Logo, company name

Description automatically generated

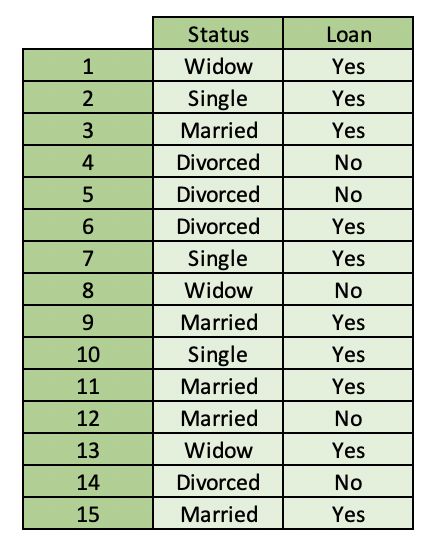
**Introducción a la Ciencia de Datos**

Parcial 09 de Octubre 2023

1. Suponga que usted trabaja como científico de datos para una empresa de e-commerce que quiere mejorar su sistema de recomendación. El conjunto de datos de la empresa incluye información sobre las interacciones de los clientes con los productos, como compras, clicks y búsquedas. Qué tipo de features pueden crearse a partir de este tipo de datos?
2. Su cliente le proporciona un dataset para que usted entrene un algoritmo de clasificación binaria. El cliente le sugiere que quiere evaluar la performance predictiva usando *accuracy*. Usted, tras analizar la distribución del *target* o columna objetivo, le aconseja no usar esta métrica. Qué es lo que usted notó particular en la distribución y qué métrica aconsejaría utilizar?
3. Un modelo fue entrenado para predecir el valor de propiedades en el rubro de bienes raíces con datos del 1990. Usted utiliza este modelo para predecir valores de propiedades actuales y realiza a posteriori un análisis de los residuos obteniendo el diagrama que se ve en el anexo. Interprete el mismo y justifique que acción se debe tomar.
4. Por qué cree que es importante lidiar con outliers? Justifique su respuesta dando al menos tres razones.
5. Por qué razones cree que las gráficas presentes en el anexo no son apropiadas? Justifique su respuesta.
6. Suponga que usted posee un dataset de 3 columnas numéricas continuas, la cual una de ellas presenta missing values. Como imputaría los valores faltantes de la misma utilizando KNN? Describa detalladamente el proceso incluyendo las fórmulas que considere necesarias.
7. Explique los distintos componentes de una serie de tiempo en base al ejemplo ilustrado en el anexo.
8. Calcular la Information Gain para la columna Status presente en la tabla del anexo.
9. Usted trabaja en un nuevo hospital. El director del mismo le solicita que cree la base de datos del mismo. Realice el diagrama Entidad-Relación del mismo identificando al menos 5 entidades, especificando las relaciones entre las mismas, incluyendo atributos y keys.
10. Corredores de distintos países corren 10 carreras y obtienen los tiempos que se muestran en la tabla del anexo. La desviación estándar para los corredores del país 1 es de 103 segundos mientras que para los corredores del país 2 es de 101 segundos. Estableciendo un nivel de significancia alpha = 0.05, existe una diferencia significativa en la media de tiempos de ambos corredores?

La tabla del anexo posee valores negativos del z-score porque es de la cola izquierda, pero es igual para los valores positivos del z-score si fuera la cola derecha.

Table

Description automatically generated 

Anexo

A graph with blue dots

Description automatically generated A newspaper with a map of the united states

Description automatically generated

A screenshot of a graph

Description automatically generated A graph of a graph of a job loss

Description automatically generated with medium confidence

Table

Description automatically generated