Ciencia de Datos

Julio Correa

www.indexar.cl

Agenda

1. Contexto
2. Discusión
3. Ejemplos
    a. Seteo Google Colab
    b. Ejemplo clasificación usando KNN
4. Referencias

De qué hablamos?

Nube de palabras encontrada en Google con las siguientes palabras:

Ciencia datos nube palabras

Fuente: https://www.researchgate.net/figure/Figura-1-Nube-de-palabras-del-concepto-que-los-estudiantes-tienen-de-la-estadistica_fig1_324573704

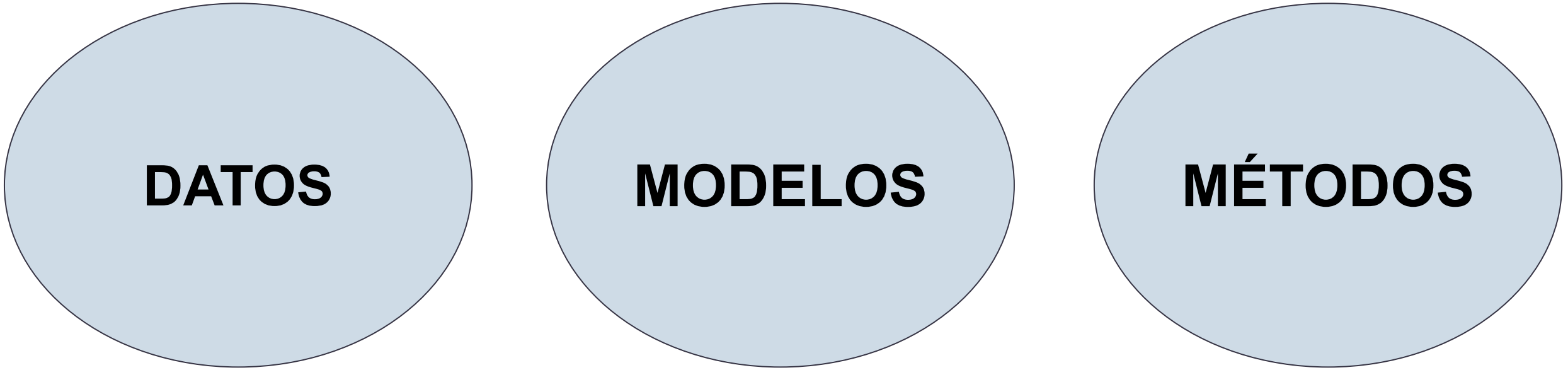
Definición de IBM [1]

Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from large volumes of **data** in various forms either structured or unstructured, which is a continuation of some of the data analysis fields such as **data** mining and predictive analytics, as well as knowledge discovery and **data** mining (KDD).

Data Science is about turning data into insights.

La ciencia de datos es un campo interdisciplinario de procesos y sistemas para extraer conocimiento y aprendizajes de grandes volúmenes de datos, ya sean estructurados o desestructurados. Esto corresponde a una continuación de algunos campos de la analítica de datos como la minería de datos o la analítica predictiva, como también del descubrimiento de conocimiento (KDD).

¿Qué hay en juego?



Elementos en tensión!

Los datos son una de las formas en las que la naturaleza y los fenómenos se expresan…

Y la naturaleza es más compleja de lo que quisiéramos….

Ejemplo

Las redes posibles que se pueden armar en un grupo de N personas es:

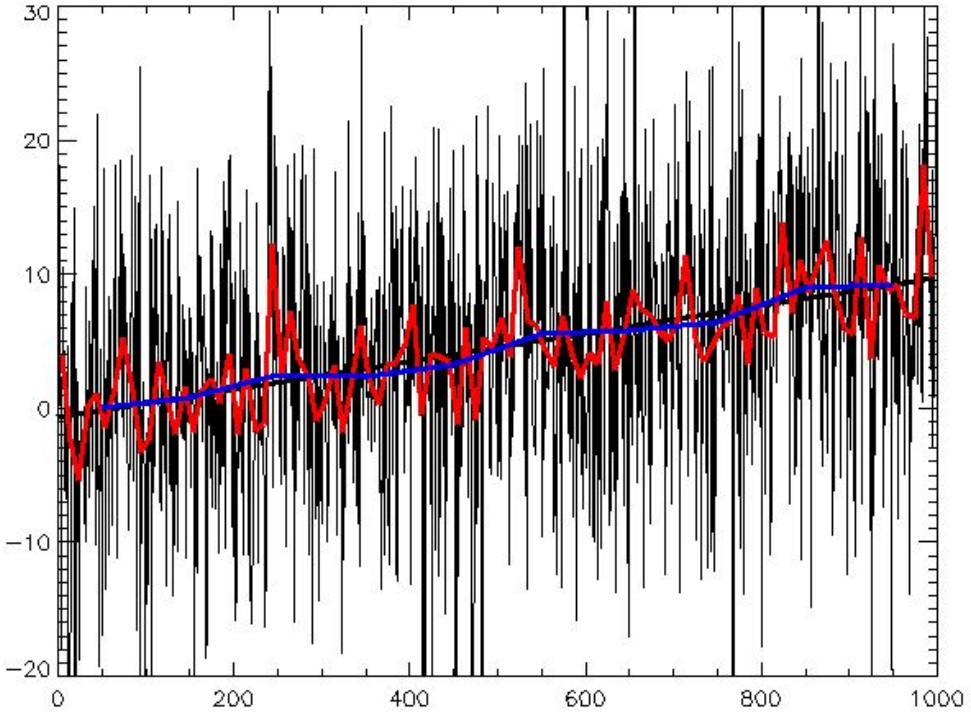
$$N = 5 \longrightarrow 2^{N(N-1)/2} = 2^{5*4/2} = 1.024$$

$$N = 6 \longrightarrow 2^{N(N-1)/2} = 2^{6*5/2} = 32.768$$

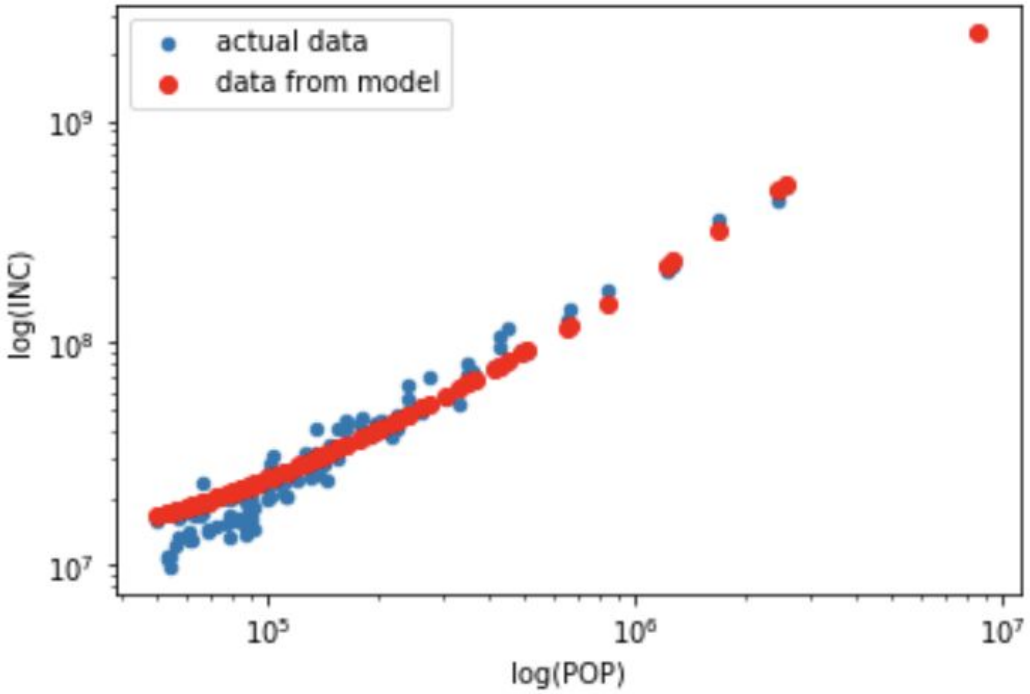
$$N = 7 \longrightarrow 2^{N(N-1)/2} = 2^{7*6/2} = 2.097.152$$

$$N = 8 \longrightarrow 2^{N(N-1)/2} = 2^{8*7/2} = 268.435.456$$

Dinámica y ‘esparcida’

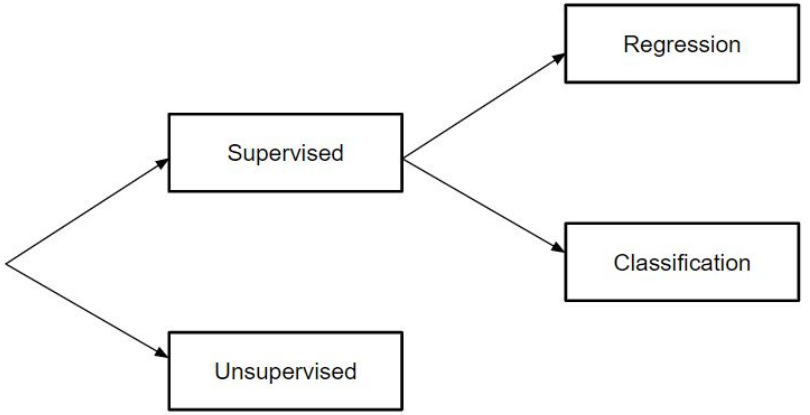


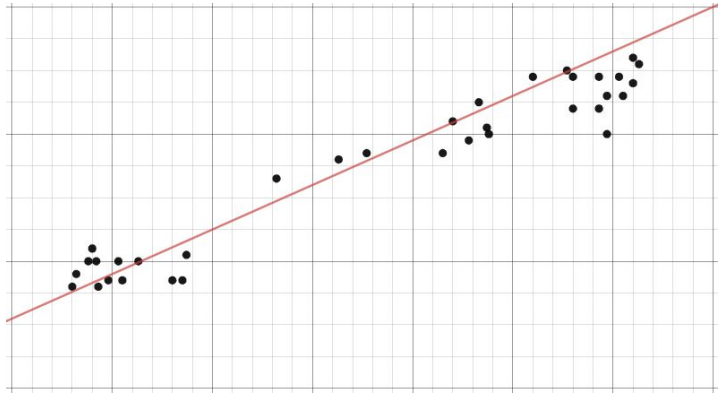
Fuente: https://commons.wikimedia.org/wiki/File:Random-data-plus-trend-r2.png



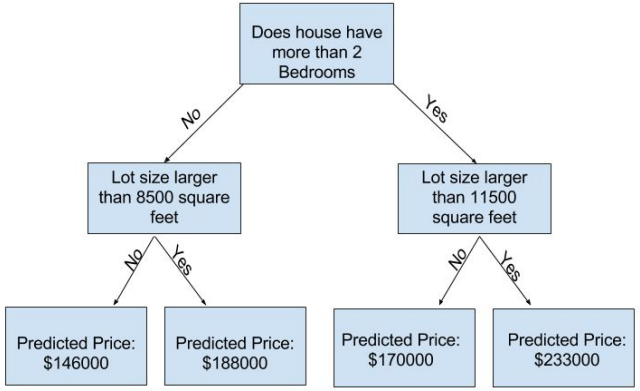
Fuente: Elaboración propia (proyecto urban scaling)

Modelos (i)

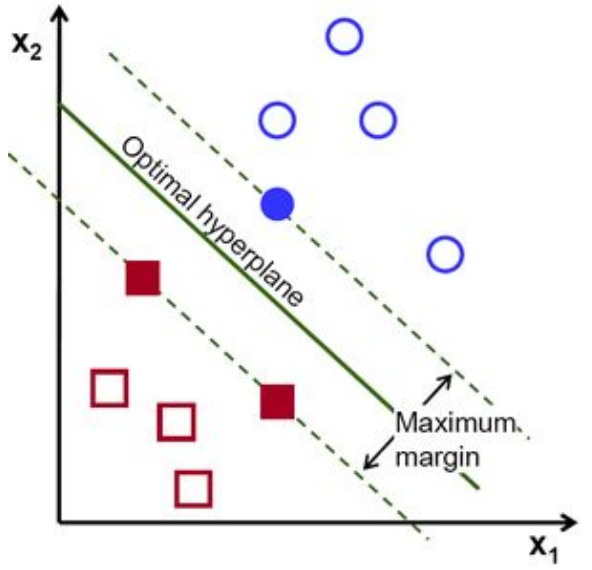




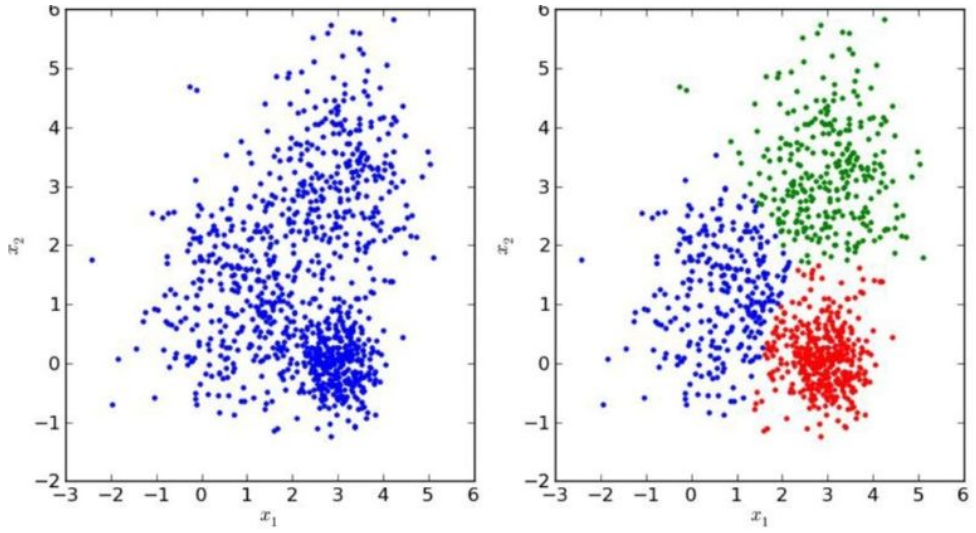
Regresión lineal
(imagen tomada de Towards Data Science)



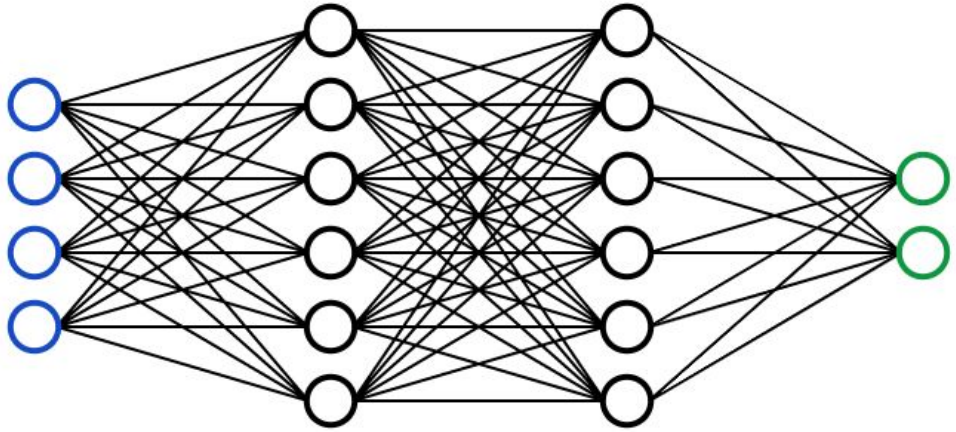
Árbol de decisión
(imagen tomada de Kaggle)



SVM
(imagen tomada de Towards Data Science)



Clustering (K-means)
(imagen tomada de Towards Data Science)

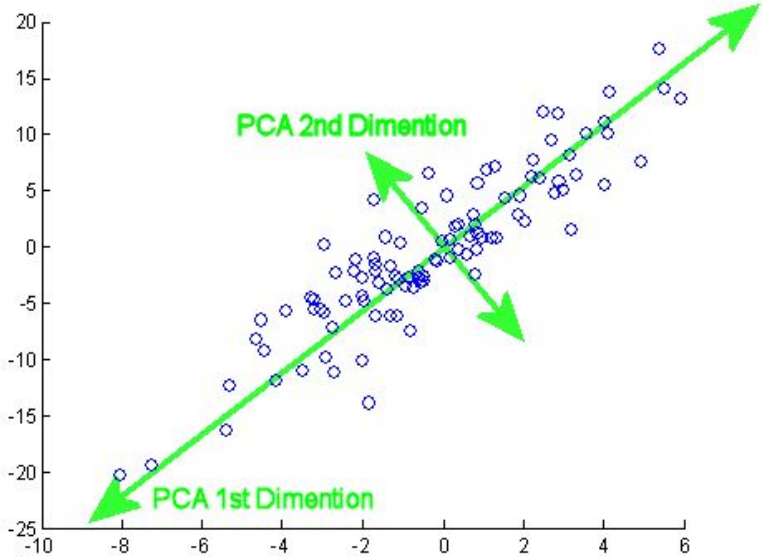


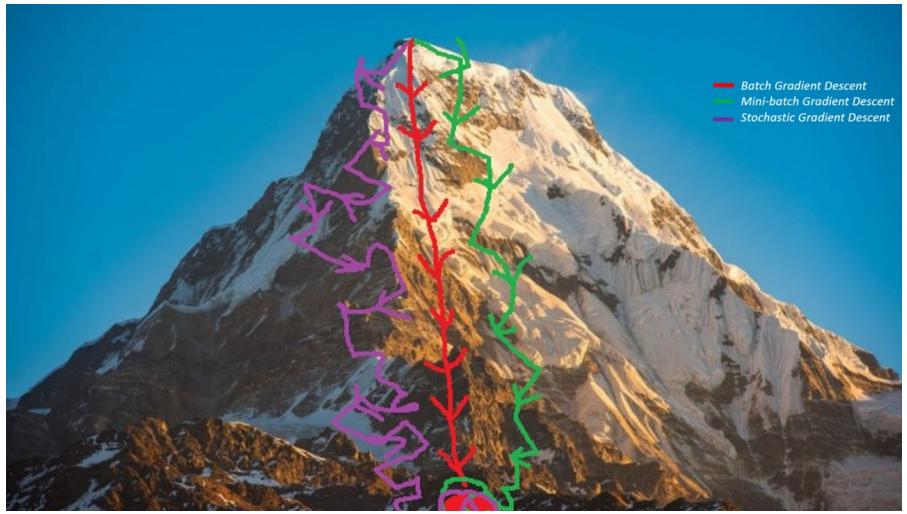
Red neuronal
(imagen tomada de Kaggle)

Métodos (i)

Álgebra lineal:

- Valores y vectores propios
- Análisis de la componente principal





Cálculo:

- Diferenciación
- Direccionalidad del gradiente

Fuente: https://towardsdatascience.com/gradient-descent-algorithm-and-its-variants-10f652806a3

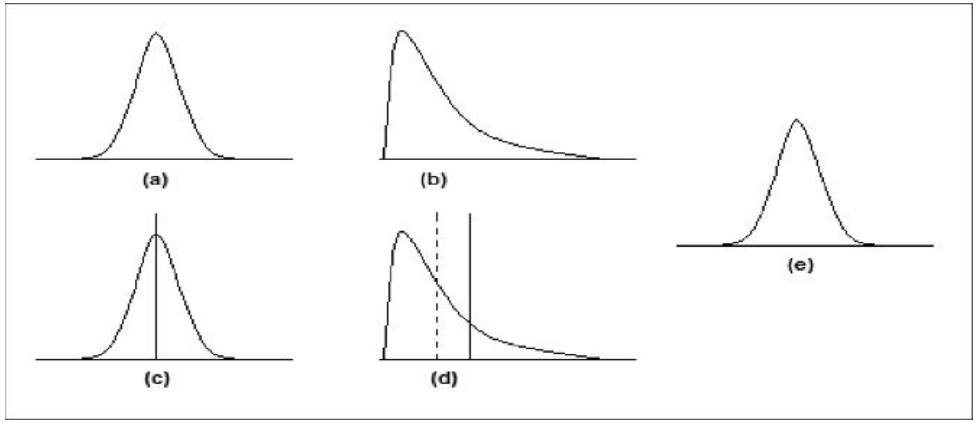
Métodos (ii)

Probabilidades.

- Teorema de Bayes
- Distribuciones de probabilidad

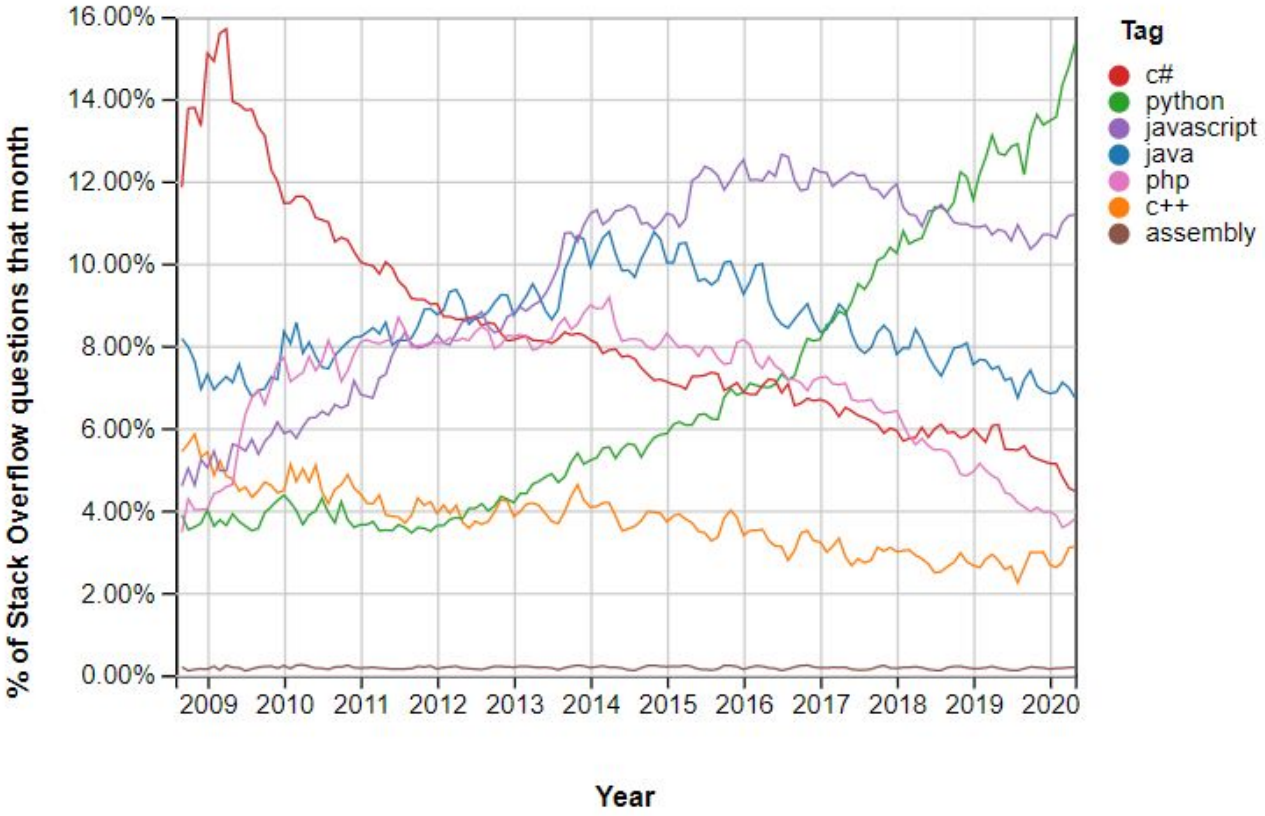
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

a. Fórmula del teorema de Bayes

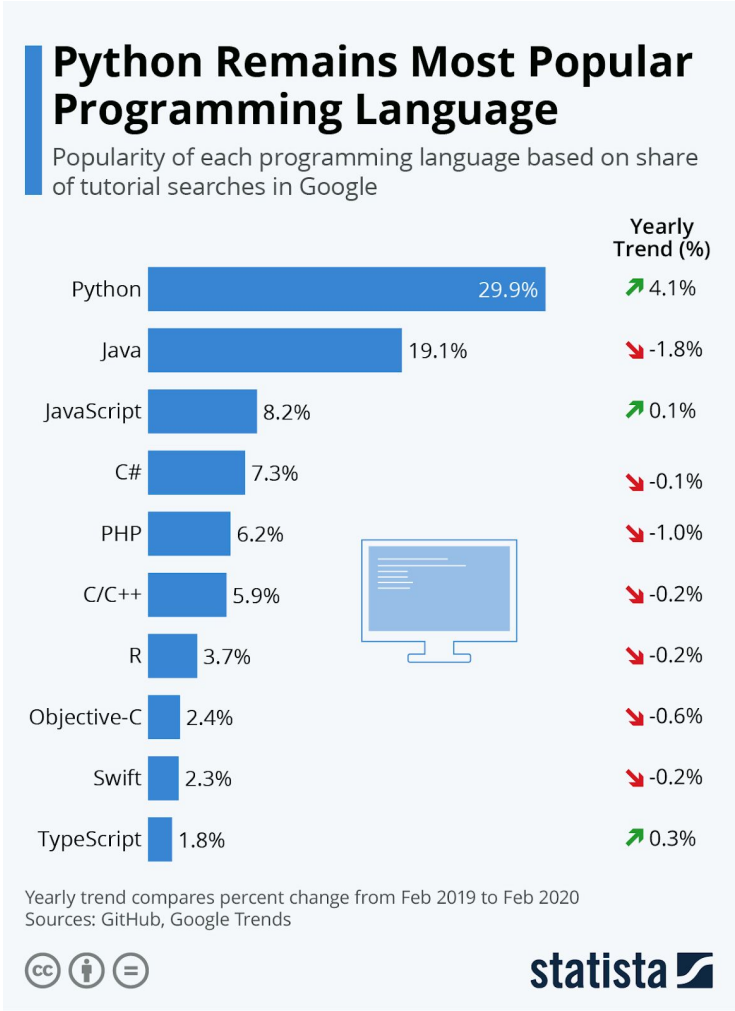


b. Distintas distribuciones de probabilidad

Métodos (iii)

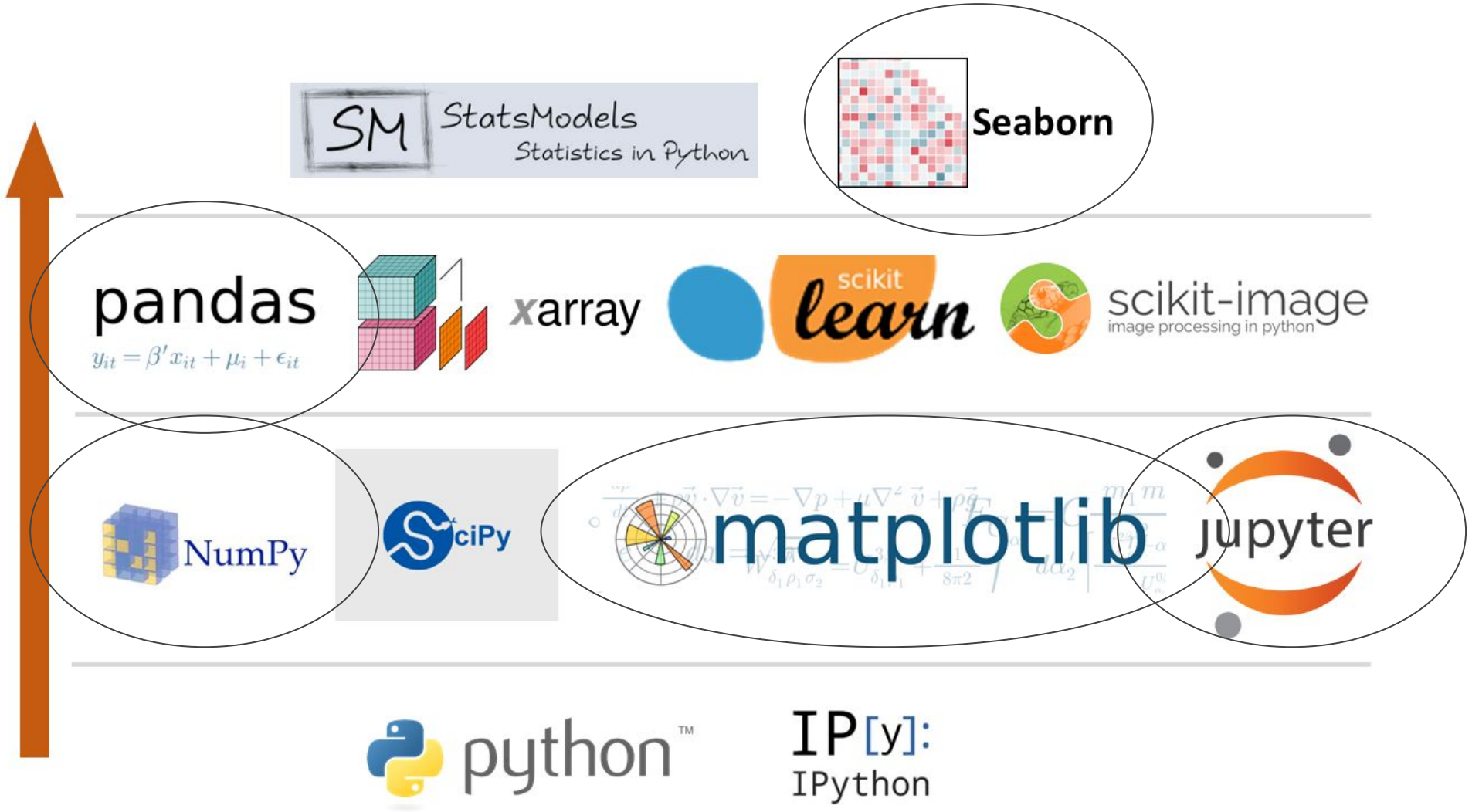


Source: https://www.stxnext.com/what-is-python-used-for/



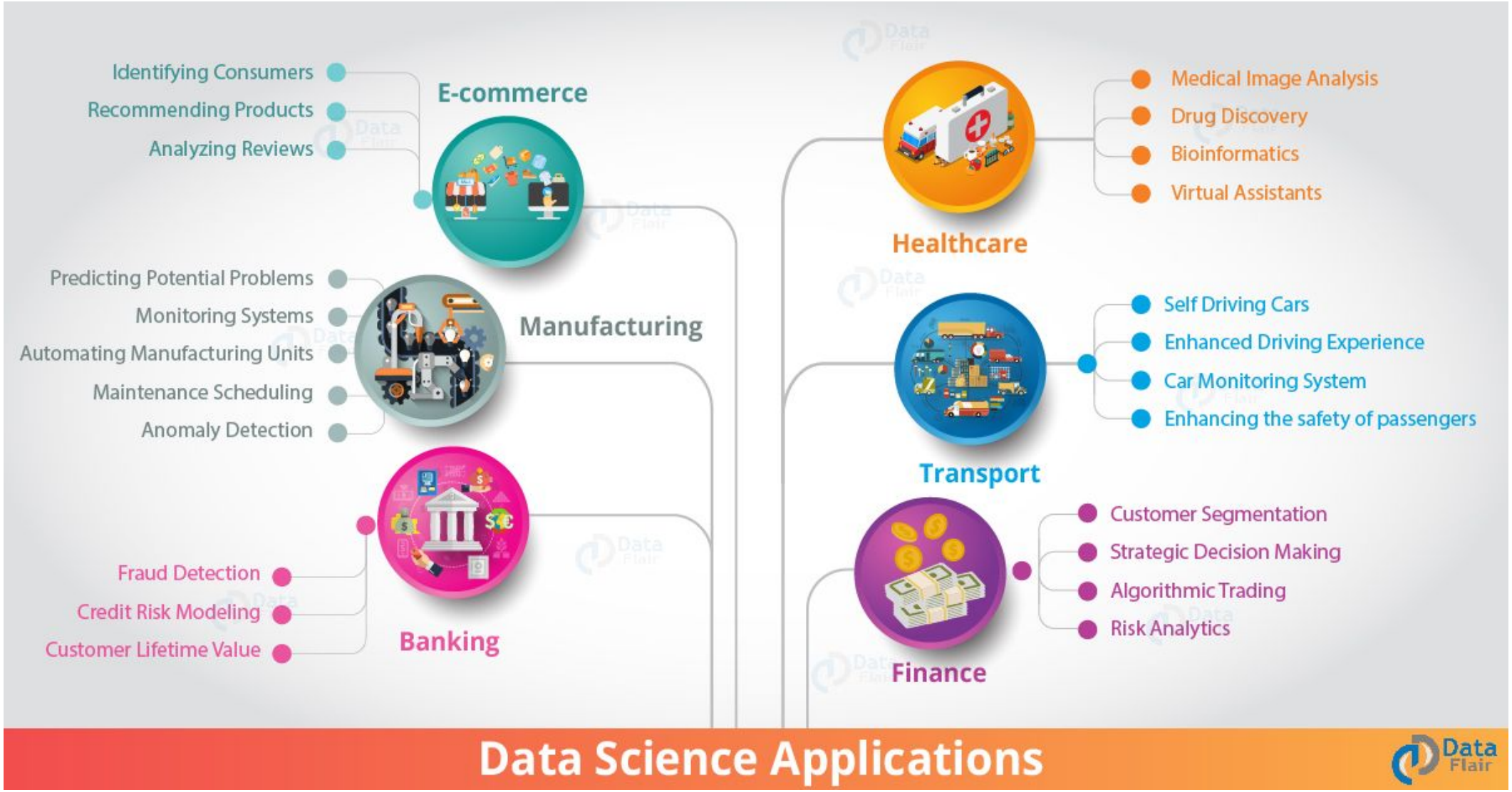
Source:
https://www.statista.com/chart/21017/most-popular-programming-languages/

Métodos (iv)



Source: http://chris35wills.github.io/courses/pydata_stack/

Aplicaciones



Ejemplos.

REFERENCIAS

[1] http://www.researchmethods.org/DataScienceDataScientists.pdf