Primer Pre Entrega

Alumno: Oviedo Nelso Alejandro

Data Science CoderHouse [61750]

Introducción

El presente análisis tiene como objetivo explorar un conjunto de datos musicales para identificar patrones y relaciones entre varias variables clave. En particular, se busca responder preguntas como: ¿Cuáles son los principales factores que influyen en la popularidad de una canción? ¿Cómo podemos identificar patrones únicos dentro de diferentes géneros musicales? ¿Es posible predecir características musicales basadas en el análisis de los datos? Estas preguntas guían el análisis hacia una mejor comprensión de las dinámicas de la industria musical y refuerzan el enfoque del trabajo. Este estudio está motivado por la necesidad de comprender cómo las características musicales influyen en la popularidad de las canciones y cómo el género musical puede predecir ciertas propiedades. La audiencia beneficiada incluye profesionales de la industria musical, analistas de datos interesados en entretenimiento y desarrolladores de plataformas de streaming que buscan optimizar sus recomendaciones basadas en datos.

Fuente de Datos

Los datos utilizados en este trabajo fueron obtenidos de Kaggle y se encuentran disponibles en el siguiente enlace: Music Genre Classification Dataset. Este conjunto de datos contiene información detallada sobre características musicales, como popularidad, energía, tempo y género, distribuidos en 17 columnas y 17,996 filas. Los datos serán analizados para responder las preguntas planteadas y validar las hipótesis definidas.

Determinación de Condiciones Iniciales

Al comenzar con la carga de datos, verificamos la presencia de valores nulos y aseguramos la integridad de los registros. Las columnas clave que se analizaron incluyen 'Popularity', 'energy', 'loudness', 'danceability', y 'instrumentalness'.

Análisis Exploratorio

Al inspeccionar los primeros registros del DataFrame, observamos que tenemos un conjunto de 17 columnas y 17,996 filas, con algunas variables faltantes que requieren tratamiento. Posteriormente, realizamos un análisis descriptivo para conocer las distribuciones de las variables numéricas y categóricas.

Detección y Resolución de Nulos

A través del análisis de nulos, identificamos varias columnas con valores ausentes, que fueron manejados adecuadamente mediante técnicas como la imputación o la eliminación de registros, dependiendo de la gravedad y el contexto de los datos faltantes. Para mayor claridad, se incluirá un gráfico que muestre la distribución de los valores nulos en el informe final.

Variables Clave para el Análisis

Tras un análisis inicial, identificamos las variables más relevantes para entender las características musicales de las canciones, tales como 'Popularity', 'energy', 'loudness', 'tempo', y 'danceability'.

Hipótesis

1. Las canciones con niveles altos de energía tienen una probabilidad significativamente mayor de ser populares según el dataset analizado.
2. Las características acústicas como ‘danceability’ y ‘tempo’ tienen mayor impacto en la popularidad que otras variables acústicas como ‘loudness’.
3. Los géneros musicales con valores de energía más altos, como el rock y la música electrónica, concentran una mayor popularidad en comparación con otros géneros como el jazz o la música clásica.

Análisis de Relación entre Popularidad y Energía

Al analizar la correlación entre las variables 'Popularity' y 'energy', encontramos que existe una relación débil. La popularidad de una canción tiene una pequeña influencia sobre su energía. Esto sugiere que la popularidad está más influenciada por otros factores como el gusto del público, las tendencias culturales y la exposición en medios de comunicación.

Relación entre el Género Musical y las Variables

Utilizando un boxplot, visualizamos la distribución de la variable 'Popularity' por género musical ('Class'). Como era de esperarse, algunos géneros, como el pop, muestran una mayor concentración de popularidad, mientras que géneros como el jazz o la música clásica presentan una menor concentración. Esto sugiere que los géneros más comerciales tienen una mayor posibilidad de alcanzar una alta popularidad, lo cual es un patrón común en la industria musical.

Distribución de Popularidad por Género

- En los géneros más populares, como el pop y el rock, la popularidad de las canciones tiende a concentrarse en valores altos.
- En géneros menos comerciales, como el jazz, la popularidad presenta una distribución más dispersa, lo que indica que estos géneros tienen menos canciones con alta popularidad.

Energía por Género

Al analizar la variable 'energy' a través de un boxplot, observamos que los géneros con mayor energía, como el rock o la música electrónica, tienen un rango de valores mucho más alto que otros géneros como el jazz o la música clásica. Esto nos indica que ciertos géneros tienden a producir canciones con características energéticas más pronunciadas.

Distribución de Energía por Género

- Géneros como la música electrónica y el rock tienen valores de energía más altos en comparación con géneros más suaves como la música clásica y el jazz.
- Esto refleja las expectativas del público en términos de la dinámica de cada género musical.

Conclusiones Finales

1. **Hipótesis 1:** El análisis muestra que no hay una relación significativa entre la energía de una canción y su popularidad, lo que sugiere que la popularidad depende más de factores externos.
2. **Hipótesis 2:** Las variables ‘danceability’ y ‘tempo’ tienen una mayor influencia en la popularidad, respaldando la hipótesis planteada.

3. **Hipótesis 3:** El género musical tiene una influencia clara en la energía y popularidad de las canciones, especialmente en géneros como el rock y la música electrónica.

Estas conclusiones proporcionan insights valiosos para la industria musical, ayudando a optimizar estrategias de marketing y a diseñar algoritmos de recomendación más efectivos. Además, el análisis exploratorio y preprocesamiento de datos destacó la importancia de manejar adecuadamente los valores nulos y seleccionar variables clave para el análisis.

Las próximas etapas incluirán la selección y entrenamiento de modelos de aprendizaje automático para realizar predicciones más precisas y verificar la aplicabilidad de los algoritmos en producción.

Random Forest

Seguimos avanzando en la utilización del modelo **Random Forest**, el cual sigue confirmando nuestra hipótesis. Se realizaron pruebas con diferentes parámetros y se identificó que la calidad de los datos y los factores externos influyen significativamente en los resultados del modelo.

En resumen, se observó que géneros como el rock establecen un estilo y parámetros fundamentales en términos de energía y popularidad. Si bien es posible asociar ciertas características como la energía con la popularidad, se concluye que los factores externos (como el entorno cultural) tienen un peso considerable, limitando la capacidad del modelo para realizar predicciones totalmente precisas.

Regresión Logística

El modelo de regresión logística se utilizó para predecir el género musical (`Class`) a partir de variables como `tempo`, `danceability`, `energy` y `loudness`. Este modelo arrojó una **exactitud general de 32.56%**. Los resultados indican que:

- Los géneros más representados, como rock e instrumental, obtuvieron mejores métricas de desempeño.
- Géneros con menor representación en los datos (p. ej., hip hop y country) tuvieron un bajo desempeño debido al desbalance de clases.

Aunque los resultados generales fueron modestos, este modelo permitió identificar patrones y limitaciones importantes, como la necesidad de utilizar técnicas de balanceo de clases y modelos más complejos que manejen relaciones no lineales.