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

Justificación de la Selección del Clasificador:

Random Forest

Responsable: Badani Dávalos Kassandra Andrea

Asignatura: Inteligencia Artificial (INF-354)

Fecha: 9 de diciembre

La Paz – Bolivia

2024

Justificación de la Selección del Clasificador Random Forest

La selección del clasificador adecuado para un problema de clasificación depende de varios factores, incluyendo el tipo de datos, la naturaleza del problema, y las características del modelo. En este caso, la tarea es una clasificación supervisada, en la que se pretende predecir si una canción es "Popular" o "No Popular" basándose en diversas características musicales como Danceability, Energy, Loudness, y otras. En este contexto, el clasificador **Random Forest** se ha seleccionado por sus propiedades robustas y su capacidad para manejar datos complejos y desbalanceados, como los que contiene este conjunto de datos.

Random Forest en Clasificación Supervisada

Random Forest es un algoritmo de ensamblaje basado en múltiples árboles de decisión. Este clasificador ha demostrado ser eficaz en tareas de clasificación supervisada debido a su capacidad para manejar características no lineales y relaciones complejas entre variables. Según Breiman (2001), Random Forest construye múltiples árboles de decisión utilizando subconjuntos aleatorios de las características y los datos de entrenamiento, y luego promedia las predicciones de estos árboles para obtener una predicción final. Este enfoque ayuda a reducir el sobreajuste y mejora la capacidad de generalización del modelo, lo que es crucial cuando se trata de datos con variabilidad o ruido.

Una de las principales ventajas de Random Forest es su capacidad para manejar **datos desbalanceados**, como los presentes en este caso, donde la clase "Popular" tiene un número significativamente menor de instancias que la clase "No Popular". Aunque los datos están desbalanceados, Random Forest sigue siendo eficaz en la clasificación de ambas clases debido a su naturaleza de "votación" entre múltiples árboles. Esto permite que el modelo siga siendo preciso incluso en presencia de una distribución desigual de las clases (Breiman, 2001).

Relación de las Características con la Popularidad

En este conjunto de datos, los atributos como Danceability, Energy, y Loudness podrían tener un impacto significativo en la popularidad de las canciones. Por ejemplo, se observó que la categoría **Danceability** muestra una correlación moderada con los puntos totales de popularidad, lo que

sugiere que una mayor "bailabilidad" podría estar asociada con una mayor probabilidad de que la canción sea popular. Sin embargo, otras características, como Energy y Loudness, también pueden influir en la clasificación, aunque su relación con la popularidad no es tan fuerte. Random Forest tiene la capacidad de identificar y manejar estas relaciones complejas sin necesidad de una intervención explícita para la selección de características, lo que lo convierte en una herramienta poderosa para este tipo de análisis.

Ventajas de Random Forest

1. **Robustez ante el Sobreajuste**: Al utilizar múltiples árboles de decisión y combinar sus predicciones, Random Forest reduce el riesgo de sobreajuste, que es un problema común en los modelos que dependen de una única fuente de información. Esta característica es particularmente útil cuando se trabaja con datos ruidosos o con un alto número de características, como en el caso de las canciones musicales.

2. **Manejo de Datos Desbalanceados**: Como se mencionó anteriormente, Random Forest puede manejar de manera efectiva datos desbalanceados. Al crear múltiples árboles con subconjuntos aleatorios de los datos, el algoritmo minimiza el sesgo hacia las clases más frecuentes, logrando así una clasificación más equitativa entre las clases.

3. **Importancia de las características**: Otra ventaja de Random Forest es su capacidad para evaluar la **importancia de las características**. El modelo asigna una puntuación a cada característica en función de su capacidad para mejorar la precisión de la predicción. Esto permite identificar qué características, como Danceability y Energy, son más relevantes para la clasificación de la popularidad.

4. **No Requiere Preprocesamiento Extensivo**: A diferencia de otros clasificadores como las máquinas de soporte vectorial o la regresión logística, Random Forest no requiere que las características sean normalizadas o transformadas antes de ser utilizadas. Esto facilita su implementación y hace que sea adecuado para datos con diferentes escalas y tipos de variables.

Evaluación del Modelo

El rendimiento del clasificador Random Forest fue evaluado mediante una **matriz de confusión** y el **error cuadrático medio (RMSE)**. Los resultados mostraron que el modelo tiene una precisión general del 99%, con un rendimiento particularmente alto en la clasificación de las canciones "No Populares", aunque el **recall** para la clase "Popular" fue algo bajo (68%). Esto sugiere que, aunque el modelo es excelente para identificar canciones que no son populares, todavía puede mejorar en la correcta clasificación de las canciones populares.

Conclusión

El clasificador **Random Forest** ha demostrado ser una opción sólida y adecuada para abordar el problema de clasificación de la popularidad de canciones. Su capacidad para manejar datos desbalanceados, su robustez frente al sobreajuste y su habilidad para trabajar sin un preprocesamiento extensivo lo convierten en una herramienta poderosa para este tipo de análisis.

Fuentes Académicas

1. **Breiman, L. (2001).** "Random forests." *Machine Learning*, 45(1), 5-32. DOI: 10.1023/A:1010933404324

   En este artículo seminal, Breiman introduce el algoritmo Random Forest, explicando cómo su capacidad para combinar múltiples árboles de decisión mejora la precisión y reduce el sobreajuste.

2. **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. ISBN: 978-0387848570

   Este libro es una referencia fundamental en el campo del aprendizaje automático y proporciona una explicación detallada sobre los métodos de ensamblaje, incluido Random Forest, y cómo estos pueden aplicarse en problemas de clasificación.