



Diplomatura Universitaria en Ciencia de Datos

<https://exa.unne.edu.ar/diplomatura/>

Módulo 3. Análisis Exploratorio de Datos

Equipo Docente:

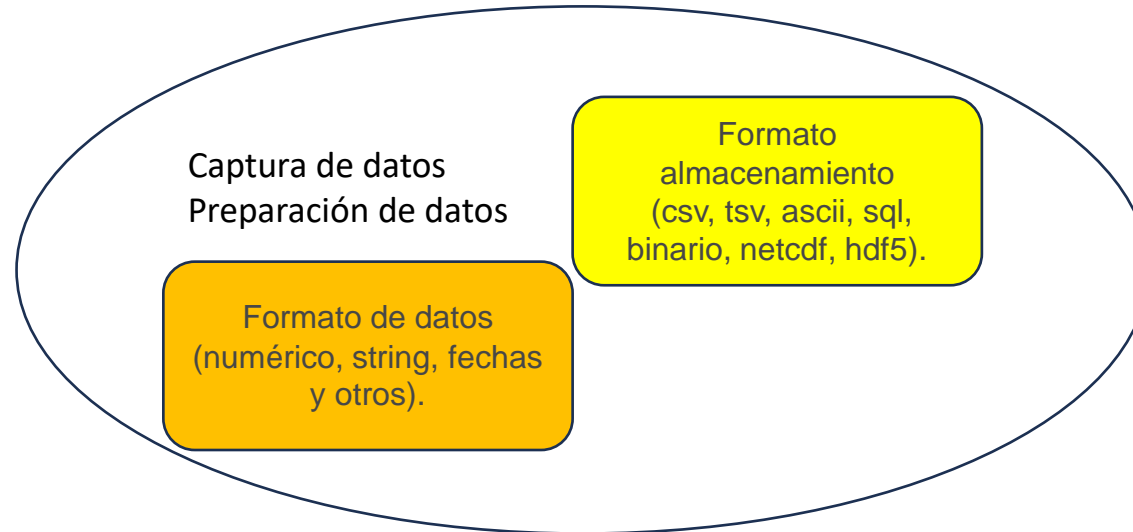
Dra. Sonia I. Mariño

Lic. Lucia del Valle Ledezma

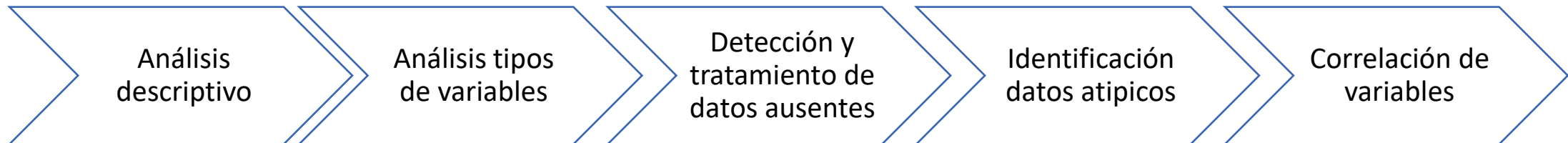
Lic. Rafael Perez

2024

Proceso EDA



ANALISIS EXPLORATORIO DE DATOS



evaluación y corrección de datos.

EDA univariado

EDA bivariado

EDA multivariado

Documentar las decisiones en el proceso

Evaluación y corrección de datos.

Involucra las fases 3 y 4, aplicable a los dataset

- Detección y tratamiento de datos ausentes
- Identificación de datos atípicos

Algunas causas:

- captura de datos, imposibilidad de obtener cierta medida u observación
- registro de datos

Ingeniería de Características

- calidad y relevancia de los datos de entrada son factores determinantes en la capacidad de un modelo para generar predicciones precisas y fiables.
- disponer de datos de calidad para procesos de entrenamiento, validación y prueba
- los datos en bruto a menudo contienen ruido, valores atípicos, valores nulos y otras anomalías que pueden influir en el rendimiento del modelo.
- Algoritmos:
 - máquinas de vector soporte y las redes neuronales, sensibles a la escala de los datos,
 - árboles de decisión, no sensibles a escala de datos.

Ingeniería de Características

Característica (feature)

- es una cantidad obtenida al codificar o transformar los datos (crudos)
- es una representación numérica de datos sin procesar.
- se derivan del tipo de datos disponibles.
- entradas que los modelos de ML / CD utilizan en entrenamiento y validación, para inferir y realizar predicciones.
- influyen en la precisión del modelo de ML / CD
 - depende de un conjunto y composición de las características.

Ingeniería de Características

Definir una característica

- Alta predicción
 - muy relacionada con la variable a predecir (etiqueta).
- Baja correlación con otras características
 - brinda información nueva, no presente en otras características
- Alta confianza
 - se calcula a partir de datos confiables y representativos del problema.
- Cálculo rápido
 - aporta factibilidad en la resolución del problema.

Ingeniería de Características

Ingeniería de características (feature engineering)

- criterios y procesos
- transformación de los datos crudos en un formato más adecuado y comprensible para el algoritmo.
- incluye el tratamiento de valores nulos, la normalización de características, la codificación de variables categóricas y la gestión de valores anómalos.
- aporta precisión en modelos ML / CD
- requiere tiempo

Ingeniería de Características

Relación entre características y modelos

- algunas características son más apropiadas para los modelos y viceversa.
- características correctas: aquellas:
 - relevantes para desarrollar la tarea en cuestión y
 - fáciles de ingerir para el modelo.

Relevancia del número de características:

- Si no hay suficientes funciones informativas, entonces el modelo no podrá realizar la tarea final.
- Si existen numerosas características o son irrelevantes, dificultades en entrenar un modelo para obtener el rendimiento esperado.

Ingeniería de Características: Tipos

Entre algunas ...

Ingeniería de Características Numéricas:

- aplica a datos numéricos. Ej. la edad, el precio o la temperatura.
- técnicas usuales: la estandarización, la normalización y la creación de variables derivadas.

Ingeniería de Características Categóricas:

- Útil para datos categóricos: el género, la ubicación o el tipo de producto.
- técnicas usuales: codificación de enteros, one-hot o dummy

Ingeniería de Características de Texto:

- aplica para datos de texto. Ej. reseñas de productos o publicaciones en redes sociales.
- técnicas usuales: tokenización, eliminación de palabras vacías (stop words) y vectorización de texto.

Ingeniería de Características de Imágenes:

- aplica para datos de imágenes, como fotografías o imágenes médicas.
- técnicas usuales: detección de bordes, extracción de características y aprendizaje de características.

Exploración de datos

- Tratar los datos faltantes
- Tratar los datos únicos
- Tratar los datos atípicos
- Crear variables
- Transformar variables categóricas - Codificación de variables
- Normalizar y Escalar

Detección y tratamiento de datos ausentes o faltantes. Detectar

Datos / valores faltantes

En proyectos de CD es esencial manejo de los datos o valores faltantes

- Errores en la entrada
- Errores en la recolección

Identificarlos y decidir su manejo evita análisis incorrectos / inexactos

Imputación, proceso de reemplazar los datos faltantes con valores estimados.

Diferenciar entre datos ausentes, valores nulos, 1 espacio, sin espacio

Distintas técnicas de imputación, como:


- Imputación con la media, mediana o moda.
- Imputación utilizando métodos como KNN o regresión.
- eliminar las filas o columnas que contengan valores faltantes en su totalidad.

Detección y tratamiento de datos ausentes o faltantes. Detectar

En librería Pandas de Python,

- los valores perdidos se representan con None y NaN (acrónimo de Not a Number).
- es un valor especial de punto flotante reconocido por todos los sistemas que utilizan la representación estándar de punto flotante IEEE.
- asigna automáticamente NaN si el valor de una celda es un string vacío "", NA o NaN.
- Si en el archivo existe otro valor como ?, se debe reemplazar

```
# Identificar el número de missing values  
df.isnull().sum()
```



	0
species	0
island	0
bill_length_mm	2
bill_depth_mm	2
flipper_length_mm	2
body_mass_g	2
sex	11

dtype: int64

Detección y tratamiento de datos ausentes. Detectar

Detectar **valores nulos, valores vacíos**

Pandas, función `isnull()`

devuelve un DataFrame donde la celda asume valores

- True (si la celda original contenía un valor faltante) o
- False (si la celda no estaba faltando un valor).

función `sum()`: determina el total de valores faltantes en la columna

`valores_faltantes = df.isnull().sum()`

Devuelve una serie con los nombres de columna como índice y el número total de valores faltantes en cada columna como valores.

Solución

- completar los valores faltantes con cierto valor
- eliminar las filas o columnas que contengan valores faltantes en su totalidad.

Identificar visualmente los valores perdidos, para encontrar patrones y vínculos existentes en estos valores en las diferentes variables.

```
import seaborn as sns
```

```
sns.heatmap(df.isnull(), cbar=False)
```

Detección y tratamiento de datos ausentes. Tratamiento

Estrategias:

- Completar los valores con la media, mediana o el valor más frecuente de la variable.
- Completar los valores que faltan por el valor que esté directamente antes o después en la fila o columna.
- Completar los datos faltantes con 0, si se trata de valores numéricos. Opción poco aconsejable, puede modificar de manera significativa los resultados.
- Eliminar las filas que presenten valores ausentes, si el conjunto de datos es suficientemente grande y no se pierde información relevante al eliminar esas filas.
- Eliminar las variables que presentan un porcentaje mayor del 50% de datos ausentes.

```
# Imputar valores faltantes en la columna 'age' con la mediana  
df['age'].fillna(df['age'].median(), inplace=True)
```

```
# Imputar valores faltantes en la columna 'embark_town' con la moda  
df['embark_town'].fillna(df['embark_town'].mode()[0], inplace=True)
```

Detección y tratamiento de datos ausentes. Tratamiento

Experimento:

Detectar los valores perdidos de las columnas y reemplazar por la mediana para evitar perder información significativa y alterar los análisis posteriores

Es recomendable:

- conocer la causa de la ausencia de datos, el tratamiento posterior.
- documentar estrategia aplicada con fines de trazabilidad del proceso, y retornar si fuera posible ante una determinada inconsistencia o debilidad en etapas posteriores del análisis de datos.

Estrategia más común: aplicar la mediana

```
# Verificar los datos faltantes de todas las columnas:  
df.isnull().sum()
```

```
species      0  
island        0  
bill_length_mm    2  
bill_depth_mm    2  
flipper_length_mm  2  
body_mass_g      2  
sex           11  
dtype: int64
```

```
[ ] # Imputación de datos faltantes con la mediana  
df['flipper_length_mm'].fillna(df['flipper_length_mm'].median(), inplace=True)  
  
# Verificar que no hay datos faltantes en 'flipper_length_mm'  
df['flipper_length_mm'].isnull().sum()
```

Detección de filas duplicadas

Determinar observaciones (casos) duplicados

- Consumen espacio de almacenamiento y ralentizan los cálculos
- Pueden distorsionar los resultados del análisis y comprometer la integridad del conjunto de datos.
- Se aplica método `Dataframe.duplicated()` // determina filas duplicadas

```
duplicate = df[df.duplicated(keep="first")] == duplicate = df[df.duplicated()]
```

1er valor es único y resto filas son duplicadas

```
duplicate = df[df.duplicated(keep="last")]
```

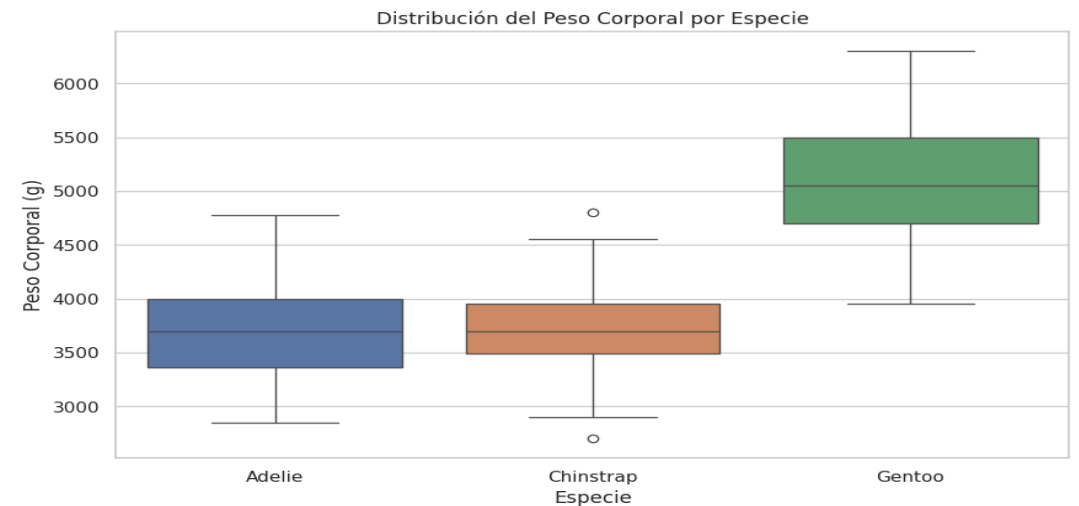
ultimo valor es único y el resto filas son duplicadas

```
duplicate = df[df.duplicated(["species", "sex"])]
```

selecciona todas filas considerando las columnas indicadas con igual valor

Tratamiento de datos atípicos o con ruido

- Datos con valor ruido, atípico u *outlier*,
 - observación significativamente distinta del resto de datos que presenta una variable, de tal magnitud que se puede considerar un valor anómalo
- Se debe detectar y tratar los valores atípicos (outliers) o anómalos para evitar inconvenientes en la descripción e interpretación de datos
- Se debe verificar que no se deben a un error de medición o de construcción del dataset
- Determinar medida de sustitución
- Graficas boxplot, facilitan la identificación de datos atípicos.
- Si se elimina o sustituyen los valores atípicos, repetir los análisis con y sin valores inusuales, para determinar:
 - si el efecto es mínimo, es razonable eliminarlos o sustituirlos.
 - si el efecto es sustancial, no ignorarlos sin justificación.



Creación de variables categóricas

Dado el estudio, necesidad de crear nuevas variables

```
# Crear una nueva característica
penguins['masa_index'] = penguins['body_mass_g'] /
penguins['flipper_length_mm']

# comprobar
print(penguins[['body_mass_g', 'flipper_length_mm', 'masa_index']].head())
```

Transformación variables categóricas

Las variables categóricas pueden ser:

Nominales:

- Se trata de símbolos o nombres que representan algún tipo, código o estado.
- Ejemplos: deportes, países, otros

Ordinales:

- Se trata de atributos que presentan un posible orden.
- Ejemplos: nivel de aceptación de un servicio, otros

Los modelos de ML/CD requieren de datos no categóricos, para el procesamiento

Se deben preparar los datos de formas específicas antes de su tratamiento con modelos de ML

Se transforman a variables que asumen como valor 0 o 1 para indicar la ausencia o presencia

Técnicas:

- Codificación de enteros, etiquetas y ordinal
- Codificación One-Hot
- Codificación Dummy o variable ficticia

Transformación variables categóricas

Los modelos de ML/CD requieren de datos no categóricos, para el procesamiento

Se deben preparar los datos de formas específicas antes de su tratamiento con modelos de ML

Codificación de enteros

codifica los valores como números enteros.

Es decir, reemplaza cada una de las k categorías de la variable por un número entero. La nueva variable se forma por números enteros pertenecientes al intervalo $[1, k]$

puede afectar rendimiento del algoritmo,

	Acceso a internet
registro 1	Frecuente
registro 2	Escaso
registro n	Intermedio

	Acceso a internet
registro 1	3
registro 2	1
registro n	2

Transformación variables categóricas

Codificaciones One-Hot y Dummy

Codificadores basados en la idea de generar variables auxiliares binarias de forma que identifiquen de manera única las diferentes k categorías de la variable X_j

Codificación One-Hot

codifica los valores como una matriz de vector binario

crea una columna para cada valor posible y coloca un 1 en la columna correspondiente.

evita la confusión de ordinalidad de números enteros

Codificación Dummy o variable ficticia

similar a **codificación** One-Hot, pero requiere una columna menos

Transformación variables categóricas

Codificación One-Hot

codifica los valores como una matriz de vector binario

crea una columna para cada valor posible y coloca un 1 en la columna correspondiente.

evita la confusión de ordinalidad de números enteros

	Acceso a internet
registro 1	Frecuente
registro 2	Escaso
registro n	Intermedio

	Acceso a internet
registro 1	3
registro 2	1
registro n	2

	Acceso_I_1	Acceso_I_2	Acceso_I_3
registro 1	0	0	1
registro 2	1	0	0
registro 3	0	1	0

Transformación variables categóricas

Codificación Dummy o variable ficticia

similar a **codificación** One-Hot, pero requiere una columna menos Redundancia.

En el ejemplo si la variable no asume valor frecuente o intermedio, es escaso.

Técnica. Aplicar variables codificadas ficticias. El número es k-1, es decir, uno menos que el número de valores posibles

	Acceso a internet
registro 1	Frecuente
registro 2	Escaso
registro n	Intermedio

	Acceso a internet
registro 1	3
registro 2	1
registro n	2

	Acceso_I_1	Acceso_I_1	Acceso_I_1
registro 1	0	0	1
registro 2	1	0	0
registro 3	0	1	0

	Acceso_I_1	Acceso_I_1
registro 1	0	1
registro 2	0	0
registro 3	1	0

Normalización y estandarización

Técnicas de pre-procesamiento de datos, mejora de la interpretación y calidad del modelo:

- ajusta las características o atributos de los datos en un rango específico (útil si están en unidades diferentes),
- asegura que todas las características tengan el mismo peso.
- evita problemas numéricos durante el entrenamiento del modelo.
- interpretación consistente, los datos están en la misma escala.

Técnicas

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Normalización

- se aplica la técnica, si se desconoce la distribución de los datos o se sabe que la distribución no es gaussiana.
- útil si los datos tienen escalas variables y el algoritmo que se utiliza no hace suposiciones sobre la distribución de los datos (k vecinos más cercanos, RNA).
- normaliza atributos numéricos entre el rango de 0 y 1, usando escalar mínimo-máximo, o escalares robustos

Técnicas

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

Estandarización

- asume que los datos tienen una distribución gaussiana
- estandariza los atributos numéricos, media = 0 y varianza = 1 utilizando un escalar estándar
- útil si los datos tienen escalas variables y el algoritmo seleccionado hace suposiciones acerca que los datos tienen una distribución gaussiana (Ej. regresión lineal, regresión logística y análisis discriminante lineal).