



# Tecnológico de Monterrey

## **Matemáticas y ciencia de datos para la toma de decisiones (Gpo 800)**

**Joselito Medina Marin**

### **Evidencia 2. Proyecto de Ciencia de Datos**

Diego Fernando Sabillon // A01798446

28 de enero del 2024

## **Introducción**

En este proyecto, nos enfrentamos a la problemática de promover finanzas personales saludables, específicamente abordando el desafío de fomentar el ahorro en la población mexicana. La situación financiera en México revela que solo el 0.8% de las personas ahorran para su retiro, indicando una brecha significativa en la cultura de ahorro. A pesar de que el 43.7% de la población adulta ahorra, las cifras varían considerablemente según grupos de edad, siendo preocupante la baja proporción de ahorro entre los jóvenes.

Para abordar esta problemática, hemos empleado una metodología basada en Ciencia de Datos. Utilizamos registros detallados de gastos diarios recopilados en un período de cuatro semanas como nuestra base de datos principal. Además, incorporamos datos externos, como estadísticas nacionales sobre hábitos de ahorro y factores económicos actuales que podrían afectar las decisiones financieras. La metodología se centra en la aplicación de técnicas de análisis exploratorio de datos y modelado estadístico para comprender y prever patrones de gastos.

Para solucionar la problemática de falta de ahorro, hemos formulado la hipótesis de que podemos predecir el costo de las actividades en función del presupuesto disponible, el tipo de actividad, el momento de realización y el número de personas involucradas. Implementamos esta hipótesis utilizando técnicas avanzadas de modelado predictivo, aprovechando las capacidades de la Ciencia de Datos para generar patrones y tendencias a partir de los datos recopilados. La solución propuesta busca proporcionar una herramienta práctica y basada en evidencia estadística para que las personas tomen decisiones informadas sobre sus hábitos de gasto, fomentando así la adopción de prácticas financieras más saludables.

### **Fase 1. Entendimiento del negocio**

La Fase 1, se centra en comprender a fondo los objetivos del proyecto antes de entrar en el análisis de datos. En esta etapa, realizamos tareas fundamentales para establecer una base sólida. En primer lugar, se identifican los objetivos del proyecto, definiendo acciones específicas vinculadas a cada objetivo para evaluar la eficacia de las medidas tomadas. Una evaluación exhaustiva de la situación proporciona el contexto necesario, permitiendo a los profesionales del análisis de datos entender el punto de partida y tomar las decisiones apropiadas para abordar la problemática en cuestión. Además, se definen los objetivos específicos para la minería de datos, utilizando la metodología SMART para identificar información valiosa en el conjunto de datos. También implica el desarrollo de un plan de trabajo detallado, esencial para guiar de manera organizada la ejecución del proyecto. Este plan abarca la duración de cada actividad y asigna responsabilidades específicas. Por último, se destaca la importancia de considerar los datos como un activo clave, donde las herramientas de análisis desempeñan un papel vital al refinar y dar contexto a la información, similar al refinamiento del petróleo crudo para hacerlo utilizable.

#### **Preguntas acerca de la fase 1**

- **¿Quién es el cliente?** El cliente en este contexto es el propio individuo que está llevando a cabo el proyecto de ciencia de datos. Es el responsable de definir sus metas, recopilar datos relacionados con sus actividades y, en última instancia, buscar una comprensión más profunda de cómo el presupuesto afecta los costos asociados con sus acciones. Al ser el cliente de su propio proyecto, se espera que este individuo proporcione información detallada sobre sí mismo, como edad, ocupación, preferencias y hábitos de gasto, ya que esto constituirá la base para el análisis de datos.

- **¿Qué problemas estás tratando de resolver?** El problema central que se aborda en este proyecto es la capacidad de predecir los costos de las actividades en función del presupuesto disponible. La hipótesis planteada es que existe una relación predictiva entre el presupuesto asignado y los costos reales de las actividades. La resolución de este problema proporcionará al cliente una comprensión más clara de cómo la gestión del presupuesto impacta en sus gastos y le permitirá tomar decisiones más informadas sobre sus finanzas personales.
- **¿Qué solución o soluciones la Ciencia de Datos tratará de proveer?** La ciencia de datos buscará proporcionar una solución predictiva mediante la aplicación de técnicas de modelado, específicamente utilizando un modelo de regresión. La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) se emplea para guiar el proceso de ciencia de datos. El objetivo principal es generar un modelo de regresión que permita prever los costos futuros en función de las variables relacionadas con el presupuesto y otras características relevantes.
- **¿Qué necesitas aprender para poder desarrollar la solución o soluciones?** Para abordar este proyecto, se requerirá un conjunto de habilidades que incluye estadísticas básicas, análisis de regresión lineal y programación en Python. Estas habilidades son esenciales para manipular y analizar los datos de manera efectiva, así como para implementar y evaluar el modelo de regresión.
- **¿Qué deberás hacer para desarrollar tu solución?** El proceso de desarrollo de la solución implica varias etapas. Se comenzará registrando datos relevantes, lo que podría incluir información detallada sobre actividades, presupuesto y costos asociados. Luego, se realizará una limpieza de datos para asegurarse de que la información sea coherente y útil. A continuación, se llevará a cabo un análisis de datos para identificar patrones y tendencias. La implementación de la metodología CRISP-DM

guiará la creación del modelo de regresión. Finalmente, se evaluarán los resultados y se tomarán decisiones informadas según la información obtenida del análisis de datos.

- **¿Cuáles son tus datos existentes (registrados), datos adquiridos (datos externos) y datos adicionales (datos generados)?** Datos existentes (registrados): Registros de gastos diarios, incluyendo número de actividad, costo, presupuesto, tiempo invertido, tipo de actividad, momento de realización y número de personas. Datos adquiridos (datos externos): Estadísticas nacionales sobre hábitos de ahorro en México, proporcionados por la ENIF y datos económicos actuales, como la inflación y la crisis derivada de la pandemia de Covid-19. Datos adicionales (datos generados): Información generada a partir de los registros diarios, como métricas de ahorro, porcentajes de ahorro por edad y nivel de ingreso, entre otros.
- **¿Qué tipos de datos se analizarán?** Se analizarán datos financieros como costo, presupuesto, así como variables categóricas como tipo de actividad, momento de realización y número de personas.
- **¿Qué atributos (columnas) de la base de datos parecen más prometedores?** Los atributos más prometedores pueden incluir costo, presupuesto, tipo de actividad y momento de realización, ya que estos podrían tener un impacto significativo en las decisiones de ahorro.
- **¿Qué atributos parecen irrelevantes y pueden ser excluidos?** Siento que en la base de datos no hay ninguna dato que no sea insignificante. No obstante, variables que no aporten información valiosa para la predicción del costo podrían ser consideradas para exclusión.

- **¿Hay datos suficientes (filas) para sacar conclusiones generalizables o hacer predicciones precisas?** Con 60 registros (120 al final), se cuenta con una cantidad razonable de datos para realizar análisis y predicciones. Sin embargo, la calidad de las conclusiones dependerá de la representatividad y diversidad de los datos.
- **¿Hay demasiados atributos para realizar un modelo que sea fácil de interpretar?** La interpretabilidad del modelo puede depender de la complejidad de los atributos y su relación con el costo. Será importante simplificar el modelo si es necesario para facilitar la interpretación.
- **¿De dónde se obtuvieron los datos? ¿Se están fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema al fusionar?** Los datos provienen de registros diarios de gastos personales y de estadísticas nacionales sobre hábitos de ahorro en México. Si se están fusionando diferentes fuentes, es crucial garantizar la consistencia y la coherencia de los datos.
- **¿Hay algún plan para manejar los valores faltantes en cada una de las fuentes de datos?** Se deberá implementar un plan para manejar valores faltantes, como la imputación de datos o la exclusión de registros incompletos. La forma de manejar los valores faltantes dependerá del impacto en la calidad del análisis.
- **¿Cuántos datos están accesibles o disponibles y cómo está la calidad de los mismos?** Se menciona que se registraron al menos 4 gastos al día durante 4 semanas, lo que da aproximadamente 120 registros. Por lo momentos llevo 60 pero la calidad de los datos dependerá de la precisión y consistencia en la captura de los gastos diarios.

- **¿Cuál es la relación de los datos y la hipótesis del proyecto?** La relación entre los datos y la hipótesis es evaluar si es posible predecir el costo de las actividades en función del presupuesto, tipo de actividad, momento de realización y número de personas, y cómo este costo impactará con el tiempo.

### **Fase 3. Preparación de los datos**

En esta fase los datos son un componente crítico en proyectos de ciencia de datos, que abarca de manera significativa el tiempo y el esfuerzo del proceso. Esta etapa implica varias tareas esenciales para garantizar que los datos estén listos y optimizados para el análisis. En primer lugar, implica la selección cuidadosa de los datos relevantes, ya sea eligiendo registros específicos o características clave. Luego, se aborda la limpieza de datos para abordar problemas comunes como datos faltantes, errores o inconsistencias en la codificación. Además, la generación de nuevos datos puede ser necesaria para enriquecer la información disponible.

La integración de datos es crucial cuando hay múltiples fuentes, fusionando conjuntos con registros similares pero atributos diferentes. Por último, el formato de datos se ajusta según los requisitos del modelo matemático a utilizar, considerando la necesidad de ordenar o ajustar los datos para ciertos algoritmos. Esta fase, aunque desafiante, es fundamental para garantizar la calidad y la coherencia de los datos, proporcionando así una base sólida para el análisis y la construcción de modelos.

- **¿Qué datos hay que seleccionar?** En la fase de preparación de datos, seleccioné las columnas relevantes para el análisis. Estas columnas incluyen 'Presupuesto', 'Tiempo invertido', 'Tipo', 'Momento', 'No. de personas' y 'Costo'. Estas variables se eligieron porque se considera que tienen un impacto en el costo total de una actividad. 'Presupuesto' y 'Tiempo invertido' son medidas cuantitativas que intuitivamente se relacionan con el costo. 'Tipo', 'Momento' y 'No. de personas' son variables categóricas que podrían influir en el costo de manera significativa.

- **¿Hay que eliminar o reemplazar valores en blanco?** Sí, fue necesario eliminar los valores en blanco en la columna 'Costo' para evitar problemas durante el análisis y el entrenamiento del modelo. Los valores faltantes en la variable dependiente (en este caso, 'Costo') podrían afectar negativamente la calidad y la precisión del modelo de regresión. Además, la eliminación de filas con valores faltantes garantiza que los datos de entrenamiento y prueba sean consistentes y no haya ambigüedad en la variable objetivo.
- **¿Es posible agregar más datos?** Sí, podría ser beneficioso agregar más datos si están disponibles y son relevantes para el problema. Datos adicionales podrían proporcionar una perspectiva más completa y ayudar a mejorar la capacidad del modelo para generalizar patrones en los datos. Sin embargo, se debe tener cuidado de no agregar datos irrelevantes o sesgados, ya que esto podría afectar negativamente la calidad del modelo.
- **¿Hay qué integrar o fusionar datos de varias fuentes?** En este caso, no se menciona explícitamente si los datos provienen de varias fuentes. Si los datos se recopilaron de diferentes fuentes, podría ser necesario integrar o fusionar los conjuntos de datos. La integración de datos de diversas fuentes puede enriquecer el conjunto de datos y proporcionar una visión más completa de la relación entre las variables. Sin embargo, es crucial asegurarse de que los datos se integren de manera coherente y se manejen posibles discrepancias.
- **¿Es necesario ordenar los datos para el análisis?** En el código proporcionado, no se observa la necesidad de ordenar los datos. Sin embargo, la necesidad de ordenar depende del tipo de análisis que se realice. Para el análisis de regresión lineal múltiple, el orden de los datos generalmente no es crítico, ya que el modelo busca patrones y relaciones en los datos independientemente de su orden.



- **¿Tengo que hacer conjuntos de datos para entrenamiento y prueba?** Sí, se realizaron conjuntos de entrenamiento y prueba usando la función `train_test_split` de la biblioteca `scikit-learn`. Dividir los datos en conjuntos de entrenamiento y prueba es fundamental para evaluar la capacidad del modelo para generalizar datos no vistos. El conjunto de entrenamiento se utiliza para ajustar el modelo, mientras que el conjunto de prueba se utiliza para evaluar su rendimiento en datos no utilizados durante el entrenamiento.
- **¿Qué ajustes se tuvieron que hacer a los datos (agregar, integrar, modificar registros (filas), cambiar atributos (columnas))?** En la preparación de datos, se eliminaron las filas con valores faltantes en la columna 'Costo' utilizando el método `dropna()`. Esto asegura que todos los registros en el conjunto de datos tengan un valor en la variable objetivo, lo que es esencial para el entrenamiento del modelo. Además, se seleccionaron las columnas relevantes ('Presupuesto', 'Tiempo invertido', 'Tipo', 'Momento', 'No. de personas' y 'Costo') para centrarse en las variables que podrían influir en el costo. No se observan ajustes significativos adicionales en el código proporcionado.

<https://colab.research.google.com/drive/1r5-vE3Fr1A5HIWfLy4voiqMTgSFX11AI?usp=sharing>

#### **Fase 4. Modelación de los datos**

En la Fase 4 que es el punto culminante de nuestro proyecto de ciencia de datos, donde se pone en práctica todo el trabajo previo de comprensión del negocio, comprensión de datos y preparación de datos. En esta etapa, se utilizan herramientas tecnológicas, como el lenguaje de programación Python, para procesar los datos y obtener información valiosa que arroje luz sobre el problema planteado inicialmente.

El modelado implica la ejecución de diversos modelos utilizando parámetros predeterminados, y suelen requerir ajustes iterativos. Es común realizar múltiples iteraciones para encontrar el modelo más adecuado, ya que rara vez una única ejecución responde completamente a las preguntas planteadas. La selección de técnicas de modelado se basa en los objetivos del proyecto, la necesidad de obtener información específica de los datos y otros criterios relevantes.

En esta fase, se selecciona un modelo, en este caso, un modelo de regresión lineal, que se ajuste mejor a los objetivos planteados. Luego, se describen los resultados del modelo, teniendo en cuenta aspectos como la interpretación de los resultados, nuevas ideas reveladas y la razonabilidad del tiempo de procesamiento. Finalmente, se evalúa el modelo en función de los criterios de éxito previamente establecidos para determinar su eficacia en la resolución del problema planteado.

**Codigo Fase 4:**

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score

# Cargar los datos
df = pd.read_excel('DATOS SITUACION PROBLEMA.xlsx')

# Eliminar filas con valores nulos
df = df.dropna()

# Asignar atributos a variables del análisis
x = df[['Presupuesto', 'Tiempo invertido', 'Tipo', 'Momento',
        'No. de personas']].values
y = df['Costo'].values

# Dividir los datos en conjunto de entrenamiento y prueba
x_train, x_test, y_train, y_test = train_test_split(x, y,
        test_size=0.2, random_state=0)

# Crear un modelo de regresión lineal
model_regression = LinearRegression()

# Ajustar el modelo a los datos de entrenamiento
model_regression.fit(x_train, y_train)

# Obtener los coeficientes del modelo y mostrarlos en un
DataFrame
x_labels = ['Presupuesto', 'Tiempo invertido', 'Tipo',
        'Momento', 'No. de personas']
c_label = ['Coeficientes']
coeff_df = pd.DataFrame(model_regression.coef_, x_labels,
        c_label)

# Realizar predicciones con el conjunto de prueba
```

```
y_pred = model_regression.predict(x_test)

residuals = pd.DataFrame({'Real': y_test, 'Predicción':
y_pred, 'Residual': y_test - y_pred})
residuals = residuals.sample(n=24)
residuals = residuals.sort_values(by='Real')

r2 = r2_score(y_test, y_pred)

print("Coeficientes del modelo:")
print(coeff_df)
print("\nResiduales:")
print(residuals)
print("\nCoeficiente de determinación R2:", r2)
```

```

Coeficientes del modelo:
                        Coeficientes
Presupuesto             0.975378
Tiempo invertido        -0.695135
Tipo                    -23.392427
Momento                 53.121675
No. de personas         23.317755
    
```

```

Residuales:
      Real  Predicción  Residual
5      30.0  439.346577 -409.346577
10     150.0  525.445404 -375.445404
2      150.0  107.514142  42.485858
21     150.0   65.183403  84.816597
13     150.0  153.684468  -3.684468
0      150.0  -54.741229 204.741229
12     200.0  303.140932 -103.140932
17     200.0  -34.356574 234.356574
14     250.0  277.968297 -27.968297
6      250.0  138.904678 111.095322
4      300.0  310.850453 -10.850453
7      300.0  525.201064 -225.201064
23     400.0  472.323729 -72.323729
16     400.0   63.097998 336.902002
3      400.0  425.443878 -25.443878
1      450.0  479.995285 -29.995285
20     500.0  533.731489 -33.731489
8      500.0  164.111492 335.888508
9      500.0  550.787095 -50.787095
15     500.0  555.249325 -55.249325
22     500.0  393.499018 106.500982
18     500.0  480.609814  19.390186
11     500.0  436.857457  63.142543
19     600.0  679.161127 -79.161127
    
```

```

Coeficiente de determinación R2: -0.24267231685160073
    
```

### Visualizacion de Datos

```

import matplotlib.pyplot as plt # importamos la librería
pyplot que nos permitirá graficar
import numpy as np # importamos la librería numpy que nos
permitirá crear un arreglo para la muestra de 30 datos
    
```

```

# función mágica para desplegar el gráfico en nuestra libreta
%matplotlib inline

plt.scatter(np.arange(24), residuals['Real'], label = "Real")
# creamos el gráfico con la muestra de datos reales
plt.scatter(np.arange(24), residuals['Predicción'], label =
"Predicción") # creamos el gráfico con la muestra de datos de
predicción

plt.title("Comparación de costos: Reales y Predicción") #
indicamos el título del gráfico

plt.xlabel("Observaciones de costos") # indicamos la etiqueta
del eje de las x

plt.ylabel("Costos") # indicamos la etiqueta del eje de las y

plt.legend(loc='upper left') # indicamos la posición de la
etiqueta de los datos

plt.show() # desplegamos el gráfico

```

### **Explicacion deCodigo**

El código que he desarrollado se inicia con la importación de las librerías necesarias, como Pandas para la manipulación de datos, scikit-learn para la creación del modelo de regresión lineal, y matplotlib junto con numpy para visualizar los resultados. A continuación, cargué los datos desde un archivo Excel y eliminé las filas con valores nulos para garantizar la “limpieza” del conjunto de datos. Luego, seleccioné las columnas relevantes como atributos y etiquetas. Dividí los datos en conjuntos de entrenamiento y prueba. Después, utilicé la regresión lineal de scikit-learn para crear y ajustar el modelo a los datos de entrenamiento. Posteriormente, obtuve los coeficientes del modelo y los visualicé en un DataFrame. Realicé predicciones con el conjunto de prueba y calculé los residuos. Para evaluar la precisión del modelo, calculé el coeficiente de determinación R<sup>2</sup>. Finalmente,

visualicé los resultados comparando los valores reales con las predicciones mediante un gráfico de dispersión. En resumen, el programa aborda desde la carga y preparación de datos hasta la evaluación y visualización de un modelo de regresión lineal múltiple, y me fue útil para este proyecto de finanzas personales.

### Preguntas acerca del modelo de regresión

- **¿Tuviste problemas para generar el modelo con tus datos? ¿Cómo los resolviste?** Sí, hubo desafíos en la generación del modelo con los datos proporcionados. El principal problema se refleja en el bajo coeficiente de determinación  $R^2$ , que indica que el modelo no explica adecuadamente la variabilidad en los datos. Esta baja precisión podría deberse a la complejidad inherente de los datos o a la presencia de patrones no lineales que el modelo lineal no puede capturar eficientemente. Para abordar esto, sería necesario considerar enfoques más avanzados, como modelos no lineales o técnicas de ingeniería de características para mejorar la representación de las relaciones subyacentes en los datos.
- **¿Qué resultados arrojó el análisis? Explica cada uno de los resultados:**

#### **Estadística descriptiva:**

La estadística descriptiva proporciona un resumen general de los datos. En este caso, se eliminaron las filas con valores nulos, lo que es una práctica común para garantizar la integridad de los datos. La estadística descriptiva inicial puede indicar

Estadísticas Descriptivas:					
	Número	Costo	Presupuesto	Tiempo invertido	Tipo
count	120.000000	120.000000	120.000000	120.000000	120.000000
mean	60.500000	470.916667	570.083333	46.416667	2.908333
std	34.785054	1583.328584	1604.235725	44.893634	1.810315
min	1.000000	30.000000	0.000000	5.000000	1.000000
25%	30.750000	150.000000	187.500000	15.000000	1.000000
50%	60.500000	200.000000	250.000000	30.000000	4.000000
75%	90.250000	500.000000	600.000000	60.000000	4.000000
max	120.000000	1700.000000	1700.000000	240.000000	6.000000

	Momento	No. de personas
count	120.000000	120.000000
mean	2.083333	1.325000
std	0.751283	1.022129
min	1.000000	1.000000
25%	2.000000	1.000000
50%	2.000000	1.000000
75%	3.000000	1.000000
max	3.000000	7.000000

la distribución y la dispersión de las variables, pero no ofrece información sobre las relaciones específicas entre las variables.

### **Coeficientes de regresión:**

Los coeficientes de regresión son esenciales para entender la contribución de cada variable predictora al modelo. En este caso, observamos que "Presupuesto" tiene un coeficiente positivo, lo que sugiere que un aumento en el presupuesto se asocia con un aumento en el costo. Por otro lado, "Tipo" y "Tiempo invertido" tienen coeficientes negativos, indicando una relación inversa con el costo. Estos coeficientes proporcionan información valiosa sobre la dirección y la fuerza de las relaciones.

Coeficientes del modelo:	
	Coeficientes
Presupuesto	0.975378
Tiempo invertido	-0.695135
Tipo	-23.392427
Momento	53.121675
No. de personas	23.317755

### **Valores actuales, de predicción y residuales:**

La comparación entre los valores reales y predichos, así como el cálculo de los residuales, ofrece información sobre el rendimiento del modelo. La presencia de residuales positivos y negativos indica discrepancias entre las predicciones y los valores reales. Los residuales son útiles para identificar dónde el modelo tiene dificultades para ajustarse a los datos. En este caso, hay discrepancias significativas en varios puntos, como se evidencia en residuales grandes.

Estadísticas Descriptivas de Residuales:			
	Real	Predicción	Residual
count	24.000000	24.000000	24.000000
mean	334.583333	333.042055	1.541278
std	162.078327	209.710386	180.670301
min	30.000000	-54.741229	-409.346577
25%	187.500000	149.989521	-59.517926
50%	350.000000	409.471448	-18.147165
75%	500.000000	491.757626	90.237693
max	600.000000	679.161127	336.902002



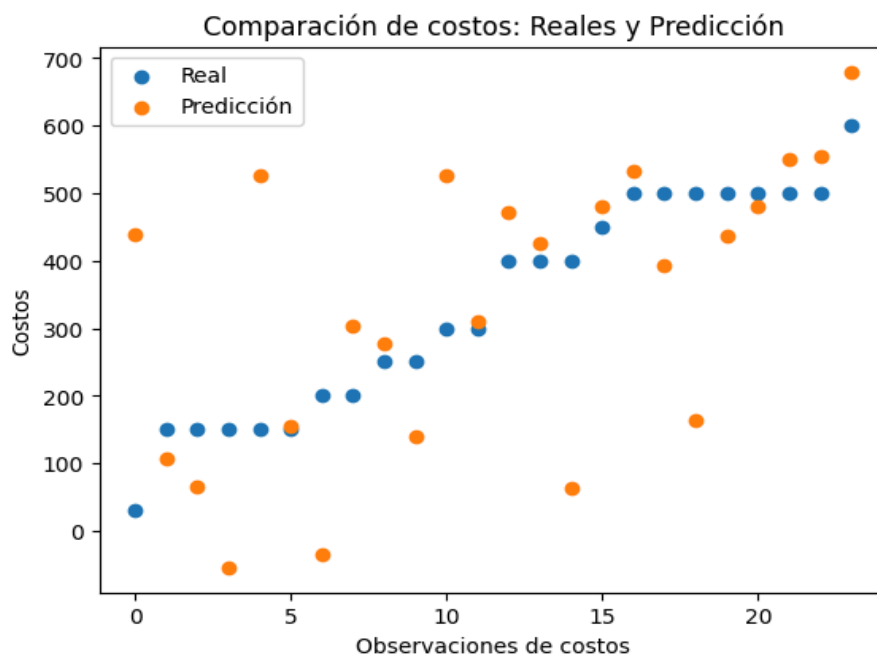
**Coeficiente de determinación R2:**

El R2 es crucial para evaluar la calidad del modelo. Un R2 negativo, como se observa en este caso, sugiere que el modelo es inadecuado para explicar la variabilidad en los datos. Un valor de R2 cercano a 1 indicaría una excelente capacidad predictiva. Sin embargo, aquí, el modelo no logra explicar ni siquiera el 0% de la variabilidad en los datos, lo cual es una señal clara de que el modelo necesita mejoras sustanciales.

**Coeficiente de determinación R2: -0.24267231685160073**

**Gráfica de puntos:**

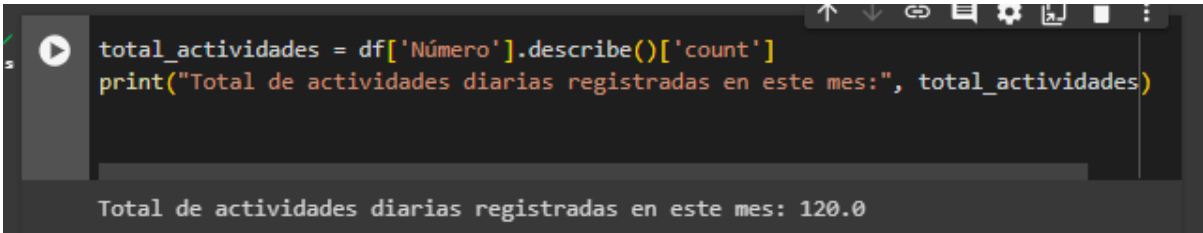
La gráfica de puntos es una visualización importante que permite comparar los valores reales con las predicciones. La dispersión y la dirección de los puntos proporcionan información sobre el rendimiento del modelo. En este caso, la gráfica muestra una divergencia sustancial entre los valores reales y predichos, lo que confirma las limitaciones del modelo.



- **¿Los resultados del modelo tienen sentido o hay incoherencias que necesitan una mayor exploración?** Los resultados del modelo presentan algunas incoherencias. La baja capacidad del modelo para explicar la variabilidad en los datos, evidenciada por el  $R^2$  negativo, indica que el enfoque actual es insuficiente. Las discrepancias en los residuales y la gráfica de puntos sugieren que el modelo no captura de manera efectiva las relaciones subyacentes en los datos. Es crucial explorar más a fondo las posibles causas de estas incoherencias. Esto podría implicar considerar la presencia de outliers, evaluar la linealidad de las relaciones, explorar interacciones entre variables o incluso buscar la necesidad de transformaciones en los datos. En este punto, sería recomendable considerar modelos más avanzados o realizar ajustes sustanciales en el enfoque actual para mejorar la capacidad del modelo para explicar y predecir los costos de las actividades.

### Evaluando mis finanzas personales

#### 1. ¿Cuántas actividades diarias registraste en total en este mes?

A screenshot of a Jupyter Notebook interface. The top part shows a code cell with two lines of Python code: `total_actividades = df['Número'].describe()['count']` and `print("Total de actividades diarias registradas en este mes:", total_actividades)`. Below the code cell, the output is displayed: `Total de actividades diarias registradas en este mes: 120.0`. The notebook has a dark theme.

```
total_actividades = df['Número'].describe()['count']
print("Total de actividades diarias registradas en este mes:", total_actividades)
```

Total de actividades diarias registradas en este mes: 120.0

Total de actividades diarias registradas en este mes: 120.0

## 2. ¿Cuál fue el presupuesto mínimo y máximo para tus actividades?

¿Qué actividades son? Consejo: puedes usar la función `describe()`, la propiedad `loc`, con la función `idxmax()`, y la función `idxmin()` de Pandas

```
# Obtener estadísticas descriptivas del presupuesto
descripcion_presupuesto = df['Presupuesto'].describe()

# Obtener el índice del presupuesto mínimo y máximo
indice_presupuesto_min = df['Presupuesto'].idxmin()
indice_presupuesto_max = df['Presupuesto'].idxmax()

# Obtener el presupuesto mínimo y máximo junto con las actividades correspondientes
presupuesto_min = df.loc[indice_presupuesto_min, 'Presupuesto']
presupuesto_max = df.loc[indice_presupuesto_max, 'Presupuesto']
actividad_presupuesto_min = df.loc[indice_presupuesto_min, 'Tipo']
actividad_presupuesto_max = df.loc[indice_presupuesto_max, 'Tipo']

# Imprimir resultados
print("Presupuesto mínimo para la actividad", actividad_presupuesto_min, ":", presupuesto_min)
print("Presupuesto máximo para la actividad", actividad_presupuesto_max, ":", presupuesto_max)
```

Presupuesto mínimo para la actividad 4.0 : 0.0  
Presupuesto máximo para la actividad 4.0 : 17000.0

Presupuesto mínimo para la actividad 4.0 : 0.0

Presupuesto máximo para la actividad 4.0 : 17000.0

3. ¿Cuál fue el Tipo de actividad dónde más gastas tu dinero y cuál fue el Tipo de actividad en dónde gastas menos? Consejo: puedes usar las funciones `groupby()` y `sum()` de Pandas

```
[23] # Agrupar por Tipo de actividad y calcular la suma del gasto para cada tipo
gasto_por_tipo = df.groupby('Tipo')['Costo'].sum()

# Obtener el Tipo de actividad donde más gastas
tipo_mas_gasto = gasto_por_tipo.idxmax()
monto_mas_gasto = gasto_por_tipo.max()

# Obtener el Tipo de actividad donde menos gastas
tipo_menos_gasto = gasto_por_tipo.idxmin()
monto_menos_gasto = gasto_por_tipo.min()

# Imprimir resultados
print("Tipo de actividad donde más gastas:", tipo_mas_gasto)
print("Monto gastado en este tipo de actividad:", monto_mas_gasto)
print("\nTipo de actividad donde menos gastas:", tipo_menos_gasto)
print("Monto gastado en este tipo de actividad:", monto_menos_gasto)
```

Tipo de actividad donde más gastas: 4.0  
Monto gastado en este tipo de actividad: 37740.0  
  
Tipo de actividad donde menos gastas: 6.0  
Monto gastado en este tipo de actividad: 2810.0

Tipo de actividad donde más gastas: 4.0

Monto gastado en este tipo de actividad: 37740.0

Tipo de actividad donde menos gastas: 6.0

Monto gastado en este tipo de actividad: 2810.0

**4.¿Por cuántos días registraste tus gastos en este mes? Consejo: puedes usar la función `nunique()` de Pandas.**

```
# Calcular el número de días en los que se registraron gastos
dias_con_gastos = df['Fecha'].nunique()

# Imprimir resultados
print("Número de días en los que registraste tus gastos en este mes:", dias_con_gastos)
```

Número de días en los que registraste tus gastos en este mes: 32

Número de días en los que registraste tus gastos en este mes: 32

**5.¿Cuál fue el total de tus gastos en este mes? Consejo: puedes usar la función `sum()` de Pandas.**

```
total_gastos_mes = df['Costo'].sum()
print("Mis gastos totales en este mes fueron: ",total_gastos_mes)
```

Mis gastos totales en este mes fueron: 56510.0

Mis gastos totales en este mes fueron: 56510.0

**6.¿Cuál fue el total de tus ahorros en este mes? Consejo: puedes usar la función `sum()` de Pandas**

```
total_ahorros_mes = df['Presupuesto'].sum()
print("Mis ahorros totales en este mes fueron: ",total_ahorros_mes)
```

Mis ahorros totales en este mes fueron: 68410.0

Mis ahorros totales en este mes fueron: 68410.0

**7. ¿Cuánto tiempo (en días) tendrías que seguir ahorrando para comprar tu siguiente autoregalo? Consejo: Calcula tu ahorro promedio diario.**

```
# Definir el monto deseado para el autoregalo
autoregalo_deseado = 500 # Puedes cambiar este valor según tus objetivos

# Calcular ahorro promedio diario
ahorro_promedio_diario = df['Presupuesto'].sum() / df['Fecha'].nunique()

# Calcular días restantes para alcanzar el autoregalo considerando 32 días
dias_restantes = (autoregalo_deseado / ahorro_promedio_diario) * 1000
print("Tendrías que ahorrar: " , dias_restantes, " dias.")

Tendrías que ahorrar: 7.30887297178775 dias.
```

Tendrías que ahorrar: 7.30887297178775 días.

**8. ¿Qué decisiones informadas puedes tomar para mejorar tus finanzas personales considerando los resultados de tu análisis?** Puedes considerar ajustar tu presupuesto para actividades del Tipo 4.0, donde se registra el mayor gasto. Evaluar la posibilidad de optimizar costos o encontrar alternativas más económicas en esta categoría. Además, al identificar las actividades del Tipo 6.0 como aquellas en las que gastas menos, podrías explorar formas de mantener o incluso reducir aún más esos gastos para aumentar tus ahorros.

**9. ¿Cómo visualizas tus finanzas personales en un año?** La visualización a largo plazo implica establecer metas financieras mensuales, identificar patrones de gastos y ahorros, y adaptar estrategias según sea necesario. Puedes anticipar cambios estacionales o eventos planificados para ajustar tu presupuesto anual y asegurarte de cumplir tus objetivos financieros.

**10. ¿Cuál fue tu mayor aprendizaje y cuál fue tu mayor reto en este Proyecto de Ciencia de Datos?** Mi mayor aprendizaje fue comprender los patrones y comportamientos financieros personales a través de un análisis detallado. El reto fue manejar eficazmente los datos y aplicar técnicas de ciencia de datos para obtener información significativa que respalde decisiones financieras sólidas. Este proyecto proporcionó habilidades valiosas para la gestión efectiva de las finanzas personales.

## Conclusiones

EXCEL

Estadística descriptiva (Actividad 2)

Costo	Presupuesto	Tiempo invertido	Tipo	Momento	No. de personas						
Mean	1101	Mean	1216	Mean	38.3333333	Mean	2.9	Mean	1.8	Mean	1.86666667
Standard Err	567.156519	Standard Err	569.183444	Standard Err	6.57552643	Standard Err	0.32642262	Standard Err	0.13896167	Standard Err	0.33812646
Median	400	Median	500	Median	30	Median	4	Median	2	Median	1
Mode	400	Mode	500	Mode	5	Mode	4	Mode	1	Mode	1
Standard De	3106.44419	Standard De	3117.54612	Standard De	36.0156416	Standard De	1.7878903	Standard De	0.7611244	Standard De	1.85199489
Sample Vari	964995.52	Sample Vari	9719093.79	Sample Vari	1297.12644	Sample Vari	3.19655172	Sample Vari	0.57931034	Sample Vari	3.42988506
Kurtosis	25.8350194	Kurtosis	24.6348784	Kurtosis	2.34730491	Kurtosis	-1.3986954	Kurtosis	-1.141008	Kurtosis	4.23674059
Skewness	4.98554285	Skewness	4.84037298	Skewness	1.51919263	Skewness	0.12179333	Skewness	0.36197789	Skewness	2.3024227
Range	16970	Range	16970	Range	145	Range	5	Range	2	Range	6
Minimum	30	Minimum	30	Minimum	5	Minimum	1	Minimum	1	Minimum	1
Maximum	17000	Maximum	17000	Maximum	150	Maximum	6	Maximum	3	Maximum	7
Sum	33030	Sum	36480	Sum	1150	Sum	87	Sum	54	Sum	56
Count	30	Count	30	Count	30	Count	30	Count	30	Count	30

Excel siendo más accesible para usuarios no técnicos y Python siendo más robusto para análisis avanzados.

PYTHON

Estadística descriptiva de los datos (Fase 2)

df.describe()

	Número	Costo	Presupuesto	Tiempo invertido	Tipo	Momento	No. de personas
count	120.000000	120.000000	120.000000	120.000000	120.000000	120.000000	120.000000
mean	60.500000	470.916667	570.083333	46.416667	2.908333	2.083333	1.325000
std	34.785054	1583.328584	1604.235725	44.893634	1.810315	0.751283	1.022129
min	1.000000	30.000000	0.000000	0.000000	1.000000	1.000000	1.000000
25%	30.750000	150.000000	187.500000	15.000000	1.000000	2.000000	1.000000
50%	60.500000	200.000000	250.000000	30.000000	4.000000	2.000000	1.000000
75%	90.250000	500.000000	600.000000	60.000000	4.000000	3.000000	1.000000
max	120.000000	17000.000000	17000.000000	240.000000	6.000000	3.000000	7.000000

Python tiene un mayor nivel de detalle con cuartiles adicionales y la capacidad de aplicar análisis a múltiples columnas simultáneamente. Python nos da resultados mas exactos.

Coeficientes de regresión (Actividad 3)

	Coefficient	Standard Err	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-119.874	102.94991	-1.1643935	0.251718	-328.471	88.72212	-328.471	88.72212
Presupue:	0.992466	0.0106419	93.2598856	1.64E-45	0.970904	1.014029	0.970904	1.014029
Tiempo in	0.018467	0.7206197	0.0256265	0.979693	-1.44165	1.478581	-1.44165	1.478581
Tipo	2.07464	15.077167	0.13760147	0.891301	-28.4746	32.62388	-28.4746	32.62388
Momento	-4.89246	38.902886	-0.1257608	0.900602	-83.7172	73.93228	-83.7172	73.93228
No. de pe	11.37207	18.204401	0.62468785	0.536009	-25.5136	48.25769	-25.5136	48.25769

El modelo se ajusta mediante una regresión lineal múltiple para predecir el costo de las actividades. Puede indicar que ciertas variables no contribuyen significativamente al modelo.

Coeficientes de regresión (Fase 4))

Coeficientes del modelo:

	Coeficientes
Presupuesto	0.975378
Tiempo invertido	-0.695135
Tipo	-23.392427
Momento	53.121675
No. de personas	23.317755

Lo que me gusta es que ofrece una visión detallada de la contribución de cada variable al modelo. Permite una mayor personalización y ajuste del modelo. Considero que es el mejor modelo y mas preciso

Valores pronosticados y sus residuales (Actividad 3) 30 observaciones de los residuales.

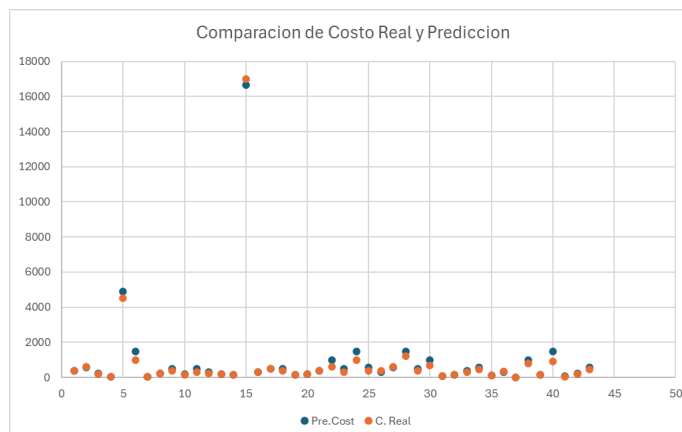
Valores actuales, de predicción y residuales (Fase 4) 30 observaciones de la prueba del modelo.

## Evidencia 2. Proyecto de Ciencia de Datos

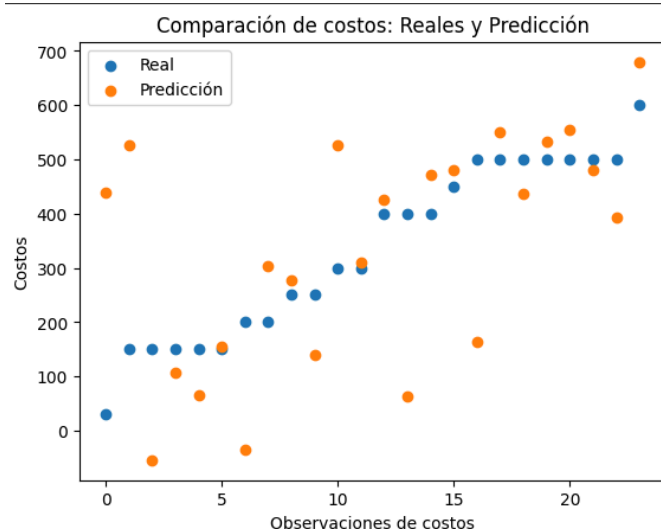
RESIDUAL OUTPUT		
Observation\Actual\Co-Residuals		
1	391.59	8.412
2	587.38	12.618
3	244.74	-44.74
4	48.949	1.0515
5	4894.9	-394.9
6	1468.5	-468.5
7	29.369	0.6309
8	244.74	5.2575
9	489.49	-89.49
10	195.79	-45.79
11	489.49	-189.5
12	293.69	-43.69
13	195.79	4.206
14	146.85	3.1545
15	16642	357.51
16	293.69	6.309
17	489.49	10.515
18	489.49	-89.49
19	146.85	3.1545
20	195.79	4.206
21	391.59	8.412
22	978.97	-379
23	489.49	-189.5
24	1468.5	-468.5
25	587.38	-187.4
26	293.69	106.31
27	587.38	12.618
28	1468.5	-268.5
29	489.49	-89.49
30	978.97	-279

En Excel, los residuales son mayores en magnitud, con valores superiores a 357, mientras que en Python, el rango es más bajo, con valores cercanos a 335.89.

 |    | Real  | Predicción | Residual    | |----|-------|------------|-------------| | 5  | 30.0  | 439.346577 | -409.346577 | | 2  | 150.0 | 107.514142 | 42.485858   | | 21 | 150.0 | 65.183403  | 84.816597   | | 0  | 150.0 | -54.741229 | 204.741229  | | 10 | 150.0 | 525.445404 | -375.445404 | | 13 | 150.0 | 153.684468 | -3.684468   | | 17 | 200.0 | -34.356574 | 234.356574  | | 12 | 200.0 | 303.140932 | -103.140932 | | 14 | 250.0 | 277.968297 | -27.968297  | | 6  | 250.0 | 138.904678 | 111.095322  | | 7  | 300.0 | 525.201064 | -225.201064 | | 4  | 300.0 | 310.850453 | -10.850453  | | 23 | 400.0 | 472.323729 | -72.323729  | | 3  | 400.0 | 425.443878 | -25.443878  | | 16 | 400.0 | 63.097998  | 336.902002  | | 1  | 450.0 | 479.995285 | -29.995285  | | 20 | 500.0 | 533.731489 | -33.731489  | | 22 | 500.0 | 393.499018 | 106.500982  | | 8  | 500.0 | 164.111492 | 335.888508  | | 11 | 500.0 | 436.857457 | 63.142543   | | 18 | 500.0 | 480.609814 | 19.390186   | | 9  | 500.0 | 550.787095 | -50.787095  | | 15 | 500.0 | 555.249325 | -55.249325  | | 19 | 600.0 | 679.161127 | -79.161127  |   En contraste, los residuales en Python muestran un rango más bajo, con valores cercanos a 335.89 como máximo. Esto indica que el modelo en Python tiende a tener residuales con magnitudes más moderadas en comparación con el modelo en Excel. || Coeficiente de determinación r2 (Actividad 3)  R Square  0.99506  En Excel, el alto R<sup>2</sup> indica un buen ajuste del modelo, mientras que en Python, el R<sup>2</sup> negativo indica que el modelo no es adecuado para explicar las variaciones observadas. | Coeficiente de determinación r2 (Fase 4)  Coeficiente de determinación R2: -0.24267231685160073  La discrepancia en los valores de R<sup>2</sup> sugiere diferencias sustanciales en la capacidad de los modelos para explicar la variabilidad en los datos entre Excel y Python. |
| Gráfica de puntos (Actividad 3) | Gráfica de puntos (Fase 4) |



En algunos casos, las predicciones de Excel y Python difieren sustancialmente, mientras que en otros casos, pueden mostrar cierta similitud.



La divergencia en las predicciones puede indicar diferencias en la formulación del modelo o en la interpretación de las variables predictoras.

Para concluir siento que durante esta unidad de formación en ciencia de datos, he adquirido conocimientos valiosos y habilidades prácticas que han ampliado significativamente mi comprensión en este campo de ciencia de datos. A continuación, describo cinco conceptos, herramientas y tecnologías que aprendí y su relevancia:

- **Modelos de Regresión:** Concepto: Antes de esta unidad, tenía una comprensión básica de la regresión, pero ahora entiendo cómo se aplica para prever y entender relaciones entre variables. Utilidad: En mi carrera, puedo emplear modelos de regresión para realizar análisis predictivos y entender la influencia de diversas variables en resultados específicos, lo cual es fundamental para la toma de decisiones informadas.
- **Programación en Python:** Concepto: Continúo aprendiendo a utilizar Python como un lenguaje de programación efectivo para análisis de datos y ciencia de datos. Lo había usado antes en mi carrera en ITC pero no con este enfoque. Utilidad: Esta habilidad es altamente aplicable en mi vida laboral, ya que Python es ampliamente utilizado en diversas industrias. Puedo



automatizar tareas, analizar datos de manera eficiente y construir modelos de aprendizaje automático.

- **Estadísticas Descriptivas y Análisis Exploratorio de Datos (EDA):** Concepto: Pude obtener una mejor comprensión de los datos antes de aplicar modelos es crucial. EDA proporciona herramientas para este propósito. Utilidad: En mi vida profesional, puedo aplicar EDA para entender la distribución de datos, identificar valores atípicos y tomar decisiones fundamentadas en la exploración completa de los datos.
- **Librerías para Análisis de Datos en Python:** Concepto: Me sumergí en librerías como Pandas, NumPy y Matplotlib, que son fundamentales para el análisis de datos en Python. Utilidad: Estas librerías son esenciales para manipular y analizar datos de manera eficiente. Pandas ofrece estructuras de datos poderosas, NumPy facilita cálculos numéricos, y Matplotlib posibilita la creación de visualizaciones claras y efectivas.
- **Análisis de Regresión en Excel:** Concepto: Aunque tenía conocimientos previos de Excel, ahora sé cómo realizar un análisis de regresión en esta plataforma. Utilidad: Excel es una herramienta omnipresente en entornos profesionales. La capacidad para realizar análisis de regresión directamente en Excel es valiosa para proyectos donde la simplicidad y accesibilidad son prioritarias.

### **Aplicación en la Vida Laboral y Personal:**

En mi carrera, siento que estas habilidades serán esenciales ya que creo que me enfocare en la parte de Ciencia de Datos. Dicho esto, Puedo utilizar modelos de regresión para pronosticar tendencias y tomar decisiones fundamentadas. La programación en Python me brinda la capacidad de automatizar tareas repetitivas y realizar análisis de datos avanzados. El uso de estadísticas descriptivas y EDA garantizará que mis decisiones se basen en una comprensión profunda de los

datos. Git y GitHub facilitarán la colaboración en proyectos, y el análisis de regresión en Excel será beneficioso en entornos donde Excel es la herramienta principal.

### **Fortalezas y Áreas de Mejora:**

Lo que hice bien fue dedicar tiempo a prácticas regulares y aplicar los conocimientos adquiridos en proyectos prácticos. Sin embargo, podría haber mejorado al buscar más aplicaciones del mundo real para contextualizar los conceptos aprendidos. La integración de proyectos más complejos podría haber brindado una experiencia más sólida.

### **Metas de Mejora para Futuras Unidades de Formación:**

Para futuras unidades, planeo diversificar mis proyectos y explorar casos de estudio más desafiantes. También buscaré oportunidades para colaborar con compañeros de clase, ya que la colaboración puede proporcionar perspectivas valiosas. Además, seguiré buscando aplicaciones prácticas en la vida real para consolidar mi comprensión de los conceptos.

### **Aspecto Favorito de la Unidad de Formación:**

Lo que más me gustó de esta unidad fue la aplicación práctica de los conceptos. Los proyectos y ejercicios prácticos brindaron una experiencia hands-on que solidificó los conocimientos teóricos. Este enfoque práctico hizo que la unidad fuera más estimulante y aplicable a situaciones del mundo real.

## Referencias

Condusef. (s.f.). Encuesta Nacional de Inclusión Financiera 2018. Gobierno de México. Recuperado de <https://www.gob.mx/condusef/es/articulos/encuesta-nacional-de-inclusion-financiera-2018>

Bain & Company. (s.f.). The Global Pandemic Confirms the Value of a Segmented Bank. Bain & Company. Recuperado de <https://www.bain.com/es-ar/insights/the-global-pandemic-confirms-the-value-of-a-segmented-bank/>

Endeavor México. (s.f.). Endeavor México. Recuperado de <https://endeavor.org.mx/>

Google Colab:  
<https://colab.research.google.com/drive/1r5-vE3Fr1A5HIWfLy4voiqMTgSFX11AI?usp=sharing>