Classwork 1: Classification Trees Fundamentals

Machine Learning with Tree-Based Models in R

Prof. Asc. Endri Raco, Ph.D.

November 2025

Learning Objectives

By completing this classwork, you will:

- Build a classification tree model using tidymodels framework
- Implement stratified train-test splitting to handle class imbalance
- Generate predictions and evaluate model performance using confusion matrices and accuracy
- Interpret model results in the context of a healthcare prediction problem

Time Allocation: 30 minutes

Setup (5 min) | Model Building (10 min) | Evaluation (10 min) | Interpretation (5 min)

Dataset: Pima Indians Diabetes

We will use a diabetes prediction dataset with 768 observations and the following features:

- `pregnancies`: Number of times pregnant
- `glucose`: Plasma glucose concentration
- `blood_pressure`: Diastolic blood pressure (mm Hg)
- `skin_thickness`: Triceps skin fold thickness (mm)
- `insulin`: 2-Hour serum insulin (mu U/ml)
- `bmi`: Body mass index (weight in kg/(height in m)^2)
- `age`: Age in years
- `outcome`: Target variable (yes = has diabetes, no = no diabetes)

Task 1: Setup and Data Exploration (5 minutes)

Load the required packages and examine the diabetes dataset structure.

```r
# Load required libraries
library(tidymodels)
library(dplyr)

# Load the diabetes dataset
data(diabetes, package = "modeldata")

# Task 1a: Examine the first 10 rows of the dataset


```

```
# Task 1b: Check the dimensions and structure of the data
# Hint: Use dim() and str()


# Task 1c: Calculate the proportion of each outcome class
# This helps us understand if we have class imbalance
```

Question 1.1: What percentage of patients have diabetes in this dataset? Is this dataset balanced or imbalanced?

Your Answer:

Task 2: Data Splitting with Stratification (10 minutes)

Create a proper train-test split that maintains the distribution of the outcome variable.

```
# Task 2a: Set a random seed for reproducibility
set.seed(2025)

# Task 2b: Create a stratified split with 75% training data
# Use initial_split() with prop = 0.75 and strata = outcome


# Task 2c: Extract training and test datasets


# Task 2d: Verify the split worked correctly
# Check that both sets have similar proportions of the outcome
```

Question 2.1: Why is stratification important when splitting classification data? What would happen if we didn't use stratification?

Your Answer:

Question 2.2: Calculate and report the proportion of "yes" outcomes in both your training and test sets. Are they similar?

Your Answer:

Task 3: Build and Train Classification Tree (10 minutes)

Create a decision tree model specification and train it on the diabetes data.

```
# Task 3a: Create a decision tree model specification
# Use decision_tree() with engine "rpart" and mode "classification"


# Task 3b: Fit the model using age and bmi as predictors
# Formula: outcome ~ age + bmi


# Task 3c: Print the model summary to see the tree structure
```

```
# Task 3d: Now fit a model using ALL available predictors
# Formula: outcome ~ .
```

Question 3.1: Compare the two models (age+bmi vs all predictors). Which model do you expect to perform better and why?

Your Answer:

Task 4: Predictions and Model Evaluation (10 minutes)

Generate predictions on the test set and evaluate model performance.

```
# Use your model trained on all predictors for this task

# Task 4a: Generate class predictions on the test set
# Use predict() with type = "class"


# Task 4b: Generate probability predictions
# Use predict() with type = "prob"


# Task 4c: Combine predictions with true outcomes
# Bind the class predictions and test data together


# Task 4d: Create a confusion matrix
# Use conf_mat() from yardstick


# Task 4e: Calculate accuracy
# Use accuracy() from yardstick
```

Question 4.1: Interpret your confusion matrix. How many true positives, true negatives, false positives, and false negatives did your model produce?

Your Answer:

Question 4.2: What is your model's accuracy? In the context of diabetes prediction, which type of error (false positive or false negative) would be more concerning from a healthcare perspective?

Your Answer:

Task 5: Critical Thinking (5 minutes)

Question 5.1: The lecture mentioned that accuracy alone can be misleading. Given your confusion matrix, is accuracy a sufficient metric for this diabetes prediction task? Why or why not?

Your Answer:

Question 5.2: What are two advantages and two disadvantages of using decision trees for this medical prediction task?

Your Answer:

Advantages:

Disadvantages:

Submission Instructions

1. Complete all code chunks and run them to verify they work
2. Answer all questions in the spaces provided
3. Knit this document to PDF
4. Submit the PDF file through the course management system

Grading Rubric (100 points total)

Component	Points	Criteria
Task 1: Data Exploration	15	Correct data examination and class proportion calculation
Task 2: Data Splitting	20	Proper stratified split implementation and verification
Task 3: Model Building	20	Correct model specification and training for both scenarios
Task 4: Evaluation	25	Accurate predictions, confusion matrix, and accuracy calculation
Task 5: Critical Thinking	20	Thoughtful analysis of metrics and model properties
Total	**100**	

Expected Outputs Summary

By the end of this classwork, you should have:

- A stratified train-test split with verified proportions
- Two trained classification tree models (simple and full)
- Class and probability predictions on the test set
- A confusion matrix showing model performance breakdown
- An accuracy score quantifying overall performance
- Written reflections on model evaluation and limitations

This classwork covers approximately 25% of the lecture content, focusing on classification trees, data splitting with stratification, and basic model evaluation metrics.