**1) En clase hemos visto el concepto de modelo relacional para bases de datos. Usted ha sido contratado por una universidad para crear la base de datos de la misma. La universidad consta de alumnos, profesores, materias, edificios y salones. Diagrame el modelo relacional para la universidad especificando las distintas entidades, las relaciones entre las mismas (por medio de primary y foreign keys) y atributos (al menos uno o dos que crea pertinente) para cada entidad.**

El modelo debe hacerse mediante un diagrama. El mismo debe constar de:

* Entidad Alumno.
  + Atributos:
    - id\_alumno: primary key
    - edad (otro)
    - nombre (otro)
  + Relaciones:
    - M:M con profesores (requiere tabla intermedia)
    - M:M con materias (requiere tabla intermedia)
    - Cualquiera de estas dos relaciones es válida, pero basta con una pues profesores y materias están relacionadas.
* Entidad Profesor.
  + Atributos:
    - id\_profesor: primary key
    - edad (otro)
    - profesión (otro)
  + Relaciones:
    - M:M con alumnos (requiere tabla intermedia)
    - 1:M con materias (puede ser también M:M, asumiendo que una materia puede ser impartida por más de un profesor)
* Entidad Materia.
  + Atributos:
    - id\_materia: primary key
    - id\_profesor: foreign key
    - id\_salon: foreign key
    - nombre (otro)
    - carrera (otro)
  + Relaciones:
    - M:1 con profesores (podría ser M:M también).
    - M:1 con salones (en un salón se dan muchas materias, pero una materia se da en un único salón)
* Entidad Salón.
  + Atributos:
    - id\_salon: primary key
    - id\_edificio: foreign key
    - tamaño (otro)
    - nombre (otro)
  + Relaciones:
    - 1:M con materias
    - M:1 con edificio (un edificio tiene muchos salones, pero un salón no puede estar en más de un edificio).
* Edificio:
  + Atributos:
    - id\_edificio: primary key
    - dirección (otro)
    - nombre (otro).
  + Relaciones:
    - 1:M con salones

**2) a) Usted trabaja para un banco analizando transacciones. Para ciertos clientes usted detecta transacciones muy altas las cuales pueden interpretarse como anomalías o outliers. Qué acción debe tomar usted?**

Si los outliers se dan de manera sistemática no deben ser ni eliminados ni modificados. Por el contrario, se les debe prestar particular atención y lidiar con ellos de manera separada. En este caso, las transacciones altas pueden interpretarse como fraudes o lavado de activos, no deben descartarse.

**b) Otra tarea en el banco que le asignaron es ver como se relaciona la edad de los clientes con sus ingresos. Para esto decide hacer una regresión lineal. Unos pocos clientes son millonarios siendo sus ingresos desproporcionadamente más grandes que el resto. Qué acción debe tomar usted?**

Ya que estamos entrenando un modelo y en concreto una regresión lineal (muy susceptible a outliers) se debe lidiar con ellos apropiadamente para obtener un buen ajuste y generar predicciones precisas. Una manera podría ser un top-coding de los outliers. Otra manera podría ser entrenar un modelo sin estos valores y entrenar otro modelo solo con estos valores. Otra manera podría ser eliminarlos.

**3) En clase hemos hablado repetidas veces del concepto de representatividad. Cuando entrenamos modelos para generar predicciones queremos que la muestra de datos de entrenamiento represente bien la realidad a la que nos enfrentamos para que el modelo ante casos nuevos haya "visto" casos similares.**

**Usted quiere generar predicciones con dos modelos. Un modelo usa datos provenientes de la curva violeta y el otro modelo usa datos provenientes de la curva roja. Suponga que puede seleccionar un número limitado de datos en total, pero usted puede decidir cuántos datos de cada curva tomar. Qué debería hacer usted y por qué?**

La proporción debería ser mayor en favor de la curva roja que de la curva violeta. La curva roja presenta valores muy extremos, tiene kurtosis muy alta. La curva violeta no posee valores extermos, sino que se distribuye de manera uniforme, kurtosis muy baja. Por lo tanto, ya que el modelo debe ser capaz de predecir casos no vistos, es más probable que si hacemos un muestreo aleatorio de la curva violeta cubramos fácilmente todos los casos posibles. En cambio, si hacemos un muestreo aleatorio de la curva roja esto es más improbable por la presencia de valores extermos. Teniendo la posibilidad de elegir la cantidad de muestras es más conveniente tomar más de la curva roja que de la violeta.

**4) Usted trabaja en un hospital. Tiene a su disposición miles de estudios médicos de pacientes, por lo que su dataset contiene muchas columnas. Le asignan investigar si existen grupos de pacientes con las mismas enfermedades. Qué problema cree que puede surgir? Explique el problema y una manera de resolverlo.**

El problema presente es la "Maldición de la dimensionalidad". A medida que aumenta la dimensionalidad de nuestro dataset nuestro espacio se vuelve menos denso o más esparso, por lo que es más difícil encontrar clusters/grupos. Una solución es usar una técnica de reducción de dimensionalidad como PCA, kPCA, t-SNE, UMAP, etc.

**5) Usted es el ministro Salinas y quiere incentivar a la población a vacunarse. Sabe que una de cada cien personas tiene Covid en el país. Qué gráfica haría? Comente al menos 3 decisiones en su diseño.**

Gráficas recomendadas:

* A single large number/word
* Icon array

Opciones de diseño:

* Resaltar con un color el 1 de 100.
* Generar contraste.
* Si es un icon array usar iconos de personas
* Otros

**6) Usted tiene un dado de 12 caras y sospecha que esta trucado. Para ver esto, usted tira el dado 120 veces y observa los valores que salieron y anota en una tabla la cantidad de veces que salio cada valor:**

**valores observados = [11, 15, 18, 11, 4, 5, 8, 15, 7, 8, 12, 6]**

**Luego, usted decide usar un test Chi-cuadrado para ver si el dado esta trucado o no. Plantee las hipótesis del test y determine si efectivamente el dado esta trucado o no, calculando el valor de Chi-cuadrado. Utilice alfa = 0.05 y la tabla adjunta en el parcial.**

Total de experimentos realizados: 120

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Valores observados: [11, 15, 18, 11, 4, 5, 8, 15, 7, 8, 12, 6]

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Valores estimados: [10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10]

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Grados de libertad: (2-1)x(12-1) = 11

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Chi: 21.4

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

p-value: 0.029455991090165515

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Dependent (reject H0)

**7) Por qué es importante la limpieza de datos? De 3 ejemplos de limpieza de datos.**

Por qué es importante:

* GIGO (Garbage In Garbage Out). Si un modelo se entrena con malos datos dará malos resultados.
* Ciertos modelos aceptan solo un tipo de datos particular. Por ejemplo, puede no aceptar strings, datetimes, booleans, etc.
* Los datos en su formato original pueden no tener sentido alguno para nuestros modelos. Ejemplos:
  + Un dato ordinal expresado como texto no tiene relevancia, pero sí cuando es expresado de manera numérica.
  + En NLP los textos en crudo no tienen sentido, debe hacerse una Bag of words.
* Algunos modelos no aceptan valores faltantes.

Ejemplos de limpiezas:

* Imputar/remover valores faltantes.
* Tratar/remover outliers.
* Cambiar data types.
* Uso de distintos transformers (One hot encoder, Ordinal encoder, Target encoder, String Indexer, Vector Assembler, etc).
* Otros.

**8) Usted trabaja para la NASA. El robot Mars Rover recaba datos de la superficie marciana y es alimentado por energía solar para funcionar. La NASA dispone una estación meteorológica que recaba información sobre el clima de la zona donde se encuentra el robot, como viento, temperatura, nubosidad, lluvia, etc. A fin de mes, usted analiza los datos que recolectó Mars Rover y ve que faltan algunos valores. Qué tipo de missingness cree que son esos valores? Cómo justifica la falta de valores?**

En este caso, el missingness del MARS (no pun intended) Rover es MAR (Missing At Random). La falta del valor se debe a variables observadas. En este caso, el robot funciona con energía solar. Si no hay mediciones se debe a un malfuncionamiento del mismo, el cual puede ser explicado por variables observadas, como la nubosidad (las nubes taparon el sol y el robot quedó sin batería, no pudiendo seguir recolectando información).

**9) Suponga que usted quiere predecir una variable objetivo, la cual es continua, a partir de diez variables de entrada, también todas continuas. Usted dispone de estas diez variables, pero no sabe si son relevantes para predecir la variable objetivo. Cómo determinaría usted qué variables deberían permanecer y qué variables descartar?**

En los modelos lineales, cuanto mayor correlación exista entre las variables de entrada y la variable objetivo mejor serán nuestras predicciones. Puede suceder que existan variables no muy correlacionadas a la clase a predecir y variables muy correlacionadas entre sí (redundancia). Puede hacerse un gráfico de correlaciones para ver esto y seleccionar qué columnas deben permanecer.

Una mejor opción sería usar un algoritmo como PCA para obtener un número concreto de componentes principales. Se pierde muy poca informacion, se reduce la redundancia y se obtienen nuevas variables no correlacionadas entre sí.

**10) Por qué, cuando no se sabe la distribución de la cual provienen los datos, conviene siempre guiarse por gráficas y no por macro-estadísticas (promedio, moda, desviación estándar, etc)? Ilustre con un ejemplo.**

Distribuciones muy distintas pueden tener las mismas macroestadísticas. Conocida si la distibución, nos da información extra pero de por sí no son suficientes. Ver cuarteto de Anscombe.