Proyecto final Data Science

Darío Cancelli

Temario

- Metas y audiencia:
- Objetivos:
- Análisis exploratorios de los datos:
- Modelos de Machine Learning
- Insights

Metas y audiencia

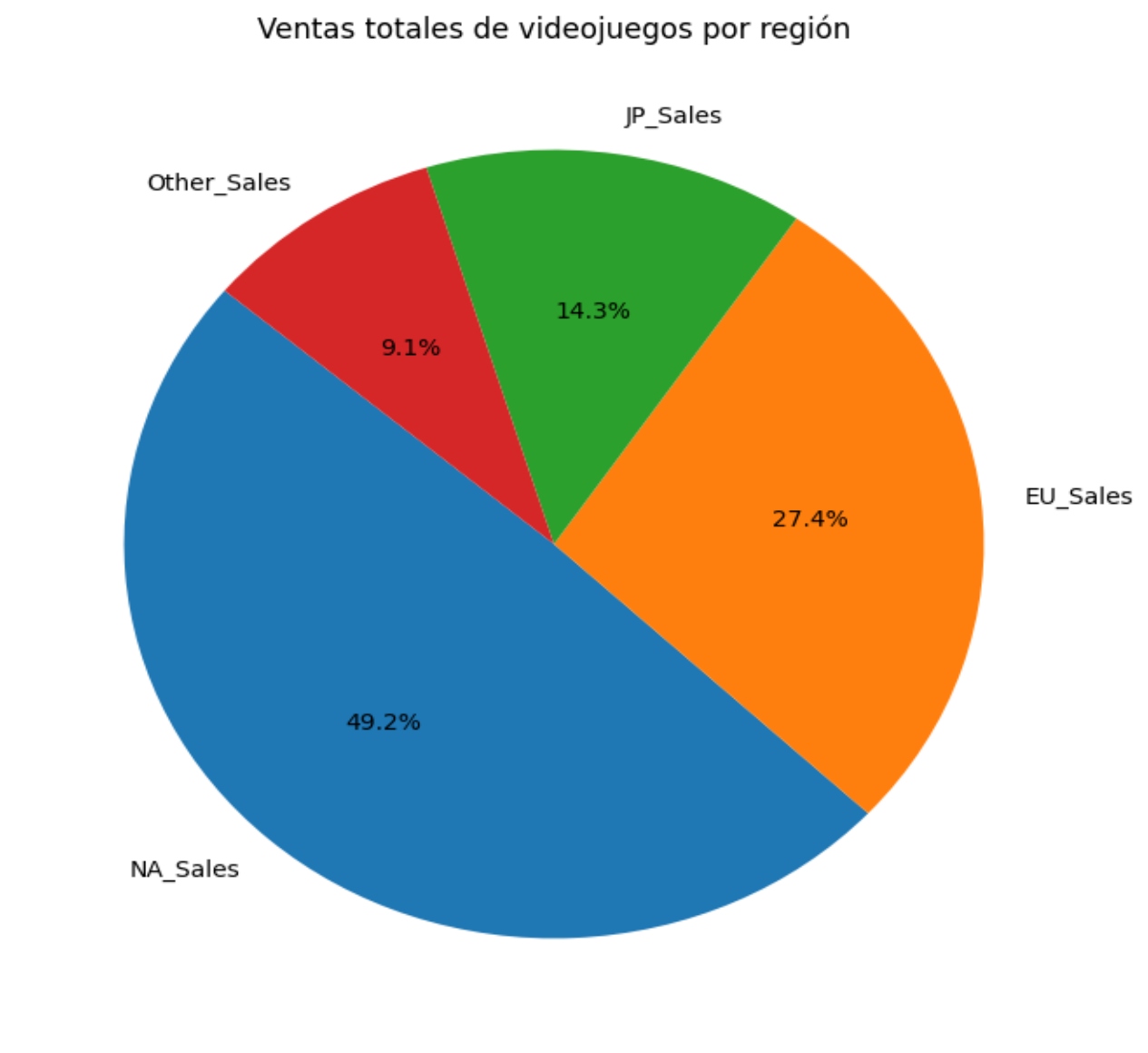
Lo que me llevó a la elección de un dataset que trate sobre videojuegos se debe a primera instancia a mi interés hacia los mismos. Al ser un apasionado, trabajar con datos relacionados con ellos hizo que pudiera combinar mi interés personal con mi trabajo, lo que hizo que el proceso sea mucho más ameno y disfrutable. Otro punto es la variedad, ya que el dataset cuenta con una gran cantidad de datos, que van desde información sobre los jugadores, hasta año de lanzamiento, plataforma, entre otros. Esto me proporciono una amplia gama de posibilidades para el análisis y extracción de información útil. Además, como la industria de los videojuegos está en constante crecimiento y es una de las más grandes del mundo del entretenimiento, trabajar con datos de videojuegos me proporciono una perspectiva única sobre esta industria en evolución.

Este proyecto fue planeado para una amplia gama de audiencias, tanto como para aquellos individuos que deseen explorar y realizar conclusiones para proyectos futuros, como a empresas interesadas en comprender en profundidad como se lleva a cabo un proyecto de Ciencia de Datos.

Objetivos:

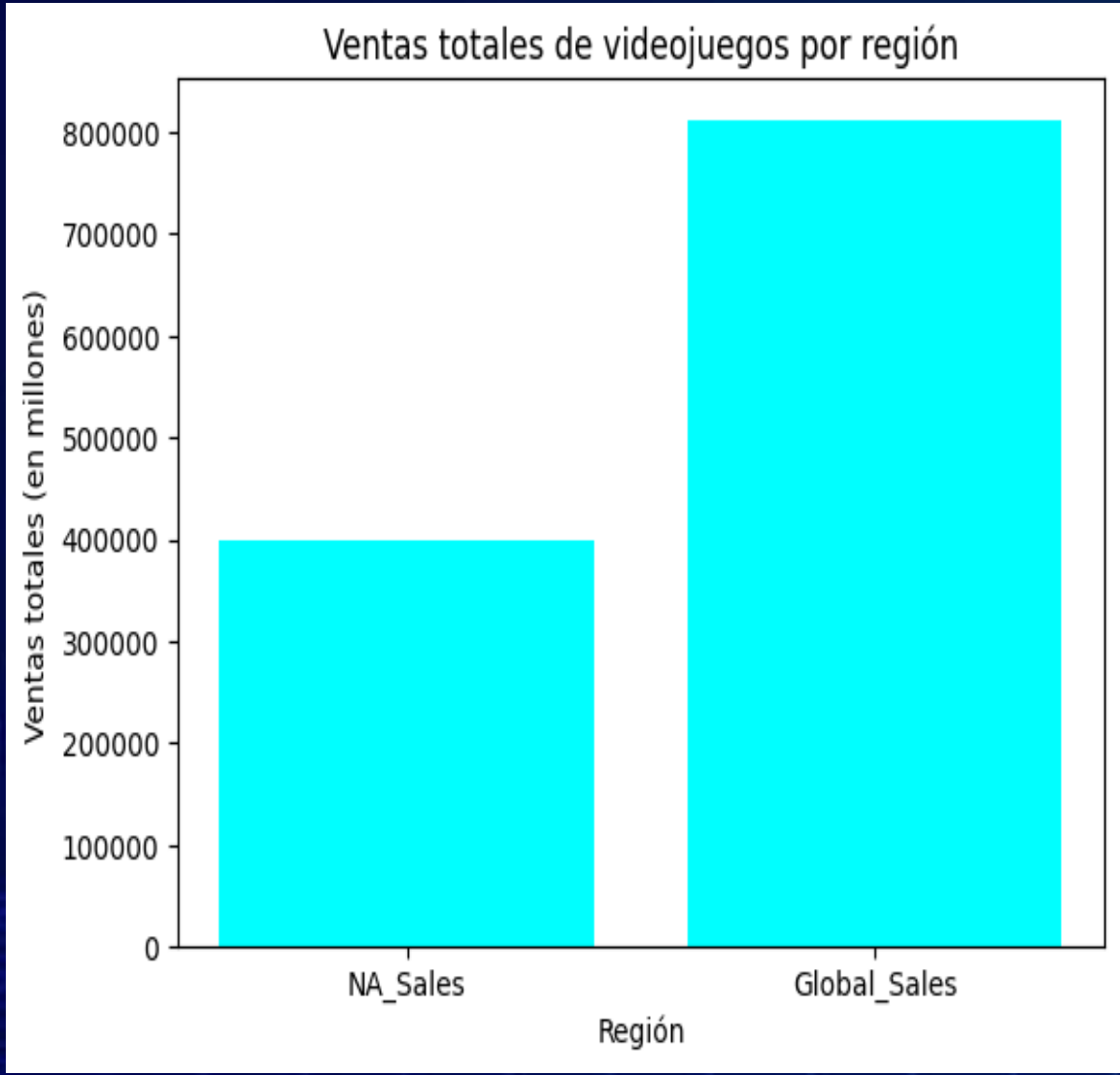
Se aplicarán dos algoritmos, uno que pueda predecir las ventas globales de un juego basado en diferentes características, ya que sería una herramienta valiosa en la toma de decisiones estratégicas dentro de esta industria, ayudando a mejorar la eficiencia y el rendimiento comercial. También un algoritmo que identificara grupos de videojuegos con ventas similares, lo que le permitiría a la empresa comprender mejor el comportamiento del mercado global, mejorando la adaptación de estrategias de marketing y desarrollo de juegos.

Análisis exploratorio de los datos



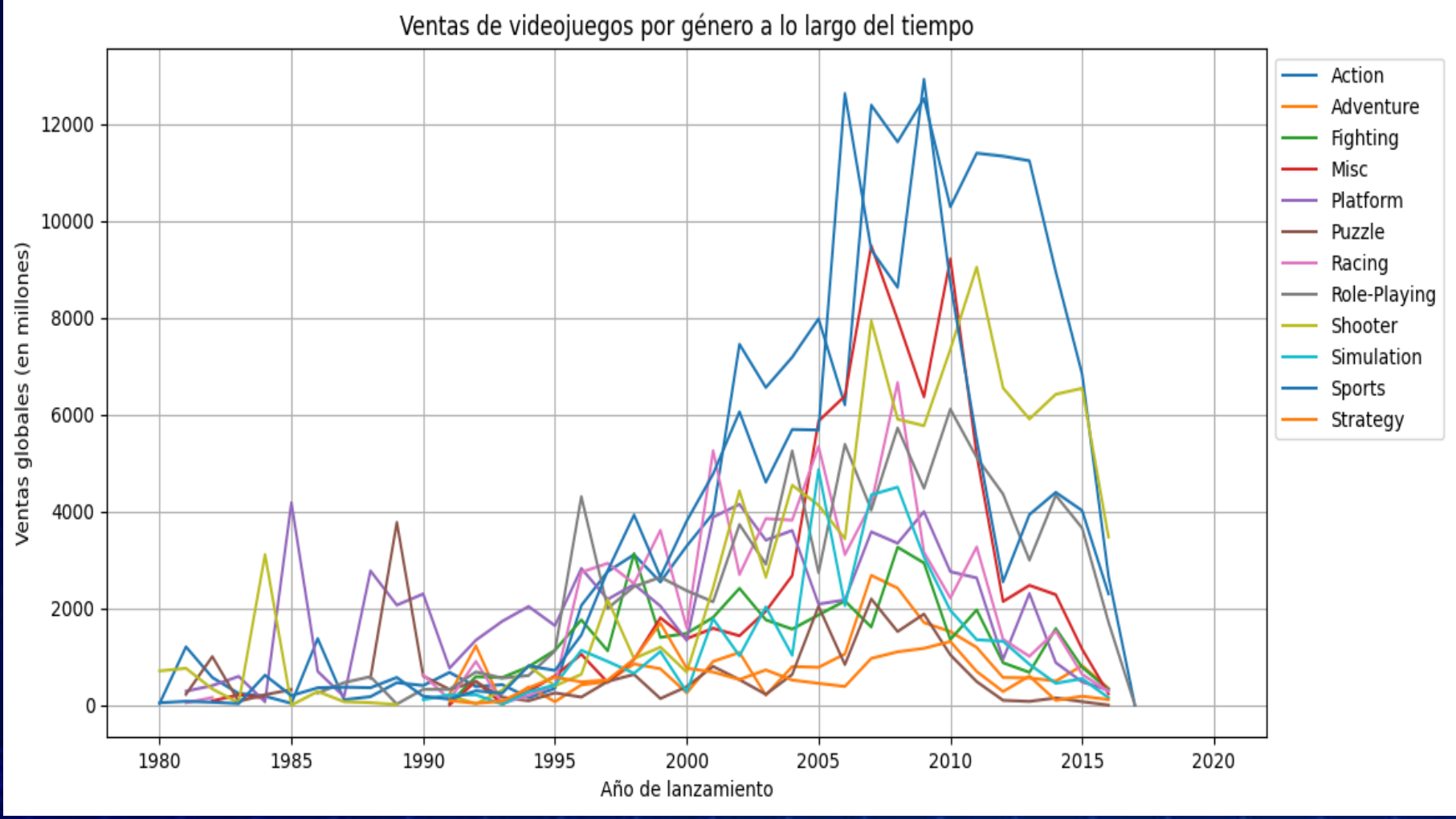
Se observa que “NA_Sales” (Norte América) cuenta con una mayor cantidad de ventas a nivel regional.

Análisis exploratorio de los datos



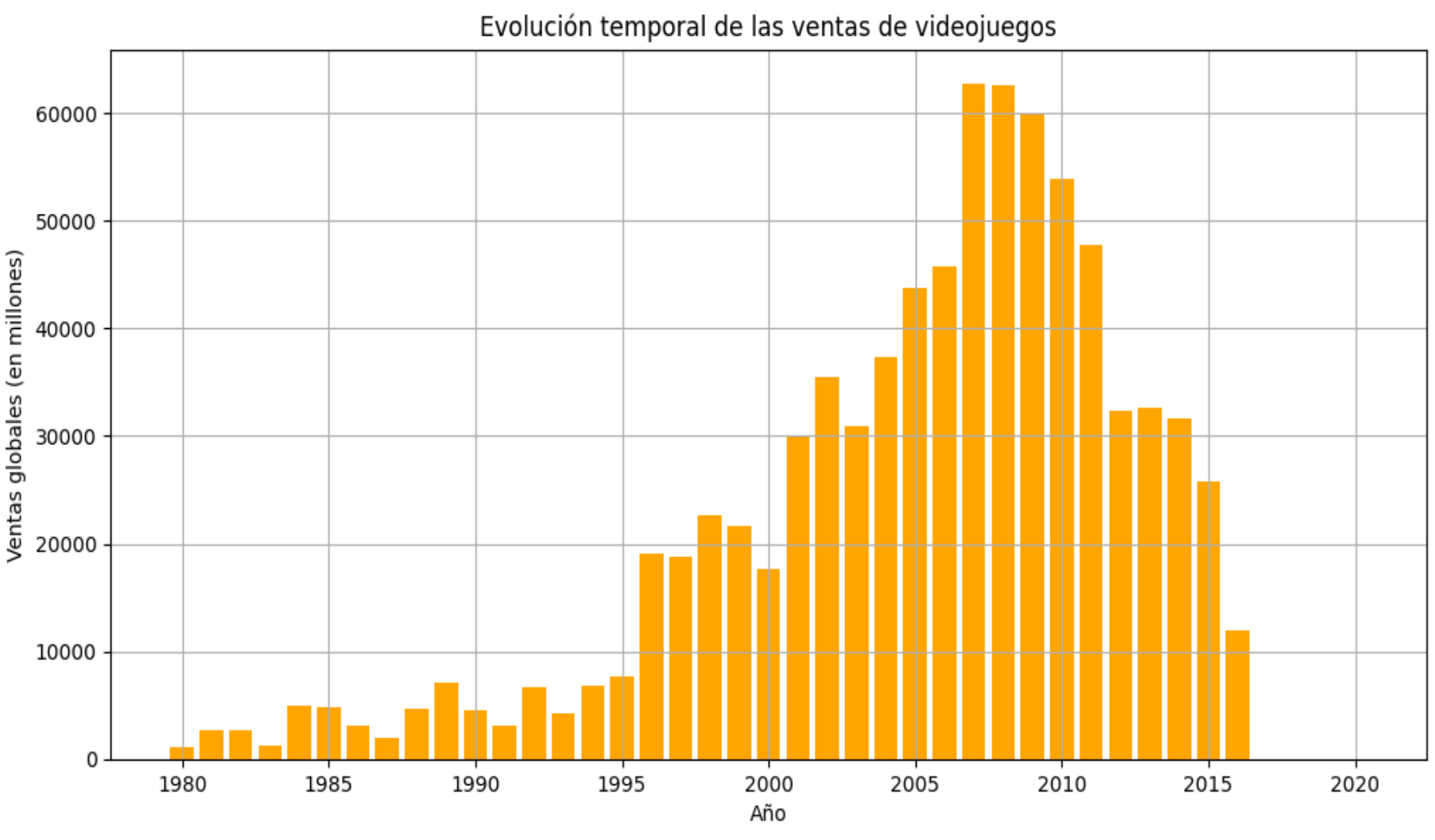
Las ventas en Norte América son las más grandes a nivel regional y son el 50% de las ventas a nivel mundial. Esto puede deberse a diversos factores, como por ejemplo el tamaño del mercado, la industria desarrollada de los videojuegos y la cultura que existe de los mismos en el país, tambien hay que tener en cuenta las tendencias del mercado y las estrategias de marketing especificas que se utilizan en esta región.

Análisis exploratorio de los datos



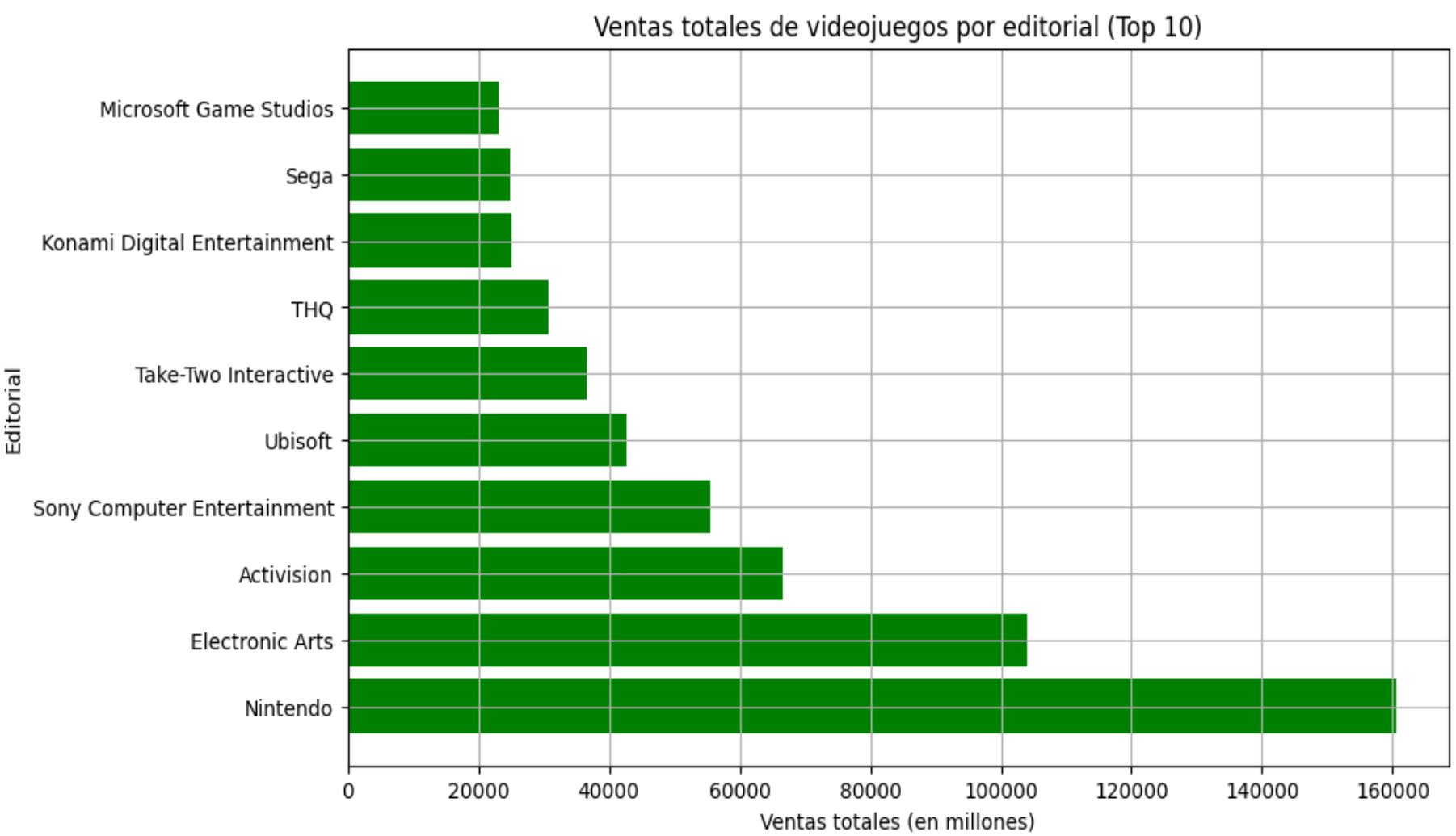
Se observa que a medida que pasan los años algunos géneros de videojuegos experimentaron picos de popularidad, tambien parece ser que algunos géneros tienden a tener un comportamiento cíclico, suben sus ventas y luego bajan y se repite. Por ultimo los picos de popularidad pueden deberse a el lanzamiento de una nueva consola o a la popularidad generada por el marketing y los medios de comunicación.

Análisis exploratorio de los datos



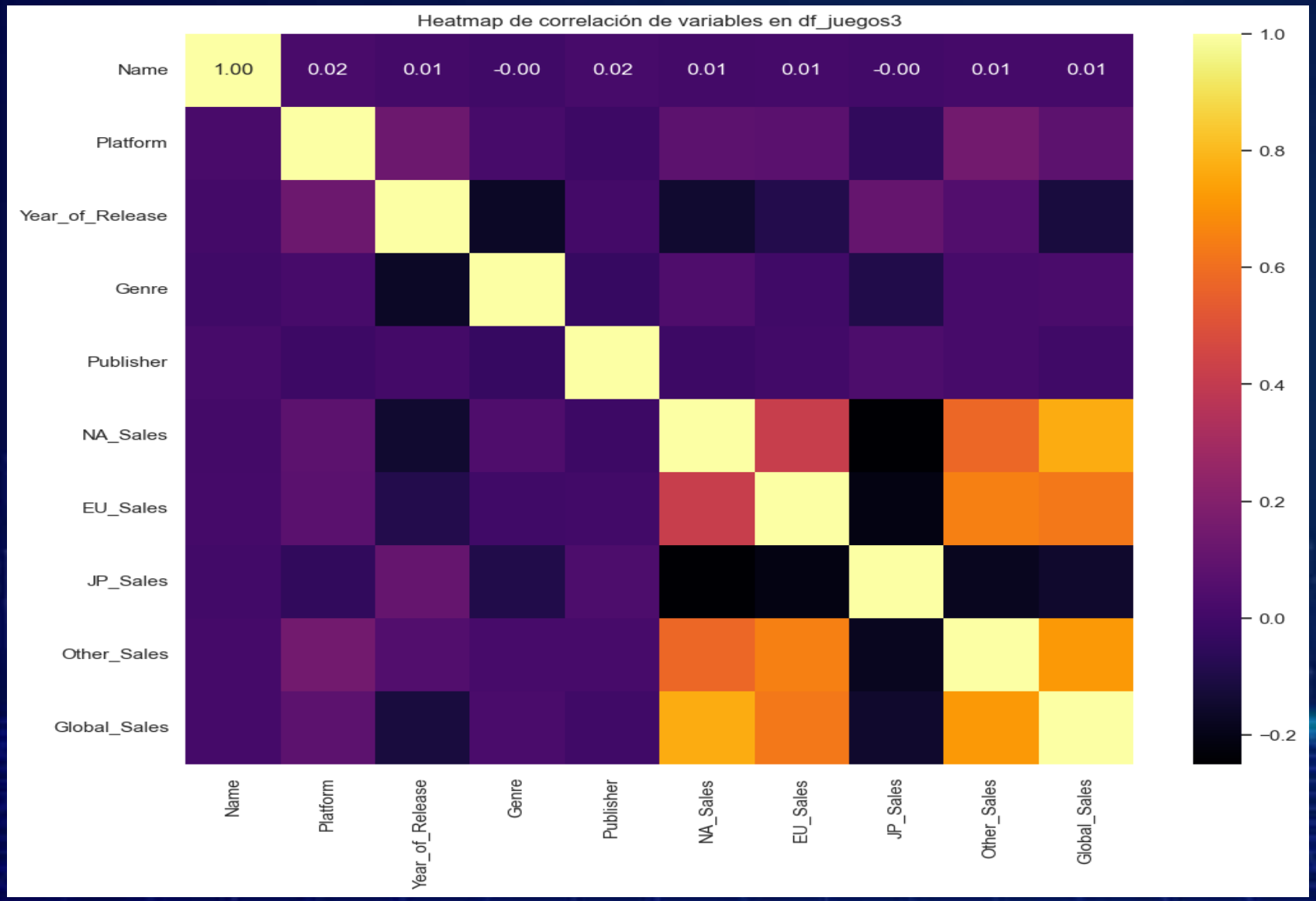
Se observa que los años en los que se lanzan nuevas consolas o se introducen avances tecnológicos importantes pueden estar asociados con picos en las ventas de videojuegos, además es posible que los acuerdos de exclusividad con determinadas plataformas o títulos exclusivos pueden tener un impacto significativo en las ventas de videojuegos en ciertos años, lo que puede generar un mayor interés por parte de los consumidores.

Análisis exploratorio de los datos



Realizando este top 10 concluí que un pequeño grupo de editoriales concentra la mayor parte de las ventas totales de videojuegos. Esto sugiere que algunas empresas tienen una posición dominante en la industria y ejercen una influencia significativa sobre el mercado, estas suelen ser aquellas con larga trayectoria en el mercado, lo que podría indicar que la reputación y el prestigio sugiere éxito comercial.

Análisis exploratorio de los datos

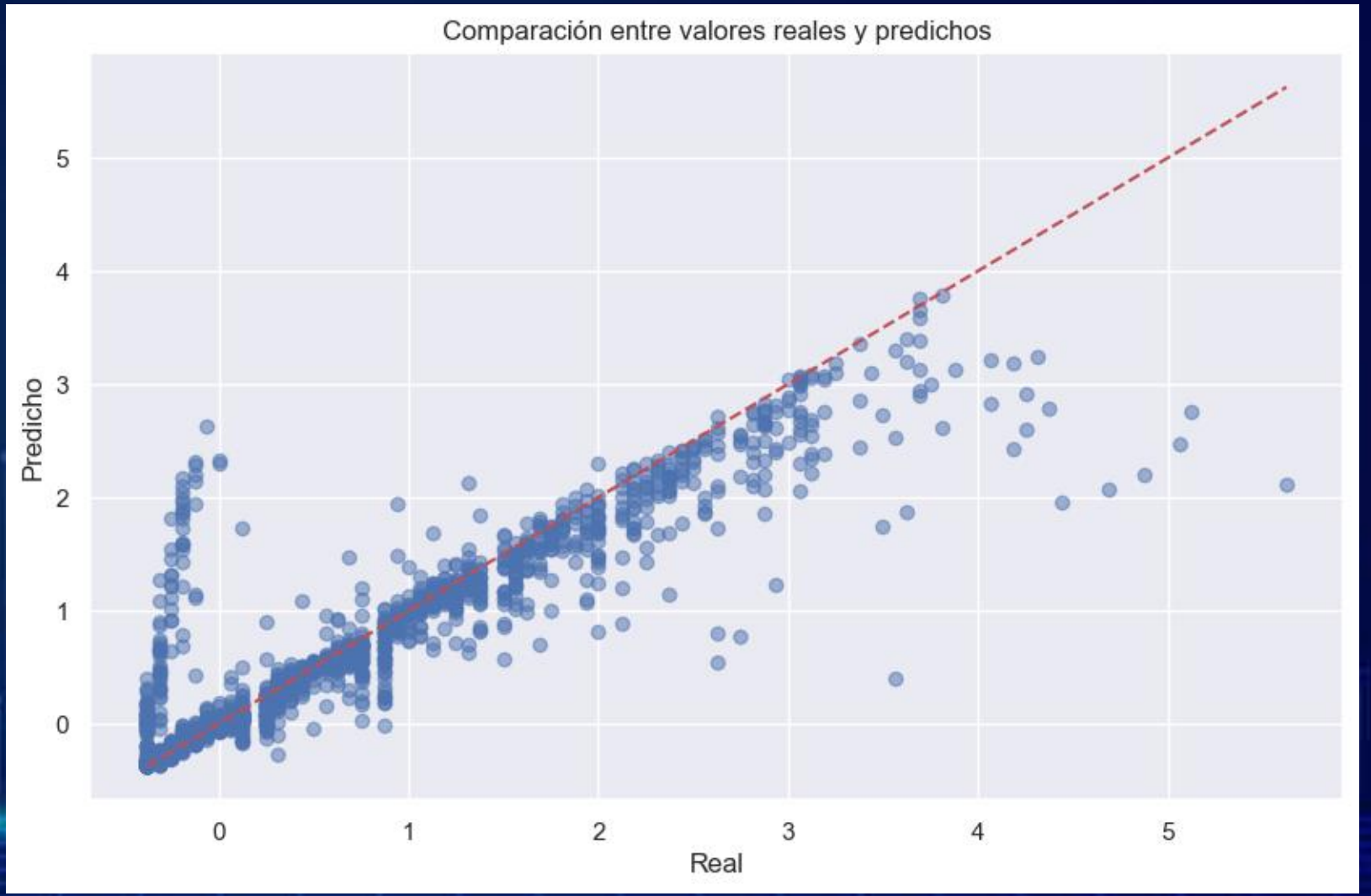


Se puede apreciar que la mayoría de variables correlacionan poco entre sí. En proporción hay muchas más correlaciones negativas que positivas, eso quiere decir que a medida que una variable crece la otra decrece, con un análisis detallado se podría saber el porqué de estos sucesos y tomar acciones para el desarrollo de la empresa.

Modelos de machine learning

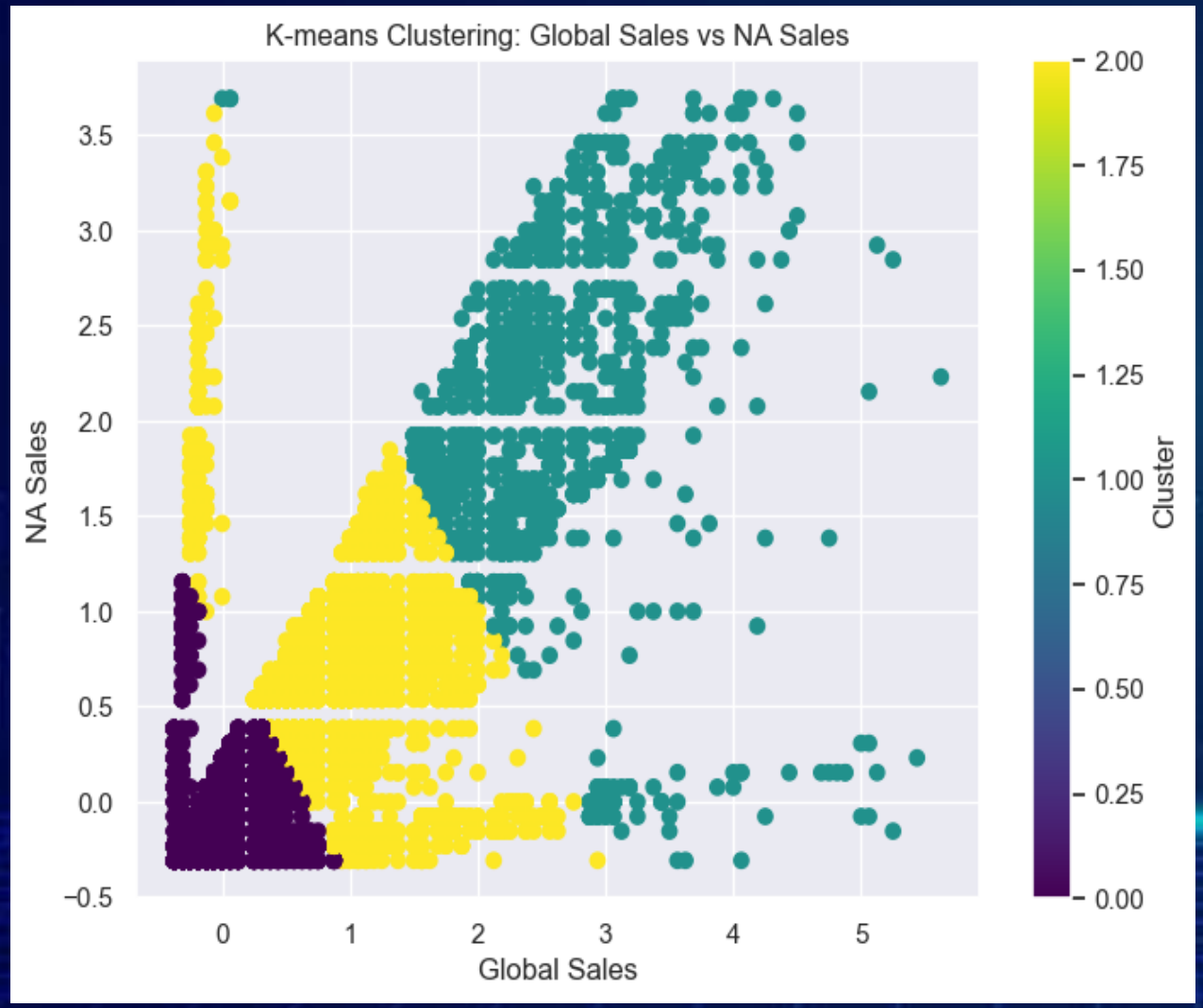
Para el primer algoritmo se realizó una regresión logística y un regresor de bosques aleatorios dando el segundo los mejores parámetros

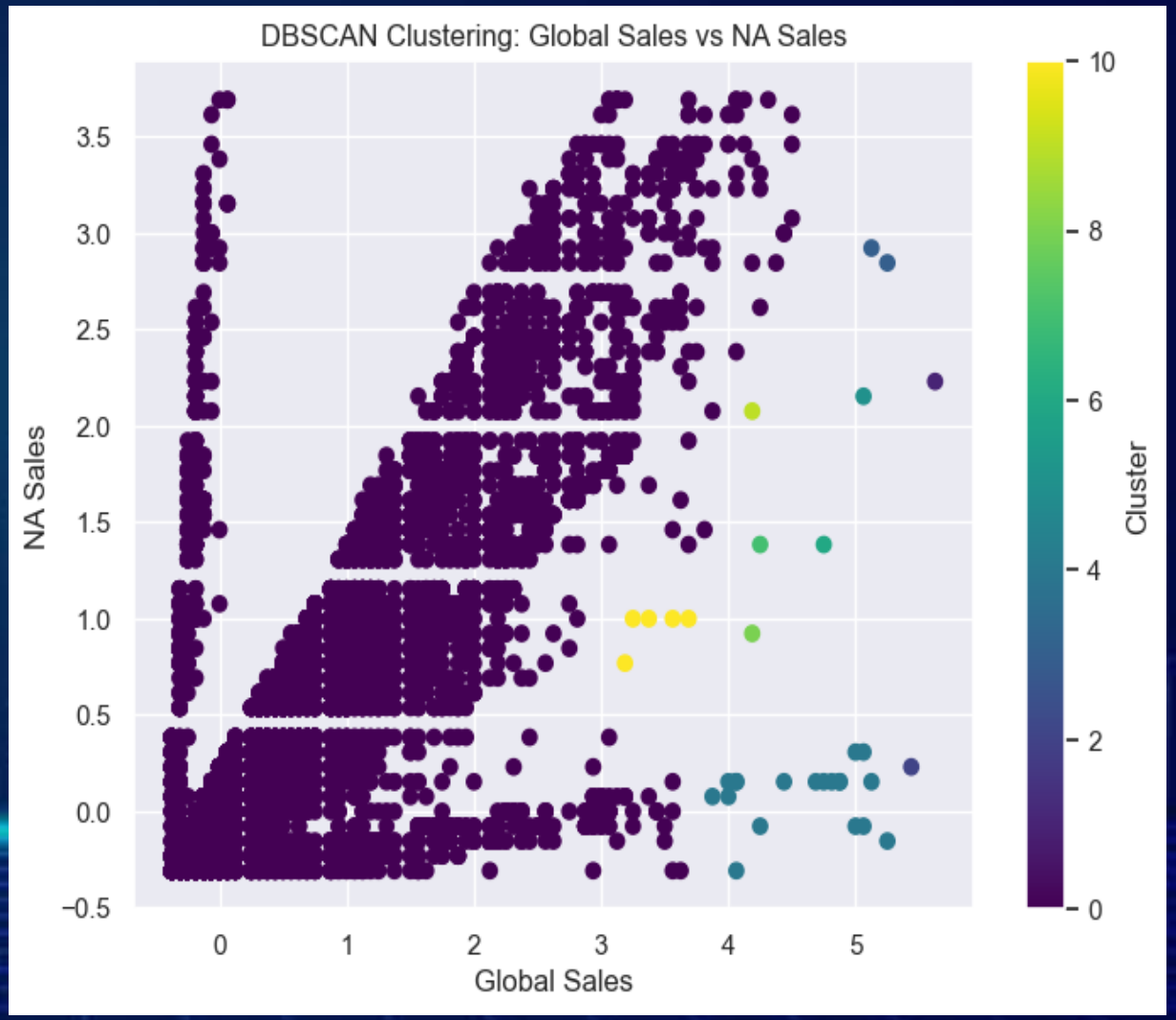
- Score: 0.8423543740274071
- Mean Squared Error (MSE): 0.14714291748339647
- Mean Absolute Error (MAE): 0.15847859145614046
- R^2: 0.8423543740274071
- Adjusted R^2: 0.8417358863928154



Modelos de machine learning

Para el segundo algoritmo utilice k-means y DBSCAN, siendo k-means el que dio mejores resultados.





Conclusiones

Primer algoritmo: El score explica aproximadamente el 84.3% de la variabilidad en los datos de salida utilizando las características proporcionadas. Esto sugiere que el modelo tiene un buen ajuste y puede capturar la relación entre las variables de entrada y de salida de manera significativa. El MSE es relativamente bajo, lo que indica que las predicciones del modelo tienen errores cuadráticos medios pequeños en comparación con la varianza total en los datos de salida. Esto sugiere que el modelo tiene una precisión razonable en la predicción de los valores de salida. El MAE también es bajo, lo que indica que las predicciones del modelo tienen desviaciones medias pequeñas de los valores reales. Esto refuerza la idea de que el modelo tiene una precisión aceptable en sus predicciones. Por último, el R cuadrado y el R cuadrado ajustado son bastante similares, lo que sugiere que el ajuste del modelo no está siendo penalizado significativamente por el número de características incluidas. Esto indica que el modelo no está haciendo overfiting y que la inclusión de las características actuales en el modelo es razonable.

Segundo algoritmo: El primer algoritmo k-means, performo mejor que el dbscan debido a que se puede examinar visualmente cómo se agrupan los datos y qué tan separados están los clusters entre sí. Esto puede proporcionar información sobre la distribución de las ventas y posiblemente identificar patrones o segmentos de mercado, además la visualización permite evaluar rápidamente la calidad del agrupamiento y si los clusters parecen significativos o no.