Data Science – Project Guidelines

Form a team of at most four members for your project. You must share your team members' names, email addresses, and roll numbers **by 21st Aug 2023** through this Google form. You will be assigned a TA who can advise you if you have any doubts.

A project will start with data collection followed by data preparation (encoding, handling outliers, missing values, dimensionality reduction, sampling etc.). Next, the project has two significant aspects: a group must perform at least one.

1. Get inferences from the data using statical analysis such as hypothesis testing. Know your tests and the reason for using them for your problem. Conduct a comprehensive comparison with other tests.
2. Train and validate some ML models (regression, classification, clustering, decision tree etc.) on the data set. Know your ML model and the reason for using it for your problem. If it is possible to apply other models, then compare your model comprehensively with others.

The guidelines for your data science projects are as follows:

1. If you have your data, then go to step 2; else, collect your data from the repository—preferably mixed data type of size at least 3M (instance x attribute).
2. Understand your data. For example, you should understand various features (columns) of the dataset, output, or the responses (if available). Explore existing analysis on the data. Formulate a problem statement.
3. Prepare your data (handle outliers, missing values, dimensionality reduction, etc.). You might need to apply sampling on the data or the features based on the data size.
4. Do preliminary analysis and (or) visualization on the dataset as much as needed and get a summarization of your inferences.
5. Do A or (and) B. Groups doing either A or B can fetch a maximum of 80% of the total marks.
6. Write a report in Google Docs detailing where you have collected the dataset, how many rows and columns there are, and what the features are. How did you prepare your data, and why do you follow such and such steps? What are the changes you got after data preparation? What are the existing analysis on the data? What are your basic questions about your data, and how are you going to address those questions through an analysis? The report must have plots, charts, bar graphs, pie charts, and so on, and submit it. Report your inference methods / ML algorithm done for analysis and what is the purpose of doing. Based on your problem statement, what are your inferences from the output? What aspects could not be addressed by your analysis and so on? The report should contain all the codes written by you, a readme file.
7. There will be at most three presentations. In the first presentation, explain your dataset, work done on the dataset, your problem statement, and its motivation in 5 mins; report submission due date is **01st Sep 2023**. In the second presentation, state the problem statement, how you prepared your data, and present your progress in 7 mins; the report submission due date is **09th Oct 2023**. In the third presentation, state the problem statement and show your result and inference in 10 mins; the report submission due date is **24th Nov 2023**. Submissions must be on time to avoid penalties; they will not be evaluated if submitted too late (more than one day).
8. Generally, you will be graded based on how much you could dig into and infer from the data. Do not copy-paste the codes from any Internet source or your friends. Any means of dishonesty will result in complete rejection.

All the best!