1.- ¿Qué es la ciencia de datos?

Define una práctica multidisciplinar con muchos significados.

El término "científico de datos" tiene múltiples significados y connotaciones. A diferencia de otras disciplinas, no tiene una definición precisa debido a su naturaleza multidisciplinaria y evolutiva. Originalmente, el término se utilizaba antes de que la ciencia de datos se estableciera como una disciplina formal. Hoy en día, aquellos que se autodenominan científicos de datos provienen de diversas áreas de investigación, lo que hace que su definición sea flexible y relativa.

La ciencia de datos no se define claramente como una disciplina, sino más bien como una práctica. Al igual que en los primeros días de la arqueología, cualquiera que trabajara con datos de manera científica podría llamarse a sí mismo científico de datos. Esta práctica implica trabajar con datos utilizando enfoques científicos, experimentación y preguntas empíricas para obtener conocimiento significativo. En este contexto, se enfatiza el método científico sobre los datos en sí.

La formación y antecedentes profesionales de los científicos de datos varían ampliamente. Personas con formación en estadísticas, análisis de datos, ciencias biológicas, ingeniería de sistemas, negocios y finanzas, entre otros, pueden autodenominarse científicos de datos. A medida que la demanda de científicos de datos continúa creciendo, se busca estandarizar las habilidades necesarias, pero esto aún no se ha logrado por completo.

Un enfoque central de la ciencia de datos es la aproximación empírica, que implica hacer preguntas a los datos, realizar experimentos y ajustar enfoques para ganar comprensión. Un ejemplo se da a través del proceso de encender una ducha en diferentes habitaciones de hotel. En lugar de depender de suposiciones, se utilizan experimentos y ajustes para lograr el resultado deseado. Los científicos de datos emplean esta aproximación empírica constantemente al trabajar con datos, formulando preguntas y ajustando enfoques para obtener un mejor entendimiento y mejorar la calidad de las preguntas.

En resumen, ser un científico de datos implica trabajar con datos utilizando métodos científicos y aproximaciones empíricas para obtener conocimiento significativo. La disciplina es fluida y abarca profesionales con diversos antecedentes. A medida que la ciencia de datos sigue evolucionando, se enfatiza la importancia de centrarse en el enfoque científico y la práctica empírica para obtener un mejor entendimiento a partir de los datos.

Los paquetes estadísticos y las herramientas informáticas.

La ciencia de datos requiere la utilización de paquetes estadísticos y herramientas informáticas para llevar a cabo tareas de almacenamiento, depuración y análisis de datos. Aunque es importante estar familiarizado con estas herramientas, el énfasis debe ponerse

en el método científico más que en las herramientas en sí. Las herramientas se pueden clasificar en tres categorías principales: almacenamiento, depuración y análisis.

1. Almacenamiento de Datos: Para almacenar datos, se pueden usar hojas de cálculo, bases de datos y sistemas clave-valor. Algunas herramientas comunes incluyen Hadoop, Cassandra y POST REST SQL. Hadoop es especialmente popular para el manejo de big data, utilizando un sistema de archivos distribuido en un clúster de servidores.

2. Depuración de Datos: La depuración es crucial para trabajar eficientemente con los datos. Herramientas como editores de texto, scripts y lenguajes de programación como Python y SCALLOP se utilizan para facilitar la manipulación y transformación de datos.

3. Análisis de Datos: Existen paquetes estadísticos que ayudan en el análisis de datos. Algunos ejemplos son el paquete de código abierto R, SBSS y las bibliotecas de datos de Python. Estas herramientas permiten realizar análisis estadísticos, visualización de datos y creación de gráficos y diagramas.

En relación al "big data", es importante destacar que la ciencia de datos no se limita exclusivamente a él. Si bien hay una conexión estrecha, la ciencia de datos aplica el método científico para estudiar los datos en general, no solo los que entran en la categoría de big data.

La depuración de datos es un aspecto esencial y a menudo consume una gran cantidad de tiempo. Los científicos de datos dedican una parte significativa de su tiempo a limpiar y preparar los datos para su análisis. Una vez que los datos están limpios, se pueden usar herramientas como R o Python para realizar análisis estadísticos y visualización de datos.

En resumen, las herramientas informáticas y paquetes estadísticos son instrumentos valiosos en la práctica de la ciencia de datos, pero el enfoque principal debe ser el método científico y el proceso de hacer preguntas empíricas a los datos para obtener conocimiento significativo. Las herramientas son herramientas, y es crucial priorizar el análisis y el entendimiento sobre la adquisición de nuevas herramientas o software.

Crea conocimiento sobre el negocio.

La ciencia de datos se diferencia de la eficiencia operativa en las organizaciones al ser exploratoria y basarse en el método científico para obtener conocimiento útil para el negocio. A través de preguntas como qué sabemos de los clientes, cómo mejorar un producto, qué hace la empresa mejor que la competencia y qué sucedería si la empresa dejara el mercado, se busca profundizar en el conocimiento organizacional. Sin embargo, muchas entidades no están preparadas para hacer este tipo de preguntas, ya que a menudo están enfocadas en objetivos y respuestas inmediatas.

Las habilidades de pensamiento crítico y cuestionamiento son esenciales para construir conocimiento organizacional. A menudo, las empresas se centran en la producción de resultados y ven las preguntas como obstáculos. Sin embargo, hacer preguntas interesantes y explorar diferentes posibilidades es fundamental para ganar conocimiento.

Un ejemplo ilustrativo es el de una web que conecta compradores de autos con concesionarias. Aunque tenían acceso a gran cantidad de datos almacenados en un clúster Hadoop, la verdadera dificultad radicaba en decidir qué hacer con esa información. La recolección de datos es relativamente fácil y barata, pero la parte científica y exploratoria de la ciencia de datos es donde muchas organizaciones encuentran desafíos.

La ciencia de datos comienza con una pregunta y luego involucra experimentación y análisis. La idea es aprovechar los datos mediante preguntas interesantes, realizar experimentos para obtener respuestas y presentar los resultados en informes bien diseñados. Es fundamental comprender que recopilar datos es solo una parte de la ecuación, y que la verdadera aportación de la ciencia de datos radica en cómo se exploran, se analizan y se traducen en insights valiosos para el negocio.

2.- Trabaja con bases de datos relacionales.

Haz conexiones con bases de datos relacionales.

Los científicos de datos trabajan con una variedad de fuentes de datos, incluyendo bases de datos antiguas, hojas de cálculo, imágenes y videos. Es importante entender cómo se almacenan los datos en las organizaciones, y la mayoría de ellas ofrecen diversas opciones de almacenamiento que pueden ser modernas o más tradicionales. Comprender la evolución de las tecnologías de bases de datos es fundamental para tener una base sólida en el tema.

Las bases de datos modernas se originaron en la década de 1960 con la misión espacial del Apolo. La NASA colaboró con IBM para desarrollar un sistema de gestión de información llamado IMS, que se parecía a una hoja de cálculo moderna con columnas y filas. Con el tiempo, se crearon las bases de datos relacionales, que dividieron los datos en grupos de tablas. Cada tabla era similar a una hoja de cálculo, y se establecían relaciones entre estas tablas para organizar los datos de manera más eficiente.

El lenguaje SQL (Structured Query Language) fue desarrollado en la década de 1970 como una forma de consultar y acceder a los datos en bases de datos relacionales. SQL permitía extraer información de múltiples tablas relacionales y presentar los resultados en una tabla virtual llamada Vista. SQL se convirtió en uno de los lenguajes más utilizados y sigue siendo una habilidad valiosa en el campo de la ciencia de datos.

El diseño de bases de datos relacionales requiere planificación cuidadosa, ya que es necesario definir cómo se estructuran las tablas y cómo se relacionan entre sí. Los esquemas, que representan las relaciones entre las tablas, son esenciales para el diseño de la base de datos. Sin embargo, este diseño puede ser desafiante, ya que los cambios en la estructura de los datos pueden requerir modificaciones en el diseño de la base de datos.

Con el tiempo, se añadieron numerosas funciones y características a las bases de datos relacionales, dando lugar a los sistemas de gestión de bases de datos relacionales (RDBMS). Empresas como IBM, Microsoft y Oracle todavía ofrecen asistencia y desarrollo en este tipo de sistemas, que son utilizados en una amplia variedad de aplicaciones y entornos.

En resumen, las bases de datos relacionales son una parte esencial de la ciencia de datos y el almacenamiento de datos en las organizaciones. Comprender su evolución, diseño y el lenguaje SQL es fundamental para trabajar con eficacia en la manipulación y extracción de datos en entornos de ciencia de datos.

Guardar información en almacenes de datos mediante ETL.

En el campo de la ciencia de datos, trabajar con almacenes de datos y realizar procesos ETL (Extract, Transform, Load) es fundamental para manipular y analizar la información. Algunos conceptos clave:

Almacén de Datos (EDW): Un almacén de datos es un tipo especial de base de datos relacional diseñada para el análisis de información. A diferencia de las bases de datos tradicionales que se optimizan para el procesamiento de transacciones en línea (OLTP), los almacenes de datos se dedican al procesamiento analítico en línea (OLAP), lo que implica analizar datos históricos y generar informes.

Bases de Datos OLTP y OLAP: Las bases de datos OLTP se utilizan para transacciones en tiempo real, como compras en línea, y deben ser rápidas y eficientes en el procesamiento de datos. Por otro lado, las bases de datos OLAP se utilizan para análisis y generación de informes, lo que implica acceder a grandes conjuntos de datos y realizar operaciones complejas.

Proceso ETL: ETL es un proceso crucial en la ciencia de datos. Significa Extract, Transform, Load (Extraer, Transformar y Cargar). Las organizaciones realizan ETL para mover datos desde diversas fuentes (extracción), transformarlos en el formato necesario (transformación) y cargarlos en un almacén de datos o base de datos (carga). Esta transformación puede incluir limpieza, unión de datos y conversión de formatos.

Interacción con Almacenes de Datos: En el ejemplo dado, cuando una empresa adquiere una página web exitosa que vende productos, puede combinar los datos de esa página con sus propios datos en su almacén de datos corporativo. Esto involucra realizar un proceso ETL para extraer los datos, transformarlos según las necesidades del almacén y luego cargarlos en él.

Hadoop y Almacenes de Datos: En algunos casos, las organizaciones consideran a Hadoop como una alternativa a los almacenes de datos tradicionales. Pueden reescribir sus procesos ETL para cargar los datos en un clúster Hadoop a medida que eliminan gradualmente los almacenes de datos existentes. Esta estrategia a menudo se elige para reducir costos.

Desafíos y Consideraciones: Trabajar con almacenes de datos y realizar procesos ETL puede ser complejo y requiere un profundo entendimiento de los datos y su estructura. Los equipos de data science suelen enfrentar desafíos en la limpieza, unión y transformación de datos antes de poder realizar análisis significativos. Comprender los términos y desafíos asociados puede facilitar la obtención de los datos necesarios para el análisis.

En resumen, los almacenes de datos y el proceso ETL son elementos clave en la ciencia de datos para la manipulación, análisis y generación de informes a partir de grandes conjuntos de datos. Es fundamental comprender cómo funcionan estos procesos para trabajar de manera eficaz con los datos en un entorno de ciencia de datos.

Olvídate del pasado con NoSQL.

NoSQL (Not Only SQL) es una alternativa a las bases de datos relacionales que se ha desarrollado para abordar los desafíos de las aplicaciones contemporáneas y las necesidades de los equipos de data science. Las bases de datos relacionales han sido fundamentales en las transacciones digitales y los almacenes de datos tradicionales han sido esenciales para las analíticas corporativas. Sin embargo, las bases de datos relacionales presentan limitaciones en términos de flexibilidad y capacidad de manejo de datos complejos y heterogéneos.

La principal característica de NoSQL es su enfoque no relacional. A diferencia de las bases de datos relacionales que requieren un esquema definido antes de ingresar los datos, en NoSQL no es necesario seguir una estructura predefinida. Esto permite manejar datos con formatos diversos, como audio, texto y video, sin la necesidad de una planificación exhaustiva. En lugar de organizar los datos en tablas con relaciones, en NoSQL se pueden almacenar los datos como "agregados", que contienen toda la información relevante en una sola transacción.

Las ventajas de NoSQL para los equipos de data science son notables. Algunas de estas ventajas incluyen:

1. Flexibilidad: NoSQL permite agregar, modificar o eliminar campos sin necesidad de rediseñar toda la base de datos. Esto es especialmente útil en situaciones donde los requisitos y la estructura de los datos cambian con frecuencia.

2. Escalabilidad: Las bases de datos NoSQL están diseñadas para funcionar en clústeres de servidores, lo que permite manejar grandes volúmenes de datos y soportar aplicaciones que crecen en escala corporativa.

3. Rendimiento: Al eliminar la necesidad de definir relaciones complejas y seguir una estructura rígida, las bases de datos NoSQL pueden ofrecer un rendimiento más rápido en ciertos escenarios.

4. Tratamiento de datos no estructurados: NoSQL es adecuado para manejar datos no estructurados o semi-estructurados, como registros de redes sociales, registros de sensores, texto libre, etc.

5. Simplicidad de desarrollo: La falta de un esquema fijo y la capacidad para almacenar datos como agregados hacen que sea más sencillo desarrollar aplicaciones que evolucionan con el tiempo.

NoSQL no es una negación completa de SQL, sino más bien una respuesta a las limitaciones de las bases de datos relacionales. NoSQL abarca una variedad de tecnologías y enfoques, como bases de datos de documentos, bases de datos clave-valor, bases de datos de columnas, etc. Cada tipo de base de datos NoSQL tiene sus propias características y casos de uso apropiados.

En resumen, NoSQL ha surgido como una solución valiosa para los equipos de data science que necesitan manejar grandes volúmenes de datos de manera flexible y escalable. Su diseño no relacional, la capacidad de almacenar datos sin estructuras fijas y su aptitud para el tratamiento de datos heterogéneos lo convierten en una herramienta poderosa para abordar los desafíos actuales en el campo de la ciencia de datos y la analítica.

Aborda los problemas de los datos masivos.

"Big Data" se refiere al conjunto de datos que es tan grande, complejo y diverso que supera la capacidad de las herramientas tradicionales de procesamiento y gestión de datos. Esta sección aborda cómo abordar los problemas y desafíos que conlleva el manejo de datos masivos:

El término "big data" a menudo se confunde con la ciencia de datos. Sin embargo, la ciencia de datos utiliza el método científico para analizar los datos disponibles, y no es necesario que los datos sean masivos para poder hacer preguntas y obtener información valiosa. A pesar de ello, los datos masivos ofrecen una fuente rica de información que puede ser aprovechada para responder a preguntas que conjuntos más pequeños de datos no podrían abordar.

El concepto de "big data" se originó como un "problema de datos masivos", lo que implica que la magnitud de la información es tan grande que excede la capacidad de almacenamiento y procesamiento del hardware y las herramientas convencionales. A menudo se utilizan las cuatro "V" para determinar si los datos se consideran realmente "big data":

1. Volumen: Los datos deben ser de un volumen significativamente alto, generalmente en el rango de petabytes o exabytes.

2. Variedad: Los datos provienen de diversas fuentes y pueden ser de diferentes tipos, como texto, imágenes, videos, etc.

3. Velocidad: Los datos llegan a gran velocidad y deben ser procesados en tiempo real o cerca de tiempo real.

4. Veracidad: Los datos deben ser confiables y precisos para ser útiles en el análisis.

En el contexto del "big data", la idea es que los datos son tan extensos y complejos que se necesita un enfoque diferente para gestionarlos y extraer información valiosa. Sin embargo, es importante destacar que no todas las empresas que trabajan con datos grandes tienen un problema real de "big data". Cumplir con las cuatro "V" es crucial para determinar si se enfrenta a un verdadero problema de datos masivos.

Es necesario tener en cuenta que "big data" no es sinónimo de ciencia de datos. Aunque los datos masivos permiten hacer preguntas más complejas y profundas, no todas las preguntas relevantes requieren conjuntos de datos extremadamente grandes. La ciencia de datos se trata de formular las preguntas correctas y aplicar métodos analíticos adecuados para extraer información valiosa de los datos disponibles, sin importar su tamaño.

Un ejemplo de un problema de "big data" es el de los autos sin conductor. Estos vehículos deben recopilar y procesar enormes cantidades de datos en tiempo real, incluyendo videos, sonidos, informes de tráfico y datos de GPS, para tomar decisiones rápidas y seguras en la carretera. En este caso, el procesamiento de datos en tiempo real y la toma de decisiones basada en una amplia variedad de fuentes de datos representan un desafío real de "big data".

En resumen, "big data" se refiere a conjuntos de datos extremadamente grandes y complejos que presentan desafíos específicos en términos de almacenamiento, procesamiento y análisis. Aunque no todas las empresas tienen un problema real de "big data", la ciencia de datos puede ser aplicada para hacer preguntas interesantes y extraer información valiosa de los datos disponibles, independientemente de su tamaño.

3.- Conoce los distintos tipos de datos.

Organiza la información con los datos estructurados.

Los datos utilizados en proyectos de ciencia de datos pueden clasificarse en diferentes tipos según su estructura y naturaleza. La elección de cómo almacenar estos datos es fundamental para garantizar un análisis eficiente y preciso. A continuación, se describen los tres tipos principales de datos: estructurados, semiestructurados y no estructurados.

Datos Estructurados: Son datos que tienen un formato específico y siguen un orden predefinido. Se asemejan a los ladrillos y el cemento en el mundo de las bases de datos. Un ejemplo común son las hojas de cálculo de oficina, donde cada entrada debe seguir reglas estrictas de formato. Los datos estructurados se basan en un "modelo de datos", que define la estructura de los campos individuales, como si serán campos de texto, números o fechas.

La consistencia en el formato es esencial en los datos estructurados para garantizar una interpretación y análisis coherentes.

Datos semiestructurados: Estos datos no siguen un formato rígido, pero tienen cierta organización que permite identificar elementos y relaciones dentro de ellos. A menudo se almacenan en formatos como JSON o XML, que permiten una mayor flexibilidad en la representación de la información. Un ejemplo es la información en redes sociales, que puede incluir texto, imágenes y etiquetas.

Datos No Estructurados: Son datos que carecen de un formato fijo o de una organización definida. Ejemplos comunes son imágenes, videos y audios. Estos tipos de datos no tienen una estructura intrínseca que permita su almacenamiento y análisis directo en bases de datos relacionales.

La elección de cómo almacenar los datos depende del tipo de datos con el que se está trabajando y de los objetivos del análisis. Las bases de datos relacionales son ideales para almacenar datos estructurados, como registros de transacciones o información tabular. Sin embargo, carecen de flexibilidad para manejar datos semiestructurados o no estructurados de manera eficiente.

Por otro lado, las bases de datos NoSQL ofrecen una mayor flexibilidad para manejar diferentes tipos de datos, incluidos los datos semiestructurados y no estructurados. Esto es especialmente útil en proyectos que requieren la combinación de diferentes fuentes de datos, como redes sociales, imágenes y videos. Aunque trabajar con datos no estructurados en bases de datos NoSQL puede ser más desafiante a la hora de crear informes y realizar análisis, brinda la capacidad de gestionar una variedad más amplia de datos.

En última instancia, el equipo de ciencia de datos debe considerar el tipo de datos con el que trabajan y elegir la tecnología de almacenamiento adecuada para maximizar la eficiencia en el análisis y la obtención de información valiosa. La ciencia de datos se basa en aplicar el método científico a los datos disponibles, y la elección del enfoque de almacenamiento es una parte fundamental de este proceso.

Comparte datos semiestructurados.

Los equipos de ciencia de datos se encuentran con una variedad de tipos de datos en sus proyectos, y la elección de la tecnología adecuada para almacenar y trabajar con estos datos es esencial. Las bases de datos relacionales son ideales para datos estructurados, que tienen un formato específico y siguen un orden predefinido. Con un modelo de datos estricto, los datos estructurados encajan en el esquema de la base de datos, y se pueden crear informes fácilmente utilizando lenguajes de consulta como SQL.

Sin embargo, la mayoría de las aplicaciones y conjuntos de datos no son tan simples y pueden contener datos semiestructurados. Los datos semiestructurados son aquellos que tienen cierta organización pero no siguen un formato fijo. Un ejemplo es la información en redes sociales, donde hay elementos comunes como emisor y destinatario, pero el

contenido y los nombres pueden variar. Estos datos se almacenan en formatos como XML o JSON.

La necesidad de trabajar con datos semiestructurados a menudo surge cuando se deben compartir datos con sistemas que tienen esquemas ligeramente diferentes. Por ejemplo, imaginemos una tienda en línea de zapatillas que necesita intercambiar datos con una compañía de envíos. Aunque ambas fuentes utilizan datos estructurados, pueden tener nombres de campos y formatos ligeramente diferentes. Para resolver esto, se intercambian datos semi estructurados que incluyen tanto la información como los nombres de campo necesarios.

El uso de datos semiestructurados permite a los equipos de ciencia de datos hacer preguntas más interesantes y combinar diferentes fuentes de datos. Por ejemplo, si se quiere evaluar la satisfacción del cliente, se pueden descargar datos semiestructurados de las redes sociales y combinarlos con los datos estructurados de los clientes para obtener una imagen completa. Si se detecta insatisfacción, se puede ofrecer una solución como un cupón de resarcimiento.

En la actualidad, los formatos de intercambio de datos semiestructurados más comunes son XML y JSON. JSON es especialmente popular en servicios web y es probable que una página web de zapatillas reciba datos en formato JSON de un proveedor de envíos.

En resumen, los datos semi estructurados son una parte esencial de la ciencia de datos moderna, ya que permiten a los equipos trabajar con datos más diversos y realizar análisis más sofisticados al combinar diferentes fuentes de información. La capacidad de gestionar tanto datos estructurados como semiestructurados es esencial para obtener conocimientos más profundos y valiosos.

Recopila datos no estructurados.

En la ciencia de datos, los equipos trabajan con tres tipos principales de datos: estructurados, semiestructurados y no estructurados. Hasta ahora hemos discutido los dos primeros tipos, y ahora vamos a profundizar en los datos no estructurados, que abarcan una gran cantidad de información diversa y que según algunos analistas constituyen aproximadamente el 80% de los datos.

Los datos no estructurados son aquellos que no siguen un formato fijo ni una estructura coherente. Ejemplos de datos no estructurados incluyen mensajes de voz, fotos, videos, documentos de texto, presentaciones y contenido en línea, como los resultados de búsqueda en motores como Google y Bing. La característica clave de estos datos es que carecen de un esquema o modelo de datos establecido, lo que dificulta su organización y análisis.

La búsqueda en motores como Google y Bing es un ejemplo de cómo se aborda el desafío de trabajar con datos no estructurados. Estos motores utilizan algoritmos y técnicas avanzadas para buscar y mostrar resultados relevantes que incluyen texto, videos,

imágenes y más. Trabajar con datos no estructurados es un reto interesante en la ciencia de datos y requiere soluciones creativas.

Para manejar datos no estructurados de manera eficiente, se utilizan tecnologías y herramientas modernas. Las bases de datos NoSQL son particularmente útiles para almacenar archivos grandes, como audio, video, imágenes y texto. Estas bases de datos permiten almacenar todos estos archivos en un solo clúster y escalar horizontalmente según sea necesario. Además, herramientas de big data como Hadoop, MapReduce y Apache Spark se utilizan para procesar y analizar grandes cantidades de datos no estructurados.

Imaginemos que eres parte de un equipo de ciencia de datos que trabaja con una tienda en línea de zapatillas. Tu objetivo es obtener una visión completa de los clientes y sus motivaciones. Puedes recopilar datos no estructurados de diversas fuentes, como redes sociales, para obtener información adicional sobre los clientes. Por ejemplo, si un cliente comparte un video corriendo una maratón, puedes usar esa información para enviarle una felicitación personalizada. También puedes analizar publicaciones de amigos y otros datos no estructurados para identificar patrones y comportamientos que te ayuden a comprender mejor a los clientes y a ofrecerles promociones específicas.

A medida que recopilas y analizas más datos no estructurados, puedes hacer preguntas más sofisticadas sobre los clientes. Por ejemplo, podrías determinar si tus clientes son viajeros frecuentes, si tienen una inclinación competitiva o si disfrutan salir a cenar. Estas preguntas te permiten personalizar tus ofertas y mejorar la interacción con los clientes.

En resumen, los datos no estructurados representan una gran parte del panorama de datos y ofrecen oportunidades emocionantes para las empresas en términos de comprensión del cliente, personalización y toma de decisiones informadas. Aunque trabajar con estos datos puede ser un desafío, las tecnologías y herramientas modernas están disponibles para ayudar a los equipos de ciencia de datos a aprovechar al máximo su potencial.

Criba los datos que no utilices.

Cuando se trabaja con datos no estructurados, surge la pregunta de si es necesario filtrar o eliminar parte de los datos. Este es un desafío importante, ya que el enfoque correcto puede influir en la calidad de los análisis y las preguntas que se pueden formular como equipo de ciencia de datos. La decisión de mantener o descartar datos puede tener argumentos a favor y en contra.

Algunos analistas de datos argumentan que siempre es posible encontrar nuevas preguntas que hacer con los datos, incluso si en un principio no se conocían. Además, el costo de almacenar grandes cantidades de datos no estructurados es relativamente bajo, lo que hace que mantenerlos sea económicamente viable. Sin embargo, esto también puede llevar a la acumulación de datos innecesarios, lo que se conoce como "ruido en los datos". Cuanta más información irrelevante haya, más difícil será encontrar resultados interesantes.

Por otro lado, algunos analistas sugieren que es recomendable eliminar datos no esenciales. La acumulación de datos innecesarios puede dificultar la identificación de patrones y tendencias significativas. Un exceso de datos también puede hacer que los informes sean menos claros y precisos. Algunos científicos de datos se enfrentan al desafío de gestionar esta "basura" en los clústeres de big data.

Un ejemplo concreto ilustra este desafío. Imagina una empresa que recopila datos sobre el comportamiento de los usuarios en su sitio web, incluidos clics en imágenes y anuncios. La empresa implementa un sistema de etiquetas para registrar estos datos. Sin embargo, con el tiempo, este sistema de etiquetas acumula una gran cantidad de datos, muchos de los cuales resultan obsoletos o difíciles de comprender. Algunos miembros del equipo argumentan que es necesario limpiar y organizar estos datos para evitar el "ruido", mientras que otros consideran que no es una prioridad debido a la relativa facilidad y bajo costo de almacenamiento.

La decisión de mantener o eliminar datos no estructurados es compleja y depende de varios factores, como los objetivos del análisis, el presupuesto y la capacidad técnica del equipo. No existe una respuesta única y correcta, pero es esencial que el equipo de data science tome una decisión coherente y establezca una política de retención de datos desde el principio. Si los datos se mantienen sin filtrar, se deben desarrollar técnicas y herramientas adecuadas para limpiar y organizar los datos antes del análisis. Por otro lado, si se opta por la eliminación selectiva de datos, es importante asegurarse de que el proceso sea transparente y bien documentado para evitar la pérdida de información valiosa en el futuro.

4.- Aplica la estadística descriptiva.

Empieza con la estadística descriptiva.

La estadística descriptiva es una herramienta crucial en el arsenal de un equipo de ciencia de datos. Este proceso implica recoger, depurar y almacenar datos, formular preguntas y crear informes utilizando conceptos y técnicas estadísticas para comprender mejor la información. La estadística es esencial para contar historias a partir de los datos, aunque es importante recordar que las estadísticas son una herramienta para comunicar una historia, no el objetivo final de la historia en sí.

Un ejemplo humorístico ilustra cómo las estadísticas pueden ser utilizadas para contar historias de manera engañosa. Un chiste sobre un elefante escondido en un árbol destaca cómo es posible esconder "elefantes" (interpretaciones sesgadas o inexactas) en los datos si no se examinan adecuadamente.

Las estadísticas son similares a la narración, en el sentido de que pueden llenarse de hechos, ficciones y fantasías. En el ámbito político, por ejemplo, los candidatos pueden usar estadísticas descriptivas para respaldar sus argumentos. Un candidato podría afirmar que el salario promedio aumentó en $5000 en los últimos cuatro años, mientras que otro podría afirmar que las familias de clase media han perdido $10,000 en ingresos durante el

mismo período. Ambos candidatos utilizan estadísticas para contar historias diferentes que respaldan sus agendas.

Las estadísticas descriptivas involucran conceptos como la media y la mediana. La media es el promedio de un conjunto de valores, mientras que la mediana es el valor medio en una distribución de datos. La mediana es útil para evitar que valores extremadamente altos o bajos distorsionen la imagen general, lo que podría ocurrir al calcular la media. Si hay una gran diferencia entre la media y la mediana, puede indicar un sesgo en los datos.

En el equipo de data science, es crucial cuestionar y analizar las historias que se presentan utilizando estadísticas. Siempre es importante revisar las justificaciones detrás de las afirmaciones estadísticas y buscar diferentes formas de describir los datos. Al cuestionar y explorar los datos de manera crítica, es posible identificar "elefantes escondidos" y obtener una comprensión más precisa de la información. Las estadísticas cuentan múltiples historias y es esencial estar atento a las diferentes interpretaciones que pueden surgir.

Entiende la probabilidad.

La probabilidad es un componente esencial de la estadística que permite calcular las posibilidades de que un evento ocurra. En otras palabras, la probabilidad mide los resultados posibles de un evento y cuán probable es que ocurra uno en particular. Un ejemplo simple es el lanzamiento de una moneda, donde la probabilidad se utiliza para predecir en qué lado caerá. La distribución probabilística es un componente clave en la estadística de probabilidad y se relaciona con la frecuencia relativa de los resultados.

En el caso de lanzar un dado de seis caras, hay seis resultados posibles y la probabilidad de que salga un número específico es de uno en seis. Esto se traduce en un 17 por ciento de probabilidad para cada número. Además de calcular la probabilidad de eventos individuales, también se puede calcular la probabilidad de una secuencia de eventos, como sacar el mismo número dos veces seguidas. Por ejemplo, la probabilidad de sacar un número específico en dos lanzamientos consecutivos del dado sería el 17 por ciento del 17 por ciento, lo que equivale al tres por ciento.

La probabilidad es una herramienta valiosa para equipos de data science en la analítica predictiva. Se utiliza para calcular la probabilidad de que un cliente tenga cierto comportamiento, lo que es útil para tomar decisiones informadas y estratégicas. Por ejemplo, en una empresa de biotecnología, se podría utilizar la probabilidad para predecir la probabilidad de que un paciente participe en un ensayo clínico. Esta información es importante para optimizar la participación de los pacientes y asegurarse de que los ensayos clínicos estén bien poblados.

La probabilidad también puede ayudar en la toma de decisiones difíciles. En el ejemplo mencionado, la empresa de biotecnología debe equilibrar la probabilidad de participación de los pacientes con la precisión de los datos obtenidos. Por ejemplo, si un análisis de sangre es más preciso pero lleva a menos participantes debido al rechazo de los pinchazos, la empresa debe decidir si es mejor tener más participantes o datos más precisos. El análisis de probabilidad puede ayudar a determinar la mejor estrategia.

Trabajar con probabilidad puede llevar a resultados inesperados y plantea preguntas interesantes. Es importante estar dispuesto a explorar diferentes direcciones y afrontar giros inesperados en el análisis. La ciencia de datos aplica el método científico a los datos y, a veces, esto puede llevar a revelaciones sorprendentes y valiosas. En última instancia, la probabilidad es una herramienta poderosa para comprender y tomar decisiones basadas en datos

Busca correlaciones para mejorar resultados.

La correlación es una herramienta importante en el análisis estadístico y se utiliza para medir el grado de relación entre dos variables. Las empresas la utilizan para predecir comportamientos y tomar decisiones informadas. Por ejemplo, plataformas como Netflix y Amazon utilizan la correlación para ofrecer recomendaciones personalizadas a sus usuarios.

La correlación se mide en una escala de uno a cero. Una correlación de uno indica una relación fuerte y positiva entre las variables, mientras que una correlación de cero significa que no hay relación entre ellas. Además, la correlación puede ser positiva o negativa. Una correlación positiva indica que a medida que una variable aumenta, la otra también lo hace. Por ejemplo, la altura y el peso de una persona tienen una correlación positiva, ya que generalmente, cuanto más alta es una persona, más pesa. Por otro lado, una correlación negativa indica que a medida que una variable aumenta, la otra disminuye. Por ejemplo, el peso de un automóvil y su rendimiento de kilometraje de combustible tienen una correlación negativa, ya que a medida que el peso del automóvil aumenta, su eficiencia de combustible disminuye.

La ciencia de datos utiliza la correlación para analizar las relaciones entre variables en un conjunto de datos. Los equipos de data science buscan correlaciones para comprender mejor las relaciones entre diferentes variables y cómo influyen entre sí. Los programas informáticos pueden calcular correlaciones utilizando fórmulas como el coeficiente de correlación, que proporciona un número que indica la fuerza y la dirección de la relación entre las variables.

Un ejemplo práctico de cómo se utiliza la correlación en la ciencia de datos es el caso de LinkedIn. La plataforma utiliza correlaciones entre las conexiones de los usuarios para sugerir posibles conexiones adicionales. Los equipos de data science buscan patrones y relaciones entre las conexiones, como la escuela a la que asistieron, los trabajos compartidos y los intereses comunes. Estas correlaciones ayudan a LinkedIn a brindar recomendaciones relevantes a los usuarios para conectarse con profesionales afines.

La correlación también puede poner a prueba suposiciones y revelar relaciones que pueden no ser evidentes a simple vista. Por ejemplo, podría haber una correlación negativa entre el tiempo que un cliente pasa en un sitio web y su nivel de satisfacción, ya que los clientes insatisfechos podrían pasar más tiempo buscando soluciones. Estas relaciones pueden ser valiosas para tomar decisiones estratégicas y mejorar la experiencia del cliente.

En resumen, la correlación es una herramienta poderosa en el análisis de datos que permite identificar relaciones entre variables y comprender cómo interactúan entre sí. Se utiliza para tomar decisiones informadas y crear recomendaciones personalizadas en diversos campos, desde el entretenimiento hasta la gestión de clientes.

La correlación no implica causalidad.

Es importante comprender que la correlación no implica causalidad. Aunque la correlación puede ayudar a identificar relaciones entre variables, no necesariamente significa que una variable sea la causa de cambios en la otra. Esta es una regla general en estadística y análisis de datos. Puede haber factores ocultos o terceras variables que influyan en la relación entre las dos variables que estás analizando.

Un ejemplo ilustrativo es el de una comunidad de jubilados en Florida. A pesar de que haya una alta correlación entre esa comunidad y lesiones graves o muertes, la verdadera causa no es la peligrosidad del lugar, sino la edad avanzada de los residentes. Esta es una clara muestra de cómo una correlación puede llevar a conclusiones erróneas si no se investiga adecuadamente la causalidad subyacente.

En el contexto de un equipo de data science, es fundamental buscar la verdadera causa detrás de las correlaciones identificadas. Por ejemplo, si se observa un aumento en las ventas de calzado deportivo en enero, el equipo debe hacer preguntas pertinentes para entender por qué está ocurriendo esto. Preguntas como si los clientes tienen más ingresos en enero, si se deben a propósitos de año nuevo o si hay otros factores en juego son esenciales para comprender la causalidad.

Para determinar la causalidad, el equipo debe realizar un análisis más profundo y examinar todas las posibles explicaciones. Pueden crear informes detallados que exploren diferentes hipótesis y causas potenciales. En el ejemplo del aumento de ventas de calzado deportivo, el equipo pudo identificar que la causa real era la presencia de nuevos corredores novatos con propósitos de año nuevo.

Sin embargo, es importante reconocer que encontrar la verdadera causalidad puede ser un desafío. Las relaciones espurias, que son relaciones aparentes pero falsas, pueden llevar a conclusiones incorrectas si no se aplican métodos científicos rigurosos. Por lo tanto, es crucial formular preguntas adecuadas, ser consciente de los sesgos y prejuicios, y considerar todas las posibles variables que podrían estar influyendo en la relación.

En resumen, aunque la correlación puede ayudar a identificar patrones y relaciones en los datos, es fundamental recordar que no implica causalidad. Los equipos de data science deben seguir un enfoque científico riguroso para explorar la causalidad y evitar relaciones espurias. Preguntas bien formuladas, análisis exhaustivos y consideración de múltiples variables son esenciales para llegar a conclusiones precisas y útiles.

Combina las técnicas de la analítica predictiva.

La analítica predictiva es una extensión del análisis de datos que se enfoca en utilizar resultados pasados para predecir eventos futuros. Es una forma de utilizar la información recopilada y los patrones identificados para tomar decisiones concretas y realizar predicciones sobre lo que podría ocurrir. Mientras que la ciencia de datos se centra en analizar y comprender los datos, la analítica predictiva toma esos resultados y los convierte en acciones anticipadas.

Un ejemplo claro de analítica predictiva es la predicción del tiempo realizada por los meteorólogos. A través del análisis de datos históricos y la identificación de correlaciones entre diferentes variables atmosféricas, los meteorólogos pueden predecir patrones climáticos y condiciones climáticas futuras. Estas predicciones son posibles gracias al uso de probabilidades y correlaciones para inferir qué eventos climáticos podrían ocurrir.

En la actualidad, con el aumento en la disponibilidad de datos masivos y la mejora en las tecnologías de análisis, la analítica predictiva se ha vuelto más precisa y valiosa. Las organizaciones pueden recopilar y analizar grandes volúmenes de datos para crear modelos predictivos complejos. Por ejemplo, en el caso de una página web de zapatillas, se pueden utilizar millones de tuits relacionados con la actividad de correr para identificar patrones de comportamiento, influenciadores en el mundo del running y eventos clave. Estos datos pueden utilizarse para dirigir estrategias de marketing, identificar oportunidades de negocio y tomar decisiones informadas.

Sin embargo, es importante destacar que la analítica predictiva no puede realizarse de manera efectiva sin un análisis riguroso de los datos históricos. No se trata solo de saltar a las predicciones sin un entendimiento profundo de los datos subyacentes. El análisis de datos es esencial para comprender las tendencias, las correlaciones y las relaciones causales que podrían influir en los resultados futuros.

En un equipo de data science, es fundamental enfocarse en la calidad del análisis de datos como base para las predicciones. Las buenas preguntas sobre los datos, el uso adecuado de herramientas estadísticas y la comprensión de los patrones históricos son fundamentales para crear modelos de analítica predictiva precisos y confiables. No se debe subestimar la importancia del análisis en el proceso de generar predicciones útiles y valiosas.

En resumen, la analítica predictiva es un enfoque que utiliza resultados pasados para predecir eventos futuros. Es una extensión de la ciencia de datos que convierte los conocimientos en acciones concretas. Sin embargo, la analítica predictiva requiere un análisis sólido de los datos históricos para generar predicciones precisas y valiosas. El entendimiento profundo de las correlaciones, patrones y relaciones causales es esencial para tomar decisiones informadas basadas en los datos.

5.- Evita las dificultades comunes en el tratamiento de datos.

Céntrate en el conocimiento.

El "clúster de los sueños" es una analogía que hace referencia a la trampa en la que algunas organizaciones caen al centrarse en la adquisición de hardware costoso y la recolección masiva de datos, en lugar de enfocarse en el conocimiento y la analítica efectiva. Esta analogía se basa en la película "Campo de sueños" en la que un granjero construye un campo de béisbol para que fantasmas de jugadores antiguos vengan a jugar. De manera similar, algunas organizaciones creen que si invierten en hardware y recolectan grandes cantidades de datos, automáticamente obtendrán conocimiento y resultados valiosos.

Sin embargo, esta mentalidad puede llevar a un enfoque erróneo en el proceso de ciencia de datos. La recolección de datos masivos y la inversión en hardware son solo una parte de la ecuación. La verdadera esencia de la ciencia de datos radica en el análisis profundo y la capacidad de hacer preguntas pertinentes a los datos. Simplemente tener una gran cantidad de datos y un clúster de análisis no garantiza la obtención de resultados significativos.

Un ejemplo citado es el caso de la Biblioteca del Congreso de los Estados Unidos, que recolectó una enorme cantidad de tuits sin un plan claro de cómo utilizar esa información. Esta situación es común en muchas organizaciones, donde se invierte en infraestructura y programas, pero se pasa por alto la importancia de tener un equipo de data science que pueda realizar un análisis riguroso y generar conocimiento valioso.

La analítica predictiva y la ciencia de datos requieren un cambio de perspectiva. No se trata solo de invertir en hardware, sino de invertir en la formación y experiencia del equipo de data science. Es crucial que este equipo pueda hacer preguntas significativas, explorar los datos de manera profunda y aplicar el método científico en su enfoque.

Además, no hay una única herramienta o enfoque en la ciencia de datos. Es importante permitir que el equipo utilice una variedad de programas y técnicas, ya que diferentes herramientas pueden ser más efectivas para diferentes tipos de análisis. La flexibilidad y la capacidad de adaptarse a diferentes enfoques son esenciales para obtener resultados significativos.

En resumen, el "clúster de los sueños" es una metáfora que advierte sobre la trampa de centrarse únicamente en la recolección masiva de datos y la inversión en hardware, en lugar de enfocarse en el análisis profundo y la generación de conocimiento valioso a través de la ciencia de datos. El conocimiento y la capacidad de hacer preguntas pertinentes son fundamentales para el éxito en este campo, y la inversión en formación y experiencia del equipo de data science es esencial para lograr resultados significativos.