Resultados

1. Resumen de lo realizado

En este proyecto se trabajó con el conjunto de datos **online_gaming_insights**, que contiene información sobre jugadores de videojuegos en línea: características demográficas, hábitos de juego (tiempo y frecuencia), logros, dispositivos utilizados, motivaciones y variables relacionadas con el gasto dentro del juego. El objetivo general fue **modelar y analizar el nivel de compromiso (engagement) y el comportamiento de compra de los jugadores** utilizando diferentes técnicas de *Machine Learning*.

En el **análisis exploratorio de datos (EDA)** se calcularon los porcentajes de valores nulos y de ceros en las variables numéricas, verificando si estos ceros tenían sentido de acuerdo con el contexto (por ejemplo, ausencia de compras). Se generaron descripciones estadísticas básicas y gráficos (histogramas, barras y mapas de correlación) para entender la distribución de las variables y la relación entre ellas. Se identificaron valores atípicos (outliers) y, en lugar de eliminar registros completos, se optó por **recortar (clipping)** los valores extremos a rangos razonables, con el fin de reducir su influencia sin perder información.

También se construyó un **mapa de calor de correlación** entre variables numéricas. No se encontraron pares de variables con correlaciones tan altas como para justificar su eliminación, por lo que se decidió conservar la mayoría de las características. Adicionalmente, se aplicó el algoritmo de **clustering DIANA** (Divisive Analysis) como técnica jerárquica para identificar grupos de jugadores con comportamientos similares, complementando el análisis con las conclusiones obtenidas previamente con K-Means en el EDA.

En la fase de **preparación de datos**, se eliminaron columnas de identificación o poco informativas (PlayerID, Location), se imputaron los valores faltantes (utilizando la mediana en variables numéricas), se codificaron las variables categóricas mediante **One-Hot Encoding** y se estandarizaron las características con un **Scaler**. La variable de interés relacionada con el engagement se transformó en un valor numérico ordenado (por ejemplo, 1 = Low, 2 = Medium, 3 = High). Con el fin de reducir la dimensionalidad y evitar trabajar con un número muy grande de columnas dummy, se aplicó **PCA** conservando un porcentaje alto de varianza explicada. Finalmente, se generaron los archivos de **Train, Validation y Test** para las tareas de clasificación y regresión, siguiendo las proporciones indicadas (80/20 y luego 80/20 sobre Train).

En la carpeta **experimentos** cada integrante implementó un algoritmo diferente:

- **RIPPER** (notebook JuanB_RIPPER.ipynb), como modelo de reglas para clasificación.

- **DIANA** (notebook Luis_Santos_DIANA.ipynb), usado para profundizar en la segmentación jerárquica de jugadores.
- **Histogram Gradient Boosting** (notebook Zoet_HistogramGradientBoosting.ipynb), como modelo de boosting para variables numéricas y categóricas.
- **Bagging** con árboles de decisión, tanto para clasificación como para regresión del engagement (notebook nikholas_bagging.ipynb).

En todos los casos se trabajó con los **conjuntos de entrenamiento, validación y prueba** previamente generados, se variaron al menos tres hiperparámetros por algoritmo y se construyeron tablas comparativas con las medidas de rendimiento en *train* y *validation*.

2. Modelo seleccionado y predicción de 5 registros

Tras revisar las tablas de resultados de cada notebook de la carpeta experimentos, se comparó el desempeño de los algoritmos en términos de **error en validación** (para clasificación y, cuando aplica, para regresión) y de la capacidad de generalizar al conjunto de prueba. De manera general:

- **RIPPER** generó reglas interpretables y útiles para entender patrones de decisión, pero su rendimiento fue más modesto y mostró cierta sensibilidad al ruido en los datos.
- **DIANA** se utilizó como técnica de segmentación jerárquica. Es útil para entender grupos de jugadores, pero no se planteó como modelo predictivo principal, sino complementario al análisis exploratorio.
- **Histogram Gradient Boosting** obtuvo buenos resultados y aprovechó bien las variables numéricas, aunque su configuración requirió más cuidado para evitar sobreajuste.
- **Bagging con árboles de decisión** logró un equilibrio adecuado entre rendimiento en entrenamiento y validación, con errores relativamente bajos y estables, tanto en clasificación como en regresión del engagement.

De acuerdo con las tablas comparativas del notebook nikholas_bagging.ipynb, el modelo que presentó **mejor comportamiento global para predecir el engagement** (considerando error bajo en validación y buena generalización en test) fue el **Bagging** utilizando un árbol de decisión como estimador base y una configuración de **valores intermedios de los hiperparámetros** (por ejemplo, un número moderado de estimadores y proporciones intermedias de muestras y características), lo que permitió evitar tanto el sobreajuste como el infraajuste.

Por esta razón, el modelo seleccionado para el punto 9 es el **Bagging aplicado a la regresión del engagement**. Con los mejores hiperparámetros encontrados, se re-entrenó el modelo sobre el conjunto de entrenamiento completo y se utilizaron **cinco perfiles de jugador** sintéticos, que representan casos muy distintos entre sí, para ilustrar las predicciones:

Jugador	Descripción del perfil (resumen)	Engagement predicho (aprox.)
1	Joven, muchas sesiones por semana, sesiones largas, alto número de logros, realiza compras con frecuencia.	Engagement alto (cerca de 3)
2	Adulto, pocas sesiones por semana, sesiones cortas, casi sin logros y sin compras.	Engagement bajo (cerca de 1)
3	Joven, tiempo de juego medio, varios logros, pocas compras pero constante conexión semanal.	Engagement medio (alrededor de 2)
4	Nuevo jugador, pocas sesiones registradas, sin logros ni compras y uso esporádico del juego.	Engagement bajo (cerca de 1)
5	Jugador con tiempo de juego moderado, varios logros alcanzados, compras ocasionales y uso estable del juego.	Engagement medio-alto (entre 2 y 3)

Estos cinco casos muestran cómo el modelo responde ante perfiles de jugadores claramente diferenciados en su comportamiento.

3. Resultados obtenidos

A nivel global, los resultados del proyecto pueden resumirse en tres bloques: **análisis exploratorio y segmentación**, **preparación y transformación de datos**, y **desempeño de los modelos de aprendizaje automático**.

1. **Análisis exploratorio y segmentación**
    a. El análisis de nulos, ceros y outliers permitió identificar problemas de calidad de datos y tratarlos de forma controlada (imputación y recorte).
    b. Los gráficos y el mapa de correlación mostraron que, si bien hay relaciones entre algunas variables, no se observaron correlaciones tan altas como para forzar la eliminación de columnas.
    c. Los algoritmos de clustering (K-Means en el EDA y DIANA en los experimentos) permitieron identificar grupos de jugadores con distintos patrones, en particular diferencias en la **intensidad de juego** y en la relación entre tiempo invertido y logros.
2. **Preparación y transformación de datos**
    a. La eliminación de columnas no informativas y la codificación de variables categóricas redujeron el ruido y permitieron trabajar con modelos que exigen variables numéricas.
    b. El uso de **PCA** ayudó a condensar la información de muchas características en un número menor de componentes, manteniendo la mayor parte de la varianza y haciendo el entrenamiento más eficiente.
    c. La separación en **Train, Validation y Test** aseguró que la evaluación de los modelos fuera justa y que se pudiera comparar su rendimiento sin sesgo.

3. **Desempeño de los modelos de aprendizaje automático**
    a. **RIPPER** produjo reglas claras y fácilmente interpretables, útiles para explicar por qué ciertos jugadores se clasifican en determinadas categorías, aunque con un desempeño moderado frente a otros modelos más complejos.
    b. **DIANA** ayudó a entender la estructura jerárquica de los grupos de jugadores, profundizando en la segmentación vista en el EDA, pero no se usó como modelo de predicción principal.
    c. **Histogram Gradient Boosting** mostró buen ajuste a los datos y capturó relaciones no lineales, con resultados competitivos, aunque requirió un ajuste cuidadoso de sus hiperparámetros para evitar sobreajuste.
    d. **Bagging con árboles de decisión** consiguió un equilibrio adecuado entre error de entrenamiento y validación, con errores similares en ambos conjuntos y un buen rendimiento en el conjunto de prueba. Esto indica una **buena capacidad de generalización** y justifica su selección como modelo final para la predicción del engagement.

En conjunto, los resultados muestran que el pipeline construido permite pasar de datos crudos a un modelo capaz de **estimar el nivel de engagement de nuevos jugadores** a partir de sus características.

4. Conclusiones

1. El proyecto permitió construir un proceso completo de *Machine Learning* que incluye **limpieza, análisis exploratorio, transformación, segmentación y modelado**, cumpliendo con los requerimientos establecidos para el proyecto 4.
2. El análisis exploratorio evidenció que variables como **frecuencia de juego, duración de las sesiones, logros e historial de compras** están estrechamente relacionadas con el nivel de engagement, mientras que otras variables (como la mera identificación del jugador o su localización textual) no aportan valor al modelo.
3. Los algoritmos de **segmentación (K-Means y DIANA)** permitieron identificar grupos de jugadores con comportamientos similares, lo que puede servir como base para estrategias de personalización y marketing en el contexto de juegos en línea.
4. La comparación de varios algoritmos en la carpeta experimentos mostró que no todos los modelos se comportan igual frente a este tipo de datos:
    a. Modelos más simples y explicables, como RIPPER, aportan interpretabilidad pero pierden algo de precisión.
    b. Modelos más potentes, como Histogram Gradient Boosting y Bagging, capturan mejor las relaciones complejas entre las variables.
5. Entre los algoritmos probados, el **modelo de Bagging con árbol de decisión como estimador base** fue el que presentó **mejor equilibrio entre rendimiento y**

estabilidad, por lo que se seleccionó como modelo final para la predicción del engagement.

6. Las predicciones realizadas sobre cinco perfiles de jugadores muy diferentes entre sí mostraron resultados coherentes con la intuición: jugadores intensivos, con logros y compras frecuentes tienden a clasificarse con engagement alto, mientras que jugadores con poco tiempo de juego y sin logros ni compras tienden a clasificarse con engagement bajo.

5. Posibles mejoras

A partir de lo realizado, se identifican varias posibilidades de mejora para trabajos futuros:

- **Profundizar en la búsqueda de hiperparámetros** utilizando técnicas como Grid Search o Random Search con validación cruzada, en lugar de probar solamente un conjunto limitado de combinaciones, para cada uno de los algoritmos evaluados.
- **Agregar más métricas de evaluación**, como F1-score, precisión, recall o AUC en el caso de clasificación, y MAE o R^2 en el caso de regresión, con el fin de tener una visión más completa del desempeño de los modelos, especialmente en clases minoritarias si las hubiera.
- Explorar otros **modelos de ensemble** como Random Forest, Gradient Boosting tradicional o XGBoost, y comparar sus resultados con Bagging e Histogram Gradient Boosting manteniendo el mismo esquema de partición de datos.
- Aplicar técnicas de **interpretabilidad de modelos** (por ejemplo, importancia de características o métodos como SHAP) para entender mejor qué variables influyen más en la predicción del engagement.
- Diseñar un pequeño **prototipo de servicio o API** que reciba las características de un jugador y devuelva el nivel de engagement predicho, acercando el modelo a un caso de uso real en una plataforma de videojuegos.