Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 121-135
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos originales
Temática: Inteligencia artificial
Recibido: 07/05/2023 | Aceptado: 15/08/2023 | Publicado: 30/09/2023

Identificadores persistentes:
DOI: 10.48168/innosoft.s12.a110
ARK: ark:/42411/s12/a110
PURL: 42411/s12/a110

Aplicación de modelo de regresión lineal para predecir el índice de popularidad en la plataforma Spotify

Linear Regression application to predict the popularity index in Spotify

Cesar Vasquez Alvarez [1], Edith M. Coaquira Cuevas [2], Emerson D. Mendoza Hilasaca [3], Jeffrey J. Pinto Ñaupa [4]

[1] Universidad Nacional de San Agustín. cvasqueza@unsa.edu.pe
[2] Universidad Nacional de San Agustín. ecoaquiracu@unsa.edu.pe
[3] Universidad Nacional de San Agustín. emendozahi@unsa.edu.pe
[4] Universidad Nacional de San Agustín. jpinton@unsa.edu.pe

[*] Autor para correspondencia: cvasqueza@unsa.edu.pe

Resumen

En la actualidad los servicios de música en streaming se han convertido en uno de los principales medios de consumo de música alrededor del mundo. Spotify ofrece servicios de transmisión de música y abarca más de treinta millones de canciones. Cada año hay un incremento en la producción de canciones por lo cual es más difícil que una canción se establezca como un hit en el mercado. El presente trabajo tuvo como objetivo aplicar la técnica de modelado de Regresión Lineal para encontrar una tendencia del conjunto de datos sobre el índice de popularidad de las canciones en la plataforma Spotify, de esta manera predecir un resultado con nuevos datos que ingresen. Se aplicó una metodología cuantitativa basada en datos medibles que se tomaron como datasets. Como resultado se obtuvo un error cuadrático medio de 94.79 y una varianza de 0.20. La conclusión del trabajo es que el dataset utilizado no fue el ideal acorde a nuestro objetivo.

Palabras clave: Python, Regresión Lineal, Predicción

Abstract

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 121-135
https://revistas.ulasalle.edu.pe/innosoft

Currently, streaming music services have become one of the main means of music consumption around the world. Spotify offers music streaming services and covers more than thirty million songs. Every year there is an increase in the production of songs so it is more difficult for a song to establish itself as a hit in the market. The objective of this work was to apply the Linear Regression modeling technique to find a trend of the data set on the popularity index of songs on the Spotify platform, in this way predict a result with new data that enters. A quantitative methodology was applied based on measurable data that were taken as datasets. As a result, a mean square error of 94.79 and a variance of 0.20 were obtained. The conclusion of the work is that the dataset used was not the ideal according to our objective.

Keywords: *Python, Linear Regression, Predict*

Introducción

A lo largo del tiempo, la industria musical ha pasado por revoluciones tecnológicas y comunicativas. La industria musical está experimentando un fuerte crecimiento gracias al avance de Internet y el surgimiento de nuevas tecnologías de la comunicación, como son las redes sociales o diversas plataformas de contenido, como Spotify o YouTube [1]. La forma de consumo de música ha cambiado y en la actualidad, los servicios de música de streaming se han convertido en uno de los principales medios en los que se consume música alrededor de todo el mundo y su crecimiento no muestra señales de estancamiento. En la investigación [2] menciona los servicios más populares de plataformas de música, entre los cuales Spotify lidera el ranking seguido de sus principales competidores Apple Music y Amazon Music.

Smith et al. [3] mencionan a Spotify como una empresa de software el cual ofrece servicios de transmisión de música, Spotify es lanzado en el 2008 y en los siguientes 10 años tuvo un crecimiento continuo convirtiéndose en un icono para una nueva generación de organizaciones ágiles. Sus oficinas de desarrollo e investigación cuentan con una forma de trabajo y estructuras organizativas con un diseño que promueve la colaboración, innovación y autonomía.

El éxito de las canciones son regidas una serie de factores tal como lo presenta Interiano et al [4], hicieron el análisis de más 500000 canciones lanzadas en el Reino Unido entre los años 1985 y 2015 para ello usaron la técnica de Machine Learning “bosques aleatorios” para predecir el éxito de las canciones que le permitió cuantificar la contribución de las características puramente musicales en el éxito de las canciones. Entonces sugiriendo una escala temporal de la dinámica de la moda en la música exitosa. También mencionan el descubrimiento de varias tendencias multi décadas

que tuvieron relevancia para comprender la dinámica del éxito el cual define como ocupar puestos en las listas principales, correlacionar el éxito con las características acústicas y explorar la previsibilidad del éxito.

En el campo laboral de empresas de servicios multimedia descritas por Braga [5] tales como Spotify,Apple,Amazon o youtube music han tenido un crecimiento drástico en los últimos años principalmente debido a que el mundo se ha vuelto cada vez más digital con el tiempo, desarrollando asi tecnologías de rápido desarrollo y facil acceso, a lo que el autor también nos brinda un panorama de la constante competencia de estas empresas por liderar el mercado pero entre estas empresas está Spotify ya que destaca por ser una compañía pionera del rubro y de momento ninguna plataforma hace temblar su dominio.

En el caso de Lopes y Simoes [6] estan de acuerdo que industria de la música ha experimentado enormes cambios en relación con sus hábitos de producción, distribución y consumo debido al desarrollo exponencial de las plataformas de streaming, ellos a través de estas buscaron analizar qué factores influyen en la intención de adquirir un servicio de streaming de música y, en consecuencia, en su recomendación pues debido a la variedad de competencia viene a ser un algo complejo y multidimensional a esto también se cuenta el factor de las estrategias que emplea cada empresa para captar más clientes además de la implementación de algoritmos que buscan la satisfacción del cliente pues cada plataforma usan varios métodos que resuelvan estas dudas.

Para Twomey y Kroll [7], el método de Regresión Lineal, juega un papel importante en estudios que involucren datos cuantitativos, si bien el método no es ideal, el uso de varias técnicas pueden ayudar a superar sus debilidades. En adición Sravani y Bola [8] indican que regresion lineal es un método perteneciente a algoritmos de Machine Learning, se basa en aprendizaje supervisado, este método define la relación entre 2 variables (una dependiendo de la otra) ajustando la línea de regresión a los datos, por otro lado Regresión lineal Múltiple utiliza múltiples variables independientes unas de otras.

Este trabajo tiene como objetivo usar el algoritmo de Regresión Lineal para encontrar una tendencia del conjunto de datos sobre el éxito de canciones en Spotify y de esta manera poder predecir un resultado con nuevos datos que puedan entrar.

Materiales y métodos

Estado del Arte

Existen trabajos como los de Interiano et al [4] en la cual realizaron un análisis de una gran repertorio de músicas lanzadas los años 1985 y 2015 en el Reino Unido para comprender la dinámica del éxito, correlacionar el éxito con las características acústicas y explorar la previsibilidad del éxito. Haciendo uso de la técnica de Machine Learning "bosques aleatorios" teniendo en cuenta la función de sus características acústicas y agregando variables permitió cuantificar la contribución de las características puramente musicales en el éxito de las canciones y sugirió la escala temporal de la dinámica de la moda en la música popular.

Por otra parte Hérnandez y Beltrán [9] desarrollaron el modelo de dos redes neuronales que identifican las transiciones de las diferentes partes de la estructura de las piezas musicales y las diferencias entre las transiciones para etiquetarlas, hicieron uso de técnicas de aprendizaje profundo y aprendizaje automático con Pytorch sus resultados obtenidos fueron similares al estado del arte que tomaron como caso similar.

Asimismo Quin et al.[10] realizan una predicción de tendencias de géneros musicales según la influencia y la similitud de la música a través del análisis de datos, haciendo uso de métodos como Deepwalk, Cosine Similarity, modelo de similitud musical basado en el Análisis de Componentes Principales y la distancia Euclidiana. Los resultados que obtuvieron fueron que las tendencias del género musical se resumen y predicen a partir de la influencia y similitud de la música.

Se tomaron como referencia las investigaciones anteriores mencionadas porque se enfocan en un análisis de diferentes puntos de vista de la música, también aporta experiencia a nuestro trabajo de investigación la cual se centra en el análisis y predicción del índice de popularidad a partir de características o atributos de un tema musical.

Marco teórico

Inteligencia Artificial

La Inteligencia Artificial (IA) es una rama de las ciencias computacionales encargada de estudiar modelos de cómputo capaces de realizar actividades propias de los seres humanos en base a dos de sus características primordiales: el razonamiento y la conducta [11].

Data reduction

Es la transformación de información digital numérica o alfabética derivada empírica o experimentalmente en una forma corregida, ordenada y simplificada [12]. El propósito de la reducción de datos puede ser doble: reducir la cantidad de registros de datos mediante la eliminación de datos no válidos o producir datos de resumen y estadísticas en diferentes niveles de agregación para varias aplicaciones.

Data cleaning

Es el proceso de detectar, corregir o eliminar registros inexactos de un conjunto de registros, tabla o base de datos y se refiere a identificar partes incompletas, incorrectas, inexactas o irrelevantes de los datos y luego reemplazar, modificar, o eliminar los datos sucios o gruesos. Data cleaning se puede realizar de forma interactiva con herramientas de gestión de datos, o como procesamiento por lotes a través de secuencias de comandos o un firewall de calidad de datos [13].

Variables dummy

Permiten medir el efecto de una característica de determinados individuos en determinada muestra [14]. También son definidas como una variable utilizada para explicar valores cualitativos en un modelo de regresión, son variables las cuales suelen tomar valores binarios [15].

Herramientas

Regresión Lineal

Según la página MathWorks [16] la define como una técnica de modelado estadística la cual es empleada para describir una variable de respuesta continua como una función de una o varias variables predictoras. Asimismo ayudan a comprender y predecir el comportamiento de sistemas complejos, como también analizar datos experimentales, financieros y biológicos.

Por otra parte Gupta, García y Chin [17] mencionan sobre el amplio uso que tiene en la estimación estadística también sobre los beneficios que tiene su modelo lineal como son la simplicidad y facilidad de uso del método, asimismo recalcan que su principal inconveniente es su alta desviación del modelo lo cual hace que pueda producir resultados deficientes.

Python

Es un lenguaje de alto nivel de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código, se utiliza para desarrollar aplicaciones de todo tipo [18]. Es un lenguaje de programación multiparadigma, ya que soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma.

Google Colab

Es una herramienta que te permite ejecutar scripts de Python a través de los servidores de Google. Esto te permite ejecutar celdas de código como si se tratara de un cuaderno de Jupyter Notebook. Pero no sólo eso, Google Colab es perfecto para implementar algoritmos de aprendizaje máquina, ya que no te limitas a los recursos de tu computadora. Esto quiere decir que tienes a tu disposición el GPU y las TPU's de Google para potencializar el cómputo de tu proyecto [19].

Numpy

Es una biblioteca para el lenguaje de programación Python que da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas [20].

Pandas

Pandas es un paquete de Python para la ciencia de datos y Machine Learning, ofrece estructuras de datos poderosas, expresivas y flexibles que facilitan la manipulación y análisis de datos [21]. Es una biblioteca de código abierto que proporciona herramientas de análisis y manipulación de datos de alto rendimiento utilizando sus potentes estructuras de datos.

Matplotlib

Matplotlib es una biblioteca Python open source que permite crear visualizaciones de datos [22]. Esta biblioteca sirve para generar gráficas a partir de datos contenidos en listas, vectores, en el lenguaje de programación Python y en su extensión matemática NumPy.

Skelearn

La librería scikit-learn, también llamada sklearn, es un conjunto de rutinas escritas en Python para hacer análisis predictivo, que incluyen clasificadores, algoritmos de clusterización, etc. Está basada en NumPy, SciPy y matplotlib, de forma que es fácil aprovechar el código que usan estas librerías [23].

Seaborn

Es una librería de visualización de datos para Python desarrollada sobre matplotlib. Ofrece una interfaz de alto nivel para la creación de atractivas gráficas. Además, está íntimamente integrada con las estructuras de datos de pandas, lo que permite utilizar el nombre de los Data Frames y campos directamente como argumentos de las funciones de visualización [24].

Resultados y discusión

Tratamiento de datos

En procesos de aprendizaje automático,el tratamiento de datos o también llamado pre procesamiento tiene como objetivo manipular y/o transformar los datos sin procesar para evitar alterar el aprendizaje y futuras predicciones [25].

El dataset utilizado para este trabajo contiene inicialmente 2000 datos y 18 variables, entre los cuales, al momento de graficarlas se pudo apreciar que contenían valores anómalos los cuales claramente tenían que ser tratados.

Transformación de datos

Primero, importamos el dataset y las librerías necesarias.

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 121-135
https://revistas.ulasalle.edu.pe/innosoft

```
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
%matplotlib inline
from mpl_toolkits.mplot3d import Axes3D
from matplotlib import cm
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
from sklearn import linear_model
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score

from google.colab import drive
drive.mount('/content/drive')

data = pd.read_csv("/content/drive/My Drive/songs_normalize_new(2).csv")
```

Figura 1. Importación de librerías y dataset

Segundo, identificamos las variables adecuadas y necesarias para aplicar el modelo, las cuales fueron todas aquellas que contienen valores continuos, eliminando así las variables Nombre del Artista, Nombre de la canción, Explícito y Género.

```
data.drop(['artist','song','explicit','genre'],1)
```

Figura 2. Eliminación de variables irrelevantes

Tercero, identificamos las variables dependientes e independientes.

Variables Independientes:

x_1: Duración

x_2: Año

x_3: Bailabilidad

x_4: Energía

x_5: Tonalidad

x_6: Volumen

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 121-135
https://revistas.ulasalle.edu.pe/innosoft

x_7: Modalidad

x_8: Pronunciación

x_9: Acústica

x_{10}: Instrumentalidad

x_{11}: En vivo

x_{12}: Valencia

x_{13}: Tempo

Variables Dependientes:

y_1: Popularidad

Cuarto, graficamos las variables por medio de histogramas para buscar posibles valores anómalos.

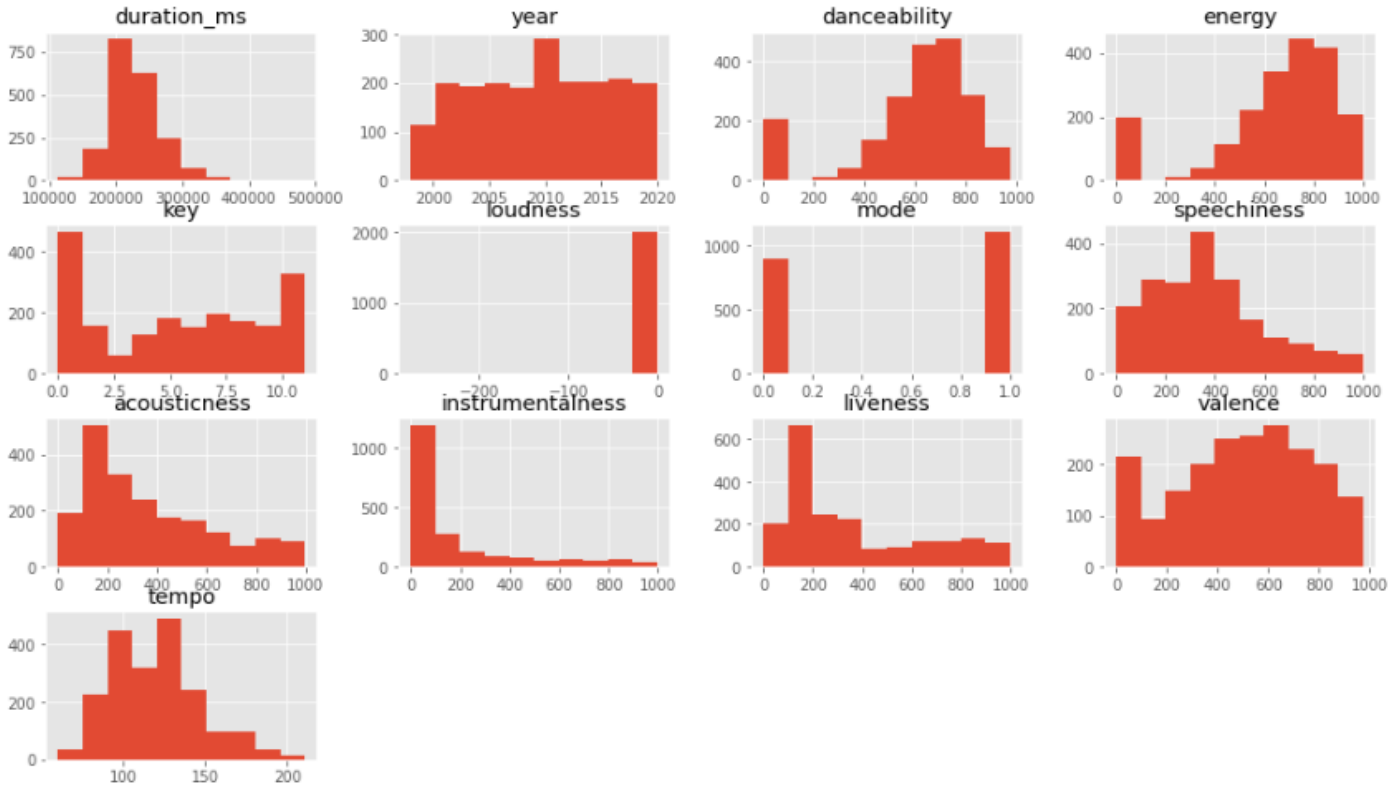

Figura 3. Histograma de variables independientes

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 121-135
https://revistas.ulasalle.edu.pe/innosoft

Quinto, procedemos a eliminar conjuntos de valores que estén alejados del grupo o iguales a cero, como en las variables Bailabilidad (danceability), Energía (energy), Tonalidad (key), Instrumentalidad (instrumentalness), como también a disminuir el rango de las variables Duración (duration_ms), Volumen (loudness) para tener una mejor distribución.

```
filtered_data = data[(data['duration_ms'] < 390000)  & (data['loudness'] > -15) & (data['instrumentalness'] >= 100)
& (data['danceability'] >= 180) & (data['energy'] >=180)& (data['key'] >=1.5)]
```

Figura 4. Data Cleaning y Data Reduction

364.56 ecm 0.04 varianza

Sexto, procedemos a crear y entrenar el modelo de Regresión Lineal.

```
#CREAMOS EL MODELO
model=linear_model.LinearRegression()
#ENTRENAMOS (predictoras, salida)
model.fit(XY_train,z_train)
#PREPARAMOS PARA LA PREDICCIÓN
z_pred=model.predict(XY_train)
```

Figura 5. Creación y entrenamiento del modelo

Séptimo, calculamos métricas para analizar el estado del modelo, Error Cuadrático Medio (MSE, por sus siglas en inglés) y la varianza.

```
# Error cuadrático medio cerca o =0
print("Error cuadrático medio: %.2f" % mean_squared_error(z_train, z_pred))
# Evaluamos el puntaje de varianza
print('Varianza: %.2f' % r2_score(z_train, z_pred))
```

```
Error cuadrático medio: 364.56
Varianza: 0.04
```

Figura 6. Métricas

Octavo, observamos que nuestro MSE es demasiado elevado por lo que decidimos agregar la variable género al modelo. Esta variable al ser de carácter nominal tuvo que ser tratada con Dummy Coding, generando así 11 columnas adicionales a las 13 ya tratadas.

```
# La columna genre es nominal (aplicamos dummy coding)
genre_dummy=pd.get_dummies(filtered_data["genre"],prefix="Género")
filtered_data=pd.concat([filtered_data,genre_dummy],axis=1)
```

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 121-135
https://revistas.ulasalle.edu.pe/innosoft

Figura 7. Dummy coding

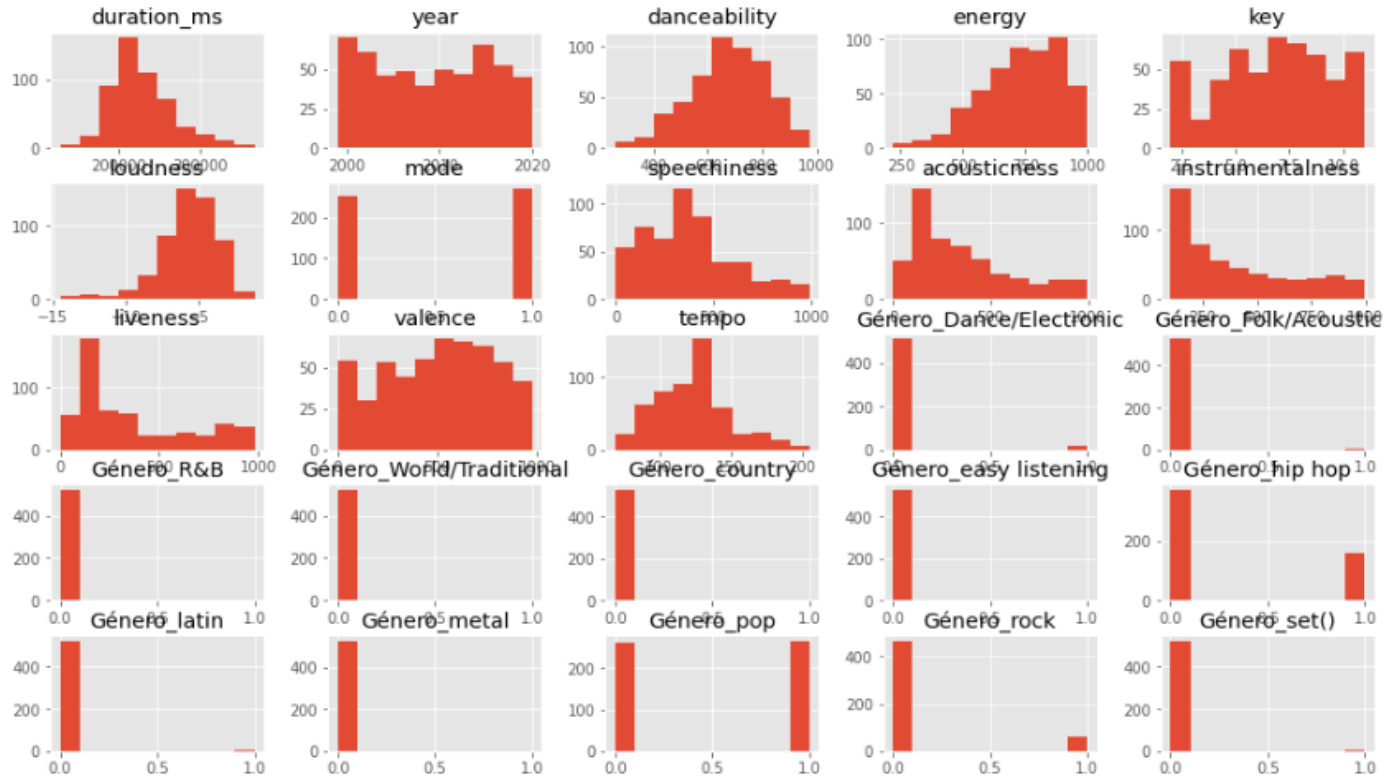

Figura 8. Histogramas del total columnas

Noveno, calculamos y revisamos las métricas nuevamente.

```
Error cuadrático medio: 347.80
Varianza: 0.08
```

Figura 9. Métricas con Dummy Coding

Observamos que se pudo disminuir el MSE, pero aún así sigue siendo demasiado alto.

Décimo, tras una serie de pruebas de datos se incluyó la variable Explícito aplicando un mapeo para transformar sus valores de Verdadero y Falso como string a valores booleanos. Además, en la variable tonalidad, sus valores iguales a 0 se igualaron a su media para evitar eliminarlos y disminuir aún más el número de datos.

```
data2["Explicito"]=list(map(lambda ele:ele=="VERDADERO",filtered_data["explicit"]))
```

Figura 10. Mapeo de la variable Explícito

Revista Innovación y Software
Vol. 4, No. 2, Mes Septiembre-Febrero, 2023
ISSN: 2708-0935
Pág. 121-135
https://revistas.ulasalle.edu.pe/innosoft

```
data2["Tonalidad"]=list(map(lambda x: 5.378000 if x==2 else x, filtered_data["key"]))
```

Figura 11.Cambio de valores de 0 a media

Obteniendo con estas nuevas variante, nuevas métricas

```
Error cuadrático medio: 94.79
Varianza: 0.20
```

Figura 11. Métricas finales

Consideramos estos valores aún muy altos pero si hacemos una comparación con las métricas iniciales se logró una reducción del 75% del MSE.

Finalmente, procedemos a realizar una predicción, que si bien no será exacta (por el valor del MSE) será mejor que la predicción que se pudo obtener con el modelo inicial.

```
z_Dosmil = model.predict([[300000, True,2020,0.7,0.1,8,-10,1,0.5,0,0.5,150,0,0,0,0,0,0,0,0,1,0]])
print(int(z_Dosmil))

-466
```

Figura 12. Predicción

Conclusiones

1. Regresión lineal es un modelo que se usa para predecir la dependencia entre dos o más variables continuas.
2. Para desarrollar un modelo de regresión lineal ideal se debe procurar tener un dataset con valores continuos y no tan dispersos. De no ser así, procurar tener más de 2000 datos para que al realizar data cleaning y/o data reduction no se tenga al final, por ejemplo un aproximado de 400 datos como en este trabajo.
3. En este trabajo, al aplicar dummy coding a variables nominales se pudo lograr una mejora en las métricas del modelo.
4. Consideramos que si se hubiera tratado la variable “Nombre del Artista” con algún método nuestro modelo mejoraría considerablemente.
5. Este conjunto de datos no es ideal para usar un modelo de Regresión Lineal.

Contribución de Autoría

Cesar Vasquez Alvarez: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Edith Coaquira Cuevas:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Emerson Mendoza Hilasaca:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Jeffrey Pinto Ñaupa:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición.

Referencias

[1]Gómez Herrero, R. (2021). Evolución de la Industria Musical. Siglo XX-Siglo XXI. UVaDOC Principal. https://uvadoc.uva.es/handle/10324/48012 [Accessed: June 22, 2022].

[2]García Pizarro, A. (2021). El auge de la música en streaming. UVaDOC Principal. https://uvadoc.uva.es/handle/10324/51809 [Accessed: June 22, 2022].

[3] D. Smite, N. B. Moe, G. Levinta and M. Floryan, "Spotify Guilds: How to Succeed With Knowledge Sharing in Large-Scale Agile Organizations," in IEEE Software, vol. 36, no. 2, pp. 51-57, March-April 2019, doi: 10.1109/MS.2018.2886178.

[4] Interiano, M., Kazemi, K., Wang, L., Yang, J., Yu, Z., & Komarova, N. L. (2018). Musical trends and predictability of success in contemporary songs in and out of the top charts. Royal Society Open Science, 5(5), 171274. doi:10.1098/rsos.171274

[5] M. M. Braga, "Spotify vs. Apple : a battle of titans", doctoral thesis, Universidade de Catolica Portuguesa, 2021.

[6] M. Lopes Barata y P. Simões Coelho, "Music streaming services: understanding the drivers of customer purchase and intention to recommend", ScienceDirect, Volume 7, Issue 8, agosto de 2021, art. n.º e07783.

[7] Golbaz, S., Nabizadeh, R., & Sajadi, H. S. (2019). Comparative study of predicting hospital solid waste generation using multiple linear regression and artificial intelligence. Journal of Environmental Health Science and Engineering, 17(1), 41-51.

[8] Sravani, B., & Bala, M. M. (2020, June). Prediction of student performance using linear regression. In 2020 International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE.

[9] Hernández Oliván, C., & Beltrán Blázquez, J. R. Análisis musical mediante inteligencia artificial.

[10] C. Qin, H. Yang, W. Liu, S. Ding and Y. Geng, "Music Genre Trend Prediction Based on Spatial-Temporal Music Influence and Euclidean Similarity," 2021 36th Youth Academic Annual Conference of Chinese Association of Automation (YAC), 2021, pp. 406-411, doi: 10.1109/YAC53711.2021.9486510.

[11] López Takeyas, B. (2007). Introducción a la inteligencia artificial. Instituto Tecnológico de Nuevo Laredo. http://itnuevolaredo.edu.mx/takeyas/Articulos/Inteligencia%20Artificial/ARTICULO%20Introduccion%20a%20la%20Inteligencia%20Artificial.pdf

[12] Contributors to Wikimedia projects. (2009, 29 de julio). Data reduction - Wikipedia. Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Data_reduction

[13] Contributors to Wikimedia projects. (2005, 31 de diciembre). Data cleansing - Wikipedia. Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Data_cleansing

[14] Alonso, J. C., & Muñoz, A. (2014). Interpretacion de variables Dummy en modelos log-lin. Cali, Colombia: Departamento de Economía, Universidad Icesi.

[15] Variable ficticia - Definición, qué es y concepto | Economipedia. Economipedia. https://economipedia.com/definiciones/variable-ficticia.html

[16] ¿Qué es la regresión lineal? MathWorks - Creadores de MATLAB y Simulink - MATLAB y Simulink - MATLAB & Simulink. https://la.mathworks.com/discovery/linear-regression.html (accedido el 13 de agosto de 2022).

[17] M. R. Gupta, E. K. Garcia and E. Chin, "Adaptive Local Linear Regression With Application to Printer Color Management," in IEEE Transactions on Image Processing, vol. 17, no. 6, pp. 936-945, June 2008, doi: 10.1109/TIP.2008.922429.

[18] Colaboradores de los proyectos Wikimedia. (2002, 13 de febrero). Python - Wikipedia, la enciclopedia libre. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/Python

[19] ¿Qué es Google Colaboratory? (s. f.). 330ohms. https://blog.330ohms.com/2021/08/10/que-es-google-colaboratory/

[20] Colaboradores de los proyectos Wikimedia. (2012, 4 de febrero). NumPy - Wikipedia, la enciclopedia libre. Wikipedia, la enciclopedia libre. https://es.wikipedia.org/wiki/NumPy

[21] Introducción a la Librería Pandas de Python. (s. f.). Aprende IA. https://aprendeia.com/introduccion-a-la-libreria-pandas-de-python-parte-1/

[22] Matplotlib: Funciones principales. Cursos de Programación de 0 a Experto © Garantizados. https://unipython.com/matplotlib-funciones-principales/

[23] Espacio de recursos de ciencia de datos. (s. f.). Espai de recursos de ciencia de dades. http://datascience.recursos.uoc.edu/es/preprocesamiento-de-datos-con-sklearn/

[24] Seaborn presentación. (s. f.). Interactive Chaos. https://interactivechaos.com/es/manual/tutorial-de-seaborn/presentacion

[25] Gao, J. (2012). Data preprocessing.