Exploratory Data Analysis:

1) Total 57 Mocked Primary Care Consultations, with 7 clinicians and 57 actors posing as patients.
2) Total duration of approximately 8 hours and 38 Minutes.
3) Utterance level transcriptions with Start and end times, stored in TextGrid format.

Data Pre-Processing Needed :

1) Both Doctor and Patient files are given separately (one recorded on Laptop and one recorded on Phone).
    a) Need to combine the channels, tool used : SOX
2) Since Utterance-level transcriptions are given, the data is stored in TextGrid format along with start and end times. Two type of formats needed here:
    a) **.rttm** : For computing Diarization Error rate
    b) **.txt** : Which contains start time, end time and speaker label, needed for audacity analysis.
3) For Pyannote fine-tuning, we need the data structure according to it's format to predict the outcome and compute DER from the reference using : pyannote.metrics
    a) GitHub - pyannote/pyannote-metrics: A toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems
    b) This includes creating appropriate folder structure, creating train, dev and test set.
    c) Placing .rttm files in the expected format, and creating a .uem file for each of the audio, which denotes the time range in the audio file considered to compute the Diarization Error Rate.

TASK 1) : Zero-Shot Pre-trained Model Evaluation
Took two of the most popular models in the Industry currently,

1) NEMO (GitHub - NVIDIA/NeMo: NeMo: a framework for generative AI).
2) PYANNOTE (GitHub - pyannote/pyannote-audio: Neural building blocks for speaker diarization: speech activity detection, speaker change detection, overlapped speech detection, speaker embedding

	`pyannote`	`NeMo`
Voice Activity Detection (VAD)	Pyannet derived from Syncnet	MarbleNet
Audio embedding	ECAPA-TDNN	TitaNet
Clustering	Hidden Markov Model clustering	Multi-scale clustering (MSDD)

Both explored open source algorithms have a:

- Voice Activity Detection which differentiates between speech and non-speech regions.
- Audio Embedding, representing a speaker in the form of a vector representation.
- Clustering : to represent more similar embeddings as one speaker representation.

DER Computation done on 6 Core Intel i-7 CPU

Pre-trained Models	**Dataset (TEST-FULL)**	**NEMO**	**PYANNOTE**
Diarization Error Rate (DER)	Full Primok57 Dataset (Approximately 8 hours of Data).	28.49 %	26.23 %

TASK 1 Outcomes : PYANNOTE performs slightly better overall then NEMO.

TASK 2: Fine-tune PYANNOTE on a Domain Specific dataset.

PRIMOK57, which is approximately 8 hours of data, is splitted into Train, Test and Dev, with 80%, 10% and 10% Split respectively.

Models	**Dataset (TEST)**	**NEMO Pre-trained-1**	**PYANNOTE Pre-trained**	**PYANNOTE Fine-tuned-1**	**PYANNOTE Fine-tuned-2**
Diarization Error Rate (DER)	10% of Full Primok57 Dataset	31.02 %	23.16 %	8.3%	7.3%

PYANNOTE Fine-tuned-1 : Model with segmentation and clustering threshold optimized, fine-tuned on 20 Epochs
PYANNOTE Fine-tuned-2 : Model with segmentation and clustering threshold optimized, fine-tuned on 25 Epochs