Proyecto 2 y 3

Introducción a la Ciencia de Datos

Informe

Elaborado por:

Gómez Agudelo, JUAN SEBASTIÁN – 2259474

Henao Aricapa, STIVEN – 2259603

Hernández Ortíz, VÍCTOR MANUEL – 2259520

Docente:

Ocampo Arbeláez, HÉCTOR FABIO

Sede Tuluá

Abril 2025

Índice

Análisis del Dataset

El dataset utilizado en este proyecto corresponde al conjunto de datos "Avocado Prices" disponible en Kaggle.

Este conjunto de datos recopila información sobre los precios y volúmenes de venta de aguacates en distintas regiones de Estados Unidos, entre los años 2015 y 2018.

Características principales

- **Número de registros:** 18,249 observaciones.
- **Variables numéricas:**
    - **AveragePrice:** precio promedio por unidad de aguacate.
    - **Total Volume:** volumen total de ventas.
    - **Small Bags, Large Bags, XLarge Bags:** volumen de ventas segmentado por tamaño de empaque.
    - **4046, 4225, 4770:** cantidades vendidas de tipos específicos de aguacate.
- **Variables categóricas:**
    - **type:** tipo de producto (conventional o organic).
    - **region:** área geográfica de venta.
- **Variable objetivo:**
    - AveragePrice, que se buscará predecir.

Descripción general

- El dataset combina variables categóricas y numéricas, permitiendo aplicar técnicas de codificación y normalización.
- Las variables de volumen presentan valores atípicos importantes, evidenciados mediante boxplots.
- La distribución de Total Volume muestra un sesgo positivo, con un gran número de ventas bajas y unos pocos registros de ventas extremadamente altas.
- Se observaron ligeras correlaciones entre variables de volumen y precio, aunque no siempre lineales.
- Se evidencian diferencias de precio entre aguacates convencionales y orgánicos.

Todo el análisis exploratorio detallado, incluyendo histogramas, boxplots, matrices de correlación y más, se encuentra documentado en el archivo ***AnalisisDescriptivo.ipynb***.

Técnicas de limpieza y normalización utilizadas.

Durante esta etapa del proyecto se aplicaron las siguientes técnicas de limpieza y normalización de datos:

- **Eliminación de columnas innecesarias**:
    - Se eliminó la columna Unnamed: 0, correspondiente a un índice automático sin valor analítico.
    - Posteriormente, se eliminaron Date, year y type, tras generar variables derivadas más útiles (Dias_Desde_Primer_Registro y type_ORGANIC).
- **Tratamiento de valores nulos**:
    - Se verificó mediante un mapa de calor (heatmap) que no existían valores nulos significativos en el dataset.

- **Normalización de variables categóricas**:
    - Se estandarizaron las cadenas de texto de type y region, asegurando consistencia de mayúsculas y formato.
    - Se renombraron las columnas 4046, 4225 y 4770 a Pequeño/Mediano, Grande y Extra Grande respectivamente, para mejorar la interpretación.
- **Conversión de variables de fecha**:
    - Se transformó la columna Date a tipo datetime.
    - Se creó una nueva variable Dias_Desde_Primer_Registro, representando el número de días transcurridos desde el primer registro, para facilitar el modelado temporal.
- **Detección y eliminación de valores atípicos**:
    - **Total Volume**: Se eliminaron outliers utilizando el rango intercuartílico (IQR).
    - **Total Bags**: También se aplicó el método IQR para limpiar valores atípicos.
    - **AveragePrice**: Se utilizaron puntajes Z (z-score) para identificar y eliminar valores extremos, usando un umbral de 2.5.
- **Normalización de variables numéricas**:
    - Se utilizó **MinMaxScaler** para escalar todas las variables numéricas al rango [0, 1], preservando la forma original de las distribuciones.
    - Variables normalizadas: AveragePrice, Total Volume, Total Bags, Small Bags, Large Bags, XLarge Bags, Pequeño/Mediano, Grande, Extra Grande.
- **Codificación de variables categóricas**:
    - Se aplicó **One-Hot Encoding** a la variable type, generando una nueva columna type_ORGANIC.
    - Se eliminó una de las categorías (drop='first') para evitar multicolinealidad en los modelos predictivos.

- **Exportación del dataset final**:
    - El dataset limpio y normalizado fue exportado como avocado_transformado.csv, preparado para la fase de modelado.

Modelos entrenados y comparación de desempeño

Durante esta fase se entrenaron y evaluaron cinco modelos de regresión distintos con el objetivo de predecir el precio promedio de los aguacates (`AveragePrice`). Estos modelos fueron:

- **Regresión Lineal**
- **Árbol de Decisión (Decision Tree Regressor)**
- **Random Forest Regressor**
- **Gradient Boosting Regressor**
- **Red Neuronal (MLP - Multilayer Perceptron)**

Evaluación de desempeño

Se utilizó un conjunto de prueba del 20% de los datos (hold-out) y se aplicaron las siguientes métricas para evaluar el rendimiento de los modelos:

- **MSE (Mean Squared Error)**: mide el error promedio cuadrático entre las predicciones y los valores reales.
- **R^2 (Coeficiente de determinación)**: indica qué proporción de la varianza de la variable objetivo es explicada por el modelo.

Los resultados obtenidos fueron:

Modelo	MSE	R^2
Gradient Boosting	**0.0025**	**0.9296**
Random Forest	0.0085	0.7582
Decision Tree	0.0129	0.6336
Regresión Lineal	0.0159	0.5483
Red Neuronal (MLP)	0.0159	0.5484

El modelo **Gradient Boosting** presentó el mejor desempeño tanto en MSE como en R^2, seguido por **Random Forest**.

Visualización del error

Se analizaron las distribuciones de errores residuales para cada modelo. Las distribuciones de Gradient Boosting y Random Forest muestran una forma más centrada y simétrica, lo que indica una buena capacidad predictiva. Modelos como la regresión lineal y el MLP presentaron mayor dispersión y errores más extremos.

Hallazgos obtenidos en cada fase

Fase 1 - Análisis Descriptivo

- No se encontraron valores nulos en el dataset.
- AveragePrice tiene una distribución casi normal, ligeramente sesgada a la derecha.
- La mayoría de los precios de aguacate se concentran entre **$1.00** y **$1.50**.

  Total Volume muestra una distribución muy sesgada a la derecha, con outliers mayores a **40 millones**.
- Los aguacates **orgánicos** tienen precios promedio más altos y mayor variabilidad que los **convencionales**.
- La venta de aguacates en empaques **Small Bags** domina sobre **Large Bags** y **XLarge Bags**.
- Se detectaron outliers importantes en Total Volume y Total Bags.
- Existe una alta correlación positiva entre Total Volume y Total Bags ($r \approx 0.99$).
- Existe una correlación negativa moderada entre AveragePrice y las variables de volumen.
- Los precios del aguacate presentan estacionalidad anual, con picos en enero-febrero (relacionados al Super Bowl).
- De 2015 a 2017 se registró un aumento general de precios, con una ligera baja en 2018.
- Regiones como **HartfordSpringfield**, **San Francisco** y **New York** presentan los precios promedio más altos.

Nota: Para más detalles sobre el análisis descriptivo y visualizaciones, consultar el archivo ***AnalisisDescriptivo.ipynb***.

Fase 2 - Limpieza y Normalización de Datos

- Se eliminó correctamente la columna innecesaria Unnamed: 0, mejorando la claridad del dataset.
- No se detectaron valores nulos en el dataset original, evitando la necesidad de imputaciones.
- Las variables categóricas type y region fueron normalizadas, corrigiendo inconsistencias de formato en los textos.
- La creación de la variable Dias_Desde_Primer_Registro permitió representar la dimensión temporal de forma numérica y continua.
- Se identificó una alta cantidad de outliers en las variables Total Volume y Total Bags, los cuales fueron eliminados usando el método del rango intercuartílico (IQR).
- La variable AveragePrice presentaba valores extremos que fueron detectados y eliminados mediante el uso de puntajes Z (z-score) con un umbral de 2.5.
- Tras la limpieza de outliers, las distribuciones de las principales variables conservaron su tendencia natural, pero con menor dispersión y mayor estabilidad.
- La normalización con **MinMaxScaler** reescaló todas las variables numéricas al rango [0, 1], mejorando la comparabilidad entre características.
- Se aplicó **One-Hot Encoding** a la variable type, generando una nueva representación numérica (type_ORGANIC) adecuada para modelos de machine learning.
- La estructura final del dataset es más homogénea, consistente y lista para ser utilizada en el entrenamiento de modelos predictivos.

Nota: Todo el detalle del proceso de limpieza, eliminación de outliers y normalización se encuentra documentado en el archivo ***LimpiezaNormalizacion.ipynb***.

Fase 3 - Implementación de Modelos Predictivos

- El modelo **Gradient Boosting** fue el más preciso, con un **MSE de 0.0025** y un **R^2 de 0.9296**, siendo el que mejor explica la variabilidad de los precios.
- **Random Forest Regressor** también obtuvo buenos resultados, con un **R^2 de 0.7582**, destacando por su robustez frente a outliers.
- El modelo de **regresión lineal** alcanzó un R^2 bajo (**0.5483**), lo que indica que no capta adecuadamente las relaciones no lineales del problema.
- La **red neuronal MLP** tuvo un rendimiento similar a la regresión lineal, con R^2 de **0.5484**, mostrando que no aportó una mejora significativa.
- El **árbol de decisión individual** logró un R^2 de **0.6336**, pero con mayor tendencia al overfitting en comparación con ensambles como Random Forest.
- La distribución de errores de los modelos Gradient Boosting y Random Forest fue más centrada y simétrica, lo que refleja mejor ajuste a los datos.
- Modelos basados en árboles (especialmente Gradient Boosting) demostraron ser más adecuados para este problema de regresión.
- Se evidenció la importancia de evaluar múltiples métricas y visualizar los errores residuales para entender el comportamiento real del modelo.

Nota: Todos los detalles, código y visualizaciones de esta etapa se encuentran documentados en el archivo ***ModelosPredictivos.ipynb***.

Conclusiones

- Este proyecto demostró cómo el análisis exploratorio, la limpieza de datos y el modelado predictivo pueden integrarse para resolver un problema real: predecir el precio del aguacate a partir de características del mercado.
- La calidad del análisis descriptivo permitió entender la complejidad del dataset, revelando patrones temporales, diferencias por tipo y región, y relaciones entre el volumen y el precio.
- La fase de limpieza y normalización fue clave para asegurar la estabilidad de los modelos. La eliminación de outliers y la transformación adecuada de variables permitió entrenar modelos más precisos y generalizables.
- El entrenamiento de múltiples modelos mostró diferencias notables en capacidad predictiva. Mientras que la regresión lineal sirvió como baseline, modelos como Gradient Boosting capturaron relaciones no evidentes, logrando una predicción robusta y precisa.
- Se evidenció que la complejidad de los datos (interacción entre tipo de aguacate, región, empaque y tiempo) exige modelos capaces de manejar relaciones no lineales y estructuras de datos con muchas dimensiones (dummies, fechas, etc.).
- El proyecto no solo cumplió con el objetivo de predicción, sino que dejó herramientas analíticas listas para extenderse a escenarios reales como predicción por región, optimización de precios o análisis estacional.

Posibles Mejoras

- **Incorporación de variables exógenas**: incluir eventos externos como feriados, clima o demanda internacional (por ejemplo, cercanía al Super Bowl o festividades mexicanas) podría mejorar aún más la precisión del modelo.
- **Modelos de series temporales**: dado que hay una clara estacionalidad, sería valioso probar modelos como ARIMA, Prophet o LSTM, que capturan la evolución de precios a lo largo del tiempo de forma secuencial.
- **Feature engineering adicional**: construir nuevas variables como media móvil del precio, cambio porcentual entre semanas o volumen acumulado mensual permitiría al modelo entender dinámicas de mercado más profundas.
- **Validación cruzada más robusta**: se podría utilizar una validación k-fold estratificada o incluso TimeSeriesSplit para asegurar que el modelo no se sobreentrena con ciertos periodos.
- **Mejor tuning de hiperparámetros**: el uso de GridSearch fue adecuado, pero podría mejorarse con RandomizedSearch o herramientas como Optuna para una búsqueda más eficiente.
- **Interpretabilidad avanzada**: técnicas como SHAP o LIME podrían explicar qué variables afectan más al precio del aguacate por observación, dando más valor al modelo en contextos reales.
- **Evaluación regional personalizada**: segmentar el modelo por región o tipo de aguacate y entrenar modelos por subgrupo podría mejorar la precisión local.
- **Pipeline reproducible**: implementar un pipeline completo con scikit-learn.pipeline aseguraría mayor robustez y escalabilidad si se desea poner en producción el modelo.