Análisis de Datos

Clase 1 - Introducción y herramientas de SW

Presentación

- Nahuel Pelli
  - Ingeniero en Telecomunicaciones, Instituto Balseiro
  - Maestría en Ciencia de datos, UdeSA
  - Contacto:
    - Email: nahuelpelli91@gmail.com

Programa de la materia

- Clase 1: Introducción al análisis de datos. Softwares necesarios. GIT, Python. Introducción a los frameworks clásicos Numpy, scipy, pandas y matplotlib.
- Clase 2: Análisis básico de media, desvío estándar, oblicuidad (skewness, curtosis, cuantiles, IQR
- Clase 3: Datos, características (features) e ingeniería de features. Tipos de variable de entrada y salida: continuas y categóricas (nominal y ordinal)
- Clase 4:Codificación one-hot y dummy. Otros típos de codificación (binaria, hashing, etc.). Valores faltantes. Normalización de datos. Transformación de datos.

Programa de la materia

- Clase 5: Variables Aleatorias. Teoría de la información. Entropía. Entropía cruzada. Divergencia KL. Ejemplos. Tests estadísticos univariados para Machine Learning. Introducción a test estad´ısticos, definición de p-value, z score. Tests de normalidad. Tests de correlación. Tests de independencia. Análisis de varianza (ANOVA). Ejemplos.
- Clase 6: Aplicación de test estadísticos univariados para ML. Ejemplos prácticos. Reducción de dimensionalidad. Normalización
- Clase 7: Taller práctico.
- Clase 8: Evaluación final.

Métodos de evaluación

- Si bien la materia cuenta con un pequeño examen multiple choice. Además vamos a contar con un trabajo práctico final integrador.
- El trabajo es grupal. Los grupos serán de 2 alumnos.
- Vamos a facilitarle el enunciado a partir de la clase 2 y la idea es que vayan avanzando de a poco con lo visto en clase.
- En la clase 8 va a haber una presentación de 15 min por grupo para exponer lo encontrado. Deben considerar una ventana de al menos 5 minutos para preguntas.

1. ¿Qué es el análisis de datos?

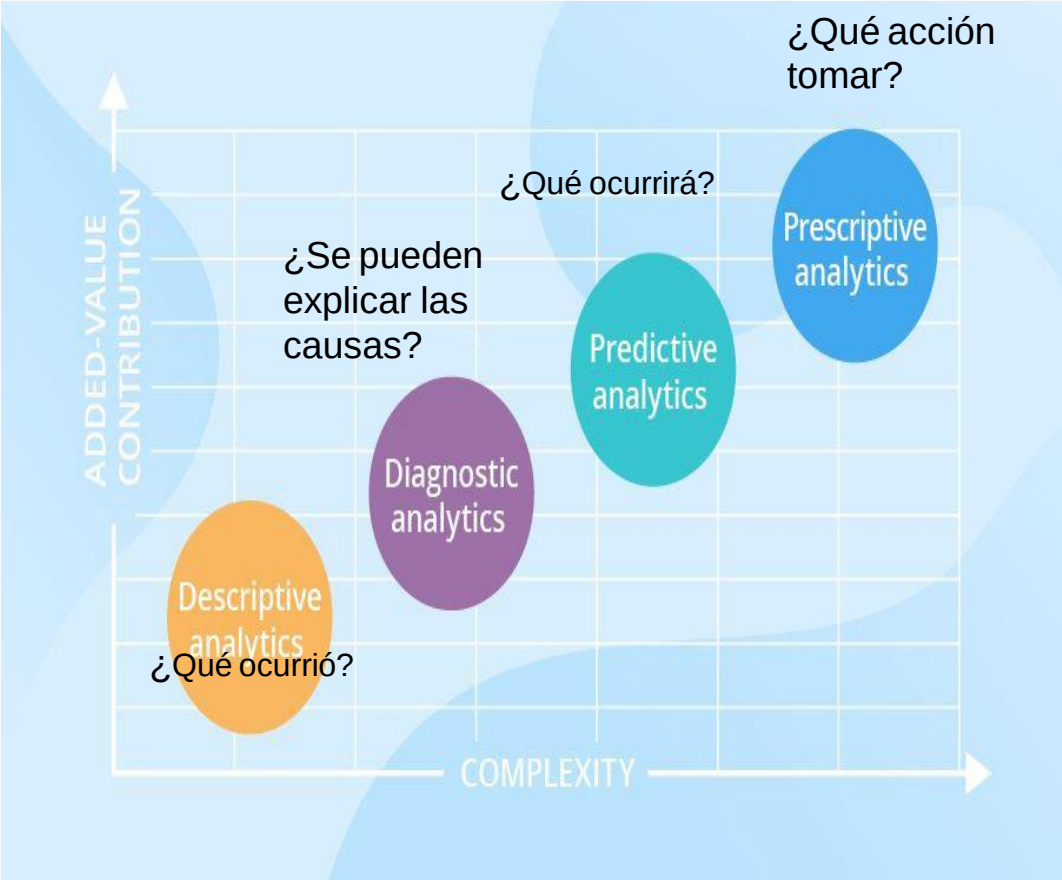
-Es el proceso de explorar y analizar conjuntos de datos con el objetivo de hacer predicciones y contribuir a la toma de decisiones apoyada en datos.

- Analizar datos
- Tomar decisiones

2. Aplicaciones

- Análisis de fraude
- Salud
- Administración de inventario
- Logística
- Marketing
- Planificación urbana

3. Tipos de análisis de datos



- Descriptivo
- Diagnóstico
- Predictivo
- Prescriptivo

3. Tipos de análisis de datos (ejemplos)

- Análisis descriptivo:
  - Estudiar la cantidad de unidades vendidas de un producto y el beneficio obtenido.
- Análisis de diagnóstico:
  - Hallar la correlación entre la contaminación del agua y una enfermedad.
- Análisis predictivo:
  - Estimar la demanda a futuro de un bien o servicio.
- Análisis prescriptivo:
  - Establecer los parámetros óptimos de una cadena de producción y distribución para suplir una demanda.

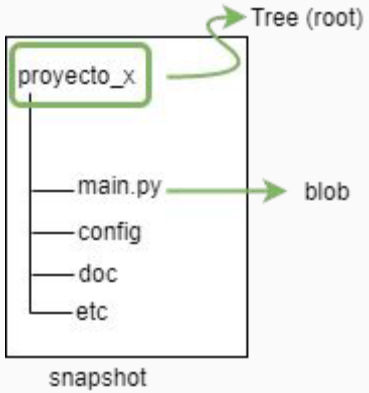
4. Pasos en el proceso de análisis de datos

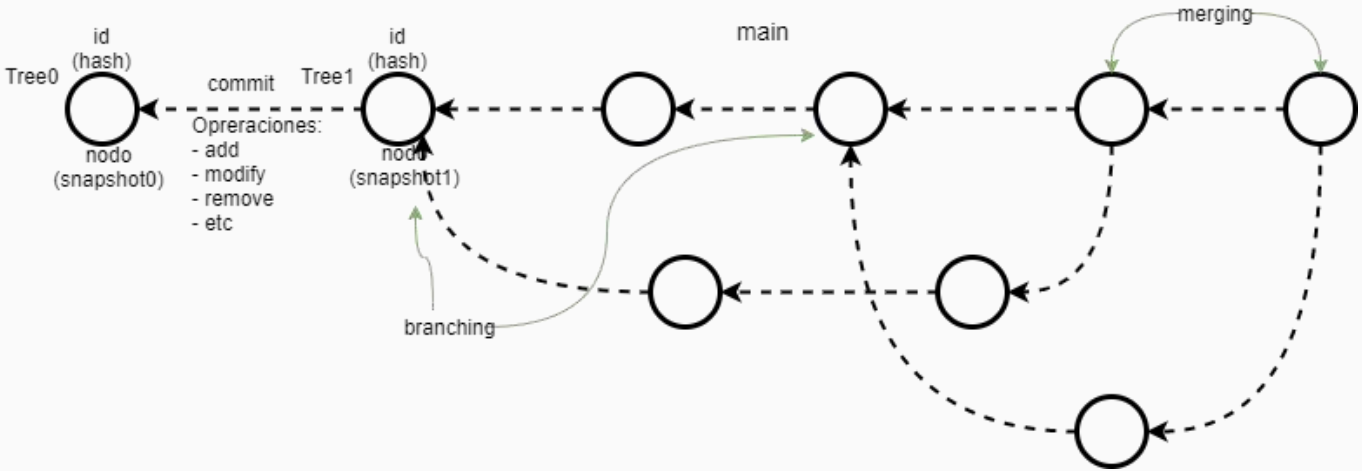
1. Obtención de datos
2. Preparación de datos
3. Exploración de datos
4. Desarrollo de modelos
5. Interpretación de resultados

5. Python para análisis de datos

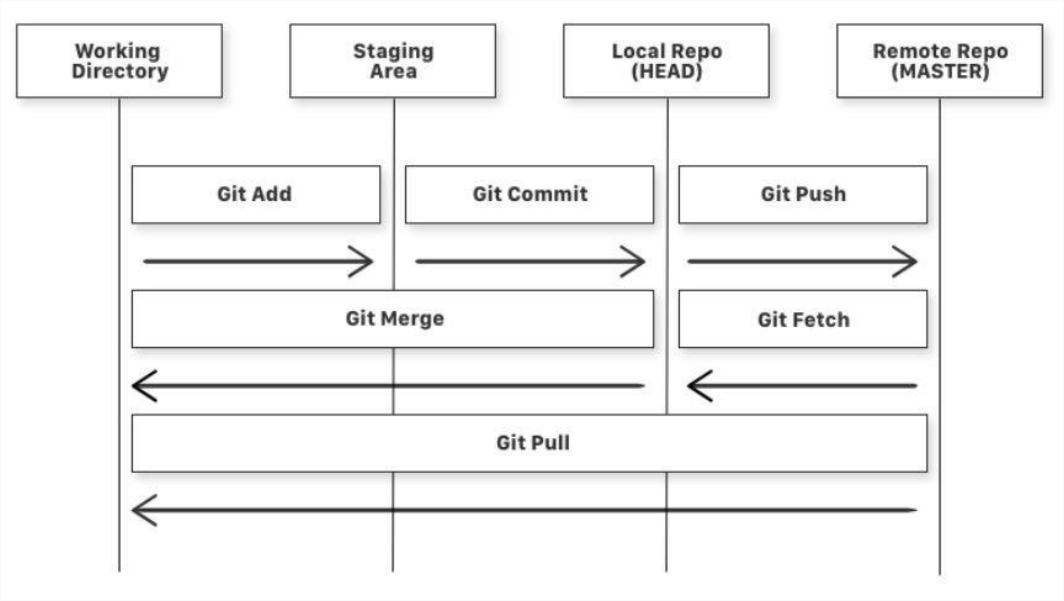
6. Git

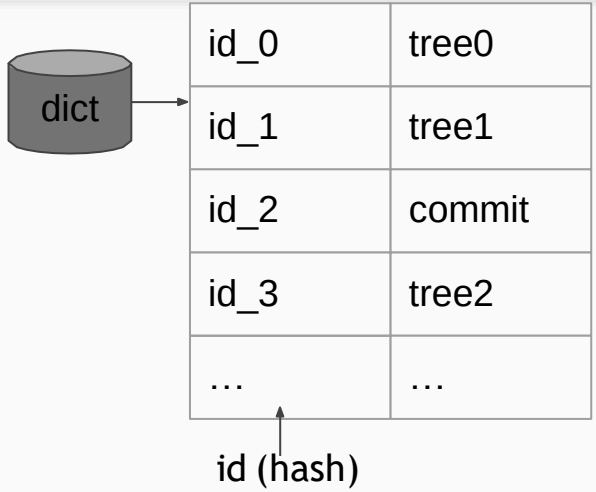
- Git es un sistema de control de versiones distribuido.
- Modelo de datos: grafo dirigido acíclico (DAG)





7. Git





TAGS

Fuente

8. Git

Algunos comandos útiles

- git clone <url> #hace una copia local de un repo
- git add <filename> #agrega un archivo al *staging area*
  git add . # agrega todos los archivos al *staging area*
- git commit -m “msj” # hace el commit
- git push # envía los commit al repo remoto
- git pull # trae los cambios del repo remoto
- git checkout -b <my branch name> # crea una nueva rama y te posiciona allí

Cheat Sheet + Instalación

9. Python

- Lenguaje de alto nivel muy utilizado en el ámbito de ML.
- Lenguaje no tipado de scripting e identando.
- Existen múltiples frameworks que nos facilitan el uso de Python. Entre ellos Conda (o anaconda) es uno de los mas utilizados:
  - Ambientes virtuales (conda,virtualenv,etc.).
    - Posibilidad de tener múltiples configuraciones de paquetes.
  - Sistema de paquetes conda.
    - Amplia comunidad de usuarios.
    - Librerías para Fortran, Python, C/C++, R, etc.

10. Hands-on

- Plan:
  - Repaso de GIT como sistema de control de versiones.
    - Modelo de datos de GIT.
  - Repaso de Python con algunos conceptos de Ingeniería de SW y algoritmos y estructuras de datos.
  - Introducción/repaso de Pandas.
  - Trabajo Práctico Nro. 1: Temperaturas en Europa en los últimos 500 años.