Módulo 2 Análisis y Reporte sobre el desempeño del modelo.

José Andrés Orantes Guillén - A01174130

Resumen

En este documento se describe el proceso llevado a cabo para la creación de un modelo de Machine Learning utilizando un framework, con el objetivo de implementar una solución que forma parte del Portafolio de Implementación.

1. Introducción

Con el avance acelerado de las tecnologías, el desarrollo de modelos de *machine learning* se ha convertido en una herramienta crucial para resolver problemas complejos en diversas áreas, como la clasificación de datos. Durante el Módulo de *Machine Learning*, se exploraron diferentes tópicos y modelos, entre ellos el algoritmo *Random Forest*, conocido por su capacidad para realizar tareas tanto de regresión como de clasificación [1]. El objetivo de este documento es aplicar el modelo Random Forest para clasificar canciones en géneros musicales y analizar el desempeño de los modelos. A lo largo del documento, se describirá el proceso de selección de características, entrenamiento del modelo y evaluación de su desempeño.

2. Planteamiento del Problema

Se busca principalmente realizar el entrenamiento de dos modelos de *Machine Learning* y su evaluación. Para realizar esto, se plantea el uso de un conjunto de datos de 30,000 canciones de Spotify, el cual se explica con mayor detalle en la siguiente sección. Se determinará cómo entrenar de manera efectiva el modelo *Random Forest* utilizando un conjunto de características representativas de las canciones. Además se buscará de que forma se pueden obtener los mejores parámetros para la clasificación de canciones. La evaluación del desempeño del modelo se realizará considerando el grado de bias, varianza y nivel de ajuste del modelo. El objetivo del documento es identificar el rendimiento en la tarea de clasificación del modelo y realizar mejoras en base al análisis realizado. Esto permitirá entender mejor las fortalezas y limitaciones del algoritmo en el contexto específico de la clasificación de género de canciones.

3. Conjunto de Datos

Para realizar los modelos se tomará un conjunto de datos obtenido en Kaggle, el cual cuenta con 32,834 registros, y 23 columnas.

Variable	Traducción	Clase	
track$_id$	Id de la pista	character	
track$_name$	Nombre de la pista	character	
track$_artist$	Artista de la pista	character	
track$_popularity$	Popularidad de la pista	double	
track$_album_id$	Id del álbum	character	
track$_album_name$	Nombre del álbum de la pista	character	
track$_album_release_date$	Fecha de lanzamiento del álbum de la pista	character	
playlist$_name$	Nombre de la lista de reproducción	character	
playlist$_id$	Id de la lista de reproducción	character	
playlist$_genre$	Género de la lista de reproducción	character	
playlist$_subgenre$	Subgénero de la lista de reproducción	character	
danceability	Bailabilidad	double	
energy	Energía	double	
key	Clave	double	
loudness	ruido	double	
mode	escala	double	
speechiness	cantidad de palabras	double	Speechiness detecta la presencia de p
acousticness	Acusticidad	double	
instrumentalness	Instrumentalidad	double	
liveness	En vivo	double	
valence	Valencia	double	
tempo	tempo	double	
duration$_ms$	duración en milisegundos	double	

Cuadro 1: Tabla de descripciones de conjunto de datos

4. Limpieza y Exploración del Conjunto de Datos

Al revisar las columnas del conjunto de datos se eliminaron: track_id, track_name, track_artist, track_album_id, track_album_name, track_album_release_date, playlist_name, playlist_id y playlist_subgenre. Debido a que no parecen relevantes, sin embargo, se mantiene en observación track_album_release_date, pues podría ser que el año de estreno de una canción influya en el género en el que se desarrolló el artista. Con esas columnas se hizo el mapa de calor de la figura 1, el cual ayuda a mostrar correlación entre géneros y *features*. Para incluir los géneros se realizó *One-Hot Encoding* con las variables a predecir.

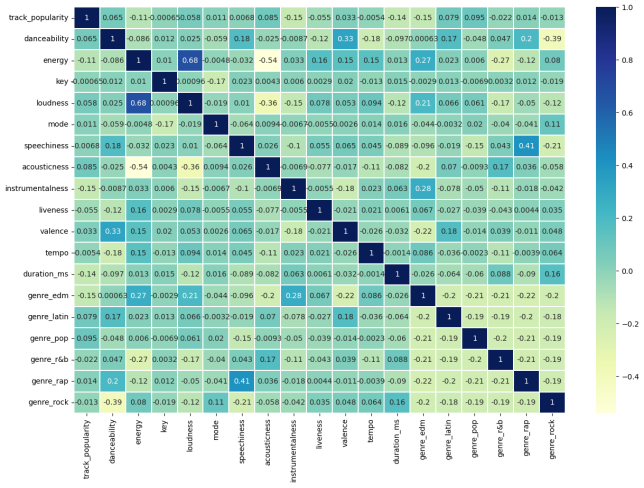

Figura 1: Mapa de Calor de Features y Labels de Géneros musicales

Las correlaciones más fuertes con el género EDM son energy, ruido e instrumentalidad. Para el género latino, las correlaciones más destacadas son bailabilidad y positividad. En el caso del género pop, se observa una correlación negativa con la cantidad de palabras utilizadas. Para el R&B, hay una correlación negativa con la

energía y el ruido, y una correlación positiva con la acusticidad. El rap muestra una correlación positiva con la bailabilidad y la cantidad de palabras. Finalmente, el rock presenta una correlación negativa con la bailabilidad y la cantidad de palabras utilizadas, y una correlación positiva con la duración de la canción. Además, se observa una correlación significativa entre el ruido y la energía de la canción, así como una correlación negativa entre la energía de la canción y la acusticidad.

Para conocer más sobre los datos se realizaron gráficas de caja y bigotes, además de obtener las medidas de tendencia central.

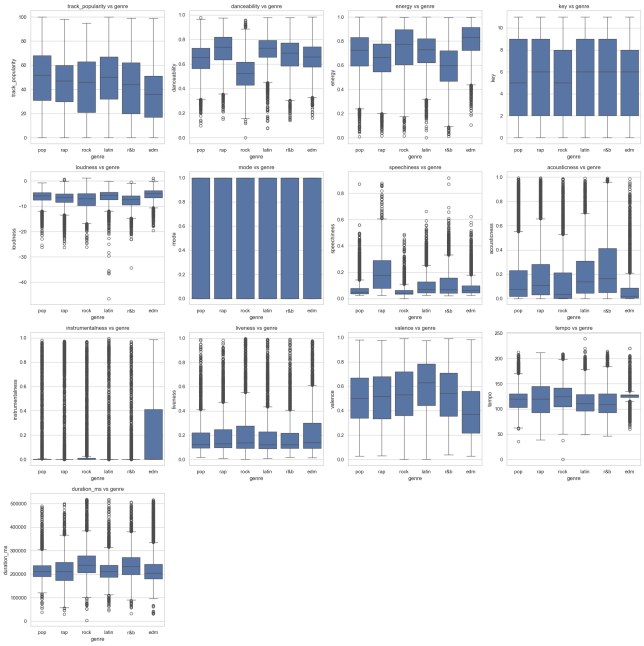

Figura 2: Gráficos de Caja y bigotes de cada *feature* contra los géneros

En conjunto de las medidas de tendencia y la gráfica 2 se obtuvieron **hallazgos** que se presentarán a continuación:

- **Popularidad** El análisis revela que el género EDM presenta el promedio de popularidad más bajo en comparación con los demás géneros, mientras que el género pop ostenta el mayor promedio de popularidad, seguido por el latino. En términos de popularidad máxima, el rock es el género con el valor más bajo.
- **Bailabilidad** Los géneros que destacan por su alta bailabilidad promedio son el rap y el latino. No obstante, los géneros con mayor bailabilidad máxima son el pop y el latino. El género rock, por otro lado, tiene la canción con la menor bailabilidad.
- **Energía** Las canciones del género EDM tienen el promedio de energía más alto, así como la menor desviación estándar, lo que indica una consistencia en la alta energía de este género. En contraste, el R&B presenta el promedio de energía más bajo. La canción con la mayor energía se encuentra en el género latino.
- **Clave** Las claves máximas y mínimas en las canciones se encuentran predominantemente en C y B, respectivamente. Además, la mediana de las claves se sitúa entre F y F/G.
- **Ruido** El género R&B muestra un promedio de ruido percibido relativamente bajo, mientras que el EDM tiene el promedio más alto. Los géneros con las canciones más ruidosas son el rock y el EDM, mientras que la canción menos ruidosa pertenece al género latino.
- **Escala** La mayoría de las canciones se encuentran en escala mayor, con el rock destacándose por tener la mayor cantidad de canciones en esta escala.
- **Locuacidad** El rap tiene la mayor cantidad promedio de palabras por canción. Sin embargo, la canción con la mayor cantidad de palabras se encuentra en el R&B. El género con la menor cantidad promedio de palabras es el rock, que también presenta una canción sin palabras y la canción con más palabras en este género tiene menos palabras que las canciones más extensas de otros géneros.

- **Acusticidad** El puntaje de acusticidad en general es bajo, lo que sugiere que hay poca confianza en la acusticidad de las canciones de todos los géneros.
- **Instrumentalidad** En general, todos los géneros incluyen canciones con letras, ya que el tercer cuartil de todas las categorías está por debajo de 0.5. Sin embargo, las canciones EDM suelen ser más instrumentales en comparación con otros géneros, donde la instrumentalidad es menos común.
- **En vivo** Las canciones en el conjunto de datos generalmente no son grabaciones en vivo.
- **Valencia** Las canciones del género latino tienen el promedio de valencia más alto, indicando una mayor sensación de positividad. En contraste, el género EDM tiene el promedio de valencia más bajo, sugiriendo una mayor tendencia hacia sonidos negativos. La canción más positiva pertenece al género rock.
- **Tempo** El género EDM presenta el mayor promedio de tempo, seguido por el rock. Se observa un valor atípico en el rock con un tempo de 0, asociado a una pista titulada "Hi, How are you doing?"de Dreams Come True, que será eliminada del análisis. El género con la canción de mayor tempo es el latino.
- **Duración en milisegundos** La duración promedio de las canciones varía entre 3 y 4 minutos. En promedio, las canciones del rock son las más largas, mientras que las del rap son las más cortas. La canción más larga es del género latino, seguida por el rock.

Tras encontrar estos datos se eliminó la variable "*liveness*", debido a que no se considera relevante para la predicción de datos.

5. Modelo

Se decidió hacer uso del algoritmo *Random Forest*, el cual está basado en árboles de decisión y usa votación o el promedio para decidir la predicción final.

5.1. Preprocesamiento

Primero, se realizó un escalado de los datos, ya que podría ser más sencillo para el algoritmo realizar su entrenamiento. Posteriormente, se dividieron los datos en conjuntos de entrenamiento, prueba y validación. La división se hizo con los siguientes porcentajes: 70 %, 20 % y 10 %, respectivamente.

5.2. Entrenamiento de Modelos

Con el objetivo de tener los mejores resultados posibles para el modelo se hizo uso de grid search, con híperparametros específicos. Se usó la siguiente configuración de *Grid Search*:

- n_estimators: 100, 200,500
- criterion: gini, entropy
- max_features: sqrt, log2
- max_depth: 10,25,50
- min_samples_split: 2, 5

Con esta configuración la mejor combinación de hiperparámetros obtenida fue:

- n_estimators: 500
- criterion: gini
- max_features: sqrt
- max_depth: 50
- min_samples_split: 5

5.3. Evaluación del Modelo

Debido a que se usan datos categóricos, estos se tienen que cambiar para hacer la evaluación de los modelos. A continuación se presentan los equivalentes numéricos de los diferentes géneros:

- rock: 0
- latin: 1
- pop: 2
- edm: 3
- r&b: 4
- rap: 5

En la tabla 6 se puede notar una comparación entre los puntajes obtenidos por los datos de Entrenamiento y validación.

	Exactitud	Puntaje F1
Entrenamiento	0.944	0.944
Validación	0.544	0.537
Prueba	0.564	0.558

Cuadro 2: Comparación de Métricos

Se puede notar que en los puntajes de Exactitud y Puntaje F1 de los datos de entrenamiento son mayores a los de validación y prueba.

	rock	latin	pop	edm	r&b	rap
rock	3406	3	23	0	12	1
latin	1	3331	118	43	39	53
pop	47	86	3590	80	75	27
edm	2	51	142	3936	32	34
r&b	10	55	113	23	3553	78
rap	16	56	15	19	28	3884

Cuadro 3: Matriz de confusión de Datos de entrenamiento

	rock	latin	pop	edm	r&b	rap
rock	366	7	40	12	40	5
latin	27	203	76	53	86	92
pop	86	53	182	85	77	38
edm	10	41	89	423	23	42
r&b	47	47	81	17	246	130
rap	22	45	25	33	68	367

Cuadro 4: Matriz de confusión de Datos de validación

	rock	latin	pop	edm	r&b	rap
rock	789	24	88	35	86	14
latin	47	432	168	101	123	162
pop	132	132	364	182	168	73
edm	39	68	156	856	39	60
r&b	85	71	120	34	498	223
rap	39	90	57	71	144	767

Cuadro 5: Matriz de confusión de Datos de Prueba

Las tablas 3, 4 y 5 son matrices de confusión de la clasificación en los datos de entrenamiento, validación y prueba, estas proporcionan otra forma de observar los datos. Se concluye que el modelo tiene sesgo medio y varianza alta, debido a las buenas clasificaciones que realiza en el entrenamiento, pero la calidad de las clasificaciones baja con los otros conjuntos de datos.

5.4. Refinamiento

Para realizar la mejora de los datos, los hiperparámetros de grid search se cambiarán y la cantidad de características será menor, puesto que se sospecha que la cantidad de características presentes hacen que el modelo sea muy específico. Se eliminarán las características acousticness, instrumentalness, key y mode. Estos se eliminan debido a que no se consideran importantes, por lo que su eliminación podría favorecer la generalización del modelo. La nueva configuración de *Grid Search* es la siguiente:

- n_estimators: 500, 600
- criterion: gini, entropy
- max_features: sqrt, log2
- max_depth: 10,50,75
- min_samples_split: 2, 5,10

Con esta configuración la mejor combinación de hiperparámetros obtenida fue:

- n_estimators: 500
- criterion: gini
- max_features: sqrt
- max_depth: 50
- min_samples_split: 10

En la tabla 6 se puede notar una comparación entre los puntajes obtenidos por los datos de Entrenamiento y validación.

	Exactitud	Puntaje F1
Entrenamiento	0.909	0.908
Validación	0.555	0.549
Prueba	0.567	0.561

Cuadro 6: Comparación de Métricos

Con los cambios realizados se pueden notar aumentos y cambios dentro de los datos, sin embargo, estos no son realmente significativos. Sigue siendo evidente el sesgo medio y la varianza alta en este modelo. Aunque no hay ni overfitting ni underfitting, siguen faltando cosas para un buen ajuste.

6. Conclusiones

Para tener mejores resultados, habiendo hecho cambios, se buscaría aumentar la complejidad del modelo. No se debe olvidar que se trata de un problema de clasificación de 6 categorías, por lo que se puede esperar el uso de un modelo complejo. También se puede intentar el uso de otro modelo que logre realizar la clasificación como una red neuronal. Además se puede considerar volver a tener las características que se tenían con la primera prueba y agregar la de año, pues podría ser que el año permita al algoritmo entender los patrones de comportamiento de los géneros, puede haber correlación entre año de lanzamiento y género. El aumento del grid search también podría ser una técnica que permita la mejora del modelo.

7. Referencias

[1] IBM. *What is random forest?* URL: https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.. (accessed: 31.08.2024).

8. Anexos

El código utilizado para el análisis se encuentra en el siguiente repositorio, de igual forma se pueden ver más datos estadísticos dentro del código: Repositorio del Portafolio de Implementación