# PES University, Bengaluru
## UE18CS312 - Data Analytics

**Session: Aug – Dec 2020**
**Project Guidelines**

**Phase 1:**

**(a) [Weeks 1-3] Formation of a team and selecting a team identity**

(b) **[Week 4] Dataset Selection and Problem Statement:**

- 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

- Rather than start with a problem statement and then look for data, you could consider finding a data set and then looking at what other inferences can you draw from this data other than the problem it was meant for/ challenge posted online.
- Support from student mentors
    - [WoW Series] Lecture 1 on Design Thinking by Ms. Richa
- Exceptions
    - To those planning on 'collecting data' – ensure you plan this suitably, so you have enough time to actually work on the data and try multiple approaches and
    - to those working on time series or text or images/ video or multimodal data, ensure you have enough of a background to be able to complete the project in time.

      Both categories of projects: for EDA + Visualization, you can practice on the worksheets or any data set you is amenable to this

**(c) [Week 5] Setting up 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.?
- Support from student mentors
    - [WoW Series] Lecture 2 on Kaggle and Google Colab by Ms. Bharani U. Kempaiah, Mr. Ruben John and Ms. Bhavya Charan

- [WoW Series] Lecture 3: Why Github, how do we go about this, what can we do with this? By Mr. Mayank Agarwal and Mr. Tanay Gangey
- [WoW Series] Lecture 4: Preparing R markdown/ Jupyter Notebook files – the what, why and how by Ms. Mainaki Saraf

**(d) [Weeks 6-7] 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
- Student mentor support
  - [WoW Series] Lecture 5: What is a literature survey report? How do we go about writing this? What is plagiarism and how do we avoid this ? by Ms Greeshma Karanth and Ms. Diya Sateesh
  - [WoW Series] Lecture 6: What is Overleaf? How can we use this to create a professional-looking report? By Ms. Bharani U. Kempaiah, Mr. Ruben John and Ms. Bhavya Charan

**(e) [Weeks 8-9] Refine the literature survey report, continue to experiment with data**
Prepare for/ write ISA 1

(f) **[Week 10] Phase 1 report (literature survey) + Github 'link' due by 7:00pm on October 12, 2020** (you are welcome to submit this earlier – a form will be sent out in Week 8).

**Phase 2:**

**(a) [Weeks 10-11]** Model design and testing
- Design model/ refine model parameters
- Run cross validation tests and make a note of the results; how can we do better?

**(b) [Week 12]** Run comparisons, test any other models that need to be tested

**(c) [Week 13-14]** Prepare for/ write ISA 2

**(d) [Week 15]** Wrap up the model building/ testing and work on presenting and interpreting results

**(e) [Week 16]** Comment the code, record the video presentation, complete writing the final report

**(f) [Week 17] Phase 2 (final) report + documented code/ (sample) data to reproduce the results with readme + 5 min video presentation (recording) due by 7:00pm on November 30, 2020**

*****

**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**

4-5 page report in 2 column IEEE Conf format

[~1 page] Introduction to the context of the problem (why is it important?)

[~1 page] What have others done to solve it - 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?

[~0.5-1 page] Proposed problem statement with the specific issue you intend to address

[~1-2 pages] 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.

## 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

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.