RESULTADOS GENERALES MODELO DE PREDICCIÓN DE VENTAS

ANDREA ARBOLEDA TOBON

OBJETIVO

Desarrollar un modelo que permita predecir el comportamiento de venta de artículos en determinadas tiendas mediante el análisis de los datos y aplicación de modelos predictivos como el de regresión lineal y arboles de decisión.

METODOLOGÍA

Examinación de los datos – Limpieza – análisis exploratorio – balanceo de datos – aplicación de modelos — entrenamiento y predicción – métricas y rendimiento - análisis de resultados.

DATA

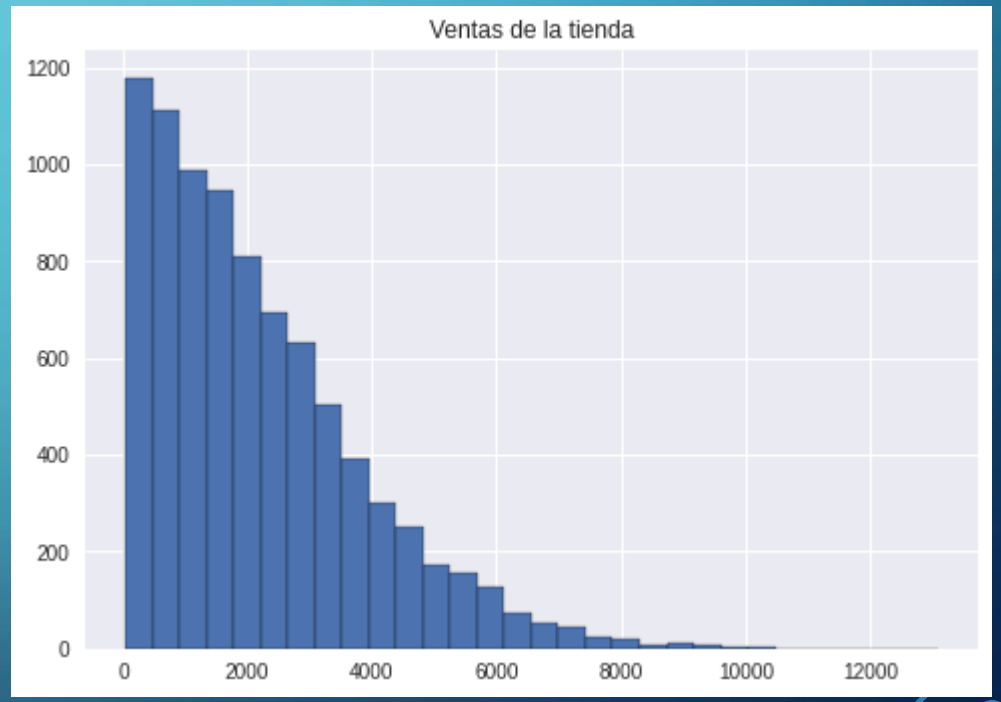
- Esta es la fuente de donde se descargaron los datos:

https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/

Se trata de un problema práctico de ventas de supermercado, con 12 columnas y 8.523 filas.

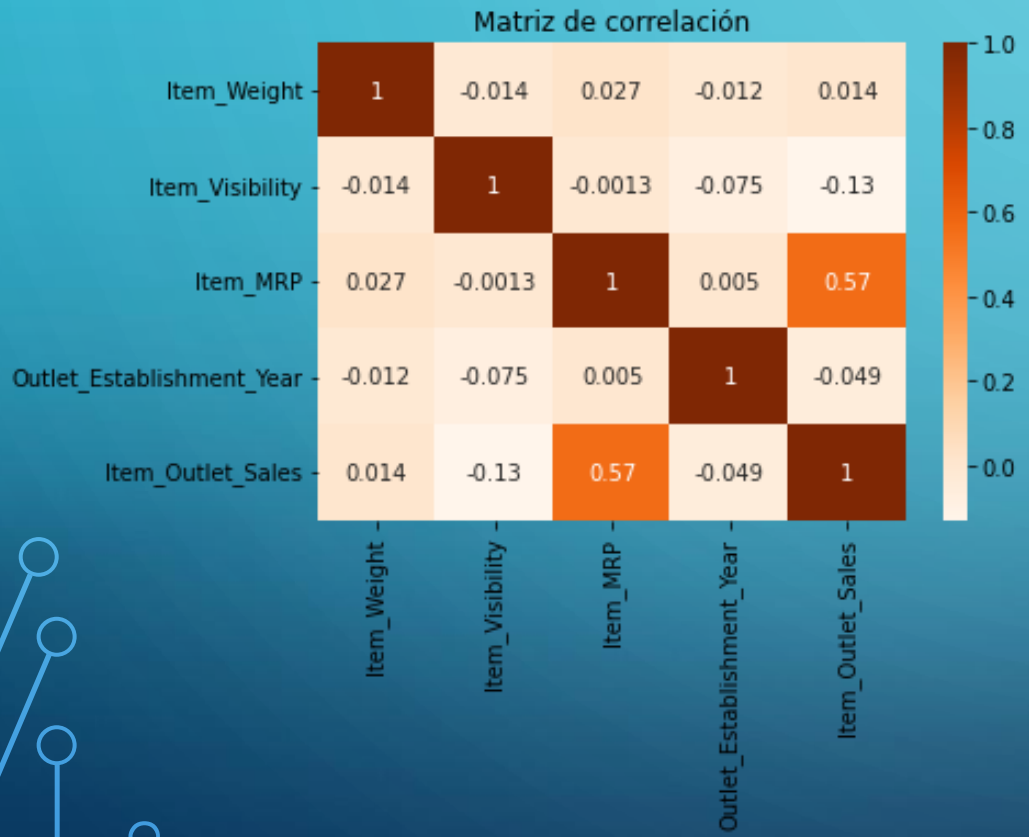
- La variable a predecir es Item_Outlet_Sales y tenemos características como el peso, contenido de grasa, porcentaje de visibilidad, categoría, precio de venta, tamaño y tipo de tienda, entre otras.

El gráfico muestra el comportamiento general de los datos de ventas concentrados entre 0 y 3000 $.



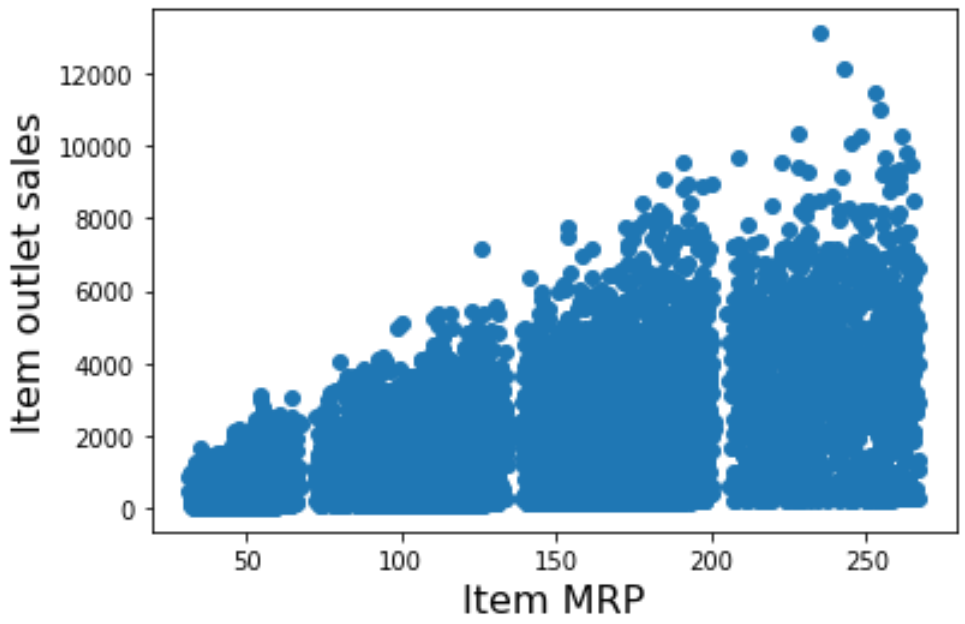
ANÁLISIS EXPLORATORIO:

- Mapa de calor de la correlación entre las características:



El gráfico muestra una correlación moderada entre el precio máximo de venta y las ventas del producto, dado que el coeficiente de correlación es positivo se infiere que si aumenta MRP también aumentaran las ventas generales del producto.

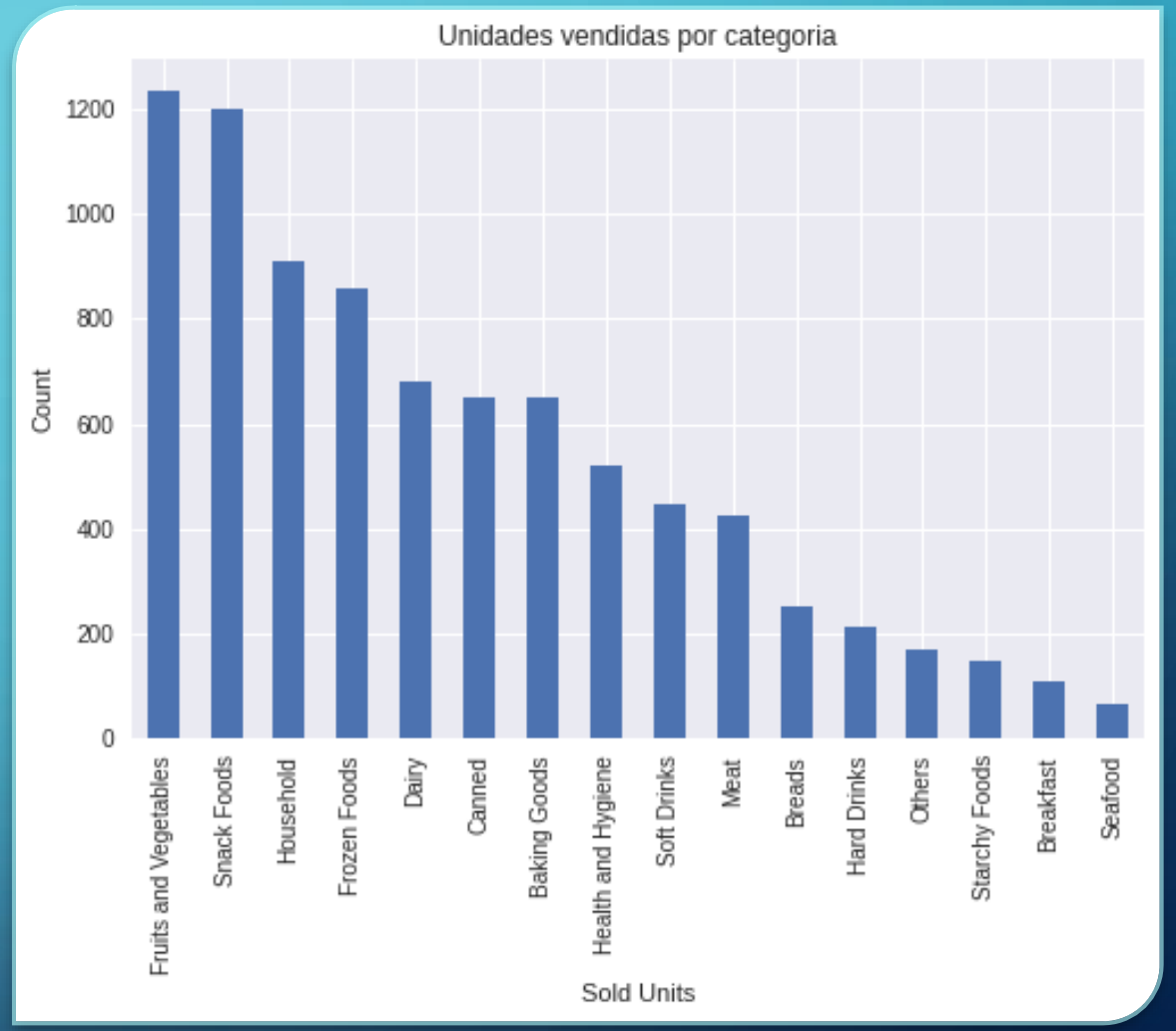
También vemos que existe una correlación casi nula entre la visibilidad del producto y las ventas de la tienda, es decir, un producto que ocupe mayor espacio físico en la tienda no vende más.



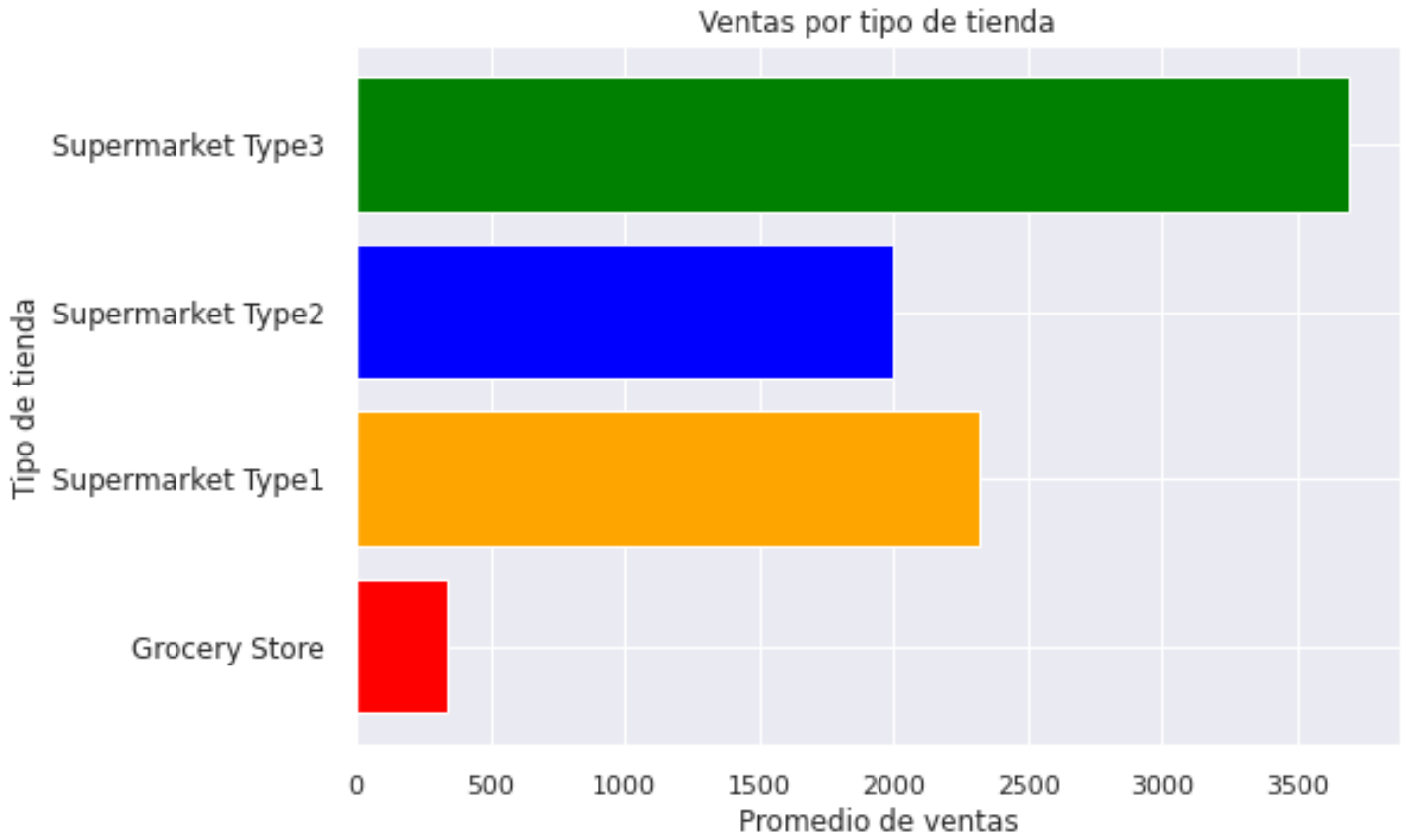
- Dada la correlación positiva que se muestra en la gráfica anterior, podemos ver en el siguiente gráfico de dispersión el comportamiento de los datos entre estas dos variables y su relación directa positiva.

CATEGORÍA MAS VENDIDA:

- El siguiente grafico nos muestra que la categoría mas vendida es “frutas y vegetales”, y la segunda “snacks”.
- Las categorías que tal vez requieran mayor fuerza de ventas son aquellas que aparecen en menor cantidad como desayunos y comida de mar.

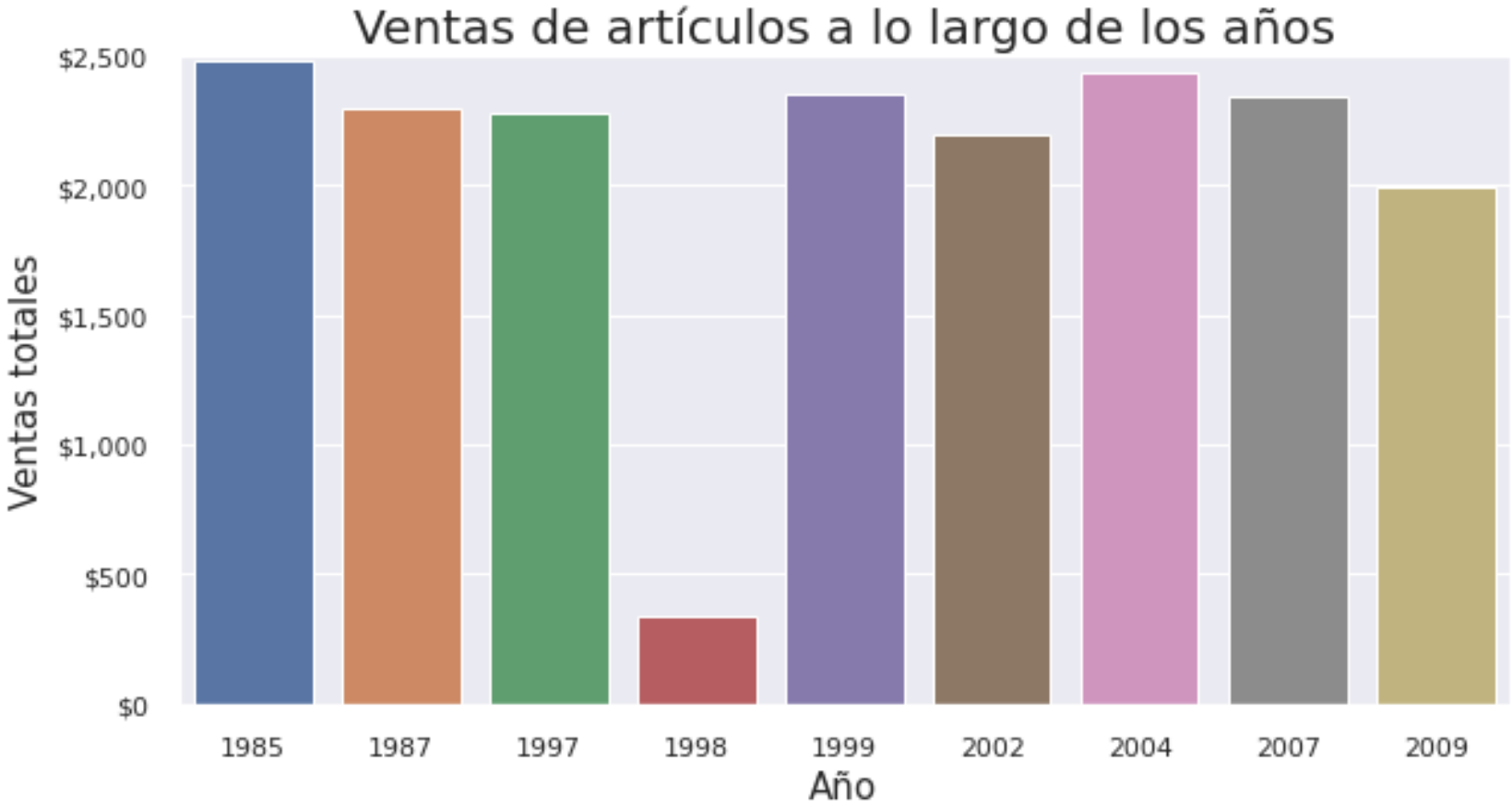


TIENDA CON EL MEJOR PROMEDIO DE VENTAS:



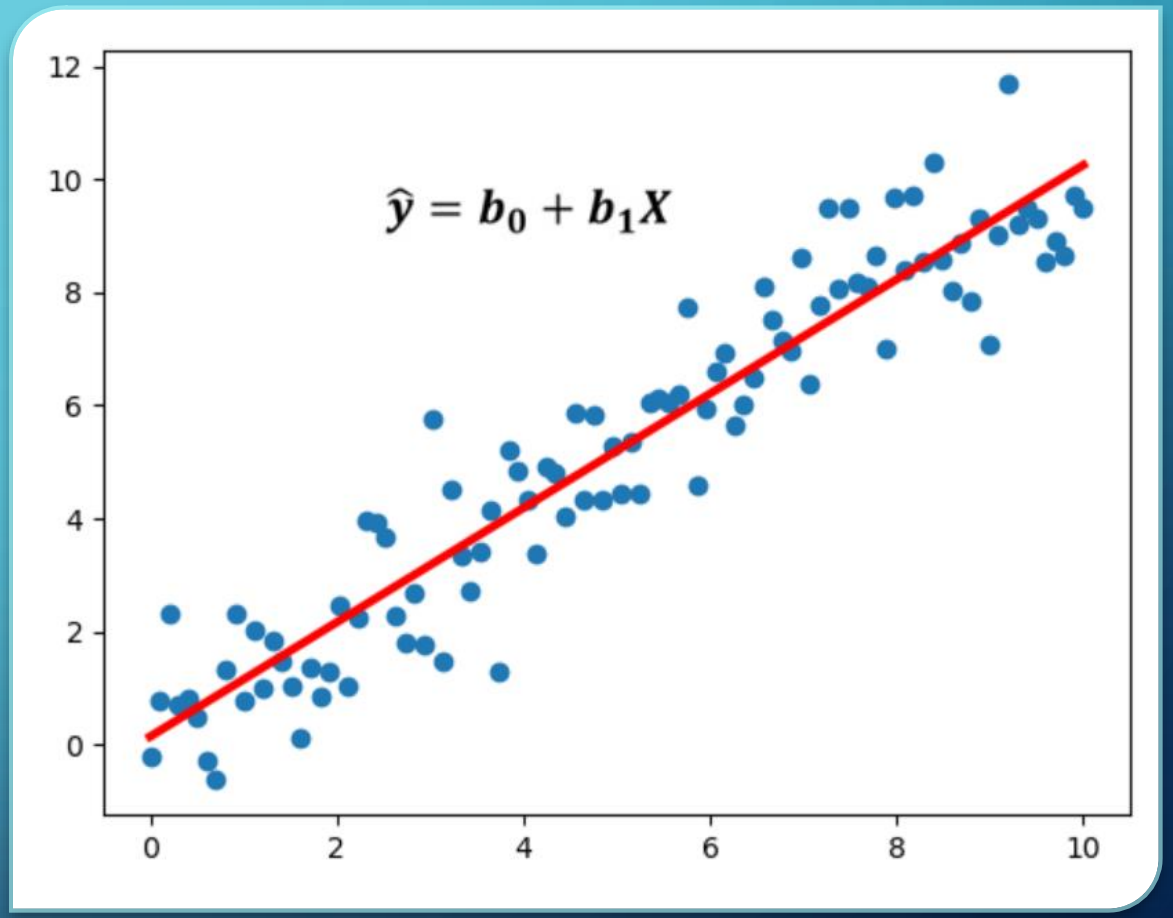
VENTAS A TRAVÉS DE LOS AÑOS:

- El gráfico nos muestra que las ventas muestran una cierta consistencia a través de los años a excepción del año 1998 donde las ventas son significativamente bajas.



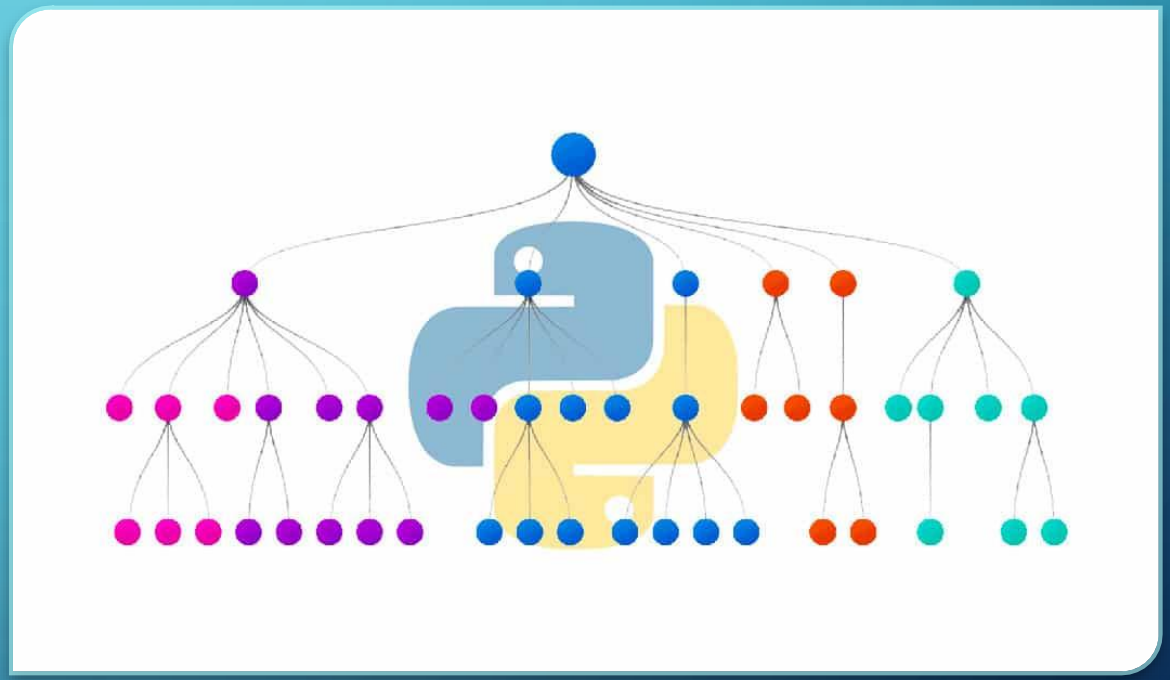
MODELOS:

- **Modelo de regresión lineal:**
- El valor de R^2 para el modelo de entrenamiento es: 0.562
- El valor de R^2 para el modelo de prueba es: 0.567
- Estos valores de R^2 muestran un bajo rendimiento en la predicción usando la regresión lineal.



MODELO DE ARBOLES DE DECISIÓN:

- Con este modelo obtuvimos un mejor rendimiento dado los resultados de R^2:
- R^2 para el modelo de entrenamiento: 0.604
- R^2 para el modelo de prueba: 0.595.
- Comparado con el modelo de regresión lineal, aconsejaría utilizar el modelo de árboles de decisión para la predicción.



RESULTADOS Y HALLAZGOS

Después de un análisis exploratorio y una serie de pruebas para encontrar el mejor modelo para predecir las ventas de los productos, se concluye que:

- Ninguno de los 2 modelos aplicados muestra un buen rendimiento en la predicción que se desea realizar, se debe seguir indagando y probar otros modelos.

- La regresión logística muestra menor capacidad predictiva para este caso en particular.
- El modelo de arboles de decisión muestra mayor capacidad puesto que puede arrojar un R^2 cercano al 60% para los datos de entrenamiento y de prueba.
- No se encuentran relaciones fuertes entre variables, solo una relación moderada entre el precio máximo de ventas y el total de ventas de los productos.