a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

DATA 542 Project (Group work)

Team size: 2-3 students. We will form the groups.

Dataset:

Become familiar with the AIDev dataset here. The link includes link to the dataset, initial paper releasing the dataset, and also open questions that you can explore about the data.

The project is open-ended. You can select your own research questions (RQ) or choose the ones from the above link and explore them using the dataset. The RQs are what you are exploring about the dataset, towards a specific goal.

Your tasks

1) Select or define three research questions. **The questions require combining at least two of the features from the dataset and should be comprehensive, so you cannot e.g., use the dataset statistics or apply a simple group operation and respond to an RQ.**
2) Define a methodology to answer the questions and implement them. In other words, what are you analyzing and how do you apply your analysis to answer those questions. Release your code on GH, and provide any instructions required to access them. The code should be well written and documented.
3) Write a short report (5-7 pages) about the methodology to answer each research question, as well as the obtained results and interpretations. Use ACM journal template to write your reports. Use \documentclass[acmsmall,screen,review]{acmart} to adjust the font, and remove the preamble about the journal. You can open the template and share among your teammates with a free Overleaf account.

Code reuse

Should you decide to reuse code (written by you or others or GenAI) or the papers published on this dataset:

- indicate its copyright and how your usage does not violate it,
- give proper credits,

- how did you use the published papers, e.g., their code is used, their RQs is adopted, the results is compared with them.

Usage of GenAI: The use of GenAI is discouraged. However, if you still decide to use online sources or GenAI to generate your code, first, make sure it is correct and executable. Second, make sure to include the resources you have used and/or mention that you used GenAI in your report. There is no penalty of using available code, the only drawback can be that you might not be engaged as writing the code yourself for your future references. You have to clearly state how and in what capacity you have used GenAI and the models you used.

How to submit: Submit your PDF file to Canvas, which should also include the links to your GH repository and the dataset you used for testing. Use the Association for Computing Machinery (ACM) - Small Standard Format Template (Overleaf) for your report. Remove the ACM, journal, and copyright information. The report should be between 5-7 pages, including plots and references.

- Clearly state the role of each team member in your reports.

GH repository:

The repository should include readme file explaining the project you are working, the structure of the repo and execution instructions. The repository should be clean and at a professional level. See the AIDev GH repo for an example of the expectations.

Deadlines:

Milestone	Description	Deadline	What to submit	Grade
Milestone 1	Meet your group, explore the dataset and define RQs and methodology	**Friday** Nov 21th **8 PM**	Submit max 1 page report about the RQs and how to answer them	5
Milestone 2	Complete half of the project	**Friday** Dec 5th **8 PM**	Submit maximum 4 pages report , including link to your GH report	10
Milestone 3	Complete the project	**Thursday** Dec 11th **8 PM**	Submit full report as described above	15
Total				30

a place of mind
THE UNIVERSITY OF BRITISH COLUMBIA

Grading rubric

	Weights	Subtotals
Report		40
Research questions and methodology	10	
Obtained results	10	
Interpretation of the results	10	
Concise and clear writing	10	
Code static		20
Follow conventions	10	
Comments	10	
Code execution		20
Syntax-error free, runs	10	
Takes reasonable time to run (so if you need to optimize your code or for loops, do so)	10	
Adhering the project requirements		20
Using complex analysis	10	
Code/ideas reuse statements	10	
Total	100	100