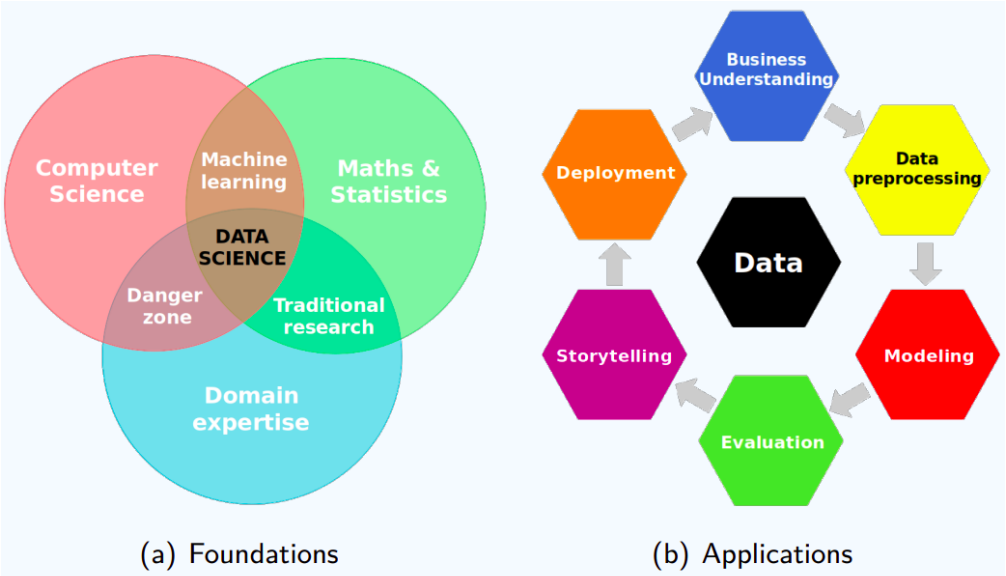
Datos

Minería de Datos - Grado en Matemáticas

DSLAB

2025-09-24

Datos en la Ciencia de Datos



(a) Foundations (b) Applications

Tipos de datos

- Según la **estructuras**: Datos estructurados vs no estructurados
- Según el **comportamiento en el tiempo**: Datos estáticos vs datos dinámicos

Datos estructurados vs no estructurados

- Datos **estructurados**: poseen longitud, tipo, formato y tamaño definidos. Se organizan en formatos de bases de datos, por ejemplo, tablas.

	A	B	C	D	E	F	G	H	I
1	rowid	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
2	1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
3	2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
4	3	Adelie	Torgersen	40.3	18	195	3250	female	2007
5	4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007
6	5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
7	6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
8	7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
9	8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
10	9	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007
11	10	Adelie	Torgersen	42	20.2	190	4250	NA	2007
12	11	Adelie	Torgersen	37.8	17.1	186	3300	NA	2007

- Datos **no estructurados**: Carecen de formato específico. Documentos de texto, vídeo, datos de redes sociales, correos electrónicos, etc. Se almacenan en su formato original y requieren un procesamiento para ser analizados.

Datos estáticos vs dinámicos

- **Estáticos**: no varían a lo largo del tiempo. Ejemplo: censo, datos de natalidad.
- **Dinámicos**: evolucionan con el tiempo. Ejemplo: base de datos de una tienda con productos y precios

Obtención de datos

Recopilación de información en un dominio específico.

Obtención datos --> Procesamiento de datos

Algunas técnicas de obtención de datos:

- Encuestas y entrevistas
- Toma de muestras
- Web scraping
- Sensores y dispositivos IoT
- etc

Fuentes de datos (¡para practicar!)

- UCI Machine Learning repository
- OpenML
- Kaggle
- KEEL dataset repository (Artículo de referencia)
- Penn Machine Learning Benchmarks (Artículo de referencia)
- Eurostat
- Datos abiertos del Gobierno de España

Datos en R

- R incluye en sus librerías distintos conjuntos de datos
- Librería "datasets" contiene bastantes. Para ver la lista completa, basta ejecutar `library(help = "datasets")`

Datos en R

```r
# install.packages('palmerpenguins')
library(palmerpenguins)

str(penguins)
```

```
tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
 $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

¿Con qué tipo de datos vamos a trabajar?

- Datos tabulares: $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$
- Observaciones (filas): items, instancias, puntos, elementos, objetos, etc.
  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$
- Variables (columnas): atributos, características (del inglés *features*)
  $\mathbf{f}_j = \mathbf{x}_j = (x_{1j}, \dots, x_{nj})$

Las variables explicativas de forma matricial:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

Datos

Primer paso: **Entender los datos**

- ¿Cuál es la dimensión de los datos? ¿Cuál es el número de filas (instancias) y de columnas (variables)?
- ¿Qué significan las variables?
- ¿Hay datos erróneos?
- ¿Hay datos faltantes?

¡Practiquemos un poco más con estos datos en R!

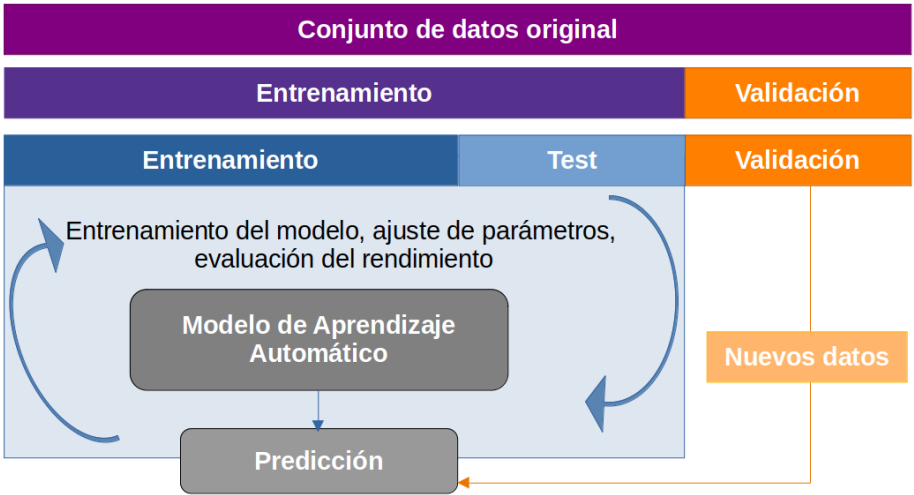
Particiones de los datos

¿Con qué datos entrenamos y evaluamos los modelos de Machine Learning?

Particiones de los datos

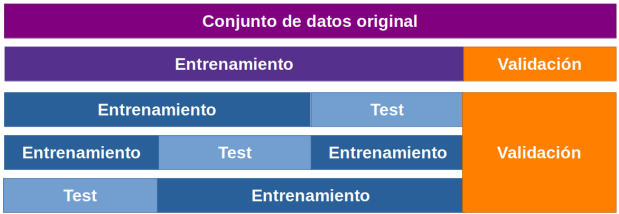
- **Entrenamiento (Training)**: Muestra para entrenar el modelo, el modelo aprenderá el comportamiento de los datos con esta muestra
- **Test**: Para probar el modelo entrenado y comparar el rendimiento en entrenamiento y test. En base a los resultados, se puede cambiar de modelo o realizar ajustes sobre él (reentrenar el modelo)
- **Validación (Validation)**: Para reflejar el comportamiento del modelo en un entorno real con nuevos datos. ¡No se usa para reentrenar!

Particiones de los datos



Particiones de los datos

- Construcción de las particiones: Train 60% - Test 20% - Validación 20% (aproximadamente)
- Los % anteriores dependerán del volumen de los datos y los objetivos del problema
- *k*-fold cross validation. Se obtienen k valores del error –> media y desviación



Particiones de los datos: Muestreo

- ¿Por qué funcionan bien las particiones?
- Muestreo aleatorio
- Muestreo estratificado –> guiado por la variable objetivo

Bases de datos

- **Relacionales**. Siguen el modelo entidad-relación, también llamado modelo relacional, en donde cada una de las tablas (o entidades) presenta algún tipo de enlace con otras (relaciones).
  - SQL: Structured Query Language
- **No relacionales** (no SQL). Representar datos de forma más flexible

Infraestructuras para datos

- Bases de datos relacionales y no relaciones
- Almacenamiento de datos en la nube
- Almacenamiento en memoria
- Almacenamiento distribuido
  - Federated learning

Calidad de los datos

- Clave en cualquier proyecto que involucre datos –> influye directamente en la confiabilidad y el valor de los resultados
- Precisión
- Integridad
- Consistencia
- Relevancia
- Actualización
- Limpieza
- Documentación

Ética, privacidad y seguridad en los datos

- La ética, privacidad y seguridad en los datos son aspectos entrelazados y fundamentales para garantizar que la recopilación, el análisis y el uso de datos se realicen de manera responsable y en beneficio de la sociedad
- ¿Algún ejemplo de falta de ética?
- ¿Algún ejemplo de falta de privacidad?

https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases

Referencias

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., ... & Beard, N. (2019). Integrating ethics within machine learning courses. *ACM Transactions on Computing Education (TOCE)*, *19*(4), 1-26.