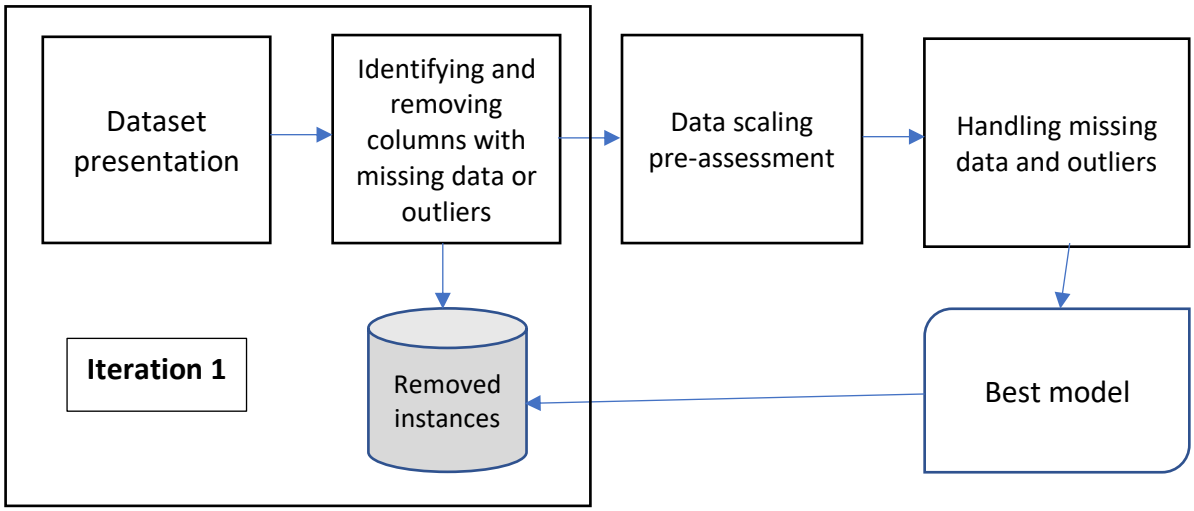
COMP 3400–6981 Data Preprocessing Techniques

Project - Iteration 2 (P2)

Fall 2022

This deliverable is due on the December 2nd, 2022 at 10:00 PM Newfoundland time. No submissions done outside D2L will be marked (e.g., email). Please organize yourself with your team to submit the document on time. Grade deductions for late submissions will be applied, check the syllabus for the detailed grade decrease.

Your deliverable for the first iteration is going to be a **Jupyter Notebook** with several cells that must describe and discuss with your code commented. **Properly commenting your code will grant you 10 marks in the evaluation (This is not a bonus; this is part of your assessment!).** Make sure you discuss your function's purpose, attributes and how it would behave. That's essential for your growth as a data scientist. Record a 10-15 minute presentation of the work you designed detailing your main findings and plots. **The quality of your video recording will grant you 15 marks**. The following figure provides an overview of the tasks that must be performed in this iteration (Parts 1 to 3).



Part 1 – Addressing instructor's comments from Iteration 1 (10 marks)

The team must prepare a report and submit a PDF answering how comments made in iteration 1 were addressed in your second submission. The document must contain the list of comments made by the instructor and the respective answer with the decisions made by the team. If no comment was made regarding your first iteration, you will start this iteration with 10 marks already (Good for you!).

Part 2 – Data scaling pre-assessment (27 marks)

The second part of your project will consist of exploring data scaling techniques. The following tasks must be performed in this part:

- Explore all data scaling techniques presented in the course and decide which one of them will be removed or kept in your experimental design. You must create (or refer to previous) plots and/or statistical indicators to justify your decisions.
- As a way to condense your findings, you may plot the results, side by side, of the original and the data scaling techniques used. Make sure you discuss the figure in-depth enough to justify your decisions.

Part 3 – Handling missing data and outliers (38 marks)

The third part of your project must advance your pre-processing strategy to handle the missing data and outliers selected in the Iteration 1 (remember that here you have at least two variables for each task, if you have more than that you must handle them as well). Besides, the techniques used in this part must also be combined with the techniques used in Part 2 of this iteration. To receive all marks in this part, you must:

- Decide a baseline strategy that will be used as a starting experiment where the use of all aforementioned strategies are likely to improve. A baseline strategy could be something as simple as a centrality measure (e.g., mean, median or mode) to replace missing values or upper and lower quartiles to replace outliers.
- You must create models using the two algorithms for classification or regression discussed in class (e.g., Linear or KNN). You should engineer an experiment where you will be able to decide the best combination **Data scaling technique + (Classifier, Regressor)** that will lead to a potential best result. Therefore, to assess results, you must rely on data that you know the ground truth. Remember that producing a single experimental result might not give you confidence to make a decision. You must discuss why you chose the combination. Don't forget to justify which evaluation metric you will use to make such a decision. The quality of your code will be essential to receive full marks in this experiment. Make sure you optimize your code by generalizing your problem using functions to put your data in a proper pipeline of analysis.
- Finally, fill the missing data and replace the outliers using the decided (Classifier, Regressor + Data scaling technique). Make sure the rows are properly scaled before you forecast the values.

Part 4 – Supervised Learning problem design and experimentation **(25 marks)**
(Graduate students only) (10 extra marks to undergrads to be added in Iteration 1)

In this last part of your project, you must design an experiment that can be framed as a supervised learning task. Using your data set that was fixed in the former tasks, you must determine a goal to be achieved (a label or value to be predicted) and contextualize why it is worth it to forecast this attribute. At this task, your project must:

- Decide a baseline model that could be improved. A baseline model may be the model trained with all attributes available. Make sure all other pre-processing decisions discussed in the former parts are considered in your experimental design.
- After, you must use any pre-processing strategy available to you that may enhance the quality of your results. Those strategies may include engineering novel features for your data set, using data discretization, feature selection, dimensionality reduction, etc. It is up to your team to decide and explore these pre-processing strategies and find a way to improve your results.

Grading scheme for Part 4

The grades for this part will be ranked based on the quality of the presentation (e.g., quality of the code, legibility, and documentation), creativity, and demonstrated expertise in using one or more of the several strategies listed above. Submissions may be ranked with the same value (position). After ranking all submissions, the grades will be assigned as follows:

Top 25% submissions – The submissions at this portion will receive all marks (25 marks).
Between 50% and 75% - The submissions at this portion will receive 18 marks.
Between 25% and 50% - The submissions at this portion will receive 10 marks.
Below 25% - The submissions at this portion will receive 5 marks.