# Dataset Wine – Normalización en Notebook en la nube de IBM

Se provee una implementación en un jupyter notebook en donde se realizarán las tareas requeridas. La solución fue implementada en Watson Studio (IBM Cloud), la primera celda del notebook refiere a la forma de cargar el dataset.

```
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

df = pd.read_csv(body)
df.head(10)
```

Out[18]:

| | Class | Alcohol | Malic acid | Ash | Alcalinity of ash | Magnesium | Total phenols | Flavanoids | Nonflavanoid phenols | Proanthocyanins | Color intensity | Hue | OD280/OD315 of diluted wines | Proline |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 14.23 | 1.71 | 2.43 | 15.6 | 127 | 2.80 | 3.06 | 0.28 | 2.29 | 5.64 | 1.04 | 3.92 | 1065 |
| 1 | 1 | 13.20 | 1.78 | 2.14 | 11.2 | 100 | 2.65 | 2.76 | 0.26 | 1.28 | 4.38 | 1.05 | 3.40 | 1050 |
| 2 | 1 | 13.16 | 2.36 | 2.67 | 18.6 | 101 | 2.80 | 3.24 | 0.30 | 2.81 | 5.68 | 1.03 | 3.17 | 1185 |
| 3 | 1 | 14.37 | 1.95 | 2.50 | 16.8 | 113 | 3.85 | 3.49 | 0.24 | 2.18 | 7.80 | 0.86 | 3.45 | 1480 |
| 4 | 1 | 13.24 | 2.59 | 2.87 | 21.0 | 118 | 2.80 | 2.69 | 0.39 | 1.82 | 4.32 | 1.04 | 2.93 | 735 |
| 5 | 1 | 14.20 | 1.76 | 2.45 | 15.2 | 112 | 3.27 | 3.39 | 0.34 | 1.97 | 6.75 | 1.05 | 2.85 | 1450 |
| 6 | 1 | 14.39 | 1.87 | 2.45 | 14.6 | 96 | 2.50 | 2.52 | 0.30 | 1.98 | 5.25 | 1.02 | 3.58 | 1290 |
| 7 | 1 | 14.06 | 2.15 | 2.61 | 17.6 | 121 | 2.60 | 2.51 | 0.31 | 1.25 | 5.05 | 1.06 | 3.58 | 1295 |
| 8 | 1 | 14.83 | 1.64 | 2.17 | 14.0 | 97 | 2.80 | 2.98 | 0.29 | 1.98 | 5.20 | 1.08 | 2.85 | 1045 |
| 9 | 1 | 13.86 | 1.35 | 2.27 | 16.0 | 98 | 2.98 | 3.15 | 0.22 | 1.85 | 7.22 | 1.01 | 3.55 | 1045 |

In [19]:
```
#df[["Alcohol"]] = pd.to_numeric(df_data_1.Alcohol, errors='coerce')
#3, 4 5 y 6 con una única línea de código
# Mediante el describe, podemos ver que todos los atributos son numéricos. De lo contrario, podríamos utilizar el código comentado anteriormente para convertir el tipo a numérico.
df.describe(include='all')
```

Out[19]:

| | Class | Alcohol | Malic acid | Ash | Alcalinity of ash | Magnesium | Total phenols | Flavanoids | Nonflavanoid phenols | Proanthocyanins | Color intensity | Hue | OD280/OD315 of diluted wines | Proline |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 | 178.000000 |
| mean | 1.938202 | 13.000618 | 2.336348 | 2.366517 | 19.494944 | 99.741573 | 2.295112 | 2.029270 | 0.361854 | 1.590899 | 5.058090 | 0.957449 | 2.611685 | 746.893258 |
| std | 0.775035 | 0.811827 | 1.117146 | 0.274344 | 3.339564 | 14.282484 | 0.625851 | 0.998859 | 0.124453 | 0.572359 | 2.318286 | 0.228572 | 0.709990 | 314.907474 |
| min | 1.000000 | 11.030000 | 0.740000 | 1.360000 | 10.600000 | 70.000000 | 0.980000 | 0.340000 | 0.130000 | 0.410000 | 1.280000 | 0.480000 | 1.270000 | 278.000000 |
| 25% | 1.000000 | 12.362500 | 1.602500 | 2.210000 | 17.200000 | 88.000000 | 1.742500 | 1.205000 | 0.270000 | 1.250000 | 3.220000 | 0.782500 | 1.937500 | 500.500000 |
| 50% | 2.000000 | 13.050000 | 1.865000 | 2.360000 | 19.500000 | 98.000000 | 2.355000 | 2.135000 | 0.340000 | 1.555000 | 4.690000 | 0.965000 | 2.780000 | 673.500000 |
| 75% | 3.000000 | 13.677500 | 3.082500 | 2.557500 | 21.500000 | 107.000000 | 2.800000 | 2.875000 | 0.437500 | 1.950000 | 6.200000 | 1.120000 | 3.170000 | 985.000000 |
| max | 3.000000 | 14.830000 | 5.800000 | 3.230000 | 30.000000 | 162.000000 | 3.880000 | 5.080000 | 0.660000 | 3.580000 | 13.000000 | 1.710000 | 4.000000 | 1680.000000 |

**Normalizar MinMax y estandarizar**

In [20]:
```
df_class = df["Class"]
df = df.drop(columns=["Class"])
df_min_max = df.copy()
df_normalized = df.copy()
```

In [21]:
```
# 8 Normalización (transformación Z)
df_normalized =(df_normalized-df_normalized.mean())/df_normalized.std()
df_normalized["Class"] = df_class
df_normalized.head()
```

Out[21]:

| | Alcohol | Malic acid | Ash | Alcalinity of ash | Magnesium | Total phenols | Flavanoids | Nonflavanoid phenols | Proanthocyanins | Color intensity | Hue | OD280/OD315 of diluted wines | Proline | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1.514341 | -0.560668 | 0.231400 | -1.166303 | 1.908522 | 0.806722 | 1.031908 | -0.657708 | 1.221438 | 0.251009 | 0.361158 | 1.842721 | 1.010159 | 1 |
| 1 | 0.245597 | -0.498009 | -0.825667 | -2.483841 | 0.018094 | 0.567048 | 0.731565 | -0.818411 | -0.543189 | -0.292496 | 0.404908 | 1.110317 | 0.962526 | 1 |
| 2 | 0.196325 | 0.021172 | 1.106214 | -0.267982 | 0.088110 | 0.806722 | 1.212114 | -0.497005 | 2.129959 | 0.268263 | 0.317409 | 0.786369 | 1.391224 | 1 |
| 3 | 1.686791 | -0.345835 | 0.486554 | -0.806975 | 0.928300 | 2.484437 | 1.462399 | -0.979113 | 1.029251 | 1.182732 | -0.426341 | 1.180741 | 2.328007 | 1 |
| 4 | 0.294868 | 0.227053 | 1.835226 | 0.450674 | 1.278379 | 0.806722 | 0.661485 | 0.226158 | 0.400275 | -0.318377 | 0.361158 | 0.448336 | -0.037767 | 1 |

In [22]:
```
# 7 MinMax
df_min_max =(df_min_max-df_min_max.min())/(df_min_max.max()-df_min_max.min())
df_min_max["Class"] = df_class
df_min_max.head()
```

Out[22]:

| | Alcohol | Malic acid | Ash | Alcalinity of ash | Magnesium | Total phenols | Flavanoids | Nonflavanoid phenols | Proanthocyanins | Color intensity | Hue | OD280/OD315 of diluted wines | Proline | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.842105 | 0.191700 | 0.572193 | 0.257732 | 0.619565 | 0.627586 | 0.573840 | 0.283019 | 0.593060 | 0.372014 | 0.455285 | 0.970696 | 0.561341 | 1 |
| 1 | 0.571053 | 0.205534 | 0.417112 | 0.030928 | 0.326087 | 0.575862 | 0.510549 | 0.245283 | 0.274448 | 0.264505 | 0.463415 | 0.780220 | 0.550642 | 1 |
| 2 | 0.560526 | 0.320158 | 0.700535 | 0.412371 | 0.336957 | 0.627586 | 0.611814 | 0.320755 | 0.757098 | 0.375427 | 0.447154 | 0.695971 | 0.646933 | 1 |
| 3 | 0.878947 | 0.239130 | 0.609626 | 0.319588 | 0.467391 | 0.989655 | 0.664557 | 0.207547 | 0.558360 | 0.556314 | 0.308943 | 0.798535 | 0.857347 | 1 |
| 4 | 0.581579 | 0.365613 | 0.807487 | 0.536082 | 0.521739 | 0.627586 | 0.495781 | 0.490566 | 0.444795 | 0.259386 | 0.455285 | 0.608059 | 0.325963 | 1 |

In [23]:
```
from sklearn.model_selection import train_test_split
# 9 Creación de train y test
df_normalized_y = df_normalized["Class"]
df_normalized_x = df_normalized.drop(columns=["Class"])
#Realizo la division con el dataset normalizado con transformación z, pero se podría hacer análogamente para el minmax. Se aplica un shuffle al dataset antes de realizar el split
#con la seed 42
x_train, x_test, y_train, y_test = train_test_split(
    df_normalized_x, df_normalized_y, test_size=0.10, random_state=42
)
```

In [27]:
```
train = x_train.join(y_train)
train.head()
```

Out[27]:

| | Alcohol | Malic acid | Ash | Alcalinity of ash | Magnesium | Total phenols | Flavanoids | Nonflavanoid phenols | Proanthocyanins | Color intensity | Hue | OD280/OD315 of diluted wines | Proline | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **9** | 1.058578 | -0.882918 | -0.351810 | -1.046527 | -0.121938 | 1.094330 | 1.122011 | -1.139816 | 0.452690 | 0.932547 | 0.229909 | 1.321588 | 0.946649 | 1 |
| **114** | -1.134008 | -0.847112 | 0.486554 | 0.899835 | -1.102159 | 0.423244 | 0.261028 | 0.547563 | -0.962506 | -0.930899 | -0.120091 | 0.814539 | -1.149205 | 2 |
| **18** | 1.465069 | -0.668085 | 0.413653 | -0.896807 | 0.578221 | 1.605634 | 1.902902 | -0.336302 | 0.470162 | 1.570950 | 1.192408 | 0.293405 | 2.963114 | 1 |
| **66** | 0.134736 | -1.187265 | -2.429493 | -1.345967 | -1.522254 | 1.094330 | 1.152045 | -0.818411 | 1.203967 | 0.104349 | 0.711158 | 0.800454 | -0.777667 | 2 |
| **60** | -0.826061 | -1.106702 | -0.315359 | -1.046527 | 0.088110 | -0.391646 | -0.940343 | 2.154591 | -2.063214 | -0.771298 | 1.279908 | -1.326335 | -0.212422 | 2 |

In [28]:
```
test = x_test.join(y_test)
test.head()
```

Out[28]:

| | Alcohol | Malic acid | Ash | Alcalinity of ash | Magnesium | Total phenols | Flavanoids | Nonflavanoid phenols | Proanthocyanins | Color intensity | Hue | OD280/OD315 of diluted wines | Proline | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **19** | 0.787585 | 0.683574 | 0.705257 | -1.286079 | 1.138347 | 0.646939 | 1.001874 | -1.541573 | 0.120730 | 0.018078 | 0.011159 | 1.053978 | 0.311541 | 1 |
| **45** | 1.489705 | 1.525003 | 0.267850 | -0.178150 | 0.788268 | 0.886613 | 0.621440 | -0.497005 | -0.595603 | 0.078468 | -0.382591 | 1.011724 | 1.057792 | 1 |
| **140** | -0.086987 | 0.423984 | 1.215566 | 0.450674 | -0.261969 | -1.206537 | -1.531017 | 1.351077 | -1.469181 | -0.197599 | -0.820091 | -0.424915 | -0.466465 | 3 |
| **30** | 0.898446 | -0.748647 | 1.215566 | 0.899835 | 0.088110 | 1.126287 | 1.222125 | -0.577356 | 1.378682 | 0.276890 | 1.017408 | 0.138473 | 1.708777 | 1 |
| **67** | -0.776789 | -1.044043 | -1.627580 | 0.031458 | -1.522254 | -0.295777 | -0.029303 | -0.738059 | -0.962506 | -0.163090 | 0.711158 | 1.222995 | -0.752263 | 2 |