variables. La regresión lineal podría simplificar el exceso de relaciones en escenarios en los que varios factores afecten al resultado. Al seleccionar el algoritmo de regresión lineal, se invoca un caso especial del algoritmo de árboles de decisión con parámetros que restringen el comportamiento del algoritmo y requieren ciertos tipos de datos de entrada. Además, en un modelo de regresión lineal el conjunto de datos completo se utiliza para calcular las relaciones en el paso inicial. Los requisitos para este tipo de modelo son los siguientes. Una columna de una sola clave, cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas. Una columna de predicción, se requiere al menos una columna de predicción. Se pueden incluir varios atributos de predicción en un modelo, pero deben ser de tipos de datos numéricos continuos, no se puede utilizar un tipo de datos de fecha y hora como atributo de predicción, aunque el almacenamiento nativo para estos datos sea numérico. Las columnas de entrada deben contener datos numéricos continuos y se les debe asignar el tipo de datos adecuado. En un modelo de regresión lineal el contenido incluye metadatos. La fórmula de regresión y estadísticas sobre la distribución de los valores de entrada. Una vez procesado el modelo, los resultados se almacenan como un conjunto de estadísticas junto con la fórmula de regresión lineal, que se puede utilizar para calcular tendencias futuras. Además de crear un modelo de regresión lineal seleccionando el algoritmo de regresión lineal, si el atributo de predicción es un tipo de datos numérico continuo, también se puede crear un modelo de árbol de decisión que contenga regresiones. En este caso, el algoritmo dividirá los datos cuando encuentre puntos de separación adecuados, pero en cambio creará una fórmula de regresión para algunas regiones de datos.

- **Algoritmo de regresión logística**
- La regresión logística es una técnica estadística conocida que se usa para modelar los resultados binarios. Es un método estadístico que se utiliza para determinar la contribución de varios factores a un par de resultados. Se puede utilizar una red neuronal modificada para modelar las relaciones entre las entradas y los resultados. Se mide el efecto de cada entrada en el resultado y se ponderan las diversas entradas en el modelo acabado. El nombre regresión logística procede del hecho de que la curva de los datos se comprime mediante una transformación logística para minimizar el efecto de los valores extremos. Los requisitos para un modelo de regresión logística son los siguientes. Una columna de una sola clave, es una columna numérica o de texto que identifica cada registro de manera única. No están permitidas las claves compuestas. Se necesita como mínimo una columna de entrada que contenga los valores que se utilizan como factores en el análisis, puede tener tantas columnas de entrada como se desee. Al menos necesita una columna de predicción de cualquier tipo de datos, incluidos datos numéricos continuos. Los valores de la columna de predicción también se pueden tratar como valores de entrada al modelo o se puede especificar que solo se utilicen para las predicciones. No se admiten tablas anidadas en las columnas de predicción, pero se pueden usar como entradas. Una de las ventajas de la regresión logística es que el algoritmo es muy flexible, puede tomar cualquier tipo de entrada y admite varias tareas analíticas diferentes. Los casos de uso que se apliquen pueden ser por ejemplo usar datos demográficos para realizar predicciones sobre resultados, el riesgo de contraer una determinada enfermedad, explorar y ponderar los factores que contribuyen a un resultado, digamos, buscar los factores que influyen en que los clientes vuelvan a visitar un establecimiento, o clasificar los documentos, el correo electrónico u otros objetos que tengan muchos atributos.

- Test de capítulo
- ¿Qué algoritmo proporciona varios algoritmos optimizados para la previsión en el tiempo de valores continuos? Series de tiempo
- ¿Cuál es el algoritmo que ayuda a calcular una relación lineal entre una variable independiente y otra dependiente? Regresión lineal
- ¿Qué algoritmo se considera como un método estadístico que se utiliza para determinar la contribución de varios factores a un par de resultados? Regresión logística

- **El análisis de datos en el mundo de la empresa**

Si seleccionas líneas de la transcripción en esta sección, irás a la marca de tiempo en el vídeo El análisis de datos es el proceso de inspección, limpieza, transformación y modelado de datos para la toma de decisiones empresariales. Los analistas de datos trabajan con cualquiera de estos procesos o con todos ellos, apoyándose en herramientas muy diversas. En la actualidad, dada la cantidad y variedad de los datos con los que trabajamos a diario, ha aumentado la necesidad de que los analistas recopilen y procesen la información de forma eficiente. El análisis de datos es crucial en todas las industrias. Aprende cuál es tu papel dentro de este escenario, y prepárate para cumplirlo. Con este curso de LinkedIn Learning, puedes llegar a descubrir si ya eres un analista, o convertirte en analista de datos. Aprenderás a mejorar considerablemente la calidad de los análisis que realizas aplicando las pautas y sugerencias que veremos. Soy Ana María Bisbé, consultora y formadora Business Intelligence, Microsoft Data Platform MVP, y Microsoft Partner en Power BI, y me gustaría que me acompañaras en este viaje que te ayudará a convertirte en un buen analista de datos, capaz de interpretar los datos que ves, y por tanto, serás capaz de identificar en ellos patrones, anomalías, así como las necesidades de limpieza y transformación.

¿Empezamos?

Definición de análisis de datos y analista de datos

Hoy en día existe una creciente necesidad de realizar análisis de datos, por lo que se necesitan muchos analistas de datos. Entonces, ¿qué es el análisis de datos? Hay muchas definiciones, pero en última instancia es el proceso de inspección, limpieza, transformación y modelado de datos para la toma de decisiones empresariales. Entonces, puedes preguntarte lo obvio: ¿qué es un analista de datos? Los analistas de datos trabajan con cualquiera de estos procesos o con todos ellos. Yo estoy trabajando constantemente con mi equipo, mostrándole los resultados de estos procesos, o simplemente configurando datos para la toma de decisiones que debe ocurrir en cada proyecto. El trabajo de los analistas de datos es tan diverso como las descripciones de los puestos de trabajo que encontrarás en la web si haces una búsqueda entre las ofertas que existen. Los analistas de datos trabajan con datos que se presentan en formas muy diversas, y las herramientas son tan diversas como los datos que encontramos. Los analistas de datos a menudo pueden comenzar con filas de base de datos, hojas de cálculo, o incluso, archivos CSV. Convierten estos registros en resultados más significativos para que otros los interpreten. Recuerda, una imagen vale más que mil palabras, o en el caso de los datos, un gráfico vale más que mil líneas. Los analistas de datos tienen la gran suerte de poder crear a veces excelentes visualizaciones de datos. ¿Por qué se necesitan analistas? Cada vez disponemos de más datos, vivimos en el mundo del big data, nuevos estilos, tipos y orígenes de datos, incluyendo los dispositivos electrónicos que generan y almacenan gran cantidad de información. La demanda supera la oferta, se necesitan más personas con habilidades analíticas profundas, personas que sepan cómo afrontar los cambios en tratamiento de datos. Se necesitan, en resumen, analistas de datos con formación que les permita desarrollar su trabajo con buena calidad y eficiencia.

- **Descubrir si ya eres analista**
- Podrías comenzar preguntándote si ya eres un analista de datos. Yo creo que hay muchas personas que ya son analistas, pero simplemente no lo saben, porque no tienen un título que lo acredite, pero realizan procesos de análisis de datos todo el tiempo. Puedes comenzar a identificarte como analista de datos si estás todo el tiempo sumergido en los datos, si creas gráficos, si vives en hojas de cálculo, si encuentras la forma de mejorar los procesos a través de los datos, o simplemente parece que puedes identificar patrones con mayor facilidad que otros. Para mí, este trabajo es una bendición. Casi todos los proyectos de datos necesitarán algún tipo de limpieza de datos para prepararlos para los informes; a veces tu trabajo puede ser, sencillamente, limpiar los datos. Todos los analistas de datos deberían aprender las mejores formas de limpiar adecuadamente los datos. Como analista, es posible que diseñes informes, y compartas tus descubrimientos iniciales con tu equipo, y en ocasiones, será necesario realizar un análisis más profundo. Los datos no siempre nos llegan de la manera que lo necesitamos. Los proyectos de datos requieren una limpieza de datos; esa es una de las razones por las que somos tan necesarios los analistas. La limpieza es una parte activa en la mayoría de los proyectos de datos. En algún momento descubrirás que hay personas y herramientas que tienen procesos completos basados en cuestiones relativas a la calidad y a la limpieza de datos. Ahora, es importante entender que la limpieza de datos no significa cambiar los valores, significa remodelar los datos, como, por ejemplo, separar o combinar columnas de nombres y apellidos y ordenar el resultado. Agregar cálculos que no existen, para enriquecer de esta forma el modelo, o eliminar columnas y filas que no son necesarias para el análisis. La visualización de datos puede servir al analista de datos con dos propósitos distintos. Por una parte, puede mostrarle de forma inmediata y visual al analista, los problemas existentes en sus datos, y por otra parte, a las organizaciones a las que servimos, puede ofrecerles mejores formas de interpretar o visualizar sus datos. Un gráfico de barras, por ejemplo, con un valor alto o bajo, le da a quien toma las decisiones un excelente punto de partida. Como analista de datos, cada conjunto de datos que encuentres te brinda la oportunidad de ofrecer a la organización una mejor herramienta para la toma de decisiones.

- **Entender los roles en el análisis de datos**
- La ciencia de datos tiene una definición muy amplia. Creo que siempre será un objetivo móvil, porque los datos evolucionan, las herramientas que utilizamos evolucionan y las habilidades requeridas también evolucionan. En mi opinión, la escasez que tenemos hoy de personas capacitadas para ejercer como científicos de datos se debe a que se relaciona la palabra "datos" con la palabra "ciencia". Y esto puede provocar que haya personas que se sientan descalificadas de inmediato. Puede que ya estés preparado o que estés transitando el camino hacia la ciencia de datos. Si es el caso, hay que mantenerse en este camino y seguir avanzando. Por otra parte, puede ser que tu camino no pase por la ciencia de datos. Tampoco pasa nada, lo importante es tener claras las bases necesarias para ser analista de datos. O quizás ir más allá, y ser científico de datos. Un analista de datos es una pieza clave para cualquier equipo de ciencia de datos; y digo más, es una pieza clave para cualquier equipo por muy multidisciplinario que sea. Los analistas de datos a veces trabajan de forma independiente, o puede encontrarse en un equipo que cuenta con otros analistas, investigadores, estadísticos y expertos de negocio. He tenido la oportunidad de realizar análisis de datos en muchos sectores de la economía y la vida. De todos ellos, recuerdo con mucho cariño proyectos realizados en colaboración con personas pertenecientes a equipos médicos, técnicos del sector turístico y equipos que están directamente integrados en la confección textil. Es increíblemente bueno tener la oportunidad de trabajar con personas con estas habilidades. Aprendí muchísimo de ellas, y al mismo tiempo, pude aportar mi experiencia como analista para dar respuesta a sus necesidades. Es posible que tu rol requiera que traigas datos al equipo, pero podría expandirse fácilmente, y que tu función sea, además, hacer análisis iniciales sobre los datos que has traído, o incluso, crear las visualizaciones que sustentan el análisis que has realizado. Puedes encontrar que tu equipo compartirá algunas habilidades comunes entre sí, pero también tendrán sus propios datos, sus propias habilidades y pensamientos únicos como aporte individual de cada uno. Un buen equipo trabaja de forma coordinada hacia el mismo objetivo; analizar, construir medidas y brindar información que pueda mejorar los resultados de la entidad que solicita el análisis.

- **Descubrir las habilidades del analista de datos**
- Vamos a hablar sobre las habilidades de los analistas de datos y el punto del que partimos todos ante un análisis. Y nos preguntamos entonces, ¿cómo se definen los datos? He trabajado con datos de todas las formas, tamaños y formatos, y una cosa es coherente, siempre varía según la organización y los objetivos de cada proyecto. Una de las razones es que los datos los definen las personas de manera muy diferente. Realmente varía de persona a persona. ¿En qué es en lo primero que piensa una persona cuando escucha la palabra "datos"? La respuesta es muy variada, en dependencia de cada persona, su formación, conocimientos y experiencias. Es clave tener en cuenta que debido a que se definen de forma diferente, entonces, los roles en la ciencia de la información también se definen de forma diferente. Hay roles grandes y pequeños en la ciencia de la información y no podemos olvidar los roles que aún no existen. A medida que se va desarrollando la tecnología y el ser humano es cada vez más innovador, los datos continúan creciendo, por lo que la oferta y la demanda seguirán aumentando y los roles estarán en un estado de constante cambio. También es importante recordar que cada organización definirá su rol y sus herramientas. Algunos de los roles que podemos encontrar en la ciencia de la información son: análisis, investigación, gobierno de datos, ciencia de datos y gestión de proyectos. Las herramientas con las que se trabaja en la ciencia de datos están mejorando constantemente. Como parte de ese desarrollo y crecimiento se crean nuevas herramientas para tratar con los datos de forma más eficiente. Dependiendo del tamaño de los datos, su complejidad y lo que estemos haciendo con ellos, se determinará qué herramienta podemos y debemos utilizar. Y sobre las habilidades relacionadas con la técnica de la información, puedo decir que dependiendo de la función a realizar por el analista, así será el requisito de las habilidades técnicas con las que deba contar, los conjuntos de datos estarán cambiando con el tiempo, por lo que surgirán nuevos conjuntos de datos y las habilidades también necesitarán cambiar con el tiempo. Ciertamente, no hay una lista exacta, pero si recién estás comenzando o si descubres que ya eres un analista en sus inicios, necesitas saber acerca de la minería de datos y las consultas a orígenes de datos. Es importante que sepas cómo modelar los datos y entenderás mejor cómo son, por ejemplo, las bases de datos si sabes cómo diseñar un modelo. Y nunca está de más entender los diagramas de flujos de trabajo y, por supuesto, siempre necesitarás dominar las técnicas de la visualización. No dejes de aprender sobre las herramientas de datos para determinar dónde encajas, cuál es tu perfil y en algún momento, estarás listo para sumergirte y aumentar tu experiencia en un área o en una herramienta determinada.

- Test de capítulo
- Hay dos cosas que necesitas saber si estás comenzando o si eres un analista en sus inicios: consultas a orígenes de datos y minería de datos
- De los roles de la lista inferior, uno NO se corresponde a la ciencia de la información. ¿Podrías decir cuál? Recopilación de la información

- **Aprender a identificar los datos**
- Aprender a identificar los datos es una habilidad fundamental si vas a trabajar con datos, y especialmente si vas a construir sistemas de datos. Hay que ver los datos, especialmente, cuando no es obvio. Los datos realmente están en todas partes y casi siempre hay más allá de lo que inicialmente se piensa. Comencemos con un ejemplo sencillo, una silla. Piensa en qué atributos podemos identificar cuando vemos una silla. ¿Distinguimos que tiene un color?, ¿identificamos de qué material está hecha? Si vamos más allá y pensamos en que además de ser simplemente una silla es un mueble, obtendríamos aún más datos. Están ahí, solo tenemos que profundizar más para verlos. ¿Es un mueble de interiores o de exteriores?, ¿está diseñada nuestra silla para un sitio específico de la casa, oficina u otro lugar, como puede ser un hospital? Solo tenemos que mirarlo con nuestra lupa para descubrir datos. La silla tiene además dimensiones, como altura y peso; los valores de estos atributos afectan, por ejemplo, a la forma y el coste de envío, y esto nos lleva aún más lejos. Nos lleva a los proveedores de sillas. ¿Dónde están?, ¿dónde se ubican esos proveedores?, ¿cuántos proveedores tenemos? y ¿cuántas sillas tienen en stock? ¿Has tenido esto en cuenta para definir, por ejemplo, el número de serie? A este objeto se le pueden asociar otros atributos, como pueden ser las fechas. ¿Qué pasa con todas las fechas relacionadas con la silla? Por ejemplo, el día en que se ordenó, la fecha en que se envió, la fecha en que se entregó. Una vez más, identificar los datos que no son tan obvios, puede ser de extraordinaria ayuda en todos los procesos y todos los proyectos de datos, y muy especialmente cuando comienzas tu carrera como analista. Prueba tu lupa de datos con los elementos con los que trabajas ahora, y prueba a ver cuántos datos identificas. Esto te ayudará a hacer mejores preguntas en las reuniones donde se definen y se describan los datos.

- **Aprender sobre campos y tipos de datos**
- Es importante entender cómo se ven los datos. Los datos se ven en dependencia de quién los está mirando. La mayoría de las personas que no son analistas pensarán en los valores de los datos, pero un analista ve los datos y piensa en tipos de datos, campos y valores. Entonces, ¿qué es un tipo de datos? El tipo de datos define la estructura de dato que tienen los valores que se almacenan en una columna o campo de la base de datos. Por ejemplo, una columna con tipo de datos de fecha, contendrá solo fecha. Si el tipo de dato es numérico, contendrá solo números. Los tipos de datos también definen las acciones que se pueden realizar con los valores de datos de este tipo en concreto. Por ejemplo, si necesito agregar 30 días, querré estar segura, de que estoy trabajando con un tipo de datos de fecha. Los datos tienen muchos tipos y cada programa o herramienta tiene diferentes maneras de manejarlos y darles formato. Ahora comencemos con lo básico que todos necesitamos saber. Los tipos de datos más comunes son: Texto o cadena, Fecha, Fecha y Hora, Número, y uno muy especial que identificamos en datos con valores de tipo sí o no, verdadero o falso, o incluso encendido y apagado, que es el tipo de datos Booleano. Cuando observamos por primera vez los datos de cualquier conjunto de datos, generalmente donde nos detenemos es en los encabezados del campo. Por ejemplo, nombres, apellidos, edad, los valores que vemos en el campo como, por ejemplo, Elena para el nombre, Crespo para el apellido y 20 para edad. Pero, como analista de datos, recuerda que debes profundizar lo más posible para ver qué tipos están detrás de estos datos. La combinación de los encabezados de campo y todos los valores juntos, crean registros o conjuntos de registros. Esto probablemente te resulte familiar, las hojas de cálculo con las que sueles trabajar probablemente ya estén estructuradas con registros como este. La estructura, en forma de registros, nos da mucha flexibilidad al trabajar con herramientas de datos. Dependiendo de los datos con los que vas a trabajar, un punto de partida podría ser obtener un formato de conjuntos de registros. Las herramientas de datos con las que trabajarás, esperarán los datos en este formato de registros.

- Lidiar con los datos que no tenemos
- No siempre tenemos todo lo que necesitamos. Entonces, ¿qué pasa con los datos que no tenemos? En un mundo perfecto, todos los datos que necesitamos existen y están almacenados en algún lugar para que podamos acceder a ellos. Tenemos que reconocer que el mundo real no es perfecto y que algunos datos, simplemente, no existen. Entonces, ¿qué hacemos si los datos no están allí? Siempre recomiendo que se intente encontrar una forma de calcularlos a través de funciones o fórmulas. Y entonces la gente me pregunta: «¿Qué fórmulas debería aprender?» Hay algunas funciones comunes que son básicas e indispensables. A estas se unirán otras y otras muchas. Por necesidad, acabaremos aprendiendo mucho a lo largo de nuestra carrera en el mundo de los datos. Las más básicas son las funciones matemáticas, como sumar, restar, multiplicar y dividir. También recomiendo cosas como la concatenación y estructuras condicionales como si o if en inglés y las declaraciones de casos. La forma de escribir las funciones, la cantidad y tipo de parámetros que se esperan recibir dependen del software que estemos utilizando. Aprender a concatenar textos entre sí o textos con números y fechas, por ejemplo, permite unir campos para crear un único campo final. Un ejemplo muy común en el que se suele concatenar es cuando se necesita crear un único campo con los nombres y apellidos. Suele ser más fácil para trabajar con este campo que con ambos campos por separado. Una vez concatenados, se pueden ordenar por nombres y apellidos o por ambos, definiendo la precedencia en el orden, según se haya creado. Otro caso a tener en cuenta es la forma en que se obtienen los valores, basados en una condición. Hablamos de utilizar una función si o if. Yo diría que está funcionalidad está disponible en todas y cada una de las herramientas y que en todos los análisis se necesita, por lo general, este tipo de campos, creado a partir de evaluar una condición. En ese caso, a partir de evaluar una prueba lógica tendríamos que definir qué sucedería si el resultado es verdadero y qué sucedería si es falso. La evaluación con función SI o IF puede ser muy útil. Las declaraciones de caso tipo case son excelentes, especialmente si tienes la necesidad de comprobar varias pruebas lógicas. También es importante saber que los datos no siempre se almacenan en la forma en que los necesitamos. Por tanto, aprender a convertirlos en otros tipos de datos puede ser muy útil. Casi siempre se puede convertir cualquier cosa que se haya almacenado, por lo que vale la pena aprender las fórmulas y funciones de conversión. Ten en cuenta que las matemáticas solo funcionan con números y fechas. Si una fecha se almacena como texto, lo conviertes a tipo fecha para poder realizar cálculos específicos creados a partir de un tipo de datos fechas. Si una cantidad se almacena como texto, lo conviertes a un tipo de datos numérico, decimal o moneda, para luego poder utilizarlo en cálculos y agregaciones. Los datos que no tenemos, los creamos a través de cálculos basados en los que sí tenemos.

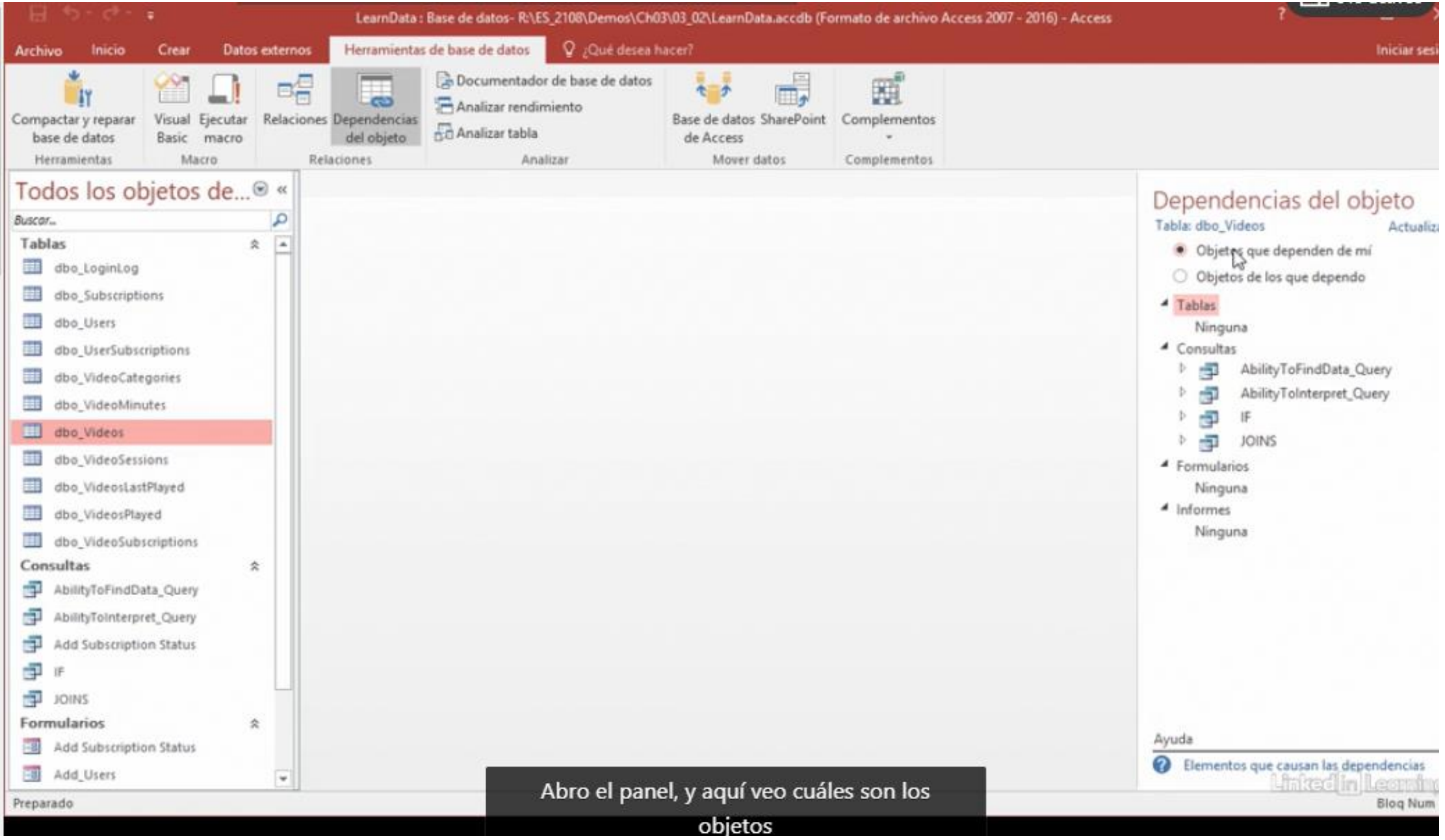
- **Aprender la sintaxis para mejorar la comunicación**
- Durante años, no importa qué programa hayan estado aprendiendo mis alumnos, siempre me han preguntado cuál es la mejor forma de aprender sintaxis. Comencemos por aprender de qué se trata y, si eres un principiante, las mejores formas de hacerlo. No es diferente a aprender la ortografía y la gramática de cualquier idioma. Utilizamos el lenguaje y las palabras para comunicar lo que queremos hacer, dónde queremos mirar. La sintaxis nos permite comunicar lo que queremos hacer con el programa. Las computadoras necesitan sintaxis. Si desea que su programa funcione y haga algo, entonces, la programación se ejecuta en segundo plano. Realmente, nada en la computadora existe sin el lenguaje que se compone de la sintaxis. Es importante, cuando estés aprendiendo sintaxis, que sepas que no es algo que no se le da bien a todo el mundo. Cada programa tiene su propio lenguaje fundamental, y, al igual que el software, evoluciona. ¿Cuál es la mejor ayuda para aprender sintaxis? Por regla general, es el menú de ayuda de cualquier programa con el que estés trabajando. Ese, es el punto de partida. Lo primero, es mirar la ayuda. Cuando no estés seguro de por dónde empezar, busca por la acción que necesitas y el nombre del programa. Así, por ejemplo, unir palabras en Excel o unir campos y números en SQL. Es más fácil encontrar la sintaxis cuando ya conoces otros programas. Por ejemplo, si sabes cómo escribir una fórmula de concatenación en Excel, puedes buscar la forma de concatenar en Access, en Tableau o en Power Bi. Veamos un ejemplo de sintaxis desde una hoja de cálculo de Excel en blanco. Voy a utilizar el acceso directo Control más coma para que me devuelva la fecha de hoy. Para el ejemplo de sintaxis, voy a la ficha Fórmulas y selecciono Insertar función. Cuando no estés seguro de la sintaxis o, incluso, de qué función deberías utilizar, entonces, te recomiendo que vengas a esta ventana a buscarla primero. Como queremos trabajar sobre una fecha escojo la categoría Fecha y hora. Me desplazo hasta la función FIN MES. Esta es una gran función. Me dice que devuelve el número de serie del último día del mes, antes o después del número especificado de meses. Así que la sintaxis es, abro paréntesis, Fecha inicial, meses y cierro paréntesis. Si quiero aprender más sobre esta función antes de usarla, haré clic en Ayuda sobre esta función. He animado a mis alumnos, a través de los años, a leer siempre sobre las funciones antes de utilizarlas. Entonces, se muestra una ventana con un artículo sobre esta función en particular, y me da la descripción de lo que hará la función y una descripción detallada sobre su sintaxis. Cierro esta ventana, y selecciono Aceptar. Como Fecha de inicio selecciono la fecha que había escrito antes, y digamos que quiero ver "tres meses". La ventana me ayudó a crear la sintaxis sin que yo tuviese que saber, precisamente, de sintaxis. Con fórmulas comunes, ya vas a saber la sintaxis sin necesidad de hacer este paso. Ahora, me está dando el número de serie de la fecha. Vamos a cambiar esto. Para ello, nos vamos a ir a la ficha Inicio, al grupo Número, y, aquí, pasamos de General a Fecha corta. Así es que el último día del mes, tres meses después del 15 de enero, es el 30 de abril. Sin irme de esta celda, presiono F2 para ver la sintaxis. F2 sobre una celda te sirve para ver la sintaxis de la fórmula utilizada en esa celda. Es muy útil si estás viendo la hoja de cálculo de otra persona, y quieres analizar cómo está estructurada tu fórmula. Aprender la sintaxis es uno de los obstáculos clave de cualquier lenguaje de programación que aprenderás. Solo puedes usarlo para fórmulas y funciones que escribas como analista de datos. Si eliges ser programador para ganarte la vida, recuerda, "la sintaxis es la vida".

- **Crear nuevas columnas de datos con Power BI**
- En pantalla tenemos la ventana del editor de consultas Power Query en Power BI. Vemos la lista de empleados con sus datos y la fecha de inicio de sesión en la plataforma de cursos. No tenemos una columna específica que indique si el empleado vio un curso o no. Sabemos que las personas que iniciaron sesión, pero que no vieron ningún curso, se identifican porque el nombre del curso y el inicio de sesión aparecen vacíos. Estas columnas tienen filas con valores nulos. Lo vemos simplemente gracias al perfil de calidad que indica parte de la banda de color negro, lo que significa que hay espacios. Vamos a crear una nueva columna desde el menú Agregar columna, grupo General, opción Columna condicional. Le llamamos «Visto o no», y vamos a comprobar si el nombre de la columna, Inicio de sesión es igual a null, eso significa que el curso no se ha visto. Si no es null, hay datos y se ha visto. Aceptamos. Hemos creado una columna condicional, que es lo que hacemos con un if o un sí. Actualizo en tipo de columna, es Texto, y nos devuelve dos valores: Visto o No visto. Bien, lo importante es detenerse a mirar con lupa lo que sí tenemos para poder sacar mayor provecho de ello. Vamos entonces a concatenar nombres y apellidos. Nos vamos al menú Transformar porque no queremos una nueva columna que mantenga las originales, sino que las elimine como columnas individuales tras concatenarlas. Selecciono el nombre, y con la tecla Control, Apellido. Desde la ficha Transformar, grupo Columna de texto, opción Combinar columnas. Indicamos un separador, en este caso Espacio, y un Nuevo nombre para la columna, en este caso Nombres y apellidos, y aceptamos. Tenemos la nueva columna y las originales desaparecieron. Los analistas nunca saben a qué se enfrentan hasta que comienzan a trabajar con los datos reales, por eso hay que mirarlos con mucho cuidado. Como tercera y última tarea, necesitamos obtener el tiempo que ha transcurrido desde el inicio de sesión hasta hoy, es decir, la antigüedad de esta fecha, y lo necesitamos en días. Nos vamos al menú Agregar columna porque, en este caso, queremos mantener esta columna original. Desde aquí al grupo de Fecha y hora, nos vamos a Fecha y la primera opción es Antigüedad. Esta columna es de tipo Duración, por ello tiene una estructura de día, horas, minutos y segundos. Como lo que queremos son los días nos vamos al menú para columnas de tipo Duración y seleccionamos Días, y ya tenemos el valor en días que era lo que necesitamos. Sabiendo que desde el principio tendrás escenarios en los que tendrás que calcular datos que no tienes, aprovecha al máximo lo que sí tienes y las capacidades de la herramienta con las que estás trabajando.

- Test de capítulo
- Aprender a identificar los datos es especialmente útil si vas a construir sistemas de datos
- ¿De qué color se muestra la banda del perfil de calidad cuando en las columnas hay filas con valores nulos? Negro
- La sintaxis nos permite comunicar lo que queremos hacer con el programa.

- **Aprender a interpretar los datos existentes**
- Si vas a ser un analista de datos, entonces, es importante que sepas cómo interpretar los datos que ves. Por lo general, creo que es mejor obtener datos sin formato e ir identificando el formato más adecuado para cada uno de ellos. Empecemos con un concepto simple, algo con lo que todos podemos relacionarnos: es un menú de un restaurante. Durante años se ha utilizado este ejemplo para ayudar a mis alumnos a aprender a interpretar los datos. Echemos un vistazo al menú de un restaurante. El menú en si es un informe, y todos los platos se componen de datos. Un analista de datos trabajará con datos y reglas de negocio, y eso no es diferente a trabajar con una receta de un plato del menú que incluye ingredientes e instrucciones. Cuando estás listo para hacer un nuevo plato a partir de una receta, ¿qué es lo primero que haces? ¿Haces la lista de los ingredientes que tienes, o que no tienes? Y cuando trabajas con datos, siempre debes crear una lista que te indique qué tienes, qué necesitas encontrar o qué necesitas crear. Sé que a veces voy a supermercados sin una lista, pero nunca comienzo un proyecto de datos, sin hacer la lista. Una vez que hayas definido tu lista de datos, es el momento de ir a comprar los ingredientes o los datos que faltan. Al igual que en el pasillo de supermercado puedes examinar los datos en sus diferentes sistemas u hojas de cálculo. Puedes encontrar un millón de cosas que necesitas mientras compras, pero recuerda que debes mantenerte enfocado en la lista para que puedas completar la primera receta. Veamos un ejemplo. Si quieres, puedes abrir el archivo Aprender a interpretar datos PDF desde tu archivo de ejercicios, para que puedas hacerlo conmigo. Echemos un vistazo a este informe que está en ingles. Contiene elementos como apellidos, nombres, nombres de vídeos e identificador de vídeos. Todo está en ingles. Estoy buscando en qué momento cada una de las personas ve un vídeo, así es que dejame mostrarte un ejemplo de mi lista. Para ello, voy a abrir un documento Word vacío y empezaré a escribir mi lista. Empiezo por la lista de campos que tengo. Estos son los campos que tengo, y luego voy a escribir la lista de los campos que estoy buscando y que no tengo. Aquí lo tengo. Es hora de ir a comprar por los pasillos del supermercado, o lo que es lo mismo recorrer mi base de datos. Estoy buscando algo como nombres de usuarios y nombres de vídeos. Así es que abro la base de datos Acces que se llama Learn data e inmediatamente veré su contenido. Veo que tengo Users, que contiene los datos de usuario. Perfecto. Tengo los nombres, el UserID, tengo lo que necesito. Esto es una tabla que voy a utilizar. También tengo una tabla Vídeos. Muy bien, contiene las fechas, las horas, todo lo que me hace falta. Pues aquí tengo las dos tablas que necesito y con ellas regreso a mi fichero Word para completar la lista. Con esto, he terminado la lista. Una vez que hayas descubierto el menú, hayas determinado el plato, hayas encontrado la receta y hayas completado tu compra, estás listo para hacer el plato. Los datos no son diferentes, descubren la necesidad, determinan las reglas comerciales de los datos. Haz tu compra, y luego crea tu conjunto de datos.

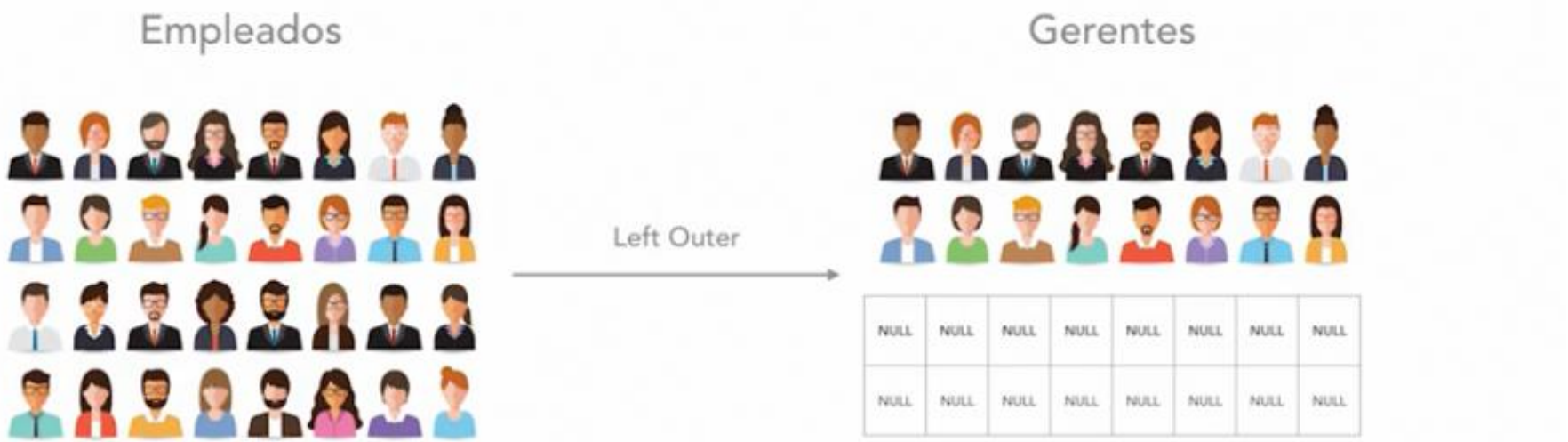
- **Encontrar datos existentes**
- La capacidad de encontrar datos existentes es crítica para un analista de datos. Puede ayudarte a comenzar a crear nuevos conjuntos de datos y a ampliar los conjuntos de datos existentes. A través de los años, he descubierto que trabajar hacia atrás es la clave. Verás, con mucha frecuencia, que el dato que necesitas para trabajar ya existe en un informe, un formulario, sitio web o programa de contabilidad. Veamos un ejemplo. Aquí tenemos un informe que contiene una tabla dinámica con distintos datos. Este informe que contiene una tabla dinámica está creado a mano; es muy frecuente. He visto que muchos de estos informes se crean a mano por los usuarios, y me pregunto: ¿Cómo se ha generado este informe? Nuestro objetivo será determinar de dónde provienen los datos para poder ampliarlos o tal vez automatizar la forma en la que se generan. Todo proviene de algún lugar; así es que busquemos de dónde vienen los datos, y veamos si podemos trabajar hacia atrás para encontrar el origen. Haz muchas preguntas, y recuerda que la perseverancia es la clave. Regresemos al archivo Excel. Vemos que además de la tabla dinámica, tenemos una pestaña con los datos. Quiero buscar en el sistema para determinar de dónde provienen estos datos. De esa manera, no tendría que reconstruirlo todo manualmente, y podría regenerar la tabla dinámica. Por eso, voy a abrir la base de datos. Aquí tengo la tabla Videos. Es muy importante, para mí, entender cuál es la dependencia de los objetos. Desde la ficha Herramienta de base de datos, obtengo este resultado. Abro el panel, y aquí veo cuáles son los objetos que dependen de mí, y los objetos de los que dependo. Me interesa más ver los objetos que dependen de mí. En este caso, vamos a ver la consulta AbilityToFindData Query. Ya puedo cerrar el panel Dependencias de objeto, y veo que en la consulta se utilizan las columnas VideoID, UserID, VideoName. Veamos si coinciden con las de Excel. Pues, sí, vemos que aparecen las mismas columnas; no importa que no estén en el mismo orden, lo importante es que las columnas existen. Nuevamente, desde el entorno Access, Vamos a exportar la consulta. Para ello, nos vamos a Datos externos, y los vamos a exportar a Excel. Selecciono las dos opciones, Exportar datos con formato y diseño y Abrir el archivo de destino al finalizar la operación de exportación, y el botón Aceptar. Aquí tenemos los dos libros de Excel abiertos. Por un lado, este es el resultado de la consulta que acabamos de hacer, y este es el archivo que teníamos antes, donde estaba la tabla dinámica. Contiene los mismos datos, y como dije antes, el orden da igual; lo importante es que nosotros tengamos ya de donde sacar los datos. Una de las primeras cosas que descubrirás es que gran parte de lo que necesitas, puede ser que ya exista. No dudes en hacer la ruta a la inversa para encontrar ese origen, y poderlo reutilizar.



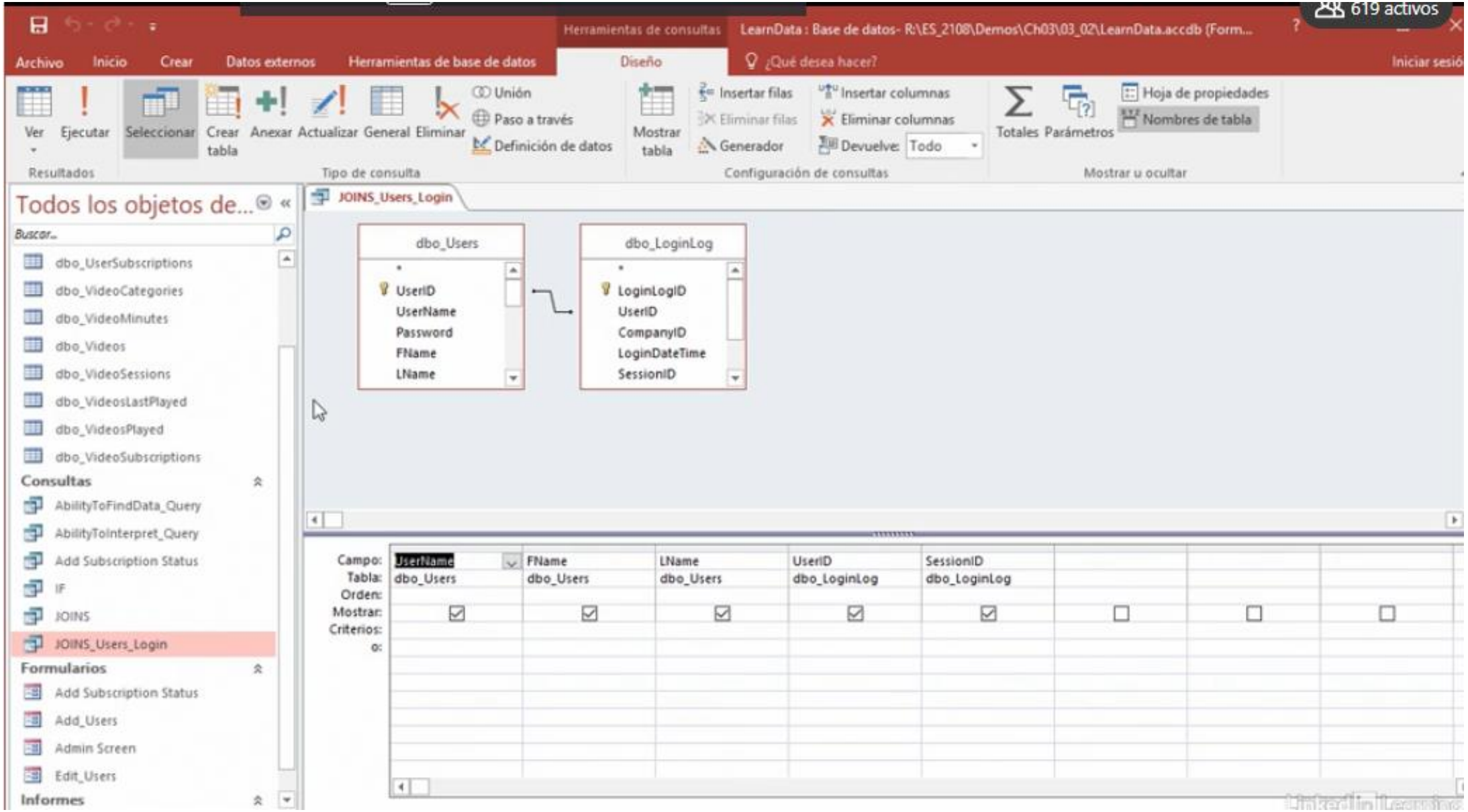
Abro el panel, y aquí veo cuáles son los objetos

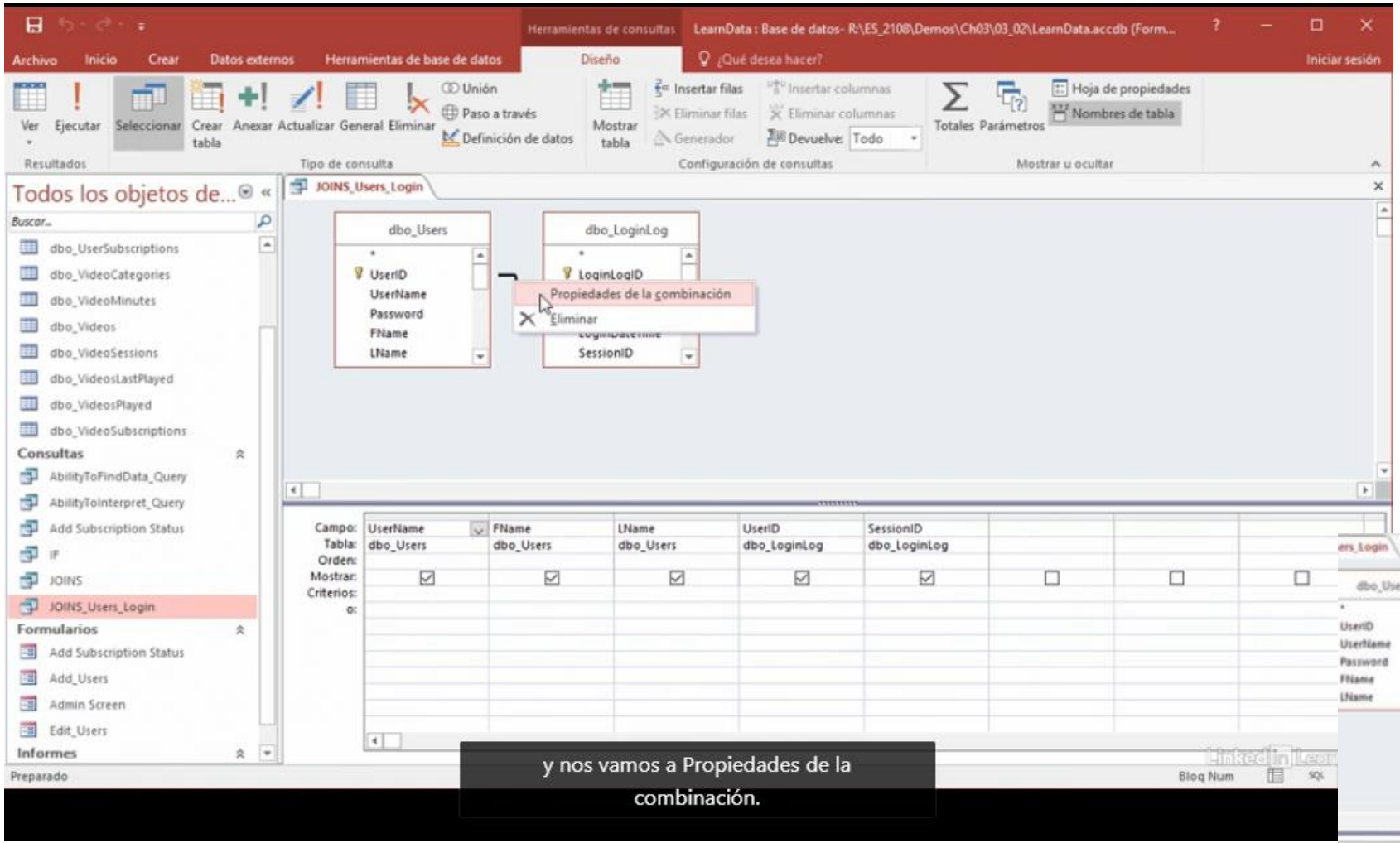
- **Entender las uniones entre datos**
- Como analista de datos encontrarás que trabajas con datos que se almacenan en tablas. Las tablas se relacionan, se unen. Es importante que entiendas las uniones. Entender la unión entre dos tablas del modelo de datos es vital para cualquier analista eficaz. En nuestro ejemplo, imaginemos que tenemos una tabla que tiene todos los empleados, tenemos otra tabla de empleados que son en realidad los gerentes; una combinación nos permitirá vincular los datos de ambas tablas. Creamos combinaciones de tablas para vincular datos de dos o más tablas y en una base de datos relacional estas uniones generalmente se proporcionan cuando abres una consulta y agregas esas tablas, de lo contrario, nosotros creamos las uniones. Las formas para crear una unión difieren de una herramienta a otra. Como concepto, los tipos más comunes son: uniones internas o Inner Join y externas, Outer Join. Las uniones internas son las predeterminadas. Como resultado, se obtienen los registros coincidentes de ambas tablas. Esto es genial siempre y cuando sea lo que tú quieres. Por ejemplo, si queremos ver toda la información de los empleados y sus gerentes, podemos hacer la unión predeterminada o Inner Join y reunir toda la información de los registros de los empleados junto con los gerentes. Esto nos da una lista completa de los datos de los gerentes. Una unión externa izquierda o Left Outer Join o resumido, Left Join, nos permitirá ver a todos los empleados, en este caso, es la tabla de la izquierda y la información correspondiente para cada gerente, ya que busca si hay una coincidencia en la tabla de la derecha. Imagina que en el informe del empleado, aquellos empleados que no son gerentes devolverán un valor null o nulo en las columnas correspondientes a la tabla Gerente. Veamos un ejemplo. En la base de datos vamos a seleccionar la vista Diseño de la consulta JOINS_Users_Login. En esta consulta tenemos los datos de los usuarios y la información de inicio de sesión. Están relacionados con la consulta predeterminada que es el Inner Join. Ejecutamos la consulta y vemos que solo tenemos a los usuarios que han iniciado sesión. Pero el requisito del informe es mostrar a todos los usuarios, incluidos a los que no han iniciado sesión. La única forma de lograr este informe es ajustar el tipo de unión. Por eso vamos a regresar a la vista Diseño. Regresamos a la vista Diseño. Vamos a la combinación, con clic derecho nos vamos a Propiedades. Desde la vista Diseño vamos a la unión, hacemos clic derecho y nos vamos a Propiedades de la combinación. Lo que nos interesa es unir todos los registros de la tabla de usuarios y solo aquellos registros de login donde los campos combinados sean iguales. Seleccionamos Aceptar. Vamos a ejecutar y aquí podemos ver que si nos movemos, efectivamente, vamos a encontrar usuarios que no han iniciado sesión. Aquí los tenemos. Estas personas no han iniciado sesión, por eso tenemos un valor nulo en el UserID y en el SessionID. Si quiero ver una lista de los usuarios que no han iniciado sesión, puedo hacer clic derecho en cualquiera de ellos y filtrarlo por la condición Igual a En blanco. Si mi objetivo es siempre ver esta lista, es decir, la lista de los usuarios que no han iniciado sesión, lo que podemos hacer es otra cosa en la consulta. Vamos a la vista Diseño para crear un criterio. En la celda correspondiente a Criterios para UserID escribo “es Nulo”. Significa que ese usuario todavía no ha iniciado sesión. Ejecuto la consulta y vemos cómo únicamente tenemos aquellos usuarios que no han iniciado sesión. La única forma de lograr algunos informes es uniendo los conjuntos de datos y sobre todo, ajustando los tipos de unión. Cuanto más trabajes con las uniones, más te familiarizarás con ellas y las comprenderás. Dedica todo el tiempo que puedas, sobre todo al principio, para aprender sobre las uniones y cómo se aplican en las distintas herramientas. Y recuerda, el tipo de unión determina lo que aparece en tu conjunto de datos resultante.

Uniones de tablas – LEFT OUTER

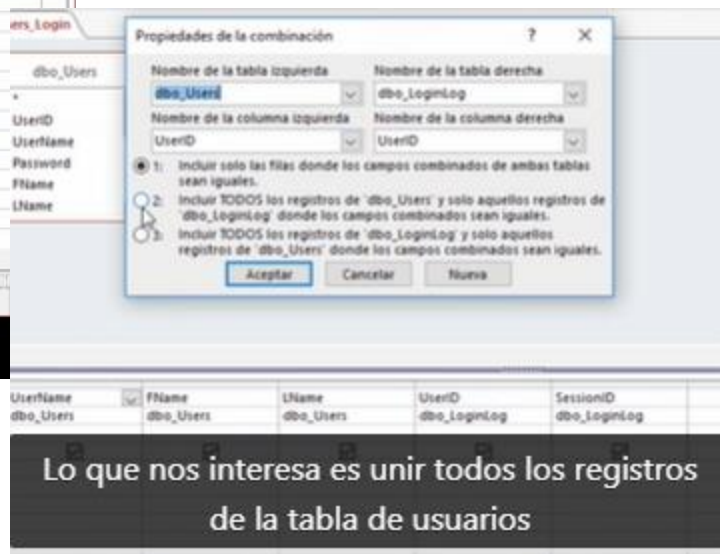


ya que busca si hay una coincidencia en la tabla de la derecha.

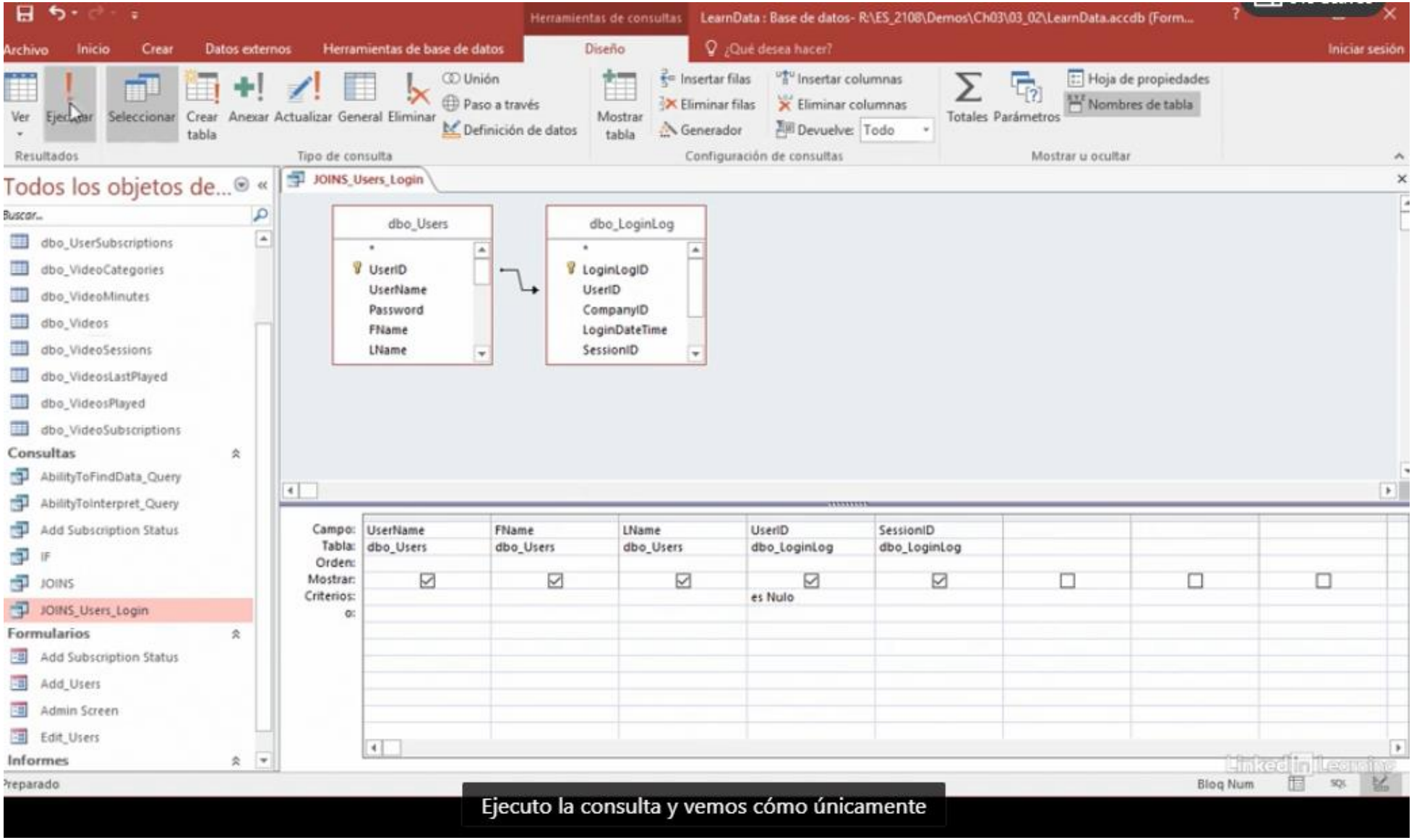

619 activos
Herramientas de consultas
LearnData : Base de datos- R:\ES_2108\Demos\Ch03\03_02\LearnData.accdb (Form...
Archivo
Inicio
Crear
Datos externos
Herramientas de base de datos
Diseño
¿Qué desea hacer?
Iniciar sesió
Ver
Ejecutar
Resultados
Seleccionar
Crear tabla
Anexar
Actualizar
General
Eliminar
Unión
Paso a través
Definición de datos
Tipo de consulta
Mostrar tabla
Insertar filas
Eliminar filas
Generador
Insertar columnas
Eliminar columnas
Devuelve: Todo
Configuración de consultas
Totales
Parámetros
Hoja de propiedades
Nombres de tabla
Mostrar u ocultar
Todos los objetos de...
Buscar...
dbo_UserSubscriptions
dbo_VideoCategories
dbo_VideoMinutes
dbo_Videos
dbo_VideoSessions
dbo_VideosLastPlayed
dbo_VideosPlayed
dbo_VideoSubscriptions
Consultas
AbilityToFindData_Query
AbilityToInterpret_Query
Add Subscription Status
IF
JOINS
JOINS_Users_Login
Formularios
Add Subscription Status
Add_Users
Admin Screen
Edit_Users
Informes
JOINS_Users_Login
dbo_Users
*
UserID
UserName
Password
FName
LName
dbo_LoginLog
*
LoginLogID
UserID
CompanyID
LoginDateTime
SessionID
Campo:
Tabla:
Orden:
Mostrar:
Criterios:
o:
UserName
FName
LName
UserID
SessionID
dbo_Users
dbo_Users
dbo_Users
dbo_LoginLog
dbo_LoginLog



y nos vamos a Propiedades de la combinación.



Lo que nos interesa es unir todos los registros de la tabla de usuarios


Herramientas de consultas
LearnData : Base de datos- R:\ES_2108\Demos\Ch03\03_02\LearnData.accdb (Form...
Archivo
Inicio
Crear
Datos externos
Herramientas de base de datos
Diseño
¿Qué desea hacer?
Iniciar sesión
Ver
Ejecutar
Seleccionar
Crear tabla
Anexar
Actualizar
General
Eliminar
Unión
Paso a través
Definición de datos
Mostrar tabla
Insertar filas
Eliminar filas
Generador
Insertar columnas
Eliminar columnas
Devuelve: Todo
Totales
Parámetros
Hoja de propiedades
Nombres de tabla
Resultados
Tipo de consulta
Configuración de consultas
Mostrar u ocultar
Todos los objetos de...
Buscar...
dbo_UserSubscriptions
dbo_VideoCategories
dbo_VideoMinutes
dbo_Videos
dbo_VideoSessions
dbo_VideosLastPlayed
dbo_VideosPlayed
dbo_VideoSubscriptions
Consultas
AbilityToFindData_Query
AbilityToInterpret_Query
Add Subscription Status
IF
JOINS
JOINS_Users_Login
Formularios
Add Subscription Status
Add_Users
Admin Screen
Edit_Users
Informes
JOINS_Users_Login
dbo_Users
UserID
UserName
Password
FName
LName
dbo_LoginLog
LoginLogID
UserID
CompanyID
LoginDateTime
SessionID
Campo:
Tabla:
Orden:
Mostrar:
Criterios:
o:
UserName
FName
LName
UserID
SessionID
dbo_Users
dbo_LoginLog
es Nulo
Preparado
Bloq Num
Ejecuto la consulta y vemos cómo únicamente

- **Entender los datos y el flujo de trabajo**
- Como analista de datos querrás aprender a encontrar datos dentro de un flujo de trabajo. Comienza por entender qué significa el flujo de trabajo y las herramientas que lo admiten, como pueden ser los diagramas de flujo y los diagramas de flujo de trabajo. Los diagramas de flujo de trabajo deben llevarte paso a paso a través de un proceso. Te ayudarán a visualizar cómo funciona el proceso y te indicarán el siguiente paso en cada momento. Estos diagramas pueden ayudar a cualquier analista a encontrar datos dentro de un proceso. Al iniciar cualquier proyecto, normalmente pido estos diagramas o la documentación que los debe acompañar, y si no existen, empiezo a hacer una versión para al menos poder empezar a trabajar. Un flujo de trabajo bien documentado nos permite ser más efectivos; sin el diagrama es fácil omitir accidentalmente un paso, porque a veces estamos tan acostumbrados a hacer el proceso que no nos damos cuenta que falta un paso. Un diagrama puede ayudar a todos a trabajar con precisión hasta el final. Veamos cómo podría ser un flujo de trabajo y cómo podría verse el programa a su lado. Comencemos con el diagrama de la izquierda y comparémoslo con la aplicación de la derecha. Así que tenemos los datos de usuarios y una pregunta: ¿es un usuario nuevo? Si seguimos el bloque de respuesta afirmativa Yes, nos dice que hagamos clic y agreguemos un nuevo usuario. A la derecha en la aplicación, en el número uno tenemos el botón para agregar usuarios. ¿Es este el primer paso? Podría ser el primer paso en el programa, pero el primer paso real son los datos. ¿De dónde vino el usuario?, ¿alguien nos llamó?, ¿nos lo enviaron por correo electrónico?, estas cosas hay que definirlas muy bien. Luego, pasar al paso dos parece bastante fácil, agrega la información del usuario nuevo y luego agrega la suscripción. Esto debería plantear otra pregunta para el analista. ¿Cómo sabemos qué suscripción? Espero que ya estés empezando a entender el punto que quiero explicar. Volvamos al diagrama y sigamos ahora la respuesta negativa No. El flujo de trabajo dice que si el usuario no es nuevo, tenemos que ir al paso Editar usuario, pero lo que no nos muestra es que necesitamos encontrar a ese usuario para poder editarlo. Ahí es donde la documentación que tenemos puede ser muy útil. Los diagramas de flujo de trabajo pueden ayudarte a crear tu plantilla de partida. Si no existen, hay que crearlos, es un tiempo muy bien empleado. Imagina cuánto llegarás a saber con un diagrama de flujo de trabajo bien documentado.

- Limpieza de datos
- Ojalá viviera en un mundo donde cada vez que alguien me entregue un proyecto de datos, los datos estén listos para que yo simplemente cree informes y visualizaciones sorprendentes. Pero esto nunca ocurre, porque los datos no están limpios. Piensa en las empresas que tienen 20, 30, o incluso 100 años de edad y pregúntate con qué frecuencia han tenido cambios en sus procesos durante ese tiempo. ¿Y qué impacto puede o no haber tenido en sus datos? Con el tiempo, los analistas aprendemos a limpiar los datos. No hay una única receta para esto, pero al menos descubrirás los mejores métodos para solucionar los desafíos a los que nos enfrentamos. Entonces, ¿qué es la limpieza de datos? Puedes leer una definición completa en un millón de sitios, como Wikipedia. En general, se trata de estandarizar los datos, eliminar lo que no es necesario para el informe y corregir los valores cuando son inconsistentes. Los pasos del proceso que seguirás puede incluir además la eliminación de datos con error cuando sepas que no son válidos. Se trata de validarlos con respecto a la verdad conocida. E incluso estandarizar los datos, como, por ejemplo, comprobar que para todos los estados de Estados Unidos se utilicen abreviaturas de dos letras. Esto no significa de ninguna manera el conjunto completo de cosas que harás, pero espero que sea suficiente para empezar. Es posible que descubras además que solo estas extrayendo una cantidad selecta de datos o haciendo que los códigos sean significativos, como puede ser convertir F a Mujer y M a Hombre. Puedes encontrar que realizas una gran cantidad de ordenación o incluso que divides columnas enteras de datos en campos separados para que puedas hacer aún más efectiva la ordenación. Es importante que siempre tengas un objetivo en mente, que es que debes tener un conjunto de datos de alta calidad y sentirte seguro acerca de los resultados y la legibilidad.

- Test de capitulo
- Según la formadora, por lo general es mejor obtener datos formateados e ir clasificándolos según el formato más adecuado para cada uno de ellos. FALSO En realidad, lo mejor es obtener datos sin formato e ir identificando el más adecuado para cada uno de ellos.