MAJOR PROJECT - BITS F464 - Fall 2020

We all love to follow a sport, be it indoors or outdoor. What if we could predict who would win a game? Well, this is a domain where many ML engineers and researchers are currently working. Many betting platforms like betway and many fantasy team apps like Dream11 have a lot of data analytics behind their algorithms that predict the odds of winning for a particular team.

Task description:

Your task is to predict a winner for a specific "game." One sport from the given five sports will be allotted to you -

1. Cricket
2. Football
3. Basketball
4. Volleyball
5. Kabaddi

You are responsible for every aspect of this problem, from data collection to metrics selection. We are providing you a list of subtasks(/checklist) that you can check out as you move through this process.

Sub-Tasks -

- Collect data for the sport which is allotted to you. It can be a league as well. For example, you can collect data of teams playing in the UCL League/NBA/IPL.
- Select the features that you feel are relevant and/or engineer new features and apply pre-processing of your choice. An example of a feature could be the form of a team, i.e., how well the team performed in the last five games.
- Select appropriate models for this task and test their performance on the data you have collected. Feel free to use any model taught in the course.
- Tabulate the model's performance on the metrics that you feel are appropriate for the task and data.

Marking Criteria -

- The source of the data - Data collection is a very important component of any ML problem. We hence value the effort that goes into collecting quality and relevant data for the allotted task. This means that if you put effort into collecting data from true sources of the sporting events, your hard work would be valued over others who pick data directly from sources like Kaggle, etc.
- The originality and creativity of the features that you select/generate - Another important aspect of ML is feature selection. You are required to use at least five unique features for the allotted sport (Unique here means that you simply cannot take total scores for five different players and classify them as five features. For example, in CS:Go, five unique

features can be the number of kills, number of deaths, number of assists, number of no scopes, number of Deagle shots).

- Models selected - You are required to experiment with the appropriate choices of models for this task. A part of the score would be dependent on whether the models you chose were appropriate/good for the task, and the justification you provide for the same. You are required to analyze at least two unique models (For example - two neural networks with different architectures are not unique, but a NN and DT are two unique models).
- Metrics selected - You would be judged on the choice of metrics too. Specifically, whether the metrics are suitable for the task or not (**and the scores you achieve on these metrics**).

Resources -

Web Scraping :

- https://www.dataquest.io/blog/web-scraping-tutorial-python/
- https://realpython.com/python-web-scraping-practical-introduction/
- https://www.youtube.com/watch?v=aIPqt-OdmS0
- If the data is in table format on the webpage, you can use pandas as well to scrape the data (for example, I used pandas to scrape your public and private leaderboards from Kaggle). For this, first, download the HTML locally and then open it in pandas using pd.read_html().

Some sources to collect the data from (not exhaustive feel free to use any data source of your choice) -

- https://www.uefa.com/uefachampionsleague/statistics/
- https://www.iplt20.com/stats/2020/most-runs
- https://www.cricketworldcup.com/tournament-stats
- https://www.prokabaddi.com/teams/u-mumba-profile-5
- https://in.global.nba.com/statistics/?_ga=2.178542752.589391114.1603182770-745267402.1603182770
- https://www.volleyball.world/en/volleyball/worldcup/2019/women/womenstatistics

Submission Format

One member of the group has to submit a zip file. The name of the zip file should be grp_number.zip. For example, group number 10's zip file would be "10.zip" without the quotes.

Each zip file should have the following dir structure -

```
grp_number.zip
  |
  | --- src
  |        | --- file1.py
  |        | --- file2.py
  |
  | --- data
  |        | --- file1.csv
  |        | --- file2.csv
  |
  | --- report.pdf
  |
  | --- contributions.txt
  |
  | --- requirements.txt
  |
  | --- README.txt
  |
```

Description of the file structure -

- src - This folder should contain all the code files that you have used for the project. Remember to use python files and not notebooks (you can convert your notebooks to .py files by saving/downloading them as python files)
- data - This folder should contain all the data files (CSV files, etc.).
- report.pdf - The final report of your project. The format of the report should be similar to any research paper with the following sections -
    - Abstract, Introduction, Related works, Dataset, Methodology, Results, Conclusion, and References. (Recommended Sections again, but your report should answer all the tasks properly)
    - It is recommended (not compulsory) to write the paper in Latex so that you can upload it later on arxiv (Note - marks won't be deducted for the report not being in Latex)
- contributions.txt - This file should contain all the members' contributions, i.e., what part was worked upon by whom.
- requirements.txt - This file should contain all the dependencies of your code.
- README.txt - This file should provide a step-by-step guide to run your code.

Extras -

- FAQ document - https://docs.google.com/document/d/1lzs5V4UdkVXplFdkE_QGMHxC8m52luE1OS30vzDZsRQ/edit?usp=sharing

Currently, this document is empty. If you guys have any general doubts (not specific to your project, e.g., Some doubts regarding scraping, etc., or doubts regarding submission format, etc.), put them up in this doc; TAs will answer them periodically.

- Additionally to this FAQ document, we will be giving a 10 min slot to each group to discuss doubts specific to their approach to the problem (although we won't be telling if your approach is correct or not).