### **Descripción del Proceso Realizado**

Este proyecto aborda el análisis, preprocesamiento y modelado de un conjunto de datos musicales, con el objetivo de clasificar las canciones según su género. Se utilizan técnicas de aprendizaje supervisado, no supervisado y reducción de dimensiones, con un enfoque en la implementación de un **clasificador Random Forest** y un análisis mediante **Componentes Principales (PCA)**. A continuación, se describe el flujo del proceso realizado en el código:

#### **1. Importación de Librerías**

En la primera sección del código, se importan las librerías necesarias para la manipulación de datos, visualización y modelado. Entre ellas destacan:

* **Pandas** y **NumPy** para la manipulación de datos.
* **Matplotlib** y **Seaborn** para la visualización de gráficos.
* **Scikit-learn** para la implementación de modelos de machine learning, como Random Forest y PCA.
* **Imbalanced-learn** para la aplicación de técnicas de sobremuestreo, como SMOTE.

#### **2. Lectura y Limpieza de los Datos**

El conjunto de datos es leído desde un archivo CSV ubicado en un repositorio de GitHub. Posteriormente, se inspecciona el dataset utilizando el método info() para verificar la estructura y tipos de los datos. Durante el proceso de limpieza, se realizan varias acciones:

* **Eliminación de columnas irrelevantes**: Se eliminan columnas como track\_id, artist, y otras que no aportan valor para el análisis y modelado.
* **Manejo de valores nulos**: Se eliminan las filas con valores nulos en la columna genre, que es la variable objetivo del análisis.
* **Transformación de valores**: La columna genre se transforma para clasificar las canciones en dos categorías: "Rock" y "No Rock", basándose en la presencia de la palabra 'Rock' en el género.

#### **3. Análisis Exploratorio de Datos (EDA)**

En esta fase, se exploran las características de los datos para entender su distribución y posibles relaciones. Se crean varios gráficos para visualizar:

* **Distribución de la variable key**: Se muestra la frecuencia de cada clave musical.
* **Distribución de la variable mode**: Se visualiza la cantidad de canciones en cada modo.
* **Distribución del género**: Se observa el número de canciones por género musical.

Además, se realiza un análisis de las características numéricas, como tempo, year, duration\_ms, entre otras, mediante histogramas y diagramas de caja para identificar posibles valores atípicos.

#### **4. Manejo de Valores Atípicos y Preprocesamiento**

El siguiente paso es el tratamiento de valores atípicos. Para ello, se utiliza la técnica de **capping** (o recorte) en varias columnas numéricas del dataset. Esta técnica asegura que los valores fuera de un rango razonable sean ajustados a un valor límite, evitando que afecten negativamente el modelo.

Se realiza el **balanceo de clases** utilizando **SMOTE** (Synthetic Minority Over-sampling Technique), una técnica de sobremuestreo que genera muestras sintéticas para las clases minoritarias en el dataset, con el fin de obtener una distribución equilibrada de las clases.

Finalmente, se aplica **estandarización** a las características numéricas para ponerlas en una escala común y mejorar el rendimiento de los modelos de machine learning.

#### **5. Clasificación con Random Forest**

Se implementa un **modelo de Random Forest** para clasificar las canciones según su género. El proceso de modelado se realiza de la siguiente manera:

* **División del dataset**: El conjunto de datos se divide en entrenamiento, validación y prueba utilizando la función train\_test\_split.
* **Entrenamiento y evaluación**: El modelo se entrena con el conjunto de entrenamiento y se evalúa en los conjuntos de validación y prueba. Se reportan métricas como la **precisión**, el **reporte de clasificación** y la **matriz de confusión** para evaluar el desempeño del modelo.

Se evalúa la importancia de las características mediante el método feature\_importances\_, lo que permite identificar qué variables son más relevantes para la clasificación.

#### **6. Codificación de Etiquetas y One-Hot Encoding**

Para mejorar la representación de las clases o solamente para visualizacion, se utilizan dos técnicas de codificación:

* **Label Encoding**: Se transforma la variable genre en números enteros, asignando un valor único a cada clase.
* **One-Hot Encoding**: Se crea una representación binaria de la variable genre, donde cada clase se representa como una columna diferente con valores 0 o 1.

También se aplica **discretización** a las características numéricas restantes utilizando el método **KBinsDiscretizer**, que agrupa los valores en intervalos.

#### **7. Clasificación con Árboles de Decisión**

Además de Random Forest, se implementa un **modelo de Árbol de Decisión** para la clasificación del género musical. Se entrena el modelo y se evalúa utilizando las métricas estándar de clasificación. Se visualiza la **matriz de confusión** para comparar las predicciones con los valores reales.

### **8. Evaluación del Modelo con 100 Splits**

En este procedimiento, se realiza una **evaluación repetida del modelo** utilizando el clasificador **Random Forest**. El proceso consiste en dividir el conjunto de datos en Académico (primera ejecucion) 80(train)/20(test) – Investigación 50/50 (segunda ejecución) de manera **aleatoria** en **100 repeticiones**. Cada repetición se ajusta a un modelo, se evalúa en el conjunto de prueba y se calcula la precisión. Al final, se calcula la **mediana de las precisiones** obtenidas en todas las iteraciones.

El propósito de realizar 100 splits es obtener una estimación más robusta y estable del rendimiento del modelo, reduciendo el sesgo que puede surgir de una sola división de los datos. Esto proporciona una mejor indicación de la capacidad del modelo para generalizar a datos no vistos.

#### **9. Análisis de Componentes Principales (PCA)**

Se realiza una **reducción de dimensionalidad** utilizando **PCA** para reducir el número de características en el dataset. PCA es una técnica de aprendizaje no supervisado que permite identificar las componentes principales que explican la mayor parte de la varianza en los datos. Se lleva a cabo lo siguiente:

* Se calculan los **valores propios** y **vectores propios** de la matriz de covarianza.
* Se visualiza la **varianza explicada** por cada componente principal.
* Se proyectan los datos originales en un número reducido de componentes principales (12, 11, 10, etc.) y se evalúa el rendimiento de un clasificador Random Forest en los datos reducidos.

Se realiza un análisis de precisión del modelo después de aplicar PCA en diferentes dimensiones. Se comparan los resultados del modelo entrenado en los datos originales con los entrenados en los datos reducidos, lo que permite observar cómo la reducción de dimensiones afecta al rendimiento del clasificador.

#### **10. Aprendizaje No Supervisado (K-Means)**

Se aplica el algoritmo de **K-Means** para realizar un análisis de clustering no supervisado. Se utiliza el **método del codo** para determinar el número óptimo de clusters, y posteriormente se visualizan los resultados del clustering en el espacio reducido de 2 dimensiones utilizando PCA.

### **Conclusiones**

Este proyecto aborda un enfoque integral para el análisis y clasificación de canciones utilizando técnicas de machine learning. Se exploran varias etapas, desde la limpieza y preprocesamiento de los datos hasta la implementación de modelos de clasificación supervisada (Random Forest y Árboles de Decisión), pasando por la reducción de dimensionalidad con PCA y el clustering con K-Means. El uso de técnicas como SMOTE, PCA y la validación cruzada permite mejorar el desempeño y asegurar la robustez de los modelos generados.