|  |  |  |
| --- | --- | --- |
| **Nombre**: Ruben Alejandro Deambrossi | | **Matrícula**: 3012348 |
| **Nombre del curso:**  Aprendizaje Automático | **Nombre del profesor**:  Rolando Evelio Pérez Versón | |
| **Módulo**: 3 | **Actividad**: Evidencia 3 | |
| **Fecha**: 14-Agosto-2022 | | |
| **Bibliografía**: | | |

**Título**: Evidencia 3.

**Introducción**:

El siguiente reporte corresponde a la Evidencia 3 de Aprendizaje Automático.

**Desarrollo**:

1. Sistema de Recomendación de Canciones.

*Github Notebook*:

Pasos realizados:

1. Se importo un dataset con características de diversas canciones. Se analizo dicho dataset y se hizo un poco de clean up eliminando aquellos elementos que no tenían datos.
2. Se analizo el dataset y se determinaron las características que servirán para la clusterizacion del mismo. Se determino que características a evaluar serian las siguientes:

['danceability', 'energy', 'loudness', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo' ]

1. Se escalaron los datos antes de usar el modelo no supervisado.
2. Para el modelo no supervisado se escogió K-mean.
3. Para determinar el numero de cluster adecuado se utilizó el método de Silhouette, corriendo K-mean con diversos grupos y analizando el score.
4. Según el método de Silhouette los clusters óptimos serian 2 o 6, se procedió a la elección de 6 clusters.
5. Se entreno K-mean con 6 clusters.
6. Se creo un nuevo dataset el cual contenía las características de las canciones, el cluster label determinado por K-mean y el nombre de la canción.
7. Para el sistema de recomendación, se pide al usuario ingresar el nombre de una canción, luego se busca dicha canción en el dataset y se extrae las características y el cluster al que pertenece. Posteriormente se filtra todo el dataset para obtener las canciones que corresponden al mismo cluster de la canción del usuario. Se eliminan del dataset a la canción que corresponda a la seleccionada por el usuario. Se calcula la distancia euclidiana entre todos los puntos de datos del dataset a la canción del usuario. Se agrega dicha distancia al dataset como una nueva columna. Para finalizar se ordena el dataset resultante en base a la nueva columna de distancia de manera ascendente y se seleccionan las primeras 5 canciones que son las que tienen mayor similitud a la seleccionada por el usuario.

**Conclusiones**

El agrupamiento con K-mean funciona bastante bien, lo único confuso fue que el método de silhouette indicara que 2 grupos es lo mas optimo pero para fines prácticos de la evidencia se selecciono 6 grupos. La única limitación para este sistema es que la canción que seleccione el usuario debe estar dentro del dataset, tal vez una mejora seria extenderlo para que pudiera clasificar nuevas canciones en base a los grupos descubiertos por K-means. Para la recomendación de nuevas canciones, básicamente se determinó la distancia mínima usando la distancia euclidiana para determinar que otras canciones están más próximas al punto de dato del usuario en el plano.