Diplomatura Universitaria en Ciencia de Datos

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

Módulo 3. Análisis Exploratorio de Datos

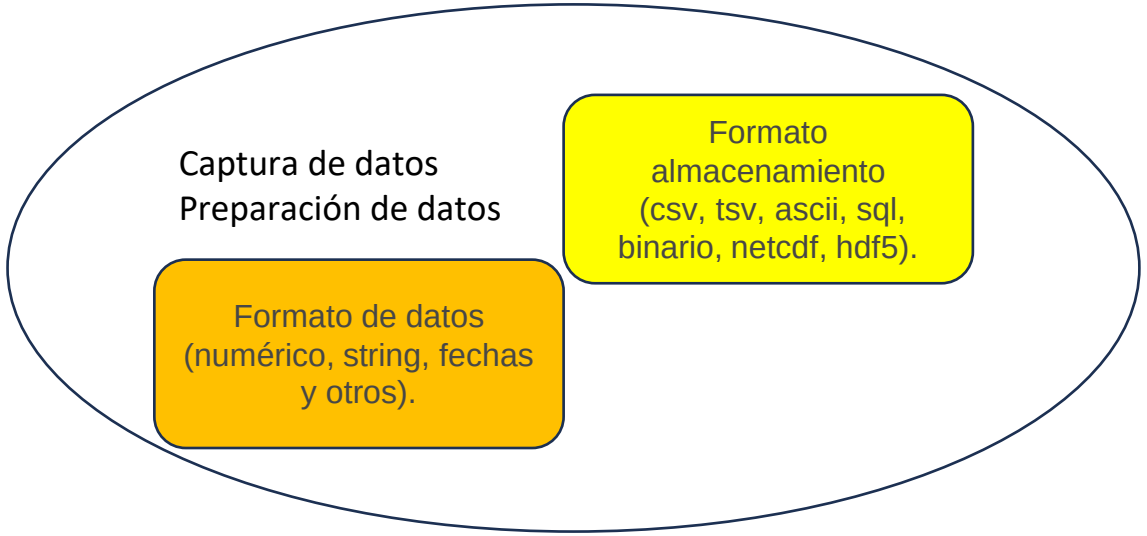
Equipo Docente:

Dra. Sonia I. Mariño

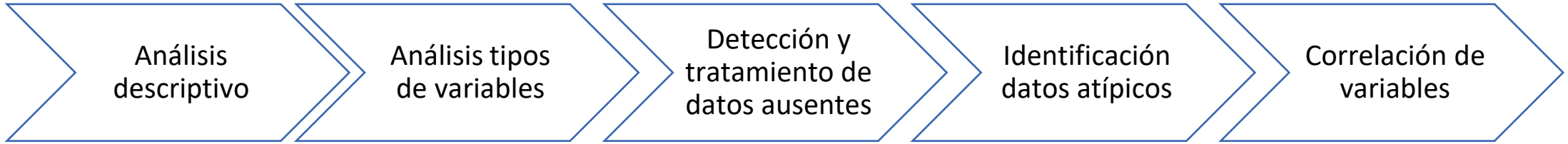
Lic. Lucia del Valle Ledezma

Lic. Rafael Perez

Proceso EDA



ANÁLISIS EXPLORATORIO DE DATOS



evaluación y corrección de datos.

EDA univariado

EDA bivariado
EDA multivariado

Documentar las decisiones en el proceso

EDA, niveles

Nivel 1 – EDA descriptivo, univariado

- Estudio centrado en una variable
- Visualización univariante de datos, estadísticas de resumen.

Nivel 2 – EDA inferencial, bivariado

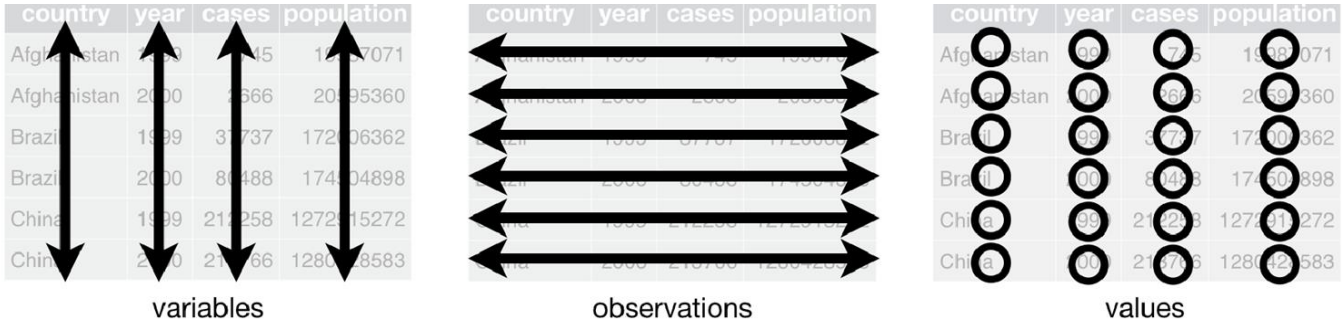
- Estudia una variable en función de otra.
- Visualizaciones bivariantes, estadísticas de resumen

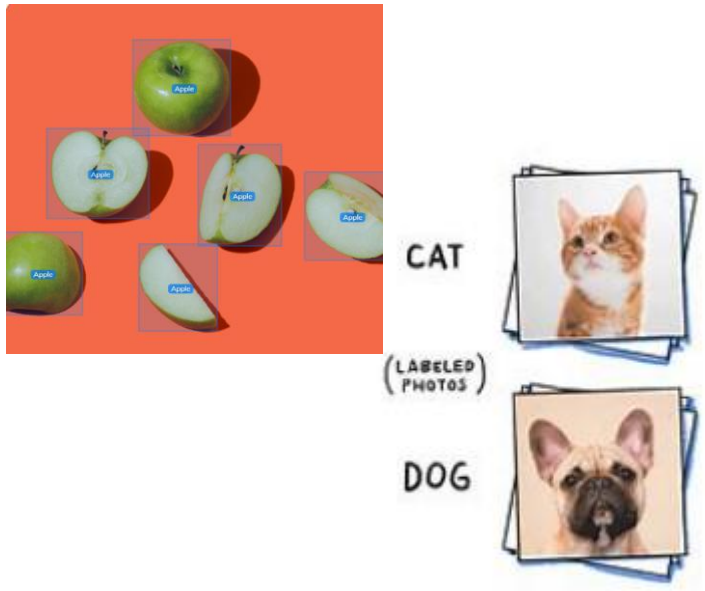
Nivel 3 – EDA modelización, multivariado

- Centrado en los indicadores disponible para estudiar un fenómeno determinado.
- Clasificación cruzada, análisis de varianza y regresiones simples.
- Visualizaciones multivariadas, mapear y comprender las interacciones entre variables

Datos estructurados vs. no estructurados

- Datos
- Procesos
- Modelos





BD tidy **(Leek 2015):**
Cada variable medida debe estar en una columna.
Cada observación distinta de la variable debe estar en una fila diferente.

Datos estructurados vs. no estructurados

Datos estructurados

- se organizan en un formato fijo y predefinido. Disponibles en archivos tipo texto, base de datos, hojas de cálculo. organizados en columnas (etiquetadas) y filas

Datos no estructurados

- sin formato predefinido o específico. Objetos con datos sin organización, sin una estructura comprensible. El valor de estos datos se obtiene de un procesamiento: ej. identificar y almacenar organizadamente el texto al aplicar algoritmos.
- mayor dificultad para su procesamiento respecto a los datos estructurados.
- datos cualitativos, principalmente de naturaleza categórica y característica.

Datos semi estructurados

- presentan un nivel medio de organización y clasificación.
- utilizan metadatos para agruparse y almacenarse.

Datos no estructurados

Ejemplos, datos no estructurados disponibles en:

- Documentos en archivos de texto
- Archivos PDF
- Archivos de registro y de datos de aplicaciones como .dll
- Datos en la web: Instagram, Facebook y Twitter
- Imágenes
- Archivos de audio o grabaciones telefónicas
- Vídeos
- Correos electrónicos

Datos no estructurados. Caso de aplicación

Caso de aplicación de Datos no estructurados de texto: Análisis de sentimientos

- Objetivo: conocer y comprender la actitud general, positiva, negativa o neutra, expresada por el autor de un texto.
- Proceso de utilizar técnicas computacionales para evaluar y determinar la polaridad emocional asociada con un conjunto de datos de texto.
- Identificar temas, sentimientos y relaciones en los datos de texto, para comprensión y análisis Se basa en fundamentos de distintas disciplinas: lingüística, psicología, procesamiento del lenguaje natural.

Ej. Algunas aplicaciones de análisis de sentimientos:

- seguimiento de comentarios en redes sociales.
- reseñas de productos: nivel de satisfacción.
- evaluación de opiniones de clientes
- estudios de mercado: resultado de un análisis de satisfacción de un producto y/o servicio.
- evaluación de campañas diversas: éxito o no éxito.

Datos no estructurados.

Aplicación: Análisis de sentimientos

Técnicas

- Enfoques basados en Support Vector Machines (SVM), Naive Bayes o RN.
    - Aprendizaje de patrones y relaciones a partir de datos etiquetados.
- Enfoque de Procesamiento del Lenguaje Natural (NLP)
    - basados en el léxico: métodos que aplican diccionarios o léxicos predefinidos que contienen palabras y puntuaciones de sentimiento asociadas.
    - semántico: basado en similitudes semánticas mediante la utilización de embeddings.
- Enfoques híbridos:
    - Combinación de enfoques basados en el léxico con modelos ML, mejorarían resultados.

Fases asociadas al análisis del texto

1. Recopilar el conjunto de datos.
2. Transformación de datos no estructurados a estructurados. Etiquetar los datos, asociar la representación de diferentes emociones o sentimientos.
3. *Preprocesar los datos: eliminar ruido (por ejemplo eliminar caracteres extraños y stopwords) y normalizar el texto (tokenización, lematización, stemming, etc.). Identificar número de párrafos o comentarios, identificar número de palabras por comentarios, identificar frecuencia de las palabras,*
4. Dividir los datos, en conjuntos de entrenamiento, (si corresponde de validación) y prueba.
5. Entrenar un algoritmo de aprendizaje automático. (puede ser un algoritmo lexicográfico o semántico)
6. Evaluar el rendimiento del modelo: mediante métricas, ¿Qué se quiere medir?, tipo de aprendizaje y sus métricas específicas.
7. Ajustar y mejorar el modelo según sea necesario. Descubrimiento y exploración de los resultados ,.resuelve las necesidades del negocio???

Aceptado el modelo, puede aplicarse para inferir nuevos textos y determinar el sentimiento expresado en ellos.

Fase 1: Fase de Exploración de datos

EDA, actividades:

- frecuencia del número de palabras positivas en tuits marcados como positivos,
- frecuencia del número de palabras negativas en tuits marcados como negativos,
- frecuencia del número de palabras negativas en tuits marcados como positivos,
- frecuencia del número de palabras positivas en tuits marcados como negativos
- Balance de clases: grafica frecuencia tuits positivos y tuits negativos.

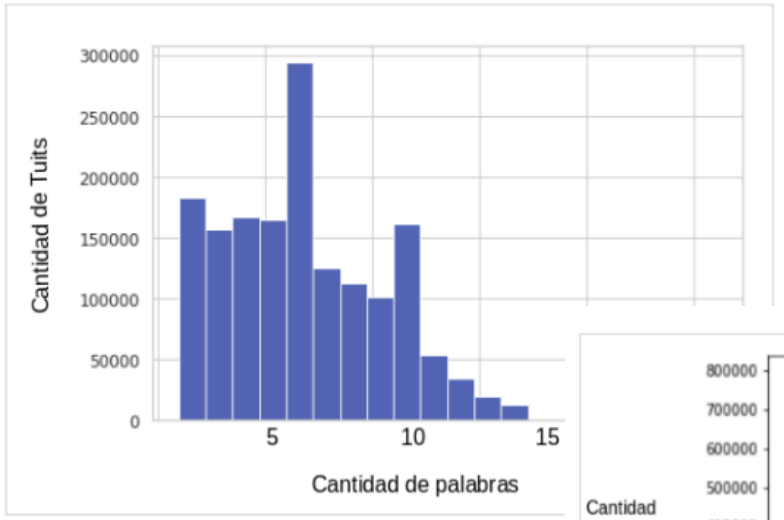
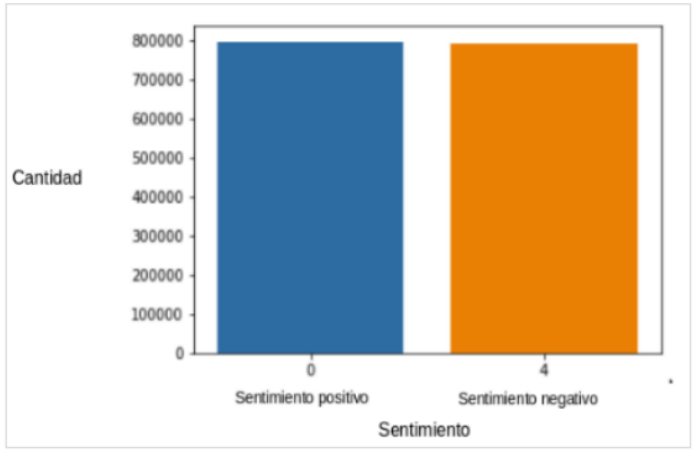


Figura 3. *Cantidad de palabras*



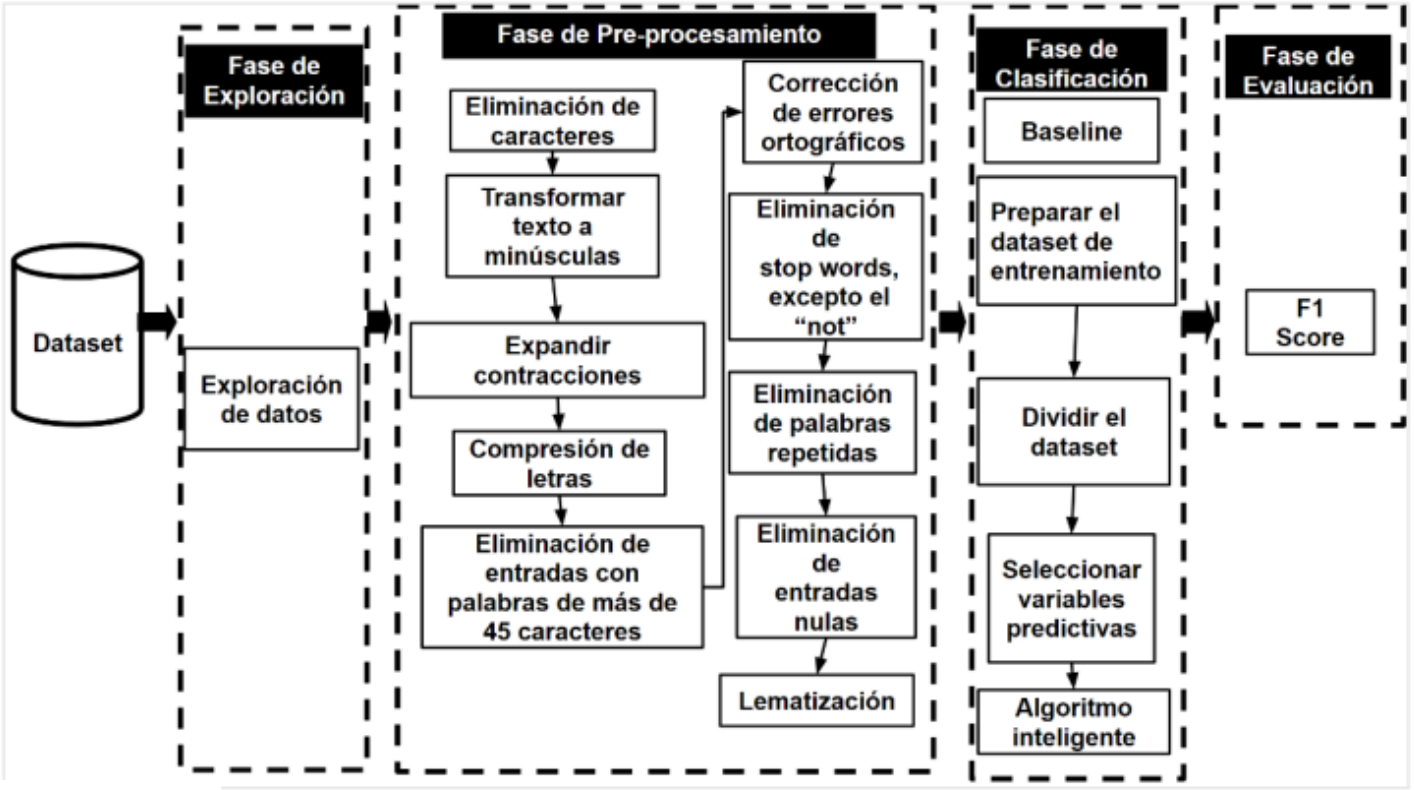


Figura 1. *Esquema general del enfoque metodológico propuesto*

Lovera, F. A. & Cardinale, Y. (2023). Análisis de sentimientos en Twitter: Un estudio comparativo. Revista Científica de Sistemas e Informática, 3(1), e418. https://doi.org/10.51252/rcsi.v3i1.41

Fases asociadas al preprocesamiento del texto.

- Normalización de texto, estandarizar para garantizar la uniformidad en el conjunto de datos.
  - convertir texto a minúsculas, eliminar signos de puntuación y manejo de caracteres especiales, eliminar espacios, eliminar números, otros.

- Reducir el ruido en el conjunto de datos y centrar el análisis en contenido significativo
  - dividir el texto en tokens individuales.
  - eliminar las palabras irrelevantes, palabras comunes no informativas. “Stop Words”.

- Lematización, derivación y Stemming:
  - reducir las palabras a sus formas raíz.
  - se utilizan para comprender el sentimiento asociado.

Visualización

Distintas técnicas de visualización de datos, facilitan identificar las tendencias y patrones en los datos analizados

- gráficos de barras,
- gráficos de dispersión.
- mapas de calor.

Etiquetado de datos

Etiquetado de datos:
cuello de botella humano de la IA
Costo

Etiquetado vs. anotación

- Etiquetado,
  - asigna etiquetas a datos
- Anotación,
  - Asigna metadatos o información adicional a los datos etiquetados.
- Ejemplos ?

Etiquetado de datos

Etiquetado manual vs. etiquetado automático

- Aplicar algoritmos de ML para agilizar y simplificar el proceso de etiquetado.
- El sistema aprende a reconocer patrones importantes en los datos para asignar etiquetas relevantes sin intervención humana.
- Riesgos:
  - rapidez vs. precisión en el etiquetado
  - manipular datos complejos o subjetivos.

Híbrido.

Etiquetador de datos

Perfiles: In-housing vs outsourcing vs crowdsourcing

Se necesita cierto trabajo manual.

Estrategia de etiquetado:

- in-house,
- outsourcing,
- crowdsourcing

Basada en:

- requerimientos y condiciones de la organización.
- tiempo, costo, calidad, seguridad de datos…

	Outsource	In-house	Crowdsource
Time required	Average	High	Low
Price	Average	Expensive	Cheap
Quality of labeling	High	High	Low
Security	Average	High	Low

Algunas prácticas de etiquetado de datos

- **Definir las pautas de etiquetado**
  - establecer criterios específicos para lograr precisión y consistencia en el proceso.
- **Proporcionar una formación integral**
  - formar en criterios y directrices. Establecer requisitos, asegurar la precisión en el etiquetado.
  - establecer ejemplos, escenarios para comprender la tarea.
- **Revisión de datos etiquetados**
  - seguimiento de pautas en el proceso de etiquetado.
  - detectar errores o diferencias en el proceso de etiquetado, permite errores y corregirlos.
- **Equilibrar la calidad y la cantidad**
  - equilibrar la calidad y la cantidad de los datos etiquetados.
  - garantizar disponibilidad de datos etiquetados de alta calidad.

Algunos ejemplos de etiquetado

En un texto se pueden aplicar distintas etiquetas

1. Etiqueta de sentimiento

- evalúa actitudes y emociones. Ej: positivo, negativo o neutral.

2. Etiqueta de intención,

- se asocia con la necesidad o motivo del texto. Ej. intención de solicitud, de confirmar, negar, orden

3. Etiqueta semántica

- asigna varias etiquetas al texto que asocian pensamientos y entidades, como personas, lugares o temas.

4. Etiqueta de relación

- Descripción de relaciones dentro de diversas partes del documento.

5. Etiqueta de texto en NLP,

- ML aplicado a lectura, interpretación, análisis y presentación de texto de relevante.