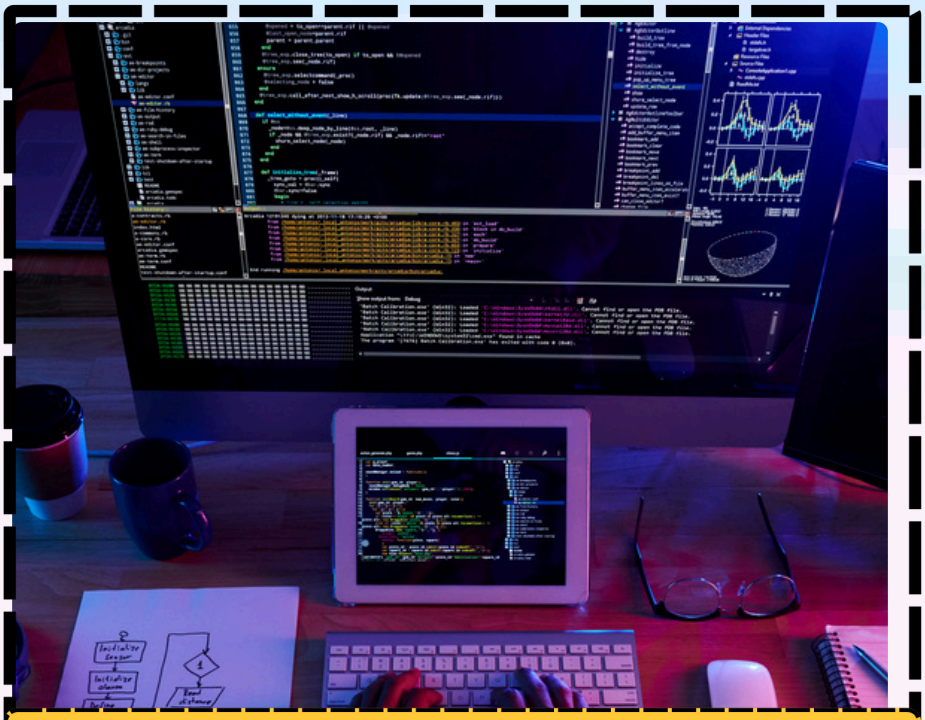
TIC

Lección 3: Tareas del análisis de datos



En esta sesión se detallarán las tareas del análisis de datos, explorando las tareas fundamentales que los científicos de datos realizan para extraer información valiosa de conjuntos de datos.

Tareas del análisis de datos

Previamente se ha estudiado el rol que desempeña cada cargo en análisis de datos, también se han observado algunas tareas del análisis y el paso a paso en una ruta de procesamiento de datos. En la literatura y la documentación, a estas rutas de procesamiento se les conoce como data pipelines. Cuando se crea un data pipeline hay que revisar qué tipos de datos son los que están siendo analizados. No es lo mismo analizar una serie de tiempo que datos de una tabla de datos con muchas características. Tampoco es igual analizar imágenes que textos. Por eso es necesario realizar ciertas tareas asociadas al proceso de datos que ayuden a orientar los análisis y seleccionar los algoritmos y modelos apropiados. No todos los problemas se resuelven con los mismos modelos, en especial, con el auge del aprendizaje profundo de máquina y los modelos de inteligencia artificial, se debe considerar que estos modelos no siempre serán la solución adecuada a problemas de ciencia de los datos.

TIC

Exploración de Datos:

La exploración de datos es la primera etapa esencial en cualquier proyecto de análisis de datos. En las actividades de ejemplo, nos sumergimos en el conjunto de datos de los Mundiales de Fútbol para entender su estructura y características clave. Comenzamos identificando las variables presentes, analizando los tipos de datos que contienen y examinando la distribución general. La visualización juega un papel crucial en esta exploración, permitiéndonos obtener una visión inicial de patrones, tendencias y posibles anomalías. A través de gráficos y diagramas, como histogramas y diagramas de dispersión, podemos revelar información valiosa sobre la naturaleza de los datos y plantear preguntas específicas que guiarán el análisis subsiguiente. Antes de realizar cualquier proceso de analítica de datos es crucial el análisis exploratorio de los datos para conocer qué información hay disponible, que variables están involucradas y qué modelan los datos existentes. También es importante contar con asesoría de expertos en el dominio de los datos. Es decir, si se están analizando datos financieros de una empresa, el contexto de las operaciones y las decisiones que muestran los datos se puede obtener indagando al experto en el tema. Con eso se pueden identificar características en los datos y se puede tener una idea previa antes de la exploración.

TIC

Obtención de datos

Existen diferentes fuentes de información de las que se pueden extraer conjuntos de datos, normalmente en ciencia de datos se trabajan con datos previamente recopilados por lo que no es necesario ir hasta la base donde se miden y se toman variables, sino que se consultan bases de datos o archivos que contienen la información de interés. Una forma de consultar datos es a partir de archivos que los almacenan. Existen varios formatos como son: CSV: El formato CSV (Comma-Separated Values) es uno de los más simples y ampliamente utilizados, donde los datos están organizados en filas y columnas separadas por comas. JSON: El formato JSON (JavaScript Object Notation) es común en el intercambio de datos en la web, ya que es fácilmente legible por humanos y fácilmente procesable por máquinas al representar la información en formato de objeto XML: Similar al formato JSON, este formato contiene etiquetas que permiten a un programa fácilmente separar la información en columnas y características. Existen otros formatos propietarios que usan ciertos programas como puede ser Excel o HDF5 que son formatos de almacenamiento con estructuras muy particulares de ciertos programas.

TIC

También permiten el almacenamiento de los datos junto con análisis, cálculos y otras características que no es posible tener en formatos como CSV, JSON y XML. Los datos pueden ser obtenidos a través de protocolos de comunicación como HTTP que permite la consulta de información desde internet. Para adquirir datos se necesita un servidor que reciba solicitudes HTTP. Una vez la solicitud llega, el servidor responderá con un código que indica si la solicitud fue exitosa o no. En caso de ser exitosa en la respuesta se podrá encontrar un JSON o un XML con la información. Otra forma de almacenar los datos es mediante motores de bases de datos. Para leer los datos se realizan consultas. Las consultas se hacen a través de un estándar llamado SQL (structured query language) que es una serie de reglas para interactuar con una base de datos relacional. Cuando se hace una consulta se pueden obtener los datos en múltiples formatos para trabajar con ellos y hacer tareas de análisis. Solicitudes web: de igual forma como se realizó la descarga del dataset de iris, funcionan las solicitudes web. Una solicitud tiene una dirección web (URL) que identifica el recurso que se quiere obtener, cuando se envía la petición un servidor en internet devuelve un archivo con la información solicitada.