Charla introductoria a Ciencia de Datos (DS) con Python

Sebastian Chaparro Cuevas
Estudiante de maestría en Ingeniería
de Sistemas y Computación

Esquema de trabajo en DS

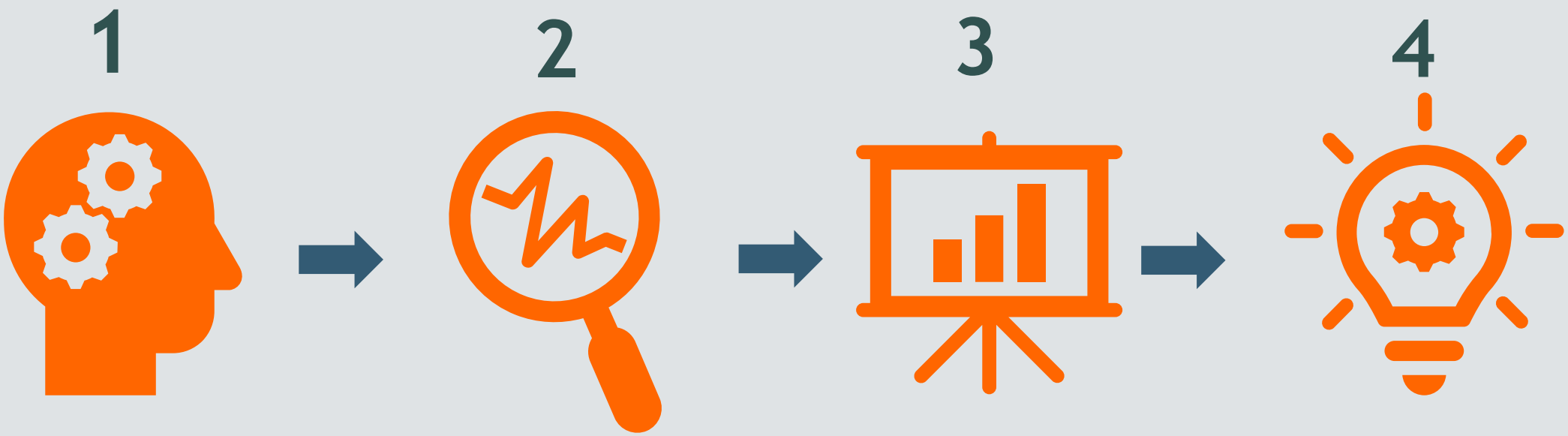
1.  Adquisición de datos, proveniente de distintas fuentes.

2.  Preparación de datos (es necesario limpiarlos, estructurarlos para poder trabajar con ellos).

3.  Exploración de datos: interpretación mediante descripciones o visualizaciones.

4.  Experimentación con los datos y realización de predicciones.

Esquema de trabajo en DS



Esquema de trabajo en DS
Usando aprendizaje supervisado

Paso 1

Fuentes de datos

Pueden provenir de una empresa privada para toma de decisiones y ser restringidos.

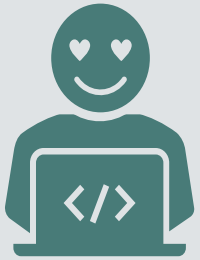
Pueden ser abiertos para toma de decisiones (acceso libre).

Pueden ser recolectados por expertos en un domino, de resultados experimentales o a través de encuestas.

Paso 1

Recopilación de información donde se incluyen *variables objetivo* (etiquetas) y *variables independientes* (características), estos datos se denominarán: conjunto de entrenamiento.

¿Quién es el encargado?



El *ingeniero de datos* se encarga de definir la forma de almacenamiento de los datos y de su mantenimiento.

Los datos se almacenarán en Bases de datos relacionales o no relacionales para guardar la información.

Esquema de trabajo en DS
Usando aprendizaje supervisado

Paso 1

Ejemplo

Alimentos S.A.S contrata una empresa especializada en recolección de datos, está interesada en saber cual será el salario y el cargo al que una persona debería aspirar según su perfil, para ello obtienen la información a través de encuestas.

Adquisición de datos

	Edad	Sexo	Barrio	Nivel educativo	Años de experiencia	Cargo	Salario
Manuel	25	M	Estrada	Universitario	2	Arquitecto	-
Juan	-	-	-	Técnico	25	Programador	2’000.000
Laura	34	F	Candelaria	Universitario	5	Secretaria	1’000.000
David	-	N	-	Bachillerato	100	Digitador	800.000
Sofía	20	F	Colina	Técnico	1	Recepcionista	1’000.000
Pedro	40	M	Norte	Bachillerato	15	Abogado	2’500.000
María	32	F	Álamos	-	8	-	500.000

Esquema de trabajo en DS
Usando aprendizaje supervisado

Paso 2

Se deben **estructurar los datos** porque:

Es necesario identificar variables categóricas y numéricas.

No se suministran algunos datos o son erróneos.

Se pueden filtrar datos que no resulten relevantes.

¿Quién es el encargado?

El *analista de datos*

Limpia, filtra, transforma y procesa los datos que el Ingeniero de datos le suministra.

Comprende e interpretar los datos recopilados.

Describe de manera intuitiva como se relacionan los datos.

Esquema de trabajo en DS
Usando aprendizaje supervisado

Paso 2

Ejemplo

	Edad	Sexo	Barrio	Nivel educativo	Años de experiencia	Cargo	Salario
Manuel	25	M	Estrada	Universitario	2	Arquitecto	-
Juan	-	-	-	Técnico	25	Programador	2’000.000
Laura	34	F	Candelaria	Universitario	5	Secretaria	1’000.000
David	18	N	Restrepo	Bachillerato	100	Digitador	800.000
Sofía	20	F	Colina	Técnico	1	Recepcionista	1’000.000
Pedro	40	M	Norte	Bachillerato	15	Abogado	2’500.000
María	32	F	Álamos	-	8	Escritorio	500.000

	Edad	Sexo	Barrio	Nivel educativo	Años de experiencia	Cargo	Salario
Manuel	25	M	Estrada	Universitario	2	Arquitecto	2’000.000
Juan	50	M	Egipto	Técnico	25	Programador	2’000.000
Laura	34	F	Candelaria	Universitario	5	Secretaria	1’000.000
David	20	M	Restrepo	Bachillerato	1	Digitador	800.000
Sofía	20	F	Colina	Técnico	1	Recepcionista	1’000.000
Pedro	40	M	Norte	Bachillerato	15	Abogado	2’500.000
María	32	F	Álamos	Universitario	8	Ingeniera	500.000

Esquema de trabajo en DS

Usando aprendizaje supervisado

Paso 3

Realización de un análisis exploratorio de los datos, se generan hipótesis.

La visualización es fundamental, para este proceso junto a estadísticas descriptivas (comprensión de los datos).

¿Quién es el encargado?

El *científico de datos* debe tener buenas bases de estadística.

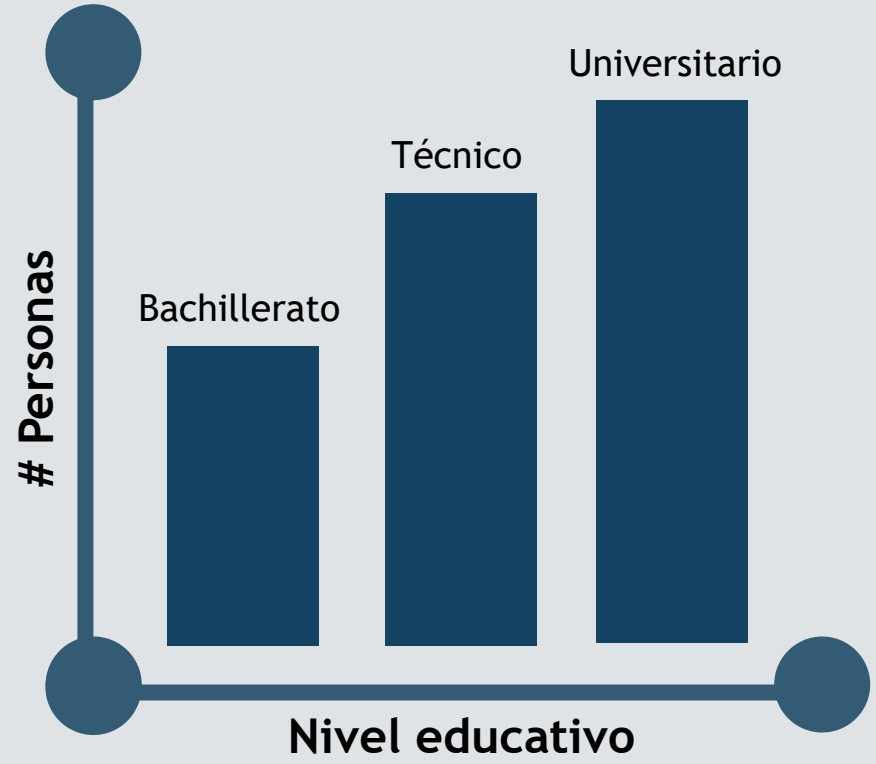
Realiza diversos experimentos y analiza los datos para generar nuevo conocimiento.

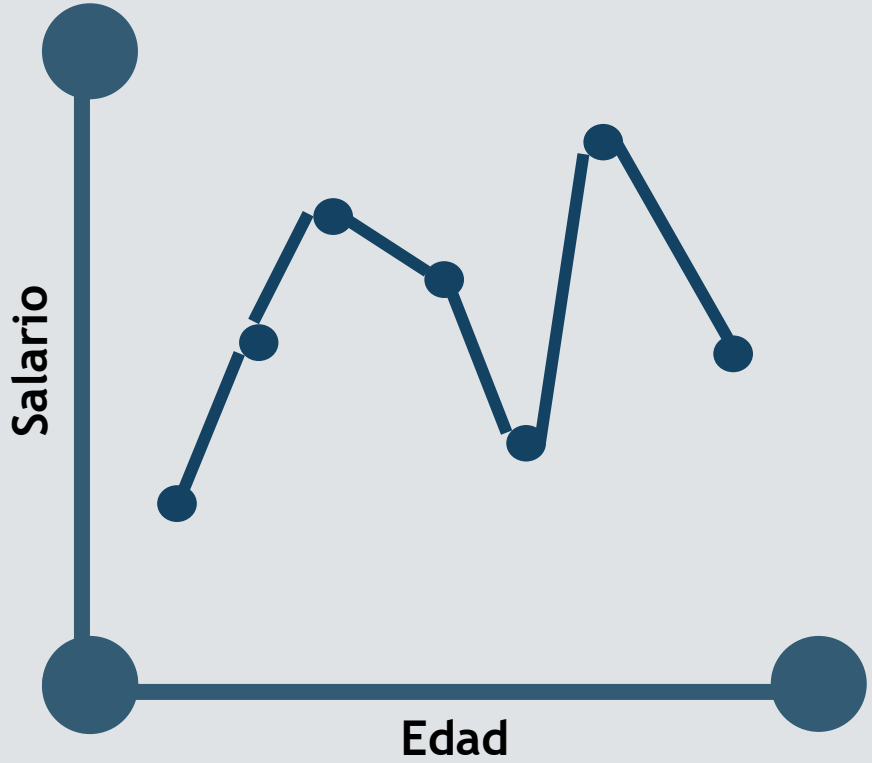
Esquema de trabajo en DS

Usando aprendizaje supervisado

Paso 3

Ejemplo





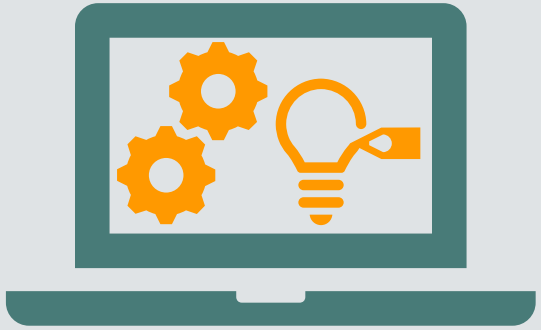
Esquema de trabajo en DS
Usando aprendizaje supervisado

Paso 4

*¿Para qué queremos los datos? (*Es necesario tener una pregunta bien definida (Narrow Learning)).

Si la etiqueta (variable que queremos predecir) es una categoría, el problema es un problema de clasificación.

Si la etiqueta (variable que queremos predecir) es un número real, el problema es un problema de regresión.



¿Quién es el encargado?

El *ingeniero en aprendizaje de máquina* se encarga de realizar a nivel de producción predicciones sobre los datos (requiere utilizar modelos de Aprendizaje de Máquina convencionales).

Esquema de trabajo en DS
Usando aprendizaje supervisado

Paso 4

Ejemplo

¿Cuál es el salario de las personas dadas sus características?

¿Cuál es el cargo que se adapta mejor a la persona?

Para ello es importante:

Un conjunto de datos de entrenamiento.
Un conjunto de datos a los cuales queremos realizar una predicción.

Referencia bibliográfica

El contenido propuesto se basó en los siguientes recursos de Data-camp:

- https://learn.datacamp.com/skill-tracks/python-fundamentals
- https://campus.datacamp.com/courses/data-science-for-everyone/
- https://learn.datacamp.com/skill-tracks/machine-learning-fundamentals-with-python
- https://learn.datacamp.com/courses/data-manipulation-with-pandas
- https://learn.datacamp.com/courses/introduction-to-data-visualization-with-seaborn