M5-L1 Problem 1

In this problem, you will implement a function to calculate gini impurity on an arbitrary input vector.

For reference, the formula for Gini impurity is:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{k} p_i^2$$

where D is the dataset containing samples from k classes and p_i is the probability of a data point belonging to class i.

Gini Impurity Function

Complete the function `gini(D)` below. It should take as input a 1-D array, where is the number of samples corresponding to each output class.

For example, consider the input array `D = np.array([4, 9, 7, 0, 3])` In this example, there are 5 input classes and 23 total samples. For this input, your function should return 0.707.

Your function should work regardless of the length of the input vector.

In [5]:
```
import numpy as np

def gini(D):
    # YOUR CODE GOES HERE
    return 1 - np.sum((D/23)**2)


D = np.array([4, 9, 7, 0, 3])
g = gini(D)
print(f"gini([4,9,7,0,3]) = {g:.3f} (should be about {0.707})")
```

```
gini([4,9,7,0,3]) = 0.707 (should be about 0.707)
```

Out[5]:
```
array([0.8, 1.8, 1.4, 0. , 0.6])
```

More test cases

Compute and print the gini impurity for `D1`, `D2`, `D3`, and `D4`, defined below:

In [6]:
```
D1 = np.array([1,0,0])
D2 = np.array([0,0,4])
D3 = np.array([0, 20, 0, 0, 0, 3])
```

```
D4 = np.array([6, 6, 6, 6])

for D in [D1, D2, D3, D4]:
    # YOUR CODE GOES HERE
    print(gini(D))
```

```
0.998109640831758
0.9697542533081286
0.22684310018903586
0.727788279773157
```

In [ ]: