

TRABAJO FIN DE MÁSTER
MÁSTER EN BIG DATA Y BUSINESS ANALYTICS



IMF Smart Educación

Análisis y desarrollo de modelos de aprendizaje
automático para predecir el éxito de canciones en
Spotify

Autor: D. Sergio Martínez Martínez

Tutor: D. Juan Manuel Moreno Lamparero

Murcia, mayo de 2023

Índice

Resumen

Abstract

1. Introducción

1.1.- Motivación

1.2.- Definición

2. Objetivos propuestos (generales y específicos)

3. Estudio del arte

3.1.- Big Data en el sector de la música

3.2.- Relación con proyectos previos

3.3.- Estudio de Viabilidad

3.3.1.- Alcance del proyecto

3.3.2.- Estudio de la situación actual

3.3.3.- Estudio y valoración de las alternativas de solución

3.3.4.- Selección de la solución

4. Tecnologías y herramientas utilizadas en el proyecto

4.1.- Herramientas de desarrollo

4.2.- Herramientas de gestión de proyecto.

5. Metodologías usadas

5.1.- Mapa conceptual

6. Desarrollo del contenido del proyecto

6.1.- Explicación de la estructura general del proyecto

6.2.- Cuaderno-1: Limpieza y preparado de los datos

6.2.1.- Limpieza, depuración de ruido y enriquecimiento de datos

6.2.2.- Análisis exploratorio

6.2.3.- Guardado en la base de datos

6.3.- Cuaderno-2: Modelos supervisados

6.3.1.- Regresión Lineal

6.3.2.- Regresión Lineal Múltiple con ols

6.3.3.- Conversión de 'Visita_media_Artist' a variable categórica

6.3.4.- Regresión múltiple polinomial

6.3.5.- Validación Cruzada

6.3.6.- Pruebas finales.

6.3.7.- Pruebas con LassoCv y Ridge

6.4.- Cuaderno-3: Modelos no supervisados

6.5.- Power BI

7. Resultados y conclusiones

7.1.- Objetivos alcanzados y resultados obtenidos

7.2.- Resultados obtenidos

7.2.1.- Modelos supervisados

7.2.2.- Modelos no supervisados

7.2.3.- Power BI

7.3.- Conclusiones del trabajo y personale

7.4.- Vías futuras

8. Bibliografía

Índice de ilustraciones

- Ilustración 1: Características del Big Data
- Ilustración 2: Ejemplo CSV_1 TFG
- Ilustración 3: Ejemplo de soluciones TFG
- Ilustración 4: El Streaming reaviva la industria musical (Roa, 2023)
- Ilustración 5: Comparativa de Lenguajes de programación - Elaboración propia
- Ilustración 6: Comparativa de herramientas de visualización - Elaboración propia
- Ilustración 7: Comparativa de Bases de Datos - Elaboración propia
- Ilustración 8: Aplicaciones Anaconda Navigator
- Ilustración 9: Mapa Conceptual - Elaboración propia
- Ilustración 10: Distribución Jupyter
- Ilustración 11: Uso de percentiles - Elaboración propia
- Ilustración 12: Información general dataframe final
- Ilustración 13: Estudio variables numéricas
- Ilustración 14: Gráfica Streams-Artist
- Ilustración 15: Gráfica Streams-Track_Name
- Ilustración 16: Gráfica Streams-Position
- Ilustración 17: Gráfica Streams-Region
- Ilustración 18: Gráfica Streams-Visita_media|vistas_media_Artist
- Ilustración 19: Resultados OLS
- Ilustración 20: Número óptimo de clusters
- Ilustración 21: Cluster Visitas_media_Artist | Streams
- Ilustración 22: Columna "Genre" sin simplificar
- Ilustración 23: Gráficas Géneros | Streams (Sin simplificar)
- Ilustración 24: Gráfica de Géneros simplificados | Mes
- Ilustración 25: Gráfica Género simplificado | Mes | España
- Ilustración 26: Clustering sobre Géneros simplificados.
- Ilustración 27: Página inicial Power BI
- Ilustración 28: Power BI enero 2017
- Ilustración 29: Power BI 2018
- Ilustración 30: Página Custom Power BI
- Ilustración 31: Gráfica géneros- Región Global
- Ilustración 32: Gráfica de géneros Italia
- Ilustración 33: Clustering resultados
- Ilustración 34: Pop - Género más popular
- Ilustración 35: Géneros más populares en España

Índice de tablas

Tabla 1: Correlación entre variables numéricas
Tabla 2: Tabla VIF
Tabla 3: Regresión Lineal - Prueba 1
Tabla 4: Regresión Lineal - Prueba 2
Tabla 5: Regresión Lineal - Prueba 3
Tabla 6: Regresión Lineal Múltiple OLS
Tabla 7: Valores de 'Visita_media_Artist'
Tabla 8: Regresión Lineal – Con 'Visita_media_Artist' categórica
Tabla 9: Regresión Lineal Múltiple OLS – Con 'Visita_media_Artist' categórica
Tabla 10: Regresión múltiple polinomial - Prueba 1- Con variable categórica
Tabla 11: Regresión múltiple polinomial - Prueba 2- Sin variable categórica
Tabla 12: Regresión múltiple polinomial - Prueba de diferentes grados de polinomio
Tabla 13: Regresión múltiple polinomial - Validación cruzada 1
Tabla 14: Regresión múltiple polinomial - Validación cruzada 2
Tabla 15: Regresión múltiple polinomial - Validación cruzada - Variables numéricas normalizadas y Lasso
Tabla 16: Prueba de grados - Modelo final – Lasso
Tabla 17: Prueba con grado 4- Modelo final – Lasso
Tabla 18: Modelo final con LassoCv
Tabla 19: Modelo final con Ridge
Tabla 20: Objetivos finales completados
Tabla 21: Resumen de los resultados - Modelos supervisados

RESUMEN

La finalidad de este proyecto consiste en la utilización de una serie de técnicas de aprendizaje automático relacionadas con Data Science sobre un conjunto de datos de Spotify. En concreto se realizará en primer lugar un proceso de preparación, procesado y análisis de datos, seguido del uso de diferentes modelos, tanto supervisados como no supervisados, sobre nuestro conjunto de datos para más tarde finalizar con una parte visual destinada al usuario final con una serie de gráficas que ayuden a comprender nuestros datos.

Para ello, contaremos con un archivo CSV que contiene las 200 canciones más escuchadas de 53 países en los años 2017 y comienzos del 2018. Sobre este archivo se realizarán numerosas actualizaciones, las cuales iremos guardando conforme sea necesario en una base de datos que tendremos a nivel local.

Se comenzará por hacer una preparación y limpieza del CSV, seguida de su almacenamiento en una base de datos. Todo esto con el objetivo de usar dichos datos para la elaboración de varios modelos de predicción, tanto supervisados como no supervisados.

Para los modelos supervisados su finalidad será predecir el número de Streams por canción, desarrollando para ello varios modelos con el fin de contrastarlos y elegir los que mejores resultados arrojen. En el caso de los modelos no supervisados su objetivo será poder agrupar y clasificar la información de nuestro CSV de forma que podamos entender mejor el comportamiento de los datos.

Finalmente, también habrá una parte completamente visual en el cual se mostrarán diferentes gráficas con las distintas modificaciones que se han ido realizando sobre nuestros datos base. Cabe resaltar que durante todo este proceso se guardará una copia en la nube la cual se puede encontrar en (Martínez, 2023)

ABSTRACT

The purpose of this project is to use a series of machine learning techniques related to Data Science on a Spotify dataset. Specifically, it will begin with a data preparation, processing, and analysis process, followed by the use of different supervised and unsupervised models on the dataset. The project will conclude with a visual component aimed at the end user, presenting a series of graphs to help understand the data.

To achieve this, we will have a CSV file with the top 200 songs from 53 countries in the years 2017 and early 2018. This file will undergo numerous updates across our work, which will be saved as necessary in a local database.

The project will start with the preparation and cleaning of the CSV file, followed by its storage in our database. The purpose of these steps is to use the data for the development of various prediction models of both types, supervised and unsupervised.

For the supervised models, the objective will be to predict the number of streams per song, developing multiple models to compare and select the best-performing one. In the case of unsupervised models, the goal will be to cluster and classify the information from our CSV in order to understand better the data's behavior.

Finally, there will also be a visual part, highlighting different graphs with the various modifications made to our base data. It should be noted that during this entire process a copy of our project will be saved in the cloud which can be found in (Martínez, 2023).

PARA MÁS INFORMACIÓN PUEDE ENCONTRARME EN:
FOR MORE INFORMATION, YOU CAN FIND ME AT:



LinkedIn:

<https://www.linkedin.com/in/sergio-mart%C3%ADnez-a94255269/>



Gmail:

sersergio65@gmail.com