Práctica 1

Inteligencia de Negocio

Amin Kasrou Aouam

UNIVERSIDAD
DE GRANADA

2020-11-10

Índice

Experiments

We will first try to gather information about our dataset, by evaluating the statistics of our attributes.

```
from pandas import read_csv
from sklearn.preprocessing import LabelEncoder


def replace_values(df):
    columns = ["BI-RADS", "Margin", "Density", "Age"]
    for column in columns:
        df[column].fillna(value=df[column].mean(), inplace=True)
    return df


def process_na(df, action):
    if action == "drop":
        return df.dropna()
    return replace_values(df)


def encode_columns(df):
    encoder = LabelEncoder()
    encoder.fit(df["Shape"])


def parse_data(source, action):
    df = read_csv(filepath_or_buffer=source, na_values="?")
    processed_df = process_na(df, action)
    return processed_df
```

```
df = parse_data("../data/mamografia.csv", "drop")
print(df.describe())
```

```
           BI-RADS         Age      Margin     Density
count  847.000000  847.000000  847.000000  847.000000
mean     4.322314   55.842975    2.833530    2.909091
std      0.703762   14.603754    1.564049    0.370292
min      0.000000   18.000000    1.000000    1.000000
25%      4.000000   46.000000    1.000000    3.000000
50%      4.000000   57.000000    3.000000    3.000000
75%      5.000000   66.000000    4.000000    3.000000
max      6.000000   96.000000    5.000000    4.000000
```

We observe that **margin** and **density** are the columns with the most unknown values. The age group of our cohort is middle aged, the BI-RADS score is mostly in the suspicious category, the density is mostly low and the margin belongs to the microlobulated/obscured category.

We'll try to impute values, instead of dropping them, when they're invalid.

```
df = parse_data("../data/mamografia.csv", "replace")
print(df.describe())
```

```
           BI-RADS         Age      Margin     Density
count  961.000000  961.000000  961.000000  961.000000
mean     4.296142   55.487448    2.796276    2.910734
std      0.705555   14.442373    1.526880    0.365074
min      0.000000   18.000000    1.000000    1.000000
25%      4.000000   45.000000    1.000000    3.000000
50%      4.000000   57.000000    3.000000    3.000000
75%      5.000000   66.000000    4.000000    3.000000
max      6.000000   96.000000    5.000000    4.000000
```