Unidad 2: Preparación de la información
2.1: Análisis Exploratorio

Docente: Pablo Torres Tramón[1,2]

[1]Facultad de Ingeniería
t.pabloandrestorres@uandresbello.edu
[2]Ponencias originales elaboradas por:
Mailiu Díaz Peña y Alejandro Figueroa

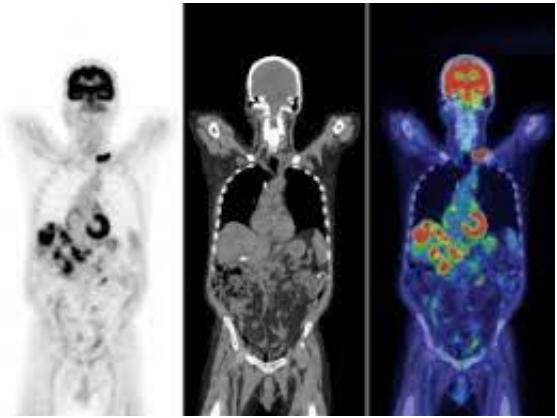
Minería de Datos
Otoño 2023

Índice

Table of Contents

Introducción

¿Cómo construimos un modelo para la detección del cáncer si solo tenemos imágenes y sus respectivas clases?



Condiciones típicas del mundo real:

- Definir las características

Condiciones típicas del mundo real:

- Definir las características
  - Dominio

Condiciones típicas del mundo real:

- Definir las características
  - Dominio
  - Trabajo relacionado

Introducción

Condiciones típicas del mundo real:

- Definir las características
  - Dominio
  - Trabajo relacionado
  - Ingeniería de características (Deep learning)

Introducción

Condiciones típicas del mundo real:

- Definir las características
  - Dominio
  - Trabajo relacionado
  - Ingeniería de características (Deep learning)
- Datos incompletos

Introducción

Condiciones típicas del mundo real:

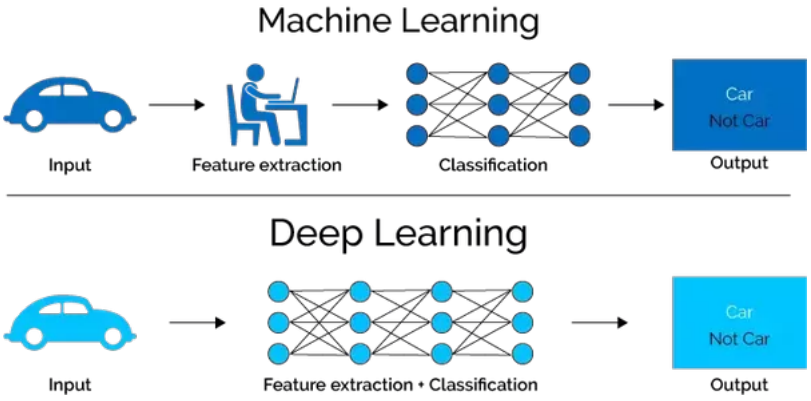
- Definir las características
  - Dominio
  - Trabajo relacionado
  - Ingeniería de características (Deep learning)
- Datos incompletos
- Ruido

Introducción

Condiciones típicas del mundo real:

- Definir las características
  - Dominio
  - Trabajo relacionado
  - Ingeniería de características (Deep learning)
- Datos incompletos
- Ruido
- Inconsistentes

Introducción



Introducción [Ars21]

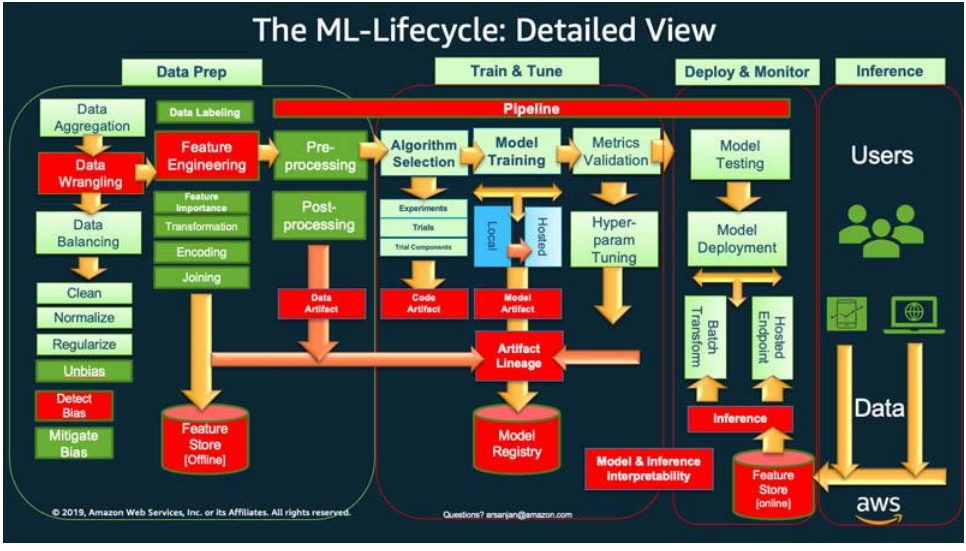
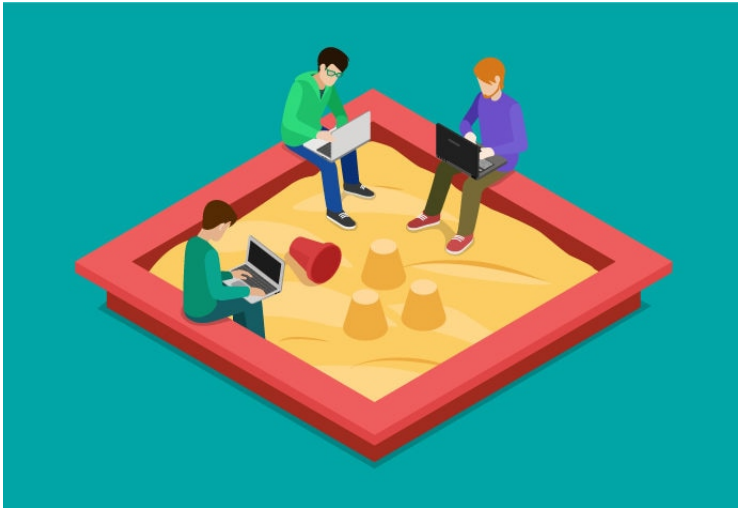


Table of Contents

Necesitamos un lugar para jugar



Nuestro Sandbox ideal

Herramientas:

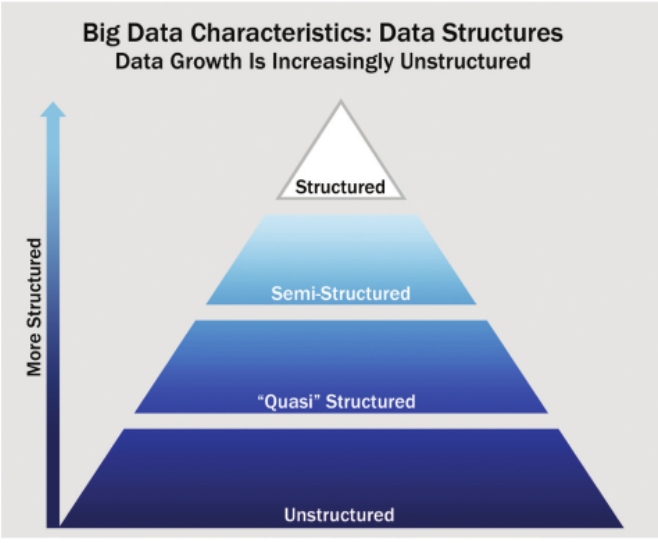
1. Herramientas para la manipulación de datos
2. Visualización
3. Poder de computo
4. Almacenamiento de datos pre-procesados
5. Acceso a la data de origen:

Nuestro Sandbox ideal

Herramientas:

1. Herramientas para la manipulación de datos
2. Visualización
3. Poder de computo
4. Almacenamiento de datos pre-procesados
5. Acceso a la data de origen:
6. **¿Cuál es nuestra data de origen?**

Tipos de datos [Ser17]



Formatos

Tipo	Descripción	Ejemplos
Estructurado	Representación específica y consistente	`sql, csv, xls`
Semi-Estructurado	Representación autodescrita	`xml`
Quasi-Estructurado	Representación autodescrita pero inconsistente	`html`
No-estructurado	Representación debe ser reconocida desde los datos	`txt, jpg, png`

Extract, Transform and Load (ETLs)

ETLs: Son pequeños códigos que permiten extraer la data desde el origen hasta el Sandbox.

- **Extract**: Extraer la data desde su origen

Extract, Transform and Load (ETLs)

ETLs: Son pequeños códigos que permiten extraer la data desde el origen hasta el Sandbox.

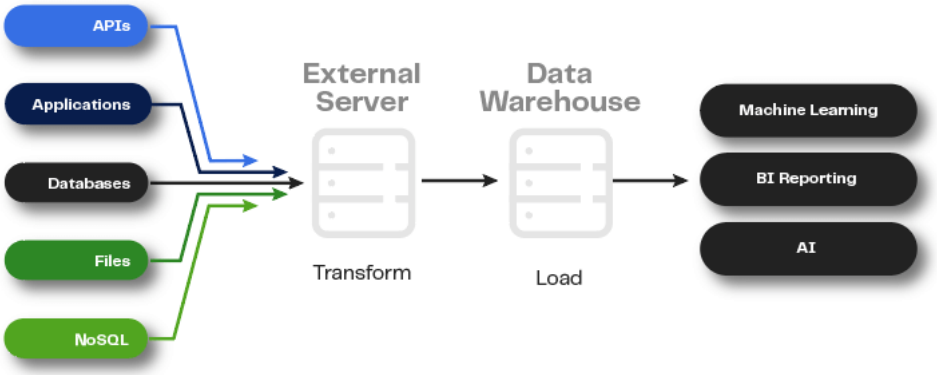
- **Extract**: Extraer la data desde su origen
- **Transform**: Transformar y/o interpretar la data de acuerdo a la tarea

Extract, Transform and Load (ETLs)

ETLs: Son pequeños códigos que permiten extraer la data desde el origen hasta el Sandbox.

- **Extract**: Extraer la data desde su origen
- **Transform**: Transformar y/o interpretar la data de acuerdo a la tarea
- **Load**: Almacenar la data transformada en un repositorio de fácil acceso

Diagrama ETL [Mat21]



Extracción de características

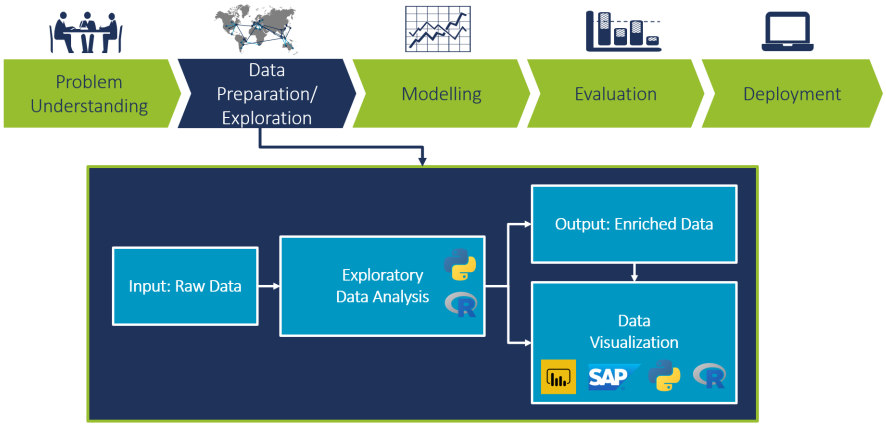
Ejercicio 4.1

Escriba un ETL en `Python` para extraer las características para un dataset que solo contiene imágenes [1].

[1] https://archive.ics.uci.edu/ml/datasets/Shoulder+Implant+X-Ray+Manufacturer+Classification

Table of Contents

The Big Picture [Lab19]



Problemas con las características

Los características extraídas suelen ser sucias:

- **Incompletas**: carecen de valores para ciertas columnas, carecen de interés o contienen solo datos agregados.

Problemas con las características

Los características extraídas suelen ser sucias:

- **Incompletas**: carecen de valores para ciertas columnas, carecen de interés o contienen solo datos agregados.
- **Ruido**: contienen errores/valores atípicos. Por ejemplo, manejar valores negativos para un atributo que maneja salarios.

Problemas con las características

Los características extraídas suelen ser sucias:

- **Incompletas**: carecen de valores para ciertas columnas, carecen de interés o contienen solo datos agregados.
- **Ruido**: contienen errores/valores atípicos. Por ejemplo, manejar valores negativos para un atributo que maneja salarios.
- **Inconsistentes**: contienen discrepancias en códigos o nombres. Por ejemplo, edad de un empleado = 30 y fecha de nacimiento = 03/07/1998.

Sin datos de calidad, no hay buenas predicciones

Objetivos del AED

- Determinar si hay algún problema con el conjunto de datos.

Objetivos del AED

- Determinar si hay algún problema con el conjunto de datos.
- Determinar si la pregunta de investigación se puede responder con los datos que tiene.

Objetivos del AED

- Determinar si hay algún problema con el conjunto de datos.
- Determinar si la pregunta de investigación se puede responder con los datos que tiene.
- Desarrollar un bosquejo de la respuesta a su pregunta.

Etapas de un AED

1. Realizar un examen gráfico y un análisis descriptivo de la naturaleza de las variables individuales.

Etapas de un AED

1. Realizar un examen gráfico y un análisis descriptivo de la naturaleza de las variables individuales.
2. Realizar un examen gráfico y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre las variables.

Etapas de un AED

1. Realizar un examen gráfico y un análisis descriptivo de la naturaleza de las variables individuales.
2. Realizar un examen gráfico y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre las variables.
3. Evaluar algunos supuestos básicos subyacentes a muchas técnicas estadísticas, por ejemplo, normalidad, linealidad y homocedasticidad (igualdad de varianza).

Etapas de un AED

1. Realizar un examen gráfico y un análisis descriptivo de la naturaleza de las variables individuales.
2. Realizar un examen gráfico y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre las variables.
3. Evaluar algunos supuestos básicos subyacentes a muchas técnicas estadísticas, por ejemplo, normalidad, linealidad y homocedasticidad (igualdad de varianza).
4. Identificar los posibles valores atípicos (*outliers*) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.

Etapas de un AED

1. Realizar un examen gráfico y un análisis descriptivo de la naturaleza de las variables individuales.
2. Realizar un examen gráfico y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre las variables.
3. Evaluar algunos supuestos básicos subyacentes a muchas técnicas estadísticas, por ejemplo, normalidad, linealidad y homocedasticidad (igualdad de varianza).
4. Identificar los posibles valores atípicos (*outliers*) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.
5. Evaluar, el impacto potencial que pueden tener los datos ausentes (*missing*) sobre la representatividad de los datos analizados.

Table of Contents

Conclusiones

- La intraestructura moderna para minería de datos obliga a plantearnos cuáles son las características adecuadas para nuestro modelo.
- Tipos de variables
- ¿Cuáles son las etapas del análisis exploratorio de datos?
- ¿Cómo se calcula la tabla de frecuencia y representación de distribuciones de frecuencia?
- ¿Cuáles son las medidas estadísticas para describir una variable?

Table of Contents

Referencias

[Ser17] EMC Education Services. *Data Science and Big Data analytics: Discovering, Analyzing, Visualizing and Presenting Data*. EMC eduction Services, 2017.

[Lab19] Camelot IT Lab. *EXPLORATORY DATA ANALYSIS. AN IMPORTANT STEP IN DATA SCIENCE*. 2019. URL: https://blog.camelot-group.com/2019/03/exploratory-data-analysis-an-important-step-in-data-science/ (visitado 01-09-2022).

[Ars21] Ali Arsanjani. *Arquitectura del ciclo de vida de un modelo de Machine Learning en AWS: una demo completa*. 2021. URL: https://aws.amazon.com/es/blogs/aws-spanish/arquitectura-del-ciclo-de-vida-de-un-modelo-de-machine-learning-en-aws-una-demo-completa/ (visitado 24-08-2022).

[Mat21] Matillion. *What is Data Extraction? Everything You Need to Know*. 2021. URL: https://www.matillion.com/resources/blog/what-is-data-extraction-everything-you-need-to-know (visitado