**Proyecto de evaluación Actividad 3**

Logotipo

Descripción generada automáticamente

**Estudiante: Juan Carlos González Ibarra**

**Correo:** [**jcarlos0284@gmail.com**](mailto:jcarlos0284@gmail.com)

**Fecha: 17 de mayo, 2024**

**Contenido**

1. Tema
   1. Planteamiento del problema
   2. Objetivos
2. Metodología de Desarrollo
3. Extracción de Datos

3.1 Fuentes de extracción de datos

3.2 Que es la extracción de datos

3.3 Extracción de datos con Python

1. Data Clean

4.1 Descripción de los datos

4.2 Carga de Datos

4.3 Preprocesar los datos

4.3.1 Fusionar data frames

4.3.2 Eliminar variables irrelevantes

4.4 Definición de la(s) Variable(s) Objetivo(s).

4.5 Almacenamiento de data frame en base de datos relacional

1. Modelado de Datos

5.1 Tipos de Modelos

5.2 Métodos de Modelado

5.3 Modelo No Supervisado

5.4 Modelo Supervisado

5.5 Proceso Modelado de Datos

5.6 Modelo de Clasificación

5.6.1 Resultados del Modelo de Clasificación

5.7 Modelo de Regresión

5.7.1 Entrenamiento del modelo

5.7.2 Validación del modelo

5.7.3 Resultados del Modelo de Regresión

5.8 Modelo de Regresión Logística

5.8.1 Entrenamiento del modelo

5.8.2 Validación del modelo

5.8.3 Resultados del Modelo de Regresión Logística

5.9 Predicción y toma de decisiones

5.9.1 Guardar el modelo

5.10 Aplicación Web de Predicción

5.10.1 Desarrollo

1. Análisis exploratorio de Datos (AED)

6.1 Análisis de Ventas Globales por año.

6.2 Análisis de Correlación de Ventas Globales por Consola y Genero

6.2.1 Relación entre Genero y Ventas Globales

6.2.2 información relevante

6.3 Análisis de Ventas Globales de Videojuegos por Plataforma

6.3.1 información relevante

6.4 Análisis de Ventas Globales por Género.

6.4.1 información relevante

6.5 Análisis de Ventas Globales por Compañía de Desarrollo.

6.5.1 información relevante

1. Resultados de AED
2. Conclusión
3. Bibliografía

**6. Análisis exploratorio de Datos (AED)**

En el análisis exploratorio se analizan los DataFrames con la información proporcionada en la sección anterior, el propósito es analizar e investigar los conjuntos de datos, resumir sus características principales, identificar relaciones, entender la distribución de la(s) variables(s) objetivo(s) y las variables para conseguir el objetivo del proyecto.

Se define a realizar lo siguiente:

* Análisis de Ventas Globales por año.
* Análisis de Correlación de Ventas Globales por Consola y Genero.
* Análisis de Ventas Globales de Videojuegos por Plataforma.
* Análisis de Ventas Globales por Género.
* Análisis de Ventas Globales por Compañía de Desarrollo.

**6.1 Análisis de Ventas Globales por año.**

En este análisis, se examina la distribución de las ventas de videojuegos a nivel global a lo largo de los años utilizando el atributo 'Fecha\_Lanzamiento' que contiene la información de la fecha en que se lanzó a la venta del videojuego y el atributo 'Ventas\_Globales' que contiene la información del total de ventas globales del videojuego, por lo que la relación de las ventas globales con el año de lanzamiento se visualiza en un gráfico lineal.

El objetivo es identificar patrones o tendencias en las ventas a lo largo del tiempo, como picos o declives en ciertos años. Esto puede ayudar a comprender la evolución de la industria de los videojuegos y tomar decisiones estratégicas basadas en las ventas históricas.

# Ventas globales por año

df\_juegos\_validos = df\_juegos[df\_juegos['Fecha\_Lanzamiento'] != 0]

ventas\_por\_año = df\_juegos\_validos.groupby('Fecha\_Lanzamiento')['Ventas\_Globales'].sum()

# Gráfico de líneas de ventas globales por año

plt.figure(figsize=(10,6))

ventas\_por\_año.plot(kind='line', linewidth=2, color='blue')  # La línea es ahora azul y más gruesa

# Obtén el año y las ventas del punto máximo

max\_year = ventas\_por\_año.idxmax()

max\_ventas = ventas\_por\_año[max\_year]

# Grafica el punto máximo

plt.scatter([max\_year], [max\_ventas], color='red')  # Dibuja un punto rojo en el valor máximo

plt.title('Ventas Globales por Año')

plt.xlabel('Año de Lanzamiento')

plt.ylabel('Ventas Globales (en millones)')

plt.xlim([1980, 2017])  # Esto define el rango del eje X

plt.xticks(range(1980, 2018, 1), rotation=90)  # Esto establece los ticks de 1980 a 2017, de uno en uno

plt.grid(True)

plt.show()

Gráfico, Gráfico de líneas

Descripción generada automáticamente

**6.1.1 información relevante**

* El promedio de ventas globales tiene un promedio de 500000 mil copias en el lapso de 37 años.
* Los picos más altos de ventas Globales están entre 2007 y 2008.
* La tendencia de ventas incremento a partir del año 2000, pero a partir de 2015 las ventas bajaron.

**6.2 Análisis de Correlación de Ventas Globales por Consola y Genero**

Se puede explorar las correlaciones entre los diferentes atributos que se tienen en el DataFrame df\_juegos para ayudar a entender mejor sus relaciones y seleccionar las características para el modelado de datos.

La matriz de correlación es una tabla que muestra el coeficiente de correlación (también conocido como Pearson's r) entre pares de variables en un conjunto de datos. Cada celda de la tabla muestra la correlación entre dos variables.

Las características de la matriz de correlación son:

* Coeficiente de correlación que es una medida estadística que describe el grado de relación entre dos variables.
* El coeficiente de correlación puede variar de -1 a +1.
* Una correlación de +1 indica una fuerte correlación positiva: esto significa que las dos variables tienden a aumentar juntas.
* Una correlación de -1 indica una fuerte correlación negativa: esto significa que a medida que una variable aumenta, la otra disminuye.
* Una correlación de 0 indica que no hay una relación lineal entre las variables.

# Seleccionamos solo las columnas numéricas

numeric\_df = df\_juegos.select\_dtypes(include=[np.number])

# Ahora calculamos la matriz de correlación solo con las columnas numéricas

correlation\_matrix = numeric\_df.corr()

# Suponiendo que 'df' es tu DataFrame y 'correlation\_matrix' es la matriz de correlación que has calculado

plt.figure(figsize=(15,10))

sns.heatmap(correlation\_matrix, annot=True)

plt.title('Correlation matrix of DataFrame')

plt.xticks(rotation=90)

plt.yticks(rotation=0)

plt.show()

Imagen que contiene Gráfico

Descripción generada automáticamente

Al visualizar la matriz de correlación con el mapa de calor, se puede obtener una vista de cómo están correlacionadas las variables en el conjunto de datos del DataFrame df\_juegos.

En el mapa de calor:

* Las celdas coloreadas representan las correlaciones entre las variables.
* El color de cada celda se determina según el valor de correlación y el mapa de color que se utiliza.
* Las correlaciones fuertes (cercanas a 1 o -1) se indican con colores más intensos, mientras que las correlaciones débiles (cercanas a 0) se indican con colores más claros.

Se visualiza una fuerte correlación positiva entre las 'Ventas\_Globales', 'Ranking' y la 'Puntuacion\_Reseña', por lo que se puede inferir que los videojuegos con puntuaciones de reseñas más altas tienden a tener más ventas globales. Esto sugiere que la calidad de un juego (según lo medido por las reseñas) puede influir en su éxito comercial. Sin embargo, la correlación no implica causalidad y se va necesita más investigación para entender la relación entre estas variables.

**6.2.1 Relación entre Genero y Ventas Globales**

Para profundizar en el análisis de correlación, se va a graficar las relaciones entre tres variables: 'Plataforma', 'Ventas\_Globales', y 'Genero'. La variable 'Plataforma' se muestra a lo largo del eje x, la variable 'Ventas\_Globales' se muestra a lo largo del eje y, y la variable 'Genero' se representa mediante el color de los puntos. Por lo tanto, cada punto en el gráfico representa un juego, y su posición en el eje x y y indica la plataforma en la que se lanzó el juego y las ventas Globales que tuvo. El color del punto indica el género del juego.

Se observa en cada punto del gráfico cómo las ventas en Globales varían según la plataforma y el género.

# Ajustar el tamaño de la figura

plt.figure(figsize=(20,8))

# Crear un scatterplot 2D donde la tercera variable ('Genero') se representa mediante el color de los puntos

sns.scatterplot(data=df\_juegos, x='Genero', y='Ventas\_Globales', hue='Plataforma')

plt.title('Relacion entre Genero y Ventas Globales')

plt.xlabel('Genero')

plt.ylabel('Ventas de Globales (en millones)')

# Rotar las etiquetas del eje x a 45 grados para una mejor legibilidad

plt.xticks(rotation=90, ha='right')

plt.yticks(rotation=0)

plt.grid(True)

plt.show()

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

**6.2.2 información relevante**

Los puntos de un color específico (videojuegos de un cierto género) están concentrados en una cierta área del gráfico, sugiere que el género tiende a tener más ventas en determinadas plataformas.

El promedio de ventas globales de cada genero están por debajo de los 10 millones de copias, por lo tanto, es de sugerir hacer una análisis de los promedios de ventas globales por genero y/o plataforma.

Los géneros que más venden son el de Accion, Racing, Platform, Role Playing y Shooter.

Las plataformas que más venden por genero son:

* Accion (Nintendo, Play Station y Xbox).
* Racing (Nintendo, Play Station y Xbox).
* Accion (Nintendo, Play Station y Xbox).
* RPG (Nintendo, Play Station y Xbox).
* Shooter (Play Station y Xbox).

**6.3 Análisis de Ventas Globales de Videojuegos por Plataforma**

Resultado del análisis de correlación anterior, se puede observar que la relación entre las 'Ventas\_Globales', 'Genero' y 'Plataforma' está por debajo de 1 millón de copias, por lo tanto, se va a realizar un análisis de la media de 'Ventas\_Globales.

En este análisis, se examina la distribución de las ventas de videojuegos según la plataforma en la que se juegan, utilizando el atributo 'Plataforma' que contiene la información de la plataforma en la que se ejecuta el videojuego y el atributo 'Ventas\_Globales' que contiene la información del total de ventas globales del videojuego, por lo que la relación puede calcular el total de ventas por Plataforma y visualizar esta información en un gráfico lineal marcando cada punto máximo en el atributo 'Plataforma' de los videojuegos más populares o exitosos en términos de ventas. Esto puede ser útil para comprender el mercado de consolas y tomar decisiones sobre el desarrollo y la distribución de juegos específicos para ciertas plataformas.

Se va a calcular y obtener la media de Ventas Globales por cada Plataforma. Se grafica las ventas globales por plataforma del mayor al menor.

# Calcular la media de 'Ventas\_Globales' para cada 'Plataforma'

media\_ventas\_por\_plataforma = df\_juegos.groupby('Plataforma')['Ventas\_Globales'].mean()

# Calcular la media de 'Ventas\_Globales' para cada 'Genero'

media\_ventas\_por\_genero = df\_juegos.groupby('Genero')['Ventas\_Globales'].mean()

# Ordenar media\_ventas\_por\_plataforma del promedio más alto al más bajo

media\_ventas\_por\_plataforma\_sorted = media\_ventas\_por\_plataforma.sort\_values(ascending=False)

# Crear un gráfico de barras para media\_ventas\_por\_plataforma\_sorted

media\_ventas\_por\_plataforma\_sorted.plot(kind='bar', figsize=(12, 6))

plt.title('Promedio de Ventas Globales por Plataforma')

plt.ylabel('Promedio de Ventas Globales')

plt.xlabel('Plataforma')

plt.xticks(rotation=90)

plt.grid(True)

plt.show()

Gráfico, Histograma

Descripción generada automáticamente

# Calcular el promedio total de ventas globales

promedio\_total\_ventas\_globales = df\_juegos['Ventas\_Globales'].mean()\*1000000

# Imprimir el resultado

print(f"El promedio total de ventas globales para todas las plataformas es: {promedio\_total\_ventas\_globales}")

El promedio total de ventas globales para todas las plataformas es: 548394067325.8527

**6.3.1 información relevante**:

* El promedio de ventas globales por plataforma está por debajo de 1 millón de copias.
* El promedio total de ventas globales para todas las plataformas es de 548394 millones de copias.
* Las plataformas que más venden son de Nintendo, Play Station y Xbox.

**6.4 Análisis de Ventas Globales por Género**.

En este análisis, se examina la distribución de las ventas de videojuegos a nivel global utilizando el atributo 'Genero' que contiene la información del género del videojuego y el atributo 'Ventas\_Globales' que contiene la información del total de ventas globales del videojuego, por lo que la relación puede calcular el total de ventas por género y visualizar esta información en un gráfico lineal marcando cada punto máximo en el atributo 'Genero' de videojuegos más populares o exitosos en términos de ventas. Esto puede ser útil para tomar decisiones relacionadas con el desarrollo y la comercialización de nuevos juegos, así como para comprender los intereses y preferencias de los jugadores.

# Ordenar media\_ventas\_por\_plataforma del promedio más alto al más bajo

media\_ventas\_por\_genero\_sorted = media\_ventas\_por\_genero.sort\_values(ascending=False)

# Crear un gráfico de barras para media\_ventas\_por\_genero

media\_ventas\_por\_genero\_sorted.plot(kind='bar', figsize=(12, 6))

plt.title('Promedio de Ventas Globales por Genero')

plt.ylabel('Promedio de Ventas Globales')

plt.xlabel('Genero')

plt.xticks(rotation=90)

plt.grid(True)

plt.show()

Gráfico

Descripción generada automáticamente

**6.4.1 información relevante**:

El promedio de ventas globales por genero están por debajo de 1 millones de copias.

El promedio total de ventas globales para todos los Géneros es de 548394 millones de copias.

Los Géneros que más venden son:

* Arcade.
* SandBox.
* Platform.
* Shooter.
* RPG.
* Racing.
* Sports.
* Accion.

**6.5 Análisis de Ventas Globales por Compañía de Desarrollo**

En este análisis, se examina la distribución de las ventas de videojuegos a nivel global según la compañía de desarrollo de los juegos, utilizando el atributo 'Compañia\_Desarrollo' que contiene la información de la compañía que desarrollo el videojuego y el atributo 'Ventas\_Globales' que contiene la información del total de ventas globales del videojuego, por lo que la relación puede calcular el total de ventas por compañía y visualizar esta información en un gráfico de lineal.

El objetivo es identificar qué compañías de desarrollo tienen mayor éxito en términos de ventas de videojuegos. Esto puede ser útil para evaluar la reputación y el rendimiento de las compañías de desarrollo y tomar decisiones sobre asociaciones o adquisiciones en la industria.

# Ventas globales por Compañia de Desarrollo

ventas\_por\_compañia = df\_juegos.groupby('Compañia\_Desarrollo')['Ventas\_Globales'].sum().sort\_values(ascending=False).head(22)

plt.figure(figsize=(12,6))

plt.plot(ventas\_por\_compañia.index, ventas\_por\_compañia.values, marker='o', linestyle='-')

plt.title('Ventas Globales por Compañía de Desarrollo')

plt.xlabel('Compañía de Desarrollo')

plt.ylabel('Ventas Globales (en millones)')

plt.xticks(rotation=90)

plt.grid(True)

plt.show()

Gráfico, Gráfico de líneas

Descripción generada automáticamente

**6.5.1 información relevante**:

La mayoría de las Compañías de Desarrollo tienen un promedio de ventas Globales debajo de 500000 mil copias.

Las Compañías de Desarrollo que más venden son:

* Nintendo.
* Electroni Arts.
* Activision.
* Sony.
* Ubisoft.

**7. Resultados de AED**

Para relacionar los análisis al modelado de datos, se resume los puntos fuertes del tipo de análisis acorde con el objetivo del proyecto:

**Análisis de Correlación de Ventas Globales por Consola y Genero:**

* Se puede examinar la correlación entre los atributos del DataFrame df\_juegos.
* Se puede identificar la correlación positiva entre las 'Ventas\_Globales', 'Ranking' y la 'Puntuacion\_Reseña'.
* Se puede evaluar qué que los videojuegos con puntuaciones de reseñas más altas tienden a tener más ventas globales y que se puede usar como variable objetivo para definir el Videojuego, considerando las reseñas y el ranking que el jugador puede tener al jugar un videojuego, y como este puede desarrollar una experiencia de juego enfocada a la experiencia educativa.

En conclusión, se pueden utilizar estos datos para determinar en qué género del videojuego se debe desarrollar el Videojuego.

**Análisis de Ventas Globales de Videojuegos por Plataforma:**

* Se puede examinar la distribución de las ventas de videojuegos según la plataforma o consola en la que se juegan.
* Se puede identificar qué plataformas son más populares en términos de ventas.
* Se puede evaluar qué plataformas son más adecuadas para el Videojuego, considerando las características técnicas y de usabilidad requeridas para ofrecer una experiencia educativa efectiva.

En conclusión, se pueden utilizar estos datos para determinar en qué plataformas se debe desarrollar y distribuir el Videojuego.

**Análisis de Ventas Globales por Género:**

* Se puede examinar la distribución de las ventas de videojuegos según el género.
* Se puede identificar los géneros de videojuegos más populares en términos de ventas.
* Se puede evaluar qué géneros son relevantes para el Videojuego, considerando el objetivo educativo o informativo del juego.

En conclusión, se pueden utilizar estos datos para determinar qué género de Videojuego tiene un mayor potencial de aceptación y éxito en el mercado.

**Análisis de Ventas Globales por Compañía de Desarrollo:**

* Se puede examinar la distribución de las ventas de videojuegos según la compañía de desarrollo.
* Se puede identificar qué compañías tienen un historial exitoso en términos de ventas de videojuegos.
* Se puede evaluar posibles asociaciones o colaboraciones con estas compañías para el desarrollo y distribución del Videojuego.

En conclusión, se pueden utilizar estos datos para seleccionar una compañía de desarrollo con experiencia en el género de juegos y plataforma seleccionados.

Al combinar estos análisis, se puede obtener una visión más completa y respaldada por datos del tipo de género y plataforma más adecuados para el desarrollo del Videojuego. Estos análisis proporcionarán información sobre las preferencias de los jugadores, las tendencias del mercado y las oportunidades para lograr el éxito en términos de ventas y aceptación del juego. Y validar la implementación del modelo predictivo en base al 'Genero', 'Plataforma' y 'Ventas\_Globales' \*\*"si el desarrollo de un Videojuego puede ser exitoso"\*\*.

**8. Conclusión**

En el análisis de los datos, se evaluaron tres tipos diferentes de modelos predictivos: un modelo de clasificación, un modelo de regresión lineal y un modelo de regresión logística. Cada uno de estos modelos tiene su utilidad y es útil en diferentes contextos.

Sin embargo, en base a los resultados obtenidos y al objetivo específico de este proyecto, se concluye que el modelo de regresión lineal es el más adecuado para predecir las ventas globales de un videojuego basado en atributos como el género, la plataforma y la compañía de desarrollo del juego. Los otros dos modelos, de clasificación y regresión logística, demostraron tener dificultades para realizar predicciones precisas en el contexto de este problema, ya que tuvieron problemas el primero para predecir el género de un videojuego y el segundo para predecir si un videojuego fuera exitoso.

A pesar de que se obtuvieron algunos resultados útiles, se hace evidente que hay margen para mejorar estos modelos. En concreto, sería útil explorar otras características que podrían influir en las ventas globales de un videojuego, como la fecha de lanzamiento del juego, las calificaciones recibidas por el juego, o incluso la presencia de ciertos personajes o temáticas en el juego. También puede ser útil probar con diferentes tipos de modelos, como los árboles de decisión, máquinas de soporte de vectores o redes neuronales artificiales, que pueden ser más eficaces en ciertos contextos.

Asimismo, sería beneficioso recolectar más datos y posiblemente ajustar los parámetros de los modelos para mejorar la precisión de las predicciones. Sin embargo, la recolección de datos adicionales siempre debe hacerse de manera ética y con el consentimiento adecuado, y debe tenerse en cuenta que ajustar demasiado los modelos puede llevar a un sobreajuste, donde el modelo se desempeña muy bien en los datos de entrenamiento, pero mal en los datos de prueba.

En resumen, aunque el modelo de regresión lineal fue el que dio los mejores resultados en este proyecto, hay mucho margen para mejorar y explorar otras opciones. El análisis de datos y la modelización son procesos iterativos, y siempre hay espacio para aprender, mejorar y probar nuevas ideas.

**8. Bibliográfia**

* Documentación oficial de Python: https://docs.python.org/3/
* Documentación oficial de Pandas: https://pandas.pydata.org/docs/
* Documentación oficial de Matplotlib: <https://matplotlib.org/stable/contents.html>
* Documentación oficial de Seaborn: <https://seaborn.pydata.org>
* Documentación oficial de Scikit-learn: https://scikit-learn.org/stable/
* Documentación oficial de Numpy: https://numpy.org/doc/stable/
* Analisis Exploratorio de Datos https://www.aprendemachinelearning.com/analisis-exploratorio-de-datos-pandas-python/
* Conceptos en Python <https://www.geeksforgeeks.org>
* Dudas <https://stackoverflow.com/questions/tagged/pandas+python>
* Analisis de datos <https://ocw.uc3m.es/course/view.php?id=230>
* Diccionarios de datos en data frame <https://github.com/nsheikh23/COVID_StockMarket_Analysis/blob/master/52_Week.ipynb>
* Procesamiento de data frames en pandas https://barcelonageeks.com/eliminar-una-o-varias-columnas-de-pyspark-dataframe/
* Data Clean <https://github.com/mramshaw/Data-Cleaning>
* Ploteo de datos <https://github.com/tomimester/python-histogram/blob/master/plot-histogram-python-pandas.ipynb>
* Data Cleaning in Python: the Ultimate Guide (2020) <https://towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d>
* Regresión lineal en Python: un análisis detallado <https://www.cienciadedatos.net/documentos/py10-regresion-lineal-python.html>
* Aplicando Machine Learning para predecir éxitos de videojuegos   
  <https://www.saturdays.ai/projects/Videojuegos.html>
* Machine Learning para la predicción de ventas https://www.aprendemachinelearning.com/regresion-lineal-en-espanol-con-python/
* McKinney, Wes. Python for Data Analysis. O'Reilly Media, Inc, 2017.
* VanderPlas, Jake. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media, Inc, 2016.