

Tratamento dos dados

```
In [ ]: import numpy as np
import pandas as pd
```

```
In [ ]: df = pd.read_csv('../datasets/raw.csv')
df.head()
```

```
Out [ ]:
```

	Data	Energia_usada	V1	V2	V3	V4	V5	V6	V7	V8	V9
0	01/01/2018 00:15	3.17	2.95	0.0	0.0	73.21	100.0	900	Dia da Semana	Segunda- feira	Carga leve
1	01/01/2018 00:30	4.00	4.46	0.0	0.0	66.77	100.0	1800	Dia da Semana	Segunda- feira	Carga leve
2	01/01/2018 00:45	3.24	3.28	0.0	0.0	70.28	100.0	2700	Dia da Semana	Segunda- feira	Carga leve
3	01/01/2018 01:00	3.31	3.56	0.0	0.0	68.09	100.0	3600	Dia da Semana	Segunda- feira	Carga leve
4	01/01/2018 01:15	3.82	4.50	0.0	0.0	64.72	100.0	4500	Dia da Semana	Segunda- feira	Carga leve

```
In [ ]: df.columns = ['data', 'consumo_energia', 'corrente_atrasada', 'corrente_p...
```

```
In [ ]: memory_before = df.memory_usage(deep=True).sum()
memory_before
```

```
Out [ ]: 11722932
```

```
In [ ]: df.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35040 entries, 0 to 35039
Data columns (total 11 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   data                                  35040 non-null  object
1   consumo_energia                      35040 non-null  float64
2   corrente_atrasada                    35040 non-null  float64
3   corrente_principal                   35040 non-null  float64
4   co2                                  35040 non-null  float64
5   potencia_atrasado                    35040 non-null  float64
6   potencia_principal                   35040 non-null  float64
7   segundos_depois_meia_noite           35040 non-null  int64
8   estado_semana                        35040 non-null  object
9   dia_semana                           35040 non-null  object
10  tipo_carga                           35040 non-null  object
dtypes: float64(6), int64(1), object(4)
memory usage: 11.2 MB
```

```
In [ ]: df.nunique()
```

```
Out[ ]: data          35040
        consumo_energia  3343
        corrente_atrasada 1954
        corrente_principal 768
        co2              8
        potencia_atrasado 5079
        potencia_principal 3366
        segundos_depois_meia_noite 96
        estado_semana    2
        dia_semana       7
        tipo_carga       3
        dtype: int64
```

```
In [ ]: df.data = pd.to_datetime(df.data, format='%d/%m/%Y %H:%M')
        df.co2 = df.co2.astype('float16')
        df.segundos_depois_meia_noite = df.segundos_depois_meia_noite.astype('int32')
        df.estado_semana = df.estado_semana.astype('category')
        df.dia_semana = df.dia_semana.astype('category')
        df.tipo_carga = df.tipo_carga.astype('category')
```

```
In [ ]: memory_after = df.memory_usage(deep=True).sum()
        memory_after
```

```
Out[ ]: 1998745
```

```
In [ ]: df.info(memory_usage='deep')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35040 entries, 0 to 35039
Data columns (total 11 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   data                                  35040 non-null  datetime64[ns]
1   consumo_energia                      35040 non-null  float64
2   corrente_atrasada                    35040 non-null  float64
3   corrente_principal                    35040 non-null  float64
4   co2                                  35040 non-null  float16
5   potencia_atrasado                    35040 non-null  float64
6   potencia_principal                    35040 non-null  float64
7   segundos_depois_meia_noite           35040 non-null  int32
8   estado_semana                        35040 non-null  category
9   dia_semana                           35040 non-null  category
10  tipo_carga                            35040 non-null  category
dtypes: category(3), datetime64[ns](1), float16(1), float64(5), int32(1)
memory usage: 1.9 MB
```

```
In [ ]: print (f'{memory_after/memory_before*100: .2f}% da memória original')
        17.05% da memória original
```

```
In [ ]: df.to_pickle('../datasets/processed.pkl')
```