Open in Colab
(https://colab.research.google.com/github/efviodo/idatha-data-science-course/blob/master/notebooks/02%20-%20DS%20-%20Introduccion%20a%20Data%20Science%20-%20Python.ipynb)

Introducción a la Ciencia de Datos

Objetivos

- Conocer conceptos básicos de Data Sience
- Entender qué problemas se resuelven en Data Sience
- Introducir una metodología de trabajo para Data Science

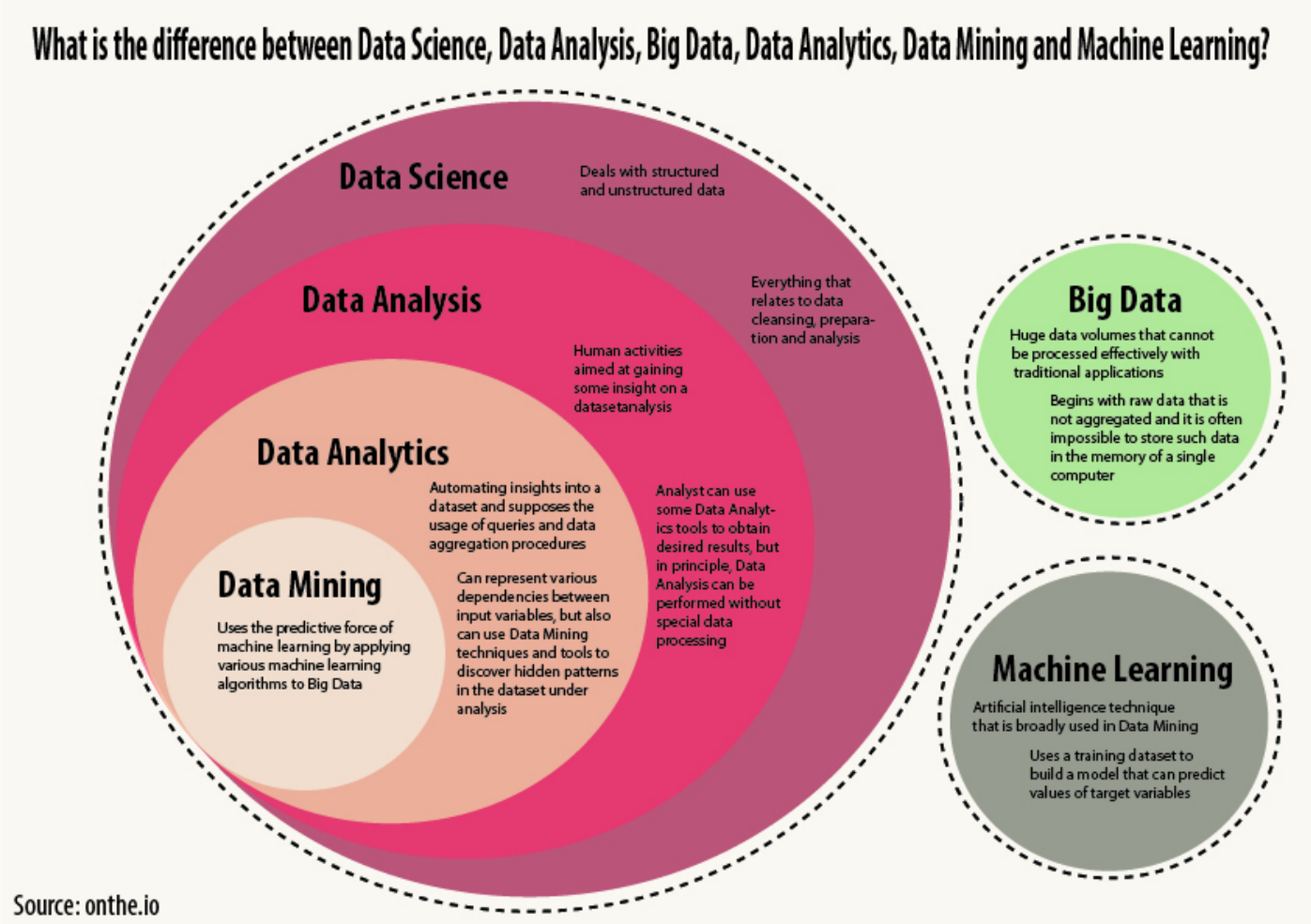
Índice

Data Sience

Inicio ▲

La ciencia de datos es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o alcanzar un mejor entendimiento de datos en sus diferentes formas (estructurados y no estructurados). Para ello se basa en algunos campos del análisis de datos como la estadística, la minería de datos, el aprendizaje automático y la analítica predictiva.

En otras palabras, puede imaginarse como un área de conocimiento que se basa en otras áreas ya desarrolladas, como la Minería de Datos, Análisis Estadístico y utiliza técnicas de Aprendizaje Automático y BigData para descubrir nueva información.



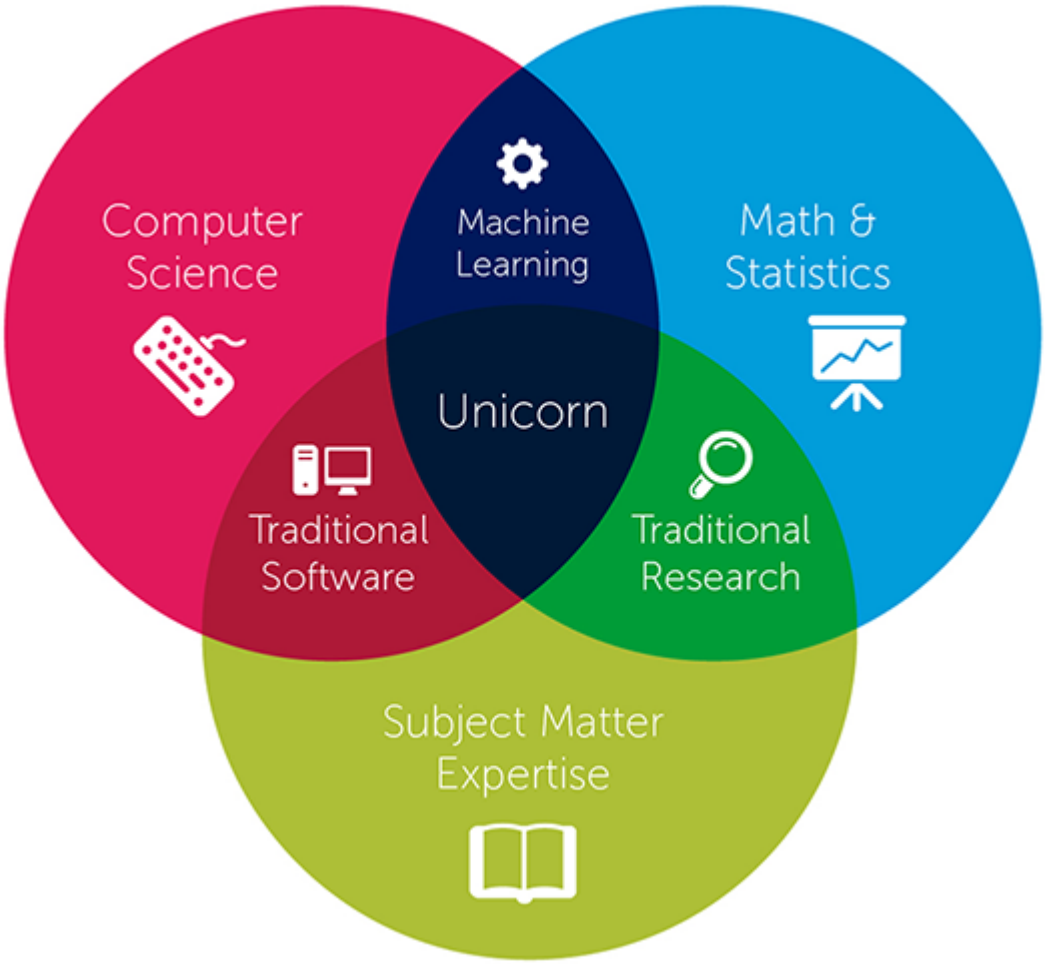
Data Scientist

Inicio ▲

El Data Scientist o Científico de Datos, es una persona capaz de analizar e interpretar datos complejos, así como utilizar tecnicas de estadística y aprendizaje automático para comprender mejor estos datos y extraer conslusiones que permitan resolver un problema de la realidad.

Combina una sólida formación teórica y práctica en las materias fundamentales asociadas al análisis avanzado de datos: pensamiento analítico, comprensión de problemas de la realidad, estadística, programación, tratamiento de bases de datos, trabajo con algoritmos y comunicación efectiva, preparado para encarar problemas de la realidad y convertirlos en soluciones utilizando datos.

Es muy común colcoar a un data scientist en la intersección de las siguientes áreas de conocimento: (i) Ciencias de la Computación, (ii) Matemáticas y Estadística y (iii) Conocimiento de un dominio específico



Conceptos Básicos

Inicio ▲

Modelo

Representación matemática de un proceso del mundo real; un modelo predictivo pronostica el resultado del futuro basado en comportamientos del pasado.

Ejemplo: Modelo de probabilidad de fuga de clientes.

Algoritmo

Conjunto ordenado de operaciones sistemáticas que permite hacer un cálculo y hallar la solución a un problema.

Ejemplo: Algoritmo de ordenamiento QuickSort (https://es.wikipedia.org/wiki/Quicksort).

Entrenamiento

El proceso de crear un modelo a partir de los datos de entrenamiento. Los datos alimentan un algoritmo de entrenamiento que aprende la representación del problema y produce un modelo. Comúnmente llamado “aprendizaje”.

Regresión

Método de predicción cuyo resultado es un número real (un valor que representa una cantidad en una recta). Por ejemplo: predecir la temperatura de un motor o la ganancia de una empresa.

Clasificación

Método de predicción que asigna una categoría predefinida a cada dato de entrada, por ejemplo, categoría de cliente según sus compras.

Target

En estadística se le llama variable dependiente. Es la salida del modelo o la variable que se quiere predecir.

Conjunto de Entrenamiento

Comunmente llamado *Training dataset*, se utiliza para encontrar relaciones potencialmente predictivas que serán utilizadas para crear un modelo. También puede encontrarse como *Corpus de entrenamiento*.

Conjunto de Verificación

Comunmente llamado *Test dataset*, es un conjunto de datos diferente al de entrenamiento, pero con la misma estructura. Se utiliza para evaluar la performance de los modelos predictivos. También puede encontrarse como *Corpus de pruebas*.

Feature

También conocida como variable independiente o variable predictora, una feature es una cantidad observable,

Aplicaciónes de Data Science

Inicio ▲

Ejemplos más comúnes:

- Mantenimiento predictivo
- Análisis de sentimiento
- Detección de intereses
- Segmentación de clientes
- Riesgo de fuga
- Detección de spam
- Predicción de demanda
- Detección de fraude

Ejemplos de aplicación por industria y vertical

Industry	Sales & marketing	Finance & risk	Customer & channel	Operations & workforce
Retail	• Demand forecasting • Loyalty programs • Cross-sell & upsell • Customer acquisition	• Fraud detection • Pricing strategy	• Personalization • Lifetime customer value • Product segmentation	• Store location demographics • Supply chain management • Inventory management
Financial services	• Customer churn • Loyalty programs • Cross-sell & upsell • Customer acquisition	• Fraud detection • Risk & compliance • Loan defaults	• Personalization • Lifetime customer value	• Call center optimization • Pay for performance
Healthcare	• Marketing mix optimization • Patient acquisition	• Fraud detection • Bill collection	• Population health • Patient demographics	• Operational efficiency • Pay for performance
Manufacturing	• Demand forecasting • Marketing mix optimization	• Pricing strategy • Performance risk management	• Supply chain optimization • Personalization	• Remote monitoring • Predictive maintenance • Asset management

Metodologías

Inicio ▲

Motivación

Necesito contar con una metodología o marco de trabajo que me permita sistematizar ciertas etapas como recolectar datos, limpiarlos, generar un modelo predictivo y determinar acciones. Un proceso estándar que traduzca un problema de la vida real en tareas abordables por un equipo de cientificos de datos.

Alternativas

Existen varias propuestas de modelos o marcos de trabajo, que ayudan a un cientifico de datos a abordar un problema de forma ordenada. En este taller vamos a trabajar con un modelo bastante conocido, que se llama CRISP-DM y es bien conocido por ser uno de los modelos más adoptados para problemas en Minería de Datos.

Cabe destcar, que existen otros modelos e incluso algunas empresas tecnológicas muy fuertes como Facebook y Uber, implementaron sus propias herramientas o plataformas para data science, basadas en sus propios modelos de trabajo.

- Microsoft TDSP (https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/), es una metodología o flujo de rabajo alternativa.
- FBLearner Flow (https://code.fb.com/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/), es una plataforma de Machine Learning de Facebook.

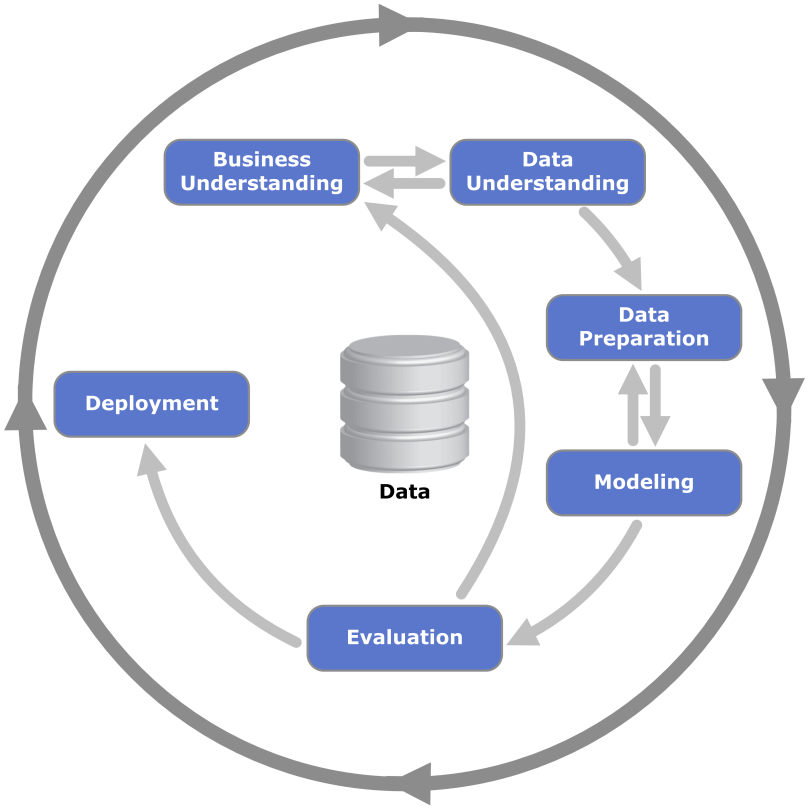
Queda a cargo del lector profundizar en cualquiera de las alternativas propuestas o buscar nuevas.

Modelo CRISP-DM

Inicio ▲

CRISP-DM es la sigla para *CRoss Industry Standard Process for Data Mining* (algo así como *Proceso Estándar Multi Industria para Minería de Datos*). Es un modelo de proceso, propuesto inicialmente para proyectos de minería de datos y que puede ser adaptado para proyectos en ciencia de datos. El proceso es independiente del sector de la industira del cual proviene el problema que queremos resolver o de las tecnologías utilizadas.

Fue presentado por primera vez en el año 2000, a través del trabajo *CRISP-DM: Towards a standard process model for data mining (https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf)* [1]. En caso de estar interesado en el trabajo podes encontrar la cita en la sección de Biblografía.



En este taller veremos solamente las primeras tres fases del modelo, quedando fuera de alcance las otras tres.

Fases del modelo CRISP-DM

El modelo CRISP-DM se divide en 6 fases o etapas, que inicialmente en un proyecto de ciencia de datos se ejecutan en un orden determinado. Luego el proceso puede retro-alimentarse, volviendo a la etapa inicial o cualquier etapa intermedia, formando un circulo de retroalimentación.

I. Comprensión del negocio

Comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición del problema de minería de datos, y un plan preliminar diseñado para alcanzar los objetivos.

Etapas:

- Determinar objetivos empresariales.
- Analizar la situación actual.
- Determinar objetivos de minería de datos.
- Planificación con duración, recursos y riesgos.

II. Comprensión de los datos

Recolección inicial de datos y procesos con actividades con el objetivo de familiarizarse con los mismos, identificar problemas en la calidad de los datos, descubrir primeros insights en los datos, o detectar subconjuntos de datos para formular hipótesis sobre datos ocultos. Hay un vinculo muy cercano entre las etapas de *Comprensión del negocio* y *Comprensión de los datos*

Etapas:

- Recopilar datos disponibles
- Explorar y describir los datos con tablas y gráficos.
- Verificar calidad de los datos.

III. Preparación de datos

Actividades para construir el conjunto de datos de entrenamiento. Estas tareas son ejecutadas en múltiples oportunidades y sin orden. Las tareas incluyen selección y transformación de tablas, registros y atributos, y limpieza de datos para las herramientas de modelado.

Etapas:

- Selección de subconjunto de datos.
- Limpieza de datos.
- Creación de nuevos atributos (ingeniería de atributos).
- Fusión y agregado de conjuntos y registros.
- Verificación de formato de datos para el modelado.
- División en conjuntos de datos de prueba y entrenamiento.

IV. Modelado

Se seleccionan y aplican varias técnicas de modelado y se calibran los parámetros para mejorar los resultados. Hay varias técnicas que tienen requerimientos específicos sobre la forma de los datos, por lo que puede ser necesario volver a la fase de preparación de datos.

V. Evaluación

Evaluación del modelo (o modelos) construidos, que parecen tener gran calidad desde una perspectiva del análisis de datos.

VI. Despliegue

Esta fase depende de los requerimientos, pudiendo ser simple como la generación de un reporte o compleja como la implementación de un proceso de explotación de información que atraviese a toda la organización.

Herramientas

Inicio ▲

Jupyter Notebooks

- Proyecto open-source basado en IPython.
- **Entorno interactivo** para la ejecución de código:
  - Versionado de notebooks
  - Celda como unidad de trabajo con un único formato (tipo de celda)
  - Edición (cortar, copiar, pegar), merge, split y desplazamiento de celdas
  - Visualización de celdas de diferentes tipos
  - Inserción de celdas arriba y abajo (shortcut: 'A' para insertar arriba y 'B' abajo)
  - Manejo de visualización de resultado (esconder, permitir scroll y borrar)
  - Administración del Kernel (interrupción, reinicio, cambio)
- Puede mostrar:
  - `Código`
  - Gráficas generadas a partir de código
  - Texto enriqucido
  - Expresiones matemáticas: $e^x = \sum_{i=0}^{\infty} \frac{1}{i!}x^i$
  - Dibujos y *rich media* (HTML, LaTeX, PNG, SVG, etc.)
- **Lenguajes soportados**: Python, R, Scala y muchos más (ver lista de kernels soportados (https://github.com/jupyter/jupyter/wiki/Jupyter-kernels)).

Lenguaje R

- Lenguaje de programación con fuerte foco en el análisis estadístico.
- Ofrece una amplia variedad de herramientas y librerias para trabajar con datos
  - **dplyr**: Libreria **R** para manipular data frames de forma simple: `select` , `filter` , etc. https://dplyr.tidyverse.org/ (https://dplyr.tidyverse.org/)
  - Herramientas para data profiling: DataExplorer (https://cran.r-project.org/web/packages/DataExplorer/vignettes/dataexplorer-intro.html), HMisc (https://cran.r-project.org/web/packages/Hmisc/index.html), entre otros.
  - Herramientas para data visualization: ggplot2 (https://ggplot2.tidyverse.org/)

Lenguaje Python

- Es otro de los lenguajes de programación elegido por la comunidad de data scientists.
- Interpretado, de fácil aprendizaje y muy ligero.
- Tiene una oferta de herramientas muy buena para data scientists:
  - Pandas (https://pandas.pydata.org/): Librería para manipulación de data frames.
  - Scikit-Learn (https://scikit-learn.org/stable/): Librería para mineria y análisis de datos, implementa algorítmos para los problemas clásicos de Machine Learning: Classification, Regression, Clustering.
  - Herramientas para data visualization: matplotlib (https://matplotlib.org/), seaborn (https://seaborn.pydata.org/), bokeh (https://bokeh.pydata.org/en/latest/).
  - Librerías para profiling de datos: Pandas Profiling (https://pypi.org/project/pandas-profiling/)

- Libreraías matemáticas: NumPy (http://www.numpy.org/), SciPy (https://www.scipy.org/),

Bibliografía

Inicio ▲

1. Wirth, R., & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining (pp. 29-39). Citeseer.