

Essential Machine Learning with Python

https://www.facebook.com/qslearningperu/?ref=page_internal



QS Learning



+51 937 012 707 / +51 915 111 457



quants.admission@gmail.com

Modelos predictivos con Machine Learning

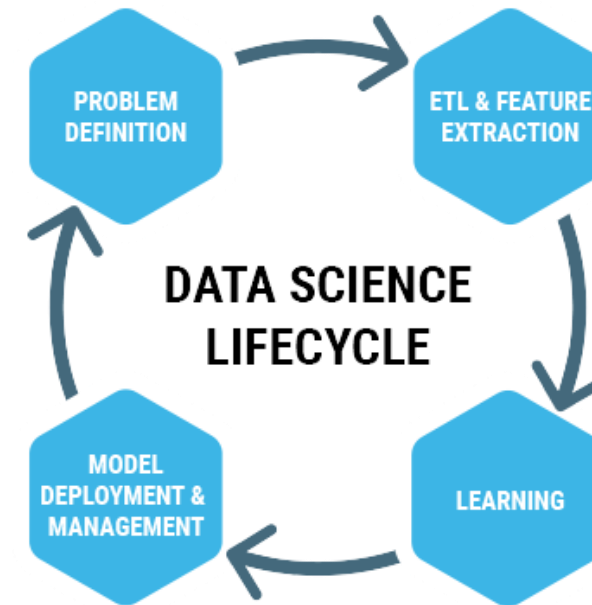
¿Que es un modelo?

No existe una definición única de modelo, y muchas veces depende de la bibliografía que consultas.

- Modelos determinísticos
- Modelos probabilísticos
- Modelos de Machine Learning

El modelado predictivo es un proceso que utiliza [la extracción de datos](#) y la [probabilidad](#) para pronosticar resultados.

Cada modelo se compone de una serie de predictores, que son variables que probablemente influyan en los resultados futuros. Una vez que se han recopilado datos para los predictores relevantes, se formula un modelo estadístico. El modelo puede emplear una ecuación lineal simple, o puede ser una red neuronal compleja, diseñada por un software sofisticado.



Proceso de modelamiento



Entendimiento del negocio

Negocio de la entidad

El entendimiento se obtiene de la indagación y corroboración de:

- ✓ Estructura legal y operativa
- ✓ Objetivos y estrategias del negocio
- ✓ Relaciones e interacciones con sus clientes, proveedores, empleados y sus alianzas
- ✓ Productos y servicios claves
- ✓ Relaciones e interacciones con sus Inversionistas y con entes del estado
- ✓ Actividades de financiación
- ✓ Litigios y reclamos
- ✓ Entre otros

Plan estratégico



Industria y ambiente de la entidad

Se obtiene mediante indagación y corroboración de:

- El ambiente de la industria en donde desarrolla la actividad la entidad
- Ambiente político, económico, social, tecnológico y ambiental.
- Ambiente legal y reglamentario.

“La forma que tiene una empresa de realizar actividades y ganar dinero”

Obtención de datos

Datos internos

- Bases de datos
- Archivos de texto (ó excel, csv, ...)
- Otros (pdfs, imágenes , ¡cualquier cosa!)

Lo mejor y más confiable siempre son los datos internos (propios)

Repositorios de datos abiertos

- [European Union Open Data Portal](#)
- [UNdata](#)
- [Open Data Inception](#)

Brokers de datos

- [Experian](#)
- [Acxiom](#)

APIs

- [Yahoo finance](#)
- [OpenWeather](#)

Web Scraping

- [Scrapy](#)
- [import.io](#)

Todo análisis se enriquece con datos externos

Tipos de variables

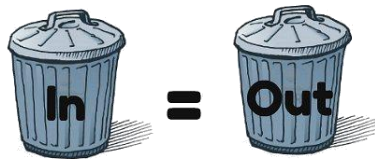
Continuas	Edad, Altura, colores RGB
Ordinales	Ratings, Niveles educativos, Muy de acuerdo/De acuerdo/En desacuerdo
Categóricas	Hombre/Mujer, Apto/No Apto, días de la semana

TIPOS DE DATOS - ESTRUCTURA

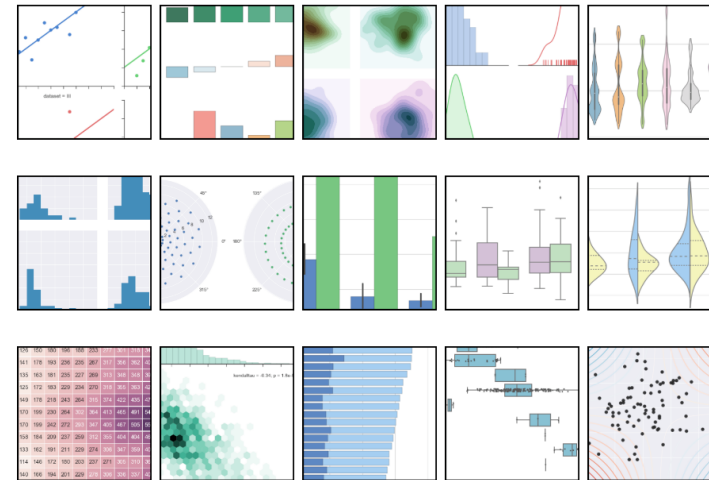
Estructurados (<10%)	Catálogo de biblioteca, bases de datos sql
Semiestructurados (<10%)	XML, JSON, CSV
No Estructurados (>=80%)	Emails, Fotos, PDF

Análisis exploratorio de datos (EDA)

- ✓ En los proyectos de Machine Learning y Data Science los algoritmos mas complejos son quienes se llevan las palmas, pero la verdad es que las “victorias” no serían posible sin un Análisis Exploratorio de Datos (EDA) llevando con calma, paciencia y buen humor.
- ✓ Esta etapa es importante para ver la calidad de los datos con los que trabajaremos, recuerda que:



- ✓ Para el EDA no existe una metodología estandarizada, quizá podamos conseguir “líneas generales”, además, el EDA irán mejorando con cada análisis que hagamos y sobre todo los adecuaremos a nuestra data.



Análisis exploratorio de datos (EDA)



Revisar fuentes

¿Dónde encuentro la data?



Lectura de datos

¿Cómo cargos la data?



Cuadratura

¿La data es la correcto?

```
In [34]: data.describe()
Out[34]:
```

	Area Code	Item Code	Element Code	latitude	longitude	Y1991	Y1992	Y19
count	21477.000000	21477.000000	21477.000000	21477.000000	17938.000000	17938.000000	17938.000000	17938.000000
mean	125.445001	2594.211028	1271.687154	25.455913	15.794445	185.262059	202.782220	202.4564
std	72.886149	148.973468	146.803079	24.803358	98.013104	1884.134336	1884.266891	1881.1741
min	1.000000	2511.000000	8142.000000	-40.900000	-172.100000	0.000000	0.000000	0.0000
25%	63.000000	2581.000000	8142.000000	6.430000	-11.780000	0.000000	0.000000	0.0000
50%	125.000000	2545.000000	8142.000000	20.580000	16.110000	1.000000	1.000000	1.0000
75%	185.000000	2182.000000	8142.000000	41.700000	46.810000	21.000000	22.000000	23.0000
max	276.000000	2581.000000	9521.000000	64.980000	179.410000	112227.000000	108135.000000	136356.0000

8 rows x 9 columns

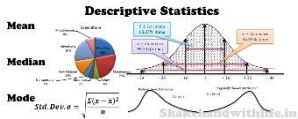
Procesado

iiiTengo huecos!!! ¿Qué hago?



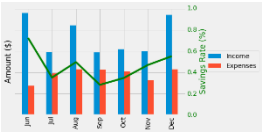
Descriptivos

¿Todos estos datos serán útiles?

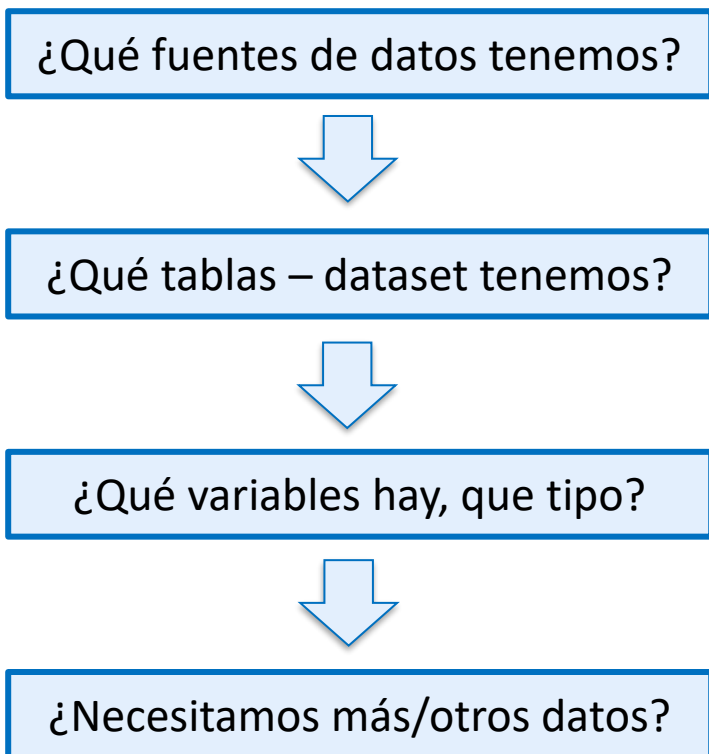


Gráficos

Una imagen vale más que mil palabras



Revisar Fuentes



Recomendación:
La data solicitada debe coincidir con la entregada, por lo cual es buena idea generar un archivo de “cuadratura”... ¿Cuadratura?

Procesamiento de datos

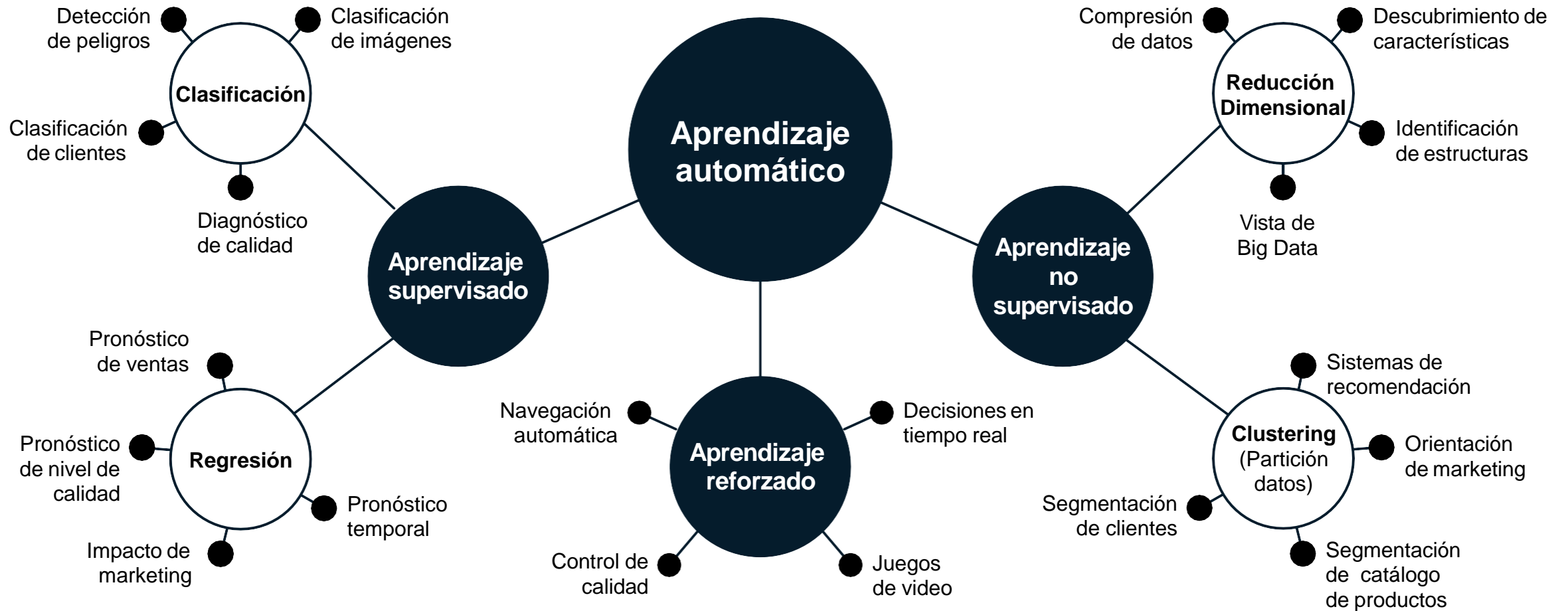
Existen muchos pasos antes de tener un set de datos útil. Entre ellos tenemos:

- Renombrar variables
- Buscar valores perdidos (missing)
- Tratar los missing
- Crear variables dummies
- Detectar outliers
- Normaliza o estandarizar variables
- Guardar nuestro data set limpio.

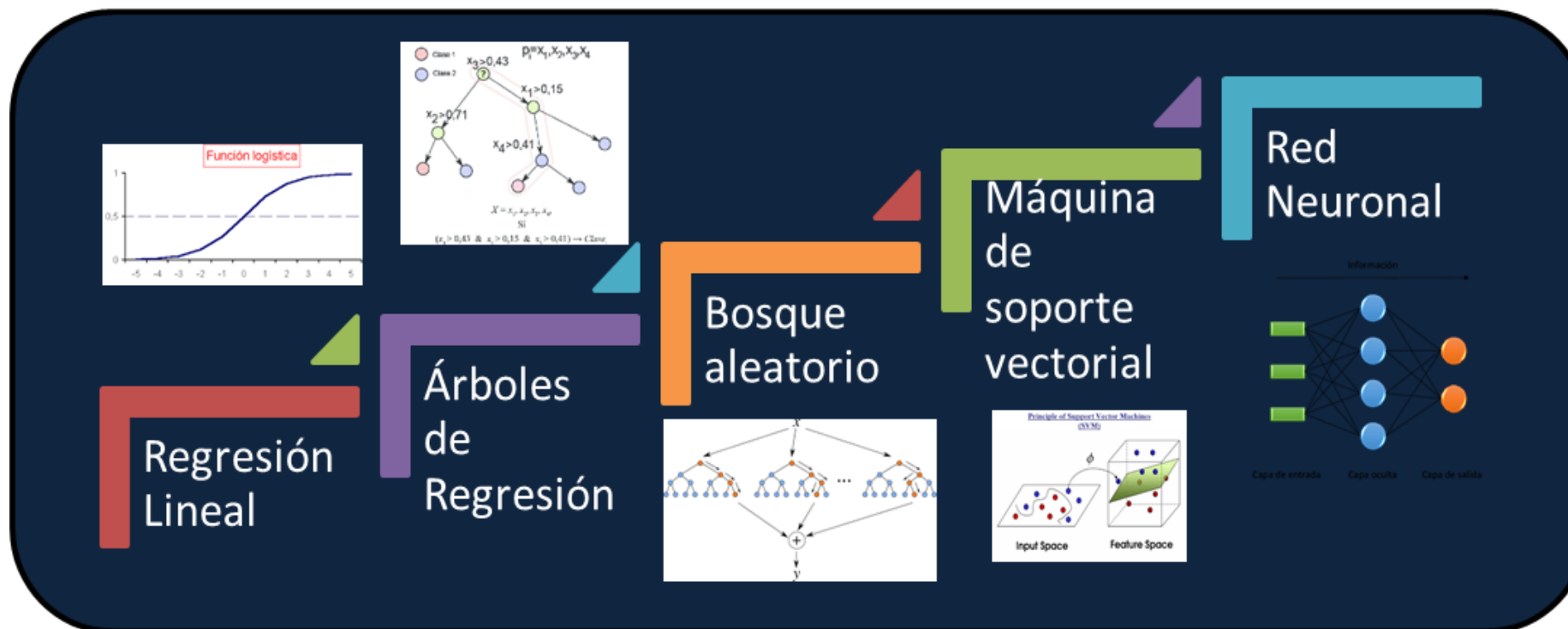


Aprendizaje automático

Aprendizaje automático



Modelos de Regresión



La exactitud y precisión de los modelos depende de siempre del set de datos con los que trabajamos, ellos definen quien es el mejor.

Ojo!!! No existe un modelo ideal, no existe un modelo único

Regresión Lineal

Modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon \quad (4)$$

Donde, p es la cantidad de predictores, X_j representa el j -ésimo predictor, β_j son los parámetros desconocidos a estimar que cuantifican la asociación entre la variable predictora y la respuesta, y ϵ es el error aleatorio.

Además, β_j se interpreta como el efecto promedio en Y de un incremento en una unidad de X_j manteniendo todos los demás predictores fijos.

La estimación de los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$ se realiza comúnmente mediante el método de Mínimos Cuadrados Ordinarios, el cual se enfoca en encontrar el valor de los coeficientes de regresión de modo tal que la suma de los cuadrados de las diferencias entre los valores observados y la línea de regresión sea mínima. Matemáticamente, esto es, minimizar la suma de cuadrados de los residuales (RSS):

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \quad (2)$$

De donde se obtiene:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (3)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Regresión Lineal


Las métricas de evaluación disponibles para los modelos de regresión son:

- Error cuadrático medio (MSE)



$$\frac{1}{n} \sum_i^n (y_i - \hat{y}_i)^2$$

- Raíz del error cuadrático medio (RMSE)

- Error absoluto promedio (Mean Absolute Error, MAE) 

$$\frac{1}{n} \sum_i^n |y_i - \hat{y}_i|$$

- Raíz del MAE (Root Mean Absolute Error, RMAE)

- Coeficiente de determinación



$$R^2 = \frac{\sum_{t=1}^T (\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^T (Y_t - \bar{Y})^2}$$
