

Entrega 1: Descripción y Formulación del Objetivo

1. Descripción general del proyecto

En este proyecto se aplicarán técnicas de **Aprendizaje Supervisado** con el objetivo de desarrollar un modelo capaz de **clasificar el género musical de una canción** a partir de sus características de audio y metadatos obtenidos de Spotify.

El dataset utilizado proviene de **Kaggle**, bajo el título “[*Spotify Data Visualization*](#)”, y contiene **2000 registros** de canciones populares comprendidas entre los años **2000 y 2019**. Cada registro incluye información sobre el artista, nombre de la canción, duración, nivel de energía, bailabilidad, acústica, entre otras variables, así como el **género musical** asociado, que será la variable objetivo.

2. Contexto y relevancia del problema

La clasificación automática de géneros musicales es un problema clásico y relevante dentro del campo del *Machine Learning* aplicado a la música y al análisis de datos multimedia. Plataformas de streaming, como Spotify o Apple Music, utilizan este tipo de algoritmos para **mejorar sus sistemas de recomendación, etiquetar contenido automáticamente y analizar tendencias culturales o de consumo**.

Comprender cómo las características sonoras (energía, ritmo, acústica, etc.) se asocian con distintos géneros permite no solo optimizar sistemas de recomendación, sino también estudiar la evolución de la música a lo largo del tiempo y los patrones que determinan el éxito de ciertos estilos.

Además, trabajar con este tipo de dataset resulta didáctico porque combina **variables numéricas y categóricas**, exige **preprocesamiento de datos** y permite aplicar **modelos supervisados de clasificación** vistos en clase.

3. Objetivos del proyecto

Objetivo general

Desarrollar un modelo de aprendizaje supervisado capaz de **predecir el género musical de una canción** en base a sus características cuantitativas y cualitativas presentes en el dataset [*Spotify Data Visualization*](#).

Objetivos específicos

- Analizar y comprender la estructura del dataset, identificando las variables más relevantes para la clasificación.
- Realizar el **preprocesamiento de los datos**, abordando valores faltantes, codificación de variables categóricas y normalización de características numéricas.

- Explorar el comportamiento de las variables mediante análisis descriptivo y visualizaciones.
- Entrenar y comparar distintos modelos de clasificación supervisada:
 - **K-Nearest Neighbors (KNN)**
 - **Árbol de Decisión (Decision Tree Classifier)**
- Evaluar el rendimiento de los modelos utilizando métricas adecuadas (precisión, recall, F1-score, matriz de confusión).
- Determinar cuál modelo ofrece mejor desempeño en la clasificación de géneros y discutir posibles mejoras.

4. Tipo de problema

Cada fila representa un ejemplo de entrenamiento (una canción), y las variables predictoras son las características cuantitativas asociadas a ella, tales como:

- **danceability**
- **energy**
- **speechiness**
- **acousticness**
- **instrumentalness**
- **liveness**
- **valence**
- **tempo**
- **loudness**, entre otras.

El modelo aprenderá a asociar patrones entre estas variables y el género correspondiente, permitiendo luego predecir el género de una canción no vista previamente.

5. Modelos candidatos

Los modelos supervisados que se aplicarán para resolver el problema de clasificación son:

1. **K-Nearest Neighbors (KNN):**
Modelo basado en la similitud entre canciones. Se espera que canciones con características similares pertenezcan al mismo género.
2. **Árbol de Decisión:**
Permite interpretar fácilmente qué variables son más importantes en la clasificación y cómo se toman las decisiones.

6. Conclusión esperada

Se espera obtener un modelo capaz de predecir con una **precisión razonable** el género musical de una canción a partir de sus características, demostrando el proceso completo de un flujo de trabajo de *Machine Learning supervisado*: desde el análisis exploratorio y preprocesamiento hasta el entrenamiento y evaluación de modelos.

El proyecto permitirá además reflexionar sobre la importancia de la calidad de los datos, la selección de características y la elección del modelo más apropiado según la naturaleza del problema.

Link del dataset: [*Spotify Data Visualization*](#).