

```
1
2 from yellowbrick.target import ClassBalance
3 import pandas as pd
4 from sklearn.model_selection import train_test_split
5
6 def trainTestSplit(dataFrame,
7                     test_size=0.2,
8                     random_state=765,
9                     stratifyColumn=None,
10                     shuffle=True):
11
12     origDataSize = len(dataFrame)
13     indent = '---> '
14     if stratifyColumn is None:
15         train, test = train_test_split(dataFrame,
16                                         test_size=
17                                         test_size,
18                                         random_state=
19                                         random_state,
20                                         shuffle=shuffle
21                                         )
22     else:
23         train, test = train_test_split(dataFrame,
24                                         test_size=
25                                         test_size,
26                                         random_state=
27                                         random_state,
28                                         stratify=
29                                         stratifyColumn,
30                                         shuffle=shuffle
31                                         )
32
33     print(f'Completed train/test split (test_size = {
34           test_size}):')
35     print(f'{indent}Original data size: {origDataSize}'
36           )
37     print(f'{indent}Training data size: {len(train)}')
38     print(f'{indent}Testing data size: {len(test)}')
```

```

32     if stratifyColumn is None:
33         print(f'{indent}Not stratified on any column')
34     else:
35         print(f'{indent}Stratified on column: {
stratifyColumn}')
36
37     return train, test
38
39
40 def classBalanceUndersample(dataFrame,
41                             columnName,
42                             alreadyBalanced=False):
43
44     #Display the initial state
45     tDf = dataFrame.copy()
46     visualizer = ClassBalance()
47     visualizer.fit(tDf[columnName])
48     visualizer.show()
49
50     if alreadyBalanced:
51         print("Classes already balanced")
52         return
53
54     # Not balanced, need to get some info to get size
to balance to
55     ttlColName = 'ttlCol'
56
57     # Find the sample size by finding which group/class
is smallest
58     tDfSize = tDf.groupby([columnName]).size().to_frame
(ttlColName).sort_values(by=ttlColName)
59     tDfSize.reset_index(inplace=True)
60     sample_size = pd.to_numeric(tDfSize[ttlColName][0])
61     sample_class = tDfSize[columnName][0]
62     print(f'Undersampling data to match min class: {str
(sample_class)} of size: {sample_size}')
63
64     # Do the sampling

```

```
65     tDf = tDf.groupby(columnName, group_keys=False).
        apply(lambda x: x.sample(sample_size))
66     tDf.reset_index(drop=True, inplace=True)
67
68     # Visualize
69     visualizer2 = ClassBalance()
70     visualizer2.fit(tDf[columnName])
71     visualizer2.show()
72
73     # Return the balance dataset
74     return tDf
75
76
77 def displayClassBalance(dataFrame,
78                         columnName,
79                         verbose=False,
80                         showRecords=5):
81     ttlColName = 'ttlCol'
82
83     visualizer = ClassBalance()
84     visualizer.fit(dataFrame[columnName]) # Fit the
85     data to the visualizer
86     visualizer.show() # Finalize and render the
87     figure
88
89     if verbose:
90         tDfSize = dataFrame.groupby([columnName]).size
91         ().to_frame(ttlColName).sort_values(by=ttlColName)
92         tDfSize.reset_index(inplace=True)
93         display(tDfSize.head(showRecords))
```