**Modulo uno**

**¿Qué es la ciencia de datos?**

Cuando tienes un modelo o hipótesis de un problema y tratas de validar esa hipótesis o modelo con tus datos.

Es cuando traduces los datos en una historia, es decir utilizas la narración para generar información, y con esa información puedes tomar decisiones estratégicas para una empresa o una institución.

El intento de trabajar con datos para encontrar respuestas a preguntas que estamos explorando.

Si tienes datos y tienes curiosidad y estás trabajando con datos y los estás manipulando, los estás explorando, el ejercicio mismo de revisarlos, analizarlos, tratar de obtener algunas respuestas de ellos es ciencia de datos.

**Fundamentos de la ciencia de datos**

Cualquier persona describe ciencia de datos de manera diferente. La mayoría concuerda que tiene un componente importante de análisis de datos.

La ciencia de datos puede ayudar a las organizaciones a comprender sus entornos, analizar problemas existentes, y revelar oportunidades previamente ocultas.

Los buenos científicos de datos son personas curiosas que hacen preguntas para aclarar las necesidades del negocio. Las siguientes preguntas son: ¿qué datos necesitamos para resolver el problema? ¿Y de dónde saldrán esos datos?

Cuando los datos han revelado su potencial, el papel del científico de datos pasa a ser el de un narrador que comunica los resultados a las partes interesadas del proyecto.

**Camino hacia la ciencia de datos**

La *curiosidad* es absolutamente necesaria. Si no tienes curiosidad, no sabrás qué hacer con los datos. *Juzgador* porque si no tienes nociones preconcebidas sobre las cosas, no sabrías por dónde empezar. *Argumentativo* porque si puedes argumentar y si puedes, puedes defender un caso.

La otra cosa que un científico de datos necesitaría es cierta comodidad y flexibilidad con las plataformas de análisis, algún software, y alguna plataforma informática.

*La capacidad de contar una historia*. Una vez que tengas tus análisis, una vez que tengas tus tabulaciones, ahora deberías ser capaz de contar una gran historia a partir de ello.

Así que primero debes determinar cuál es tu interés y cuál es tu ventaja competitiva. Tu ventaja competitiva no necesariamente serán tus destrezas analíticas. Su ventaja competitiva es su comprensión de algún aspecto de la vida en el que usted supera a otros en esa comprensión. Quizás sea cine, quizás sea comercio minorista, quizás sea salud, quizás sean computadoras.

|  |  |
| --- | --- |
| Algoritmos | Un conjunto de instrucciones paso a paso para resolver un problema o completar una tarea. |
| Modelo | Una representación de las relaciones y patrones encontrados en los datos para hacer predicciones o analizar sistemas complejos, manteniendo los elementos esenciales necesarios para el análisis. |
| Valores atípicos | Cuando un punto de datos o puntos ocurren significativamente fuera de la mayoría de los otros datos en un conjunto de datos, lo que podría indicar anomalías, errores o fenómenos únicos que podrían afectar el análisis estadístico o la modelización. |
| Análisis cuantitativo | Un enfoque sistemático que utiliza análisis matemáticos y estadísticos para interpretar datos numéricos. |
| Datos estructurados | Datos organizados y formateados en un esquema predecible, generalmente relacionados con tablas con filas y columnas. |
| Datos no estructurados | Datos desorganizados que carecen de un modelo de datos o de organización predefinido, lo que dificulta su análisis utilizando métodos tradicionales. Este tipo de datos a menudo incluye texto, imágenes, videos y otro contenido que no encaja perfectamente en filas y columnas como los datos estructurados. |

**Comprender los distintos tipos de formatos de archivo**

Los archivos de texto delimitados son archivos de texto que se utilizan para almacenar datos como texto en el que cada línea o fila tiene valores separados por un delimitador, donde un delimitador es una secuencia de uno o más caracteres para especificar el límite entre entidades o valores independientes. Se puede utilizar cualquier carácter para separar los valores, pero los delimitadores más comunes son la coma el tabulador, los dos puntos, la barra vertical, y el espacio. Los valores separados por comas, o CSV, y los valores separados por tabulaciones, o TSV, son los tipos de archivos más utilizados en esta categoría. En los CSV, el delimitador es una coma, mientras que en los TSV el delimitador es un tabulador. Cuando hay comas literales en los datos de texto y, por lo tanto, no se pueden usar como delimitadores, las TSV sirven como alternativa al formato CSV. Los tabuladores son poco frecuentes en el texto continuo. Los archivos delimitados permiten valores de campo de cualquier longitud y se consideran un formato estándar para proporcionar un esquema de información sencillo.

Hoja de cálculo Open XML de Microsoft Excel, o XLSX, es un formato de archivo Open XML de Microsoft Excel que se incluye dentro del formato de archivo de hoja de cálculo. Puede haber múltiples hojas de cálculo. y cada hoja de cálculo está organizada en filas y columnas en cuya intersección se encuentra la celda. Cada celda contiene datos. XLSX utiliza el formato de archivo abierto, lo que significa que generalmente es accesible para la mayoría de las demás aplicaciones. El lenguaje de marcado extensible o XML es un lenguaje de marcado con reglas establecidas para codificar datos. El formato de archivo XML es legible tanto para humanos como para máquinas. Es un lenguaje autodescriptivo diseñado para enviar información a través de Internet.

XML es similar a HTML en algunos aspectos, pero también tiene diferencias. Por ejemplo, un XML no utiliza etiquetas predefinidas como lo hace HTML. XML es independiente de la plataforma y del lenguaje de programación y, por lo tanto, simplifica el intercambio de datos entre varios sistemas.

El Formato de Documento Portátil, o PDF, es un formato de archivo desarrollado por Adobe para presentar documentos de manera independiente del software de aplicación, hardware y sistemas operativos (sistema operativo), lo que significa que se puede ver de la misma manera en cualquier dispositivo.

ADP. Este formato se utiliza frecuentemente en documentos legales y financieros y también se puede utilizar para rellenar datos como formularios. La Notación de Objetos de JavaScript o JavaScript Object Notation, es un estándar abierto basado en texto diseñado para transmitir datos estructurados a través de la web.

JSON es fácil de usar, es compatible con una amplia gama de navegadores, y se considera una de las mejores herramientas para compartir datos de cualquier tamaño y tipo, incluso audio y vídeo.

*Los datos no estructurados son básicamente datos que provienen principalmente de la web, donde no son tabulares. No está en filas y columnas. Es texto. A veces es video y audio. Por lo tanto, habría que implementar algoritmos más sofisticados (más sofisticados que la regresión) para extraer datos. Y, de hecho, muchas veces tomamos Datos no estructurados y pasamos una gran cantidad de tiempo y esfuerzo para obtener algo de Datos estructurados de ellos y luego analizarlos. Así que, si tienes algo que se adapte bien a tablas, columnas y filas, adelante. Esos son datos estructurados, pero si ves que se trata de un registro web o si estás intentando obtener información de páginas web y tienes un montón de páginas web, se trata de datos no estructurados que requerirían un poco más de esfuerzo para extraer la información de ellos.*

**¿Qué es la regresión?**

Si alguna vez has tomado un taxi, entenderás la regresión. Así es como funciona. En el momento en que te sientas en un taxi, en un taxi ves que hay una cantidad fija allí. Dice $ 2,50. Ya sea que el taxi se mueva o usted baje, esto es lo que le debe al conductor. Desde el momento en que te subes a un taxi, eso es una constante. Si has subido a un taxi, tendrás que pagar esa cantidad. Luego, a medida que comienza a moverse, por cada metro o cien metros, la tarifa aumenta en una cierta cantidad. Entonces, hay una fracción, hay una relación entre la distancia y la cantidad que pagarías, más allá de esa constante. y si no te estás moviendo y estás atrapado en el tráfico, entonces por cada minuto adicional tienes que pagar más, así que a medida que aumentan los minutos tu tarifa aumenta, a medida que aumenta la distancia tu tarifa aumenta y mientras todo esto sucede ya has pagado una tarifa base que es la constante, esto es lo que es la regresión, la regresión te dice cuál es la tarifa base ¿Y cuál es la relación entre el tiempo y la tarifa que has pagado y la distancia que has recorrido y la tarifa que has pagado? Porque en ausencia de conocer esas relaciones y solo saber cuánto viajó la gente y cuánto pagó, la regresión le permite calcular esa constante que no sabía que era 250 y calcularía la relación entre la tarifa y la distancia y la tarifa y el tiempo

**¿Qué convierte a alguien en un científico de datos?**

Alguien que encuentra soluciones a problemas analizando datos grandes o pequeños utilizando herramientas apropiadas y luego cuenta historias para comunicar sus hallazgos a las partes interesadas relevantes.

Mientras uno tenga una mente curiosa, fluidez en analítica y la capacidad de comunicar los hallazgos, considero que esa persona es un científico de datos.

Un científico de datos tiene que ser alguien con una mente muy curiosa, dispuesto a dedicar tiempo y esfuerzos significativos para explorar sus corazonadas.

Tienden a ser personas realmente curiosas, pensadores que hacen buenas preguntas y están bien al lidiar con situaciones no estructuradas y tratando de encontrar estructura en ellas.

*Los científicos de datos investigan y encuentran explicaciones para muchos problemas. Por ejemplo, el Dr. Murtaza Haider encontró una explicación de por qué medio millón de clientes se quejaron del transporte público en Toronto. Después de mucha investigación, encontró una relación entre los eventos climáticos inesperados y la cantidad de quejas ese día en particular.*

Granville, autor sobre ciencia de datos enumera el álgebra, el cálculo, y la formación en probabilidad y estadística como antecedentes educativos necesarios para ser un científico de datos.

Un científico de datos exitoso es una combinación de científico informático, ingeniero de software, y estadístico. Su capacidad para transformar soluciones no estructuradas en conocimientos estructurados define su destreza.

|  |  |
| --- | --- |
| **Término** | **Definición** |
| Valores separados por comas (CSV) / Valores separados por tabulaciones (TSV) | Formato comúnmente utilizado para almacenar datos tabulares como texto plano donde el valor es separado por una coma o una tabulación. |
| Tipos de archivos de datos | Una configuración de archivo informático diseñada para almacenar datos de una manera específica. |
| Formato de datos | Cómo se codifican los datos para que puedan ser almacenados dentro de un tipo de archivo de datos. |
| Visualización de datos | Una forma visual, como un gráfico, de representar datos de manera fácilmente comprensible que facilita la identificación de tendencias en los datos. |
| Archivo de texto delimitado | Un archivo de texto plano donde un carácter específico separa los valores de los datos. |
| Lenguaje de Marcado Extensible (XML) | Un lenguaje diseñado para estructurar, almacenar y permitir el intercambio de datos entre diversas tecnologías. |
| Hadoop | Un marco de trabajo de código abierto diseñado para almacenar y procesar grandes conjuntos de datos en clústeres de computadoras. |
| Notación de Objetos de JavaScript (JSON) | Un formato de datos compatible con varios lenguajes de programación para que dos aplicaciones intercambien datos estructurados. |
| Notebooks de Jupyter | Un entorno computacional que permite a los usuarios crear y compartir documentos que contienen código, ecuaciones, visualizaciones y texto explicativo. Consulta los notebooks de Python. |
| Vecino más cercano | Un algoritmo de aprendizaje automático que predice una variable objetivo-basada en su similitud con otros valores en el conjunto de datos. |
| Redes neuronales | Un modelo computacional utilizado en el aprendizaje profundo que imita la estructura y el funcionamiento de las vías neuronales del cerebro humano. Toma una entrada, la procesa utilizando el aprendizaje previo y produce una salida. |
| Pandas | Una biblioteca de Python de código abierto que proporciona herramientas para trabajar con datos estructurados y que se utiliza a menudo para la manipulación y análisis de datos. |
| Notebooks de Python | También conocido como un notebook “Jupyter”, este entorno computacional permite a los usuarios crear y compartir documentos que contienen código, ecuaciones, visualizaciones y texto explicativo. |
| R | Un lenguaje de programación de código abierto utilizado para computación estadística, análisis de datos y visualización de datos. |
| Motor de recomendación | Un programa informático que analiza la entrada del usuario, como comportamientos o preferencias, y hace recomendaciones personalizadas basadas en ese análisis. |
| Regresión | Un modelo estadístico que muestra una relación entre una o más variables predictoras y una variable de respuesta. |
| Datos tabulares | Datos que están organizados en filas y columnas. |
| XLSX | El formato de archivo de hoja de cálculo de Microsoft Excel. |

**Modulo dos**

**¿Como los macrodatos impulsan la transformación digital?**

La disponibilidad de grandes cantidades de datos y la ventaja competitiva que supone su análisis han desencadenado transformaciones digitales en muchas industrias. Netflix pasó de ser un sistema de préstamo de DVD por correo a uno de los principales proveedores de transmisión de video del mundo. El equipo MBA de los Houston Rockets utilizó datos recopilados por cámaras aéreas para analizar las jugadas más productivas, y Lufthansa analizó los datos de los clientes para mejorar su servicio.

La transformación digital no es simplemente duplicar procesos existentes en forma digital. El análisis en profundidad de cómo funciona el negocio ayuda a las organizaciones a descubrir cómo mejorar sus procesos y operaciones y aprovechar los beneficios de integrar la ciencia de datos en sus flujos de trabajo.

La transformación digital impacta cada aspecto de la organización, por lo que es manejada por los tomadores de decisiones en los niveles más altos para garantizar el éxito. El apoyo del director ejecutivo es crucial para el proceso de transformación digital, como lo es el apoyo del director de información y el rol emergente del director de datos

**Introducción a la nube**

La computación en la nube, también conocida como nube, es la entrega de recursos informáticos bajo demanda, como redes, servidores, almacenamiento, aplicaciones, servicios y centros de datos, a través de Internet mediante un pago por uso.

El término computación en la nube se puede utilizar para describir aplicaciones y datos a los que los usuarios acceden a través de Internet en lugar de hacerlo en su computadora local. Algunos ejemplos de computación en la nube incluyen usuarios que utilizan aplicaciones web en línea, empleados que utilizan aplicaciones comerciales en línea seguras para realizar su trabajo, y usuarios que almacenan archivos personales en plataformas de almacenamiento basadas en la nube, como Google Drive, OneDrive, y Dropbox.

Comencemos por comprender las cinco características esenciales de la nube. El autoservicio a pedido significa que usted obtiene acceso a los recursos de la nube como la potencia de procesamiento, el almacenamiento, y la red que necesita mediante una interfaz simple sin necesidad de interacción humana con cada proveedor de servicios. El acceso amplio a la red significa que se puede acceder a los recursos de computación en la nube a través de la red mediante mecanismos y plataformas estándar como teléfonos móviles, tabletas, computadoras portátiles, y estaciones de trabajo. La puesta en común de recursos es lo que brinda a los proveedores de la nube economías de escala, que luego transmiten a sus clientes, haciendo que la nube sea rentable. Al utilizar un modelo multiinquilino, los recursos informáticos se agrupan para atender a múltiples consumidores, y los recursos de la nube se asignan y reasignan dinámicamente según la demanda, sin que los clientes necesiten conocer la ubicación física de estos recursos. La elasticidad rápida implica que usted puede acceder a más recursos cuando los necesita y reducirlos cuando no los necesita porque los recursos se aprovisionan y liberan de forma elástica. Y un servicio medido significa que usted sólo paga por lo que usa o reserva a medida que avanza.

Los modelos de implementación de la nube indican dónde reside la infraestructura, quién la posee y la administra, y cómo los recursos y servicios de la nube se ponen a disposición de los usuarios. Hay tres tipos de modelos de implementación de nube: pública, privada e híbrida. La nube pública es cuando se aprovechan los servicios en la nube a través de Internet abierta, en hardware propiedad del proveedor de la nube, pero su uso es compartido por otras empresas. Nube privada significa que la infraestructura de la nube está destinada para uso exclusivo de una sola organización. Podría ejecutarse localmente, o podría ser propiedad, de un proveedor de servicios y estar administrado y operado por él. Y cuando usas una mezcla de nubes públicas y nube privada, trabajando juntas de manera fluida, eso se clasifica como el modelo de nube híbrida.

Ahora, veamos los tres modelos de servicios en la nube que se basan en las tres capas de una pila informática Infraestructura, Plataforma, y Aplicación. Estos modelos de computación en la nube se denominan acertadamente Infraestructura como Servicio o IaaS, Plataforma como Servicio o PaaS y Software como Servicio o SaaS.

En un modelo IaaS, puede acceder a la infraestructura y a los recursos informáticos físicos, como servidores redes, almacenamiento, y espacio del centro de datos, sin la necesidad de administrarlos ni operarlos. En un modelo PAAS, se puede acceder a la plataforma que comprende las herramientas de hardware y software que normalmente se necesitan para desarrollar e implementar aplicaciones para los usuarios a través de Internet. Y un SAAS es un modelo de entrega y licencia de software en el que el software y las aplicaciones se alojan de forma centralizada y se licencian mediante suscripción. A veces se le denomina software bajo demanda.

**Nube para científicos de datos**

La nube es una bendición para los científicos de datos principalmente porque les permite tomar sus datos, tomar su información y colocarla en la nube, colocarla en el sistema de almacenamiento central.

La nube no sólo permite almacenar grandes cantidades de datos en servidores en algún lugar de California o Nevada, sino que también permite implementar algoritmos informáticos muy avanzados y la capacidad de realizar cálculos de alto rendimiento utilizando máquinas que no son suyas.

Permite que múltiples entidades trabajen con los mismos datos al mismo tiempo. Así que puedes estar trabajando con los mismos datos que tus colegas en, digamos, Alemania, y otro equipo en India, y otro equipo en Ghana; ellos están trabajando colectivamente y pueden hacerlo porque la información, los algoritmos, las herramientas, las respuestas y los resultados, lo que sea que necesiten, está disponible en un lugar central.

El uso de la nube le permite obtener acceso instantáneo a tecnologías de código abierto como Apache Spark sin la necesidad de instalarlas y configurarlas localmente. El uso de la nube también le brinda acceso a las herramientas y bibliotecas más actualizadas sin la preocupación de mantenerlas y garantizar que estén actualizadas. La nube es accesible desde cualquier lugar y en cualquier zona horaria. Puede utilizar tecnologías basadas en la nube desde su computadora portátil, desde su tableta e incluso desde su teléfono, lo que permite colaborar con más facilidad que nunca.

Algunas grandes empresas tecnológicas ofrecen plataformas en la nube, que le permiten familiarizarse con las tecnologías basadas en la nube en un entorno prediseñado. IBM ofrece IBM Cloud, Amazon ofrece AWS y Google Cloud Platform.

**Fundamento de macrodatos**

En este mundo digital, cada uno deja huella. Desde nuestros hábitos de viaje hasta nuestros entrenamientos y entretenimiento, la creciente cantidad de dispositivos conectados a Internet con los que interactuamos a diario registran enormes cantidades de datos sobre nosotros. Incluso hay un nombre para ello: Big Data.

Big Data se refiere a los volúmenes de datos dinámicos grandes y dispares que crean las personas, las herramientas y las máquinas. Éstas son las V del big data:

La *velocidad* es la velocidad a la que se acumulan los datos. Los datos se generan extremadamente rápido en un proceso que nunca se detiene. Las tecnologías de transmisión en tiempo real o casi real, locales y basadas en la nube pueden procesar información muy rápidamente. El *volumen* es la escala de los datos o el aumento en la cantidad de datos almacenados. Los impulsores del volumen son el aumento de las fuentes de datos, los sensores de mayor resolución y la infraestructura escalable.

La *variedad* es la diversidad de los datos. Los datos estructurados encajan perfectamente en filas y columnas en Base de datos relacional, mientras que los datos no estructurados no están organizados de una manera predefinida como los tweets, las publicaciones de blogs, las imágenes, los números y los videos. La variedad también refleja que los datos provienen de diferentes fuentes, máquinas, personas, y procesos, tanto internos como externos a las organizaciones.

Los impulsores son las tecnologías móviles, las redes sociales, las tecnologías portátiles, las geotecnologías, el vídeo, y muchos, muchos más. La *veracidad* es la calidad y origen de los datos y su conformidad con los hechos y la exactitud. Los atributos incluyen consistencia, completitud, integridad y ambigüedad. Los factores que impulsan esta tendencia son el costo y la necesidad de trazabilidad.

El *valor* es nuestra capacidad y necesidad de convertir los datos en valor. El valor no es sólo ganancia. Puede tener beneficios médicos o sociales, así como satisfacción del cliente, del empleado o personal. La razón principal por la que las personas invierten tiempo en comprender el big data es para obtener valor de él.