

Despliegue de soluciones analíticas

Entrega 2 proyecto

Análisis y predicción de ingresos en el mercado de videos juegos en Stream

**Integrantes Grupo 18**

Sandra Fabiola Rodríguez Blanco

Diego Fernando González Larrote

Víctor Ernesto Velandia Suárez

Luis Álvaro Velásquez

26 de octubre de 2024

1. **Problema abordado y su contexto**

**1.1. Contexto**: El mercado global de videojuegos, con ingresos anuales superiores a $180 mil millones, está en constante evolución. Steam, una de las plataformas más grandes, enfrenta competencia intensa con más de 50,000 títulos y 120 millones de usuarios activos mensuales. Los desarrolladores y editores deben tomar decisiones críticas sobre predicción de ingresos, impacto de reseñas y estrategias de precios para maximizar ingresos y engagement. Este proyecto busca ofrecer una herramienta analítica que permita anticipar el desempeño financiero de un juego, optimizar precios y estrategias de mercado, y reducir riesgos financieros.

**1.2. Pregunta de negocio.** A fin de responder a cuáles son las variables críticas como el precio, tiempo del juego, entre otras, para mejor la toma de decisiones en el lanzamiento de videojuegos al mercado, se plantea la pregunta:

*¿Qué factores influyen de manera dominante en el éxito financiero de un videojuego en Steam?*

**1.3. Alcance.** El objetivo es desarrollar una solución analítica que incluya:

* **Análisis Descriptivo:** Identificar patrones históricos de ventas, precios y calificaciones.
* **Análisis Predictivo:** Crear modelos para predecir ingresos y clasificar juegos según sus calificaciones esperadas.
* **Despliegue de Solución:** Construir un dashboard interactivo que permita explorar resultados, realizar simulaciones y comparar juegos en función de características clave.

1. **Conjunto de datos a emplear**

Este conjunto de datos contiene información completa sobre los 1500 juegos más importantes lanzados en Steam entre el 1 de enero de 2024 y el 9 de septiembre de 2024. Se obtuvo información de 30 archivos separados y se combinó en un único conjunto de datos. Se realizaron pequeños ajustes, como alinear las fechas de lanzamiento de los juegos para mantener la coherencia.

* **nombre**: Representa el título o nombre del juego.
* **releaseDate:** La fecha de lanzamiento oficial del juego, que indica cuándo estuvo disponible para el público.
* **copiasVendidas**: El número total de unidades o copias del juego vendidas.
* **Precio**: El precio minorista original del juego en el momento de su lanzamiento.
* **Ingresos:** La cantidad de dinero generada por los ingresos del juego.
* **avgPlaytime**: la duración promedio que los jugadores han pasado jugando el juego.
* **reviewScore**: La puntuación o calificación otorgada al juego según las reseñas de los usuarios y los críticos de los video juegos.
* **publisherClass**: La clasificación del editor, que indica si el editor es AAA, AA o independiente.
* **editores:** El nombre de la empresa responsable de publicar el juego.
* **desarrolladores**: el nombre del equipo de desarrollo o la empresa que creó el juego.
* **steamId:** un identificador único asignado al juego por Steam para seguimiento y gestión.

**Referencia:** estos datos y su descripción son tomados de la siguiente dirección URL al dataset[[1]](#footnote-2).

1. **Repositorio Git en uso para el código**

La evidencia del código fuente y el dataset “Steam\_2024\_bestRevenue\_1500.csv” se ilustran en la siguiente imagen dentro con el usuario GitHub la.velasquez55:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 1. Archivo de código cargado en GitHub

La siguiente imagen ilustra un extracto del código dentro del repositorio[[2]](#footnote-3) GitHub para para el desarrollo del proyecto.

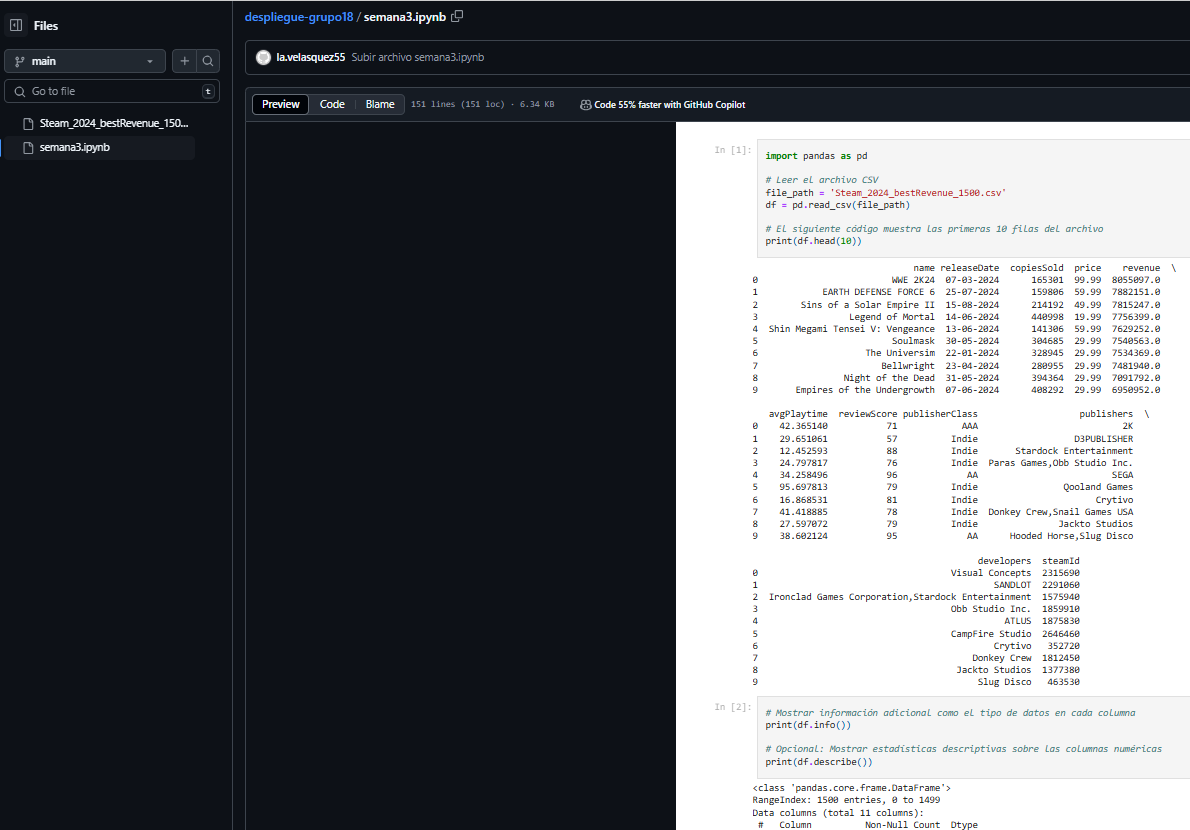
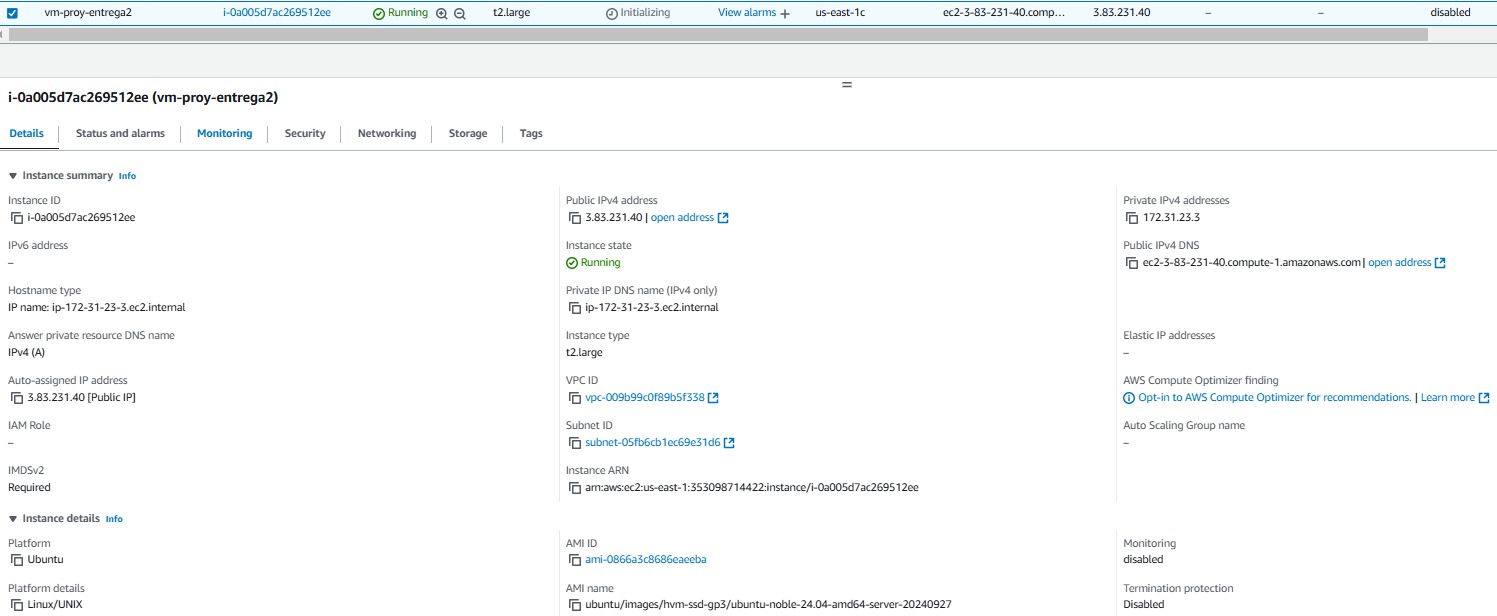


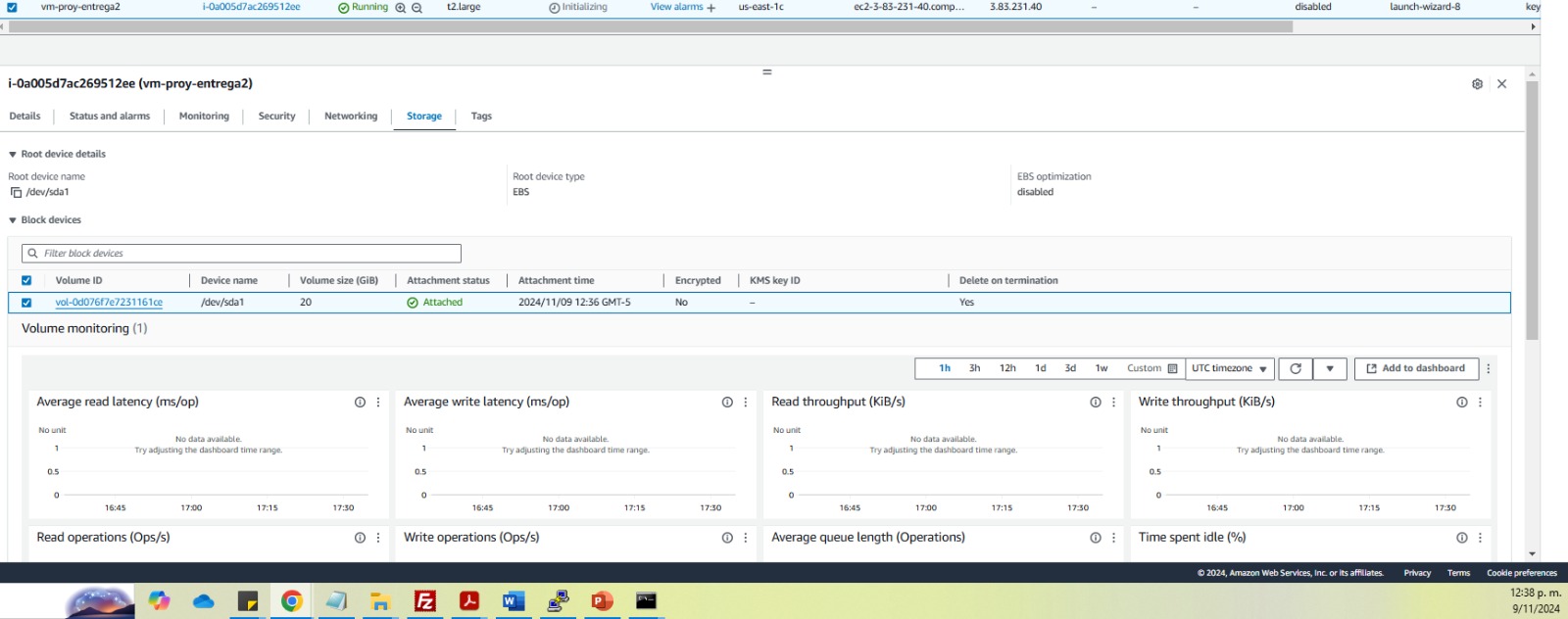
Ilustración 2. Extracto código fuente en GitHub

1. **Instalación y prueba de MLFlow en la máquina virtual**

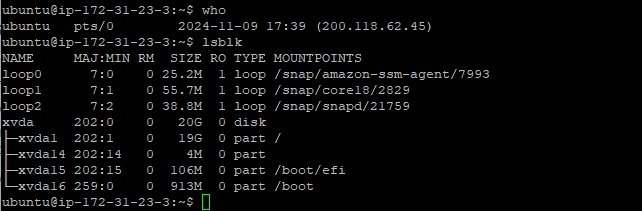
**4.1. Pantallazo de la consola AWS EC2 con la máquina en ejecución**

A continuación se presenta la instancia creada en ejecución:

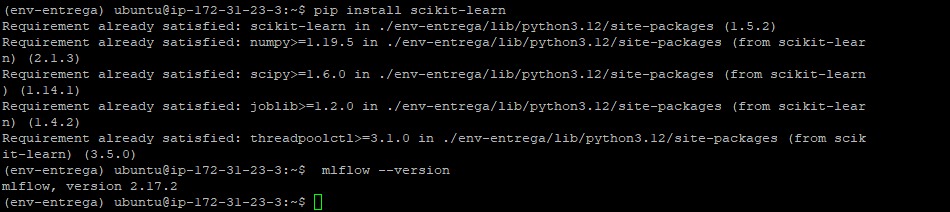




**4.2. Conexión a la máquina virtual**

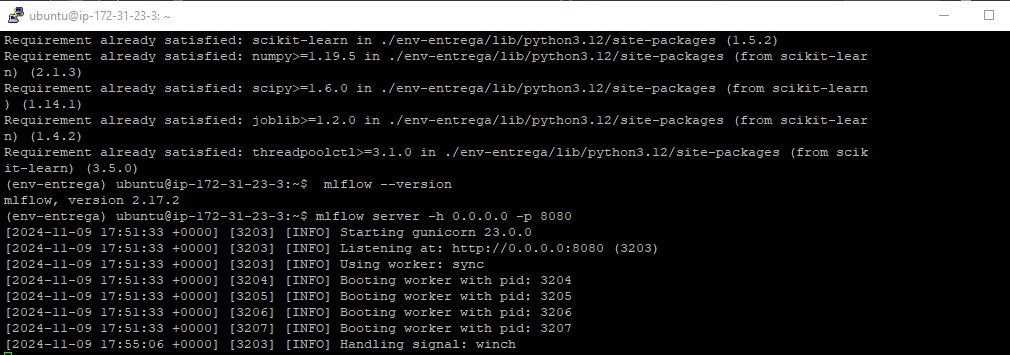
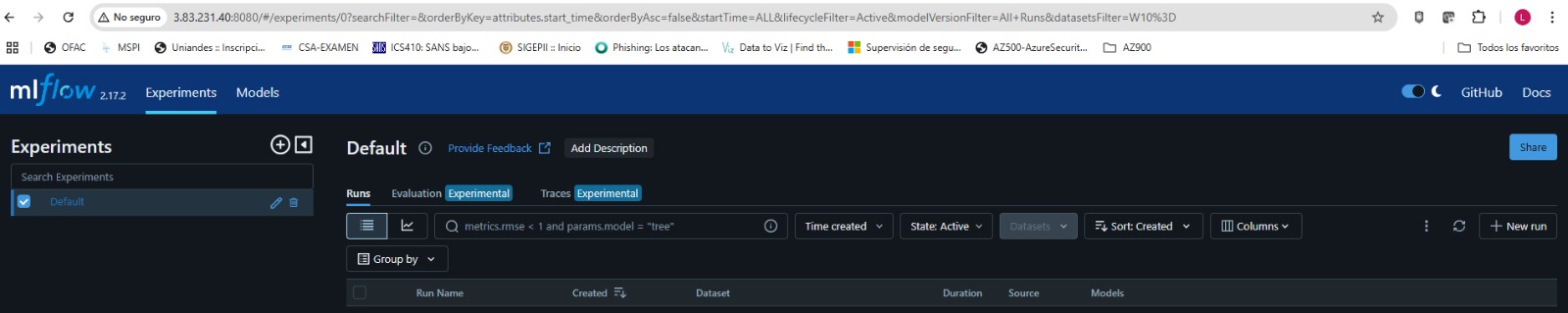


**4.3. Instalación Scikit-learn y ML-Flow**



**4.4. MLFLOW Server Operativo**

**Disponibilización del servido MLFLOW**

**Acceso desde internet al servidor MLFLOW**

1. **Desarrollo y evaluación del modelo**

**5.1. Modelo inicial Random Forest**

El modelo desarrollado utiliza un **Regresor de Bosques Aleatorios** (Random Forest Regressor), una técnica de aprendizaje supervisado que combina múltiples árboles de decisión para mejorar la precisión y reducir el sobreajuste. Este modelo es particularmente útil para capturar relaciones no lineales y complejas entre las variables.

**Metodología**

1. **Preparación de Datos**:
   * Se eliminaron columnas categóricas no necesarias (name, releaseDate, publishers, developers, steamId).
   * Se aplicó una transformación logarítmica a copiesSold para manejar alta variabilidad.
   * **One-Hot Encoding**: La columna publisherClass fue codificada para permitir su uso en el modelo.
   * Se definieron las características (X) y la variable objetivo (revenue).
2. **División de Datos**:

Los datos se dividieron en conjuntos de entrenamiento y prueba usando train\_test\_split.

1. **Definición del Modelo**:
   * + Se configuraron los hiperparámetros iniciales del Random Forest: n\_estimators, max\_depth, min\_samples\_leaf, min\_samples\_split y max\_features.
2. **Entrenamiento del Modelo**:

Se entrenó el modelo con el conjunto de entrenamiento y se evaluó con el conjunto de prueba.

**Evaluación del Modelo**: La métrica principal utilizada para evaluar y comparar los modelos es **Error Cuadrático Medio** (MSE).

1. **Versionamiento con MLflow**:
   * Se utilizaron las funcionalidades de **MLflow** para:
     + Registrar los hiperparámetros (n\_estimators, max\_depth, max\_features).
     + Almacenar el modelo entrenado.
     + Registrar métricas y realizar un seguimiento de los experimentos.
   1. **Ejecución de experimentos del modelo**

Se realizó la ejecución de los experimentos del modelo:

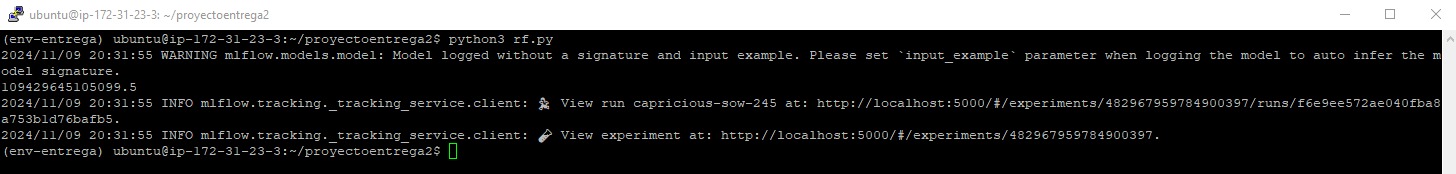
****

Ilustración 3 Ejecución del modelo

**Comparación de modelos ejecutados:**

Los experimentos fueron registrados en **MLflow** para realizar un análisis detallado de los hiperparámetros y su impacto en la métrica de error cuadrático medio (**MSE**). Los hallazgos clave de los 4 experimentos son:

**1. Observaciones Clave**

**Hiperparámetros Evaluados:**

* **maxdepth**: 25, 10, 10, 20
* **min\_samples\_leaf**: 5, 4, 4, 3
* **num\_trees**: 350, 500, 300, 300
* **min\_samples\_split**: Fijo en 2 para todos los experimentos.

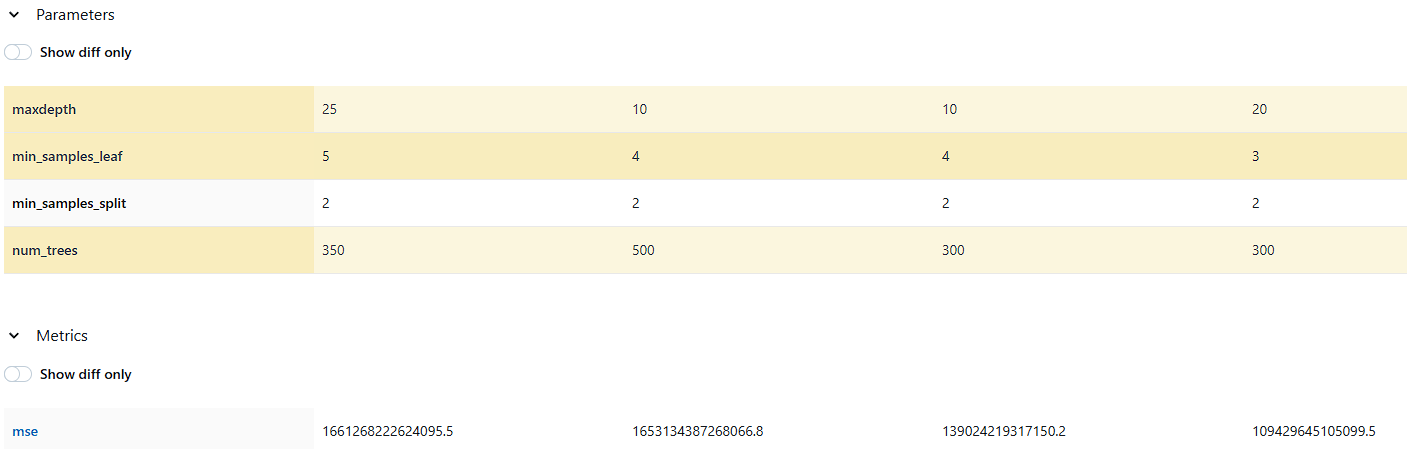
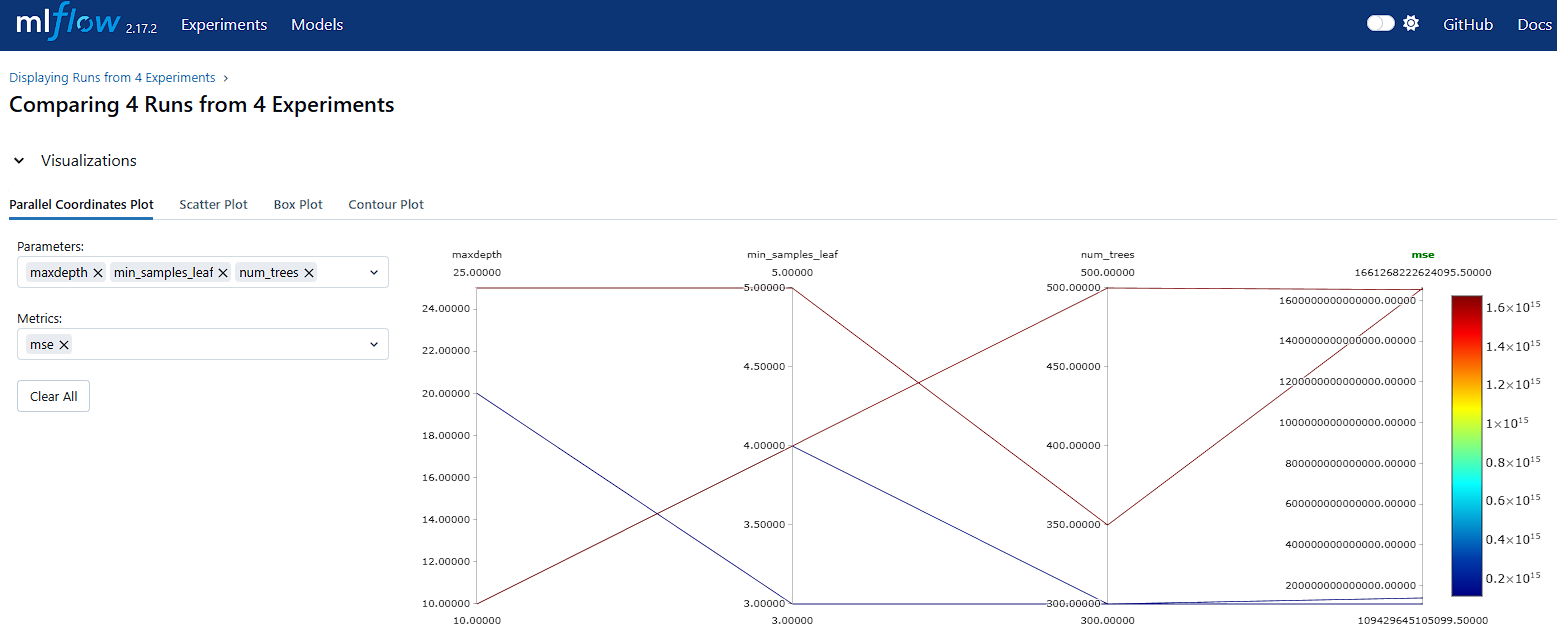
**Resultados Esperados (MSE):**

1. **Experimentación con maxdepth**:
   * La profundidad máxima del árbol es un factor importante. Los valores más altos (20-25) tienden a mejorar el ajuste, pero con resultados marginales.
   * **maxdepth=25** no mejora sustancialmente el MSE en relación con **maxdepth=20** en algunos casos.
2. **Impacto de min\_samples\_leaf**:
   * Valores más bajos (3-4) para este hiperparámetro resultan en mejores predicciones, ya que permiten al modelo ajustar mejor los datos.
   * **min\_samples\_leaf=5** mostró un aumento en el MSE, lo que indica que este valor reduce la capacidad del modelo para capturar la complejidad.
3. **Efecto de num\_trees**:
   * Se observa una ligera mejora en el MSE al incrementar el número de árboles de 300 a 500, pero la diferencia es menor en comparación con los ajustes realizados en otros hiperparámetros.
   * Los beneficios se estabilizan más allá de 350 árboles.

**Visualización de Coordenadas Paralelas:**

El gráfico muestra claramente cómo los cambios en los parámetros afectan el MSE:

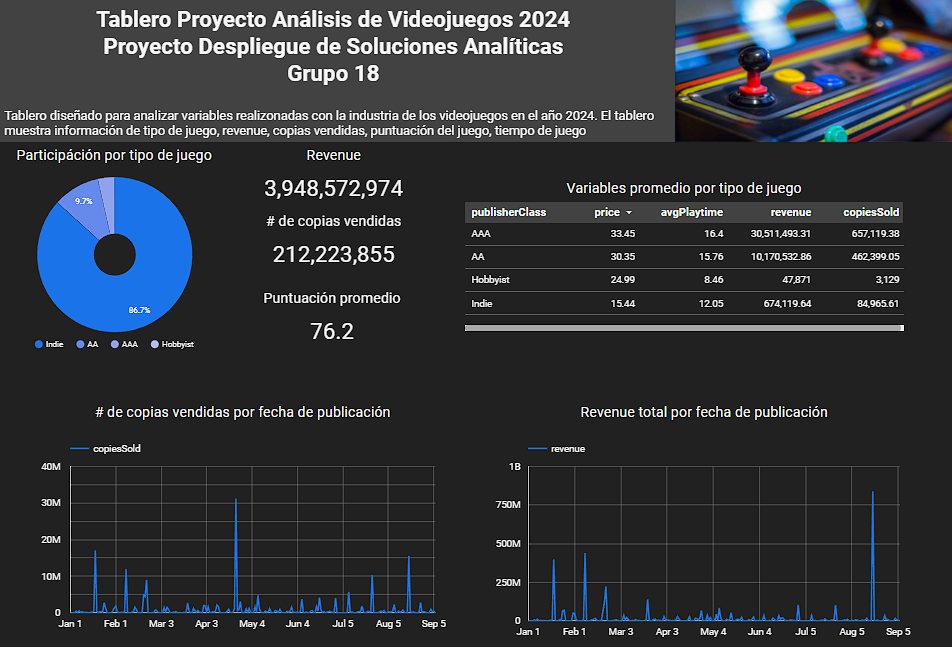
* **La combinación más óptima de parámetros** en estos experimentos es **maxdepth=20**, **min\_samples\_leaf=3**, y **num\_trees=350**, lo que produce el MSE más bajo.



**2. Conclusiones**

1. **Configuración Óptima Identificada**:
   * Basado en los experimentos, la mejor configuración observada fue:
     + maxdepth=20
     + min\_samples\_leaf=3
     + num\_trees=350
   * Esta configuración proporciona un buen balance entre precisión del modelo y tiempo de entrenamiento.
2. **Trade-off entre Complejidad y Precisión**:
   * Aumentar maxdepth más allá de 20 o incrementar num\_trees a 500 no aporta mejoras significativas en el MSE, sugiriendo que hay un punto de retorno decreciente en la complejidad del modelo.
3. **Siguientes Pasos**:
   * Se recomienda explorar ajustes adicionales a **min\_samples\_split** o utilizar herramientas como **Optuna** para un enfoque más exhaustivo de búsqueda de hiperparámetros.
   * Validar esta configuración óptima con un conjunto de validación cruzada para evaluar la estabilidad del modelo.
4. **Descripción del Tablero de Análisis de Videojuegos 2024**

El tablero desarrollado proporciona una visión integral del desempeño de los videojuegos más relevantes en Steam durante 2024. Está diseñado para ayudar a desarrolladores, editores y analistas a comprender los factores que influyen en el éxito financiero de los juegos. Este tablero combina métricas descriptivas y visualizaciones interactivas para facilitar la toma de decisiones estratégicas. La ruta de acceso al tablero es la siguiente: <https://lookerstudio.google.com/reporting/9021de7a-9788-48b0-887d-b931d84a92dd>



**Componentes del Tablero**

1. **Título y Contexto**:
   * El encabezado incluye el nombre del proyecto y una breve descripción del propósito del tablero.
   * Se especifica que el tablero se centra en variables críticas como ingresos, copias vendidas, puntuación promedio y tiempo de juego.
2. **Indicadores Clave de Desempeño (KPIs)**:
   * **Revenue Total**: Muestra los ingresos generados por todos los juegos en 2024.
   * **Número de Copias Vendidas**: Total de unidades vendidas durante el periodo analizado.
   * **Puntuación Promedio**: Promedio de calificaciones recibidas por los juegos.
3. **Gráfico de Participación por Tipo de Juego**:
   * Un gráfico circular que muestra la distribución de ingresos por categorías de juegos (AAA, AA, Indie, Hobbyist).
   * Ofrece una visión clara de cómo cada tipo de publicador contribuye al total de ingresos.
4. **Tabla de Variables Promedio por Tipo de Juego**:
   * Muestra valores promedio de variables como price, avgPlaytime, revenue, y copiesSold desglosados por la categoría del publicador.
   * Ayuda a identificar patrones de comportamiento según el tipo de juego.
5. **Gráfico de Copias Vendidas por Fecha de Publicación**:
   * Un gráfico de líneas que muestra el número de copias vendidas a lo largo del tiempo, permitiendo analizar picos en ventas y su relación con fechas clave.
6. **Revenue Total por Fecha de Publicación**:
   * Otro gráfico de líneas que detalla la evolución de los ingresos acumulados por fecha de publicación.
   * Útil para identificar tendencias estacionales o el impacto de eventos promocionales.

**Funcionalidades del Tablero**

* **Visualización Comparativa**:
  + Permite comparar el desempeño de diferentes categorías de juegos y analizar la relación entre precio, tiempo promedio de juego e ingresos.
* **Análisis Temporal**:
  + Los gráficos de líneas ayudan a entender la evolución de ventas e ingresos, permitiendo identificar patrones temporales y fechas de mayor impacto.
* **Toma de Decisiones Basada en Datos**:
  + Los usuarios pueden utilizar esta información para ajustar estrategias de precio, evaluar el impacto de promociones, y optimizar el portafolio de productos.
* **Accesibilidad e Interactividad**:
  + Diseñado con un enfoque visual claro, el tablero facilita la exploración interactiva de los datos para usuarios con diferentes niveles de experiencia analítica.

1. **Reporte de trabajo en equipo**

En la siguiente tabla se relacionan las actividades generales desarrolladas por los integrantes del Grupo 18.

|  |  |  |
| --- | --- | --- |
| **No.** | **Integrante** | **Actividades general** |
| 1 | Sandra Fabiola Rodríguez Blanco | Convocar a las reuniones de equipo para trabajar en el proyecto.  Revisar el dataset y explorar las variables de interés para el análisis de los datos. |
| 2 | Diego Fernando González Larrote | Ajuste de tablero para presentación  Ajuste de código para despliegue de experimento en MLFlow  Revisión de modelos entrenados |
| 3 | Víctor Ernesto Velandia Suárez | Convocar a las reuniones de equipo para trabajar en el proyecto.  Consolidación del reporte con los entregables y análisis de resultados  Trabajar el entorno local para el desarrollo del código del proyecto. |
| 4 | Luis Álvaro Velásquez | Convocar a las reuniones de equipo para trabajar en el proyecto.  Trabajar en el repositorio local y remoto para el desarrollo del código del proyecto.  Solucionar problemas de librerías en el entorno de trabajo Python. |

Tabla 1. Actividades generales equipo de trabajo

1. Fuente de los datos: https://www.kaggle.com/datasets/alicemtopcu/top-1500-games-on-steam-by-revenue-09-09-2024?resource=download [↑](#footnote-ref-2)
2. https://github.com/lavelasquez55/[proyectoentrega2](https://github.com/lavelasquez55/proyectoentrega2) [↑](#footnote-ref-3)