Géneros de videojuegos y su éxito en ventas: Un Análisis Regional

Autor: Lautaro Olivera

Tutor: Juan Cruz Alric

Fecha: 27/10/2023

Curso: Data Science / Comisión 41875

1. Descripción del caso de negocio

En la industria de los videojuegos, es crucial identificar los factores que inciden en las preferencias de consumo.

Vamos a analizar los datos de ventas de videojuegos desde sus inicios hasta 2016 para evaluar el impacto de la región, el género, el desarrollador y otros elementos en la industria.

2. Tabla de versionado

Existe únicamente esta versión

3. Objetivos del modelo

Considerando todas las variables en juego, nos planteamos la siguiente pregunta: ¿Influyen los géneros de videojuegos en la distribución de ventas por región?

Este análisis está dirigido a las empresas desarrolladoras y distribuidoras de videojuegos, ya que proporcionará información valiosa para adaptar sus estrategias a las preferencias regionales y destacar en esta competitiva industria.

4. Descripción de datos

El dataset "Video Games Sales" es un conjunto de datos creado por Hamed Etezadi, científico de datos. Es un dataset público que contiene los videojuegos más vendidos en la historia, desde 1977 hasta 2016, teniendo un tamaño de 16719 observaciones y 16 variables.

El mismo ha sido descargado del sitio Kaggle y se puede acceder en el siguiente Link.

A continuación se provee de una breve descripción de las variables:

• Name: Nombre del videojuego
• Platform: Plataforma en el que se publicó (Wii, Playstation 2, etc.)
• Year of Release: Fecha de lanzamiento
• Genre: Género (Acción, Deportes, Rol, etc.)
• Publisher: Empresa publicadora (Nintendo, Sony, etc.)
• NA Sales: Ventas en Estados Unidos, Canadá y México.
• EU Sales: Ventas en la Unión Europea
• JP Sales: Ventas en Japón
• Other Sales: Ventas en Latinoamérica, Oceanía, Asia (excluyendo China y Japón), África y países no incluidos en la Unión Europea.
• Global Sales: Total de ventas, sumando cada región.
• Critic Score: Reseña de la crítica, en una escala del 0 al 100.
• Critic Count: Cantidad de críticos que proporcionaron su reseña.
• User Score: Reseña del usuario/jugador, en una escala del 1 al 10, incluyendo decimales.
• User Count: Cantidad de usuarios/jugadores que proporcionaron su reseña.
• Developer: Empresa desarrolladora (Take-Two, Activision, etc.)
• Rating: Clasificación por edad

Resumen de las variables que forman parte del dataset (Crudo)					
Column	***Type***	***Non-Null***	***Nulls***	***Unique***	***Example***
Name	object	16717	2	11562	Grand Theft Auto: San Andreas
Platform	object	16719	0	31	PS2
Year_of_Release	float64	16450	269	39	2004.0
Genre	object	16717	2	12	Action
Publisher	object	16665	54	581	Take-Two Interactive
NA_Sales	float64	16719	0	402	9.43
EU_Sales	float64	16719	0	307	0.4
JP_Sales	float64	16719	0	244	0.41
Other_Sales	float64	16719	0	155	10.57
Global_Sales	float64	16719	0	629	20.81
Critic_Score	float64	8137	8582	82	95.0
Critic_Count	float64	8137	8582	106	80.0
User_Score	object	10015	6704	96	9
User_Count	float64	7590	9129	888	1588.0
Developer	object	10096	6623	1696	Rockstar North
Rating	object	9950	6769	8	M

Posteriormente, durante el proceso de manipulación de datos, se identificaron las variables que tenían una cantidad significativa de datos faltantes, las cuales se consideraron menos informativas y, por lo tanto, se descartaron del conjunto de datos.

Sin embargo, se tomó una decisión diferente con respecto a la variable 'Rating'. Se imputaron valores en base de la combinación de las variables 'Genre' y 'Publisher'. Para ello, se realizó un proceso de agrupación de datos, calculando la moda de la columna 'Rating' en base a las columnas 'Genre' y 'Publisher'. En el caso de que el valor de moda no resultase interesante o diferente, se utilizó el valor de moda únicamente de la columna 'Genre' para realizar la imputación de la columna 'Rating'

.

Se introdujeron dos nuevas variables al conjunto de datos. La primera, Sales_Category, se encarga de segmentar las ventas globales en tres categorías: por debajo del promedio, dentro del promedio y por encima del promedio. Esto ayuda a facilitar la predicción y el análisis al tiempo que prescinde de la necesidad de predecir cantidades exactas de ventas.

La segunda variable, Platform_Category, representa una reconversión de la clasificación de las plataformas de juegos. En lugar de mantener diversas plataformas individuales, ahora se agrupan en tres categorías principales: Consola, Portátil o PC. Este enfoque facilita la clasificación y contribuye a una predicción más precisa, aunque menos detallada.

Se resume que se realizaron los siguientes cambios en orden:

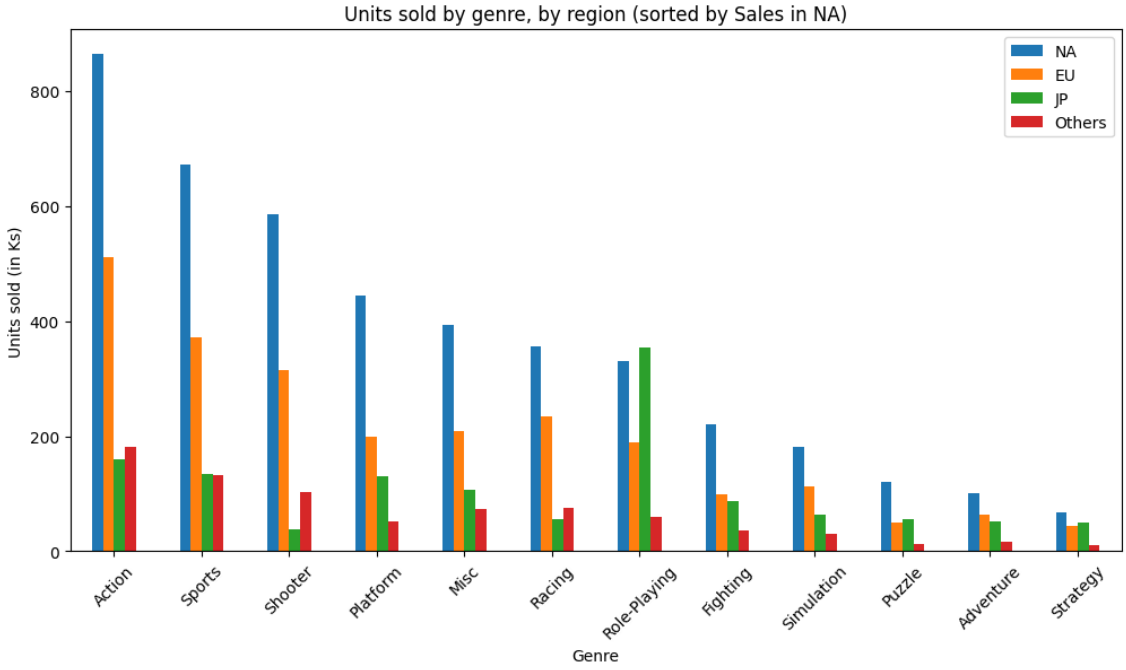
• Las variables “Critic_Score”, “Critic_Count”, “User_Score”, “User_Count” y “Developer” fueron removidas debido a la cantidad de datos nulos.
• Se agregaron valores en la variable “Rating” en base a “Genre” y “Publisher”.
• Se eliminaron todos los datos nulos restantes.
• Se generaron variables sintéticas "Sales_Category" y "Platform_Category" con el propósito de limitar las opciones del Modelo, resultando en una mayor precisión.

Resumen de las variables que forman parte del dataset (Manipulado)					
Column	***Type***	***Non-Null***	***Nulls***	***Unique***	***Example***
Name	object	16416	0	11397	Grand Theft Auto: San Andreas
Platform	object	16416	0	31	PS2
Year_of_Release	float64	16416	0	39	2004.0
Genre	object	16416	0	12	Action
Publisher	object	16416	0	579	Take-Two Interactive
NA_Sales	float64	16416	0	401	9.43
EU_Sales	float64	16416	0	307	0.4
JP_Sales	float64	16416	0	244	0.41
Other_Sales	float64	16416	0	155	10.57
Global_Sales	float64	16416	0	628	20.81
Rating	object	16416	0	4	M
Sales_Category	category	16416	0	3	High Sales
Platform_Category	object	16416	0	3	Console

5. EDA: Exploratory Data Analysis

En el análisis exploratorio se estudiaron las distribuciones de las variables numéricas y de las variables categóricas.

Unidades vendidas por géneros, por región (Ventas NA)



En el anterior gráfico se presenta una clasificación de los géneros más vendidos en todas las regiones, organizados según las ventas en NA

Se observa una notable similitud en el orden de los géneros más vendidos en las regiones de América del Norte (NA), Europa (EU) y otras regiones (Otros), especialmente en el top 3, donde los géneros más populares son Acción, Deportes y Disparos. Por otro lado, en Japón (JP), el género más vendido es Rol, mientras que el género menos vendido es Disparos.

Aunque esta información podría sugerir una respuesta directa a nuestra pregunta inicial, es importante continuar verificando la información.
Esto es necesario para asegurarnos de que no exista una respuesta inflada debido al impacto temporal de un videojuego particularmente exitoso.

Por ende, hemos reunido datos complementarios en forma de porcentajes de ventas a nivel mundial, desglosados por sector y género. Para mejorar la visualización de las preferencias de los jugadores en diversas regiones y géneros, hemos resaltado con un fondo verde los tres géneros más vendidos en cada región.

Además, estos porcentajes de ventas permiten comprender la contribución de cada región al mercado global. Resulta destacable que América del Norte (NA) aporta casi el 50% de las ventas totales, subrayando su influencia significativa en la industria.

Porcentajes de la influencia de los géneros en el mercado					
Genre	***Global_Sales***	***NA_Sales***	***EU_Sales***	***JP_Sales***	***Other_Sales***
Action	**19,49**	**9,80**	**5,80**	**1,82**	**2,07**
Sports	**14,87**	**7,62**	**4,21**	**1,53**	**1,50**
Shooter	**11,82**	**6,64**	**3,57**	0,44	**1,17**
Role-Playing	10,57	3,75	2,14	**4,01**	0,67
Platform	9,37	5,04	2,24	1,48	0,58
Misc	8,89	4,47	2,37	1,24	0,83
Racing	8,21	4,05	2,66	0,64	0,86
Fighting	5,02	2,50	1,12	0,99	0,41
Simulation	4,41	2,05	1,28	0,72	0,35
Puzzle	2,73	1,37	0,57	0,64	0,14
Adventure	2,65	1,15	0,72	0,59	0,19
Strategy	1,96	0,77	0,51	0,56	0,12
Total	100	49,21	27,18	14,67	8,88

Top 10 juegos más vendidos por región

Ahora, examinemos los videojuegos más exitosos en cada región. El "ID" se calcula en relación al puesto del juego más vendido a nivel mundial, restando 1. Esto significa que 0 representa el juego más vendido en todo el mundo, 1 el segundo más vendido, y así sucesivamente.

ID	Name	Year_of_Release	Platform	Genre	Global_Sales
0	Wii Sports	2006	Wii	Sports	82,53
1	Super Mario Bros.	1985	NES	Platform	40,24
2	Mario Kart Wii	2008	Wii	Racing	35,52
3	Wii Sports Resort	2009	Wii	Sports	32,77
4	Pokemon Red/Pokemon Blue	1996	GB	Role-Playing	31,37
5	Tetris	1989	GB	Puzzle	30,26
6	New Super Mario Bros.	2006	DS	Platform	29,80
7	Wii Play	2006	Wii	Misc	28,92
8	New Super Mario Bros. Wii	2009	Wii	Platform	28,32
9	Duck Hunt	1984	NES	Shooter	28,31

ID	Name	Year_of_Release	Platform	Genre	NA_Sales
0	Wii Sports	2006	Wii	Sports	41,36
1	Super Mario Bros.	1985	NES	Platform	29,08
9	Duck Hunt	1984	NES	Shooter	26,93
5	Tetris	1989	GB	Puzzle	23,20
2	Mario Kart Wii	2008	Wii	Racing	15,68
3	Wii Sports Resort	2009	Wii	Sports	15,61
14	Kinect Adventures!	2010	X360	Misc	15,00
8	New Super Mario Bros. Wii	2009	Wii	Platform	14,44
7	Wii Play	2006	Wii	Misc	13,96
18	Super Mario World	1990	SNES	Platform	12,78

ID	Name	Year_of_Release	Platform	Genre	EU_Sales
0	Wii Sports	2006	Wii	Sports	28,96
2	Mario Kart Wii	2008	Wii	Racing	12,76
10	Nintendogs	2005	DS	Simulation	10,95
3	Wii Sports Resort	2009	Wii	Sports	10,93
19	Brain Age: Train Your Brain in Minutes a Day	2005	DS	Misc	9,20
7	Wii Play	2006	Wii	Misc	9,18
6	New Super Mario Bros.	2006	DS	Platform	9,14
16	Grand Theft Auto V	2013	PS3	Action	9,09
4	Pokemon Red/Pokemon Blue	1996	GB	Role-Playing	8,89
15	Wii Fit Plus	2009	Wii	Sports	8,49

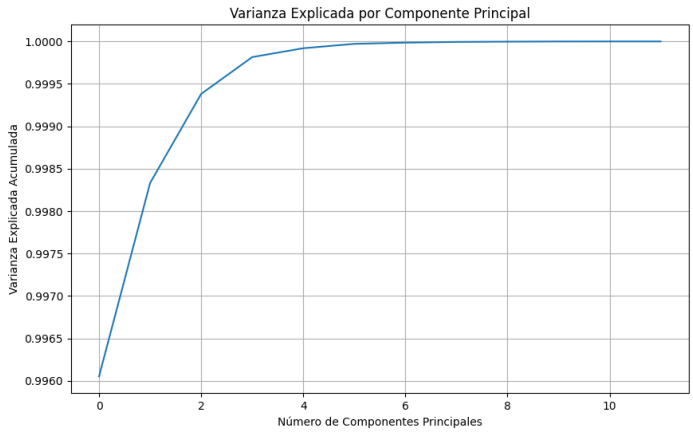
ID	Name	Year_of_Release	Platform	Genre	JP_Sales
4	Pokemon Red/Pokemon Blue	1996	GB	Role-Playing	10,22
12	Pokemon Gold/Pokemon Silver	1999	GB	Role-Playing	7,20
1	Super Mario Bros.	1985	NES	Platform	6,81
6	New Super Mario Bros.	2006	DS	Platform	6,50
20	Pokemon Diamond/Pokemon Pearl	2006	DS	Role-Playing	6,04
27	Pokemon Black/Pokemon White	2010	DS	Role-Playing	5,65
25	Pokemon Ruby/Pokemon Sapphire	2002	GBA	Role-Playing	5,38
43	Animal Crossing: Wild World	2005	DS	Simulation	5,33
26	Brain Age 2: More Training in Minutes a Day	2005	DS	Puzzle	5,32
215	Monster Hunter Freedom 3	2010	PSP	Role-Playing	4,87

ID	Name	Year_of_Release	Platform	Genre	Other_Sales
17	Grand Theft Auto: San Andreas	2004	PS2	Action	10,57
0	Wii Sports	2006	Wii	Sports	8,45
48	Gran Turismo 4	2004	PS2	Racing	7,53
16	Grand Theft Auto V	2013	PS3	Action	3,96
2	Mario Kart Wii	2008	Wii	Racing	3,29
3	Wii Sports Resort	2009	Wii	Sports	2,95
349	Pro Evolution Soccer 2008	2007	PS2	Sports	2,93
6	New Super Mario Bros.	2006	DS	Platform	2,88
7	Wii Play	2006	Wii	Misc	2,84
10	Nintendogs	2005	DS	Simulation	2,74

Al *examinar el top 10 en cada ubicación, podemos concluir que no hay un solo videojuego que haya tenido un impacto lo suficientemente significativo como para dominar por completo la lista de los géneros más vendidos. Además, se aprecia una notable diversidad de géneros en cada región, con "JP" y "Otros" siendo las únicas regiones con preferencias más definidas: "JP" muestra una preferencia por los juegos de rol, mientras que "Otros" tiende a favorecer los juegos de deportes.*

Además, en términos generales, los videojuegos más vendidos suelen estar asociados a plataformas de Nintendo. Se destaca la presencia de un único juego de Microsoft en el top, con unos pocos títulos de PlayStation.

Análisis de Componentes Principales (PCA)



Luego de verificar la cantidad de componentes necesarios para explicar la mayor cantidad del dataset, se podría concluir que prácticamente Publisher, insinuando lo valiosa que es esta variable para el consumidor, al momento de decidir cual juego consumir.

Column	Component 1	Component 2
Platform	0.000846	-0.974330
Year_of_Release	-0.001032	-0.217157
Genre	0.001113	-0.001400
Publisher	-0.999998	-0.000619
NA_Sales	-0.000021	-0.002623
EU_Sales	-0.000034	-0.002671
JP_Sales	-0.000092	0.003744
Other_Sales	-0.000014	-0.001312
Global_Sales	-0.000160	-0.002855
Rating	0.000424	-0.013922
Sales_Category	-0.000007	0.000660
Platform_Category	-0.000211	0.057358
Explained Variance	0.996052	0.002279
Total Explained Variance	0.996052	0.998331

El primer componente principal está fuertemente influenciado por la variable “Publisher”, mientras que el segundo componente está influenciado por “Platform” y ”Rating”.

En conjunto, estas dos componentes principales explican aproximadamente el 99.83% de la variabilidad en los datos originales.

Esto significa que estos componentes son altamente efectivos para resumir la información en los datos, lo que facilita el análisis y la interpretación de la estructura de los datos.

6. Algoritmo Elegido

Variable “Genre”:

Para aplicar modelos de Machine Learning para la variable “Genre”, fue necesario codificar las variables categóricas utilizando el método de Label Encoding en las columnas "Platform", "Publisher" y "Rating". En cuanto a las variables categóricas restantes, es decir, "Name" y "Genre", se optó por eliminarlas. La razón detrás de esta decisión es que estamos tratando de predecir la variable "Genre", pero "Name" actúa más como una identificación única o etiqueta para cada juego, siendo una explicación literal de a cuál juego nos referimos en lugar de una característica que aporta información predictiva.

Se utilizaron los siguientes modelos para predecir la variable “Genre”:

- *RandomForestClassifier*
- *SupportVectorMachine*
- *GradientBoostingClassifier*
- *KNeighboursClassifier*
- *NeuralNetworkClassifier*

Se realizaron pruebas de los modelos con una división de Train/Test de 80/20, los cuales se ejecutaron con valores por defecto, obteniendo los siguientes resultados de precisión en los modelos:

Metrics	Random Forest	SVM	Gradient Boosting	KNN	ANN
Accuracy	0.49	0.19	0.50	0.48	0.22

Luego, seleccionamos los modelos con la mejor precisión para realizar la afinación de hiper parámetros. Utilizamos una búsqueda en cuadrícula (grid search) para encontrar los mejores valores de hiper parámetros para estos modelos. Al evaluar las métricas finales de los dos modelos mejorados, observamos lo siguiente:

Metrics	Random Forest	Gradient Boosting
Accuracy	0.50	0.57
Precision	0.50	0.56
Recall	0.50	0.57
f1	0.49	0.56

En resumen, después de llevar a cabo la optimización de hiperparámetros y evaluar el rendimiento de los modelos, el modelo más eficaz resultó ser Gradient Boosting. Sin embargo, es importante destacar que incluso con la optimización, las métricas de rendimiento siguen siendo insatisfactorias, ya que no logra superar el umbral del 60% en ninguna de sus métricas.

Variable “Sales_Category”:

Dado que la predicción del género de los videojuegos no parece ser factible mediante modelos de Machine Learning, se exploró la posibilidad de predecir la variable sintética "Sales_Category" como un enfoque más viable para el modelo. Sin embargo, para llevar a cabo esta predicción, se requirió eliminar las siguientes variables: 'Name', 'Platform', 'Global_Sales', 'NA_Sales', 'EU_Sales', 'JP_Sales', 'Other_Sales' y 'Sales_Category'. Mientras que Label Encoding fue utilizado en las variables Publisher', 'Rating', "Genre" y "Platform_Category"

La razón detrás de esta selección radica en que mantener las variables que ayudan a predecir ventas podría llevar al modelo a memorizar datos en lugar de aprender patrones para predecir "Sales_Category", por lo que la intención original de crear el modelo se vería distorsionada.

Para llevar a cabo esta tarea, se utilizaron los mismos modelos que se habían empleado previamente para la predicción de “Genre”.

Al igual que en “Genre”, se realizaron los modelos con una divisón de 80/20 en Train/Test. Los resultados fueron los siguientes:

Metrics	Random Forest	SVM	Gradient Boosting	KNN	ANN
Accuracy	0.69	0.69	0.72	0.69	0.58

Aquí existen varios modelos de métricas similares, volviendo a destacar Gradient Boosting,aunque para una comparación, decidimos tomar Random Forest como una opción para aplicar una búsqueda en cuadrícula, pudiendo encontrar los hiper parámetros más óptimos.

Metrics	Random Forest	Gradient Boosting
Accuracy	0.73	0.73
Precision	0.66	0.66
Recall	0.73	0.73
f1	0.67	0.67

En comparación, ambos modelos con los hiper parámetros corregidos, brindan exactamente los mismos resultados.

Sin embargo, es importante remarcar la mejora de aproximadamente 15% comparado a la predicción de “Genre”. Este enfoque puede ayudar a los profesionales de la industria a tomar decisiones en otros aspectos y no únicamente en el género al cuál deben apuntar, siempre enfocándose en un mercado global.

7. Futuras líneas

Se destacan ciertas limitaciones tanto en la construcción del modelo como en las métricas que se aplicaron. El dataset proporciona una gama de información valiosa para evaluar el éxito de los videojuegos, pero denota carencias.

En primer lugar, la falta de datos actualizados es evidente, ya que la información más reciente se remonta a 2016. En un mercado tan dinámico como el de los videojuegos, donde han surgido nuevos subgéneros los desarrolladores independientes (Indie Developers) han cobrado gran relevancia y la Realidad Virtual (VR) ha emergido como una alternativa viable, la falta de datos más recientes puede limitar la precisión de las predicciones.

Además, uno de los desafíos notables radica en la dificultad para estimar con precisión las edades adecuadas del público objetivo. La clasificación por edades en la industria de los videojuegos a menudo no refleja la realidad, ya que ciertos juegos dirigidos a audiencias maduras pueden ser consumidos por un público más amplio, incluyendo a niños. Esto puede resultar en ventas infladas para juegos destinados a adultos, lo que puede distorsionar el análisis de las ventas en función de la clasificación por edades.

A pesar de la eliminación de las reseñas tanto de críticos como de usuarios, que podrían haber enriquecido el modelo, se reconoce que este es un componente valioso para predecir géneros. Las opiniones de los usuarios y las críticas de la prensa especializada aportan información adicional que podría haber mejorado la precisión del modelo.

Otro elemento útil que podría haberse considerado es la "duración del videojuego". La obtención de datos directamente desde fuentes como howlongtobeat.com podría haber ayudado a determinar el género de un videojuego en función de su duración promedio y la cantidad de usuarios que completaron el juego.

En resumen, el dataset enfrenta desafíos debido a la falta de datos recientes, la diversificación de subgéneros, la importancia de la industria indie, la aparición de la Realidad Virtual y la flexibilidad en cuanto a la edad del público objetivo. A pesar de estas limitaciones, el proyecto ha logrado sus avances y revela oportunidades para futuras mejoras.

8. Conclusiones

En base a la pregunta inicial que se dio en este proyecto de análisis de ventas de videojuegos, se han obtenido diversas conclusiones significativas.

En primer lugar, se ha identificado un patrón de similitud en las preferencias de género de videojuegos en diferentes regiones. América del Norte (NA), Europa (EU) y otras regiones comparten una afinidad hacia los géneros de Acción, Deportes y Disparos como los más vendidos. Este hallazgo indica una tendencia global hacia experiencias de juego orientadas a la acción y la competencia.

Sin embargo, Japón (JP) destaca como una excepción notable, donde el género de Rol se erige como el favorito indiscutible. Esta preferencia única puede atribuirse a la tradición de juegos de rol japoneses y franquicias icónicas como Final Fantasy, Dragon Quest y Monster Hunter, que han dejado una huella profunda en la cultura de los videojuegos en Japón.

Además, se ha observado la influencia de las plataformas en las preferencias regionales de los jugadores. Los juegos asociados con plataformas de Nintendo a menudo ocupan posiciones destacadas en las listas de los más vendidos. Esto indica que la elección de la plataforma de juego también desempeña un papel importante en la determinación de las preferencias de los jugadores en cada región.

Aunque este proyecto ha arrojado luz sobre las preferencias de género de videojuegos en diferentes regiones, también ha revelado desafíos significativos. La falta de datos actualizados y la complejidad de la clasificación por edades en la industria de los videojuegos son obstáculos notables. Sin embargo, los resultados alentadores en la predicción de la categoría de ventas abren nuevas oportunidades para tomar decisiones en la industria de los videojuegos.

En resumen, las preferencias de género de videojuegos influyen en la distribución de ventas por región, pero también se ven moldeadas por factores culturales y de plataforma. La localización efectiva y la comprensión de las diferencias regionales son esenciales para el éxito en un mercado global altamente competitivo. Las lecciones aprendidas en este proyecto proporcionan información valiosa para las empresas de la industria de videojuegos en su búsqueda de adaptarse y destacar en este emocionante y cambiante mercado.