Questão 1

In [137…

```python
import pandas as pd
from sklearn.datasets import load_breast_cancer

cancer_data = load_breast_cancer()
df = pd.DataFrame(cancer_data.data, columns=cancer_data.feature_names)
df.head()
```

Out[137…

	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	...	25.38	17.33	184.60
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	...	24.99	23.41	158.80
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	...	23.57	25.53	152.50
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	...	14.91	26.50	98.87
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	...	22.54	16.67	152.20

5 rows × 30 columns

In [138…

```python
df['target'] = cancer_data.target
df.target.value_counts()
```

Out[138…

	count
target	
1	357
0	212

dtype: int64

In [139…

```
df.isnull().sum()
```

Out[139…

	0
mean radius	0
mean texture	0
mean perimeter	0
mean area	0
mean smoothness	0
mean compactness	0
mean concavity	0
mean concave points	0
mean symmetry	0
mean fractal dimension	0
radius error	0
texture error	0
perimeter error	0
area error	0
smoothness error	0
compactness error	0
concavity error	0
concave points error	0
symmetry error	0
fractal dimension error	0
worst radius	0
worst texture	0
worst perimeter	0

	0
worst area	0
worst smoothness	0
worst compactness	0
worst concavity	0
worst concave points	0
worst symmetry	0
worst fractal dimension	0
target	0

dtype: int64

In [140…

```
count_class_0, count_class_1 = df.target.value_counts()
count_class_0, count_class_1
```

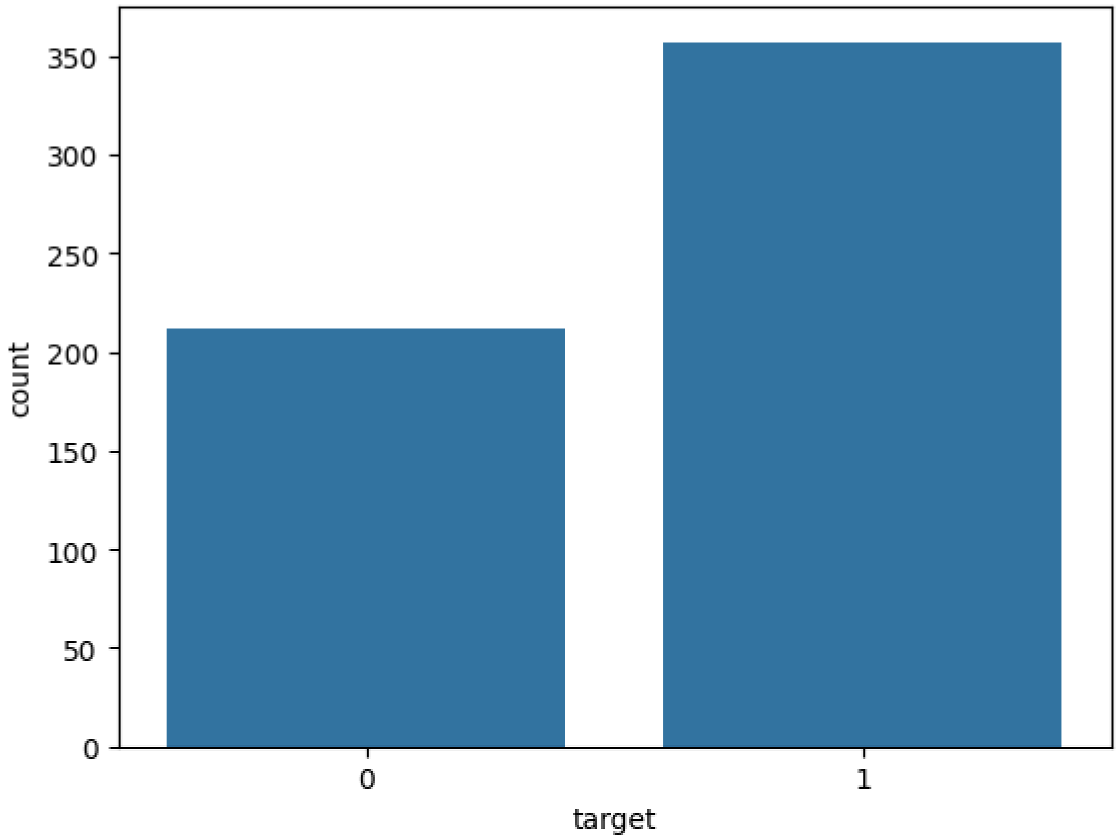
Out[140…

```
(357, 212)
```

In [141…

```
import seaborn as sns
import matplotlib.pyplot as plt

sns.countplot(x='target', data=df)
plt.show()
```



In [142…

```
target_0 = df[df['target'] == 0]
target_1 = df[df['target'] == 1]

target_0.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 212 entries, 0 to 567
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   mean radius              212 non-null    float64
 1   mean texture             212 non-null    float64
 2   mean perimeter           212 non-null    float64
 3   mean area                212 non-null    float64
 4   mean smoothness          212 non-null    float64
 5   mean compactness         212 non-null    float64
 6   mean concavity           212 non-null    float64
 7   mean concave points      212 non-null    float64
 8   mean symmetry            212 non-null    float64
 9   mean fractal dimension   212 non-null    float64
 10  radius error             212 non-null    float64
 11  texture error            212 non-null    float64
 12  perimeter error          212 non-null    float64
 13  area error               212 non-null    float64
 14  smoothness error         212 non-null    float64
 15  compactness error        212 non-null    float64
 16  concavity error          212 non-null    float64
 17  concave points error     212 non-null    float64
 18  symmetry error           212 non-null    float64
 19  fractal dimension error  212 non-null    float64
 20  worst radius             212 non-null    float64
 21  worst texture            212 non-null    float64
 22  worst perimeter          212 non-null    float64
 23  worst area               212 non-null    float64
 24  worst smoothness         212 non-null    float64
 25  worst compactness        212 non-null    float64
 26  worst concavity          212 non-null    float64
 27  worst concave points     212 non-null    float64
 28  worst symmetry           212 non-null    float64
 29  worst fractal dimension  212 non-null    float64
 30  target                   212 non-null    int64  
dtypes: float64(30), int64(1)
memory usage: 53.0 KB
```

Undersampling

In [143…

```
target_1_undersample = target_1.sample(count_class_1)
target_1_undersample.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 212 entries, 510 to 511
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   mean radius              212 non-null    float64
 1   mean texture             212 non-null    float64
 2   mean perimeter           212 non-null    float64
 3   mean area                212 non-null    float64
 4   mean smoothness          212 non-null    float64
 5   mean compactness         212 non-null    float64
 6   mean concavity           212 non-null    float64
 7   mean concave points      212 non-null    float64
 8   mean symmetry            212 non-null    float64
 9   mean fractal dimension   212 non-null    float64
 10  radius error             212 non-null    float64
 11  texture error            212 non-null    float64
 12  perimeter error          212 non-null    float64
 13  area error               212 non-null    float64
 14  smoothness error         212 non-null    float64
 15  compactness error        212 non-null    float64
 16  concavity error          212 non-null    float64
 17  concave points error     212 non-null    float64
 18  symmetry error           212 non-null    float64
 19  fractal dimension error  212 non-null    float64
 20  worst radius             212 non-null    float64
 21  worst texture            212 non-null    float64
 22  worst perimeter          212 non-null    float64
 23  worst area               212 non-null    float64
 24  worst smoothness         212 non-null    float64
 25  worst compactness        212 non-null    float64
 26  worst concavity          212 non-null    float64
 27  worst concave points     212 non-null    float64
 28  worst symmetry           212 non-null    float64
 29  worst fractal dimension  212 non-null    float64
 30  target                   212 non-null    int64  
dtypes: float64(30), int64(1)
memory usage: 53.0 KB
```

In [144…

```
df_test_undersample = pd.concat([target_1_undersample, target_0], axis=0)
```

```
df_test_undersample.target.value_counts()
```

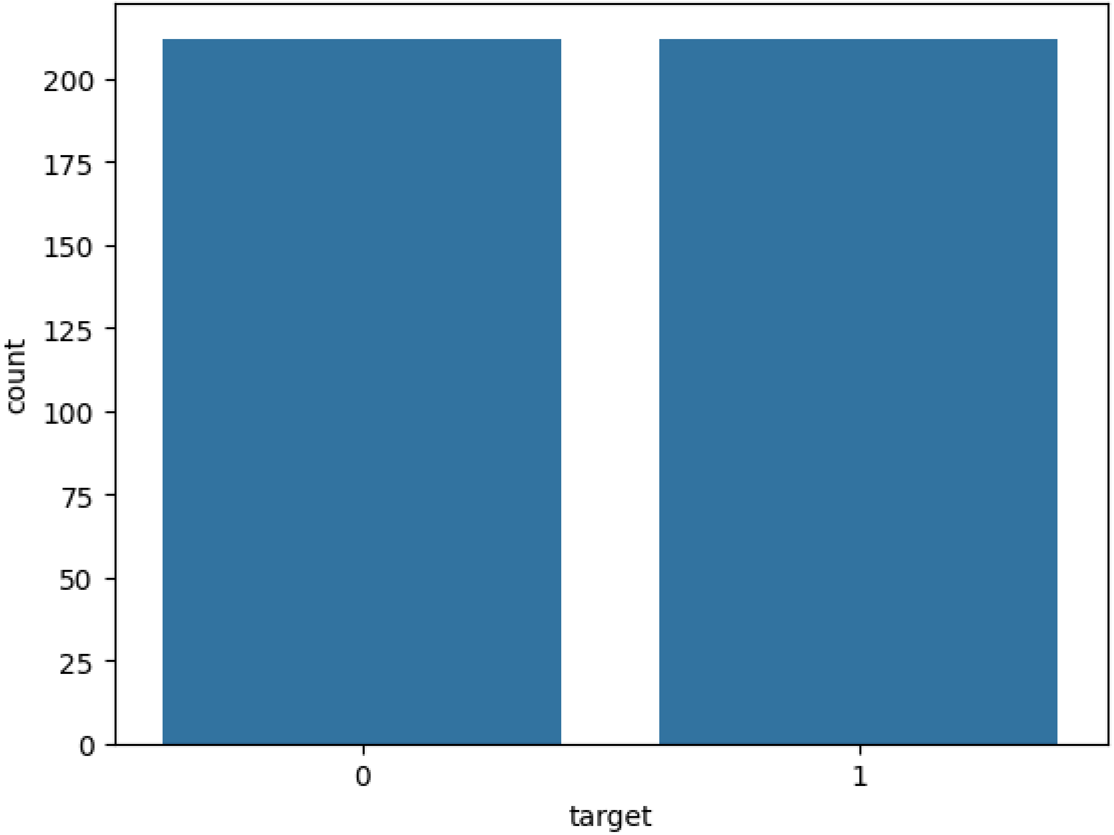
Out[144…

	count
target	
1	212
0	212

dtype: int64

In [145…

```
sns.countplot(x='target', data=df_test_undersample)
plt.show()
```



Oversampling

In [146…

```
target_0_oversample = target_0.sample(count_class_0, replace=True)
target_0_oversample.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 357 entries, 36 to 352
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   mean radius              357 non-null    float64
 1   mean texture             357 non-null    float64
 2   mean perimeter           357 non-null    float64
 3   mean area                357 non-null    float64
 4   mean smoothness          357 non-null    float64
 5   mean compactness         357 non-null    float64
 6   mean concavity           357 non-null    float64
 7   mean concave points      357 non-null    float64
 8   mean symmetry            357 non-null    float64
 9   mean fractal dimension   357 non-null    float64
 10  radius error             357 non-null    float64
 11  texture error            357 non-null    float64
 12  perimeter error          357 non-null    float64
 13  area error               357 non-null    float64
 14  smoothness error         357 non-null    float64
 15  compactness error        357 non-null    float64
 16  concavity error          357 non-null    float64
 17  concave points error     357 non-null    float64
 18  symmetry error           357 non-null    float64
 19  fractal dimension error  357 non-null    float64
 20  worst radius             357 non-null    float64
 21  worst texture            357 non-null    float64
 22  worst perimeter          357 non-null    float64
 23  worst area               357 non-null    float64
 24  worst smoothness         357 non-null    float64
 25  worst compactness        357 non-null    float64
 26  worst concavity          357 non-null    float64
 27  worst concave points     357 non-null    float64
 28  worst symmetry           357 non-null    float64
 29  worst fractal dimension  357 non-null    float64
 30  target                   357 non-null    int64  
dtypes: float64(30), int64(1)
memory usage: 89.2 KB
```

In [147…

```
df_test_oversample = pd.concat([target_0_oversample, target_1], axis=0)
```

```
df_test_oversample.target.value_counts()
```

Out[147…

	count
target	
0	357
1	357

dtype: int64

Questão 2

In [148…

```
from sklearn.datasets import fetch_openml
import numpy as np

yeast = fetch_openml(name='yeast', version=1)
```

In [149…

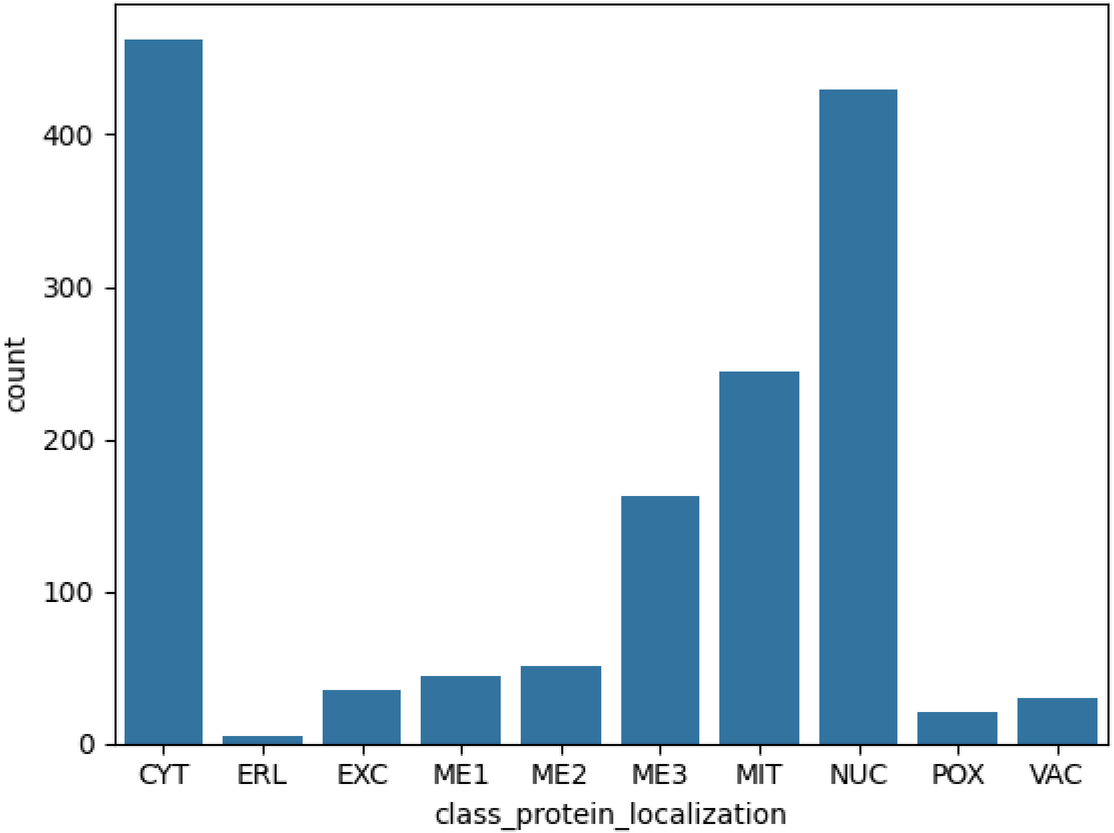
```
np.bincount(yeast.target.cat.codes)
```

Out[149…

```
array([463,   5,  35,  44,  51, 163, 244, 429,  20,  30])
```

In [150…

```
sns.countplot(x=yeast.target)
plt.show()
```



In [151…

```
from imblearn.over_sampling import SMOTE
```

In [152…

```
smote = SMOTE(k_neighbors=4) # k_neighbors >= 5 gera erro nesse dataset: ValueError: Expected n_neighbors <= n_samp

X_smote, y_smote = smote.fit_resample(yeast.data, yeast.target)
```

In [153…

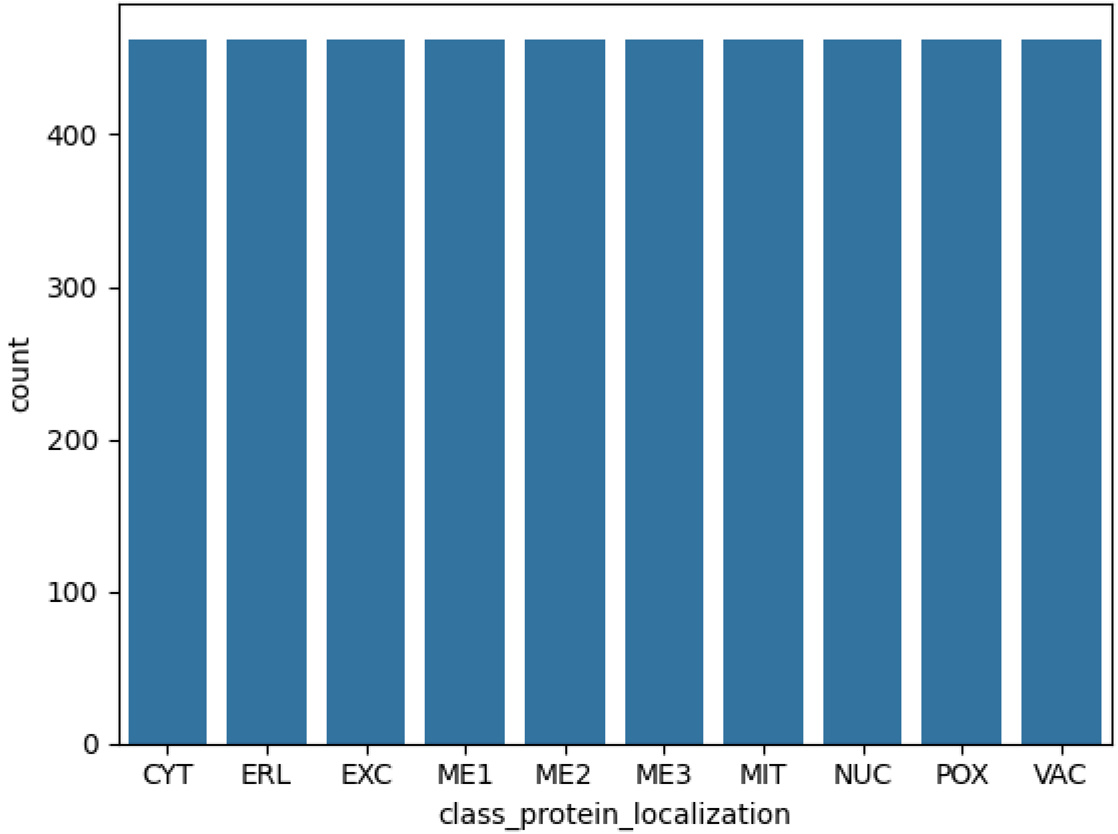
```
np.bincount(y_smote.cat.codes)
```

Out[153…

```
array([463, 463, 463, 463, 463, 463, 463, 463, 463, 463])
```

In [154…

```
sns.countplot(x=y_smote)
plt.show()
```


count
0
100
200
300
400
CYT
ERL
EXC
ME1
ME2
ME3
MIT
NUC
POX
VAC
class_protein_localization