Student ID: ______________________ NetID: ______________________

AE 03: Use your data

Discussion questions

1. For each predictive problem, decide whether simple random sampling is appropriate or not. If not, suggest an alternative.

   - Medical AI for Emergency Diagnosis
     - Dataset: 50,000 emergency room visits from 3 hospitals (Hospital A: 35,000 cases, Hospital B: 10,000 cases, Hospital C: 5,000 cases)
     - Goal: Predict whether patients need immediate surgery based on symptoms and vital signs

   - Climate change prediction model
     - Dataset: Daily temperature and weather measurements from 1980-2023 across 200 weather stations
     - Goal: Predict future temperature trends for the next 5 years

2. You’re building a model to predict house prices. Your workflow includes:

   - Calculating neighborhood average prices as a feature
   - Removing outliers (houses > $2M)
   - Normalizing all price-related features
   - Training a regression model

   At what point should you split your data? Identify potential sources of information leakage and explain how your timing prevents them.

3. You have a dataset with 1,000 observations and are deciding between 5-fold CV and 10-fold CV. For each approach, explain:

   - How many observations will be in each analysis set? _______
   - How many observations will be in each assessment set? _______
   - Which approach would likely give you a less biased estimate of model performance?
   - Which approach would likely give you more stable (less variable) results?
   - Which would you choose and why?