



SECRETARÍA DE
INNOVACIÓN

CIENCIA DE DATOS



SECRETARÍA DE
INNOVACIÓN






Agenda

Sesión 8/18

Exploración de Datos

- ¿Qué es la exploración de los datos? (minería de datos)
- Ventajas la minería de datos
- Quiénes se pueden beneficiar de la minería de datos
- Técnicas de minería de datos
- Python en la exploración de datos



"Descubrimiento de conocimientos
en bases de datos"

¿Qué es la exploración de los datos?

Es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

Es la etapa de análisis de "Knowledge Discovery in Databases" o KDD

El objetivo general del proceso de minería de datos consiste en **extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.**

¿Qué es la exploración de los datos?

Además de la etapa de análisis en bruto, supone aspectos de gestión de datos y de bases de datos, de procesamiento de datos, del modelo y de las consideraciones de inferencia, de métricas de intereses, de consideraciones de la teoría de la complejidad computacional, de post-procesamiento de las estructuras descubiertas, de la visualización y de la actualización en línea.

¿Qué es la exploración de los datos?

El término es un concepto de moda, y es frecuentemente mal utilizado para referirse a cualquier forma de datos a gran escala o procesamiento de la información (recolección, extracción, almacenamiento, análisis y estadísticas), pero también se ha generalizado a cualquier tipo de sistema informático de apoyo a decisiones, incluyendo la inteligencia artificial, aprendizaje automático y la inteligencia empresarial.

En el uso de la palabra, el término clave es el descubrimiento, comúnmente se define como "la detección de algo nuevo".

Proceso y Protocolos de la Minería de Datos

Un proceso, entendiéndose como conjunto de etapas sucesivas, típico de minería de datos consta de los siguientes pasos generales:

1. Selección del conjunto de datos
2. Análisis de las propiedades de los datos
3. Transformación del conjunto de datos de entrada
4. Selección y aplicación de la técnica de minería de datos
5. Extracción de conocimiento
6. Interpretación y evaluación de datos

Proceso y Protocolos de la Minería de Datos

Un proyecto de minería de datos tiene cinco fases necesarias que son, esencialmente:

- Comprensión: del negocio y del problema que se quiere resolver.
- Determinación, obtención y limpieza: de los datos necesarios.
- Creación de modelos matemáticos.
- Validación, comunicación: de los resultados obtenidos.
- Integración: si procede, de los resultados en un sistema transaccional o similar.

Ventajas la exploración de datos

- Descubrir información inesperada, gracias a los algoritmos.
- Analizar enormes cantidades de datos.
- Interpretación fácil de resultados sin necesidad de tener conocimientos avanzados de ingeniería en información.
- Uso de estadísticas para ver si las predicciones que se obtienen son válidas, para no usar los modelos en vano.

Quiénes se pueden beneficiar de la minería de datos

Como hemos aprendido, la minería de datos ayuda a cualquier institución, organización o empresa que genere datos y quiera analizarlos efectivamente. Permite encontrar información que no siempre resulta aparente y puede dar una ventaja competitiva si se usa de forma correcta.

Algunas de las áreas en las que resulta muy útil contar con data mining son:

- Comercio y banca : administración empresarial, segmentación de clientes, predicción y previsión de ventas, análisis de riesgo.
- Farmacia y medicina: diagnóstico de enfermedades y efectividad de tratamientos.

Quiénes se pueden beneficiar de la minería de datos

- Seguridad y detección de fraude: identificaciones biométricas, accesos a redes restringidas, reconocimiento facial, entre otras.
- Recuperar información que no es numérica: obtener texto, imágenes, voz, video y textos de materiales multimedia.
- Astronomía: identificación de nuevas galaxias y estrellas.
- Ambientales: modelos de funcionamiento de ecosistemas para observación, organización y control.
- Sociales: flujos de opinión pública, planificación de ciudades, datos demográficos y poblacionales.

Técnicas de minería de datos

Como ya se ha comentado, las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

Las técnicas más representativas son:

- Redes neuronales.
- Regresión lineal.
- Árboles de decisión.
- Modelos estadísticos.
- Agrupamiento o Clustering.
- Reglas de asociación.

Python en la exploración de datos

Podemos automatizar el proceso de manipular datos con Python. Vale la pena pasar tiempo escribiendo el código que haga estas tareas ya que una vez que se escribió, lo podemos usar una y otra vez en distintos conjuntos de datos que usen un formato similar. Esto hace a nuestros métodos fácilmente reproducibles. También resulta fácil compartir nuestro código con nuestros colegas y ellos pueden replicar el mismo análisis.

Python en la exploración de datos

Nuestros datos

Usaremos los datos de “Portal Teaching” que son subconjunto de los datos estudiados por Instituto de Monitoreo y manipulación experimental de un ecosistema del desierto de Chihuahua cerca de Portal, Arizona, EE. UU.

Usaremos los datos de Portal Project Teaching Database.

Esta sección usa el archivo surveys.csv el cual puede ser descargado desde:

<https://ndownloader.figshare.com/files/2292172>

Python en la exploración de datos

	A	B	C	D	E	F	G	H	I
1	record_id	month	day	year	plot_id	species_id	sex	hindfoot_length	weight
2	1	7	16	1977	2	NL	M	32	
3	2	7	16	1977	3	NL	M	33	
4	3	7	16	1977	2	DM	F	37	
5	4	7	16	1977	7	DM	M	36	
6	5	7	16	1977	3	DM	M	35	
7	6	7	16	1977	1	PF	M	14	
8	7	7	16	1977	2	PE	F		
9	8	7	16	1977	1	DM	M	37	
10	9	7	16	1977	1	DM	F	34	
11	10	7	16	1977	6	PF	F	20	
12	11	7	16	1977	5	DS	F	53	
13	12	7	16	1977	7	DM	M	38	
14	13	7	16	1977	3	DM	M	35	
15	14	7	16	1977	8	DM			
16	15	7	16	1977	6	DM	F	36	
17	16	7	16	1977	4	DM	F	36	
18	17	7	16	1977	3	DS	F	48	

Python en la exploración de datos

Vamos a estudiar la especie y el peso de los animales capturados en sitios dentro de nuestra área de estudio. El conjunto de datos esta guardado en un archivo .csv: cada línea tiene información sobre un solo animal y las columnas representan:

Columna	Descripción
record_id	identificador único de la observación
month	mes de observación
day	día de la observación
year	año de la observación
plot_id	ID de un sitio en particular
species_id	código de dos letras
sex	sexo del animal ("M", "F")
hindfoot_length	tamaño de pata en mm
weight	peso del animal en gramos



A Visual Studio Code

Demostración de

Uso con Python

RESUMEN DE CLASE



SECRETARÍA DE
INNOVACIÓN