



# SISTEMAS INTELIGENTES PARA CIENCIA DE DATOS

Maestría en Ciencia de Datos

Universidad de la Ciudad de Aguascalientes

# Tipos de preprocesamiento de datos

- Selección
- Tratamiento de valores faltantes
- Tratamiento de valores atípicos
- Transformación de formatos
- Normalización
- Tratamiento de duplicados
- Integración de múltiples fuentes
- Reducción de dimensionalidad
- Creación de características

# Selección de datos

- Selección de columnas
- Selección de registros
  - Muestreo
  - Por propiedades

# Tratamiento de valores faltantes

- ¿Qué hacer cuando tenemos?
  - Na, NaN, "", NULL
- Eliminar los registros
  - Sencillo
  - Se pierde información
- Imputación
  - Valor aleatorio
  - Media/Mediana general
  - Media por grupos
  - Vecino mas cercano
  - Regresión lineal
  - Algoritmos especializados de imputación

# ¿Valores atípicos? (outliers)

- Valores son inusuales o extremos en comparación con el resto
- Causados por errores en
  - Medición
  - Captura
- También pueden ser legítimos
- Métodos para detectar atípicos
  - Análisis visual
  - Boxplot
  - Análisis de densidad
  - Regresión
  - Medidas de distancia
- Eliminar los registros
- Imputación

# Transformación de formatos

- Todo lo que ya vimos de
  - Texto
  - Audio
  - Imagen
- Agrupamiento
- Separación de datos
- Codificación de variables categóricas
- Agregación de datos



# Halftime Show



# BONUS: Particiones

- Separar los datos en 2 partes
  - Entrenamiento (train) (70%-95%)
  - Prueba (test) (5%-30%)
- Resultados nunca han sido vistos por el modelo
- Aplicar transformaciones
- Evitar usar la parte de prueba
  - Utilizar parte de entrenamiento (validación)



# Normalización (estandarización)

- Datos tienen distintos orígenes e interpretaciones
  - Escalas son distintas
- Transformar los datos a una escala común
- Métodos
  - Min-Max
  - Z-score
  - Logarítmico
- Estas transformaciones afectan la interpretación

# Tratamiento de duplicados

- ¿Qué consideramos un duplicado?
  - Registro completo
  - Identificador único
  - Cierta cantidad/selección de valores
- Tratamientos
  - Conservar solo uno
    - Primero
    - Ultimo
    - Mas alto
  - Agregar los datos
  - Conservar todos

# Integración de múltiples fuentes

- ¿Qué hacemos cuando tenemos los datos en distintas tablas?
- Unión de tablas
  - Left
  - Right
  - Inner
  - Outer
- Cardinalidad
  - 1-1
  - 1-muchos
  - Muchos-1
- Revisar que las uniones son correctas
  - Duplicados
  - Faltantes

# Reducción de dimensionalidad

- Gran cantidad de variables/columnas
  - Información repetida
  - Información no importante
- Utilizar todas las variables
  - Mas no es mejor
  - Modelos mas difíciles de entrenar
- Ruido
- Relaciones espurias
- Menor desempeño
- Selección de variables
  - Conocimiento del dominio
  - Análisis estadístico
- Componentes principales

# Creación de características

- Crear nuevas características a partir de las existentes
  - Porcentajes
  - Tasas
  - Máximos/Mínimos
  - Cruces



# Notas finales

La aplicación de técnicas de preprocesamiento

Puede afectar la calidad de los resultados

Puede desbloquear nuevos análisis

Es una tarea iterativa

Puede ser fácilmente la tarea que demande mas tiempo/esfuerzo

The image features a light gray background with decorative elements in the corners. The top-left corner contains several circles in shades of teal, light green, and orange. The bottom-right corner features a cluster of circles in shades of teal, light green, and white. The text 'colab' is centered in the middle of the image.

colab