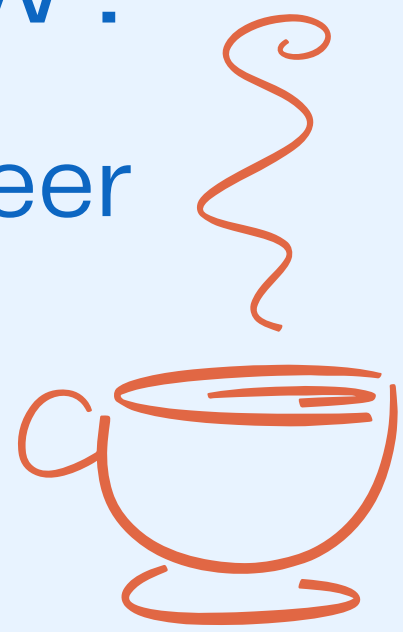
Vaishnavi's Chai & AI Series

"Is my dataset
big/balanced enough
for my image
classification use case?
How would I know?"

~ an anxious ML Engineer



Each label should have at least 1000 images

For example, if we have a dataset of fruits with classes like apples, bananas, oranges. Then, you should have at least 1000 images for each class.

Tip 1

Each primary feature of that label should have at least 10% of images of that label

For example, if you are classifying apples from oranges, the color of apples (red, green and yellow) will be it's primary feature. The dataset should've at least 100 images per color.

Tip 2

Each secondary feature of that label should have at least 5% of images of that label.

Secondary features are the features that are not always necessarily present in the class. In our case, an orange may or may not have a stem. So out of 1000 images for class orange, there should be at least 50 of oranges with stem and 50 without stem.

Tip 3

If you have a feature associated with passage of time, like shelf-life of an orange, you need at least 20% images covering it.

For example, out of 1000 images for class orange, there should be at least 200 that cover how a rotten, spoiled or insect damaged orange looks like.

Tip 4

Vaishnavi's
Chai & AI series

Join the
'chai'-n of the AI thought
& recieve some 'tea'-rific machine
learning tips for the everyday
conundrums a ML Engineer faces