Machine Learning Assignment - 1

Instructions.

- This is a simple coding assignment mainly focus on the exploratory data analysis (EDA). This assignment will introduce you to type of data we deal with in machine learning and its challenges.
- Use **Google Colab** for this assignment. Only downloaded .ipynb files with naming convention as (***name-rollno.ipynb***) will be a valid submission. If not strictly followed will fetch 0 marks, and no further discussion will be done.
- You are only allowed to use *numpy, scipy, matplotlib, pandas, PIL(Pillow), librosa* and core python libraries.
- Here in the assignment randomly selected means you have to code for selecting elements random from the data structure, hard coded selection will fetch 0 marks for the question.

Dataset download link You can use this tutorial to load Kaggle data directly in Google Colab.

Image Data

1. ⟨ **2 Marks** ⟩ Find out whether the given dataset is imbalanced, if found plot a bar plot for the number of images per class vs classes and mention the imbalanced class, and suggest methods to balance the dataset.
2. ⟨ **2 Marks** ⟩ Randomly take 8 images from entire dataset and plot there respective histograms and label the class.
3. ⟨ **2 Marks** ⟩ Find and show mean and variance of each class. What can you deduce from these mean and variance of the data.
4. ⟨ **2 Marks** ⟩ Take 4 images from RANDOM class in dataset standardize them and plot before and after images, write your observation.
5. ⟨ **2.5 Marks** ⟩ Perform transformation on images (random rotation, random croping, random scale) and plot before and after images.

Audio Data

1. ⟨ **2 Marks** ⟩ Find mean audio length of each class and check for imbalance class if any, suggest methods to balance the dataset.
2. ⟨ **2 Marks** ⟩ Plot spectrogram of randomly selected 4 audios from complete dataset.
3. ⟨ **2 Marks** ⟩ Implement Pre-Emphasis filter from scratch and plot before and after time-domain plots of randomly selected 4 audios. Explain what does PreEmphasis filter do.
4. ⟨ **2.5 Marks** ⟩ Perform upsampling and downsampling of audio named *speech-librivox-0053.wav* in CLASS 1 and plot before and after spectrograms, write your observations.
5. ⟨ **3 Marks** ⟩ Save a randomly selected chunk of 2 sec audio form a randomly selected audio in *.mp3* and *.flac* format, load, compare all 3 spectrogram (*.wav, .flac, .mp3*) and write your observation. If audio is less than 2 sec repeat the same audio.

Text Data

1. ⟨ **2 Marks** ⟩ Find average text length of English and Hindi corpus (including whitespace, punctuations etc.)
2. ⟨ **2 Marks** ⟩ Randomly select 20 parallel texts and remove puctuations and special characters.
3. ⟨ **2 Marks** ⟩ Randomly select 50 parallel texts and make dictionary for both english and hindi corpus. (Here dictionary means a key-value mapping where key is word and value is a unique number, no two keys can have same number)
4. ⟨ **2 Marks** ⟩ Find 10 most occuring words and plot there histogram with labels.