INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY

EVIDENCIA 2: PROYECTO DE CIENCIA DE DATOS

Nombre del alumno:

Gabriel Esperilla León

Matricula:

A01277955

Profesor:

Héctor Hernández De la Cerda

ASIGNATURA:

Matemáticas y ciencia de datos para la toma de decisiones

19 de noviembre de 2023

INTRODUCCIÓN

En México solamente el 0,8% de las personas ahorran para su vejez, mientras que tan solo 1 de cada 5 jovenes de entre 18 y 29 años tienen el habito del ahorro, de acuerdo con la ENIF, estas cifras solamente demustran que la sociedad mexicana, cuenta con una escasa conciencia financiera.

Es precisamente esa falta de cultura financiera la problemática a resolver en este proyecto, la pregunta seria, ¿Qué podemos hacer para tener mejores habitos finacieros y como podemos ayudarnos de la ciencia de datos para lograrlo?, el objetivo del presente trabajo es darle una solucion a esta problemática.

La metodologia que usada para el desarrollo del proyecto, es la llamada metodologia CRISP-DM (Cross-industry standard process for data mining por sus siglas en inglés), la cual es el estandar para la investigacion, mineria, analitica y ciencia de datos.

Para resolver esta problemática realizamos un programa basado en la estadistica, matematicas y ciencia de datos, que tiene como tarea predecir los gastos de las personas con base en ciertos factores, como su prespuesto, momento del gasto y el tipo de gasto entre otras variables.

Este modelo sera de ayuda para ayudar a las personas a determinar que en que gastan mas, y asi mismo el que influye para hecer esos gastos, por ejemplo, sabiendo el tipo de gaste, el momento del dia en el que se haria el gasto o el presupuesto que se tiene para hacer el gasto, se podria determinar un aproximado del valor del gasto y asi el usuario podria decir si es un buen gasto o no lo es.

FASE 1. ENTENDIMIENTO DEL NEGOCIO

La primera fase, conocida como "Entendimiento del Negocio", consiste en definir primero que nada exactamente cuáles son los objetivos de nuestro proyecto. Esto implica establecer objetivos claros, precisos y concisos, que luego nos ayudarán a determinar qué hay que hacer para alcanzar esas metas. Como resultado, cada objetivo debe ir acompañado de una acción que asegure su cumplimiento.

Sin embargo, para definir nuestros objetivos necesitaremos aplicar la metodología SMART, que implica fijar objetivos específicos con las siguientes características, ser específicos, es decir, evitar ideas vagas o mal definidas, para determinar su efectividad y deben tener indicadores concretos para garantizar su cumplimiento.

Las metas deben de ser *alcanzables,* en otras palabras, objetivos realistas considerando las adversidades y los retos que podamos enfrentar durante el transcurso del proyecto, *relevantes,* ósea que ayuden a cumplir un propósito y por ultimo los objetivos según la metodología SMART deben de tener una temporalidad definida, es decir que se debe de establecer tanto la fecha de su ejecución como la de su cumplimiento

Otro aspecto muy importante para la realización de esta fase es saber el contexto de nuestro punto de partida, es decir saber en qué situación estamos y a partir de ahí encontrar la metodología adecuada para lograr darle una solución a nuestro problema.

Cuestionario

¿Quién es el cliente?

En este caso el cliente seriamos nosotros mismos, dado que los datos recopilados tienen que ver con nuestra vida cotidiana y tienen la finalidad de ser analizados y con base en ellos crear un modelo que infiera el costo de futuros gastos dependiendo de ciertas variables.

¿Qué problemas estás tratando de resolver?

A partir de este modelo basado en la estadística, matemática y ciencias de la computación voy a poder inferir el monto de mis gastos a partir de factores como el momento del día en el que se realiza el gasto, la cantidad de personas involucradas, el presupuesto destinado para dicha actividad y el tipo de gasto que se realiza por ejemplo gastos académicos, de diversión, nutrición entre otro tipo de gastos.

La problemática que resolveré con esto será saber que hacer para mejorar mis finanzas, por ejemplo, al saber en qué tipo de cosas invierto más dinero y más tiempo, identificar que acciones haría con base al modelo de análisis para aminorar los gastos menos necesarios.

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

La solución que me brinda la ciencia de datos es un modelo que determinará mis gastos a partir de los factores ya mencionados, este modelo me permitirá mejorar mis finanzas e identificar los llamados gastos hormiga que parecen ser pequeños pero que a largo plazo representan un gran porcentaje del total de gastos.

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

Primero que nada, es necesario entender el contexto de la situación, una vez sabiendo que lo que queremos es mejorar mis finanzas necesito aprender a generar este modelo a partir de la ciencia de datos, para ello tengo que aprender

computación, matemáticas, estadística, entre otras disciplinas relacionadas con el análisis de datos.

¿Qué deberás hacer para desarrollar tu solución?

Debo plantear los objetivos reales y concisos, posterior mente partiendo de esos objetivos establecer mi plan de acción para lograr su ejecución, una vez teniendo el proyecto claro, realizar lo ya planificado, en este caso el modelo de predicción de gastos.

FASE 2. ENTENDIMIENTO DE LOS DATOS

El objetivo de esta fase es poder comprender la información que una organización posee para posteriormente determinar si es necesario hacer ajustes es decir determinar si es necesario adquirir más información de otras fuentes o modificar la información obtenida y es una etapa de suma importancia, ya que de esta fase se sustentan las fases siguientes.

Las principales tareas de esta fase son recolectar los datos (existentes, adquiridos y adicionales), describirlos con base en su cantidad y calidad, hacer un análisis de los mismos en el cual se formularán hipótesis que ayuden a manejar los datos y por último se realiza una auditoria de datos, lo cual básicamente es verificar que los datos sean verídicos.

Para hacer un análisis exploratorio de los datos, existen varias herramientas entre las que destacan la funciones, que nos permiten apreciar los datos de una manera más clara y concisa lo que hacen las funciones es mostrarnos elementos como correlaciones, tendencias, o los datos anormales dentro del conjunto total.

Una vez teniendo el análisis el problema será como presentarlo, para solucionar esto existen recursos como los gráficos que a menudo hacen más fácil el entendimiento del análisis, ya que muchas veces el público no cuenta con los conocimientos técnicos para interpretar los datos, y es ahí donde las gráficas pueden ser de gran beneficio.

Cuestionario

¿Cuáles son tus datos existentes, datos adquiridos y datos adicionales?

En este caso los datos existentes serían los que nosotros incluimos en nuestro documento de Excel como los costos, presupuestos y todos con los que ya contamos, los datos adquiridos serian aquellos que tenemos con consultar de fuentes externas para después registrarlos en Excel. Y finalmente los datos adicionales en caso de necesitarlos serian aquellos que no tenemos disponibles y queramos agregar.

¿Qué tipos de datos se analizarán?

Dentro de nuestros 150 registros serán analizados en su totalidad, es decir que por ejemplo para las datos del concepto de gasto serán cadenas porque el nombre del gasto esta dado por palabras, el costo y el presupuesto esta dado por valores numéricos que pueden ser enteros o flotantes, es decir con valor decimal, mientras que los valores de número de personas, momento del día y tipo de gasto serán simples datos de tipo entero, ya que todos ellos tienen en su dominio únicamente valores enteros.

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

Para efectuar nuestro análisis, considero que los datos que tendrán una mayor significancia en nuestro estudio serán los valores de nombres del gasto, costos, presupuesto, tipo de gasto, así como el momento en el que se efectúa el gasto.

¿Qué atributos parecen irrelevantes y pueden ser excluidos?

A mi parecer en este caso los valores de la columna de tiempo invertido y número de personas podrían se irrelevantes o poco significativos para nuestro análisis, dado que muchas veces estas variables no tienen correlación con el costo del gasto

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

Considerando que cuento con 150 registros de mis actividades el resultado puede ser significativo, pero como ya sabemos, para tener una predicción más precisa siempre será de utilidad tener la mayor cantidad de datos a tu alcance.

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

El hecho de tener muchos datos hace que interpretarlos sea un poco mas difícil, pero si los manejamos en porcentajes y los presentamos en gráficos será más fácil entenderos e interpretar su correlación.

¿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?

Para obtener estos datos me base en mis gastos diarios, es decir que solamente fue necesaria una fuente que es básicamente el registro de mis gastos diarios durante un periodo de nueve semanas.

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

Considerando que los datos faltantes no son relevantes para el estudio, creo que lo mejor es definirlos como cero ya que los datos faltantes en este caso son nulos, porque todos los datos necesarios están definidos dentro del documento de Excel y para nuestro estudio un hace falta consultar otra fuente o adquirir datos externos.

¿Cuántos datos están accesibles o disponibles y cómo está la calidad de los mismos?

El documento de Excel cuenta con un total de 1200 datos partiendo de 150 filas que representan el total de los registros durante las 9 semanas y 8 columnas que representan los atributos de cada registro excluyendo el número de registro. y de los cuales tenemos acceso al cien por ciento de ellos, y considerando que todos los datos están basados en mis compras diarias podemos inferir que la calidad y significancia de los datos es confiable.

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

La hipótesis del proyecto es que podemos inferir el costo de un gasto a partir de variables como el presupuesto, momento etcétera, teniendo esto en cuenta podemos establecer que los datos tienen una correlación fuerte con la hipótesis de nuestro proyecto.

FASE 3: PREPARACIÓN DE LOS DATOS

En que consiste la fase 3: Preparación de los datos

El objetivo de esta fase es tener los datos listos para un posterior análisis y modelado, y dependiendo de los datos obtenidos, se realizan las siguientes tareas:

Selección de datos: En esta etapa se eligen los datos pertinentes para el análisis. Esto significa evaluar qué elementos o datos son necesarios para el proyecto, asimismo, en esta fase, se pueden eliminar datos que no sean relevantes o útiles para los objetivos del proyecto.

Limpieza de datos: durante esta fase, se buscan errores en los datos proporcionados, es decir, si los datos presentan inconsistencias, o hay datos faltantes, esto se puede hacer tanto de forma manual, como haciendo uso de herramientas tecnológicas.

Generación de nuevos datos: de ser necesario se pueden agregar nuevos datos y hay dos maneras para hacerlo, se pueden agregar más registros, es decir más filas, o se pueden agregar nuevos datos dentro de una fila ya estructurada, teniendo en cuanta que mientras más datos tengamos, será un modelo más preciso.

Finalmente, tenemos las etapas de integración y formato de datos, en la primera, debemos de identificar, Si hay más de una fuente para los datos, de ser así, debemos integrarlos de forma apropiada. Esto puede implicar combinar diferentes conjuntos de datos.

Mientras que, en la fase de formato de datos, lo que hacemos es básicamente verificar, si se requiere un formato u orden particular para los datos, es decir, darle una forma específica a los datos para posteriormente, ejecutar el modelo predeterminado.

CUESTIONARIO

¿Qué datos hay que seleccionar? Por qué.

Los datos seleccionados serán, los valores que hemos ingresado en nuestro archivo Excel. Estos datos son la base, sobre la que se construye nuestro modelo de aprendizaje automático. El cual nos servirá para evaluar el costo de nuestras actividades según los atributos que hemos recopilado y registrado en el archivo de datos de Excel.

Al utilizar estos datos, podremos analizar y comprender mejor las relaciones entre los diversos atributos y el costo asociado con nuestras actividades. El aprendizaje automático nos permitirá analizar patrones y tendencias para así poder terminar el proyecto con el cual podremos predecir el costo de todas nuestras actividades.

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

En este caso, no hay inconsistencias ni datos faltantes en mi documento, por lo tanto, no es necesario limpiar los datos porque no falta información. La cantidad y la integridad de los datos de este documento son óptimas, lo que garantiza una base sólida para el análisis y el procesamiento posterior.

La no existencia de datos faltantes simplifica significativamente la tarea de trabajar con esta información porque no se requieren correcciones o crear datos, para llenar los vacíos en los datos. Esto, a su vez, aumenta la confiabilidad del modelo de aprendizaje automático de nuestro proyecto, el cual se puede llevar a cabo utilizando estos datos porque se puede confiar en su calidad e integridad.

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

Si es posible y es incluso muy recomendable, ya que como sabemos, a mayor cantidad de datos, el modelo tendrá una mayor precisión y funcionabilidad, esto quiere decir que mientras más registros de datos tenga el documento de Excel, por ende, el modelo de aprendizaje tendrá mas de donde aprender, y por ello, será más eficiente.

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

En nuestro caso no es necesario fusionar datos de varias fuentes, ya que como sabemos, todos nuestros datos se encuentran en los registros dentro del documento de Excel, y no tenemos otra fuente de información adicional a esa, la única alternativa, en la que tendríamos que hacer eso, seria si por ejemplo hiciéramos otro documento con mas datos, o una fuente de información distinta.

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

En nuestro caso no es muy necesario modificar el orden de los datos, ya que como sabemos su orden no alteraría la predicción del modelo de aprendizaje automático, y con el orden en el que están introducidos en el documento de Excel, se generaría el mismo modelo que el que se produciría si modificamos el orden.

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

SI, como se indica en la guía de preparación de datos, vamos a asignar el 20 % de los datos a las ‘Xs’ y ‘Ys’ de prueba, mientras que el 80% restante van a funcionar para el entrenamiento del modelo de aprendizaje automático, con el cual funcionara nuestro proyecto de predicción de costos con base en los atributos como tipo de gasto, presupuesto, etcétera.

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

Para crear nuestro modelo de aprendizaje automático, primero que nada dejamos únicamente los datos que son relevantes para nuestro proyecto, es decir los datos de las columnas de: “costo”, “presupuesto”, “Tiempo invertido”, “Tipo”, “Momento” y “Número de personas”, dejando a fuera a los atributos irrelevantes los cuales son: “Numero de registro”, “Fecha” y “Nombre de la actividad”.

FASE 4: MODELACIÓN DE LOS DATOS

Como su nombre lo dice, el objetivo de esta fase, es crear y validar un modelo de predicción de nuestros datos, el cual puede ser elaborado, mediante técnicas y algoritmos, como la regresión lineal, regresión logística, redes neuronales, conjunto de arboles y entre otras técnicas de estadística y aprendizaje automático.

Posteriormente, se evalúa la eficiencia del modelo, haciendo uso de ciertas métricas, como en nuestro caso, el coeficiente de regresión lineal, que para ser considerado preciso, debe tener un valor cercano al 100% es decir un valor cercano al 1.0

Después de eso tenemos que tener en cuanta para que sirve y que va a demostrar nuestro modelo creado, dependiendo de ello, vamos a decidir como representarlo, por ejemplo, si lo que vamos a amostrar es una comparación podemos hacer uso de gráficos de barras o incluso de tablas.

Sin embargo, lo que queremos mostrar son datos relacionados con la distribución acerca de algo, podemos utilizar las graficas de histogramas, histogramas en forma lineal y si lo que queremos es graficar datos que representan una composición, lo más conveniente será usar graficas circulares o gráficos de áreas

En nuestro caso al ser un modelo de relación entre variables, los gráficos que mas se adecuan a nuestra necesidad, son los gráficos de dispersión, en los que podemos ver los puntos de relación que existen entre las variables.

Cuestionario

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

Después de ejecutar el código de Python en Google colab, la principal problemática que encontré en mis resultados, fue que los coeficientes de regresión lineal, no eran muy cercanos al cien por ciento, lo cual quiere decir que nuestras predicciones no serán tan exactas ni precisas, como lo queremos.

Otro problema que identifiqué dentro de mi programa, fue los valores residuales, ya que como ya mencioné, los valores de los coeficientes, me indican que mi modelo no fue muy confiable, y esto se vio reflejado en los valores de las predicciones, que están muy alejados con respecto a los valores reales y por lo tanto los valores absolutos de la columna de residuales dan números muy elevados.

Posteriormente en la gráfica dispersión de valores reales contra valores de la predicción, se visualizan resultados muy diferentes a los esperados, ya que el modelo de predicción creo datos muy alejados a los datos del costo real, y esto lo podemos ver en la gráfica con los puntos de predicción muy separados con respecto a los datos reales.

La conclusión que obtenemos de este modelo, es que las predicciones no son confiables, al menos con esa cantidad de datos, ya que como sabemos, mientras mas datos, se generara un modelo más preciso y funcional, pero por el momento el modelo generado en Python no es confiable.

¿Qué resultados arrojó el análisis?

Los primeros resultados que fueron arrojados por nuestro programa de Python fueron los coeficientes de regresión lineal de las variables independientes, es decir del presupuesto, tiempo invertido en minutos, Tipo de gasto, Momento del día, y el numero de personas involucradas. Los datos fueron mostrados en una tabla.

FASE4. Modelacion de los datos

```
[1] import pandas as pd
```

```
[2] df = pd.read_excel('A01277955_Registro.xlsx')
```

```
[3] df.head()
```

	Número	Fecha (dd/mm/aa)	Nombre actividad	Costo mxn	Presupuesto	Tiempo invertido min	Tipo	Momento	No. de personas
0	1	2023-08-14	Colegiatura	10000	20000	1	5	1	1
1	2	2023-08-14	Renta	4200	10000	1	5	1	1
2	3	2023-08-14	Ahorro mensual	2000	5800	1	2	1	1
3	4	2023-08-14	Compra de Pan bimbo	50	3800	10	1	1	1
4	5	2023-08-14	Compra de Platano	47	3750	10	1	1	1

```
[7] df = df.iloc[:, 3:9]
```

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

	Costo mxn	Presupuesto	Tiempo invertido min	Tipo	Momento	No. de personas
0	10000	20000	1	5	1	1
1	4200	10000	1	5	1	1
2	2000	5800	1	2	1	1
3	50	3800	10	1	1	1
4	47	3750	10	1	1	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 mxn             300 non-null    int64
 1   Presupuesto           300 non-null    int64
 2   Tiempo invertido min  300 non-null    int64
 3   Tipo                  300 non-null    int64
 4   Momento               300 non-null    int64
 5   No. de personas       300 non-null    int64
dtypes: int64(6)
memory usage: 14.2 KB
```

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

```
Costo mxn               0
Presupuesto             0
Tiempo invertido min    0
Tipo                    0
Momento                 0
No. de personas         0
dtype: int64
```

```
[11] df = df.dropna()
```

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

```
False
```

```
[13] df.columns
```

```
Index(['Costo mxn', 'Presupuesto', 'Tiempo invertido min', 'Tipo', 'Momento',
       'No. de personas'],
      dtype='object')
```

```
[14] x = df[['Presupuesto','Tiempo invertido min','Tipo','Momento','No. de personas']].values #Var independientes
     y = df['Costo mxn'].values #var independiante
```

```
[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)
```

```
y_test
```

```
array([ 70, 500, 172, 100,  26,  37, 150,  50,  56,  50,  50,  45,  18,
        25, 250, 100,  40, 100,  39,  35, 449,  20, 200,  28, 100,  26,
        10,  23,  10,  50, 230, 150, 100,  45, 300, 100,  40,  47, 300,
       150,  18])
```

```
from sklearn.linear_model import LinearRegression
model_regression = LinearRegression()
```

```
[16] model_regression.fit(x_train, y_train) # aprendizaje automático con base en nuestros datos
```

```
LinearRegression
LinearRegression()
```

```
[17] x_labels = ['Presupuesto', 'Tiempo invertido min', 'Tipo', 'Momento', 'No. de personas']
     c_label = ['Coeficientes']
```

```
x_labels = ['Presupuesto', 'Tiempo invertido min', 'Tipo', 'Momento', 'No. de personas']
c_label = ['Coeficientes']
```

```
[18] coeff_df = pd.DataFrame(model_regression.coef_, x_labels, c_label)
     coeff_df
```

	Coeficientes
Presupuesto	0.401194
Tiempo invertido min	0.803653
Tipo	1.909966
Momento	-40.132375
No. de personas	48.506137

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

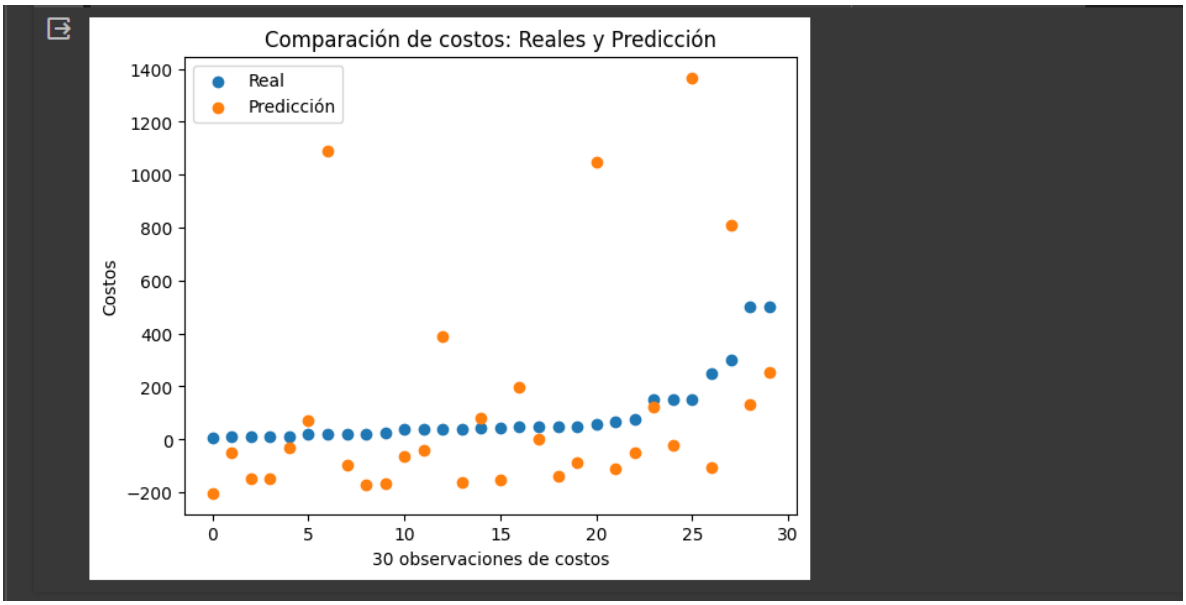
```
residuals = pd.DataFrame({'Real': y_test, 'Predicción': y_pred, 'Residual': y_test - y_pred})
residuals = residuals.sample(n = 30)
residuals = residuals.sort_values(by='Real')
residuals
```

	Real	Predicción	Residual
33	5	-203.733597	208.733597
32	10	-50.796139	60.796139
37	10	-147.169470	157.169470
19	10	-146.367083	156.367083
45	10	-33.143621	43.143621
39	18	72.659351	-54.659351
6	20	1090.499782	-1070.499782
18	20	-98.223850	118.223850
2	20	-172.845860	192.845860
23	25	-166.465052	191.465052
30	37	-62.116426	99.116426
12	37	-40.050778	77.050778
8	39	389.836762	-350.836762
48	40	-160.073827	200.073827

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

```
-9.49388818082518
```

```
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(30), residuals['Real'], label = "Real")  # creamos el gráfico con la muestra de datos reales
plt.scatter(np.arange(30), residuals['Predicción'], label = "Predicción")  # creamos el gráfico con la muestra de datos de

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

plt.xlabel("30 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
```



Explicación del código

En primer lugar, importamos la librería Pandas para poder analizar datos. Después de eso, insertamos nuestro documento en la variable "df" y luego recuperamos los datos para que solo queden datos numéricos.

Después de eso, usamos los métodos "isnull" e "info" para confirmar que no hay valores nulos. A continuación, extraemos los nombres de las columnas usando el método de columns y declaramos x con las variables independientes, mientras que y tiene el valor de la variable de costo.

Después usamos sklearn para asignar el 80% de los datos para el aprendizaje del modelo, y el otro 20% para el conjunto de prueba. Posteriormente, de la librería SciKit-Learn, importamos la clase que nos permite crear un modelo de regresión lineal al cual se asignamos los campos de x learn y y learn para que aprenda de ellos.

Después de eso desplegamos los coeficientes de cada variable independiente mediante un dataframe y usando el método ‘coef’ , después de eso con el método

predict, creamos las predicciones de los costos y las imprimimos en un dataframe con los valores reales y los residuos

Finalmente desplegamos la grafica de puntos que representan el valor real de un registro contra puntos de otro color, que representan los valores de las predicciones de dichos registros.

FASE 4: MODELACIÓN DE LOS DATOS

Como su nombre lo dice, el objetivo de esta fase, es crear y validar un modelo de predicción de nuestros datos, el cual puede ser elaborado, mediante técnicas y algoritmos, como la regresión lineal, regresión logística, redes neuronales, conjunto de arboles y entre otras técnicas de estadística y aprendizaje automático.

Posteriormente, se evalúa la eficiencia del modelo, haciendo uso de ciertas métricas, como en nuestro caso, el coeficiente de regresión lineal, que para ser considerado preciso, debe tener un valor cercano al 100% es decir un valor cercano al 1.0

Después de eso tenemos que tener en cuanta para que sirve y que va a demostrar nuestro modelo creado, dependiendo de ello, vamos a decidir como representarlo, por ejemplo, si lo que vamos a amostrar es una comparación podemos hacer uso de gráficos de barras o incluso de tablas.

Sin embargo, lo que queremos mostrar son datos relacionados con la distribución acerca de algo, podemos utilizar las graficas de histogramas, histogramas en forma lineal y si lo que queremos es graficar datos que representan una composición, lo más conveniente será usar graficas circulares o gráficos de áreas

En nuestro caso al ser un modelo de relación entre variables, los gráficos que mas se adecuan a nuestra necesidad, son los gráficos de dispersión, en los que podemos ver los puntos de relación que existen entre las variables.

Cuestionario

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

Después de ejecutar el código de Python en Google colab, la principal problemática que encontré en mis resultados, fue que los coeficientes de regresión lineal, no eran muy cercanos al cien por ciento, lo cual quiere decir que nuestras predicciones no serán tan exactas ni precisas, como lo queremos.

Otro problema que identifiqué dentro de mi programa, fue los valores residuales, ya que como ya mencioné, los valores de los coeficientes, me indican que mi modelo no fue muy confiable, y esto se vio reflejado en los valores de las predicciones, que están muy alejados con respecto a los valores reales y por lo tanto los valores absolutos de la columna de residuales dan números muy elevados.

Posteriormente en la gráfica dispersión de valores reales contra valores de la predicción, se visualizan resultados muy diferentes a los esperados, ya que el modelo de predicción creo datos muy alejados a los datos del costo real, y esto lo podemos ver en la gráfica con los puntos de predicción muy separados con respecto a los datos reales.

La conclusión que obtenemos de este modelo, es que las predicciones no son confiables, al menos con esa cantidad de datos, ya que como sabemos, mientras mas datos, se generara un modelo más preciso y funcional, pero por el momento el modelo generado en Python no es confiable.

¿Qué resultados arrojó el análisis?

Los primeros resultados que fueron arrojados por nuestro programa de Python fueron los coeficientes de regresión lineal de las variables independientes, es decir del presupuesto, tiempo invertido en minutos, Tipo de gasto, Momento del día, y el numero de personas involucradas. Los datos fueron mostrados en una tabla.

```
model_regression.fit(x_train, y_train) # aprendizaje automático con base en nuestros datos
```

```
LinearRegression()
```

+ Código + Texto

	Coeficientes
Presupuesto	0.401194
Tiempo invertido min	0.803653
Tipo	1.909966
Momento	-40.132375
No. de personas	48.506137

Incluye imagen de cada resultado y explica cada uno de los resultados:

En la siguiente tabla podemos ver los resultados de los valores reales, los valores de la predicción y de los residuales, es decir la diferencia entre los reales y la predicción, en este caso, mientras los valores residuales sean menores quiere decir que tendremos un modelo más confiable.

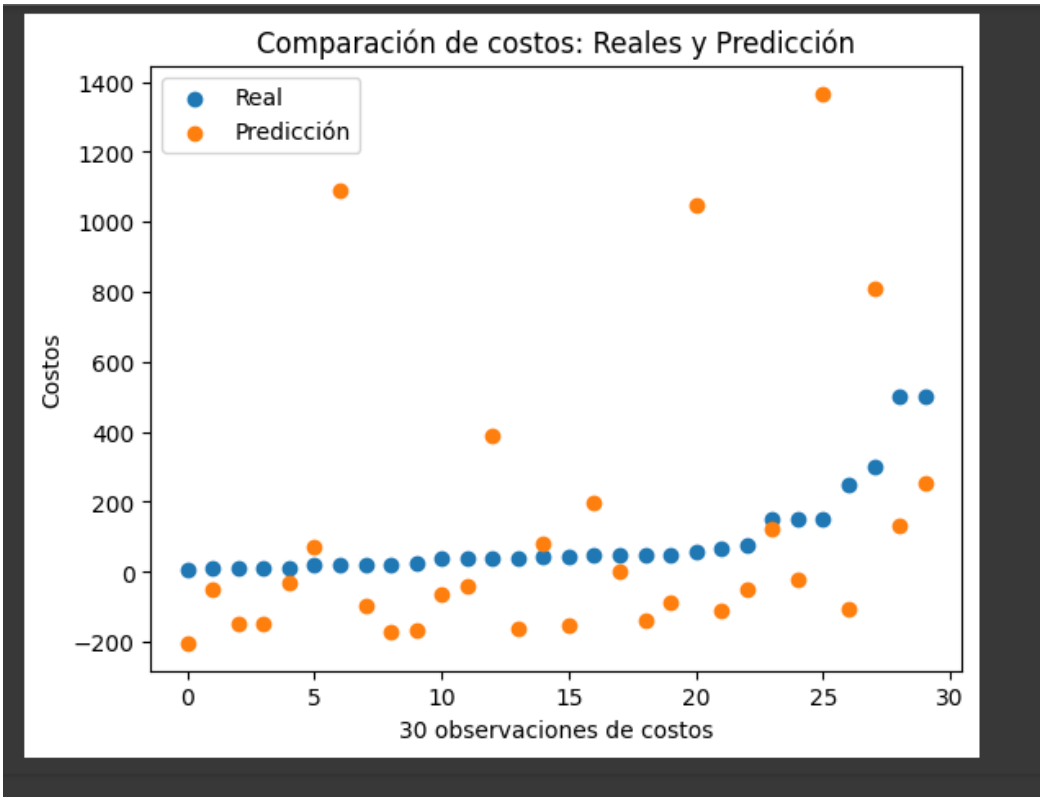
	Real	Predicción	Residual
33	5	-203.733597	208.733597
32	10	-50.796139	60.796139
37	10	-147.169470	157.169470
19	10	-146.367083	156.367083
45	10	-33.143621	43.143621
39	18	72.659351	-54.659351
6	20	1090.499782	-1070.499782
18	20	-98.223850	118.223850
2	20	-172.845860	192.845860
23	25	-166.465052	191.465052
30	37	-62.116426	99.116426
12	37	-40.050778	77.050778
8	39	389.836762	-350.836762
48	40	-160.073827	200.073827
22	45	81.845916	-36.845916
3	45	-150.445180	195.445180
51	50	199.862637	-149.862637
13	50	1.607196	48.392804
46	50	-140.816534	190.816534
35	50	-86.589236	136.589236
44	56	1046.010148	-990.010148
25	65	-109.152440	174.152440
28	75	-51.647068	126.647068
17	150	123.372537	26.627463
10	150	-19.948242	169.948242
40	150	1365.750625	-1215.750625
27	250	-108.209400	358.209400
41	300	809.490178	-509.490178
5	500	133.589147	366.410853
50	500	254.902808	245.097192

En este caso los valores resuduales son muy altos, esto quiere decir que el modelo de prediccion no es confiable.

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

```
-9.49388818082518
```

El coeficiente de determinación R2 nos entregó un valor de 9.49, lo cual es un número muy alejado del 100 por ciento, lo que significa que los resultados, o las predicciones del modelo de Python no son precisas, o que los gastos reales, son muy diferentes a los gastro generados por e modelo de predicción.



Como lo podemos ver en la gráfica de comparación entre costos (Reales y Predicción) los datos generados por el programa de Python son muy poco parecidos a los datos reales, es decir el modelo genero resultados poco confiables y esto lo vemos reflejados en esta grafica.

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

Los resultados que nos arrojó el modelo de Python, es decir la estadística descriptiva, los coeficientes de regresión, los valores residuales, el coeficiente de determinación y la gráfica de dispersión nos muestran que no podemos confiar en el modelo generado por Python, al menos no todavía, esto puede cambiar si le agregamos mas datos a nuestro registro en Excel.

Los coeficientes de regresión nos devolvieron valores muy alejados al 100 por ciento, lo cual quiere decir que no podemos asegurar que nuestra predicción sea muy cercana al valor real, esto lo confirmamos mediante la grafica de comparación entre valores reales y los valores de la predicción y también mediante los valores residuales, que nos indican que la distancia entre los costos reales y los costos resultantes en el modelo de predicción son muy grandes.

En conclusión, el modelo generado en Python no es confiable, debido a que las predicciones generadas por el mismo están muy lejos de los valores de los costos reales introducidos en Excel.

EVALUANDO MIS FINANZAS PERSONALES

¿Cuántas actividades diarias registraste en total en este semestre?

Considerando que desde el día 14 de agosto del 2023 hasta el día 19 de noviembre del 2023 hice un total de 300 registros en el documento de Excel, tenemos que se realizaron alrededor de 4 registros diarios en promedio, ya que fueron un total de 12 semanas, sin contar las 2 semanas “tec”, en las cuales no hubo registros.

```
df.describe()
```

	Costo mxn	Presupuesto	Tiempo invertido min	Tipo	Momento	No. de personas
count	300.000000	300.000000	300.000000	300.000000	300.000000	300.000000
mean	148.940000	781.580000	34.330000	2.416667	1.796667	1.653333
std	682.136269	1532.222574	37.769603	1.999094	0.704967	1.130227
min	5.000000	20.000000	1.000000	1.000000	1.000000	1.000000
25%	23.000000	200.000000	10.000000	1.000000	1.000000	1.000000
50%	45.000000	333.000000	15.000000	1.000000	2.000000	1.000000
75%	100.000000	573.750000	45.000000	4.000000	2.000000	2.000000
max	10000.000000	20000.000000	240.000000	6.000000	3.000000	8.000000

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

El registro en el que mas presupuesto tuve fue el primero con un presupuesto de 20000, mientras que las actividades con menor presupuesto, fueron la compra de chicles y de un café.

std	682.136269	1532.222574	37.769603	1.999094	0.704967	1.130227
min	5.000000	20.000000	1.000000	1.000000	1.000000	1.000000
25%	23.000000	200.000000	10.000000	1.000000	1.000000	1.000000
50%	45.000000	333.000000	15.000000	1.000000	2.000000	1.000000
75%	100.000000	573.750000	45.000000	4.000000	2.000000	2.000000
max	10000.000000	20000.000000	240.000000	6.000000	3.000000	8.000000

```
df.loc[df.loc[:, 'Presupuesto'] == 20000]
```

	Costo mxn	Presupuesto	Tiempo invertido min	Tipo	Momento	No. de personas
0	10000	20000	1	5	1	1

```
df.loc[df.loc[:, 'Presupuesto'] == 20]
```

	Costo mxn	Presupuesto	Tiempo invertido min	Tipo	Momento	No. de personas
59	18	20	10	1	1	2
105	12	20	10	1	3	1

¿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 gastos en el que mas dinero invierto es en el tipo numero 1

Y en el que menos gasto es el numero 3.

```
print(df.groupby('Tipo').get_group(1))
```

```
     Costo mxn  Presupuesto  Tiempo invertido min  Tipo  Momento  \
3           50         3800                    10     1        1
4           47         3750                    10     1        1
5          150         3703                    45     1        1
7          100         3508                    30     1        3
8           20         3408                    10     1        3
..         ...          ...                   ...   ...      ...
293         20          218                    10     1        2
295         50          300                    30     1        1
296         37          250                    10     1        2
298        150          173                    45     1        2
299        300          500                    30     1        1

     No. de personas
3                  1
4                  1
5                  2
7                  2
8                  1
..               ...
293                1
295                4
296                2
298                4
299                4

[189 rows x 6 columns]
```

```
[29]  print(df.groupby('Tipo').get_group(3))
```

	Costo mxn	Presupuesto	Tiempo invertido min	Tipo	Momento	\
9	150	3388	3	3	2	
21	40	2292	90	3	1	
31	1000	2782	20	3	1	
37	39	1552	10	3	2	
46	40	918	90	3	2	
68	40	565	90	3	1	
84	40	350	90	3	1	
107	40	200	90	3	1	
134	40	140	90	3	1	
153	50	200	90	3	1	
174	50	170	120	3	1	
196	50	100	120	3	1	
217	50	150	90	3	1	
237	50	100	90	3	2	
258	50	112	90	3	2	
297	40	213	60	3	2	

¿Por cuántos días registraste tus gastos en este semestre?

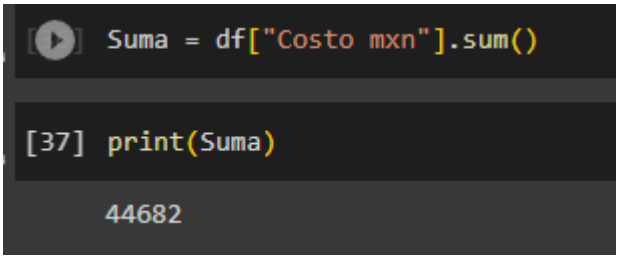
El total de días por los que realice los registros, fueron 84 dias

```
print(df.nunique)
```

```
<bound method DataFrame.nunique of      Costo mxn  Presupuesto  Tiempo invertido min  Tipo  Momento  \
0        10000        20000                     1     5        1
1         4200        10000                     1     5        1
2         2000         5800                     1     2        1
3           50         3800                    10     1        1
4           47         3750                    10     1        1
..         ...          ...                   ...   ...      ...
295         50          300                    30     1        1
296         37          250                    10     1        2
297         40          213                    60     3        2
298        150          173                    45     1        2
299        300          500                    30     1        1

     No. de personas
0                  1
1                  1
2                  1
3                  1
4                  1
..               ...
295                4
296                2
297                2
298                4
299                4

[300 rows x 6 columns]>
```

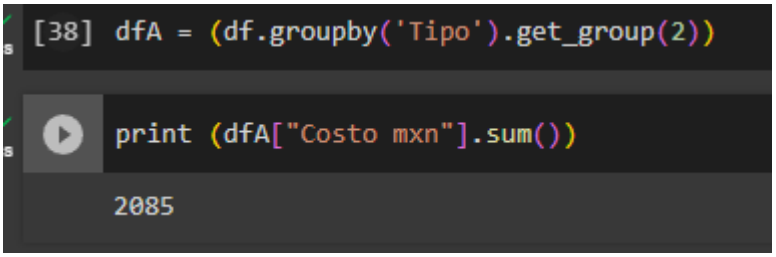
¿Cuál fue el total de tus gastos en este semestre?

El total de mis gastos fue de 44682 pesos mexicanos



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

El total de mis ahorros, fue de 2080



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

Considerando que ahorre 2085 pesos en 4 meses y me quiero comprar algo de 3000, tendría que ahorrar durante otro mes más.

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

Considero, que, si quiero mejorar mis finanzas personales, debo hacer un presupuesto inicial, antes de que empiece cada mes y antes de hacer todos mis gastos del mes, debo asignar un porcentaje a mis ahorros.

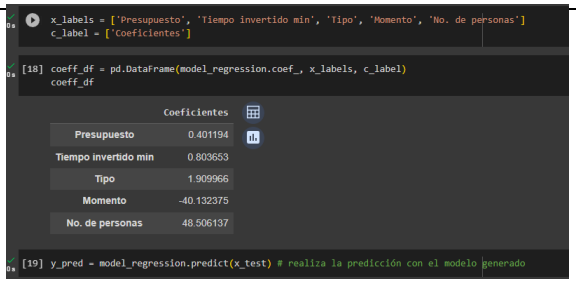
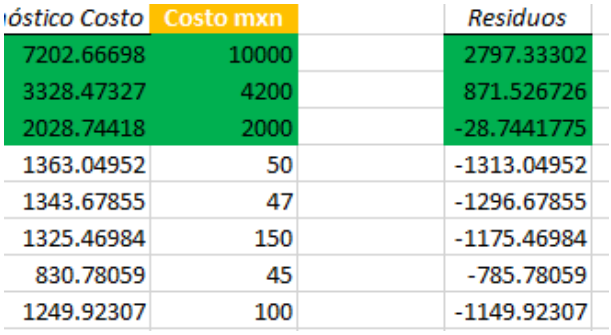
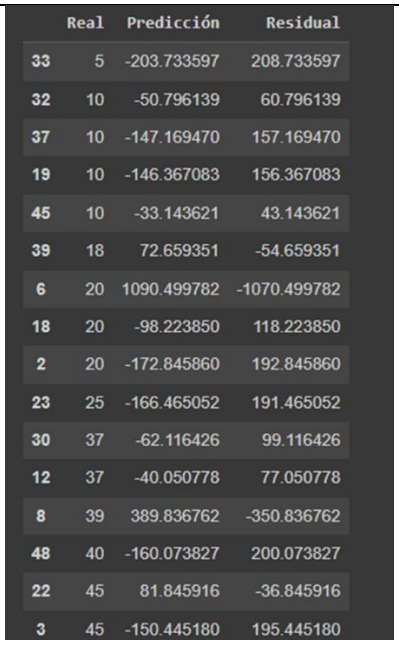
¿Cómo visualizas tus finanzas personales en un año?

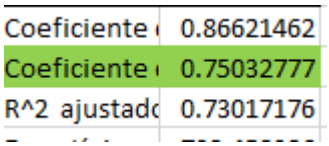
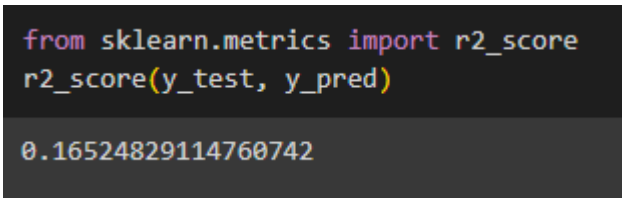
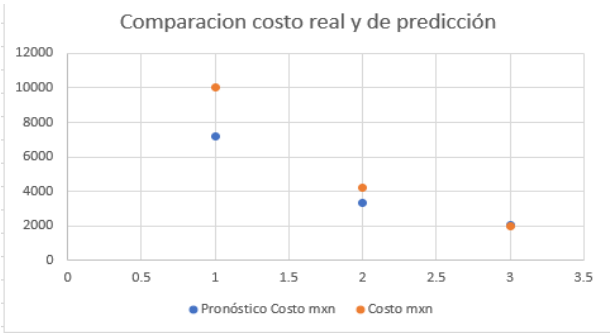
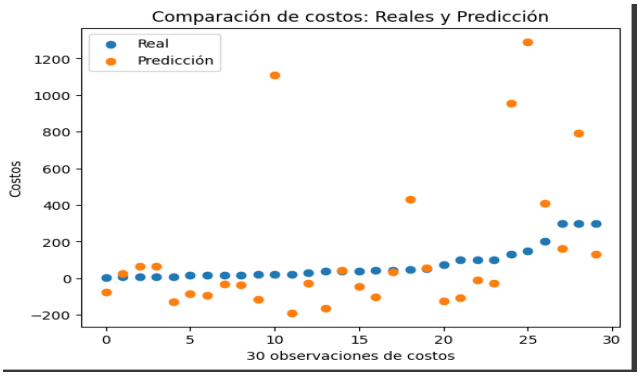
Considero que mis metas financieras para dentro de un año son establecer un porcentaje de ahorro e inversión, asi mismo tener un ahorro y una inversión considerable.

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

Creo que el mayor aprendizaje que me llevo de este proyecto, fue el hecho de aprender que el análisis de datos puede predecir el comportamiento de un fenómeno, y el mayor reto que se me presento durante el desarrollo del trabajo fue el hecho de contabilizar mis gastos durante todo el transcurso del semestre.

REFLEXION FINAL

Excel	Python
	
La tabla de exploración de Excel, me entrego datos estadísticos muy precisos y fáciles de interpretar	Mientras que Python me entrego resultados muy poco claros y difíciles de interpretar.
	
Los coeficientes que me brindo Excel fueron más cercanos a uno	Mientras que, en Python, los coeficientes fueron muy lejanos al 100%
	
Los residuales brindados por Excel son muy grandes, esto puede deberse a la poca cantidad de datos que se tenían para ese momento	En Python los residuales fueron menores, pero no tan parecidos a los valores originales.

 El coeficiente de determinación brindado por Excel es de 0.75 lo que le da una mayor fiabilidad a las predicciones.	 El coeficiente de determinación r2 en Python es mucho menor que el de Excel (0.16) lo que disminuye la confiabilidad del modelo
 La grafica proporcionada por Excel muestra únicamente tres valores reales y tres predicciones, las predicciones se parecen mucho a los valores reales pero al tratarse de una base de datos muy pequeña el modelo no puede ser mas preciso	 Mientras que la gráfica proporcionada por Python cuenta con más valores, debido a que los registros dentro del programa son mayores, aun así, los valores de las predicciones están muy lejos de los valores reales, por lo que se considera que el modelo no es muy preciso.

Responde la hipótesis inicial: ¿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?

Es posible predecir el costo de las actividades, en función del presupuesto disponible para la actividad, tipo de actividad, momento de realización y número de personas, el modelo creado en Python no brinda resultados, muy precisos, esto puede deberse a que no todas las variables son influyentes para determinar el costo.

La falta de precisión del modelo, también puede deberse a la poca cantidad de datos, y también podría aumentar el porcentaje de acierto en caso de que aumentaran el numero de datos en los registros, es decir que si es posible predecir el costo de las actividades, pero siempre y cuando la cantidad y la calidad de los datos sea la adecuada.

LINK AL PROGRAMA (GOOGLE COLAB)

https://colab.research.google.com/drive/1BHZmJs_92C-paKWqq6809uxRFOyEcts1?usp=sharing

REFERENCIAS

¿Qué es la regresión lineal? - Explicación del modelo de regresión lineal - AWS. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/linear-regression/

¿Qué es el análisis predictivo? - Explicación del análisis predictivo - AWS. (s. f.). Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/predictive-analytics/#:~:text=El%20an%C3%A1lisis%20predictivo%20consiste%20en,y%20extrapolar%20las%20tendencias%20ocultas.

¿Qué es el aprendizaje automático? (s. f.). Oracle México. https://www.oracle.com/mx/artificial-intelligence/machine-learning/what-is-machine-learning/