Lab 06: Logistic Regression

**Due date:** Thursday, April 3, 2025 submitted as Word document to Canvas ***Lab06*** link

This lab counts 9 % toward your total grade.

**Objectives:** In this lab, you will practice your skills in

1. Building a decision tree classifier
2. Examine the structure and size of the tree
3. Training data and testing data
4. Prune the Tree

**Format of answer:** Submit your answers as a **Word document** with graphs and verbal descriptions, properly labeled in the task sequence, with answers in red text and only relevant content included. **For each task, please provide R code, screenshot of the result and.**

**Accident\_data.shp** show the spatial distribution of traffic accidents in the city of Dallas. The accident is represented as 1 (Yes, a traffic accident occurred at the location) and 0 (No, no traffic accident occurred at location).

A map with red and blue dots

AI-generated content may be incorrect.

# Task 1: Load data (2 pts)

We will use the **Accident\_data.shp** to practice the skills in decision tree.

1. Load the **Accident\_data.shp** data using **sf::st\_read()** (1 pt)

*accident\_data = sf::st\_read('Accident\_data.shp')*

1. Decision tree in R handle categorical independent variable more effectively when they are stored as factors. Convert the following columns ("accident", "intersect", "BelowFreez", "Fog", "Thunder","FrozenPrec") into factors. (1 pt)

# Task 2: Train-Test Split and Initial Tree modeling (2 pts)

1. Using random sampling to split the data into 70% training and 30% testing. Use *table(training$accident)* to examine how balanced the accident classes are in training data and testing data. Provide interpretation.

*sample\_idx <- sample(nrow(accident\_data), size = 0.7 \* nrow(accident\_data))*

*train <- accident\_data[sample\_idx, ]*

*test <- accident\_data[-sample\_idx, ]*

1. Build a decision tree model using training data to model the relationship between accident occurrence and independent variables.
2. Visualizing the tree helps you interpret the decision rules. Plot the tree and label its nodes. Follow the structure from root to leaf – what does it tell you about accident likelihood under certain conditions.

# Task 3: Cross-Validation for Optimal Tree Size (2 pts)

1. Large trees may overfit the training data. Cross-validation identifies a simpler model with better generalization. Perform 5-fold cross-validation to determine the optimal tree size for the full tree in Task 2.b. (2 pt)

# Task 4: Prune Tree (1 pt)

1. Prune the initial full tree to the optimal size identified in Task 3.a. (1 pt)

# Task 5: Prediction (2 pts)

1. Using the pruned tree from Task4.a to predict accident occurrence on the test data. Measure the model accuracy based on the confusion matrix. (2 pts)