Alberto Correa, Juan Cabrera, Lucas Andrade , Maximiliano Quiroga

Prediccion Ventas Videojuegos

Contenido

[1. Introducción 2](#_Toc170679391)

[2. Análisis Exploratorio de Datos 3](#_Toc170679392)

[a. Descripción de los datos 3](#_Toc170679393)

[i. Completitud 3](#_Toc170679394)

[4](#_Toc170679395)

[b. Tratamiento de los datos faltantes 4](#_Toc170679396)

[c. Datos atípicos 4](#_Toc170679397)

[d. Estadísticas Descriptivas 5](#_Toc170679398)

[e. Visualización de Datos 5](#_Toc170679399)

[f. Selección de características 8](#_Toc170679400)

[*Eliminación de características irrelevantes* 9](#_Toc170679401)

[*Características seleccionadas* 10](#_Toc170679402)

[Variables independientes: 10](#_Toc170679403)

[Variable dependiente: 10](#_Toc170679404)

[3. Preparación de Datos 11](#_Toc170679405)

[1. División de Datos 11](#_Toc170679406)

[2. Mezcla y Estandarización: 11](#_Toc170679407)

[4. Modelado 13](#_Toc170679408)

[Enfoque Inicial: Regresión mediante Redes Neuronales 13](#_Toc170679409)

[Descripción de las Características de Entrada al Modelo 13](#_Toc170679410)

[Justificación de la Codificación Dummy 14](#_Toc170679411)

[5. Resultados 15](#_Toc170679412)

[6. 16](#_Toc170679413)

[7. Conclusión 17](#_Toc170679414)

# **Introducción**

El mundo de los videojuegos se ha convertido en un gigante cultural y económico en constante expansión. Millones de jugadores alrededor del mundo se embarcan en aventuras virtuales y experiencias inmersivas, encontrando en este mercado vibrante una fuente de entretenimiento y conexión. Esta expansión no muestra señales de desaceleración. A medida que la población mundial aumenta, también lo hace la cantidad de jugadores. Los usuarios actuales, en su mayoría jóvenes, continuarán jugando a medida que envejezcan, consolidando una base de usuarios cada vez más amplia y diversa.

La creatividad, la innovación y la pasión por el entretenimiento son los pilares del éxito de la industria. Los desarrolladores de videojuegos constantemente reinventan la forma en que jugamos, creando experiencias cada vez más atractivas, realistas y socialmente relevantes. La industria del videojuego no solo entretiene, también educa, informa y genera nuevas formas de interacción social. Los videojuegos se han convertido en una herramienta poderosa para el aprendizaje, la colaboración y el desarrollo de habilidades.

En este contexto, comprender los factores que determinan el éxito de un videojuego es crucial para todos los actores de la industria. Este informe se propone analizar en profundidad las ventas globales de videojuegos, identificar las variables que influyen en el éxito de un juego y desarrollar un modelo predictivo para ayudar a tomar decisiones estratégicas.

El objetivo inicial de este proyecto era predecir las ventas globales de videojuegos utilizando diversas características del dataset proporcionado. Durante el análisis, se descubrió que la mayoría de las variables eran categóricas, lo que complicó la predicción exacta de las ventas mediante modelos de regresión. Por ello, se decidió cambiar el enfoque a una clasificación de las ventas globales en tres categorías: alta, media y baja.

Este informe detalla el proceso seguido para realizar esta clasificación, incluyendo la preparación y transformación de los datos, el modelado, la evaluación del modelo y los resultados obtenidos. La clasificación de las ventas puede ofrecer valiosa información para los desarrolladores y editores de videojuegos al permitirles identificar qué características están asociadas con mayores niveles de éxito en el mercado.

# **Análisis Exploratorio de Datos**

## Descripción de los datos

El presente estudio se basa en un conjunto de datos que contiene información sobre 16.718 videojuegos. Las características del conjunto de datos incluyen:

* Nombre del juego
* Plataforma (consola)
* Año de lanzamiento
* Género
* Publisher (editor)
* Ventas en Norteamérica, Unión Europea, Japón y otras regiones
* Ventas globales
* Puntuación de críticos y usuarios
* Desarrollador
* Clasificación por edades

### Completitud

El análisis del conjunto de datos reveló una cantidad significativa de datos faltantes, con un total de 17.5% de datos incompletos, principalmente en las características:

* Puntuación de críticos y usuarios
* Desarrollador
* Clasificación por edades

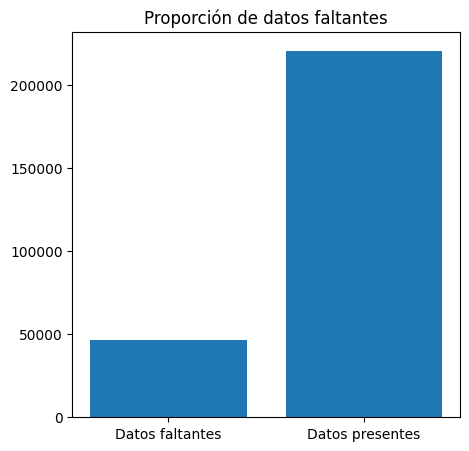


Gráfico 1. Proporción de datos faltantes en el conjunto de datos.

La distribución de los datos faltantes por característica se muestra a continuación:

Gráfico, Gráfico de barras

Descripción generada automáticamente

### 

Gráfico 2. Proporción de datos faltantes por características en el conjunto de datos.

## Tratamiento de los datos faltantes

Debido a la alta cantidad de datos incompletos en estas variables, se optó por eliminarlas del conjunto de datos. Adicionalmente, se eliminaron las filas con datos faltantes en otras características, quedando un conjunto final de 16.416 filas con 10 variables. Esta decisión se tomó considerando que la cantidad de datos faltantes era demasiado alta y que reemplazarlos por la media, la mediana o un modelo predictivo no sería significativo.

## Datos atípicos

Diagrama

Descripción generada automáticamenteSe detectaron valores atípicos en la variable "Ventas globales", correspondientes a juegos con un éxito comercial excepcional. Estos datos se consideran relevantes para el análisis, ya que representan casos de éxito dentro del mercado, por lo que no se eliminaron.

Gráfico 3. Ventas globales de videojuegos por plataforma y año de lanzamiento.

## Estadísticas Descriptivas

Se presenta un resumen estadístico del conjunto de datos, incluyendo:

* Media
* Desviación estándar
* Mínimo
* Percentiles 25%, 50% y 75%

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 0,2500 | 0,5000 | 0,7500 | max |
| Name | 16416 | 5741,67 | 3299,11 | 0,00 | 2888,00 | 5787,50 | 8624,25 | 11396,00 |
| Platform | 16416 | 15,85 | 8,37 | 0,00 | 7,00 | 16,00 | 21,00 | 30,00 |
| Year\_of\_Release | 16416 | 2006,49 | 5,881 | 1980,00 | 2003,00 | 2007,00 | 2010,00 | 2020,00 |
| Genre | 16416 | 4,91 | 3,77 | 0,00 | 1,00 | 5,00 | 8,00 | 11,00 |
| Publisher | 16416 | 293,856 | 177,63 | 0,00 | 138,00 | 325,00 | 457,00 | 578,00 |
| NA\_Sales | 16416 | 0,26 | 0,82 | 0,00 | 0,00 | 0,080 | 0,24 | 41,36 |
| EU\_Sales | 16416 | 0,15 | 0,51 | 0,00 | 0,00 | 0,020 | 0,11 | 28,96 |
| JP\_Sales | 16416 | 0,079 | 0,31 | 0,00 | 0,00 | 0,00 | 0,040 | 10,22 |
| Other\_Sales | 16416 | 0,048 | 0,19 | 0,00 | 0,00 | 0,010 | 0,030 | 10,57 |
| Global\_Sales | 16416 | 0,54 | 1,56 | 0,010 | 0,060 | 0,170 | 0,47 | 82,53 |

* Máximo

Tabla 1. Tabla de las estadísticas descriptivas de las ventas de un conjunto de videojuegos.

## Visualización de Datos

Se incluyen gráficos que permiten visualizar características relevantes del conjunto de datos, como:

### Distribución de ventas por plataforma

* La plataforma con el mayor número de ventas es **PC**.
* Las plataformas con un número de ventas similar son **PS4** y **XB1**.
* Gráfico, Gráfico de barras

  Descripción generada automáticamenteLas plataformas con el menor número de ventas son **N64**, **NES**, **SNES** y **TG16**.

Gráfico 4. Distribución de las ventas de videojuegos por plataforma.

### Evolución de las ventas por año

* Las ventas han aumentado de manera constante desde 1980 hasta 2010.
* No hay datos disponibles para las ventas después de 2010

Gráfico, Gráfico de líneas

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamente

Gráfico 5-5. Evolución de las ventas de videojuegos por plataforma (2010-2023).

### Ventas por género

* El género con el mayor número de ventas es Acción.
* Los géneros con un número de ventas similar son Aventura y Peleas.
* Los géneros con el menor número de ventas son Simulación y Deportes.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico 6. Gráfico de Distribución de las ventas por género.

### Histograma de ventas globales (sin valores atípicos)

* La mayoría de las empresas (más de 6000) tienen ventas globales entre 0.5 y 1 millón.
* Hay un grupo considerable de empresas (más de 4000) con ventas globales entre 1 y 1.25 millones.
* Hay pocas empresas con ventas globales por debajo de 0.5 millones o por encima de 1.75 millones.
* La mediana de las ventas globales es de 0.75 millones.
* La media de las ventas globales es de 1 millón.
* Gráfico, Histograma

  Descripción generada automáticamenteLa desviación estándar de las ventas globales es de 0.5 millones.

Gráfico 7. Histograma de Global\_Sales.

## Selección de características

La selección de características es un paso crucial en el proceso de aprendizaje automático, ya que permite identificar y elegir las variables más relevantes de un conjunto de datos para mejorar el rendimiento del modelo predictivo. En este caso, se presenta un análisis detallado de la selección de características realizada para predecir las ventas globales de videojuegos.

A través del análisis exploratorio de datos se identificaron dos características con un alto número de categorías que podían generar ruido en el modelo:

* Gráfico, Gráfico de barras

  Descripción generada automáticamente**Plataforma:** Originalmente presentaba 31 clases específicas para cada plataforma. Se reagruparon en tres categorías más generales: **consola de mano, consola de mesa y PC**.

Gráfico 7. Gráfico de frecuencia de Plataformas antes del agrupamiento.

* Gráfico, Gráfico de barras

  Descripción generada automáticamente**Publishers**: Inicialmente, había 581 categorías diferentes para cada Publisher. Para simplificar el análisis, se agruparon los publishers en tres categorías según el número de juegos que han lanzado:

Gráfico 8. Gráfico de frecuencia de Plataforma luego del agrupamiento.

* + **Publishers Menores**: Aquellos con menos de 100 juegos.
  + **Publishers Medios**: Aquellos con entre 100 y 700 juegos.
  + **Publishers Mayores**: Aquellos con más de 700 juegos.

Esto redujo significativamente el número de categorías de 581 a 3, facilitando el proceso de modelado y análisis.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Gráfico 9 Gráfico de frecuencia de Publishers luego de la agrupación.

## Eliminación de características irrelevantes

Se eliminaron tres características por considerarse irrelevantes para el modelo predictivo:

* **Ventas por región**: Al estar altamente correlacionadas con la variable dependiente "global\_sales" (que representa la suma de las ventas por región), se consideró que estas no aportaban información adicional.
* **Name**: Esta característica no proporcionaba información útil para el modelo, ya que no está directamente relacionada con las ventas del juego.
* **Desarrollador**: Se eliminó bajo la suposición de que el modelo busca predecir las ventas potenciales de un juego independientemente de quién lo desarrolle.

## Características seleccionadas

Tras la selección y eliminación de características, el conjunto de datos final quedó con las siguientes variables:

### Variables independientes:

* Plataforma
* Año de lanzamiento
* Género
* Publisher

### Variable dependiente:

* Ventas Globales

# **Preparación de Datos**

## Generación de conjuntos de entrenamiento, validación y prueba

### División de Datos

El conjunto de datos original se dividió en tres subconjuntos:

* **Conjunto de Entrenamiento (80%):** Este conjunto se utiliza para entrenar el modelo. El modelo aprende a identificar patrones en estos datos y a realizar predicciones.
* **Conjunto de Validación (10%):** Este conjunto se utiliza para evaluar el rendimiento del modelo durante el entrenamiento. Se utiliza para detectar posibles problemas de sobreajuste y ajustar los hiperparámetros del modelo.
* **Conjunto de Prueba (10%)**: Este conjunto se utiliza para evaluar el rendimiento final del modelo una vez entrenado. No se utiliza para entrenar ni ajustar el modelo, lo que proporciona una evaluación imparcial de su capacidad de generalizar a nuevos datos no vistos durante el entrenamiento.

### Mezcla y Estandarización:

**Mezcla:** Antes de dividir los datos, se mezclaron aleatoriamente para garantizar que cada subconjunto contenga una representación uniforme de las categorías de las variables independientes. Esto ayuda a evitar sesgos en los datos y reduce la probabilidad de que el modelo se sobreajuste a patrones específicos del conjunto de entrenamiento.

### Métricas de evaluación

Debido a que buscamos predecir un valor continuo, nos presentamos ante un modelo de regresión, por ende, las métricas de evaluación corresponderían a:

* Error absoluto medio (MAE)
* R al cuadrado (R^2
* R cuadrado ajustado (R^2 ajustado)
* Error porcentual absoluto medio (MAPE)
* Error logarítmico cuadrático medio (RMSLE)

Debido a que la escala de la variable objetivo, algunas métricas, como el RMSE y el MAE, son sensibles a la escala de la variable objetivo, por lo que se incluirá ambos, pero debido a la presencia de valores atípicos el MAE es menos sensibles a estos, por el contrario del RMSE, y debido a que mis datos tienen datos atípicos por lo que se usará el MAE.

Además, debido a que igual es necesario poder interpretar los diferentes modelos, el es una métrica fácil de interpretar, mientras que el RMSE no lo es tanto.

# **Modelado**

## Enfoque Inicial: Regresión mediante Redes Neuronales

El objetivo inicial de este proyecto fue predecir las ventas globales de videojuegos utilizando un modelo de regresión basado en redes neuronales. La elección de redes neuronales como técnica de modelado se fundamentó en su capacidad para manejar relaciones no lineales complejas y capturar patrones intrincados en los datos. Sin embargo, la dificultad para obtener una predicción precisa de ventas continuas llevó a un cambio de estrategia a clasificación, como se discutirá más adelante.

### Descripción de las Características de Entrada al Modelo

#### Características Categóricas (Codificación Dummy)

* **Plataforma (Platform)**
  + Descripción: Representa el tipo de consola en la que se ejecuta el videojuego.
  + Codificación: Para facilitar la entrada al modelo, se utilizó codificación dummy, creando variables binarias para cada tipo de plataforma posible: consola de mano, consola de mesa y PC.
* **Género (Genre)**
  + Descripción: Indica la categoría del juego, como acción, aventura, RPG, estrategia, etc.
  + Codificación: Se empleó codificación dummy, generando una variable binaria para cada género.
* **Editor (Publisher)**
  + Descripción: Designa la entidad que publica el videojuego.
  + Codificación: Se crearon variables dummy para cada editor, permitiendo al modelo diferenciar entre los efectos específicos de cada editor.

#### Características Numéricas

* Año de Lanzamiento (Year\_of\_Release)
  + Descripción: Representa el año en que se lanzó el videojuego.
  + Codificación: Se utilizó directamente como una variable numérica en el modelo.

### Justificación de la Codificación Dummy

La codificación dummy es una técnica comúnmente empleada para convertir características categóricas en una serie de variables binarias que el modelo puede interpretar. Esta transformación permite que el modelo capture el impacto de cada categoría específica sobre la variable dependiente, en este caso, las ventas globales del videojuego.

Sin la codificación dummy, las características categóricas no podrían ser procesadas correctamente por el modelo de regresión de redes neuronales, ya que este solo podría identificar el efecto general de dichas características, sin diferenciar entre las distintas categorías. La codificación dummy, por tanto, enriquece el modelo permitiendo que cada categoría influya de manera independiente en las predicciones del modelo.

### Implementación del Modelo de Regresión

Para la implementación inicial del modelo de regresión, se siguieron estos pasos:

#### Preprocesamiento de Datos

* **Escalado de Características Numéricas:** Las características numéricas, como el año de lanzamiento, se escalaron para tener media 0 y desviación estándar 1, ayudando al modelo a converger más rápidamente.
* **Codificación Dummy:** Se aplicó la codificación dummy a las características categóricas, generando un total de [número] características binarias.

### Construcción de la Red Neuronal

#### Arquitectura del Modelo:

Tabla

Descripción generada automáticamente

Ilustración 1. Modelo Regresión

* **Función de Activación:** Se seleccionó la función de activación Adam para capturar la no linealidad en los datos.

#### Entrenamiento del Modelo

* **Configuración del Entrenamiento:** El modelo se entrenó durante 100 épocas con un tamaño de lote de 32
* **Métrica de Evaluación:** Se utilizó la métrica MAE para evaluar el desempeño del modelo durante el entrenamiento.

### **Cambio de Estrategia**

Debido a la baja precisión y los altos errores obtenidos en el enfoque inicial de regresión mediante redes neuronales, se decidió cambiar a un modelo de clasificación para categorizar las ventas globales en rangos: alta, media y baja. Esta transición se aborda en las siguientes secciones del informe.

## Implementación del Modelo de Clasificación

Para la implementación inicial del modelo de regresión, se siguieron estos pasos:

### Preprocesamiento de Datos

* **Escalado de Características Numéricas:** Las características numéricas, como el año de lanzamiento, se escalaron para tener media 0 y desviación estándar 1, ayudando al modelo a converger más rápidamente.

### Construcción de la Red Neuronal

#### Tabla Descripción generada automáticamente con confianza mediaArquitectura del Modelo

Ilustración 2 Modelo de Clasificación

* **Función de Activación:** Se seleccionó la función de activación Adam para capturar la no linealidad en los datos.

#### Entrenamiento del Modelo

* **Configuración del Entrenamiento:** El modelo se entrenó durante 100 épocas con un tamaño de lote de 32
* **Métrica de Evaluación:** Se utilizó la métrica Precisión para evaluar el desempeño del modelo durante el entrenamiento.

# **Resultados**

## Resultados Modelo de regresión

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MAE | MSE | RMSE | R2 |
| Entrenamiento | 0.5 | 1.92 | 1.39 | 0.03 |
| Validación | 0.5 | 1.81 | 1.34 | 0.03 |
| Prueba | 0.54 | 4.10 | 2.02 | 0.01 |

Tabla 2 Métricas Modelo de Regresión

Gráfico, Histograma

Descripción generada automáticamente

Gráfico 10 Gráfico de Perdida Modelo Regresión

## Resultados Modelo de Clasificación

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 Score |
| Entrenamiento | 0.86 | 0.29 | 0.33 | 0.30 |
| Validación | 0.86 | 0.29 | 0.33 | 0.30 |
| Prueba | 0.87 | 0.29 | 0.33 | 0.31 |

Tabla 3 Métricas Modelo de Clasificación

Gráfico, Histograma

Descripción generada automáticamente

Gráfico 11 Grafico Perdidas Modelo de Clasificación

# **Conclusión**

El análisis exhaustivo de los datos de ventas globales de videojuegos ha proporcionado valiosos insights sobre los factores que influyen en el éxito de estos productos en el mercado. A partir de un conjunto inicial de datos, se identificaron y preprocesaron variables relevantes, destacando las siguientes conclusiones principales:

### Completitud y Tratamiento de Datos

* **Completitud de Datos:** Se observó un 17.5% de datos faltantes en características críticas como la puntuación de críticos y usuarios, desarrollador, y clasificación por edades. La eliminación de estas variables, junto con las filas incompletas, dejó un conjunto de datos más limpio y manejable con 16,416 registros.
* **Datos Atípicos:** Los valores atípicos en las ventas globales, reflejando juegos con éxito comercial excepcional, fueron retenidos para proporcionar una visión más completa del mercado.

### Visualización y Selección de Características

* **Distribución de Ventas:** Las ventas mostraron una tendencia creciente hasta 2010, con variaciones significativas según la plataforma y el género. La plataforma PC lidera en ventas, mientras que el género de acción domina en popularidad.
* **Selección de Características:** Se reagrupó la variable "plataforma" en categorías más generales y se simplificaron los publishers en tres categorías. Las ventas regionales, el nombre del juego, y el desarrollador se eliminaron debido a su baja relevancia para el modelo.

### Modelado

* **Regresión:** El enfoque inicial de regresión con redes neuronales no logró una precisión adecuada en la predicción de ventas continuas. El modelo mostró un desempeño deficiente con un R^2 muy bajo y errores relativamente altos (MAE alrededor de 0.5).
* **Clasificación:** El cambio a un modelo de clasificación para categorizar las ventas en alta, media, y baja resultó más efectivo. El modelo de clasificación logró una precisión de 0.87 en el conjunto de prueba, con métricas aceptables de precisión, recall, y F1 score, destacando su capacidad para clasificar correctamente las ventas en las tres categorías establecidas.

## Recomendaciones

* **Mejora de Datos:** Para futuros análisis, sería beneficioso obtener datos completos y reducir la cantidad de datos faltantes, especialmente en variables clave como las puntuaciones de críticos y usuarios.
* **Análisis Adicional:** Incluir características adicionales o mejorar la granularidad de las existentes podría proporcionar más precisión en las predicciones.
* **Validación Continua:** Implementar técnicas de validación más robustas y experimentar con otros algoritmos de aprendizaje automático podría mejorar aún más el modelo de clasificación.