Exploratory Data Analysis:

•First step in any machine learning project

•Analyze and investigate data sets and summarize their main characteristics

•You can google and find all information about your dataset but DON'T!

•**Dataset Overview**: Begin by getting a basic understanding of the dataset.

Example: Number of images, the resolution of images, the number of classes, and the distribution of images across these classes. Imbalanced classes or anything interesting?

•**Visual Inspection**:

Example: Manually inspect a subset of images from each class. This helps in identifying any apparent issues that could affect classification, such as inconsistent image orientations, varied lighting conditions, or irrelevant background noise.

•**Data Split Analysis**:

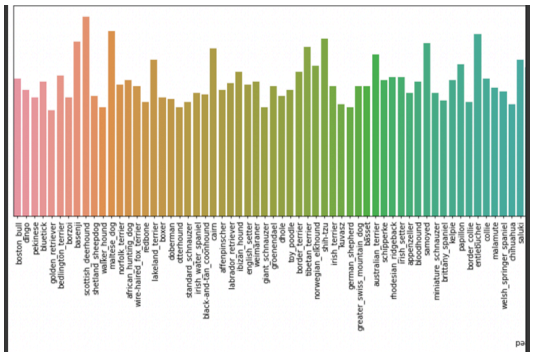
Check if train, validation, and test splits have similar distributions.

•**Pre-Processing:**

Example: Take the raw image and improve image data (also known as image features) by suppressing unwanted distortions, resizing and/or enhancing important features, making the data more suited to the model and improving performance.

Tasks (Deliverables):

1. Plot class distribution for the chosen dataset(s). If you are working with different datasets, compare the distributions. Attach your plots and code snippets.



2. Study different categories and plot sample images of some categories.

Attach sample images and code snippets.

3. Which categories are most confusing or hard to differentiate? Which categories are the easiest to differentiate?

Think how you can check qualitatively (visually) and quantitatively.

If you are working on NLP, some ideas for EDA:

Analyze the length of individual Reddit tifu posts and plot distributions across post length. This includes both the word count and character count. Is there any correlation between the upvotes, scores and length of posts?

Identify and analyze the most frequently occurring words and phrases in the TIFU stories, excluding common stopwords.