DS203-2024-S1: Exercise – 7 (Project)

- **This Exercise (Project) carries 30 marks and 30% weightage**
- **It should preferably be done in groups**
  - **A group can have maximum 4 members**
  - **Team of '1' is strongly discouraged, but not barred.**
  - **Identify one member of the team as the Group's 'anchor'**
  - **All members HAVE to register themselves at:**
    - **https://tinyurl.com/2024-S1-E7-Groups**
- **Read the instructions and the evaluation criteria carefully.**
- **Note and adhere to the submission requirements and deadlines carefully.**
- **Submissions due by: November 7, 2024, 11:55pm**

Group Registration QR Code

Report and Submission Guidelines

1. Follow the Reporting and Presentation Guidelines contained later in this document
2. The submission procedure for source / data files will be communicated / updated here in due course of time

Evaluation Criteria

Evaluation Criteria	Marks
Results • How many problems have been correctly solved? • Has there been any creative thinking and innovation while solving the problems? • Quality of Feature Engineering / Feature Creation in terms of relevance to the problem	10
Process • Are the solutions relevant, correctly applied, and backed up with proper metrics / reasons / explanations? • Are the major steps of data analysis diligently followed and correctly applied and documented (wherever required / if applicable …)?	10
Documentation • Quality of presentation: Completeness and preciseness of the final slide deck; design and readability of the slides. Are **all the above aspects** covered in the presentation?	10
Penalty • **Does the presentation contain raw and hyped-up outputs generated using LLM tools like ChatGPT / Gemini etc.**	**-10**
• Viva will be conducted, if deemed necessary, to ascertain originality of work, and to ascertain contribution by team members. • The project will be evaluated **ONLY** by reviewing the presentation. Source code / Jupyter Notebook will **NOT** be referred to, to understand your work; they will be used to *verify* the claims you have made in the presentation. Therefore, **if you forget to mention some part of your work / analysis in your presentation, it will be concluded that you have not done it**! • If you do not submit your source and data files, your project becomes unverifiable and the submission will not be given any credit.	

Problem Background and Description

About the data

- MFCC coefficients have been extracted from 115 audio files (songs) and stored into CSV files. These files are named 'nn-MFCC.csv', where 'nn' is a song number. **The song details are not accessible to you**.
- Python code segment used to create MFCC using the module `librosa`:

```
# Compute MFCC coefficients for the segment
mfccs = librosa.feature.mfcc(y=segment, sr=44100, n_mfcc=20)
```

- The MFCC coefficients, 20 of them, have been created by using a sampling rate of 44100 Hz, with a default hop size of 512 samples.
    - This means about 86 sets of 20 MFCC coefficients are generated for every 1 second duration of the song.
    - For a song with 2 minutes duration its MFCC file will have 20 rows and about 10320 columns.
    - This provides sufficient details to 'reason' about various aspects of the song and build meaningful ML models.
- The multi-part ZIP file uploaded to Moodle, MFCC-files.zip + (MFCC-files.z01 to MFCC-files.z05), contain the CSV files. Total size of the zipped files is about 220 MB.

About the audio files (songs)

The following groups of files contribute to the set of 115

- Files containing rendition of the Indian National Anthem
- Files containing Marathi 'Bhav Geet' – sung by various artistes, male and female
- Files containing Marathi Lavni songs – sung by various artistes, predominantly female
- Hindi film songs sung by Asha Bhosale
- Hindi film songs sung by Kishor Kumar
- English songs by Michael Jackson

Problems to be solved

Mandatory Problems

1. Analyze MFCC files to organize the 115 files into groups broadly corresponding to those listed above
2. Identify at least 3 files containing the National Anthem
3. Identify at least 3 files (each) containing solo songs by Asha Bhosale, Kishor Kumar, and Michael Jackson

Note: In your final presentation have Table(s) that clearly list the audio file number(s) corresponding to each of the above questions.

Optional Problems

4. Classify the songs into i) Female singers only ii) Male singers only iii) Female and Male singers
5. Select an MFCC file and precisely identify 3 distinct vocal segments (start – end seconds into the song)
6. Classify songs using specific instruments, like percussion / string / woodwind (eg. flute). Or, Specifically detect if a song uses a flute.

Solution guidelines

- Doing justice to this project involves much work, including thoroughly researching into MFCC, its structure, capabilities, and limitations.
- Do what it takes to submit well researched, well-designed, and well-reasoned solutions to the above problems – and effectively communicate them!

- Re-visit all major aspects of Data Science that you have learned so far, and check if they can be / need to be used to solve the posed problems.
- Explore feature derivation and feature engineering to solve specific problems
- It will really help to work in a team, and divide work amongst the members – to do a good job.
- It is important to have an overall **solution design** in place.

Reporting and Presentation Guidelines

1. As often mentioned, succinctly communicating your work and results is a very important part of the Data Science process. One of your important submissions is the presentation – that summarizes your approach, work, results, achievements, learnings, and possibilities. Budget adequate time for this activity and design your presentation well; last minute work will invariably be shoddy.
2. **Include the names and roll numbers of all group members on the title slide. No credit will be given to members who are not mentioned on the title slide.**
3. Provide an executive overview (1-2 slides) at the start of the presentation.
4. DO NOT use verbose paragraphs, or storytelling, to explain your steps, observations, results, and recommendations. All these should be presented precisely and point-wise.
5. Summarize your observations and results using charts / Tables and precisely explain them and draw conclusions from them. Merely including charts and Tables is not enough. Slides should be well designed. Use as many slides as required to completely convey your work.
6. If you do not include something in your presentation, it will be deemed that you have not done it. **Source code / Notebooks will NOT be reviewed to *understand* your work**.
7. Towards the end of the presentation, include slides that clearly answer all the questions posed in the ***Evaluation Criteria*** table. In addition to focussing the evaluator's attention, this will also ensure that you have covered all the expected points in the presentation.
8. Finally, include a slide or two outlining your learnings from this project, and your experiences and hurdles while doing the project.

oooOOOooo