UNIVERSIDAD MAYOR DE SAN ANDRÉS
FACULTAD DE CIENCIAS PURAS Y NATURALES
CARRERA DE INFORMÁTICA

Artículo

Análisis de Datos con PCA y K-Means Clustering

Responsable: Badani Dávalos Kassandra Andrea

Asignatura: Inteligencia Artificial (INF-354)

Fecha: 9 de diciembre

La Paz – Bolivia

2024

Resumen de la Investigación: Desarrollo de un Modelo Predictivo de Popularidad Musical utilizando el Top 200 Spotify Songs Dataset

Introducción

La música, como fenómeno artístico, social y tecnológico, ha evolucionado significativamente gracias a la digitalización y plataformas de streaming como Spotify, que ofrecen acceso a grandes volúmenes de datos sobre las canciones y sus características. Esta investigación tiene como objetivo desarrollar un modelo predictivo que permita anticipar la popularidad de una canción en función de atributos musicales, demográficos y contextuales. Para ello, se utiliza el conjunto de datos Top 200 Spotify Songs Dataset, con el atributo "Points (Total)" como indicador de popularidad, el cual se empleará para clasificar las canciones.

Objetivo: El propósito de esta investigación es desarrollar un modelo predictivo que estime la **probabilidad de que una canción sea popular** en Spotify, utilizando para ello características musicales, demográficas y de contexto. Esta predicción se basa en el análisis de datos proporcionados por el conjunto de datos **Top 200 Spotify Songs**. La popularidad de las canciones es evaluada a través de la variable "Points (Total)", que refleja su éxito dentro de la plataforma.

Acciones a realizadas:

Para lograr este objetivo, se llevarán a cabo las siguientes acciones clave:

1. **Preprocesamiento de Datos**:
    - **Escalado y normalización de variables musicales numéricas**: Esto garantiza que las características como la energía, el volumen y la bailabilidad estén en una escala comparable.
    - **Codificación de variables categóricas**: Utilizando técnicas como **one-hot encoding** para representar las variables de continente y nacionalidad, lo cual permitirá que el modelo las procese eficazmente.
    - **Manejo de valores faltantes**: Identificar y tratar las celdas vacías de los datos para evitar que afecten el desempeño del modelo.
2. **Entrenamiento del Modelo**:
    - **Selección del modelo de clasificación**: Dependiendo de la naturaleza de los datos y el objetivo, se entrenarán varios modelos como **regresión logística**, **árboles de decisión** o **redes neuronales**. Estos métodos permitirán identificar patrones en los datos que se correlacionen con la popularidad de las canciones.

    - **Entrenamiento en un conjunto de datos dividido**: Se separará el conjunto de datos en dos partes, una para entrenar el modelo (por ejemplo, 80% de los datos) y otra para evaluar su desempeño (20% restante).
3. **Evaluación del Modelo**:

    - **Métricas de clasificación**: Se utilizarán métricas como **precisión**, **recall**, **F1-score**, **AUC-ROC** y **matriz de confusión** para medir la efectividad del modelo. Estas métricas permitirán evaluar no solo la capacidad del modelo para predecir correctamente las canciones populares, sino también su capacidad para evitar falsos positivos (predecir canciones populares que no lo son).

Impacto y Aplicaciones del Modelo:

El modelo predictivo propuesto permitirá a los profesionales de la industria musical y las plataformas de streaming predecir con mayor precisión qué canciones tienen más probabilidades de volverse populares, lo cual es útil para las decisiones de promoción y marketing. Además, podrá identificar las características clave que influyen en la popularidad de las canciones, lo que podría ser de gran valor para los creadores de música en su proceso de composición y producción.

Atributos Clave para el Modelo Predictivo

Para construir el modelo, se consideran varias características que se cree afectan la popularidad de las canciones. Estos atributos son los siguientes:

- **Danceability:** Mide la facilidad con que una canción puede ser bailada. Se ha observado que las canciones con una alta puntuación en danceability tienden a ser populares en contextos sociales como fiestas y festivales.

- **Energy:** La cantidad de energía en una canción, característica típica de géneros como el pop y la música electrónica. Este atributo es clave, ya que las canciones energéticas suelen atraer a un público más amplio y diverso.

- **Loudness:** El volumen percibido de la canción. Las canciones con un volumen más alto suelen destacarse mejor en plataformas de streaming y eventos en vivo, donde el entorno competitivo es importante.

- **Valence:** La positividad de la canción, en términos de su tono emocional (alegre o triste). Las canciones con una mayor valencia positiva pueden ser más populares durante ciertas épocas del año o en contextos emocionales específicos.

- **Speechiness:** La cantidad de palabras habladas en una canción, que puede correlacionarse con géneros como el rap o spoken word, y puede atraer a diferentes segmentos del público.

- **Artist (Ind.) y Nationality: L**a popularidad de ciertos artistas y la influencia cultural de sus países de origen también juegan un papel fundamental en la popularidad de **sus** canciones.

Clasificación de Canciones Populares

Para clasificar las canciones en "Populares" y "No Populares", se utiliza el cuartil 75% de la puntuación total de popularidad (Points). Las canciones que superan este umbral se clasifican como populares. Esto se implementa con el siguiente código:

```
threshold = df_cleaned['Points (Total)'].quantile(0.75)

df_cleaned.loc[:, 'Is_Popular'] = (df_cleaned['Points (Total)'] > threshold).astype(int)
```

Selección del Clasificador: Random Forest

El clasificador Random Forest fue elegido debido a su capacidad para manejar relaciones no lineales y complejas entre las características, además de ser robusto frente al sobreajuste. Este modelo es particularmente adecuado para conjuntos de datos desbalanceados, como el caso de las canciones populares frente a las no populares. Además, Random Forest permite evaluar la importancia de las características, lo que facilita la identificación de las variables más influyentes en la popularidad de las canciones.

Primera Ejecución:

En la primera ejecución del modelo, con una división 80/20 entre datos de entrenamiento y prueba, el rendimiento general fue excelente, alcanzando una precisión del 99%. La precisión para la clase "No Popular" fue del 100%, pero para la clase "Popular", la precisión fue 81% y el recall del 68%, lo que indica que el modelo tiene dificultades para identificar correctamente todas las canciones populares. Este desbalance en los datos sugiere que el modelo podría beneficiarse de técnicas adicionales para tratar el desbalanceo de clases.

Matriz de Confusión:

La matriz de confusión mostró que el modelo es muy efectivo en identificar canciones no populares, pero algunas canciones populares fueron clasificadas incorrectamente como no populares, lo que se refleja en un número relativamente alto de falsos negativos. Para abordar este

problema, se podría considerar la implementación de técnicas de balanceo de datos, como el sobremuestreo de la clase minoritaria o el ajuste de los umbrales de decisión del modelo.

Validación Cruzada (Splits)

Se llevó a cabo una validación cruzada con 100 divisiones, obteniendo una mediana de precisión del 99.5%, lo que refuerza la estabilidad y confiabilidad del modelo, a pesar de las fluctuaciones en los datos de entrenamiento y prueba. Sin embargo, en la división 50/50 (50% entrenamiento, 50% prueba), la precisión fue del 90.3%, lo que sugiere que el modelo podría beneficiarse de una optimización adicional para manejar mejor el desbalanceo de clases.

Mejoras Potenciales

Aunque el modelo tiene un rendimiento general destacado, se identificaron áreas de mejora, especialmente en el manejo del desbalanceo de clases. Se recomienda explorar técnicas como el sobremuestreo de la clase "Popular" o el uso de métodos como SMOTE (Synthetic Minority Over-sampling Technique) para mejorar el recall y la precisión de la clase "Popular". Además, el ajuste de los hiperparámetros del clasificador Random Forest, como el número de árboles (n_estimators) o la profundidad máxima de los árboles, podría optimizar aún más el rendimiento.

Preprocesamiento de los Datos

El preprocesamiento de los datos siguió estos pasos:

1. **Creación de la variable objetivo ("Is_Popular"):** Se definió como una variable binaria para clasificar las canciones en populares y no populares, según el cuartil 75% de los "Points".

2. **División de los datos:** El conjunto de datos se dividió en un 80% para entrenamiento y un 20% para evaluación del modelo.

Análisis de Componentes Principales (PCA)

El PCA se utilizó para reducir la dimensionalidad del conjunto de datos y mejorar la eficiencia del modelo, sin perder información significativa. El PCA transforma las características originales en componentes no correlacionados, lo que permite identificar las dimensiones más relevantes. Se probaron diferentes cantidades de componentes principales (12, 10, 9, 5) y los resultados mostraron que la reducción de la dimensionalidad no afectó negativamente el rendimiento del modelo, que mantuvo una precisión del 94% incluso con menos componentes.

Interpretación Final:

- **Reducción de Dimensionalidad:** El PCA fue eficaz para reducir significativamente el número de características sin afectar el rendimiento. La mayoría de la información útil para predecir la popularidad de las canciones está contenida en un pequeño número de componentes principales.

- **Rendimiento Consistente:** No se observó una mejora sustancial al aumentar el número de componentes. El uso de un número menor de componentes (por ejemplo, 5 o 3) es más eficiente, ya que mantiene un alto rendimiento mientras reduce la complejidad computacional.

Aplicación de K-Means Clustering

Se utilizó el algoritmo K-Means para segmentar las canciones en grupos basados en características musicales comunes. Tras aplicar K-Means con 5 clústeres, se identificaron patrones musicales específicos en cada grupo. Los resultados fueron los siguientes:

- **Clúster 0:** Canciones con alta danceability y energía moderada. Este grupo parece representar canciones aptas para bailar.

- **Clúster 1:** Canciones más tranquilas, con bajos niveles de energía y loudness, probablemente géneros como baladas o música ambiental.

- **Clúster 2:** Canciones con una valencia positiva elevada y alta energía, características comunes del pop y el dance.

- **Clúster 3:** Canciones con baja danceability y energía, posiblemente representando géneros más experimentales.

- **Clúster 4:** Canciones de energía moderada, con predominancia de influencia regional de Oceanía.

Evaluación del Modelo

El rendimiento del modelo Random Forest se evaluó utilizando métricas como la precisión, recall, F1-score y el RMSE. El modelo alcanzó una precisión general del 99%, lo que indica un buen desempeño en la clasificación de canciones no populares. Sin embargo, la clasificación de

canciones populares fue menos precisa, con un recall del 68%, lo que sugiere que el modelo podría beneficiarse de ajustes para mejorar la detección de canciones populares.

Conclusión

El desarrollo de un modelo predictivo para la popularidad de canciones, utilizando el conjunto de datos de Spotify Top 200, ha demostrado que Random Forest es una herramienta robusta y efectiva para la clasificación supervisada, especialmente en conjuntos de datos complejos y desbalanceados. La aplicación de PCA permitió reducir la dimensionalidad sin sacrificar precisión, mientras que el uso de K-Means ayudó a identificar patrones musicales clave. Este enfoque tiene un gran potencial en la industria musical, especialmente en plataformas de streaming, recomendaciones personalizadas y análisis de tendencias. A futuro, el modelo podría mejorarse para optimizar la clasificación de canciones populares y lograr un recall más alto.

Referencias Académicas

- **Breiman, L. (2001). "Random forests."** ***Machine Learning*****, 45(1), 5-32. DOI: 10.1023/A:1010933404324**
- **Jolliffe, I. T. (2002).** ***Principal Component Analysis*** **(2nd ed.). Springer Series in Statistics. ISBN: 978-0387954424**
- **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** ***The Elements of Statistical Learning: Data Mining, Inference, and Prediction*** **(2nd ed.). Springer. ISBN: 978-0387848570**