PES University, Bengaluru
UE20CS312 - Data Analytics

Session: Aug – Dec 2022
Project Guidelines and Submissions
Expected deliverables, timeline and evaluation criteria

Phase 1:

(a) [Week 1] Formation of a team (recommended team size 2-3) and selecting a team identity

(b) [Week 2] Dataset selection and problem statement (team registration **due by 7:00pm IST August 26, 2022**)

- Potential sources of data:
  - KDNuggets: http://www.kdnuggets.com/datasets/index.html
  - Government data: https://data.gov.in/
  - Competitions for social good: https://www.drivendata.org/competitions/
  - Kaggle: https://www.kaggle.com/datasets
- Note:
  - It is strongly recommended that you work on publicly available data, so you do not spend too much time 'collecting data' as a part of the course project
  - To those working on time series or text or images/ video, GIS or multimodal data or any domain specific data, ensure you have enough of a background to be able to complete the project in time

(c) **[Week 3] Setting up of Github Accounts + EDA and Visualization**

- How many rows and attributes?
- How many missing data and outliers?
- Any inconsistent, incomplete, duplicate or incorrect data?
- Are the variables correlated to each other?
- Are any of the preprocessing techniques needed: dimensionality reduction, range transformation, standardization, etc.?
- Does PCA help visualize the data? Do we get any insights from histograms/ bar charts/ line plots, etc.?

(d) [Week 4] ISA 1

(e) [Weeks 5] Literature review + initial solution approach:

- Look for papers on Google Scholar and other online sources to answer the following questions:
  - What have others done to solve this problem? What other approaches can we explore on this data set?
    Or
  - How have others solved a similar problem? Can we apply any of those solution strategies to the problem we have selected?
  - Exception: If you are working on a problem for which there is no ready precedent, but know the kind of approaches you want to use, then look for papers that talk of those approaches.
- Refine your problem statement
  - What is the specific problem we are going to solve?
  - What are the questions we are going to attempt to answer?
  - What are the challenges with this data set (based on the initial exploratory analysis + coarse solution approach (trying library functions, etc., to build a simple model)

  - What solution approaches would be reasonable to attempt?
  - How is my solution approach different from what is already out there?
  - What is the use of solving this problem?
- Write a literature survey report

(f) [Weeks 6] Complete/ submit literature review report (with EDA + visualization) + Github 'link' due by **7:00pm on September 16, 2022**

(g) [Week 7] ISA 2

Phase 2:

(a) [Week 8] Model design and testing (midterm report (3-4 lines of update) on the results of the initial models, inferences, next steps planned) **7:00pm on September 30, 2022**

(b) [Week 9] Continue to test model/ refine model parameters
Run cross validation tests and make a note of the results; how can we do better?

(c) [Week 10] ISA 3

(d) [Week 11] Prepare a ppt and complete peer review – ppt for peer review and peer review report due by **7:00pm on October 21, 2022**

(e) [Week 12] Run comparisons, test any other models that need to be tested; answer peer review questions, prepare final report

(f) [Week 13] ISA 4

(g) [Week 14] Wrap up the model building/ testing and complete the final report with interpretation of results; comment the code, record the 5 min video presentation **(final) report + documented code/ (sample) data to reproduce the results with readme + 5 min video presentation (recording) due by 7:00pm on November 11, 2022**

(h) [Week 15] Prepare for ISA 5

(i) [Week 16] ISA 5

(j) **[Weeks 17-19]** Prepare for ESA and do well on the exams

*****

Formats and suggested content

Both the literature review and final report must be in **2 column IEEE format**
Templates are available here (doc) and here (LaTeX); for bibliography use Paperpile with GoogleDocs or Mendeley with MS Word and BiBTeX with LaTeX (BiBTeX is available with Overleaf).
.

Phase 1 report/ Literature survey

2-3 page report in 2 column IEEE Conf format
[~0.5-0.75 page] Introduction to the context of the problem (why is it important?)

[~1 page] What have others done to solve it – paraphrase and critique others' approach and cite the work
(a) assumptions made, if any
(b) approach used - a summary
(c) summary of the results reported
(d) any limitations reported?
(e) any lacuna in their approach/ evaluation that you inferred?
[~0.5 page] Proposed problem statement with the specific issue you intend to address
[~0.5-1 page] How is your approach (or the type of problem you are looking at) different from what has already been done? (or if you are attempting to improve upon someone else's work, explain in what way it distinguishes itself from what has been reported)

Final report

4-5 page report in 2 column IEEE Conf format
[0.5-1 page] Introduction and background – what is the problem area? Why is it important? What is the specific problem you seek to solve?
[0.5 -1 page] Previous work – A brief review of only the most relevant predecessor work; what limitations have you identified that you seek to address in your work? What are the assumptions you have made about the data/ problem area or the scope of the problem you seek to solve?
[1.5 - 2 pages] Proposed solution – an overview of the various components of your solution (preprocessing + building a model + evaluation)
This is to be followed by a detailed explanation of each component and what you have
[1.5-2 pages] Experimental results and a detailed explanation of all the insights you have gained into the data (on what cases does the model work well? When does it fail?)
[0.5 page] Conclusions
[Not included in page count] Contribution of each team member + References + [optional] Anything interesting you would like us to know (either some interesting technical find or about the problem domain or just the experience of working on the project, etc.) Also, an Appendix with any further visualizations or tables of comparison not included in the main paper.

Plagiarism check (recommended): Please submit your report to librarian@pes.edu marking a copy to your course instructor; upload the similarity report with your final report (you can refine your report to reduce the similarity if you complete this check well in time).

Policy on plagiarism and project submission: The similarity report is recommended to be within 15%. A similarity of score >= 40% and/ or resubmission of work from any other team in the past or current batch will result in 0 marks being awarded for the project component. Failing to submit any project component will result in no marks awarded for that component.

Video

[1 min] What problem have you selected and what data set are you using to solve the problem?
[1 min] Why is what you have done important/ useful?
[1 min] What is the approach you have taken?
[1 min] How did you evaluate your solution/ the algorithm you implemented?
[1 min] Anything interesting that you inferred about the data or learnt through the process? Also, the specific role of each member of the team.
+ 1 min buffer – anything beyond 6 minutes will not be evaluated.

Evaluation criterion for final submission components

1. Problem statement

- Design criterion, choice of assumptions, constraints, novelty of application, understanding of the data used/ problem domain

2. Technical content
   - Difficulty level/ time and effort invested
   - How much support was available?
   - Any improvements over existing approaches/ solutions or an attempt to solve a new problem?
   - Design of experiments and interpretation of results
   - Readability of the code + reproducibility of results
3. Correctness + Completion
   a. Have all components been submitted? Were any links broken and required follow-up?
   b. Any obvious errors in the assumptions or application of model, etc.
   c. Extent of completion
   d. Quality of inferences and analysis of the results
4. Presentation (both reports + video and code + data)
   - Score on plagiarism check on both reports (15% or less acceptable)
   - Clarity (report and video)
   - Aesthetics of the presentation
   - Cohesion as a team and contribution of each member
5. Timeliness of the submission of every component (team formation + literature review (with data source and Github link) + final report + code + data (not requite separately if files are on Github and made accessible to us) + 5 min video presentation

   Note: All submissions can be made prior to the deadline; forms will be released a week or two before they are actually due.

*****

Submission components and deadlines (time due on/before 7:00pm IST)

Week	Task	Marks	Due
1	Register team (members + team identity)	0.5	**19/8**
2-3	Register dataset + problem		**26/8**
3-6	Literature review with basic EDA + link to the Git repo summarizing relevant work/ models explored in the past	1	**16/9**
4	Midterm report through a Google form summarizing solution approach and progress on the project	0.5	**30/9**
5	Ppt + demo for peer review + (completion of peer review) + submission of peer review report	2	**21/10**
6	Submission of final report + answer to questions from peer review + code to justify/ reproduce results	4	**11/11**
7	5 minute video presentation of the work (including a demo)	2	

Worksheets (averaged for each Unit; 10 marks per Unit, best 4 out of 5; scaled to 8)

Unit	LH (lab hours) as per course information – recommended deadline)				Final due (Hard deadline)
1	LH 2 – 11/8	LH 5 – 17/8	LH 10 – 24/8	LH 13 25/8	**27/8**
2	LH 19 - 7/9	LH 25 – 13/9	LH 28 – 15/9		**17/9**
3	LH 35 – 28/9	LH 40 – 6/10			**8/10**
Kaggle	**15/10**				**15/10**
4	LH 52 – 20/10	LH 58 – 27/10			**5/11**
5	LH 69 – 17/10				**17/11**