Entrega 1: Exploración de Datos **Inteligencia Artificial**

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

Abstract

Looking forward to developing a multimedia platform that recommends songs based on the tastes of the users themselves, we seek to determine what type of algorithm would be ideal for this task. Therefore, we will rely on different techniques and methods such as K-Means Clustering, the CRISP-DM method, and several more. In the data analysis, the most relevant thing was to find the correlations of the different datasets that we handled, this allowed us to reach coherent results in comparison with existing information on the Internet, such as the fact that there is a high correlation between the popularity and the year of launch of the song as well as the popularity and the acoustics of the songs. With what has been achieved, we can ensure that the platform will: improve the user experience, improve the recommendation process and reduce the rate of abandonment of platform users.

Palabras claves

Análisis exploratorio - Segmentación - Recolección de datos - Procesamiento de datos - Limpieza de datos - Relevancia de la información.

Introducción

Con el objetivo de recomendar canciones a los usuarios de una plataforma multimedia, tomando en cuenta sus gustos y los de otros usuarios, se busca implementar una aplicación que pueda cumplir con los objetivos anteriormente mencionados mediante el uso de técnicas de inteligencia artificial. Lo interesante de este proyecto se debe principalmente al hecho de poder satisfacer los gustos musicales de los usuarios, recomendando música que posiblemente les guste teniendo en consideración factores como lo son sus artistas más escuchados, su género de mayor interés, sus canciones más oídas, la popularidad de la canción, del artista y del género Al poder acercar al usuario con nuevas canciones de su interés se logrará una mayor probabilidad que el usuario continúe utilizando esta plataforma multimedia y la prefiera antes que a otras.

Preguntas de interés

- ¿De qué manera se puede recomendar una canción a un usuario, con base al género de la canción?
- ¿Qué algoritmo de segmentación es el más adecuado para un sistema de recomendación de música?
- ¿Se puede recomendar una canción a un usuario, con base a la popularidad de la canción, del género y del artista?

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.

Métricas de progreso

Métrica 1: % de documentación hecha
Métrica 2: % de backend realizado
Métrica 3: % de frontend realizado
Métrica 4:% de pruebas implementadas

Antecedentes

Los datos de referencia para el proyecto se tomarán de un dataset de Kaggle. 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.

Nombre: Building Music Recommendation System using Spotify Dataset
Año: 2021
Objetivo: 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.
Método: Clustering con K-Means: divide los géneros en este conjunto de datos en diez grupos en función de las características numéricas de audio de cada género.
Resultados: Los géneros similares tienden a tener puntos de datos que se encuentran cerca unos de otros, mientras que los tipos de canciones similares también se agrupan.
Similitudes: Se utiliza el algoritmo de K-Means para la resolución de problemas.
Diferencias: Hay un enfoque diferente de los datos.

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.

Se desea desarrollar una aplicación que permita hacer recomendaciones de canciones a los usuarios de una plataforma multimedia, tomando en cuenta sus gustos de este usuario y de otros usuarios.

Las recomendaciones de las pistas de música actualmente se basan 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:

No va acorde a los gustos de los usuarios, cuando se lanza una nueva canción se debe ligar a otras de manera manual y no se mantiene un registro de la música favorita de cada usuario.

1.2.Business Objectives

Mejorar las recomendaciones de canciones a los usuarios de una plataforma multimedia.

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 de música, los usuarios, gracias al voz a voz, provocarán 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 quisieran agregar sus canciones a la aplicación web al ver un mercado potencial en esta.

-

1.3.Business Success Criteria

- Recomendar música basada en el consumo musical individual.

2.Assess Situation

2.1. Inventory of resources

No se espera 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 el comportamiento de los usuarios. No se espera tener restricciones de acceso, legales o financieras durante el desarrollo del proyecto.

2.3. Risk and contingencies

- Existe un riesgo a nivel de programación debido a la poca experiencia que se tiene con las tecnologías aquí implementadas (Se capacitará a los desarrolladores para cubrir las problemáticas).
- Existe un riesgo a nivel de datos en caso de que los dataset ofrecidos por kaggle no sean suficientes y que no representen bien el comportamiento de los usuarios en cuanto a sus gustos (Se buscarán nuevas fuentes de datos).
- Existe un riesgo a nivel de resultados si el sistema de recomendación no hace bien su trabajo (Realización de pruebas al código con sus diferentes escenarios para llegar al problema y corregirlo).

2.4. Terminology

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

2.5 Costs and benefits

No se esperan costos en la recolección de datos dado que los datos necesarios para este proyecto se tomarán la plataforma de Kaggle, se espera que los costos operativos sean bajos solo incluyendo el mantenimiento del software y este proyecto puede ser beneficioso ya que a lo largo de su desarrollo se adquiere convicción en el manejo de datos y se da una mejor comprensión de ellos, lo que puede producir nuevas ideas donde se use la analítica.

3.Determine Data Mining Goals

3.1. Data mining goals

- El tipo de minería de datos que se va a utilizar es Segmentación.

3.2. Data mining success criteria

- Se requiere que el proyecto sea utilizable por lo que se toma en cuenta el despliegue satisfactorio de los resultados del modelo como criterio

1.Produce Project Plan

4.1. Project plan

a. Obtención y procesamiento de los datos

Se utilizará un dataset para la visualización de datos, se realizará la limpieza de datos a las variables que lo requieran (usualmente a variables continuas y discretas) y se entrenará el modelo para poder empezar a hacer las respectivas segmentaciones.

b. Construcción del modelo

Se describirá el comportamiento del modelo y su interpretación y se documentará conclusiones respecto a los patrones en los datos (si hay alguno).

c. Diseño del test

Se diseñarán técnicas de evaluación para modelos, se definirán criterios de evaluación y se diseñarán test de prueba para cada objetivo de minería de datos.

d. Evaluación del modelo

- Se ejecutarán los test y se evaluarán los resultados de acuerdo a los criterios.
- Se compararán los resultados de la evaluación con las expectativas del proyecto.
- Se listarán los resultados de acuerdo a criterios de éxito y evaluación y se seleccionarán los mejores.

- Se interpretarán los resultados en términos de negocio y se comprobará si la información es nueva, útil y fiable.
- Se analizará si se pueden realizar ajustes a las técnicas de modelado para poder llegar a mejores resultados.

e. Despliegue

- Desplegar los componentes de software en la Web.

4.2. Initial assessment of tools and techniques

Se utilizará Kaggle como la plataforma para la obtención del dataset y Python 3.X como el lenguaje de programación.

Resultados

Ver Jupyter Notebook.

Análisis de resultados

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 que cantidad de canciones son explicitas, 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, a partir de esta fecha tenemos muchos más datos para poder trabajar en la recomendación de canciones.

En los gráficos 9, 10 y 11 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 12 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.

Hablando un poco de outliers (valores atípicos) 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.

Conclusiones y trabajo futuro

En este informe, se realizó un análisis visual para comprender mejor los datos de un dataset enfocado a la recomendación de música. Se aprendió principalmente como hacer un análisis exploratorio y de qué se que se compone el mismo. Así como también a identificar la relevancia de los datos para hacer uso de los mismos u optar por remover ciertos datos. Este trabajo fue de mucho aprendizaje en cuanto al uso y la importancia de las técnicas de análisis exploratorio para entender los datos que se van a tratar posteriormente en el modelo de machine learning. Los siguientes pasos que se tomarán en el proyecto estarán enfocados a la realización de los protocolos de evaluación, al entrenamiento de los modelos, los resultados obtenidos y el plan de despliegue.

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, diciembre 17). Kaggle.com; Kaggle. https://www.kaggle.com/code/vatsalmavani/music-recommendation-system-using-spotify-dataset/notebook