

Instituto Tecnológico y de Estudios Superiores de Monterrey  
Escuela de ingeniería y ciencias



Matemáticas y ciencia de datos para la toma de decisiones

Nombre del profesor  
**Ing. Germán Domínguez Solís**

Evidencia 2. Proyecto de Ciencia de Datos

Integrante:

Rene Miguel Macias Olivar | A00836714

Fecha de creación: 19 de noviembre de 2022

Fecha de entrega: 20 de noviembre de 2022.

Google Colab:

[https://colab.research.google.com/drive/1WmaBQQOJg\\_gJkxsfxT2Hj9atOdQ-d9lR?usp=sharing](https://colab.research.google.com/drive/1WmaBQQOJg_gJkxsfxT2Hj9atOdQ-d9lR?usp=sharing)

## Introducción

La problemática para resolver con este proyecto es ¿Cómo alcanzar mis sueños con finanzas personales saludables usando Ciencia de Datos? Ya que según estudios realizados por la ENIF el 0,8% de las personas en México ahorran para su retiro y en general solo el 43% de los mexicanos ahorran en general en cuentas de bancos o de otra manera más informal y este proyecto me ayudara a saber la cantidad de dinero que llevo a gastar durante mi día a día y saber si estoy gastando de más o si podría llegar ahorrar ese dinero que gasto de más.

Para este proyecto utilizamos la metodología CRISP DM de ciencia de datos el cual se compone de 6 fases por las cuales tendremos que ir analizando nuestro problema y buscando una solución, además de tener que plantearnos hipótesis y realizando ajustes a nuestros objetivos para poder llegar a la solución de nuestra situación problema.

Primero que todo construimos nuestra base de datos a partir los gastos diarios que realizamos a lo largo de todo el semestre, después nos planteamos una hipótesis la cual en este caso es “¿puedo predecir el costo de mis actividades en función del presupuesto disponible para la actividad, tipo de actividad, momento de realización y número de personas, y estimar cómo este costo me impactará con el paso del tiempo?” la cual nos ayudara como guía a lo largo del proyecto y después aplicamos la metodología CRISP DM, gracias al cual nuestro proyecto tendrá que pasar por 6 diferentes etapas las cuales consistirán en el análisis, orden, limpieza ,separación de mis base de datos y la creación de un modelo que nos permita modelar los datos el cual me permitirá avanzar en el análisis de datos el cual no llevara a saber si la hipótesis planteada anteriormente es correcta o no.

## Fase 1. Entendimiento del negocio

### Entendimiento del negocio

Todo negocio en la actualidad lleva consigo una planificación y trabajo detrás de ella, lo cual les ha permitido tener una posición dentro del mercado, ayudándolos a tener una base sólida pero para entender como fue desarrollado este negocio tenemos que verlo desde sus raíces, es decir desde que se plantearon y definieron los objetivos y metas del negocios, desarrollando un plan de trabajo, sucesivo a esto ya con la idea definida comenzaría una etapa de recolección y entendimiento de datos acerca del tema principal del negocio, buscando que estos sean relevantes e importantes para la realización del objetivo ,sucesivamente realizando una limpieza de datos, separándolos de estos los datos relevante e irrelevantes, construyéndolos e integrándolos a nuestro plan de trabajo y con estos datos ya filtrados comenzar a diseñar pruebas y construir el modelo del negocio planteándonos preguntas que sean relevantes para los diferentes sectores que quisiéramos que abarque nuestro negocio para que después poderlo poner a prueba analizando los resultados obtenidos finalizando con la implementación en su totalidad del negocio si el análisis de datos fue satisfactoria. Planeando un monitoreo y mantenimiento de este periódicamente, revisando si cada vez que el negocio cumpla su objetivo satisfactoriamente se llana logrado los objetivos anteriormente planteados, llevando un reporte de estos.

#### 1. ¿Quién es el cliente?

El cliente es aquel al que se le quiere comunicar y dar a entender el objetivo del negocio, convenciéndolo que participe en su ejecución, este objetivo puede ser muy variado como desde la venta de un producto, hasta la búsqueda de una inversión en algún activo. En este caso nosotros, en mi caso sería un universitario de 18 años foráneo que estudia ingeniería en tecnologías computacionales, que le gusta jugar video juegos y salir con sus amigos.

#### 2. ¿Qué problemas estás tratando de resolver?

Buscamos resolver es ver si es posible predecir los costos de mis actividades, en función de una base de datos, la cual contenga datos acerca de mis gastos, actividades, horarios, presupuesto, número de personas involucradas y fecha en la cual se llevó a cabo dicho gasto

3. ¿Qué solución o soluciones la Ciencia de Datos tratará de proveer?

Generar un análisis de regresión mediante el análisis de datos de una fuente de estos.

4. ¿Qué necesitas aprender para poder desarrollar la solución o soluciones?

Estadística básica, análisis de regresión, utilización de Excel, tratamiento de base de datos y análisis de datos, ya que con todos estos conceptos y conocimientos podre saber cómo se comportan las variables de la base de datos y seguir sus trayectorias para poder llegar a conocer si es predecible el costo que llevara consigo una actividad.

5. ¿Qué deberás hacer para desarrollar tu solución?

Primeramente, se tendría que realizar una recolección de datos, la cual ya está en proceso desde la semana 1 del semestre, debemos analizar esta base de datos para conocer cuales datos son relevantes y cuales no, descartando los que no son relevantes y aplicar un tratamiento. Sucesivamente aplicar la regresión de datos mediante las herramientas que nos proporciona Excel y lo aprendido en esta materia.

## Fase 2. Entendimiento de los datos

Entendimiento de los datos: Primero que nada, tendríamos que definir en que consiste el entendimiento de los datos y esta es una etapa que nos permite entender de una forma más sencilla los datos para determinar si posteriormente será necesario algún ajuste dentro de estos datos, esta etapa es fundamental para las próximas etapas ya que aquí se filtran los datos para que más adelante no se generen errores. Esta etapa se divide en 4 secciones, las cuales son: Recolección, aquí se recolectan los datos de diferentes fuentes, las cuales son los datos existentes, datos adquiridos y datos adicionales, como segunda sección tenemos la descripción de los datos, estos datos se pueden describir de muchas maneras pero hay que ponerle especial énfasis en la cantidad y la calidad de estos, como tercera sección tenemos el análisis exploratorio, la cual consiste en explorar los datos, en la cual se pueden formular hipótesis que ayuden a dar forma a los datos y de última sección tenemos la auditoría, esta sección se encarga de arreglar los errores, muchas veces los datos traen errores consigo pero una forma de evitar que estos pasen a la siguiente fase, se realiza una auditoría.

1. ¿Cuáles son tus datos existentes (registrados), datos adquiridos (datos externos) y datos adicionales (datos generados)?

Mis datos existentes son los costos que he generado al gastar en alguna actividad, el presupuesto, el tiempo en que realizo la actividad, el tipo de actividad que realizo y el número de personas que participan en ella, en cambio los datos adquiridos son los que me brinda Excel, como la moda, la mediana, el máximo y mínimo de una actividad etc. Todos estos datos generados por el análisis de datos de Excel y los datos adicionales no existen dentro de mi tabla de gastos

2. ¿Qué tipos de datos se analizarán?

Los que están digitados dentro del Excel, como los gastos, la fecha, el tiempo, el número de personas involucradas, el tipo de actividad y el presupuesto.

3. ¿Qué atributos (columnas) de la base de datos parecen más prometedores?

Yo creería que las más útiles y prometedoras sería las columnas de gastos y de presupuesto, ya que en estas 2 se base casi en totalidad el análisis de datos que he realizado en tareas anteriores.

4. ¿Qué atributos parecen irrelevantes y pueden ser excluidos?

Los atributos irrelevantes dentro de mi tabla de gastos pueden ser el número de personas involucradas y el tiempo en que realizo la actividad ya que estas 2 variables no influyen dentro de mi análisis de datos que he realizado anteriormente.

5. ¿Hay datos suficientes (filas) para sacar conclusiones generalizables o hacer predicciones precisas?

Si hay datos suficientes, ya que si solo utilizamos las columnas de gastos y presupuesto podemos determinar una media de mis gastos diarios y cuánto dinero estaba dispuesto a gastar en dicha actividad.

6. ¿Hay demasiados atributos para realizar un modelo que sea fácil de interpretar?

No, hay suficientes atributos para realizar un modelo fácil de interpretar ya que te muestra variables que pueden ayudar al lector a comprender con mayor facilidad los datos digitalizados dentro de la tabla

7. ¿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 fueron obtenidos 100% de mi persona ya que son un registro diario de mis actividades en la cuales realizo algún gasto monetario, pero existen datos que solo se pueden obtener mediante fórmulas dadas por Excel.

8. ¿Hay algún plan para manejar los valores faltantes en cada una de las fuentes de datos?

Si estas representar algún problema para el análisis de datos se les asigna el valor de 0 para que no sean relevantes o simplemente se eliminan de la base de datos.

9. ¿Cuántos datos están accesibles o disponibles y cómo está la calidad de estos?

Todos los datos están disponibles para ser analizados dentro de mi tabla de datos y la calidad de estos es alta ya que son registrados diariamente al finalizar el día.

10. ¿Cuál es la relación de los datos y la hipótesis del proyecto?

Se relacionan de manera directa ya que sin estos datos no podemos obtener la información necesaria para corroborar si la hipótesis está en los correcto, ya que no tendríamos la suficiente información para realizar el análisis de datos necesario.

### Fase 3. Preparación de los datos

La preparación de los datos es la fase que le sigue a las fases anteriores comprensión de negocios y comprensión de datos, en la cual se podría decir que es la fase en la cual más esfuerzo, tiempo y trabajo se tendría que poner, dependiendo de que tipo de proyecto se trabaje la preparación de datos se puede dividir en diferentes tareas como estas pueden ser la selección de datos, limpieza de datos, integración de datos formato de los datos y división de datos en conjuntos para entrenamiento y prueba .

Primero se tendría que iniciar seleccionando los datos pertinentes obtenidos en la fase anterior mediante una de las dos formas: la selección de elementos a cuál implica tomar decisiones sobre lo que se desea tener y la selección de atributos la cual implica tomar decisiones sobre el uso de características.

De segundo tendríamos la limpieza de datos la cual implica un análisis detallado de los problemas en los datos que se han elegido incluir para el análisis, unos de los problemas más frecuentes en esta sección son los datos perdidos, errores en los datos, inconsistencias en los datos y metadatos faltantes.

Después tenemos la generación de datos, este paso no es total mente necesario, pero hay ocasiones en la cual nos sería útil, necesitando crear una nueva columna de datos a partir de derivar atributos de otros datos o generar estos registros.

Sucesivamente a esto realizaríamos la integración de datos, este paso consiste en funcionar o agregar datos a una base de datos ya existente los cuales se complementen y por último tenemos al formato de datos el cual consiste es darles forma y orden a los datos para que al momento de construir un modelo no se nos dificulte la organización.

1. ¿Qué datos hay que seleccionar? Por qué.



Solo tenemos que seleccionar datos numéricos ya que nuestro análisis de regresión es numérico, además así podremos predecir o determinar un análisis de predicción 100% numérico y evitar generar errores dentro del código que estén relacionados con strings.

2. ¿Hay que eliminar o reemplazar valores en blanco? Sí / No / Por qué.

Si, No podemos permitir que existan valores en blanco o vacíos dentro de nuestra base de datos, tienen que ser solo números, no se pueden permitir comas ni puntos ni otro tipo de carácter que no sea numérico, esto podría generar un error o un cálculo incorrecto dentro del proceso de análisis.

3. ¿Es posible agregar más datos? Sí / No / Por qué.

Si, en realidad es altamente recomendable agregar más datos, para poder obtener resultados más precisos al momento de realizar el análisis ya que tendremos más datos de los cuales podamos obtener resultados, podemos agregar tantas filas como columnas como queramos.

4. ¿Hay que integrar o fusionar datos de varias fuentes? Sí / No / Por qué.

En este caso en concreto no es necesario ya que toda la información proviene desde una sola fuente la cual sería mi base de datos sobre mis gastos diarios, pero si fuera en un caso distinto en la cual se utilizan varias bases de datos para una sola fila o columna si es necesario integrarlas o fusionarlas.

5. ¿Es necesario ordenar los datos para el análisis? Sí / No / Por qué.

En este caso ya tenemos las columnas definidas y ordenadas por su tipo, pero en el caso de las filas es necesario no ordenarlas ya que no queremos no crear sesgo o tendencia al crear el modelo.

6. ¿Tengo que hacer conjuntos de datos para entrenamiento y prueba? Sí / No / Por qué.

Si es necesarios, ya que utilizaremos un 80% de los datos que hemos recopilado y obtenido para entrenar a nuestro modelo, para que este sea altamente confiable y preciso y el otro 20% de los datos obtenidos se utilizaran para poner a prueba nuestro modelo y determinar si ya está listo y funcionando correctamente

7. ¿Qué ajustes se tuvieron que hacer a los datos (agregar, integrar, modificar registros (filas), cambiar atributos (columnas))?

En mi caso solo tuve que eliminar uno datos en unas columnas los cuales se encontraban en blanco, ya que se me olvido llenar esos espacios, pero de ahí en adelante no se me presento ningún problema y pude realiza esta fase sin ninguna complicación, no tuve que integrar ninguna base de datos externa ya que solo trabajé con una y solo realiza esa modificación anteriormente mencionada.

## Fase 4. Modelación de los datos

- Revisa detenidamente las páginas **Aprende sobre: Regresión lineal** y **la Guía para: Implementación y visualización de una regresión lineal múltiple en Python**.
- Escribe tu código para realizar la implementación y visualización de la regresión para modelar los datos que registraste hasta esta semana (como se indica en la guía).
- Ejecuta tu Código.
- Realiza capturas de pantalla de tu programa mostrando cada uno de los resultados (como se indica en la guía).
- Pega las capturas en un documento Word.
- Explica con tus palabras como funciona tu programa (el procedimiento, las librerías) indicando lo que realizaste en la programación.
- Extension: 100 palabras o más.

The screenshot shows a Jupyter Notebook titled 'Fase 2 entendimiento de los datos'. The code cells contain the following:

```
[1] import pandas as pd
[4] df = pd.read_excel('datos.xlsx')
[5] df.head()
```

The output of the first cell shows a DataFrame with 5 rows and 9 columns:

	Número	Fecha (dd/mm/aa)	Nombre actividad	Costo	Presupuesto	Tiempo invertido	Tipo	Momento	No. de personas
0	1	2022-07-08 00:00:00	Supermercado	600.0	700.0	120.0	1.0	1.0	2.0
1	2	2022-07-08 00:00:00	Tabla Periódica	80.0	100.0	15.0	5.0	1.0	1.0
2	3	2022-07-08 00:00:00	Almuerzo	100.0	120.0	60.0	1.0	2.0	1.0
3	4	2022-08-08 00:00:00	Líber	120.0	200.0	50.0	6.0	2.0	2.0
4	5	2022-08-08 00:00:00	almuerzo	80.0	100.0	60.0	1.0	2.0	2.0

```
[6] df = df.iloc[:,3:9]
[7] df.head()
```

The output of the second cell shows the DataFrame with columns selected from index 3 to 8:

	Costo	Presupuesto	Tiempo invertido	Tipo	Momento	No. de personas
0	600.0	700.0	120.0	1.0	1.0	2.0
1	80.0	100.0	15.0	5.0	1.0	1.0
2	100.0	120.0	60.0	1.0	2.0	1.0
3	120.0	200.0	50.0	6.0	2.0	2.0
4	80.0	100.0	60.0	1.0	2.0	2.0

```
[8] df.info()
```

The output of the third cell shows the DataFrame information:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype
---  -
 0   Costo                5 non-null      float64
 1   Presupuesto          5 non-null      float64
 2   Tiempo invertido     5 non-null      float64
 3   Tipo                 5 non-null      float64
 4   Momento              5 non-null      float64
 5   No. de personas      5 non-null      float64
```

## Evidencia 2. Proyecto de Ciencia de Datos

```
Fase 2 entendimiento de los datos ☆
File Edit View Insert Runtime Tools Help All changes saved

Files
- sample_data
- datos.xlsx

[1] df.isnull().sum()

Costo      18
Presupuesto 18
Tiempo invertido 18
Tipo       18
Momento    18
No. de personas 18
dtype: int64

[11] df=df.dropna()

[12] df.isnull().values.any()

False

[13] df.columns

Index(['costo', 'Presupuesto', 'Tiempo invertido', 'Tipo', 'Momento',
      'No. de personas'],
      dtype='object')

[14] x = df[['Presupuesto', 'Tiempo invertido', 'Tipo', 'Momento', 'No. de personas']] # valores = variables independientes
y = df['costo'] # valores = variable dependiente

[15] from sklearn.model_selection import train_test_split

[16] x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=0)

[17] y_test

array([[100., 30., 80., 10., 100., 100., 10., 80., 100., 200.,
        100., 100., 10., 100., 100., 10., 30., 600., 30., 80., 10.,
        10., 80., 80., 30., 100., 100., 10., 10., 100., 100., 100.,
        30., 80., 20., 100., 600., 30., 100., 100., 10., 30., 100.,
        30., 80., 100., 100., 100., 10., 20., 20., 100., 600., 100.,
        80., 80.]])

Disk ██████████ 85.14 GB available

✓ Done completed at 10:16 AM
```

The screenshot shows a Jupyter Notebook environment with the following content:

**File Explorer (Left):**

- config
- sample\_data
- datos.xlsx

**Code Cells:**

```
[1] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300 entries, 0 to 299
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   Costo                  282 non-null    float64 
1   Presupuesto            282 non-null    float64 
2   Tiempo invertido        282 non-null    float64 
3   Tipo                   282 non-null    float64 
4   Momento                282 non-null    float64 
5   No. de personas        282 non-null    float64 
dtypes: float64(6)
memory usage: 34.2 KB
```

```
[9] df.describe()
```

	Costo	Presupuesto	Tiempo invertido	Tipo	Momento	No. de personas
count	282.000000	282.000000	282.000000	282.000000	282.000000	282.000000
mean	95.517730	104.421906	42.783668	1.416440	2.191409	1.241135
std	107.342443	121.117495	36.550959	1.260718	0.633089	0.525616
min	2.000000	2.000000	5.000000	1.000000	1.000000	1.000000
25%	30.000000	30.000000	10.000000	1.000000	2.000000	1.000000
50%	100.000000	100.000000	45.000000	1.000000	2.000000	1.000000
75%	180.000000	180.000000	60.000000	1.000000	3.000000	1.000000
max	800.000000	1000.000000	350.000000	6.000000	3.000000	5.000000

```
[10] df.isnull().sum()
```

```
Costo      18
Presupuesto 18
Tiempo invertido 18
Tipo       18
Momento    18
```

**Status Bar (Bottom):** 86.16 GB available

The screenshot shows a Jupyter Notebook with the following code and output:

```
[18] from sklearn.linear_model import LinearRegression
      model_regression = LinearRegression()

[19] model_regression.fit(x_train, y_train) # aprendizaje automatico con base en nuestros datos
      (LinearRegression())

[20] x_labels = ['Presupuesto', 'Tiempo invertido', 'Tipo', 'Momento', 'No. de personas']
      x_label = ['Coeficientes']

[21] coeff_df = pd.DataFrame(model_regression.coef_, x_labels, x_label)
      coeff_df
```

Output for [21]:

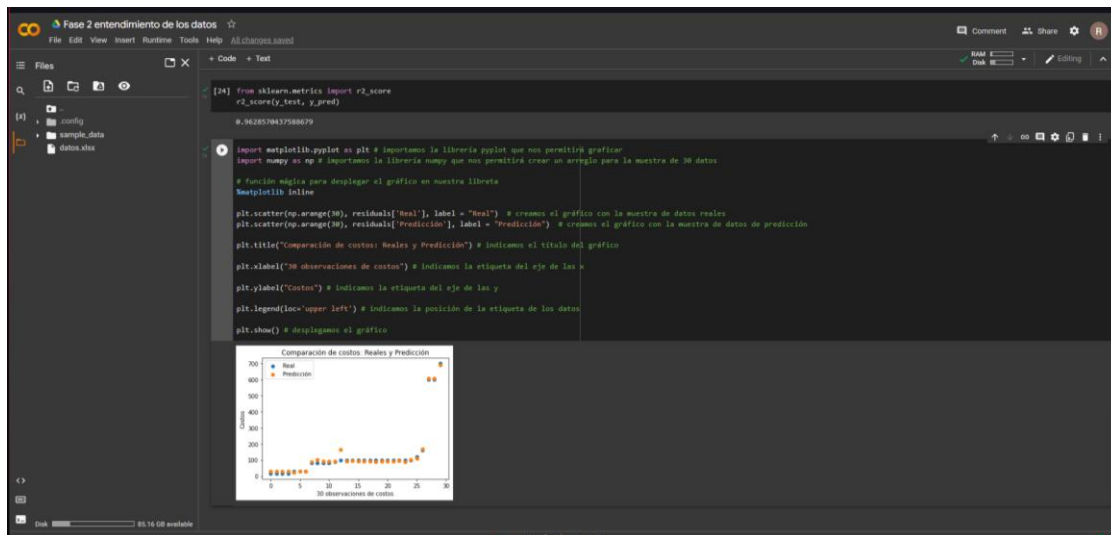
Coeficientes	
Presupuesto	0.826628
Tiempo invertido	0.147940
Tipo	-2.675966
Momento	-3.552759
No. de personas	1.976183

```
[22] y_pred = model_regression.predict(x_test) # realiza la prediccion con el modelo generado

[23] residuals = pd.DataFrame({'Real': y_test, 'Prediccion': y_pred, 'Residual': y_test - y_pred})
      residuals = residuals.drop(columns=['Real'])
      residuals = residuals.sort_values(by='Real')
      residuals
```

Output for [23]:

	Real	Prediccion	Residual
16	16.0	30.386030	-14.386030
49	16.0	30.386030	-14.386030
13	16.0	30.386030	-14.386030
29	16.0	30.386030	-14.386030



Primero iniciamos cargando el archivo que contenga nuestra base de datos para que después importemos nuestras librerías con el comando import y se la asignemos a una variable, creamos otra variable de tipo data frame para cargar el Excel en ella y usamos head() para comprobar que los datos se hallan cargado correctamente.

Seleccionamos los datos que queramos utilizar en este caso de la fila 3 a la 9 y todas las columnas con el comando iloc y usamos head() para corroborar su estado. Con el comando info() vemos que ahora solo tenemos datos tipo int y float 64 y usamos en comando describe() para obtener estadísticas de nuestra tabla.

Utilizamos el comando isnull y sum para encontrar valores nulos para después borrarlos con dropna, volvemos a buscar con isnull y values.any para corroborar.

Ordenamos los datos según el nombre de sus columnas con el comando columns, donde asignaremos sus valores a las variables x la cual contiene los datos de entrada y a la variable y que contiene los de salida y los dividimos en 20% 80% para hacer el entrenamiento y la prueba.

Con scikit learns dividiremos los datos 20% de prueba y 80% de entrenamientos. Después importamos la librería de linearregression y creamos su función. Usamos el comando fit() para ajustar nuestros datos al modelo, creamos dataframe para guardar los valores de x en ella y con el comando coef obtendremos su valores. Utilizamos el modelo predict para predecir nuestros valores. Después con los residuales de la resta de valores reales y predicción creamos una tabla con los 3 valores. Calculamos el coeficiente R2 para determinar que tan confiable es nuestro modelo, entre mas cerca a uno mejor. En mi caso el valor es 0.9 por lo cual es un modelo altamente confiable.

importamos las librerías matplotlib y numpy, usamos la función. scatter( ) para crear el gráfico de dispersión de puntos, se graficará un diagrama de puntos en el cual se observa alta precisión entre los valores reales y de predicción.

Google Colab:

[https://colab.research.google.com/drive/1WmaBQQOJg\\_gJkxsfxT2Hj9atOdQ-d9lR?usp=sharing](https://colab.research.google.com/drive/1WmaBQQOJg_gJkxsfxT2Hj9atOdQ-d9lR?usp=sharing)

Modelación de los datos.

Por estar cerca del final ya comenzando con nuestra modelación de datos, en el cual nos apoyamos de herramientas tecnológicas las cuales nos facilitan la estructuración de nuestro modelo, para que finalmente podamos resolver nuestro problema planteado al inicio de nuestro proyecto. Usualmente cuando ejecutamos nuestro modelo no sale a la primera, hay ocasiones en las cuales tenemos que regresar a fases anteriores a modificar ciertos parámetros o modificar datos en nuestra base de datos, casi nunca al primer intento se responde satisfactoriamente nuestro problema.

Para llegar a elegir o crear un modelo que sea apropiado para la modelación de nuestros datos tendrá que llevar consigo ciertas consideraciones como son: Los tipos de datos disponibles, si estos son de interés o útiles para el modelo, Si nuestro modelo está de acuerdo con los objetivos planteados, Requisitos específicos de la modelación, si el modelo necesita datos específicos o alguna otra variable. Además, es altamente recomendable llevar una descripción del modelo, llevar notas y registrar sobre el modelo,

si se presenta algún error y como se solucionó o si necesita alguna acción en específico para que el usuario pueda correr el programa sin errores. Terminando con la evaluación del modelo, basándose en los criterios que se determine, después de haberlo evaluado es necesario clasificarlos según su orden de criterios.

1. ¿Tuviste problemas para generar el modelo con tus datos? ¿Cómo los resolviste?

No tuve problemas para generar el modelo con tus datos, en mi opinión no tuve ninguna complicación al momento de crear el modelo con mis datos ya que mi base de datos se encuentra completa sin valores nulos o ceros que pueden llegar a generar algún tipo de error dentro del modelo, además de haber seguido los pasos indicados por el profesor los cuales eran explicados de manera clara lo cual lo hacía fácil de entender y de aplicar a mi propia base de datos.

2. ¿Qué resultados arrojó el análisis?

El análisis arroja resultados satisfactorios, las predicciones realizadas como los valores reales son bastante acertados, El coeficiente de  $R^2$  es bastante alto incluso es casi 1 ya que este arroja un el numero 0.96 lo cual sería un 96% de efectividad, lo cual nos lleva a pensar que el modelo realizado es altamente confiable y que los datos utilizados para el entrenamiento y prueba de este modelo son de buena calidad y no cuentan con errores de valores erróneos o nulos que puedan llegar a afectar las predicción y al coeficiente  $R^2$  al momento de ser calculados.

Priemor aqui podemos ver la estadistica descriptiva de mis datos que he recopilado a lo largo de este semestre. En el cual se muestra estadisíticas de mis datos tales como el minimo, el maximo, el promedio, etc. de mis diferentes tipos de datos recopilados.

[9] df.describe()

	Costo	Presupuesto	Tiempo invertido	Tipo	Momento	No. de personas
count	282.000000	282.000000	282.000000	282.000000	282.000000	282.000000
mean	95.517730	104.421986	42.783688	1.418440	2.191489	1.241135
std	107.342443	121.117495	36.550059	1.260718	0.635989	0.525516
min	2.000000	2.000000	5.000000	1.000000	1.000000	1.000000
25%	30.000000	30.000000	10.000000	1.000000	2.000000	1.000000
50%	100.000000	100.000000	45.000000	1.000000	2.000000	1.000000
75%	100.000000	100.000000	60.000000	1.000000	3.000000	1.000000
max	800.000000	1000.000000	360.000000	6.000000	3.000000	5.000000

Después tenemos los coeficientes de regresión los cuales representarían el cambio medio de mis datos con respecto a la variable predictora.

coeff\_df

Coeficientes	
Presupuesto	0.826828
Tiempo invertido	0.147940
Tipo	-2.675906
Momento	-3.552759
No. de personas	1.976103

aquí mostrando mis datos reales contra las predicciones realizadas por el modelo y el residual que queda de la resta de los valores reales y predictivos

	Real	Predicción	Residual
16	16.0	30.386030	-14.386030
49	16.0	30.386030	-14.386030
13	16.0	30.386030	-14.386030
29	16.0	30.386030	-14.386030
1	30.0	21.378056	8.621944
44	30.0	29.646331	0.353669
25	30.0	29.646331	0.353669
51	80.0	87.669994	-7.669994
2	80.0	101.189808	-21.189808
34	80.0	92.108188	-12.108188

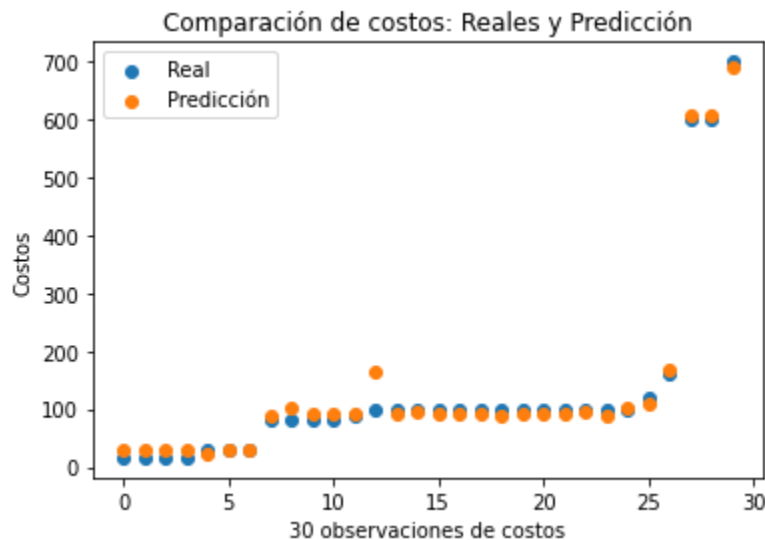


Después teniendo el coeficiente R2 que tiene un valor de 0.96 el cual equivaldría a un 96% de efectividad de mi modelo.

```
[24] from sklearn.metrics import r2_score
      r2_score(y_test, y_pred)

0.9628570437588679
```

Por último, tenemos la gráfica de puntos reales y predictivos que nos muestra de forma gráfica el comportamiento de mis datos y como tienen valores muy cercanos a los predichos por el modelo.



- ¿Los resultados del modelo tienen sentido o hay incoherencias que necesitan una mayor exploración?

Los resultados del modelo tienen sentido, ya que son predicciones muy cercanas a los de su valor real, lo cual lo hace altamente confiable y con un muy bajo margen de error, además de que al presentar un coeficiente R2 tan cercano a 1 no dice que es de gran fiabilidad.

## Evaluando mis finanzas personales

1. ¿Cuántas actividades diarias registraste en total en este semestre?

El numero total de actividades que he registrado durante este semestre es de 300 actividades registradas

```
[27] total_actividades = df["Número"].count()
total_actividades

300
```

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

El presupuesto mínimo fue de 2 pesos y este asignado a la actividad de impresiones, por el contrario, el presupuesto máximo es de 1000 pesos y este asignado a la actividad de supermercado

```
[28] df["Presupuesto"].max()
1000

[29] df.loc[df["Presupuesto"].idxmax(), "Nombre actividad"]
'Supermercado'

[30] df["Presupuesto"].min()
2

[31] df.loc[df["Presupuesto"].idxmin(), "Nombre actividad"]
'impresiones'
```

3. ¿Cuál fue el Tipo de actividad dónde más gastas tu dinero y cuál fue el Tipo de actividad en dónde gastas menos?

El tipo de actividad en la que mas gasto dinero es la numero 1 alimentación/ Salud con un total de 24173 pesos y el tipo de actividad en la que menos gasto dinero es la numero 2 ahorro/ inversión.

```
[32] df.groupby("Tipo").sum().iloc[0:1:2].sort_values(by="Costo")
# Costo más bajo y más alto por tipo de actividad
```

Tipo	Costo
2	200
5	699
6	1310
4	1780
1	24173

4. ¿Por cuántos días registraste tus gastos en este semestre?

Registre mis gastos durante 99 días

```
[33] total_dias = df["Fecha (dd/mm/aa)"].nunique()
total_dias
```

99

5. ¿Cuál fue el total de tus gastos en este semestre?

Mi total de gastos fue de 28162 pesos.

```
[34] total_gastos = df["Costo"].sum()
total_gastos
```

28162

6. ¿Cuál fue el total de tus ahorros en este semestre?

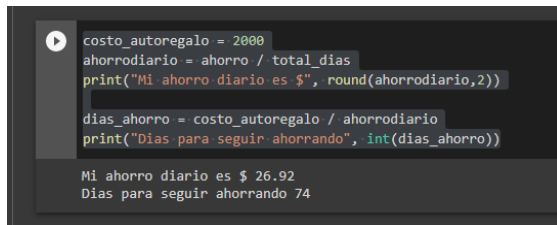
Mis ahorros totales durante este semestre fueron de 2665 pesos

```
[35] ahorro = df["Presupuesto"].sum() - df["Costo"].sum()
ahorro
```

2665

7. ¿Cuánto tiempo (en días) tendrías que seguir ahorrando para comprar tu siguiente autorregalo?

Tendria que seguir ahorrando por 74 dias para poder comprar mi siguiente autorregalo



```
costo_autoregalo = 2000
ahorro_diario = ahorro / total_dias
print("Mi ahorro diario es $", round(ahorro_diario,2))

dias_ahorro = costo_autoregalo / ahorro_diario
print("Días para seguir ahorrando", int(dias_ahorro))
```

Mi ahorro diario es \$ 26.92  
Días para seguir ahorrando 74

8. ¿Qué decisiones informadas puedes tomar para mejorar tus finanzas personales considerando los resultados de tu análisis?

Economizar las compras que realizo durante el día y reducir el número de compras compensándolas con otras que tendrían un impacto económico menor y una de eficiencia mayor.

9. ¿Cómo visualizas tus finanzas personales en un año?

Si sigo a este paso solo podré tener un par de miles ahorrados en cuestión de un año, pero si mejoro mi plan de ahorro y de gastos podría llegar a subir mucho mi tasa de ahorro y aumentar en un par de miles de pesos.

10. ¿Cuál fue tu mayor aprendizaje y cuál fue tu mayor reto en este Proyecto de Ciencia de Datos?

Mi mayor aprendizaje durante este proyecto fue el de aprender a trabajar aplicando ciencia de datos, ya que antes de llevar esta materia nunca había tenido contacto con la ciencia de datos a este nivel, además de aprender cosas nuevas y nuevas metodologías de trabajo, pero al mismo tiempo se me presentó un gran reto el cual fue el entendimiento de cómo funcionaba las librerías de pandas dentro de mi programa, al principio se me hizo bastante confuso pero después de observar el proceso del profesor se me facilitó el entendimiento y la implementación.

- Explica con tus palabras como funciona tu programa (el procedimiento, las librerías) indicando lo que realizaste en la programación.

```
[25] import pandas as pd
df = pd.read_excel('datos.xlsx')
df.head()
```

	Número	Fecha (dd/mm/aa)	Nombre actividad	Costo	Presupuesto	Tiempo invertido	Tipo	Momento	No. de personas
0	1	2022-07-08 00:00:00	Supermercado	600	700	120	1	1	2
1	2	2022-07-08 00:00:00	Tabla Periodica	80	100	15	5	1	1
2	3	2022-07-08 00:00:00	Almuerzo	100	120	60	1	2	1
3	4	2022-08-08 00:00:00	Uber	120	200	50	6	2	2
4	5	2022-08-08 00:00:00	almuerzo	80	100	60	1	2	2

```
[26] df.columns
Index(['Número', 'Fecha (dd/mm/aa)', 'Nombre actividad', 'Costo',
      'Presupuesto', 'Tiempo invertido', 'Tipo', 'Momento',
      'No. de personas'],
      dtype=object)
```

```
[27] total_actividades = df["Número"].count()
total_actividades
300
```

```
[28] df["Presupuesto"].max()
1000
```

```
[29] df.loc[df["Presupuesto"].idxmax(), "Nombre actividad"]
'Supermercado'
```

0s completed at 1:00 PM

```
[30] df["Presupuesto"].min()
2
```

```
[31] df.loc[df["Presupuesto"].idxmin(), "Nombre actividad"]
'Impresiones'
```

```
[32] df.groupby("Tipo").sum().iloc[0:1:2].sort_values(by="Costo")
# Costo más bajo y más alto por tipo de actividad
```

	Costo
Tipo	
2	200
5	699
6	1310
4	1780
1	24173

```
[33] total_dias = df["Fecha (dd/mm/aa)"].nunique()
total_dias
99
```

```
[34] total_gastos = df["Costo"].sum()
total_gastos
28162
```

```
[34] total_gastos = df["Costo"].sum()
total_gastos
28162
```

```
[35] ahorro = df["Presupuesto"].sum() - df["Costo"].sum()
ahorro
2665
```

```
costo_autoregalo = 1499
ahorrodiario = ahorro / total_dias
print("Mi ahorro diario es $", round(ahorrodiario,2))

dias_ahorro = costo_autoregalo / ahorrodiario
print("Dias para seguir ahorrando", int(dias_ahorro))

Mi ahorro diario es $ 26.92
Dias para seguir ahorrando 55
```

Comenzamos este código importando las librerías de panda y subiendo nuestra base de datos de nombre “datos” a Google colab y vemos si se a cargado correctamente con

el comando head (), definimos columnas y después calculamos el total de actividades realizadas con el comando

```
f["Número"].count(),
```

con la función Max determinamos el presupuesto máximo y con el comando

```
df.loc[df["Presupuesto"].idxmax(),"Nombre actividad"]
```

sabremos a que actividad esta asignado ese presupuesto y lo mismo con el presupuesto mínimo con la función min y el comando

```
df.loc[df["Presupuesto"].idxmin(),"Nombre actividad"]
```

después determinamos los gastos totales según su tipo de actividad con el comando

```
f.groupby("Tipo").sum().iloc[0:,1:2].sort_values(by="Costo"),
```

después con el comando

```
total_dias = df["Fecha (dd/mm/aa)"].nunique() total_dias
```

para saber durante cuantos dias registre datos y el comando

```
total_gastos = df["Costo"].sum()
```

para saber el total de gastos de todos mis registros, para determinar el ahorro usamos el comando ahorro

```
= df["Presupuesto"].sum() - df["Costo"].sum()
```

y para determinar el dentro de cuanto tiempo puedo realizar un autorregalo uso las siguientes líneas de código

```
costo_autoregalo = 2000
```

```
ahorrodiario = ahorro / total_dias
```

```
print("Mi ahorro diario es $", round(ahorrodiario,2))
```

```
dias_ahorro = costo_autoregalo / ahorrodiario
```

```
print("Dias para seguir ahorrando", int(dias_ahorro))
```

## Reflexión final

- Compara los procedimientos y resultados de la regresión realizada en Excel en la semana 4 y en Python en la semana 14.

Para esto, realiza una tabla comparativa para explicar las diferencias, incluye imagen y explicación de cada resultado en Excel y Python. ¿Cuál te pareció mejor, por qué?

	Excel	Python
Estadística descriptiva	En el caso de excel en este apartado no me gusto mucho como desplegaba los resultados pero en general los datos y la informacion despelgada era correcta	Me pareció mejor como desplegaba la información Python ya que la mostraba de una manera más simple y fácil de entender.
Coeficientes de regresión	En el caso de este apartado no me gusto el modo de obtención del coeficiente de regresión ya que se tiene que realizar varios ajustes a la base de datos teniendo que eliminar datos y repetir procesos	Por el contrario, el modo de obtención de estos coeficientes en Python es mucho más sencillo y fácil de entender, sin tener la necesidad de limpiar muchos datos. Mucho más recomendable
Valores pronosticados y sus residuales	En esta comparación no encontré pro y contras en ninguno de los casos, en ambas aplicaciones son fáciles de entender y fácil de manejar, y ambos arrojan resultados muy similares.	
Coeficiente de determinación $r^2$	Para la obtención de este coeficiente no podría elegir si una de las dos opciones es mejor que la otra ya que en ambos casos pude obtener un coeficiente muy cercano a uno y en ambos casos es muy fácil el procedimiento para obtenerlo.	

Gráfica de puntos	Sucede lo mismo en este caso, los dos procesos son sencillos de realizar y fáciles de entender y su resultado es satisfactoria así que diría que ambos procesos son igual de eficientes y recomendables.
-------------------	--

Estadística descriptiva:

Excel

Costo	Presupuesto		Tiempo invertido		Tipo		Momento		No. de personas		
Mean	135.0625	Mean	155	Mean	53.4375	Mean	2.604166667	Mean	1.875	Mean	1.5208333
Standard Error	21.1997509	Standard Error	23.06712376	Standard Error	8.114807095	Standard Error	0.317542066	Standard Error	0.0923732	Standard Error	0.089272
Median	100	Median	100	Median	45	Median	1	Median	2	Median	1
Mode	100	Mode	100	Mode	60	Mode	1	Mode	2	Mode	1
Standard Deviation	146.8761827	Standard Deviation	159.8137214	Standard Deviation	56.22103273	Standard Deviation	2.19999597	Standard Deviation	0.6399801	Standard Deviation	0.6184945
Sample Variance	21572.61303	Sample Variance	25540.42553	Sample Variance	3160.804521	Sample Variance	4.83998227	Sample Variance	0.4095745	Sample Variance	0.3825355
Kurtosis	6.570561256	Kurtosis	7.197005535	Kurtosis	19.03059063	Kurtosis	-1.372937626	Kurtosis	-0.4863816	Kurtosis	-0.3473487
Skewness	2.749226246	Skewness	2.81066258	Skewness	3.843222343	Skewness	0.742509886	Skewness	0.1111695	Skewness	0.7625139
Range	590	Range	680	Range	355	Range	5	Range	2	Range	2
Minimum	10	Minimum	20	Minimum	5	Minimum	1	Minimum	1	Minimum	1
Maximum	600	Maximum	700	Maximum	360	Maximum	6	Maximum	3	Maximum	3
Sum	6483	Sum	7440	Sum	2565	Sum	125	Sum	90	Sum	73
Count	48	Count	48	Count	48	Count	48	Count	48	Count	48

Python

	Costo	Presupuesto	Tiempo invertido	Tipo	Momento	No. de personas
count	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000
mean	93.873333	102.756667	42.216667	1.393333	2.186667	1.240000
std	104.683552	117.879101	35.734103	1.226228	0.642806	0.519583
min	2.000000	2.000000	5.000000	1.000000	1.000000	1.000000
25%	30.000000	30.000000	10.000000	1.000000	2.000000	1.000000
50%	95.000000	100.000000	45.000000	1.000000	2.000000	1.000000
75%	100.000000	100.000000	60.000000	1.000000	3.000000	1.000000
max	800.000000	1000.000000	360.000000	6.000000	3.000000	5.000000

Coefficientes de regresión

Excel

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-9.021723173	14.68205994	-0.614472575	0.541015669	-38.33540165	20.2919553	-38.33540165	20.2919553
Presupuesto	0.911105688	0.027166088	33.53834727	3.50388E-43	0.856866844	0.965344533	0.856866844	0.965344533
Tiempo invertido	0.084813665	0.076604226	1.10716692	0.272240694	-0.068131607	0.237758937	-0.068131607	0.237758937
Tipo	2.150946191	2.122865118	1.013227912	0.31465148	-2.08749077	6.389383151	-2.08749077	6.389383151
Momento	10.42650786	5.538514993	1.882545749	0.064170963	-0.631494111	21.48450983	-0.631494111	21.48450983
No. de personas	-16.34013132	8.177206546	-1.998253465	0.049812057	-32.66645095	-0.013811678	-32.66645095	-0.013811678



## Python

Coeficientes	
Presupuesto	0.825525
Tiempo invertido	0.167082
Tipo	-2.478968
Momento	-3.127532
No. de personas	2.058202

## Valores pronosticados y sus residuales

## Excel

RESIDUAL OUTPUT			
Observation	dicted Co	Residuals	
1	625.756	-25.7556	
2	89.3937	-9.39366	
3	107.272	-7.27239	
4	178.787	-58.7873	
5	89.3937	-9.39366	
6	107.272	-7.27239	
7	89.3937	10.6063	
8	446.968	153.032	
9	89.3937	10.6063	
10	89.3937	30.6063	
11	71.5149	-11.5149	
12	89.3937	10.6063	
13	178.787	-78.7873	
14	89.3937	10.6063	
15	268.181	-28.181	
16	89.3937	-9.39366	
17	17.8787	-7.87873	
18	89.3937	10.6063	
19	89.3937	-49.3937	
20	178.787	-58.7873	
21	625.756	-25.7556	
22	89.3937	10.6063	
23	107.272	-7.27239	
24	107.272	12.7276	
25	53.6362	6.36381	
26	89.3937	-9.39366	
27	107.272	-7.27239	
28	53.6362	6.36381	
29	89.3937	-9.39366	
30	107.272	-7.27239	

## Python

	Real	Predicción	Residual
12	5	2.109401	2.890599
27	16	29.371184	-13.371184
18	16	29.371184	-13.371184
21	20	28.535773	-8.535773
50	20	20.280523	-0.280523
47	30	20.280523	9.719477
35	30	28.535773	1.464227
8	30	28.535773	1.464227
59	30	28.535773	1.464227
56	30	30.206595	-0.206595
44	30	28.535773	1.464227
38	80	89.878276	-9.878276
17	80	100.697776	-20.697776
7	80	92.384509	-12.384509
6	80	92.384509	-12.384509
46	80	92.384509	-12.384509
15	80	87.372043	-7.372043
23	90	89.878276	0.121724
16	100	93.005808	6.994192
14	100	101.685901	-1.685901
5	100	93.627107	6.372893
39	100	103.142612	-3.142612
53	100	95.064010	4.935990

Coeficiente de determinación  $r^2$ 

## Excel

2		
3	<i>Regression Statistics</i>	
4	Multiple R	0.986504694
5	R Square	0.973191512
6	Adjusted R Square	0.959107005
7	Standard Error	28.29772469
8	Observations	72

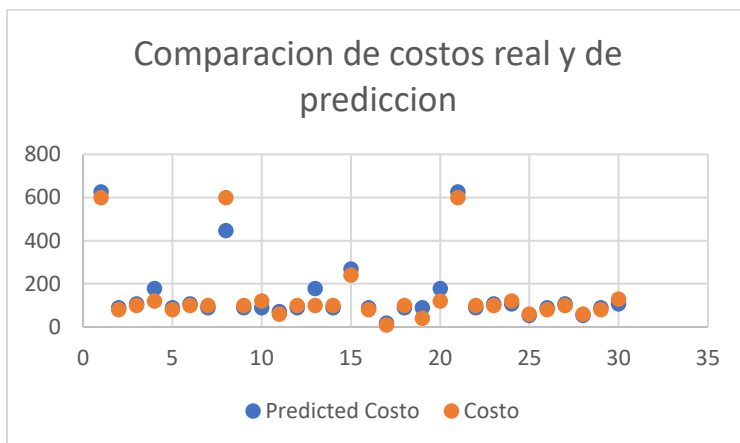
## Python

```
[23] from sklearn.metrics import r2_score
      r2_score(y_test, y_pred)

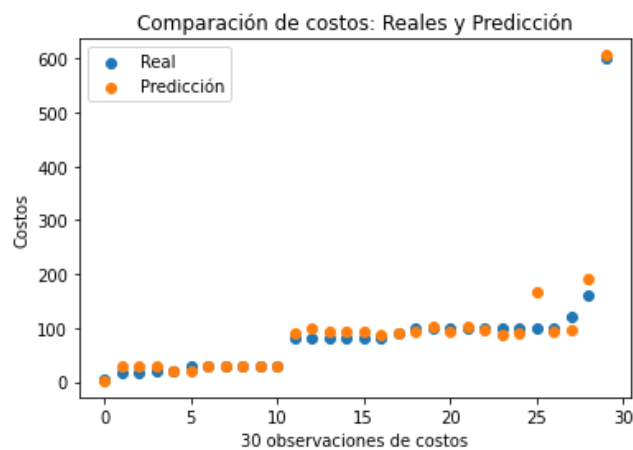
0.9518721533059494
```

## Gráfica de puntos

## Excel



## Python



¿Puedo predecir el costo de mis actividades en función del presupuesto disponible para la actividad, tipo de actividad, momento de realización y número de personas, y estimar cómo este costo me impactará con el paso del tiempo? ¿Qué tan preciso es el modelo?

Si puedo predecir el costo de mis actividades en función del presupuesto disponible para la actividad, tipo de actividad, momento de realización y número de personas, y estimar cómo este costo me impactará con el paso del tiempo, ya que es un modelo de alta precisión representado por el coeficiente  $R^2$  que en ambos casos (Excel y Python) son mayores a 95% de precisión.

## Bibliografía

IBM (2021) Introducción al CRISP-DM, IBM, recuperado el 19/11/2022 de <https://www.ibm.com/docs/es/spss-modeler/saas?topic=guide-introduction-crisp-dm>

Software DELSOL (2020) Coeficiente de determinación, Software DELSOL, recuperado el 19/11/2022 de <https://www.sdelisol.com/glosario/coeficiente-de-determinacion/>

Scikit Learn (2007-2022) Decision Tree Regression, Scikit Learn recuperado el 22/10/2022 de [https://scikit-learn.org/stable/auto\\_examples/tree/plot\\_tree\\_regression.html#sphx-glr-auto-examples-tree-plot-tree-regression-py](https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html#sphx-glr-auto-examples-tree-plot-tree-regression-py)