DATATHON UNI 2019

PRE-SELECCIÓN

El siguiente cuestionario consta de **05 preguntas** que servirán como referencia para la selección de los participantes para la I DATATHON UNI 2019. Los lenguajes de programación que deben ser utilizados para dar soluciones a cada pregunta son R y/o Python. Las respuestas a las preguntas junto al código utilizado deben de ser adjuntados por correo a datascienceresearchperu@gmail.com con copia a responsable@ucds.com y jordan.rodriguez@pichincha.pe con el asunto "**Pre-Selección DATATHON UNI 2019**", incluir nombre del equipo.

Pregunta 01:

Elaborar una función que asegure de que el nombre y apellido de las personas comiencen con una letra mayúscula en sus pasaportes. Por ejemplo, alison heck debe escribirse correctamente en mayúsculas como Alison Heck.

alison heck ⇒ Alison Heck

- **Formato de entrada:** Una sola línea de entrada que contiene el nombre completo
- **Formato de salida:** Imprima la cadena convertida.
- **Restricciones:** La cadena consta de caracteres alfanuméricos y espacios.

Nota: en una palabra, solo el primer carácter deberá estar en mayúscula. Ejemplo 12abc cuando se escribe con mayúscula permanece 12abc.

Pregunta 02:

Epidural y fiebre en el parto: Hay estudios que relacionan el uso de anestesia epidural con la aparición de fiebre materna. Además, se sospecha que la epidural afecta también a la duración del parto y al estado de salud del bebé. Para comprobar estas hipótesis se ha seleccionado una muestra de 731 mujeres embarazas que acuden al hospital a parir. Las variables recogidas para el estudio se encuentran en la base de datos epidural.sav (http://bit.ly/DataSet1-Datathon-UNI) y son las siguientes:

N°	Variable	Descripción
01	id	Identificación de la paciente.
02	edad	Edad de la madre.
03	pesonac	Peso en gramos del recién nacido.
04	tipopar	Tipo de parto (inducido, instrumental o cesárea).
05	oxitocin	Uso de oxitocina durante el parto (si/no).
06	epidural	Uso de anestesia epidural durante el parto (si/no).
07	temp1	Temperatura antes del parto.
08	temp2	Temperatura después del parto.
09	dilataci	Duración de la dilatación en minutos.
10	expulsiv	Duración del expulsivo en minutos.

11	apgar1	Puntuación del test de apgar al minuto de vida del recién nacido.
12	apgar2	Puntuación del test de apgar a los 5 minutos de vida del recién nacido.

Nota: *Para cargar un archivo sav instale el paquete foreign.*

A la vista de estos datos:

a) Realizar un análisis descriptivo de las variables edad, peso al nacer, tipo de parto y uso de oxitocina. Describir los resultados.
b) Realizar el mismo análisis distinguiendo dos grupos: uno que utiliza anestesia epidural y otro que no. Describir los resultados
c) Estudiar el porcentaje de mujeres que utiliza anestesia epidural
d) Estudiar la temperatura media de la madre antes y después del parto. ¿Existen en general diferencias significativas?, ¿Y si distinguimos por grupos según el uso de anestesia epidural?
e) ¿Existe relación entre el uso de oxitocina y la aparición de fiebre?

Pregunta 03:

Realice una inspección del archivo ListaCompanias.csv (http://bit.ly/DataSet2-Datathon-UNI) para luego realizar las siguientes tareas:

a) Generar de manera aleatoria 5 conjuntos (5 portafolios) cada uno con 10 elementos (10 empresas - columna Symbol). La condición para que un conjunto (portafolio) sea admitido es que los elementos (empresas) pertenezcan a diferentes sectores (columna Sector). Como resultado mostrar la suma de las capitalizaciones de mercado (columna MarketCap) de cada uno de los conjuntos (portafolios).
b) Escribir un script que ingrese a la ruta en Summary Quote y extraiga de Key Stats los siguientes valores: volume,open y market cap; de cada una de las empresas que componen los portafolios generados en la pregunta anterior. Son en total 50 registros, así que debe programar una versión automatizada para esto. Obtener la empresa de máximo “market cap” en cada portafolio.

Pregunta 04:

Preguntas teóricas

a) ¿Qué metodologías usas para un proyecto de Data Science?, y de la metodología elegida, ¿qué parte del proceso consideran que es el más importante? Explicar.
b) Según los tipos de modelos, ¿Cuáles son los indicadores de performance para datos desbalanceado? Explicar.
c) ¿Qué estrategias utilizarías para tratar problemas de overfitting? Explicar.
d) Se cuenta con un dataset de 2500 registros y 10 variables. ¿Considera necesario el uso de Feature Engineering? Explicar.

Pregunta 05:

Responder las preguntas sobre el siguiente conjunto de datos(https://archive.ics.uci.edu/ml/datasets/Adult).

In [1]:
```
import numpy as np
import pandas as pd
pd.set_option('display.max.columns', 100)
# to draw pictures in jupyter notebook
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
# we don't like warnings
# you can comment the following 2 lines if you'd like to
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
data = pd.read_csv('../../data/adult.data.csv')
data.head()
```

Out[2]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

a) ¿Cuántos hombres y mujeres(sex) están representados en este conjunto de datos?

Selecciona la respuesta:

- 21790 men and 10771 women
- 16346 men and 12532 women
- 21790 women and 10771 men
- 16346 women and 12532 men

b) ¿Cuál es la edad promedio (age) de las mujeres?

Selecciona la respuesta:

- 34.67
- 35.95
- 36.86
- 37.04

c) ¿Cuál es el número máximo de horas que una persona trabaja por semana (hours-per-week)? ¿Cuántas personas trabajan menos de 40 horas y de este grupo, cuál es el porcentaje de los que ganan mucho (> 50K)?

Selecciona la respuesta:

- 102 hours/week, 20 people, 41% son ricos
- 99 hours/week, 95 people, 30% son ricos
- 99 hours/week, 85 people, 29% son ricos
- 90 hours/week., 70 people, 34% son ricos

d) Calcule el tiempo promedio de trabajo (hours-per-week) para aquellos que ganan poco y mucho (salary) para cada país (native-country). ¿Cuál es el resultado para Japón?

Selecciona la respuesta:

- 41 y 48
- 46 y 43
- 44 y 48
- 41 y 40

Adjuntar el código en el correo con la respuesta