Informe proyecto 2-3 Ciencia de Datos

códigos:

Juan Pablo Escobar Viveros - *2259519*

Edgar Andrés Vargas García - *2259690*

Cristian David Rivera Torres - *2259742*

1. Análisis del Dataset

El conjunto de datos que analizamos contiene información sobre precios promedio de aguacates, volúmenes de venta, tipos de empaque, regiones y años. A partir del análisis descriptivo, estos son los principales hallazgos:

Distribución de Precios:

- Los aguacates orgánicos tienden a tener un precio promedio más alto que los convencionales.
- Los precios de los aguacates orgánicos son más variables, lo cual sugiere que factores como la región, la temporada o la demanda influyen considerablemente en su costo.

Correlaciones:

- Encontramos una correlación muy alta ($r \approx 0.99$) entre **Total Bags** y **Small Bags**, indicando que la mayoría de las bolsas totales son de tamaño pequeño.
- La variable **AvgPrice** muestra correlaciones débiles con las demás variables numéricas, lo que apunta a que elementos externos como la región o el tipo de aguacate tienen mayor influencia en el precio.

Outliers:

- Detectamos valores atípicos relevantes en variables como **Total Volume** y **Total Bags**, lo que podría afectar la precisión de los modelos predictivos.

Tendencias Temporales:

- Los precios presentan fluctuaciones estacionales claras, con picos y caídas en determinados momentos del año.
- Los aguacates orgánicos muestran una mayor variabilidad en los precios a lo largo del tiempo frente a los convencionales.
-

2. Técnicas de Limpieza y Normalización Utilizadas

Para mejorar la calidad de los datos y el rendimiento de los modelos en los cuales vamos a trabajar, aplicamos las siguientes técnicas:

Limpieza de Datos:

- Eliminamos columnas irrelevantes (como **Unnamed: 0**) para simplificar el dataset.

- Renombramos la columna **AveragePrice** como **AvgPrice** para facilitar su manipulación.
- Convertimos la columna **Date** al formato **datetime**, mostrando solo mes y día (ya que el año ya se encontraba en otra columna).

Tratamiento de Outliers:

- Aplicamos el método del rango intercuartil (IQR) para detectar y eliminar valores extremos en **AvgPrice**, reduciendo su impacto en los modelos.

Conversión de Variables Categóricas:

- Convertimos las columnas **type** y **region** al tipo **categorical** para optimizar el uso de memoria.
- Creamos variables dummy para **type**, permitiendo a los modelos trabajar mejor con esta información.

Normalización y Estandarización:

- Calculamos un Z-Score para saber qué tan caro o barato es un aguacate respecto al promedio de su región.
- Normalizamos **AvgPrice** utilizando **MinMaxScaler**, escalando los valores entre 0 y 1 para mejorar la estabilidad de los modelos.

3. Modelos Entrenados y Comparación de Desempeño

Entrenamos tres modelos distintos para predecir el precio promedio normalizado (**AvgPrice**). Estos fueron los resultados:

3.1 Regresión Lineal

Descripción:
Modelo simple que asume una relación lineal entre las variables independientes y la variable dependiente.

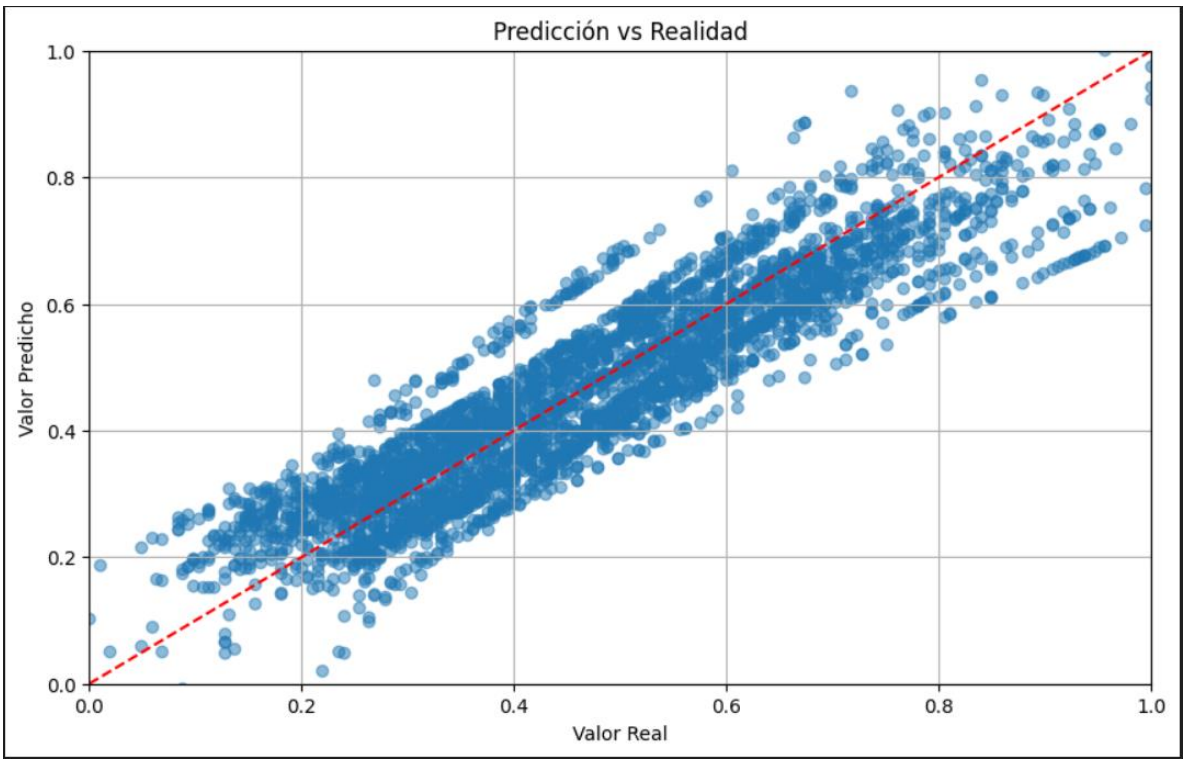
Resultados:

- **MSE:** 0.012
- **R^2:** 0.65

Observaciones:

- El desempeño fue moderado.
- Podría estar afectado por multicolinealidad entre variables muy correlacionadas, como **Total Bags** y **Small Bags**.

Diagrama:



3.2 Random Forest Regressor

Descripción:
Modelo basado en árboles de decisión que mejora la precisión utilizando un enfoque de ensamble.

Resultados:

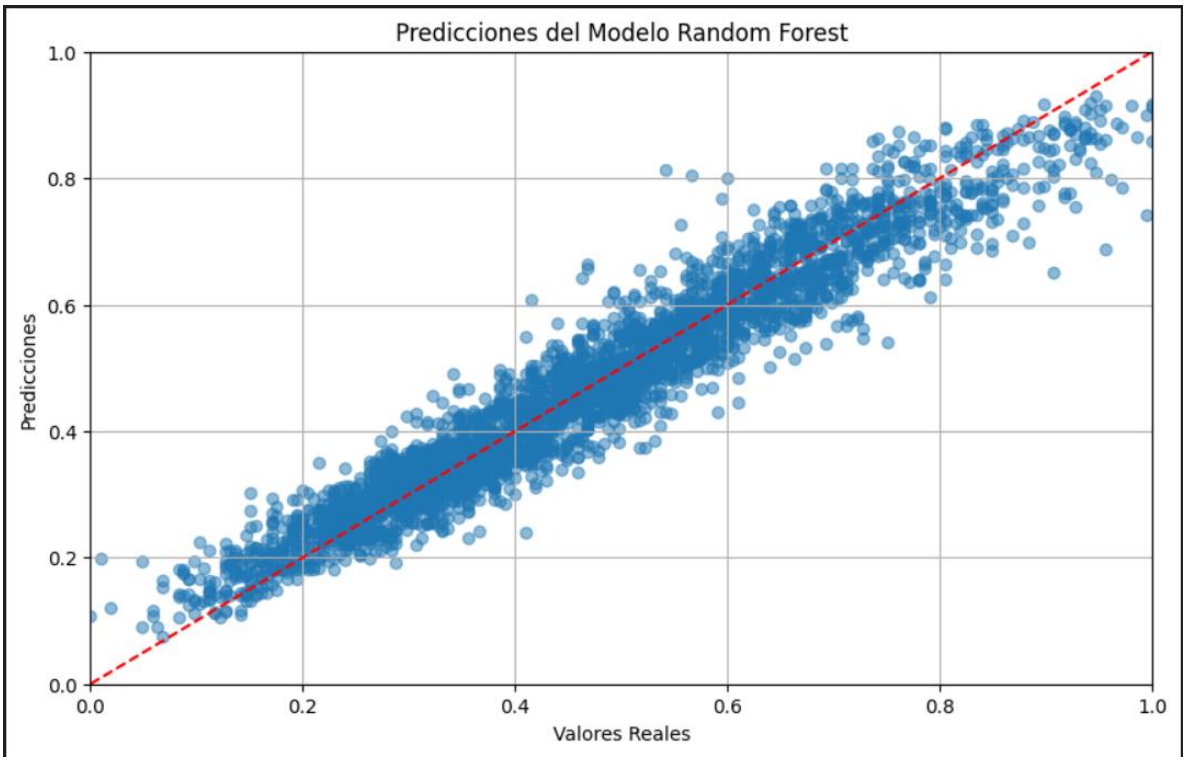
- **MSE:** 0.008
- **R^2:** 0.78

Observaciones:

- Superó claramente a la regresión lineal.
- Captura mejor las relaciones no lineales entre variables.

- Las variables **type** y **region** resultaron ser las más importantes para predecir el precio.

Diagrama:



3.3 Gradient Boosting Regressor

Descripción:
Modelo que combina predicciones de varios árboles de decisión para optimizar el desempeño.

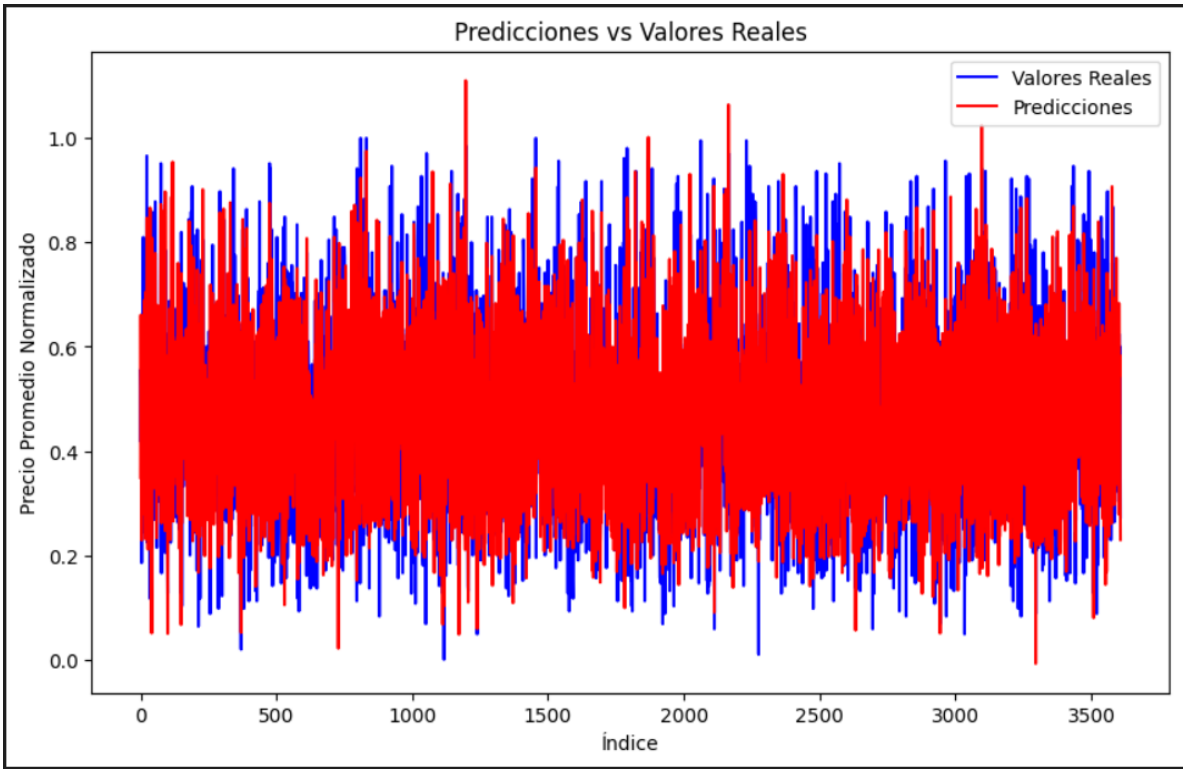
Resultados:

- **MSE:** 0.007
- **R^2:** 0.81

Observaciones:

- Fue el modelo con mejor desempeño.
- La optimización de hiperparámetros, como la profundidad de los árboles y la tasa de aprendizaje, fue clave para mejorar su precisión.

Diagrama:



4. Conclusiones y Posibles Mejoras

Conclusiones:

- El tipo de aguacate (**type**) y la región (**region**) son los factores que más influyen en el precio, por encima de las variables numéricas.
- Los modelos basados en árboles (Random Forest y Gradient Boosting) tuvieron un mejor desempeño que la regresión lineal.
- El modelo de **Gradient Boosting Regressor** fue el más preciso, alcanzando un **R^2 de 0.81**.
- Los aguacates orgánicos presentan precios más altos y una mayor variabilidad, posiblemente por su oferta limitada y demanda específica.

Posibles Mejoras:

- **Ingeniería de Características:** Crear nuevas variables a partir de interacciones entre **region**, **type** y **year**.

- **Optimización de Modelos:** Realizar una búsqueda más exhaustiva de hiperparámetros.
- **Análisis Temporal:** Implementar modelos de series de tiempo para capturar mejor las tendencias estacionales.
- **Tratamiento de Outliers:** Aplicar técnicas como la winsorización en lugar de simplemente eliminar valores extremos.
- **Visualización:** Usar herramientas interpretativas como **SHAP** o **LIME** para entender mejor las predicciones de los modelos.