Ciencia de Datos

Módulo 3

Repaso Módulo 1 y 2

UTN F r a

GyEE Gestión y Evaluación Educativa

<aprendé programando> Virtual

SECUNDARIA del futuro

Inteligencia Artificial

Fuente: https://www.pandasecurity.com/es/mediacenter/mobile-news/inteligencia-artificial/

Inteligencia Artificial es un campo de estudio que combina matemática, estadística e informática para intentar desarrollar tecnología que pueda mostrar inteligencia humana.

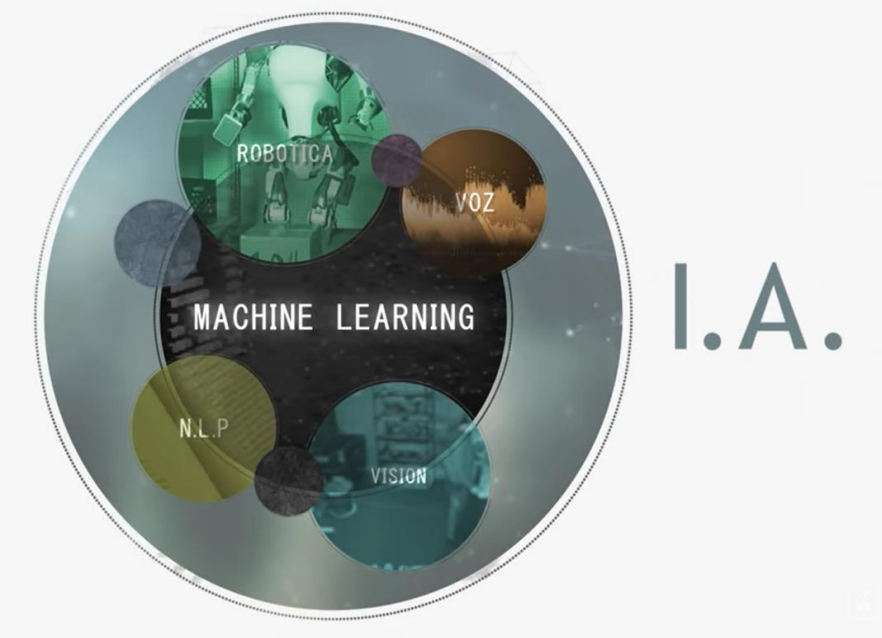
“Crear programas de ordenador o máquinas capaces de conductas que consideraríamos inteligentes si las efectuaran seres humanos”

(Kaplan, Jerry: *Inteligencia artificial. Lo que todo el mundo debe saber*, 2017)

Inteligencia artificial

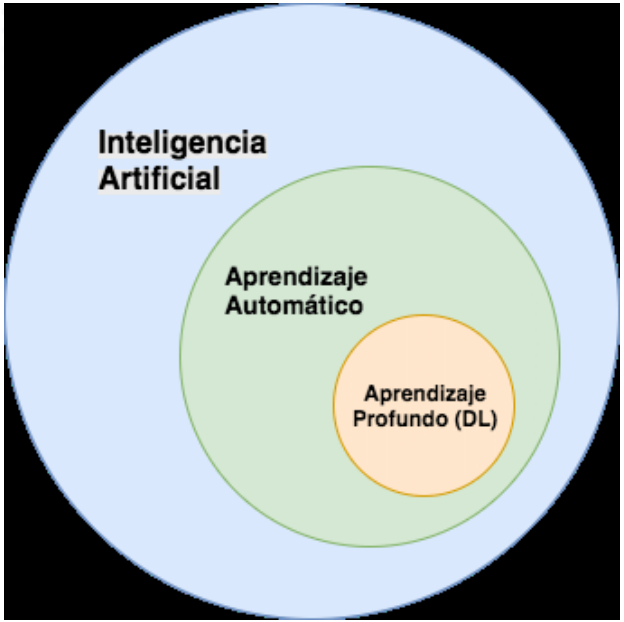
Dentro del campo de Inteligencia Artificial existen diversos subcampos:

- Robótica
- Voz (comunicación con computadoras través del lenguaje humano)
- N.L.P. (Procesamiento de lenguaje natural - comprensión de lenguaje humano)
- Visión (reconocimiento de imágenes, formas, caras, colores, etc.)



Fuente: https://www.youtube.com/watch?v=KytW151dpqU&list=PL-Ogd76BhmcC_E2RjgIIJZd1DQdYHcVf0

Aprendizaje Automático (Machine learning)



Fuente: https://www.researchgate.net/

- Planificación: técnicas para realizar determinadas acciones a través de código dándole instrucciones precisas a las máquinas.

-Machine Learning: conjunto de herramientas y técnicas que aprenden de las experiencias y extraen patrones de los datos.

Aprendizaje Automático

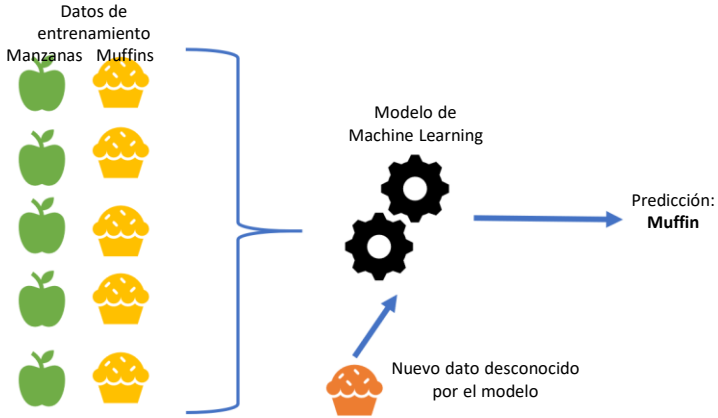
“Campo de estudio que brinda a las computadoras la capacidad de aprender sin ser programadas explícitamente” Arthur samuel, 1959

Fuente: https://www.xatakaciencia.com/computacio/

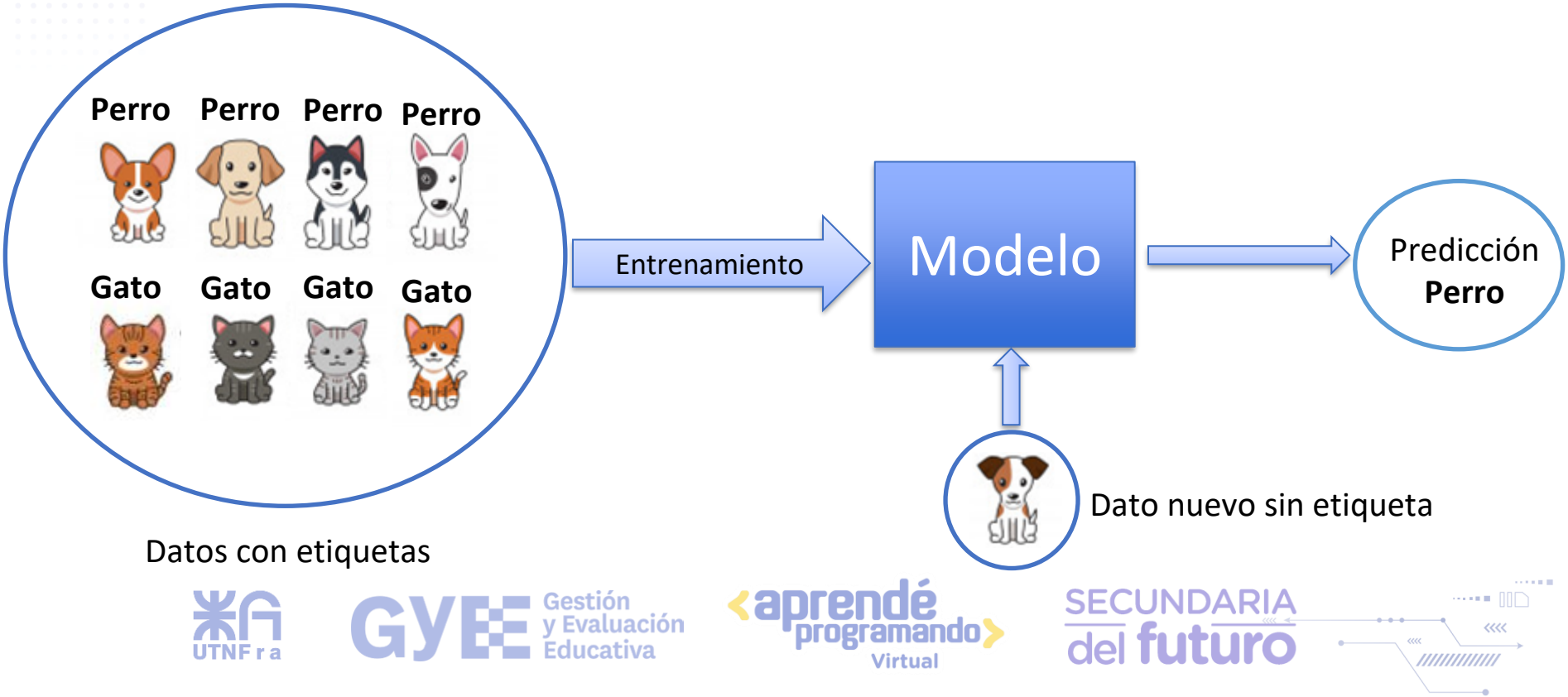
Aprendizaje Automático - Tipos

Supervisado: Se dice que un modelo de ML es supervisado cuando, los datos con los que se entrena el modelo tienen los valores que queremos predecir, a estos se le dice que están etiquetados.

Por ejemplo, si queremos predecir cuántos goles marcará nuestro equipo en base a datos de partidos anteriores, estos datos deben tener cuántos goles marcó nuestro equipo en cada uno de ellos.



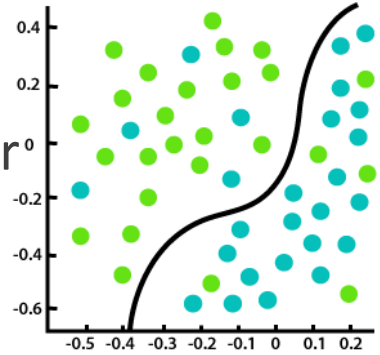
Aprendizaje supervisado



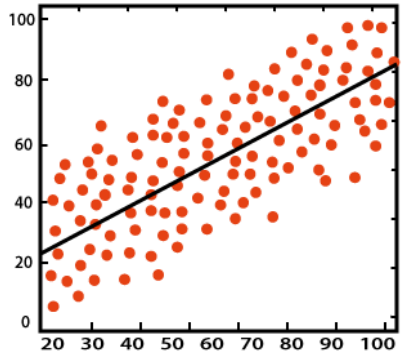
Aprendizaje Automático - Tipos

Los principales usos de los modelos supervisados son:

- Clasificación: Cuando la variable a predecir es una clase, por ejemplo: Enfermo o no enfermo, la raza de la foto de un animal, etc.
- Regresión: Cuando la variable a predecir es un valor, por ejemplo: Precio de un objeto, nota de un estudiante, probabilidad de lluvia, etc.



Clasificación



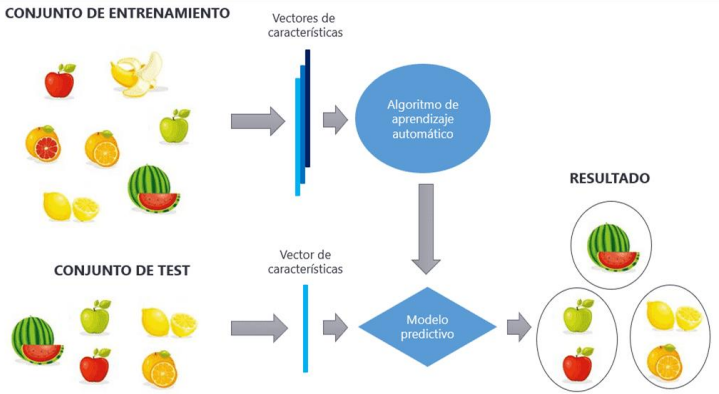
Regresión

Fuente: https://www.javatpoint.com/regression-vs-classification-in-machine-learning

Aprendizaje Automático - Tipos

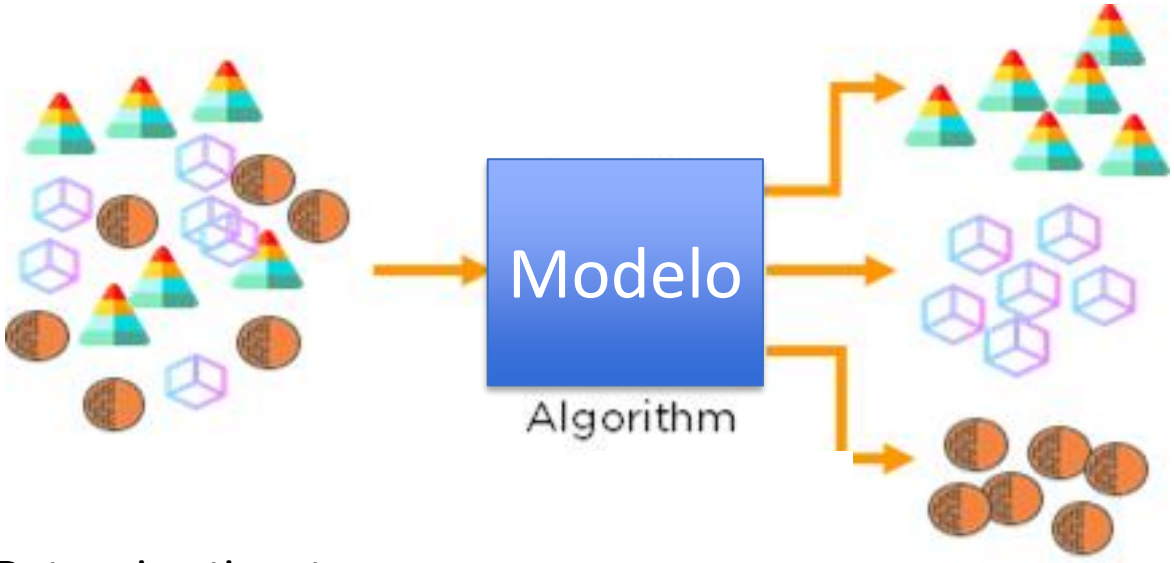
No supervisado: Estos modelos se entrenan sin información sobre el atributo que se quiere predecir, por lo que la evaluación de su desempeño es más compleja. Sueles usarse para realizar agrupaciones o clusters y como métodos de reducción de dimensionalidad.

Por ejemplo, cuando una compañía quiere hacer grupos de clientes según sus gastos mensuales, cantidades de compras, visitas a la tienda, etc.



Fuente: https://www.diegocalvo.es/aprendizaje-no-supervisado/

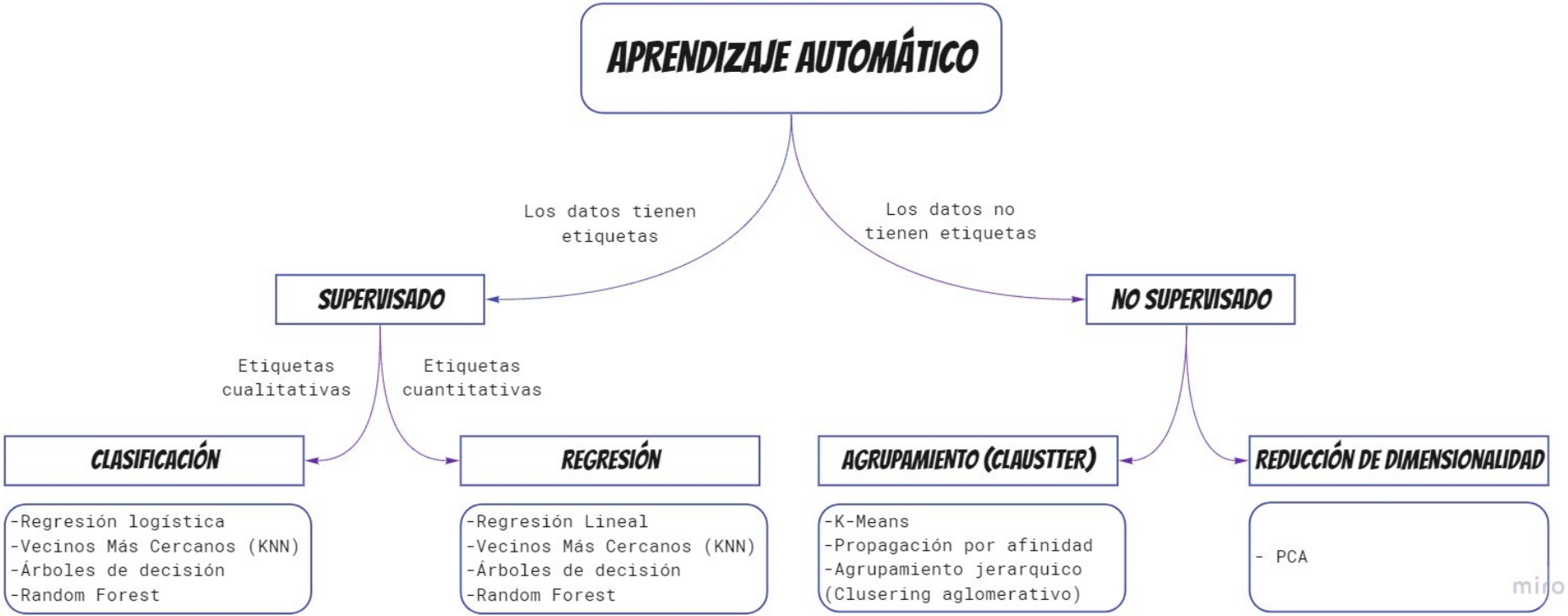
Aprendizaje No supervisado



Datos sin etiquetas

Output

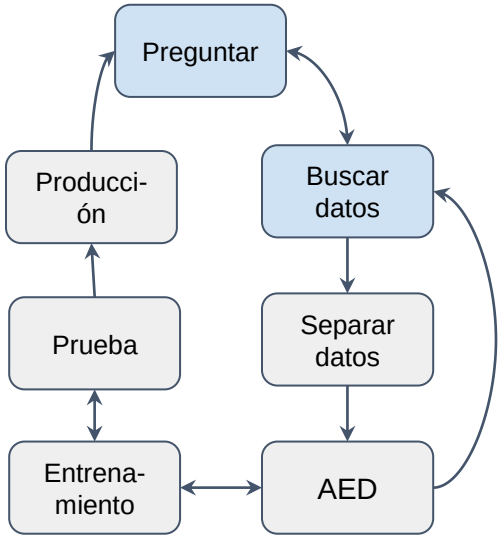
Fuente:https://es.clariba.com/machine-learning-for-business


APRENDIZAJE AUTOMÁTICO
Los datos tienen etiquetas
Los datos no tienen etiquetas
SUPERVISADO
NO SUPERVISADO
Etiquetas cualitativas
Etiquetas cuantitativas
CLASIFICACIÓN
REGRESIÓN
AGRUPAMIENTO (CLAUSTTER)
REDUCCIÓN DE DIMENSIONALIDAD
-Regresión logística
-Vecinos Más Cercanos (KNN)
-Árboles de decisión
-Random Forest
-Regresión Lineal
-Vecinos Más Cercanos (KNN)
-Árboles de decisión
-Random Forest
-K-Means
-Propagación por afinidad
-Agrupamiento jerarquico (Clusering aglomerativo)
- PCA
miro

Aprendizaje Automático - Pasos

En todo modelo de Machine Learning hay algunos **pasos comunes** que nos sirven de guia:

- **Definición del problema**: Siempre que se va a realizar un proyecto de ML es para resolver un problema específico, por ende, primero debemos de realizarnos una pregunta la cual buscaremos resolver.
- **Buscar los datos**: Por lo general debemos definir qué datos necesitamos y de dónde los podemos encontrar. Hoy en día hay muchos datos libres, sean generados por organismos gubernamentales u ONG, o en algunos casos incluso empresas privadas.
- **Análisis exploratorio de datos**: Durante esta etapa se *limpian* los datos y se realiza una exploración de los mismos, buscando características y relaciones que nos ayuden a seleccionar qué variables usar o que modelo usar.

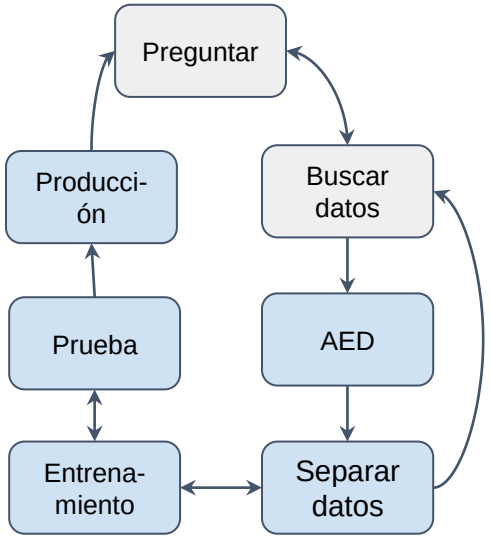


Aprendizaje Automático - Pasos

- **Dividir los datos de entrenamiento y prueba**. Esta división es de vital importancia para poder definir si el modelo tiene un buen rendimiento o no.

- **Entrenamiento del modelo**: En esta etapa se **entrenan** diferentes de modelos y se **prueban** tomando en cuenta alguna métrica previamente definida, seleccionando el modelo que mejor desempeño tenga.

- **Puesta en producción**: Finalmente se realiza la puesta en producción, este paso depende de qué clase de proyecto sea, puede que en algunos casos, colocar en producción solo signifique generar las predicciones para alguna colección de datos, o puede ser implementarlo en un sistema que reciba datos periódicamente.



Proyecto

En este curso vamos a ponernos en Grupos y realizar un Proyecto de Ciencia de Datos con un tema que elijar ustedes.

Tenemos que realizar todos los Pasos de un Proyecto de Ciencia de Datos, el primero es elegir un Tema y hacernos una Pregunta y Buscar los datos.

Algunos lugares de donde se puede extraer los datos:
https://docs.google.com/document/d/1_d7Ubr_CmqWsr5YhRcmwPfBhfi00f2mNz0TI6VX48fM/edit?usp=sharing

Extra - Calidad de los Datos

Los datos son de calidad cuando representan correctamente la construcción del mundo real a la que se refieren.
Los principales aspectos a considerar son:

- Validez
- Precisión
- Integridad
- Consistencia
- Uniformidad

Extra - Datos para entrenamiento del modelo

Antes de entrenar un modelo hay que definir **qué variables** se utilizaran, para esto es posible tener en cuenta las hipótesis, las correlaciones entre las variables, el conocimiento sobre los datos. También es posible probar distintas combinaciones e ir viendo los resultados de los modelos.

En el entrenamiento de un modelo:

- No puede haber valores nulos
- No puede haber variables categóricas

Extra - Integridad - Valores faltantes

Algunas estrategias para resolver este problema son:

Descartar

Si los valores perdidos en una columna rara vez ocurren y ocurren al azar, entonces la solución más fácil y directa es eliminar las observaciones (filas) que tienen valores perdidos.

Si faltan la mayoría de los valores de la columna y se producen al azar, una decisión típica es descartar toda la columna.

Imputar

Significa calcular el valor faltante en base a otras observaciones. Algunos métodos para resolverlo:

Por valores estadísticos - media, mediana, etc

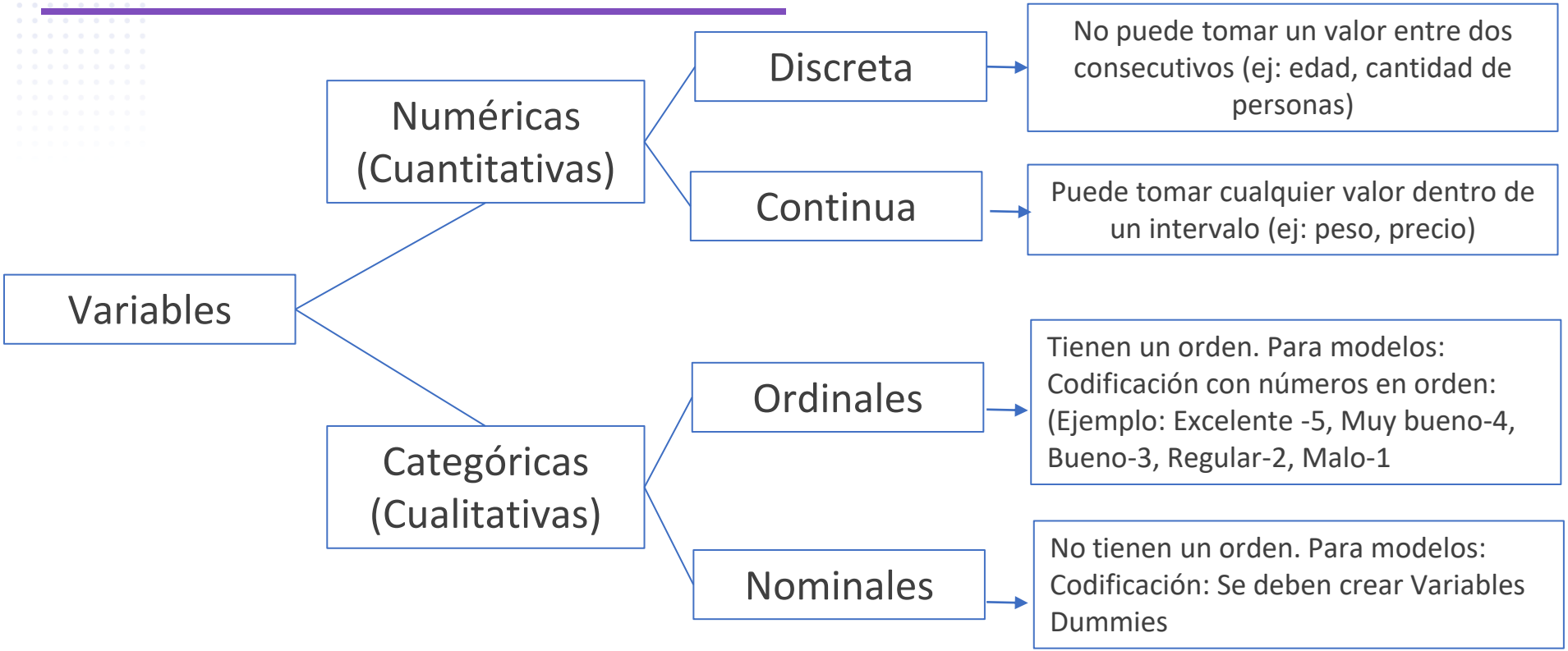
Regresión linear - en relación a otras variables

Copiando valores de otras observaciones similares - solo cuando existen mucha cantidad de datos

Marcar

Significa indicar que el dato está faltando. Algunos argumentan que completar los valores faltantes conduce a una pérdida de información, independientemente del método de imputación que usemos. Esto se debe a que decir que faltan datos es informativo en sí mismo y el algoritmo debería saberlo.

Extra - Variables



Extra - Variables dummies

Variables dummies: Variable ficticias que solo puede tomar como valor 0 y 1. Se utiliza para indicar la presencia o ausencia de valores categóricos

Al tener una variable categórica nominal: se generan columnas de tal manera que cada columna es una categoría que tiene como valor 0 y 1

Ejemplo. Variable Descripción en el dataset Clima:
En la Notebook lo realizaremos

Python:
Función de pandas get_*dummies*
o crear una función
(también hay implementación de Scikit-learn: OneHotEncoder)

	descripcion		Cold	Normal	Warm
0	Cold	0	1	0	0
1	Warm	1	0	0	1
2	Normal	2	0	1	0
3	Cold	3	1	0	0
4	Cold	4	1	0	0

Extra - Definición Datos X e y

Se divide el dataset: creando dos DataFrames: variables que se utilizaran para predecir (X) y la variable a predecir (y)

X

y

	humedad	velocidad_viento_kmh	rumbo_viento_grados	visibilidad_km	presion_mbar	lluvia	temperatura
0	0.92	11.27	130.0	8.05	1021.60	0	-0.56
1	0.73	20.93	330.0	16.10	1017.00	1	21.11
2	0.97	5.97	193.0	14.91	1013.99	1	16.60
3	0.82	3.22	300.0	16.10	1031.59	1	1.60
4	0.60	10.88	116.0	9.98	1020.88	1	2.19
5	0.32	21.46	190.0	10.35	1015.33	1	27.54
6	0.84	7.97	170.0	11.13	1009.04	1	19.98
7	0.86	14.49	30.0	15.13	1009.60	1	11.11
8	0.73	14.01	351.0	15.83	1018.39	1	8.41
9	0.81	6.44	320.0	7.86	1003.89	1	1.70
10	0.88	14.01	141.0	6.02	1021.28	0	-2.22
11	0.60	1.42	204.0	15.83	1019.52	1	21.90
12	0.87	11.03	1.0	14.91	1015.92	1	17.11
13	0.73	4.07	297.0	9.76	1013.06	1	17.77
14	0.39	7.66	35.0	9.98	1025.59	1	24.95

X

y

humedad	velocidad_viento_kmh	rumbo_viento_grados	visibilidad_km	presion_mbar	lluvia
0.92	11.27	130.0	8.05	1021.60	0
0.73	20.93	330.0	16.10	1017.00	1
0.97	5.97	193.0	14.91	1013.99	1
0.82	3.22	300.0	16.10	1031.59	1
0.60	10.88	116.0	9.98	1020.88	1
0.32	21.46	190.0	10.35	1015.33	1
0.84	7.97	170.0	11.13	1009.04	1
0.86	14.49	30.0	15.13	1009.60	1
0.73	14.01	351.0	15.83	1018.39	1
0.81	6.44	320.0	7.86	1003.89	1
0.88	14.01	141.0	6.02	1021.28	0
0.60	1.42	204.0	15.83	1019.52	1
0.87	11.03	1.0	14.91	1015.92	1
0.73	4.07	297.0	9.76	1013.06	1
0.39	7.66	35.0	9.98	1025.59	1

temperatura
-0.56
21.11
16.60
1.60
2.19
27.54
19.98
11.11
8.41
1.70
-2.22
21.90
17.11
17.77
24.95

Extra - División Datos de entrenamiento y testeo

Se divide el dataset en los datos que se utilizaran para entrenar el modelo (X_train e y_train) y los datos que se utilizaran para probar (X_test e y_test). Se utiliza la función de Scikit-Learn **train_test_split()**

X y

Train

humedad	velocidad_viento_kmh	rumbo_viento_grados	visibilidad_km	presion_mbar	lluvia	temperatura
0.92	11.27	130.0	8.05	1021.60	0	-0.56
0.73	20.93	330.0	16.10	1017.00	1	21.11
0.97	5.97	193.0	14.91	1013.99	1	16.60
0.82	3.22	300.0	16.10	1031.59	1	1.60
0.60	10.88	116.0	9.98	1020.88	1	2.19
0.32	21.46	190.0	10.35	1015.33	1	27.54
0.84	7.97	170.0	11.13	1009.04	1	19.98
0.86	14.49	30.0	15.13	1009.60	1	11.11
0.73	14.01	351.0	15.83	1018.39	1	8.41
0.81	6.44	320.0	7.86	1003.89	1	1.70
0.88	14.01	141.0	6.02	1021.28	0	-2.22

Test

humedad	velocidad_viento_kmh	rumbo_viento_grados	visibilidad_km	presion_mbar	lluvia	temperatura
0.60	1.42	204.0	15.83	1019.52	1	21.90
0.87	11.03	1.0	14.91	1015.92	1	17.11
0.73	4.07	297.0	9.76	1013.06	1	17.77
0.39	7.66	35.0	9.98	1025.59	1	24.95

X_train

humedad	velocidad_viento_kmh	rumbo_viento_grados	visibilidad_km	presion_mbar	lluvia
0.92	11.27	130.0	8.05	1021.60	0
0.73	20.93	330.0	16.10	1017.00	1
0.97	5.97	193.0	14.91	1013.99	1
0.82	3.22	300.0	16.10	1031.59	1
0.60	10.88	116.0	9.98	1020.88	1
0.32	21.46	190.0	10.35	1015.33	1
0.84	7.97	170.0	11.13	1009.04	1
0.86	14.49	30.0	15.13	1009.60	1
0.73	14.01	351.0	15.83	1018.39	1
0.81	6.44	320.0	7.86	1003.89	1
0.88	14.01	141.0	6.02	1021.28	0

y_train

temperatura
-0.56
21.11
16.60
1.60
2.19
27.54
19.98
11.11
8.41
1.70
-2.22

X_test

0.60	1.42	204.0	15.83	1019.52	1
0.87	11.03	1.0	14.91	1015.92	1
0.73	4.07	297.0	9.76	1013.06	1
0.39	7.66	35.0	9.98	1025.59	1

y_test

21.90
17.11
17.77
24.95

Extra - Entrenamiento y testeo

Entrenamiento del modelo

Se entrena el modelo con los datos de entrenamiento

X_train e y_train

X_train

humedad	velocidad_viento_kmh	rumbo_viento_grados	visibilidad_km	presion_mbar	lluvia
0.92	11.27	130.0	8.05	1021.60	0
0.73	20.93	330.0	16.10	1017.00	1
0.97	5.97	193.0	14.91	1013.99	1
0.82	3.22	300.0	16.10	1031.59	1
0.60	10.88	116.0	9.98	1020.88	1
0.32	21.46	190.0	10.35	1015.33	1
0.84	7.97	170.0	11.13	1009.04	1
0.86	14.49	30.0	15.13	1009.60	1
0.73	14.01	351.0	15.83	1018.39	1
0.81	6.44	320.0	7.86	1003.89	1
0.88	14.01	141.0	6.02	1021.28	0

y_train

temperatura
-0.56
21.11
16.60
1.60
2.19
27.54
19.98
11.11
8.41
1.70
-2.22

Testeo del modelo

Corre el modelo con los datos de Testeo para ver las predicciones que realiza y se guardan los resultados.

Se compara la predicción del modelo con los datos reales (y_test)

X_test

0.60	1.42	204.0	15.83	1019.52	1
0.87	11.03	1.0	14.91	1015.92	1
0.73	4.07	297.0	9.76	1013.06	1
0.39	7.66	35.0	9.98	1025.59	1

Modelo

y_pred

22,94
16,68
19,52
22,6

Comparación
(testeo – métrica)

y_test

21.90
17.11
17.77
24.95