

## Examen unidad 4 SAA

Nombre: \_\_\_\_\_

1. (1 pts.) Tenemos una variable “peripheral pulse” que puede tomar los siguientes valores: normal, increased, reduced, absent. ¿Qué tendríamos que hacer, suponiendo que no tenga valores nan, para efectuar la siguiente asignación?

1 = normal  
2 = increased  
3 = reduced  
4 = absent

```
dataset['peripheral_pulse'] =  
dataset['peripheral_pulse'].map({'increased':2,'normal':1,'reduced':3,'absent':4}).astype('float64')
```

2. (1 pts.) La variable lesion\_1 es de tipo numérico donde el primer número indica el lugar de la lesión y se podría considerar que dicha información es importante para predecir la supervivencia del caballo:

1 = gastric  
2 = sm intestine  
3 = lg colon  
4 = lg colon and cecum  
5 = cecum  
6 = transverse colon  
7 = retum/descending colon  
8 = uterus  
9 = bladder  
11 = all intestinal sites  
00 = none

¿Qué hay que hacer para crear otra variable con dicha información?

```
dataset['site of lesion'] = dataset['lesion_1'] // 1000
```

3. (1 pts.) La variable “pulse” tiene un 8.03% de valores nan, si en vez de inferir esos datos decidiésemos eliminar las filas correspondientes, ¿cómo se haría?

```
dataset = dataset.dropna(subset=["pulse"])
```

4. (1 pts.) ¿En la variable pulse se observa algún outlier?, explícalo

Si

5. (1 pts.) ¿Cómo harías para categorizar la variable ‘pulse’?

1: <= 88 y >= 48  
2: > 88  
3: < 48

**Examen unidad 4 SAA**

Nombre:

```
dataset.loc[dataset['pulse'] < 48, 'pulse'] = 3  
dataset.loc[dataset['pulse'] > 88, 'pulse'] = 2  
dataset.loc[(dataset['pulse'] <= 88) & (dataset['pulse'] >= 48), 'pulse'] = 1
```

6. (1 pto.) ¿Qué es la discretización? Pon un ejemplo aplicado a la práctica de horse colic.

```
dataset['rectal_temp'] = pd.qcut(dataset['rectal_temp'], 3, labels=["High", "Normal", "Low"])
```

7. (1 pto.) ¿Qué es 1 hot encoding? Pon un ejemplo aplicado a la práctica de horse colic.

El 1 hot encoding intenta paliar el posible efecto adverso de convertir valores no numéricos en una interpretación numérica: el 1 está numéricamente "más cerca" del 2 que el 3. Podemos utilizar la función `get_dummies` de pandas:

```
y = pd.get_dummies(dataset['temp_of_extremities'], prefix='1HotEnc')
```

8. (1 pto.) En la práctica se ha aplicado PCA para 15 componentes y para 2 componentes, ¿qué conclusiones puedes inferir?

9. (1 pto.) En Backward Elimination, ¿la salida es correcta? ¿Hay algún error?

```
X = dataset.drop("outcome",1)  
y = dataset["outcome"]
```

10. (1 pto.) ¿Qué diferencia existe entre correlación de Pearson y Spearman?