UNIVERSIDAD ANDINA DEL CUSCO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Informe Dataset (Entrenamiento y predicción)

ASIGNATURA: INTELIGENCIA ARTIFICIAL

DOCENTE: ESPETIA HUAMANGA HUGO

ESTUDIANTES:

CUSI RONCO JHOEL

HUARACHI PUMACHAPI JUAN ALBERTO

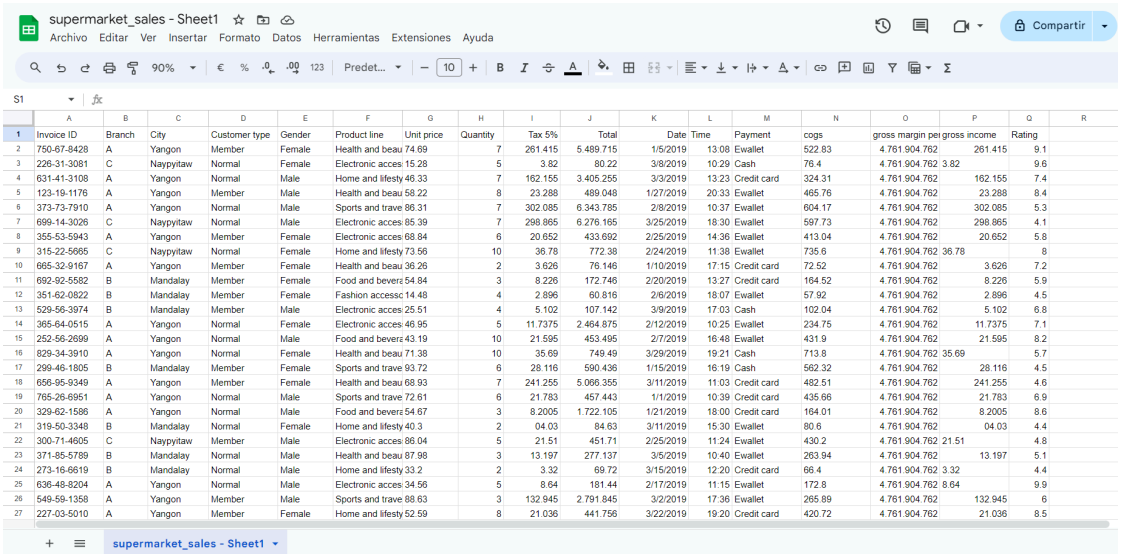
MENDOZA CHOQUEHUILLCA ULISES VALENTY

QUISPE CCOPA EVELYN

CUSCO - PERÚ

2024

1. Limpieza del DataSet



supermarket_sales - Sheet1

Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones Ayuda

Compartir

Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin per	gross income	Rating
750-67-8428	A	Yangon	Member	Female	Health and beau	74.69	7	261.415	5.489.715	1/5/2019	13:08	Ewallet	522.83	4.761.904.762	261.415	9.1
226-31-3081	C	Naypyitaw	Normal	Female	Electronic acces	15.28	5	3.82	80.22	3/8/2019	10:29	Cash	76.4	4.761.904.762	3.82	9.6
631-41-3108	A	Yangon	Normal	Male	Home and lifesty	46.33	7	162.155	3.405.255	3/3/2019	13:23	Credit card	324.31	4.761.904.762	162.155	7.4
123-19-1176	A	Yangon	Member	Male	Health and beau	58.22	8	23.288	489.048	1/27/2019	20:33	Ewallet	465.76	4.761.904.762	23.288	8.4
373-73-7910	A	Yangon	Normal	Male	Sports and trave	86.31	7	302.085	6.343.785	2/8/2019	10:37	Ewallet	604.17	4.761.904.762	302.085	5.3
699-14-3026	C	Naypyitaw	Normal	Male	Electronic acces	85.39	7	298.865	6.276.165	3/25/2019	18:30	Ewallet	597.73	4.761.904.762	298.865	4.1
355-53-5943	A	Yangon	Member	Female	Electronic acces	68.84	6	20.652	433.692	2/25/2019	14:36	Ewallet	413.04	4.761.904.762	20.652	5.8
315-22-5665	C	Naypyitaw	Normal	Female	Home and lifesty	73.56	10	36.78	772.38	2/24/2019	11:38	Ewallet	735.6	4.761.904.762	36.78	8
665-32-9167	A	Yangon	Member	Female	Health and beau	36.26	2	3.626	76.146	1/10/2019	17:15	Credit card	72.52	4.761.904.762	3.626	7.2
692-92-5582	B	Mandalay	Member	Female	Food and bevera	54.84	3	8.226	172.746	2/20/2019	13:27	Credit card	164.52	4.761.904.762	8.226	5.9
351-62-0822	B	Mandalay	Member	Female	Fashion accesso	14.48	4	2.896	60.816	2/6/2019	18:07	Ewallet	57.92	4.761.904.762	2.896	4.5
529-56-3974	B	Mandalay	Member	Male	Electronic acces	25.51	4	5.102	107.142	3/9/2019	17:03	Cash	102.04	4.761.904.762	5.102	6.8
365-64-0515	A	Yangon	Normal	Female	Electronic acces	46.95	5	11.7375	2.464.875	2/12/2019	10:25	Ewallet	234.75	4.761.904.762	11.7375	7.1
252-56-2699	A	Yangon	Normal	Male	Food and bevera	43.19	10	21.595	453.495	2/7/2019	16:48	Ewallet	431.9	4.761.904.762	21.595	8.2
829-34-3910	A	Yangon	Normal	Female	Health and beau	71.38	10	35.69	749.49	3/29/2019	19:21	Cash	713.8	4.761.904.762	35.69	5.7
299-46-1805	B	Mandalay	Member	Female	Sports and trave	93.72	6	28.116	590.436	1/15/2019	16:19	Cash	562.32	4.761.904.762	28.116	4.5
656-95-9349	A	Yangon	Member	Female	Health and beau	68.93	7	241.255	5.066.355	3/11/2019	11:03	Credit card	482.51	4.761.904.762	241.255	4.6
765-26-6951	A	Yangon	Normal	Male	Sports and trave	72.61	6	21.783	457.443	1/1/2019	10:39	Credit card	435.66	4.761.904.762	21.783	6.9
329-62-1586	A	Yangon	Normal	Male	Food and bevera	54.67	3	8.2005	1.722.105	1/21/2019	18:00	Credit card	164.01	4.761.904.762	8.2005	8.6
319-50-3348	B	Mandalay	Normal	Female	Home and lifesty	40.3	2	04.03	84.63	3/11/2019	15:30	Ewallet	80.6	4.761.904.762	04.03	4.4
300-71-4605	C	Naypyitaw	Member	Male	Electronic acces	86.04	5	21.51	451.71	2/25/2019	11:24	Ewallet	430.2	4.761.904.762	21.51	4.8
371-85-5789	B	Mandalay	Normal	Male	Health and beau	87.98	3	13.197	277.137	3/5/2019	10:40	Ewallet	263.94	4.761.904.762	13.197	5.1
273-16-6619	B	Mandalay	Normal	Male	Home and lifesty	33.2	2	3.32	69.72	3/15/2019	12:20	Credit card	66.4	4.761.904.762	3.32	4.4
636-48-8204	A	Yangon	Normal	Male	Electronic acces	34.56	5	8.64	181.44	2/17/2019	11:15	Ewallet	172.8	4.761.904.762	8.64	9.9
549-59-1358	A	Yangon	Member	Male	Sports and trave	88.63	3	132.945	2.791.845	3/2/2019	17:36	Ewallet	265.89	4.761.904.762	132.945	6
227-03-5010	A	Yangon	Member	Female	Home and lifesty	52.59	8	21.036	441.756	3/22/2019	19:20	Credit card	420.72	4.761.904.762	21.036	8.5

supermarket_sales - Sheet1

Variables innecesarias

Invoice ID: Es simplemente un identificador único de las facturas y no tiene influencia en la predicción.

cogs (Cost of Goods Sold): Este valor está relacionado con los costos internos de la empresa, pero no influye directamente en la demanda. Dado que ya se incluye el precio unitario (Unit price) y el total.

gross margin percentage: Este porcentaje refleja la relación entre el costo de los productos vendidos y los ingresos totales. Como el objetivo es predecir la demanda, este dato no influye directamente en cuántos productos se venderán.

gross income: El ingreso bruto está directamente relacionado con el precio y la cantidad vendida, pero no aporta más información que ya no esté en otras variables como Total o Quantity.

Variable dependiente::

Total de ventas o Cantidad de productos vendidos (Quantity):

Variables independientes::

Sucursal: Diferentes sucursales pueden tener diferentes patrones de demanda.

Tipo de producto (Productline): La categoría o tipo de producto afectará la demanda.

Método de pago (Payment): Los métodos de pago pueden influir en las compras (algunos métodos podrían ser preferidos en días específicos).

Fecha (Date): La demanda puede variar según la fecha (estacionalidad, días de la semana, festivos, etc.).

Tiempo de compra (Time): La hora del día puede afectar las ventas (por ejemplo, más ventas en horas pico).

Impuestos (Tax 5%): Podrías incluir el impuesto como un factor para ver si tiene influencia en la compra de productos.

DATASET ORIGINAL

```
Información del dataset original:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Branch         1000 non-null   object
 1   City           1000 non-null   object
 2   Customer type  1000 non-null   object
 3   Gender         1000 non-null   object
 4   Product line   1000 non-null   object
 5   Unit price     1000 non-null   float64
 6   Quantity       1000 non-null   int64
 7   Tax 5%         1000 non-null   float64
 8   Total          1000 non-null   float64
 9   Date           1000 non-null   object
 10  Time           1000 non-null   object
 11  Payment        1000 non-null   object
 12  Rating         1000 non-null   float64
dtypes: float64(4), int64(1), object(8)
memory usage: 101.7+ KB
None
```

DATASET LIMPIADA

```
Información del dataset limpio:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 13 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Branch         1000 non-null   object
 1   City           1000 non-null   object
 2   Customer type  1000 non-null   object
 3   Gender         1000 non-null   object
 4   Product line   1000 non-null   object
 5   Unit price     1000 non-null   float64
 6   Quantity       1000 non-null   int64
 7   Tax 5%         1000 non-null   float64
 8   Total          1000 non-null   float64
 9   Date           1000 non-null   object
 10  Time           1000 non-null   object
 11  Payment        1000 non-null   object
 12  Rating         1000 non-null   float64
dtypes: float64(4), int64(1), object(8)
memory usage: 101.7+ KB
None
```

2. Análisis del DataSet

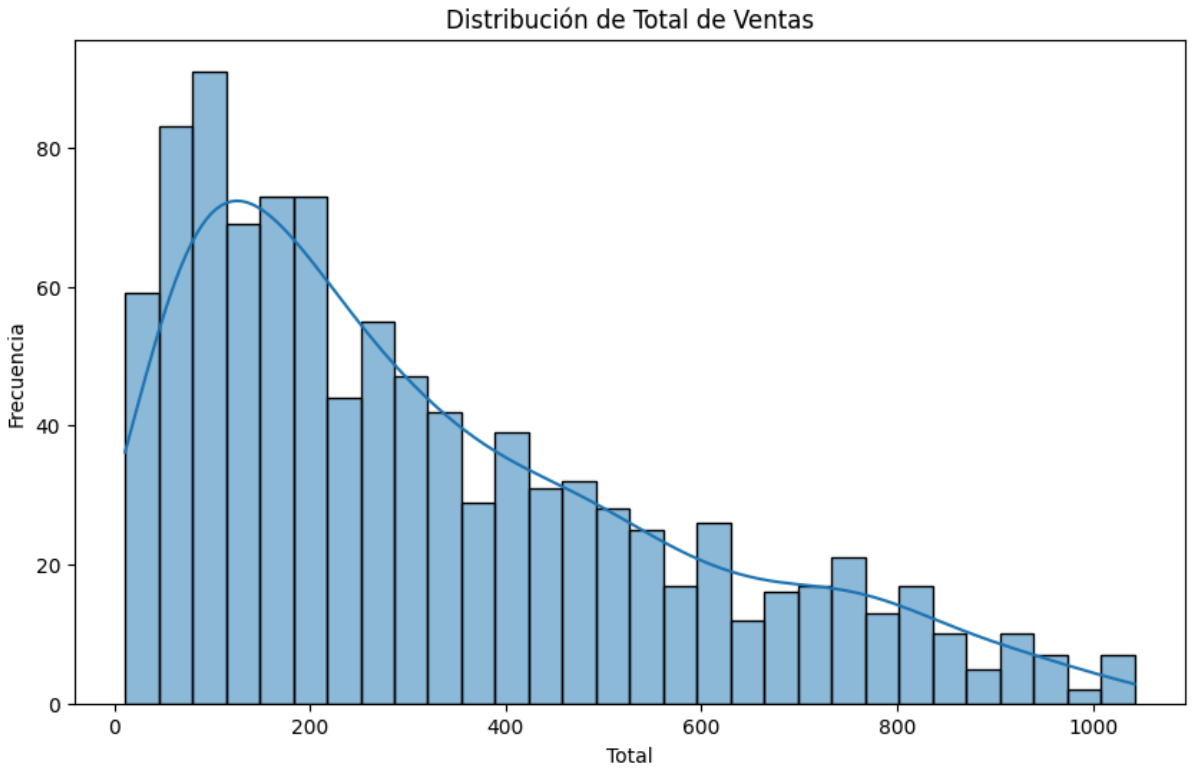
Propósito: Comprender las características y la estructura del dataset para informar el desarrollo de los modelos de predicción.

Descripción:

- **Exploración de datos:** Investigar la distribución de variables, estadísticas descriptivas y posibles correlaciones.
- **Visualización:** Utilizar gráficos para explorar relaciones entre variables, como histogramas para la distribución de precios y ventas, y diagramas de dispersión para observar correlaciones.

Estadísticas descriptivas del dataset limpio:

	Unit price	Quantity	Tax 5%	Total	Rating
count	1000.000000	1000.000000	1000.000000	1000.000000	1000.00000
mean	55.672130	5.510000	15.379369	322.966749	6.97270
std	26.494628	2.923431	11.708825	245.885335	1.71858
min	10.080000	1.000000	0.508500	10.678500	4.00000
25%	32.875000	3.000000	5.924875	124.422375	5.50000
50%	55.230000	5.000000	12.088000	253.848000	7.00000
75%	77.935000	8.000000	22.445250	471.350250	8.50000
max	99.960000	10.000000	49.650000	1042.650000	10.00000



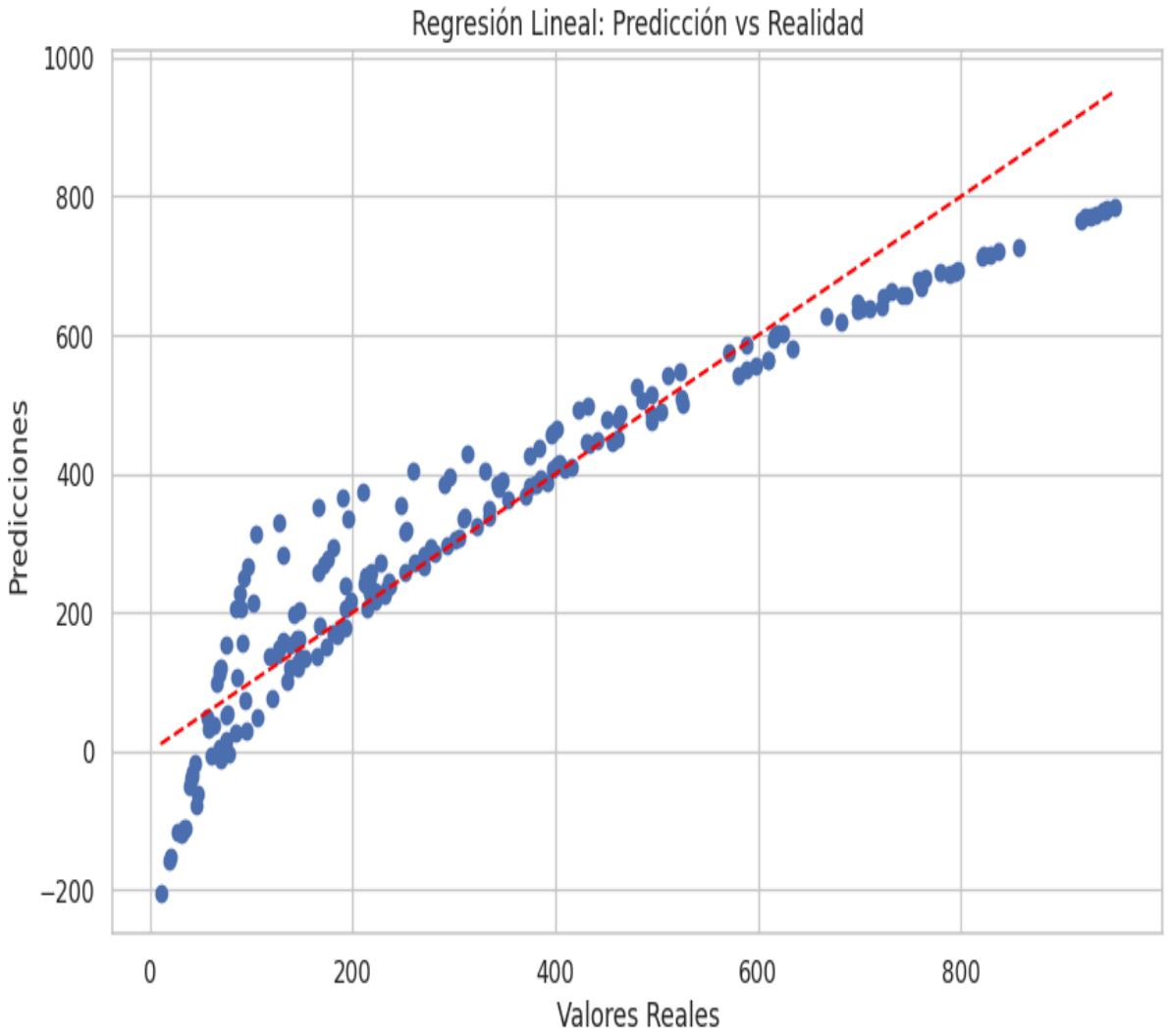
3. Implementación de Modelos de Aprendizaje Automático

Propósito: Utilizar técnicas de aprendizaje automático para construir modelos predictivos que estimen la demanda de productos.

Descripción:

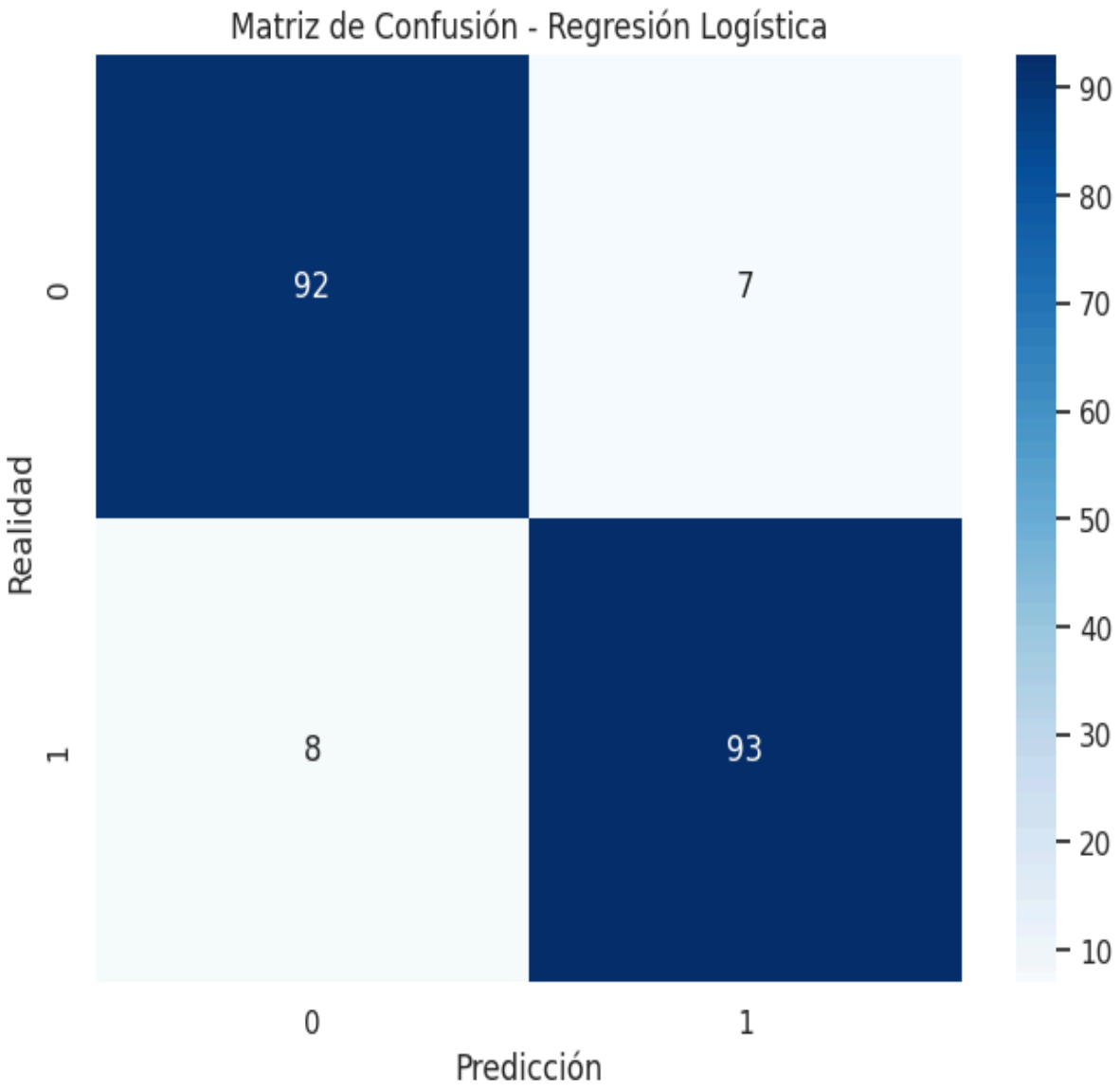
- **Modelo de Regresión Lineal (RL):** Ajustar un modelo de regresión lineal para predecir la demanda basada en características como el precio unitario.

```
MSE Regresión Lineal: 6228.045510688692
```



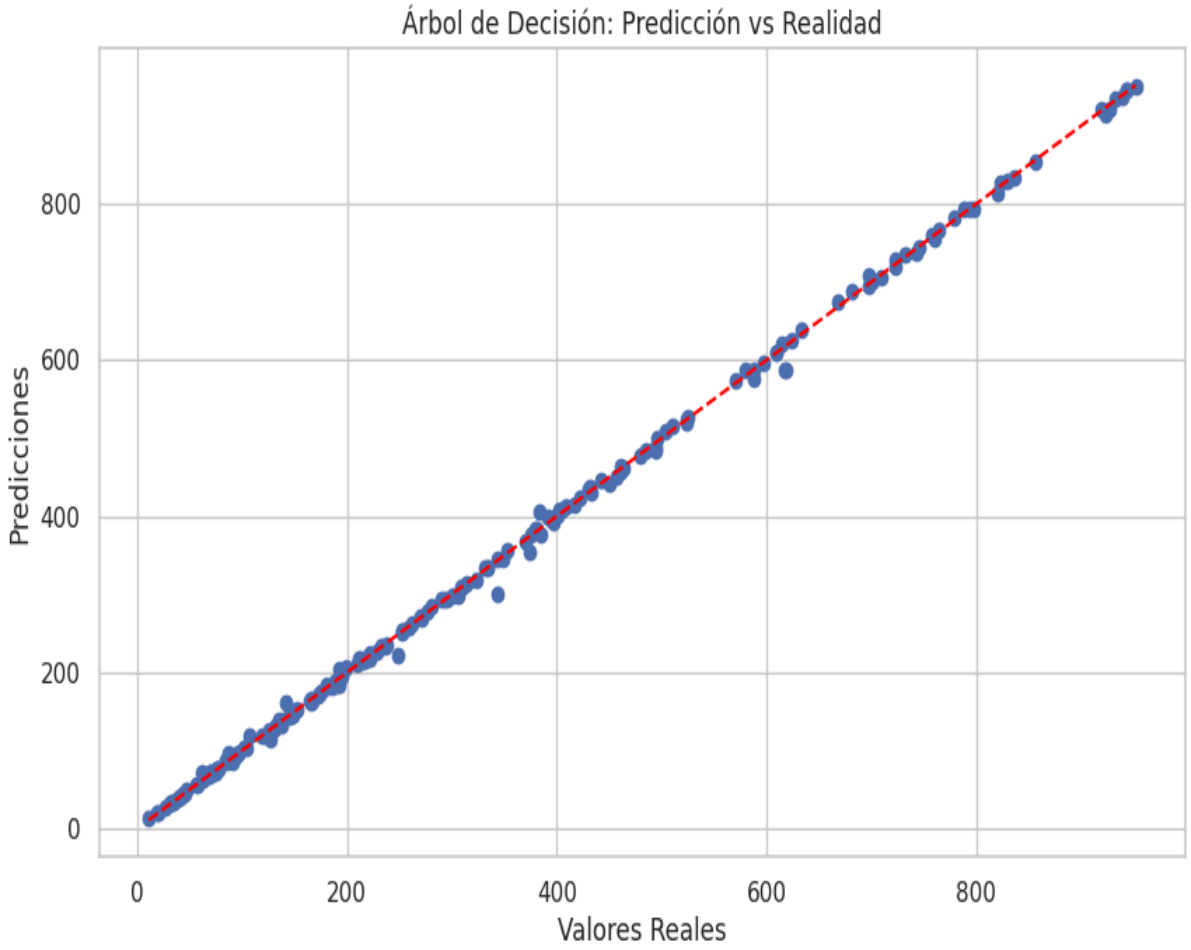
- **Modelo de Regresión Logística (RLog):** Si la variable objetivo es categórica (por ejemplo, alta/ baja demanda), aplicar un modelo de regresión logística.

```
Accuracy Regresión Logística: 0.925
```



- **Árboles de Decisión:** Implementar un modelo de árboles de decisión para capturar relaciones no lineales entre las características y la demanda.

```
MSE Árbol de Decisión: 43.20293764500004
```



4. Implementación de Modelos de Aprendizaje Automático

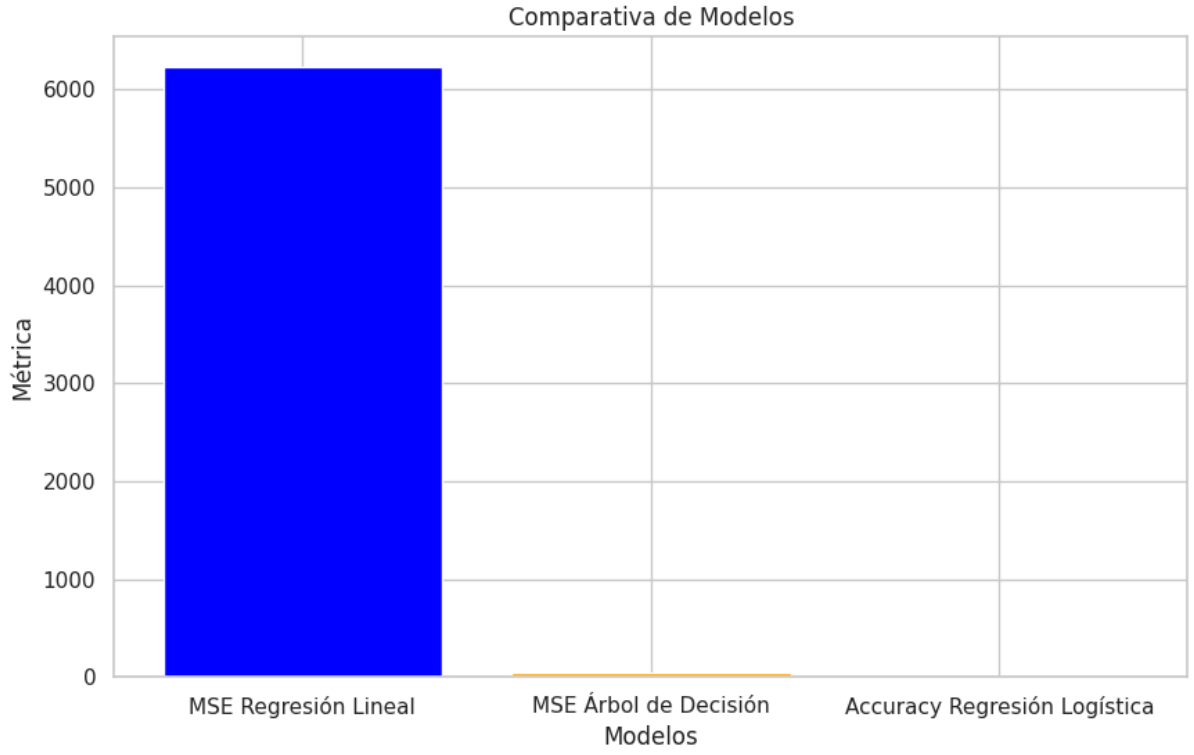
Propósito: implementa y entrena los modelos de Regresión Lineal y Árboles de Decisión, y realiza predicciones.

Descripción:

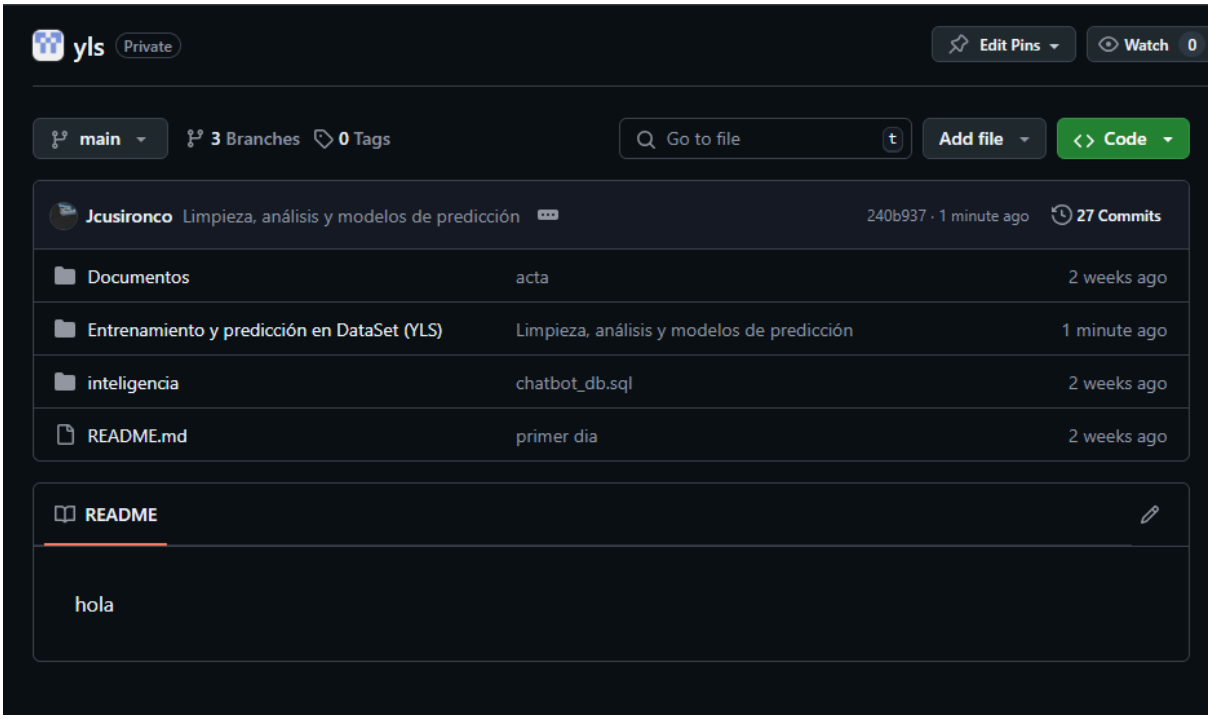
- **Regresión Lineal (RL):** Idealmente, debería tener un MSE bajo y un R2 alto. Si obtienes MSE de 0 y R2 de 1.0, puede indicar un problema con el conjunto de datos (por ejemplo, falta de variabilidad o datos demasiado simples).
- **Árboles de Decisión:** También deberías buscar un MSE bajo y un R2 alto. Los árboles de decisión pueden manejar relaciones no lineales y complejas mejor que la regresión lineal en algunos casos.

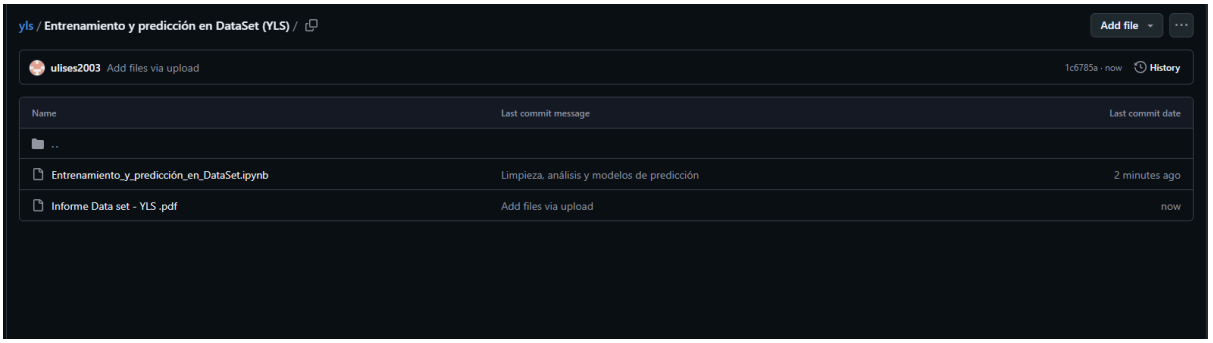
```
Comparativa de Modelos:
MSE Regresión Lineal: 6228.045510688692
MSE Árbol de Decisión: 43.20293764500004
Accuracy Regresión Logística: 0.925
```

```
El mejor modelo es: Árbol de Decisión
```



Un repositorio para los modelos y un PDF para el informe





Conclusión

1. **Regresión Lineal:** Suele ser adecuada si hay una relación lineal entre las características y la variable objetivo. Puede ser sensible a valores atípicos y no manejar bien relaciones no lineales.
2. **Árboles de Decisión:** Pueden capturar relaciones no lineales y son más flexibles. Sin embargo, pueden ser propensos a sobreajustarse (overfitting) si no se ajustan adecuadamente.