ANALISIS DE DATOS
INGENIERIA DE SISTEMAS

Autor(es):

Stiven Osorio Roldan

Mariana Álvarez Posada

Mateo Martinez Franco

Instituto Tecnológico Metropolitano - ITM

Medellín

2025

Base de datos 1:

Spotify churn:

1. Fuente

- **Origen**: Kaggle (plataforma de datasets abiertos).

2. Tipo de datos

- **Secundaria**: provienen de registros recopilados y compartidos en Kaggle, no recolectados directamente por el investigador.

3. Características básicas

- **Número de registros (filas)**: 8,000 usuarios únicos.
- **Número de atributos (columnas)**: 12 variables (ej.: edad, género, país, tipo de suscripción, tiempo de escucha, tasa de skips, dispositivo, etc.).
- **Tamaño del archivo**: ≈ 382 KB.
- **Tipo de datos**: mixto (numéricos y categóricos).

4. Nivel de documentación disponible

- Kaggle proporciona una descripción clara del propósito: **predecir churn (cancelación de suscripción)**.
- Se cuenta con **definición de variables y target** (is_churned), lo que facilita el entendimiento.
- Nivel de documentación: **medio a alto**, ya que incluye metadatos, explicación de uso y caso de negocio.

5. Posibles aplicaciones

- **Modelado predictivo**: construir modelos de Machine Learning para predecir qué usuarios cancelarán su suscripción.
- **Análisis de comportamiento**: entender patrones de uso (tiempo de escucha, skips, anuncios, offline).
- **Estrategias de retención**: identificar factores que afectan la fidelización y diseñar acciones de marketing.
- **Segmentación de clientes**: agrupar usuarios según hábitos de consumo y tipo de suscripción.

- **Optimización de producto**: mejorar experiencia en dispositivos, ajustar promociones según países o perfiles de edad.

Base de datos 2:

Documentación del Conjunto de Datos de Señales de Tránsito

1. Fuente y tipo de datos

- **Fuente:**
    - Kaggle - Car Detection Dataset
    - Roboflow - Self Driving Cars Dataset

Tipo de datos: La fuente de los datos es **secundaria**, ya que el conjunto de datos fue recopilado y preprocesado previamente por otros investigadores, en este caso, de plataformas como Kaggle y Roboflow. El tipo de datos es **imagen**, lo que es ideal para proyectos de visión por computadora.

---

2. Características básicas

- **Número de registros:** 4,969 imágenes.
- **Número de atributos:**
    - Imagen (archivo de entrada).
    - Clase asociada (etiqueta de señal de tránsito).
- **Clases (15 en total):**
    - Luz verde
    - Luz roja
    - Límite de velocidad 10
    - Límite de velocidad 20
    - Límite de velocidad 30
    - Límite de velocidad 40
    - Límite de velocidad 50
    - Límite de velocidad 60
    - Límite de velocidad 70

    - Límite de velocidad 80
    - Límite de velocidad 90
    - Límite de velocidad 100
    - Límite de velocidad 110
    - Límite de velocidad 120
    - Señal de **Parar**

- **División de datos:** El conjunto de datos está dividido en tres partes:
    - **Entrenamiento**: Usado para entrenar el modelo.
    - **Validación**: Usado para ajustar los hiperparámetros del modelo y evitar el sobreajuste.
    - **Prueba**: Usado para evaluar el rendimiento final del modelo con datos que no ha visto antes.

---

3. Nivel de documentación disponible

- **Disponible:**
    - Información sobre el número de imágenes y clases.
    - División estándar (train/valid/test).
    - Etiquetas organizadas y consistentes para cada imagen.
- **Limitaciones:**
    - La documentación no incluye metadatos adicionales (ubicación geográfica, condiciones de iluminación, clima, cámara utilizada, etc.).
    - No se detalla un análisis de balance de clases (algunas señales pueden estar sobrerrepresentadas respecto a otras).

---

4. Posibles aplicaciones

- **Navegación de vehículos autónomos:** reconocimiento en tiempo real de señales de tránsito.

- **Cumplimiento de normas de tráfico:** alertas automáticas al conductor cuando excede límites o ignora señales.
- **Programas de capacitación en seguridad vial:** entrenamiento de conductores y simulaciones educativas.
- **Infraestructura de ciudad inteligente:** integración en sistemas de CCTV/IoT para vigilancia vial en tiempo real.
- **Análisis de red vial:** uso en estudios de transporte e ingeniería civil para optimizar infraestructura y mejorar la seguridad.

Base de datos 3:

Documentación del Dataset *Iris*

1. Fuente y tipo de datos

- **Fuente original:** recolectado por Ronald A. Fisher en 1936.
- **Disponibilidad actual:** UCI Machine Learning Repository (Universidad de California, Irvine).
- **Tipo de datos:** secundarios → originalmente fueron primarios (observaciones botánicas), pero ahora se usan desde repositorios ya procesados.

---

2. Características básicas

- **Número de registros:** 150 observaciones.
- **Número de atributos:** 5 en total →
    - 4 predictivos:
        - Longitud del sépalo (cm)
        - Anchura del sépalo (cm)
        - Longitud del pétalo (cm)
        - Anchura del pétalo (cm)
    - 1 objetivo: especie (setosa, versicolor, virginica).
- **Distribución:** 50 registros por especie.
- **Tamaño del archivo:** ~4.8 KB en formato CSV.

- **Calidad de datos:** sin valores nulos ni inconsistencias.

---

3. Nivel de documentación disponible

- Publicación original en *Annals of Eugenics* (1936).
- Referencias en miles de artículos, libros y cursos de estadística y machine learning.
- Descripción completa de variables, unidades y contexto biológico.
- Disponible en múltiples formatos (CSV, ARFF, JSON) y compatible con R, Python, WEKA, etc.
- Estándar de comparación en algoritmos de clasificación supervisada.

---

4. Posibles aplicaciones

- **Modelado predictivo:** clasificación multiclase (SVM, k-NN, Naive Bayes, Árboles de decisión, Redes neuronales).
- **Análisis exploratorio:** histogramas, boxplots, pairplots, correlaciones.
- **Reducción de dimensionalidad:** PCA, LDA para visualización y separación de especies.
- **Educación:** dataset introductorio en cursos de ciencia de datos.
- **Bioinformática:** taxonomía computacional, estudios evolutivos y análisis de agrupamientos.