**INFORME PROYECTO #2 Y #3**

**JUAN CAMILO GUTIERREZ VIVEROS - 2159874**

**ANDRES FELIPE ROJAS – 2160**

**PUNTO 1 - Analisis descriptivo del dataset**

**el dataset tiene 18.249 registros y 14 columnas, lo que es una muestra amplia y representativa.**

**Ninguna columna tiene valores nulos**

**Heatmap**

**Gráfico, Gráfico de rectángulos

El contenido generado por IA puede ser incorrecto.**

**La matriz de correlación muestra cómo se relacionan las variables numéricas del dataset entre sí.**

**El color representa la fuerza y dirección de la relación:**

**Cercano a 1 → fuerte correlación positiva (aumentan juntas).**

**Cercano a -1 → fuerte correlación negativa (una sube, la otra baja).**

**Cercano a 0 → sin relación lineal relevante**

**PUNTO 2 - Limpieza y Normalización de Datos**

**Para la normalización se hicieron los siguientes procesos detallados en el archivo limpieza.ipynb**

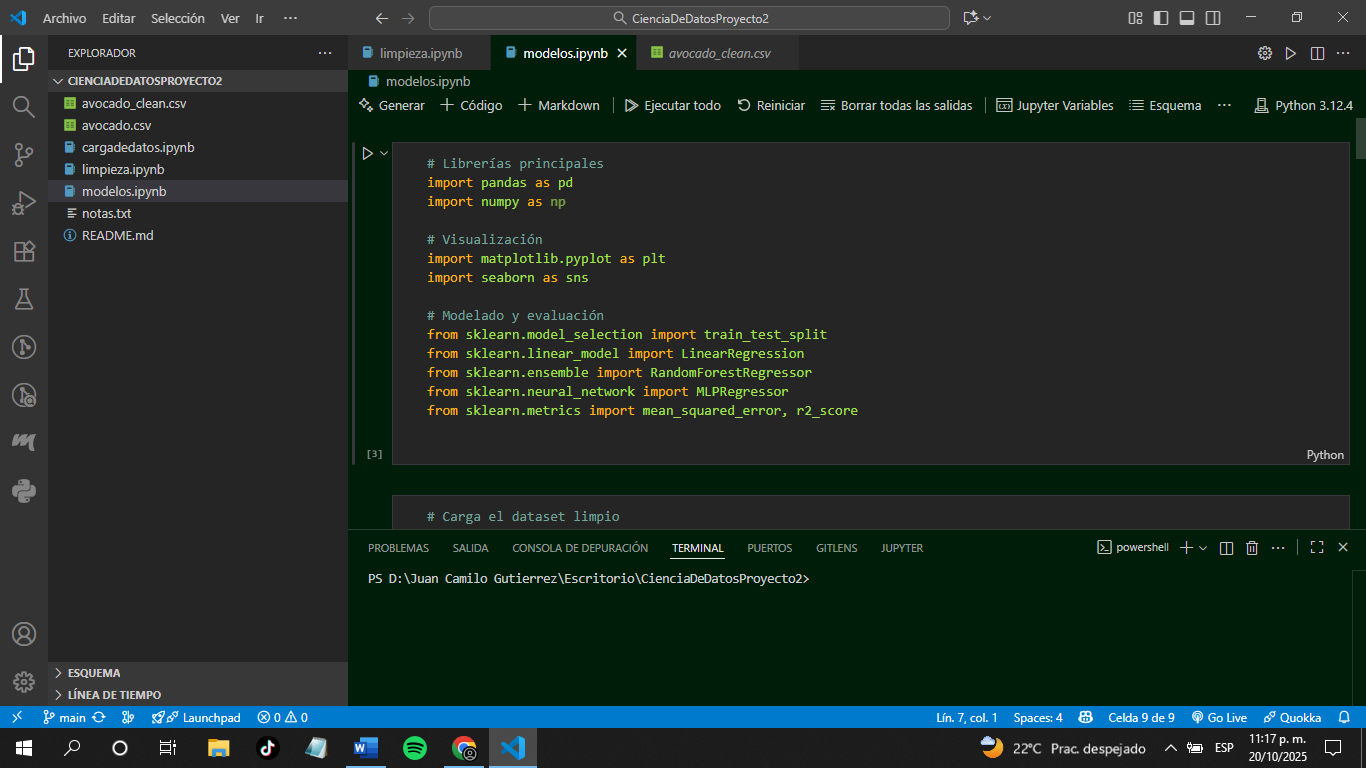
**Texto

El contenido generado por IA puede ser incorrecto.**

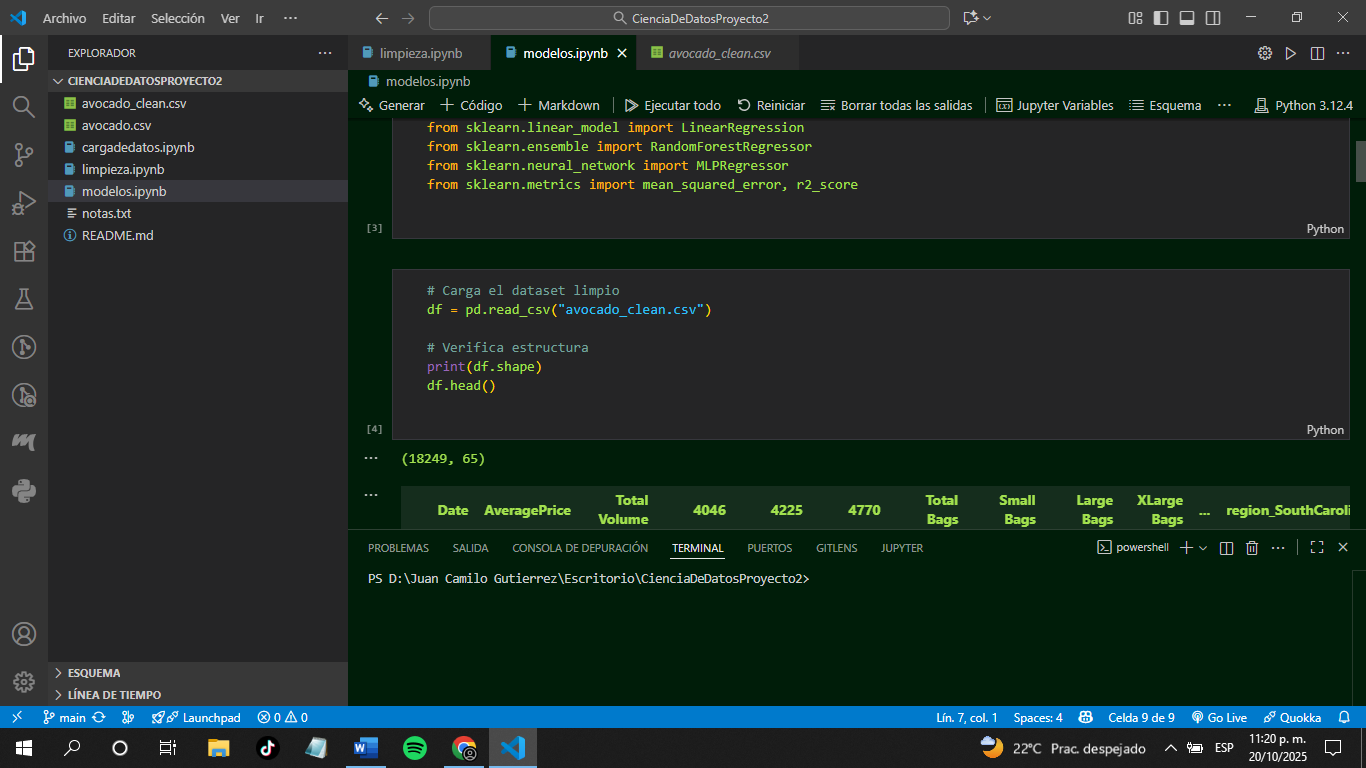
**PUNTO 3**

Se implementan los 3 modelos de machine learning, asimismo se analiza y compara sus métricas de desempeño.

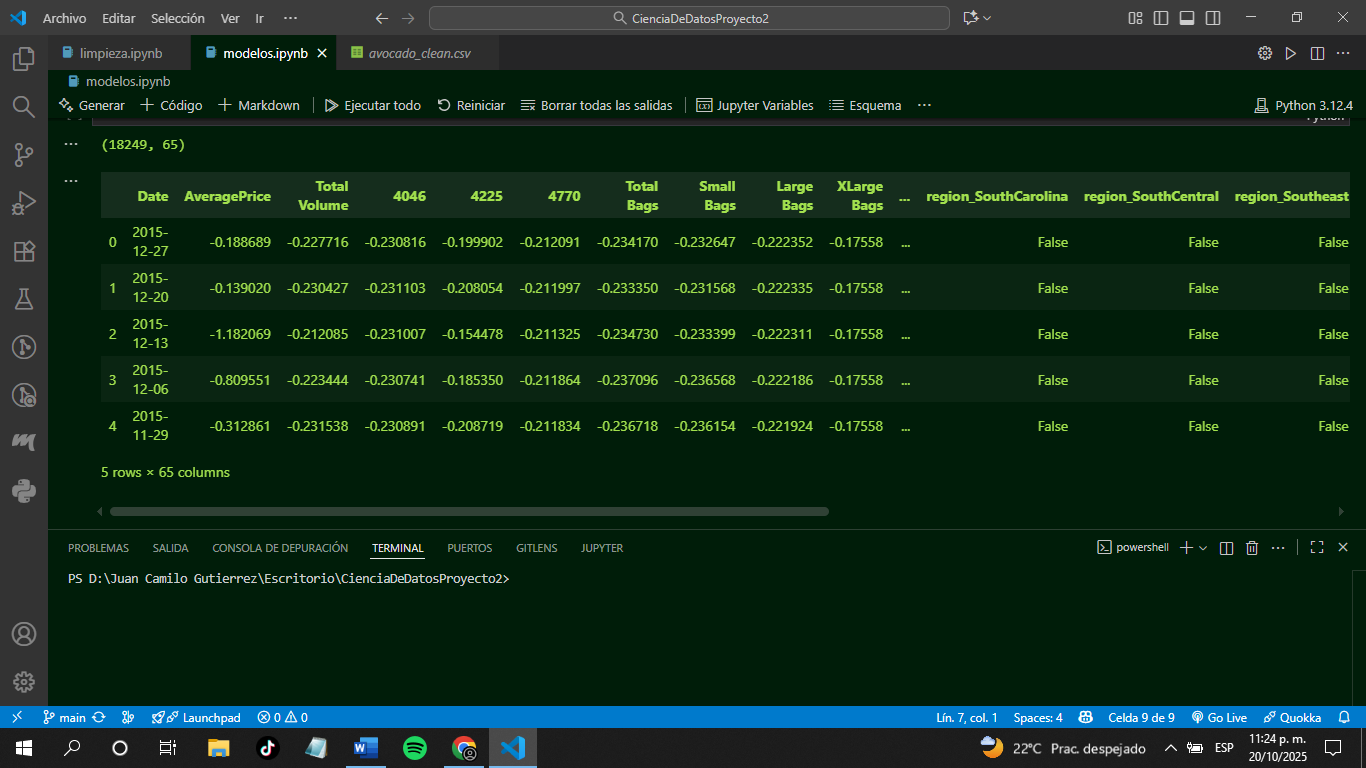
Se inicializa creando en la carpeta del proyecto el archivo *modelo.ipynb*, luego se importan las librerías necesarias (pandas, sklearn, matplotlib, etc.), se define estado aleatorio y muestra %matplotlib inline para Jupyter. Esta es la base del entorno:



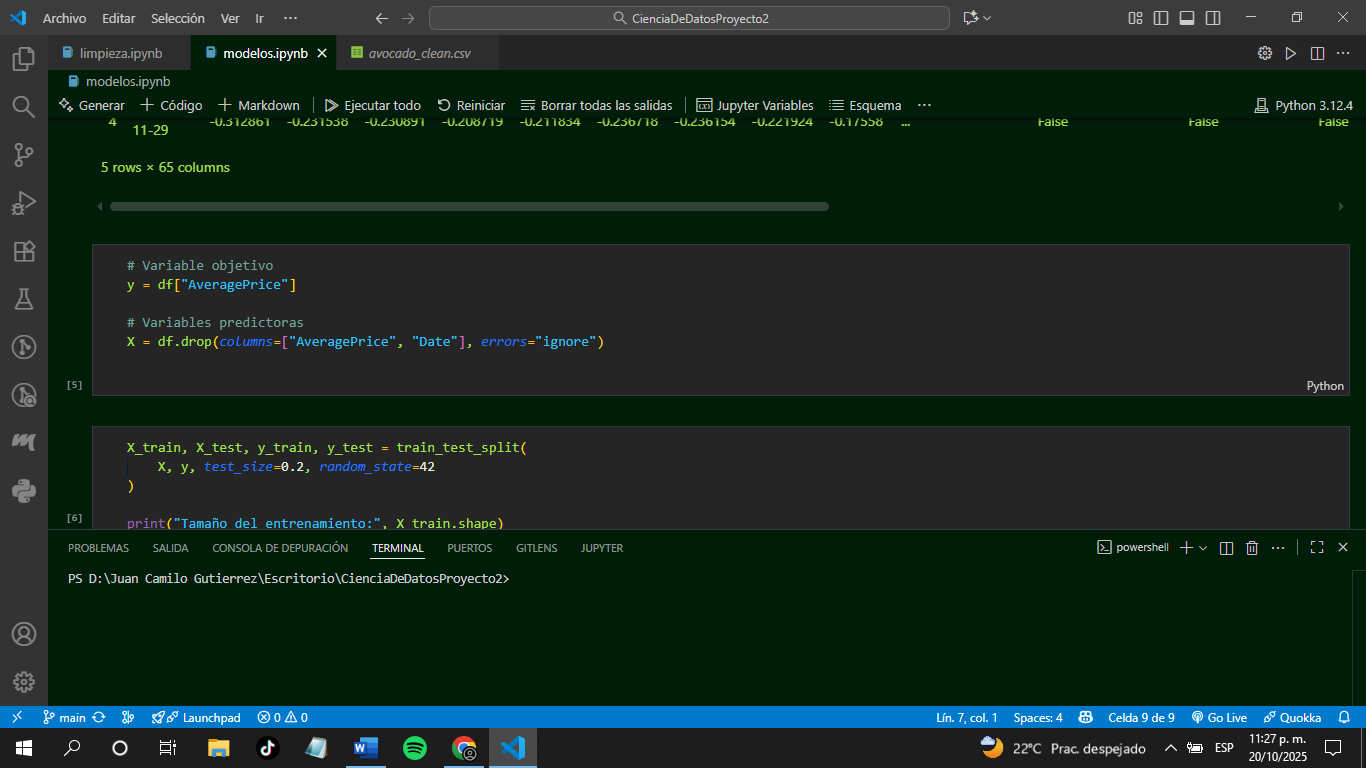
Se carga el dataset limpio (*avocado\_clean.csv*) para el respectivo uso en los modelos:



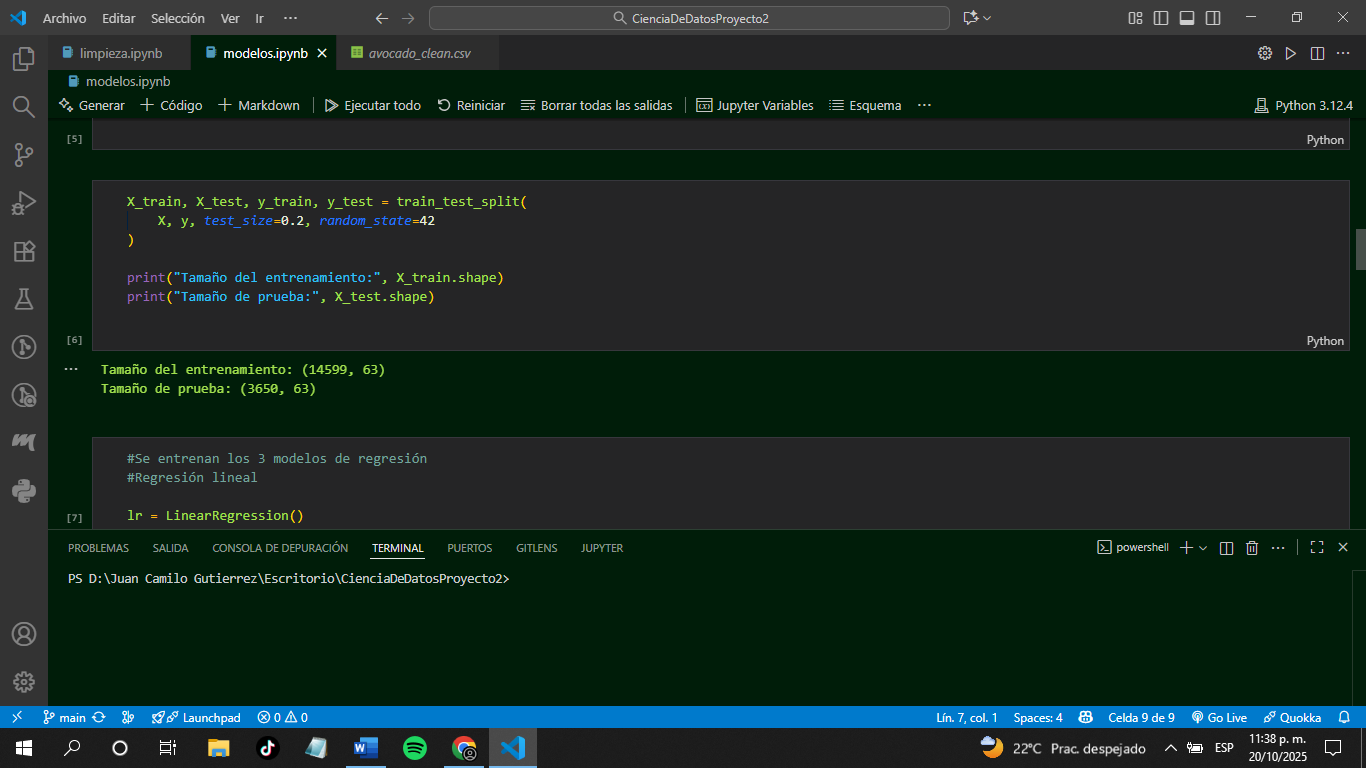
La celda anteriormente mostrada contiene *pd.read\_csv("avocado\_clean.csv"), df.head()* y *df.describe()* para comprobar forma, columnas y estadísticas básicas. Aquí se verifica que el dataset cargó correctamente:



Pequeña celda en a cual se separa la variable objetivo y = AveragePrice y las predictores X, y que se elimina Date si existe:



El siguiente fragmento de código crea X y Y, y realiza train\_test\_split (80/20). Muestra tamaños de train/test, también indica que los datos están listos para entrenar:



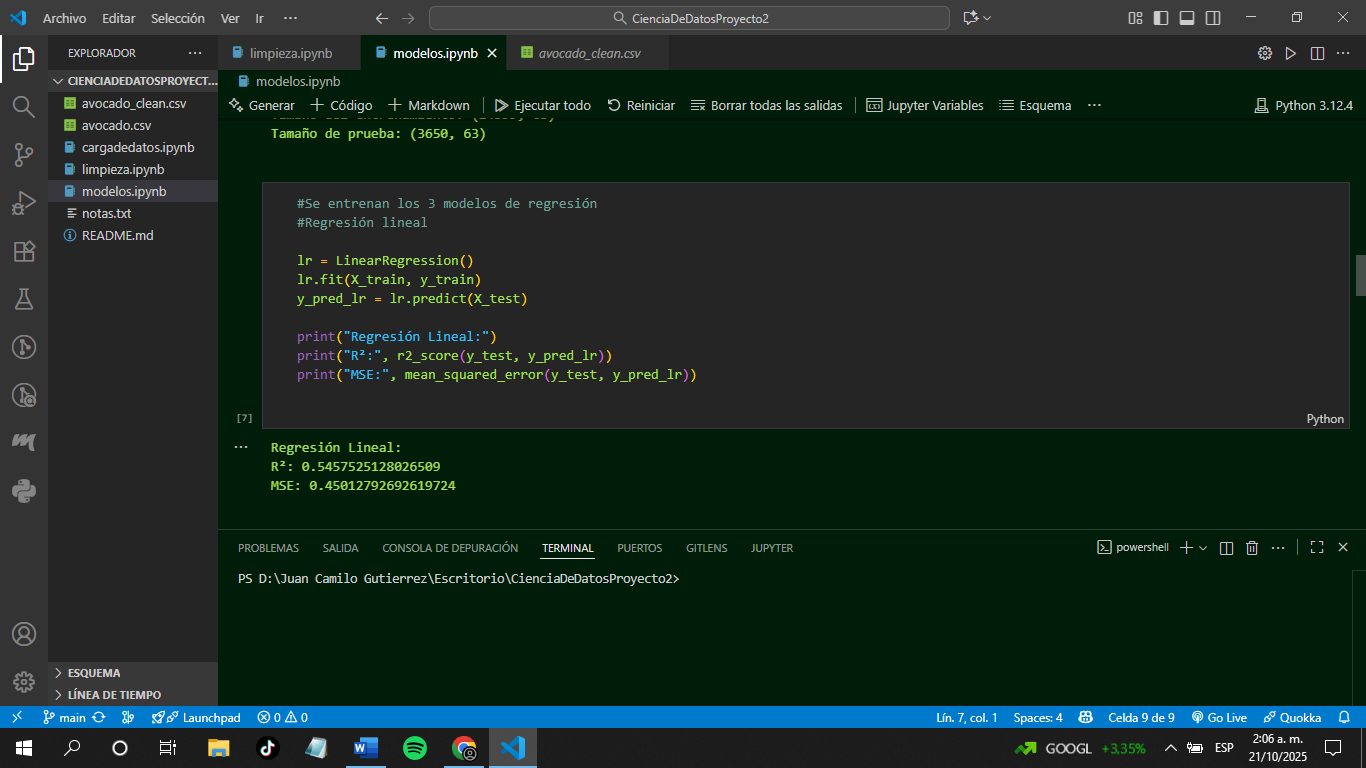
Están listas las métricas que se usarán: R², MSE, MAE.

R² (Coeficiente de determinación): mide la proporción de la varianza de la variable objetivo que el modelo logra explicar. Rango: 0 a 1. Cuanto más cercano a 1, mejor es el ajuste del modelo.

MSE (Mean Squared Error / Error cuadrático medio): mide el promedio de los errores al cuadrado entre los valores reales y predichos. Cuanto más bajo el MSE, menor es el error promedio.

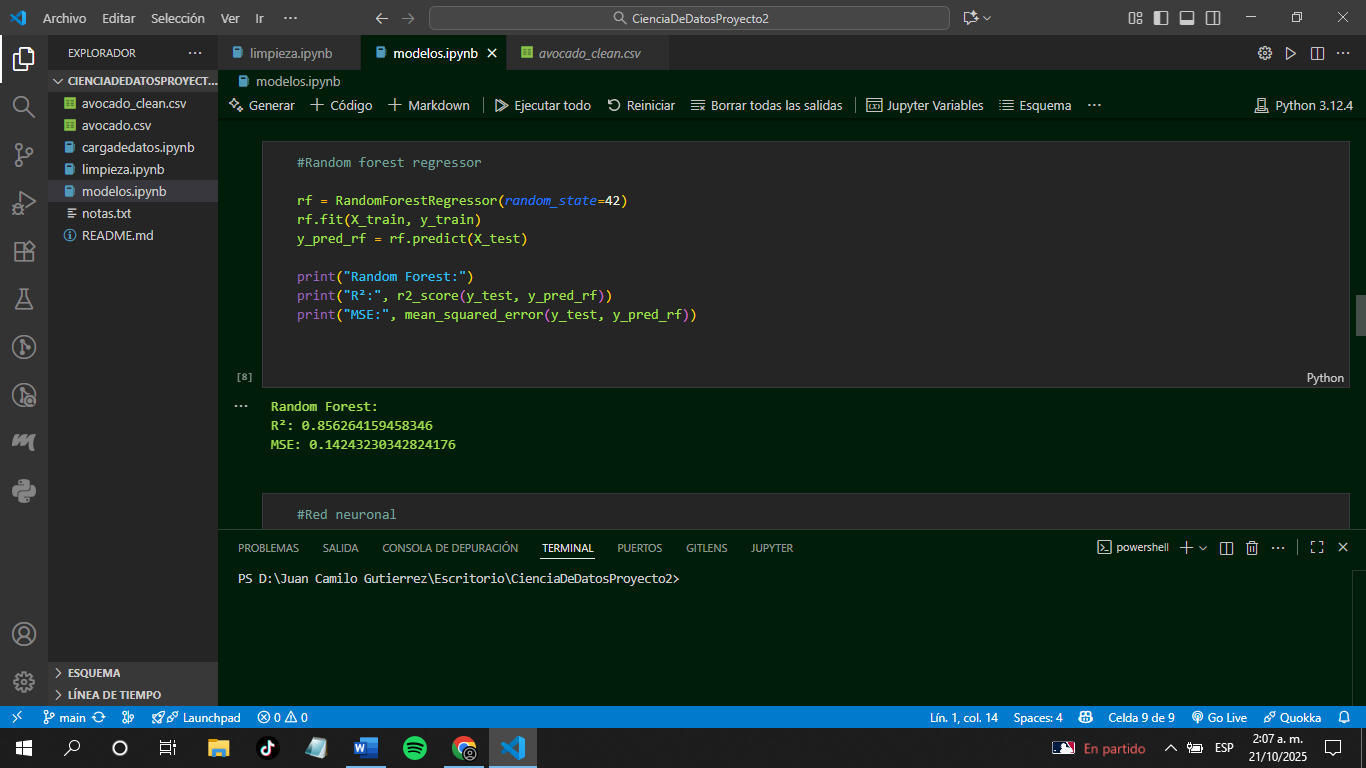
MAE (Mean Absolute Error / Error absoluto medio): mide el error promedio en las predicciones en las mismas unidades del valor real.

En esta celda se implementa el primer modelo predictivo: Regresión Lineal, el cual establece una relación lineal entre las variables independientes del conjunto de entrenamiento y la variable objetivo AveragePrice:



En los resultados obtenidos, el modelo alcanzó un R² = 0.5457 y un MSE = 0.4501, lo que sugiere que el modelo explica aproximadamente el 54.6 % de la variación del precio promedio del aguacate.  
Estos valores reflejan un rendimiento moderado, lo cual es esperado en una regresión lineal simple, ya que este tipo de modelo no capta relaciones no lineales complejas.

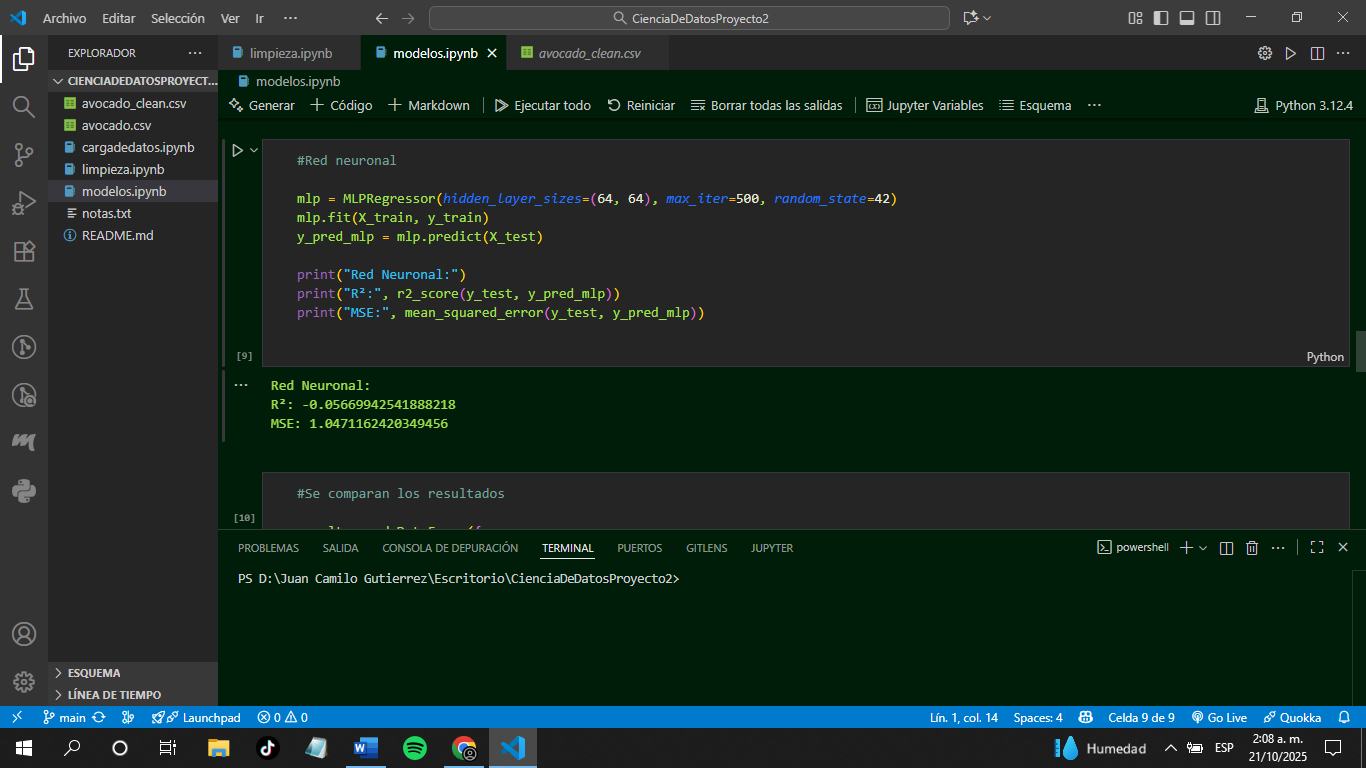
En la siguiente celda se aplica el modelo Random Forest Regressor, una técnica de aprendizaje automático basada en el ensamble de múltiples árboles de decisión.



En la evaluación del modelo, se obtuvieron los siguientes resultados:

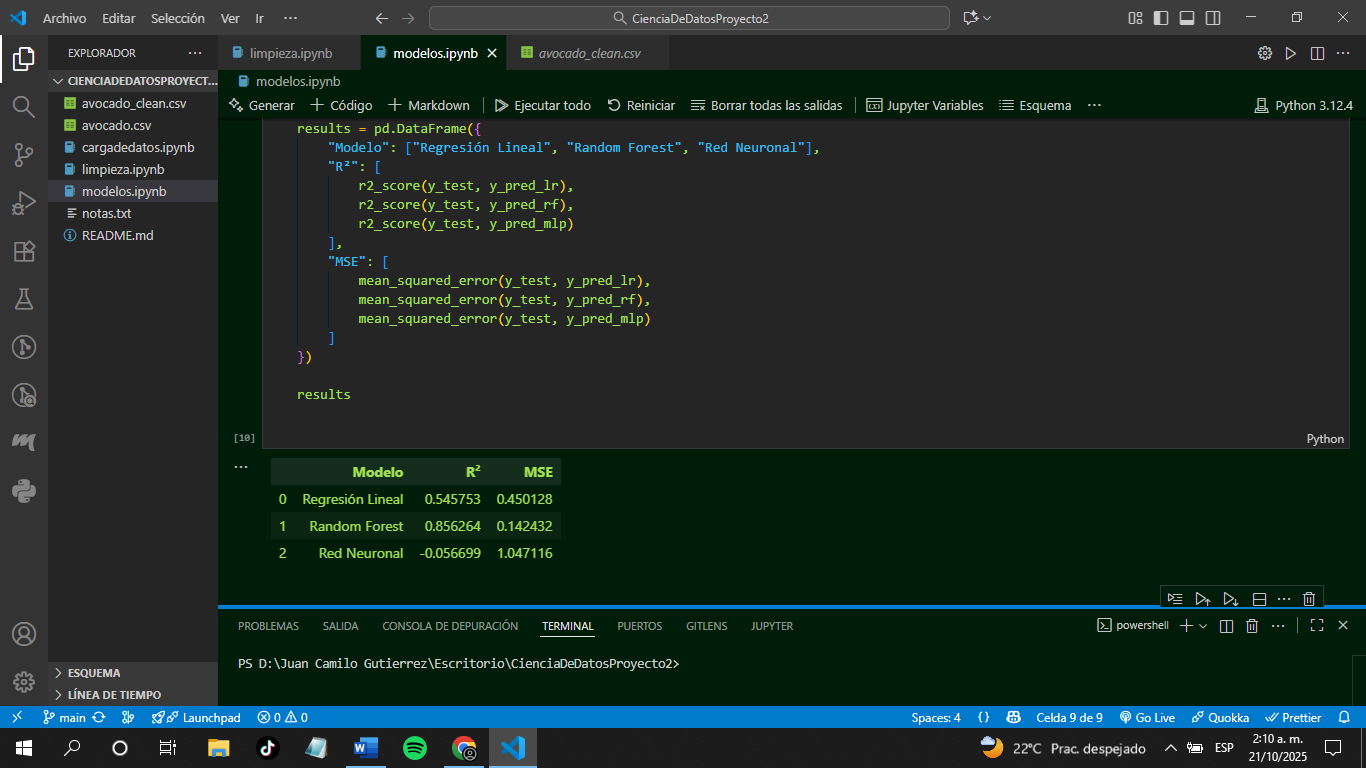
R² = 0.8562, lo que significa que el modelo explica aproximadamente el 85.6 % de la variabilidad del precio promedio de los aguacates y MSE = 0.1424, indicando un error cuadrático medio considerablemente menor que el de la regresión lineal.

Por último en esta celda se implementa el último modelo MLP Regressor (Multi-Layer Perceptron), que corresponde a una red neuronal artificial de tipo perceptrón multicapa, utilizada para resolver problemas de regresión:

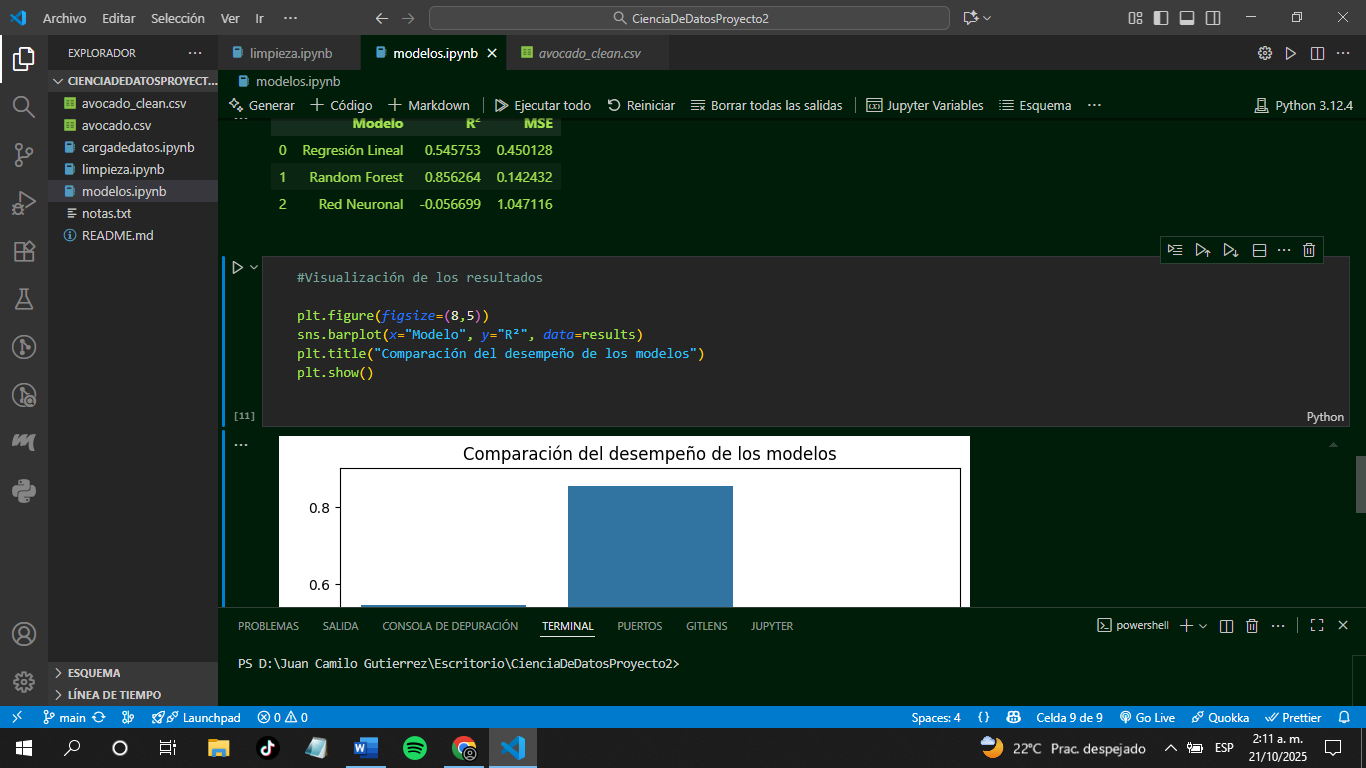


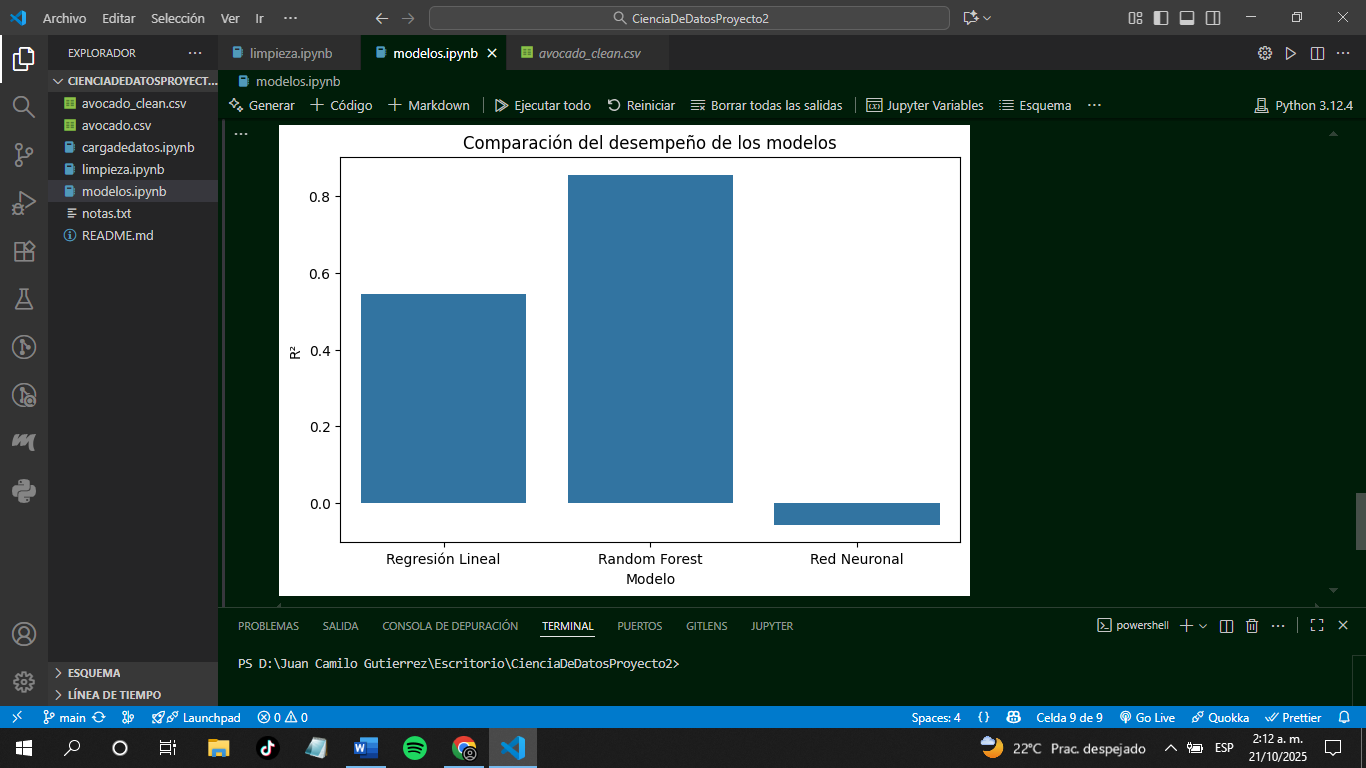
Para este modelo se obtuvo los siguientes resultados: R² = -0.0567, lo que indica que el modelo no logra explicar adecuadamente la variabilidad del precio promedio del aguacate (un R² negativo refleja un desempeño inferior al promedio) y MSE = 1.0471, que representa un error cuadrático medio alto comparado con los otros modelos.

En la siguiente tabla se planea mostrar en una tabla más organizada los resultados de las metricas de cada modelo:



Lo mismo en la siguiente celda se busca comparar el valor de R² de los 3 modelos:





En la gráfica se puede apreciar que el modelo Random Forest Regressor obtuvo el valor de R² más alto (≈ 0.86), explicando cerca del 86 % de la variabilidad en el precio promedio del aguacate.  
La Regresión Lineal alcanzó un R² ≈ 0.54, con un desempeño moderado, mientras que la Red Neuronal presentó un R² negativo, evidenciando un mal ajuste al conjunto de datos.

En conjunto, los resultados confirman que el Random Forest es el modelo más eficiente, al ofrecer la mejor precisión y capacidad de generalización frente a los demás enfoques.

**Conclusiones**

Se evidenció en la realización del proyecto que las técnicas implementadas poseen un buen potencial cuando de predecir precios en datos reales se trata.