Proyecto II
Semillero Análisis de Datos (AD)
Semestre 2024-1

Consideraciones iniciales.

Etapas en el AD:

etapa I: **descriptiva o exploratoria.**

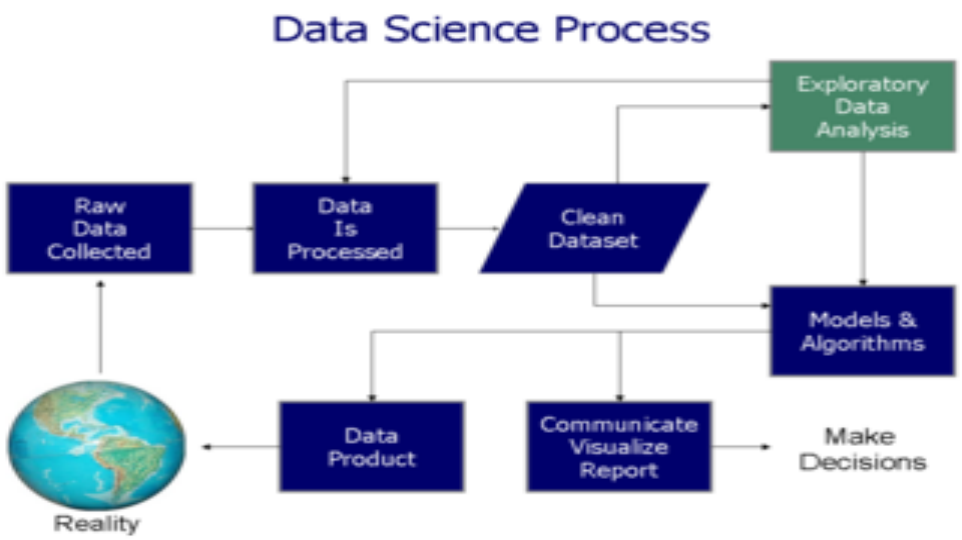
1. Adquisición de datos (SQL, búsqueda en bases de datos).
2. Procesamiento de datos (parsing – transformar de un formato a otro).
3. Exploración de datos (EDA).
4. Limpieza de datos (DC – data cleaning).
5. Métodos de estadística descriptiva.

Objetivo: obtener conocimiento sobre los datos que tenemos. Preparación de los datos para la etapa II. Sacar algunas conclusiones valiosas.

Etapa II: **predictiva o inferencial**.

1. Utilización de modelos basados en algoritmos.
2. Métodos de estadística inferencial.
2. Métodos de ML.
3. Obtener predicciones, métricas, realizar ingeniería de características.
4. Validación de modelos.
5. Presentación de resultados.

Objetivo: realizar predicciones y obtener un conocimiento más profundo sobre las correlaciones de las diferentes características (features) de los datos.



En el proyecto I vamos a trabajar en la etapa I del AD. Para eso:

1. Adquiera los datos (obténgalos en el aula en TEAMS).
2. Análisis de datos exploratorio.

Descripción de los datos. La base de datos esta formada por datos tomados de 30000 clientes de bancos que pagan por sus créditos. Las características (features) en el dataset están distribuidas en tres grupos:

1. características objetivas (variables categóricas)
   - sexo (SEX)
   - educacion (EDUCATION)

Gradación para esta variable categórica (según Taiwan):
2 – edu. Technica
1 - edu. Media.
3 – bachelor (pregrado)
4, 5, 6, 0 - edu. Superior

   - Estado civil (MARRIAGE)

1 – soltero.
2 – casado.
3 – divorciado
0 – viudo.

   - edad (AGE)
   - Realizacion del Pago (Pay_0,….,Pay6)
   - default.payment.next.month (target).

2. mediciones
   - Pagos (Bill_AMT1,….., Bill_AMT6)
3. características irrelevantes
   - ID

La característica target (es lo que se desea predecir) – será un cliente capaz de pagar el crédito bancario el próximo meses? (default.payment.next.month).

Pasos para realizar EDA:

comience importando Numpy y Pandas. Utilizaremos la base de datos UCI_Credit_Card.csv. Descárguela del aula del semilleroAD en TEAMS. Lea ese archivo con Pandas. Mire el encabezado (cinco primeras filas del archivo) y la cola del archivo.

A. Exploración y adecuación del Dataset.

1. Excluya las características irrelevantes.
2. Revise si los datos están completos. Hay datos faltantes?
3. Separe las características de la característica objetivo.
4. Analice la característica EDUCATION. ¿Qué observa?, ¿Cómo la podría transformar?
5. Analice la característica MARRIAGE. ¿Qué observa?, ¿Cómo la podría transformar?

B. Visualización.

1. Grafique la frecuencia de faltas (defaults).
2. Obtenga un recuento de la estadística (utilice el método describe).
3. Grafique las variables categóricas.
4. Grafique las variables medibles.
5. Cree box plots para los pagos.
6. Normalice todas las variables.
7. Cree violin plots para todas las variables.
8. Cree la matrix de correlación para todas las variables.
9. Utilice una regresión lineal para tratar de distinguir entre default y non-default en la categoría objetivo default.payment.next.month (target).

C. Data Cleaning.

1. De acuerdo a lo observado en los puntos 6, 7, 8, 9 determine si es necesario realizar limpieza de datos.

D. IDA.

1. Utilice train_test_split para separar el dataset original en una parte para entrenamiento (training) y otra para test (test).
1a. Utilice train_test_split para separar el dataset estandartizado en una parte para entrenamiento (training) y otra para test (test).
2. Instancie el algoritmo LogisticRegression de Sckit-Learn.

3. Cree la variable feature_names y extraiga del dataset las características más importantes.
4. Establezca el valor de los hiperparámetros del modelo, utilice np.logspace(-5, 8, 15).
5. Instancie los algoritmos LogisticRegression y RandomizedSearchCV.
6.Realice el ajuste de los datos.
7. Imprimir los parámetros del algoritmo.
8. Realice LogisticRegression con parametros C=0.00005, random_state=0.
9. Cree la matrix de confusión.
10. Determine la precisión del algoritmo con metrics_accuracy_score.