Universidad Peruana de Ciencias Aplicadas



**TB1: Segundo Entregable**

Alumno: Muchaypiña Palomino, Anthony Gabriel

Profesora: Manrique Tunque, Nérida Isabel

**Octubre de 2025**

# Proyecto de Data Science – CRISP-DM

## **Tema elegido**

Predicción del éxito de ventas de videojuegos

El proyecto analizará datos históricos de videojuegos para predecir si un juego tuvo ventas altas o bajas según sus características principales: plataforma, género, editor y año de lanzamiento. Esto permitirá analizar qué factores se relacionan con el éxito comercial de un videojuego.

## **Objetivo del proyecto**

Desarrollar un modelo de clasificación capaz de determinar si un videojuego alcanzó ventas altas o bajas (definidas como los juegos en el top 25% de ventas globales) usando variables del juego como plataforma, género, editor y año de lanzamiento, logrando un accuracy mínimo de 75–80% en el conjunto de prueba.

## **Dataset seleccionado**

Nombre del dataset: *Video Games Sales Dataset*

Fuente: [Kaggle – Video Games Sales](https://www.kaggle.com/gregorut/videogamesales)

Tamaño: Aproximadamente 16,000 registros con 11 variables. Entre las principales:

* Variables principales (predictoras)
  + Platform: Plataforma del juego (PS4, Xbox, Switch, etc.)
  + Gnere: Género del juego
  + Publisher: Editorial
  + Year\_of\_Release: Año de lanzamiento
  + NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales: Ventas regionales en millones
  + Rating: Clasificación por edades (E, T, M, etc.)
* Variable objetivo
* Global\_Sales transformada en “alta/baja” según top 25% de ventas

1. **Enlace al repositorio**

<https://github.com/Gabriell3ll/VGSALES_PROYECTO/tree/main>

1. **Asignación explícita de roles y tareas**

* Project Lead — coordinación y entrega  
  Responsable de la planificación, organización del repositorio, consolidación del informe y entrega final.
* Data Engineer — adquisición y limpieza  
  Encargado de la descarga del dataset desde Kaggle, validación del esquema, carga inicial, detección y tratamiento propuesto de valores faltantes, duplicados y tipos de datos.
* Data Analyst — EDA y visualizaciones  
  Realización de estadísticas descriptivas (numéricas y categóricas), identificación de outliers, y elaboración de visualizaciones para explorar relaciones entre variables.
* Data Scientist — modelado preliminar  
  Definición de la variable objetivo (ventas altas vs bajas), construcción del baseline de clasificación y propuesta de métricas para evaluar modelos futuros.

1. **Identificación de valores faltantes, duplicados y outliers**

* Valores faltantes — identificación

**Texto

El contenido generado por IA puede ser incorrecto.**

Comentarios:

* + Categóricas (Publisher, Genre, Platform, Rating):

En este dataset, si un valor como Publisher falta, no significa que no exista editorial, sino que la fuente no la registró. Usar la moda (el valor más frecuente) es una forma práctica de imputación porque refleja la editorial o categoría más habitual y evita generar etiquetas artificiales como “Unknown”, que no aportan información al modelo. Sin embargo, hay que reconocer que esta estrategia puede ocultar la diversidad de publishers menores, pero es útil si los faltantes son pocos.

* + Numéricas (Year):

El Year de lanzamiento está sesgado hacia décadas recientes. Imputar con la mediana reduce la influencia de valores extremos (juegos muy viejos o recientes) y mantiene coherencia temporal sin sesgar tanto como la media.

* Duplicados — identificación

Texto

El contenido generado por IA puede ser incorrecto.

Comentarios:

* Los duplicados distorsionan estadísticas y deben eliminarse.
* Conservar la primera ocurrencia mantiene la información sin redundancia.
* La clave compuesta Name+Platform+Year se usa porque un mismo juego puede aparecer en distintas plataformas y eso no es duplicado, sino un caso válido.
* Outliers — identificación

Texto

El contenido generado por IA puede ser incorrecto.

Comentarios:

Los outliers en ventas globales representan éxitos de mercado, no errores (ej. Wii Sports, GTA V). Por lo tanto, deben conservarse.

Para modelado, se sugiere:

* Transformación logarítmica para estabilizar la distribución.
* Capping (winsor) en percentiles altos si un algoritmo es sensible a extremos.

En análisis descriptivo se deben mostrar completos porque aportan información clave sobre el negocio.

* Resumen
* Imputación de categóricas con moda → refleja la categoría más común y retiene registros incompletos.
* Imputación de numéricas con mediana → robusta frente a juegos muy antiguos/recientes.
* Eliminación de duplicados → clave Name+Platform+Year evita inflar ventas.
* Outliers detectados y conservados → representan “megahits” reales, se mantienen en análisis descriptivo, pero se crean variables transformadas para modelado.
* Conclusión: El dataset queda limpio y consistente, sin perder información de negocio (los títulos más vendidos), reduciendo sesgos en imputación y duplicados.

1. **Visualizaciones entre variables**
   1. Visualización 1 — Juegos por rangos de ventas globales

Global\_Sales: ventas mundiales de cada juego en millones de copias.

* Gráfico: contamos cuántos juegos hay en cada rango de ventas (en millones). Esto es más legible que un histograma continuo.
* Insight: La mayoría de títulos venden menos de 1M; cada vez hay menos juegos conforme suben los rangos. Los "megahits" (≥20M) son raros pero relevantes para el mercado.

Gráfico, Histograma

El contenido generado por IA puede ser incorrecto.

Pantalla de celular con números

El contenido generado por IA puede ser incorrecto.

* 1. Visualización 2 — Global\_Sales por Género

Un boxplot muestra la variación de ventas por género.

* Insight: Géneros como Shooter y Sports tienen medianas más altas, lo que indica mayor propensión al éxito.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

* 1. Visualización 3 — Promedio de Global\_Sales por Plataforma (Top 10)

Se grafican las plataformas con mayor promedio de ventas.

* Insight: Consolas como Wii y PS4 muestran un desempeño promedio muy superior, lo que evidencia su peso en la industria.

Gráfico, Gráfico de barras

El contenido generado por IA puede ser incorrecto.

1. **Análisis**

## **¿En qué medida los datos responden al problema planteado?**

El dataset de **Video Game Sales** (vgchartz) resulta pertinente para abordar el problema de predecir el éxito comercial de los videojuegos porque:

* Contiene variables descriptivas fundamentales:
  + **Plataforma (Platform):** indica el ecosistema tecnológico, uno de los factores más influyentes en el volumen de ventas (ej. consolas con gran base de usuarios tienden a concentrar los títulos más vendidos).
  + **Género (Genre):** aporta información sobre las preferencias de los consumidores (ej. shooters y deportes suelen tener alta demanda).
  + **Publisher:** representa la capacidad de marketing y distribución de cada compañía, lo que afecta fuertemente la visibilidad de los juegos.
  + **Año de lanzamiento (Year):** permite analizar la evolución temporal del mercado y los ciclos de vida de consolas.
  + **Ventas regionales (NA, EU, JP, Other):** reflejan diferencias culturales y de consumo en distintos mercados.
* Contiene la variable objetivo **Global\_Sales**, expresada en millones de copias, a partir de la cual se puede derivar una clasificación binaria (ventas altas vs bajas). Esto permite operacionalizar el problema en términos de machine learning.

En ese sentido, los datos **sí permiten analizar qué factores se asocian con el éxito comercial** y posibilitan la construcción de un modelo predictivo básico. Sin embargo, presentan varias limitaciones que deben tenerse en cuenta.

1. **Limitantes**
2. **Sesgo temporal y tecnológico**
   1. Los datos abarcan títulos desde finales de los años 80 hasta aproximadamente 2016.
   2. El dataset refleja principalmente la era de los **juegos físicos en cartucho y disco**, pero no captura el auge del mercado digital (descargas, microtransacciones, juegos como servicio).
   3. Esto significa que las conclusiones son válidas para el mercado tradicional de consolas, pero **pueden no generalizarse al contexto actual**, donde dominan Steam, PS Store, Xbox Live, Epic Games y Nintendo eShop.
   4. Además, fenómenos recientes como free-to-play, battle royale y servicios de suscripción (Game Pass, PS Plus) quedan fuera del alcance.
3. **Variables incompletas o ausentes**
   1. Aunque se registran metadatos básicos, faltan variables cruciales que determinan el éxito de un juego:
      1. **Inversión en marketing**: campañas publicitarias influyen fuertemente en la visibilidad y ventas.
      2. **Precio de lanzamiento**: impacta en la accesibilidad y el volumen de ventas.
      3. **Calificaciones y reseñas**: tanto de críticos como de usuarios, son un predictor comprobado de la demanda.
      4. **Base instalada de consolas**: el tamaño del mercado de cada plataforma condiciona el potencial de ventas.
   2. La ausencia de estas variables reduce la capacidad explicativa y genera un modelo con **alto riesgo de omisión de variables relevantes**.