Logo, company name

Description automatically generated

**Introducción a la Ciencia de Datos**

Parcial abril 2021

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 para cada entidad (al menos uno o dos que crea pertinente).

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?

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?

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 anteriormente.

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?

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.

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.

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.

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

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?

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?

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.

Chart

Description automatically generated

