
# Reconocimiento de patrones

## Tarea 2

Limpieza de datos I

**Alumno:**

Pérez Rodríguez Raúl Francisco

**Octubre 2017**

## Analice los problemas de valores faltantes en el conjunto de datos Pima Indians Diabetes completo.

**Número y porcentaje de valores nulos por columna**

| variable | cantidad | porcentaje |
|---|---|---|
| emb | 0 | 0.00 % |
| gl2h | 5 | 0.65 % |
| pad | 35 | 4.55 % |
| ept | 227 | 29.55 % |
| is2h | 374 | 48.69 % |
| imc | 11 | 1.43 % |
| fpd | 0 | 0.00 % |
| edad | 0 | 0.00 % |
| class | 0 | 0.00 % |

Cinco de las ocho variables presentan valores faltantes 'gl2h' tiene un grado de impacto trivial 'pad' y 'imc' tienen un grado de impacto manejable 'ept' y 'is2h' tienen un grado de impacto crítico.

## Realice la imputación de los datos utilizando 3 aproximaciones diferentes y compare los resultados.

**Rellenando con la media**

| type | gl2h | pad | ept | is2h | imc |
|---|---|---|---|---|---|
| **count** | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| **mean** | 121.686763 | 72.405184 | 29.153420 | 155.548223 | 32.457464 |
| **std** | 30.435949 | 12.096346 | 8.790942 | 85.021108 | 6.875151 |
| **min** | 44.000000 | 24.000000 | 7.000000 | 14.000000 | 18.200000 |
| **25%** | 99.750000 | 64.000000 | 25.000000 | 121.500000 | 27.500000 |
| **50%** | 117.000000 | 72.202592 | 29.153420 | 155.548223 | 32.400000 |
| **75%** | 140.250000 | 80.000000 | 32.000000 | 155.548223 | 36.600000 |
| **max** | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 |

**Rellenando con la mediana**

| type | gl2h | pad | ept | is2h | imc |
|---|---|---|---|---|---|
| **count** | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| **mean** | 121.656250 | 72.386719 | 29.108073 | 140.671875 | 32.455208 |
| **std** | 30.438286 | 12.096642 | 8.791221 | 86.383060 | 6.875177 |
| **min** | 44.000000 | 24.000000 | 7.000000 | 14.000000 | 18.200000 |
| **25%** | 99.750000 | 64.000000 | 25.000000 | 121.500000 | 27.500000 |
| **50%** | 117.000000 | 72.000000 | 29.000000 | 125.000000 | 32.300000 |
| **75%** | 140.250000 | 80.000000 | 32.000000 | 127.250000 | 36.600000 |
| **max** | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 |

**Rellenando con la moda**

| type | gl2h | pad | ept | is2h | imc |
|---|---|---|---|---|---|
| **count** | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| **mean** | 121.686763 | 72.405184 | 29.153420 | 155.420253 | 32.457464 |
| **std** | 30.535641 | 12.382158 | 10.476982 | 118.652291 | 6.924988 |
| **min** | 44.000000 | 24.000000 | 7.000000 | 14.000000 | 18.200000 |
| **25%** | 99.000000 | 64.000000 | 22.000000 | 76.500000 | 27.500000 |
| **50%** | 117.000000 | 72.000000 | 29.000000 | 125.000000 | 32.300000 |
| **75%** | 141.000000 | 80.000000 | 36.000000 | 190.000000 | 36.600000 |
| **max** | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 |

En las variables de 'gl2h', 'pad' y 'imc' no presentan mucha diferencia en los resultados. En las variables de 'ept' y 'is2h' presentan cambios significativos en algunas de las descripciones. En 'ept' rellenando con la mediana y media no hay cambios significativos, pero rellenando con la moda presentan cambios significativos en la desviación y en el primer y tercer percentil. En el caso de 'is2h' la mayoría de las descripciones son diferentes.

## Realice una estimación de valores faltantes mediante interpolación.

Usando interpolación lineal en las variables con valores faltantes

| type | gl2h | pad | ept | is2h | imc |
|---|---|---|---|---|---|
| **count** | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 |
| **mean** | 121.492839 | 72.330729 | 29.080078 | 159.045098 | 32.465365 |
| **std** | 30.546675 | 12.207864 | 9.887474 | 111.597578 | 6.889880 |
| **min** | 44.000000 | 24.000000 | 7.000000 | 14.000000 | 18.200000 |
| **25%** | 99.000000 | 64.000000 | 22.000000 | 88.000000 | 27.500000 |
| **50%** | 117.000000 | 72.000000 | 29.000000 | 130.000000 | 32.400000 |
| **75%** | 140.250000 | 80.000000 | 36.000000 | 190.500000 | 36.600000 |
| **max** | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 |