Intuition

When thinking about experiments like those in Step 3, it helps to think in terms of information. The information is already present in your data, if you have more data, you have more information, which means you can potentially find a better model. In Step 3, you perform dimensionality reduction (DR) and then clustering. In each case, you aren't adding new information to your data; instead, you're transforming or reorganizing what's already there. So in theory, these steps alone shouldn't magically create a better model. However, DR and clustering can sometimes act as filters, processes that remove noise or unwanted variation from the data. If your data is noisy, these transformations can help highlight the underlying patterns more clearly and lead to improved models. Make sure you understand how DR or clustering can serve as useful filters!

Datasets for Fall 2025

We require the students to use these two specific datasets. If they are not explicitly using these datasets, they will receive a zero. This should hopefully not be the case but please make a note on the sheet in the comments. For further information on these datasets, please see the SL Report description or the Kaggle repos.

- **Hotel Booking Demand**: Kaggle Repository: *Hotel Booking Demand*
- **US Accidents**: Kaggle Repository: *US Accidents*

Rubric

Total Points: 100 + 5 Extra Credit

As a reminder, please give all or nothing for points. If a section has 3 points for a section, you need to either give 3 or 0 points. This will help us with ambiguity between graders and nuance.

Structured Feedback – CS7641 Fall 2025 Report

Continuing this semester, I want to put the point values by the section feedback. I have an example document I would like everyone to use as a template. You will still need to add individualized feedback for each student. Please reach out if you have any questions!

Requirements (10 points)

- **7 points total:** This is the DOCSTRING-GTusername file on Canvas. All requirements met, full reproducibility ensured.
  - **2 points:** Overleaf and LaTeX used for report. Please double check the report exists.
  - **2 points:** Github Hash. Please double check that this hash works.
  - **2 points:** Instructions to run code. The can be anything but needs to be instructions to help with packages or init to run the code. If the instructions are on Github and not the DOCSTRING, please remove these 2 points.
  - **1 point:** The document is a PDF.
- **1 point:** Report is 8 pages or less. If there is anything more, please make a note to the student for future reports.
- **1 points:** Graphs and text are legible. This means you do not need to do any zooming while on Canvas at 100%.
  - This point can be hard. Many times the student will have some cropped graphs that are hard to read at 100% zoom on a computer. If there are 3 or more graphs that are hard to read the axis or key/legend, please deduct this point.

- **1 points:** Proper References. This is a check to make sure their reference are consistent in formatting. APA, MLA, or IEEE are fine, but need to stay consistent. IEEE does have a website url in the description. This point should be taken off if the student includes something outside the format chosen or text that should be included in the main body.
  - Additionally, if the student does not have a citation outside of the course material or citing the environment, you should take off a point. The rationale here is that the student should be looking for connections to the literature to support their results.

Hypothesis (8 pts)

- **4 points:** Initial Hypothesis
  1. (+2 pts) For stating an explicit hypothesis for what the student expects in the report. This can be a wide range of potential hypotheses, e.g. the student might comment on how the labels in their data will perform better than clustering labels for classification with the NN. The report does not need to have a hypothesis for each section, rather these sections should help develop an overall narrative to keep the paper focused.
  2. (+2 pts) Explicitly stating evidence to support the claim of the hypothesis from a lecture, paper, or generally known theory. The hypothesis must be grounded in evidence before exploration.
- **4 points:** Follow Through with Hypothesis
  1. (+2 pts) Following up with the hypothesis in their discussion or conclusion. It is not enough to just state a hypothesis but follow through from a narrative perspective needs to be met at the end of the report.
  2. (+2 pts) Providing direct evidence from their report to support if the hypothesis has been met or not. This can be short, however there needs to be something explicit to help better summarize.

Step 1: Clustering for Two Datasets (20 pts)

- **10 points:** For each of EM and K-Means Clustering:
  - (+4 pts) How did you choose hyperparameters for each dataset (i.e. the number of clusters)? Two points per dataset.
  - (+2 pts) Visualization or demonstration of the results.
  - (+4 pts) What are the results? Are they meaningful? Descriptive shape of the clusters. This is where a limitation would be addressed. If they student does not mention any connotations, they should be not awarded these three points. There should be different descriptions and rationale for each dataset. If there are the same number of clusters for both datasets, only award 2/4 points.

Step 2: Dimensionality Reduction for Two Datasets (24 pts)

This might be the most involved section for reach report. You might want to double check each of the algorithms.

- **8 points:** For each of RP, PCA, and ICA:
  - (+4 pts) How did they choose the number of components? Supporting results required: Eigenvalues or explained variance for PCA, kurtosis for ICA, reconstruction error for RP.
  - (+2 pts) Visualization or demonstration of the results. This could be in a pair-plot, mapping variance, a table of descriptor values. There needs to be evidence from their experiemntation.
  - (+2 pts) What are the results? Are they meaningful? Transformed features are linear combinations of the original features. Students should analyze and interpret them. Examples include discussing loading scores, assessing the structure of the data in the new feature space, projecting data onto components to visualize patterns, etc. This is where a limitation would be addressed. If they student does not mention any connotations, they should be not awarded these two points.

Step 3: Dimensionality Reduction, then Clustering for Two Datasets (14 pts)

- There are 12 possible combinations of experiments/results. The goal of this section is for the student to demonstrate results from their experiments as well as reason on the outcome. This should be a combination of graphs and, hopefully, tables to better allocate space.
- **7 points:** For each dataset:
  - (+4 pts) Visual demonstration of performance of combination of methods. They should reselect the optimal number of clusters after DR to receive full credit. A table works best here. If there are 12 distinct plots, don't take off points but suggest the student combine graphs to save space.
  - (+3 pts) Comparative analysis between methods. Did PCA or ICA assumptions align better with the data? Did dimensionality reduction help filter noise or improve clustering results? Why or why not?

Step 4: Dimensionality Reduction on Neural Network for One Dataset (12 pts)

- **12 points (4 each):** For each linear method (RP, PCA, ICA) consider the following.
  - (+2 pts) Was there a change in performance? You should be looking for the supporting results to help with their claims.
  - (+2 pts) Was there a change in wall clock time / speed? If the student uses a different metric to judge difference, that will suffice for these points. There needs to be an alternative way to judge, measure, evaluate the results.

Step 5: Clustering on Neural Network for One Dataset (12 pts)

- **12 points (6 each):** For each cluster method (EM and K-Means):
  - (+3 pts) Was there a change in performance? Supporting results should be details for analysis. This is where the student might mention that there is a filtering effect occuring.
  - (+3 pts) Was there a change in wall clock time / speed? If the student uses a different metric to judge difference, that will suffice for these points.

Extra Credit: Non-linear Manifold Learning Algorithm (Up to 5 points)

- **5 points:** Non-linear Manifold Learning Algorithm explained for use and adds significant insights into the datasets and theoretical implications. This will include both a visualization and commentary in their analysis that might highlight clusters or grouping from the DR/clustering results.
- **3 points:** Non-linear Manifold Learning Algorithm explained for use with a visualization but only superficial insights. This might include a comment about visual inspection instead of investigating a class imbalance or how they compare to results found in main report. There must be a visualization to earn points.
- **0 points:** If the student mentions there are Non-Linear Methods and the methods might perform better than current experiments in theory. This is potentially good for their discussion but this is not enough to earn EC. An explicit visualization is needed to earn any EC.

Deductions

Formal writing requirement. Bullet lists (`itemize`, `enumerate`, or `description`) signal draft-style notes rather than polished prose. If *any* section contains a list with more than four or more items, apply a **50 point deduction to the overall report score** (single global penalty, not per occurrence).