# Proyecto final – Sistema de recomendación de música – Inteligencia Artificial

**Integrantes:** Daniela Olarte, Carlos Jimmy Pantoja, Juan Esteban Caicedo y Carlos Andrés Tafurt.

**Abstract**

Thanks to commercial music streaming services that can be accessed through smart devices, digital music is now widely available compared to the previous analog era. Putting all this digital music together is time consuming and often leads to information fatigue. Therefore, it is useful to develop a music recommendation system that can predict, then automatically search for songs on Spotify to finally suggest them appropriately to users, based on the genre of the songs they have previously listened to and their audio characteristics. Being an unsupervised learning project, the K-Means algorithm was used to group the genres of the songs, and the cosine similarity distance was used to find the similarity between the audio characteristics and genres.

**Palabras claves**

Análisis exploratorio de datos – Segmentación – Recolección de datos – Procesamiento de datos – Limpieza de datos – Relevancia de la información – K-Means – Clustering – Distancia coseno – Hiperparámetros – Vectorización

## Introducción

Gracias a los servicios de transmisión de música comercial a los que se pueden acceder a través de dispositivos inteligentes, la música digital actualmente tiene una abundante disponibilidad en comparación con la era análoga anterior. Agrupar toda esta música digital requiere bastante tiempo y suele provocar fatiga de información. Por lo tanto, es de utilidad desarrollar un sistema de recomendación de música que logre predecir, luego buscar canciones automáticamente en Spotify para finalmente sugerirlas adecuadamente a los usuarios, en función del género, artista y la popularidad de las canciones que han escuchado anteriormente al igual que sus características de audio. Al ser un proyecto de aprendizaje no supervisado, se hizo el uso del algoritmo *K-means* para la agrupación de los géneros de las canciones, y el uso de la distancia de similitud coseno para buscar la similitud entre las características de audio y géneros.

Por consiguiente, las **preguntas de interés** son las siguientes:

- ¿Es posible identificar qué clase de canciones escuchadas por el usuario tienen características en común con otras canciones para así recomendarle más de este tipo?
- ¿Son la popularidad, el género y el artista de una canción, factores determinantes para una recomendación acertada de canciones similares?

Relación entre preguntas de negocio y soluciones planteadas:

Es posible responder a las preguntas de negocio a través de un modelo de clustering con K-means planteado como solución y la forma de recomendar las canciones implementada. En efecto, recomendar canciones de música a los usuarios teniendo en cuenta características en común entre estas canciones y factores determinantes como la popularidad, el género y el artista de las canciones hace posible el poder dejar un impacto positivo en los usuarios a la hora de recomendar canciones. Al agrupar las canciones escuchadas por el usuario con alta similitud y además utilizar vector media de la lista de canciones entrante, para luego encontrar la distancia coseno de cada canción en el cluster con dicho vector, garantiza que las canciones a recomendar sean las más parecidas a las escuchadas por el usuario. El acertar en el gusto del usuario facilita la tarea de captar la atención del mismo, permite lograr una mejor retención de usuarios, y asimismo posibilita aumentar el número de suscripciones de pago en la aplicación.

## Marco teórico

- **K-means Clustering Algorithm:** K-Means Clustering es un algoritmo de aprendizaje no supervisado que se utiliza para resolver los problemas de agrupación en el aprendizaje automático o la ciencia de datos.
- **CRISP-DM:** Se trata de un modelo estándar abierto del proceso que describe los enfoques comunes que utilizan los expertos en minería de datos.
- **Segmentación:** Refiere al acto de segmentar datos de acuerdo a las necesidades de su compañía, para así refinar sus análisis basados en un contexto definido, utilizando una herramienta para análisis de cálculos varios.

Características de audio de las canciones:

- **Tempo:** El tempo de la canción. El tempo general estimado de una pista en pulsaciones por minuto (BPM). En terminología musical, el tempo es la velocidad o ritmo de una pieza dada y se deriva directamente de la duración promedio del tiempo.
- **Energy**: La energía es una medida de 0,0 a 1,0 y representa una medida perceptiva de intensidad y actividad. Por lo general, las pistas enérgicas se sienten rápidas, fuertes y ruidosas. Cuanto mayor sea el valor, más enérgica será la canción.

- **Danceability:** la bailabilidad describe qué tan adecuada es una pista para bailar en función de una combinación de elementos musicales que incluyen tempo, estabilidad del ritmo, fuerza del ritmo y regularidad general. El valor varía de 0 a 1. Cuanto mayor sea el valor, más adecuada es la canción para bailar.
- **Loudness:** los valores de sonoridad se promedian en toda la pista. Es la calidad de una canción. Va desde -60 a 0 DB. Cuanto mayor sea el valor, más fuerte será la canción.
- **Valence:** Una medida de 0.0 a 1.0 que describe la positividad musical transmitida por una pista. Las pistas con una valencia alta suenan más positivas (p. ej., felices, alegres, eufóricas), mientras que las pistas con una valencia baja suenan más negativas (p. ej., tristes, deprimidas, enfadadas).
- **Liveness:** Detecta la presencia de una audiencia en la grabación. Los valores de vivacidad más altos representan una mayor probabilidad de que la pista se interprete en vivo. Un valor superior a 0,8 proporciona una gran probabilidad de que la pista esté activa.
- **Acousticness:** una medida de confianza de 0,0 a 1,0 de si la pista es acústica. 1.0 representa una alta confianza en que la pista es acústica.
- **Speechiness**: Speechiness detecta la presencia de palabras habladas en una pista. Cuanto más parecida a la voz sea la grabación (p. ej., programa de entrevistas, audiolibro, poesía), más cercano a 1,0 será el valor del atributo. Los valores superiores a 0,66 describen pistas que probablemente estén formadas en su totalidad por palabras habladas. Los valores entre 0,33 y 0,66 describen pistas que pueden contener tanto música como voz, ya sea en secciones o en capas, incluidos casos como la música rap. Los valores por debajo de 0,33 probablemente representen música y otras pistas que no sean de voz.
- **Mode:** Las canciones se pueden clasificar en mayores y menores. 1.0 representa el modo mayor y 0 representa el menor.
- **Key:** Clave es el tono, las notas o la escala de la canción que forma la base de una canción. 12 teclas van del 0 al 11.

Métricas de progreso:

Para medir el progreso y satisfacer nuestros objetivos indicados anteriormente, los criterios de progreso que tomaremos en cuenta son:

**Métrica 1:** Obtener un valor de K clusters para el algoritmo *K-means* que corresponda al 70% de los géneros más frecuentes y populares del dataset.
**Métrica 2:** Nivel de acierto óptimo de las canciones recomendadas por género, determinado a través de pruebas manuales que verifican la correctitud.

## Antecedentes

El dataset de referencia para nuestro proyecto se tomaron de una solución previa en Kaggle titulada Building Music Recommendation System using Spotify Dataset la cual fue publicada en el año 2021. Se escogió este dataset debido a que cuenta con la información adecuada para recomendar canciones a los usuarios, es decir, datos como el género de la canción, artista más escuchado, canción más escuchada, entre otras variables que pueden ser útiles para un sistema de recomendación de música.

El objetivo del proyecto encontrado en Kaggle se basa en poder visualizar procesos para entender datos utilizando EDA (Exploratory Data Analysis) para así poder seleccionar características que son relevantes para crear Data Analysis) para así poder seleccionar características que son relevantes para crear Sistemas de Recomendación. Para esto utiliza el método de Clustering K-Means donde agrupa los géneros de este conjunto de datos en diez clusters en función de las características numéricas de audio de cada género. Sin embargo, no se indica por qué decidieron realizar la agrupación con 10 clusters. Como resultado se obtiene que los géneros similares tienden a tener puntos de datos que se encuentran cerca unos de otros al igual que los tipos de canciones los cuales también se agrupan. Con base al análisis y las visualizaciones, los géneros similares suenan de manera similar y vienen de períodos de tiempo similares y lo mismo puede decirse de las canciones dentro de esos géneros. El sistema como tal toma los puntos de datos de las canciones que un usuario ha escuchado y recomienda canciones correspondientes a puntos de datos cercanos.

Tomando en cuenta la solución de este proyecto, pudimos encontrar similitudes tal como que se utiliza el algoritmo de K-Means para la realización del clustering. Asimismo, se encontraron diferencias como el hecho de que en nuestro proyecto se utilizó un protocolo de evaluación para determinar el mejor valor de k clusters para dicho algoritmo.

Por otro lado, encontramos otro proyecto en GitHub (ver link en referencias) sobre un sistema de recomendación de música un poco más avanzado que utiliza técnicas de vectorización y el API de Spotify para obtener cualquier canción para recomendar. Nos inspiramos en esta solución en cuanto a la vectorización, y pensamos utilizar para la entrega final el API de Spotify para recomendar con base a cualquier canción que exista, sin embargo, de momento las recomendaciones las hacemos con respecto a las canciones que se encuentran en el dataset principal de alrededor 170 mil canciones.

## Metodología

Para el planteamiento del plan de este proyecto, se utilizó la metodología de CRISP-DM la cual consta de 6 fases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation y Deployment. A continuación se podrá observar los objetivos de la misma

## 1.Determine Business Objectives:

### 1.1.Output: Background

- El área problemática de la empresa es el desarrollo comercial de música.
- Las recomendaciones de las pistas de música del cliente se basan actualmente en las tendencias recientes. También, cada canción que escucha el usuario tiene ligadas otras canciones de manera predeterminada y estas se van sugiriendo aleatoriamente.

**Ventajas:**

Menor gasto (contabilidad)

**Desventajas:**

Las recomendaciones actuales de música no van acorde al tipo de canciones que escuchan los usuarios.

### 1.2.Business Objectives

- Dejar un impacto positivo en los usuarios a la hora de recomendar canciones con el fin de que prefieran esta plataforma antes que a otras.
- Dada esta retención de usuarios, aumentar la suscripción de pago en la aplicación.

**Cuestiones comerciales:**

Mejorar la experiencia de usuario, mejorar el proceso de recomendación y reducir la tasa de abandono de los usuarios de la plataforma.

**Beneficios:**

Si se está conforme con el uso de la aplicación web debido a sus recomendaciones acertadas de música, los usuarios, gracias al voz a voz, pueden provocar un aumento sustancial de los mismos y como consecuencia del anterior beneficio, se produciría una expansión del catálogo de música, debido a que nuevos artistas pueden querer agregar sus canciones a la aplicación web al ver un mercado potencial en esta.

### 1.3.Business Success Criteria

Recomendar música con base en el consumo musical individual.

## 2.Assess Situation

### 2.1. Inventory of resources

- No se esperó usar un hardware específico puesto que el despliegue será monolítico.
- Los datos se almacenan en un dataset en forma de tabla. Se dispone del acceso directo a este dataset en la plataforma Kaggle.

### 2.2. Requirements, assumptions, and constraints

Se requiere que se finalice el desarrollo de la aplicación para que pueda ser utilizable en la Web. Se supone que el dataset ofrecido por Kaggle representa de manera suficiente las canciones, géneros y artistas que escuchan los usuarios. No se espera tener restricciones de acceso, legales o financieras durante el desarrollo del proyecto.

### 2.3. Risk and contingencies

- Existió un riesgo a nivel de programación debido a la poca experiencia que se tenía con las tecnologías aquí utilizadas. Así que se capacitó a los desarrolladores para cubrir esta problemática..
- Al principio se creyó tener un riesgo a nivel de datos si el dataset encontrado en Kaggle no hubiese sido suficiente para representar correctamente las canciones, sus géneros y sus artistas, sin embargo, no se tuvo inconveniente en este aspecto y entonces no fue necesario buscar nuevas fuentes de datos.
- Existió un riesgo a nivel de resultados si el sistema de recomendación no hacía acertadamente las recomendaciones, sin embargo este no fue el caso.

### 2.4. Terminology

Pre procesamiento de datos, limpieza de datos y entrenamiento de modelos.

### 2.5 Costs and benefits

No se esperaron costos en la recolección de datos, dado que los necesarios para este proyecto se tomaron de la plataforma de Kaggle. Se espera que este proyecto sea beneficioso para el cliente ya que a lo largo de su desarrollo se adquirió convicción en el

manejo de datos, se dio una mejor comprensión de ellos, y se logró llegar a un sistema de recomendación de música completamente funcional.

## 3.Determine Data Mining Goals

### 3.1. Data mining goals

El tipo de minería de datos que se utilizó es Segmentación.

### 3.2. Data mining success criteria

Se requirió que el proyecto fuera utilizable por lo que se tomó en cuenta el despliegue satisfactorio de los resultados del modelo como criterio

## 4. Produce Project Plan

### 4.1. Project plan

**a. Obtención y procesamiento de los datos**

Se utilizó un dataset de Kaggle para la visualización de datos, se realizó la limpieza de datos a las variables que lo requerían y se entrenó el modelo para poder empezar a hacer las respectivas segmentaciones.

**b. Construcción del modelo**

Para construir el modelo, primero tuvimos que estimar el mejor valor de k clusters para que el algoritmo K-Means lo utilice en su agrupación de géneros de canciones. Para esto, nuestro procedimiento fue vectorizar los géneros para ver cuáles son los más comunes, contarlos y determinar el número de géneros que consoliden más del 70% de las canciones del dataset principal. Hallamos que son 201 géneros que representan dicho porcentaje, así que este valor pasó a ser el valor de los k clusters que usó nuestro algoritmo de K-Means. En este, nuestros hiper parámetros fueron el método de inicialización *K-means*, el número de inicializaciones aleatorias que fue de 10, el número máximo de iteraciones para una sola ejecución que fue de 300, y la semilla aleatoria que fue de 42.

Luego, de entrenar el modelo, construimos la manera de recomendar las canciones. Para esto, definimos 4 funciones: una que se encarga de obtener los datos de una canción, otra que se encarga de obtener el vector media de una lista de canciones, otra que crea un diccionario, y por último otra que es la función principal que utiliza las anteriores y que se encarga de recomendar las canciones. En esta última, definimos que recomiende 10 canciones a partir de una lista de canciones pasada por parámetro y a partir del dataset. Sin embargo, también se puede configurar por parámetro si se desea menos o más canciones recomendadas. Para realizar esta recomendación, se predice el cluster al que pertenece el vector media de la lista de canciones de entrada, y se encuentra la distancia coseno de cada canción en el cluster con el vector media. En caso de que una canción recomendada sea igual a una canción de entrada en esta recomendación, se omite.

Por último, se conectó a nuestro sistema con un API de Spotify para encontrar canciones que no estén en el dataset y poder recomendar música a través de ellas.

**c. Diseño del test**

Realizamos 4 pruebas de recomendación: una con canciones de reggaetón, otra con canciones de pop, otra con canciones de rock y

otra con canciones de electrónica

**d. Evaluación del modelo**

En todas las pruebas, el sistema recomendó acertadamente canciones similares. Por lo tanto, podemos decir que:

- Fue posible identificar las canciones que más escucha el usuario por medio de las diferentes variables del dataset donde se determina, por ejemplo, cuál es su género favorito y las características de las canciones más oídas por este.
- Se pudieron agrupar canciones con el fin de encontrar características en común y lograr recomendarle al usuario canciones que acierten con el gusto del mismo.
- El género y las características de audio son factores determinantes a la hora de recomendar una canción.

**e. Despliegue**

Desplegamos el resultado final de nuestro sistema en una aplicación Web en localhost, usando el framework Flask de Python. El front-end lo desarrollamos con la librería Dash de Plotly, que traducía código Python a HTML. El back-end representa lo implementado en el notebook.

**4.2. Initial assessment of tools and techniques**

Se utilizó Kaggle como la plataforma para la obtención del dataset y Python 3.10 como el lenguaje de programación.

Código fuente:

Ver Jupyter Notebook.

**Análisis de resultados**

AED entrega 1:

Al trabajar con cuatro datasets diferentes pudimos encontrar correlaciones entre ellos. Una de las más altas fue la relación entre la popularidad y el año, al igual que la relación entre la popularidad y la acústica de la canción. Todos los resultados encontrados fueron graficados para una mejor visualización de los mismos, en el caso anterior, se utilizó un gráfico de correlación que nos ayudó a observar tanto las relaciones directamente proporcionales como inversamente proporcionales. También se observa en qué año se produjeron menos canciones o en cuál año se produjeron más. Se observan variables estáticas en algunos periodos de tiempo, como por ejemplo, la vivacidad que no varió entre 1920-2020. A su vez, se ve la importancia de la variación en algunas variables que en el pasado fueron muy relevantes pero que en el presente no tiene relevancia alguna como lo es la acústica que en 1990 era una variable muy importante a la hora de escuchar una canción y desde 1970 hasta 2020 tornó a ser una variable insignificante.

El análisis de los datos se llevó a cabo generalizando los modelos mediante el uso de la metodología CRISP-DM la cual nos permite llevar un proceso general al iniciar para poder abarcar la mayor cantidad de análisis de información y posteriormente plantarla en gráficas que permitan su visualización y análisis. Todo el proceso de exploración y análisis se llevó a cabo sin ninguna interrupción, por ende, no tiene ninguna falla.

En la literatura según el *Libro Guinness de los Récords*, la canción White Christmas es el sencillo más vendido de todos los tiempos, con más de 50 millones de copias en todo el mundo. De acuerdo con ello y nuestros resultados, podemos ver que nuestros resultados son verídicos y confiables debido a que coinciden con la información proporcionada por el libro anteriormente mencionado, este le suma validez a los dataset empleados en el proyecto.

El dataset también nos ofrece ver qué cantidad de canciones son explícitas, esto lo observamos en el gráfico de pastel de los resultados. En este se observó que la gran mayoría de canciones son no explícitas y solo el 8,5% son explícitas o no aptas para todo público. También vimos que hay un aumento de la popularidad de las canciones a partir de 1955.

En los gráficos *Correlation matrix of 'data_by_genres.csv, Top 10 popular genres with audio features* y *Correlation matrix of 'data_by_year.csv* se puede evidenciar que cada uno de ellos posee una relación inversamente proporcional y directamente proporcional, lo que quiere decir que cada dataset con el cual están generados estos gráficos tienen aproximadamente dos relaciones más relevantes que el resto: una inversamente proporcional y la otra directamente proporcional. Asimismo, en el gráfico *Audio features timeline* se evidencia lo que dijimos anteriormente: la acústica se vuelve un valor poco relevante para el éxito de un artista debido las persona dejaron de interesarse por esta variable años atrás, sin embargo, ahora unas de las variable que sí influyen en el éxito son la valencia, energía de las canciones y la bailabilidad de las mismas.

En cuanto a los outliers que se encontraron en el dataset, pudimos identificar por medio del uso de diagramas de Cajas y Bigotes que diferentes variables como la instrumentalidad, bailabilidad, vivabilidad y entre otros, tenían presencia de estos valores, sin embargo, al analizar la relevancia de estos outliers se llegó a la conclusión que estos datos no pueden ser removidos debido a que quitaría características importantes de la canción. Por ende, los datos atípicos de nuestro dataset se tomarán en cuenta ya que esta información varía de acuerdo a la canción.

Teniendo en consideración los resultados obtenidos se puede sacar provecho de ello para la implementación de nuestro modelo de negocio y así asegurar su efectividad. Por lo que para el desarrollo de la plataforma multimedia se buscará garantizar la óptima recomendación de canciones a los usuarios teniendo en cuenta sus gustos los cuales están almacenados como datos en un dataset de alta calidad.

AED entrega 2:

Se trabajó a final de cuentas con 2 datasets únicamente: uno principal que contiene información de canciones y otro a nivel de artistas que contiene toda la información de estos. Identificamos que en el dataset principal no había una columna de géneros para cada una de las canciones sino una lista de los artistas de cada canción. Sin embargo, el otro dataset sí contenía información de los géneros de los

artistas, por lo que pudimos corresponder estos géneros a cada canción del dataset principal a partir de sus artistas, y tener uno más consolidado.

## Conclusiones

- Se logró determinar que las preguntas de negocio se pueden responder mediante el sistema de analítica propuesto.
- Se logró crear, evaluar y desplegar un sistema de recomendación de música basado en los géneros, artistas y características de audio de las canciones.
- Fue posible identificar las canciones que más escuchadas el usuario por medio de las diferentes variables del dataset donde se determina, por ejemplo, cuál es su género favorito y las características de las canciones más oídas del usuario, logrando una óptima recomendación de canción.
- Se pueden agrupar canciones con el fin de encontrar características en común y lograr recomendarle al usuario canciones que acierten con el gusto del mismo.
- La popularidad, el género y el artista son factores determinantes a la hora de recomendar una canción.
- Con este sistema de recomendación de música, se podrá dejar en los usuarios un impacto positivo en su experiencia de uso de la aplicación gracias a un mayor disfrute de la misma, lo que puede implicar una mejor tasa de retención de usuarios, frente a otras aplicaciones.
- Dada la retención de usuarios, se podrá aumentar la probabilidad de decidir optar por la suscripción premium que pueda haber en la aplicación.

## Referencias


- Guinness world records, 2007 (Bantam ed). (2007). Bantam.
- Inteligencia artificial: glosario de términos. (2018, marzo 5). Interxion.com. https://www.interxion.com/es/blogs/2018/03/inteligencia-artificial-glosario-de-terminos
- Music Recommendation System using Spotify Dataset. (2021, december 17). Kaggle.com; Kaggle. https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset/notebook
- Madhav Thaker (2020, december). Spotify recommendation system. Recuperado de https://github.com/madhavthaker/spotify-recommendation-system