

Minería de Datos

# Unidad 2: Preparación de la información 2.1: Análisis Exploratorio

**Docente:** Pablo Torres Tramón<sup>12</sup>

<sup>1</sup>Facultad de Ingeniería t.pabloandrestorres@uandresbello.edu <sup>2</sup>Ponencias originales elaboradas por: Mailiu Díaz Peña y Alejandro Figueroa

> Minería de Datos Otoño 2023

p.torres

Preparación de

Análisis Exploratorio de Datos

Conclusiones

Referencias

# Índice

### Minería de Datos

p.torres

Introducción

2 Preparación de datos

3 Análisis Exploratorio de Datos

4 Conclusiones

6 Referencias

ducción aración de

isis oratorio de

lusiones

encias

rencias

### Table of Contents



Minería de Datos

p.torres

Introducción

Introducción



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



p.torres

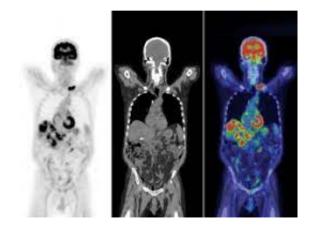
### Introducción

Preparación de datos

Análisis Exploratorio de

onclusiones

Referencias





Minería de Datos p.torres

Introducción

Preparación de datos

Exploratorio de Datos

onclusiones

eferencias

Referencias

• Definir las características



Minería de Datos p.torres

### Condiciones típicas del mundo real:

- Definir las características
  - Dominio

Introducción

Preparación de datos

Análisis Exploratorio de

Conclusiones

eferencias



Minería de Datos p.torres

p.comes

### Condiciones típicas del mundo real:

- Definir las características
  - Dominio
  - Trabajo relacionado

Introducción

Preparación de datos

Análisis Exploratorio de Datos

onclusiones

eferencias



Minería de Datos

p.torres

### Condiciones típicas del mundo real:

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

Introducción

Preparación de datos

Análisis Exploratorio de

onclusiones

eferencias



Minería de Datos

p.torres

### Condiciones típicas del mundo real:

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

Introducción

Preparación de datos

Análisis Exploratorio de

Conclusiones

Referencias



Minería de Datos

p.torres

### Condiciones típicas del mundo real:

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

Introducción

Preparación de datos

Análisis Exploratorio de Datos

Conclusiones

Referencias



Minería de Datos

p.torres

### Condiciones típicas del mundo real:

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

Introducción

Preparación de datos

Exploratorio de

Conclusiones

Referencias



Minería de Datos

p.torres

### Introducción

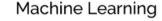
Preparación de datos

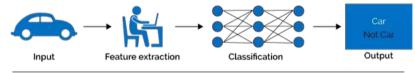
Análisis Exploratorio de

Conclusiones

Referencias

Referencias





# Deep Learning



# Introducción [Ars21]





p.torres

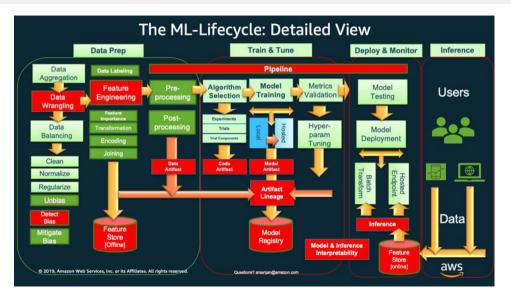
Introducción

Preparación de datos

Análisis Exploratorio de Datos

Conclusiones

Referencias



### Table of Contents



Minería de Datos

p.torres

Preparación de datos

2 Preparación de datos

# Necesitamos un lugar para jugar



Minería de Datos

p.torres

Introducción

Preparación de datos

Exploratorio o Datos

Conclusiones

Referencias



# Nuestro Sandbox ideal



Minería de Datos p.torres

itroducción

Preparación de datos

atos

onclusiones

favancias

eferencias

Herramientas:

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

# Nuestro Sandbox ideal



Minería de Datos

p.torres

### Herramientas:

- 1 Herramientas para la manipulación de datos
- Visualización
- Operation of the second of
- Almacenamiento de datos pre-procesados
- 6 Acceso a la data de origen:
- 6 ¿Cuál es nuestra data de origen?

Preparación de datos

# Tipos de datos [Ser17]



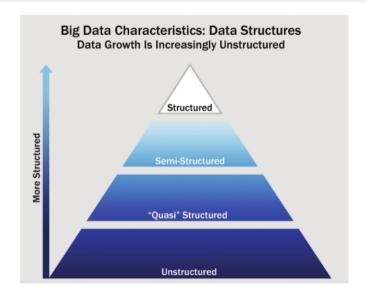
Minería de Datos p.torres

Preparación de datos

Análisis Exploratorio de

Conclusiones

Referencias



# **Formatos**



<u></u>
Universidad Andrés Bello

Minería de		
Datos		
p.torres		

### Preparación de datos

Тіро	Descripción	Ejemplos
Estructurado	Representación es- pecífica y consistente	sql, csv, xls
Semi-Estructurado	Representación auto- descrita	xml
Quasi-Estructurado	Representación auto- descrita pero inconsis- tente	html
No-estructurado	Representación debe ser reconocida desde los datos	txt, jpg, png

# Extract, Transform and Load (ETLs)



Minería de Datos p.torres

Introducción

Preparación de datos

Análisis Exploratorio de

onclusiones

erencias

eferencias

**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)



Minería de Datos p.torres

Introducción

Preparación de datos

Análisis Exploratorio de Datos

Conclusiones

eferencias

Referencias

**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)



Minería de Datos p.torres

Introducción

Preparación de datos

Exploratorio de Datos

Conclusiones

Referencias

eferencias

**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]



Minería de Datos

p.torres

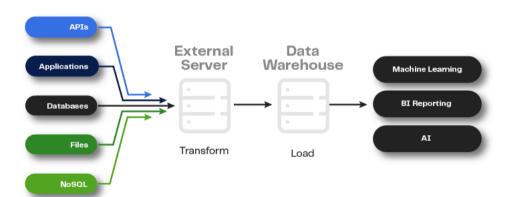
Preparación de datos

Análisis Exploratorio de

Conclusiones

. .

Poforonciae



# Extracción de características



Minería de Datos p.torres

Introducción

Preparación de datos

Análisis Exploratorio de

Conclusiones

eferencias

eferencias

### Ejercicio 4.1

Escriba un ETL en Python para extraer las características para un dataset que solo contiene imágenes <sup>1</sup>.

<sup>1</sup> https://archive.ics.uci.edu/ml/datasets/Shoulder+Implant+X-Ray+Manufacturer+Classification

### Table of Contents



Minería de Datos

p.torres

Análisis Exploratorio de Datos

- 3 Análisis Exploratorio de Datos

# The Big Picture [Lab19]





p.torres

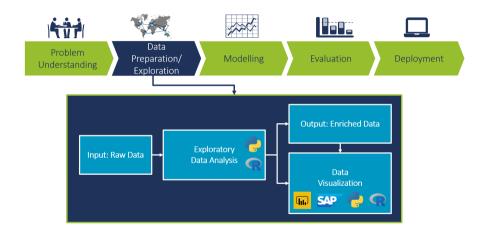
Introducción

datos Análisis

Exploratorio de Datos

Conclusiones

Referencias



### Problemas con las características



Minería de Datos p.torres

Preparación de

Análisis Exploratorio de Datos

........

eferencias

Referencias

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



Minería de Datos p.torres

Introducción

Preparación de datos

Análisis Exploratorio de Datos

Conclusiones

eferencias

Referencias

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



Minería de Datos p.torres

ntroducción

Preparación de datos

Análisis Exploratorio de Datos

Conclusiones

Referencias

eferencias

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.

# Máxima



Minería de Datos p.torres

reparación de atos

Análisis Exploratorio de Datos

anclusiones

ferencias

eferencias

Sin datos de calidad, no hay buenas predicciones

# Objetivos del AED

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



Minería de Datos p.torres

ntroducción

Preparación de datos

Análisis Exploratorio de Datos

onclusiones

eferencias

# Objetivos del AED



Minería de Datos p.torres

Introducción

Preparación de datos

Análisis Exploratorio de Datos

Canalusianas

eferencias

- 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



Minería de Datos p.torres

Introducción

Preparación de datos

Análisis Exploratorio de Datos

Conclusiones

. .

- 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.



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

Minería de Datos p.torres

p.comes

ntroducción

reparación de atos

Análisis Exploratorio de Datos

onclusiones

eferencias

- Minería de Datos p.torres

Preparación de

Análisis Exploratorio de

'amalustanaa

Datos

ferencias

- 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.



Minería de Datos p.torres

Preparación de

Análisis Exploratorio de Datos

oforoncias

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



Minería de Datos p.torres

ntroducción

Preparación de datos

Análisis Exploratorio de Datos

C . . . . I . . . I . . . . . .

Referencias

- 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.
- Sevaluar 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.



Minería de Datos p.torres

ntroducción

Preparación de datos

Análisis Exploratorio de Datos

Conclusiones

eferencias

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



### Minería de Datos

p.torres

Conclusiones

4 Conclusiones

### **Conclusiones**



Minería de Datos p.torres

Introducción

Preparación de datos

Exploratorio de Datos

Conclusiones

Referencias

- 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



Minería de Datos

p.torres

- Referencias

# Referencias



[Ser17] EMC Education Services. Data Science and Big Data analytics: Minería de Datos Discovering, Analyzing, Visualizing and Presenting Data. EMC eduction p.torres

Services, 2017. [Lab19]

[Ars21]

[Mat21]

Camelot IT Lab. EXPLORATORY DATA ANALYSIS. AN IMPORTANT STEP IN DATA SCIENCE, 2019, URL:

https://blog.camelot-group.com/2019/03/exploratory-dataanalysis-an-important-step-in-data-science/ (visitado 01-09-2022).

Ali Arsanjani. Arguitectura 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-enaws-una-demo-completa/(visitado 24-08-2022).

2021. URL: https://www.matillion.com/resources/blog/what-

ig-data-overraction-overwithing-voy-nood-to-know (visitado

Matillion. What is Data Extraction? Everything You Need to Know.

Referencias

25 / 25