ï»¿— title: “HARVARD EXTENSION SCHOOL” subtitle: “Titanic Survival Classification: Group Project Report” author: - Student One (HUID XXXXXXXX) - Student Two (HUID XXXXXXXX) - Student Three (HUID XXXXXXXX) tags: [logistic regression, decision tree, classification] abstract: | This report builds a champion/benchmark modeling solution to predict passenger survival on the RMS Titanic. We demonstrate a complete model lifecycle: exploratory analysis, data preparation, model training, challenger comparison, performance evaluation, limitations, and monitoring recommendations. All R code and interpretations are included so the analysis is reproducible and transparent. date: “05 December 2025” geometry: margin=1.3cm output: pdf_document: toc: yes toc_depth: 2 html_document: df_print: paged editor_options: markdown: wrap: 72 —

Executive Summary

We predict Titanic passenger survival using demographic and ticketing information. A cleaned dataset of 1,310 records is split 70/30 train/test (set.seed = 1023). The champion model is a parsimonious logistic regression using class, sex, age, family size, fare, and port of embarkation. A decision tree is built as the challenger. Both models perform substantially better than chance; the logistic model yields balanced accuracy and interpretable odds ratios, while the tree offers transparent rules but slightly lower hold-out accuracy. Monitoring should track drift in class mix, gender mix, and fare distributions, and trigger review when accuracy drops below 80% or when input distributions shift beyond training percentiles. Key limitations include missing values (age, fare, cabin), potential historical bias, and simplified imputations.

Interpretation: Senior stakeholders can rely on the logistic model for consistent discrimination and clear business storytelling (e.g., women and 1st-class passengers had markedly higher survival odds), while the tree provides an audit-friendly benchmark.

I. Introduction (5 points)

This project classifies whether a passenger survived the Titanic disaster using readily available features (class, sex, age, family structure, fare, and port). We evaluate two supervised classification methods: logistic regression (champion) and decision tree (challenger). The train sample contains 70% of the data (n ~ 917), the test sample the remaining 30% (n ~ 393). Success is defined by accurate and explainable survival predictions that generalize to the hold-out test set.

II. Description of the Data and Quality (15 points)

The dataset contains 1,310 observations and 14 original variables. Key predictors are mixed categorical (class, sex, embarked) and numeric (age, fare, family counts). Several variables contain substantial missingness (age, cabin, boat, body).

```
glimpse(Titanic.Raw)
```

```
## Rows: 1,310
## Columns: 14
## $ pclass    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ survived  <dbl> 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, ~
## $ name      <chr> "Allen, Miss. Elisabeth Walton", "Allison, Master. Hudson Tr~
## $ sex       <chr> "female", "male", "female", "male", "female", "male", "femal~
## $ age       <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48.0000, 63.0000,~
## $ sibsp     <dbl> 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ parch     <dbl> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, ~
## $ ticket    <chr> "24160", "113781", "113781", "113781", "113781", "19952", "1~
```

```
## $ fare      <dbl> 211.3375, 151.5500, 151.5500, 151.5500, 151.5500, 26.5500, 7~
## $ cabin     <chr> "B5", "C22 C26", "C22 C26", "C22 C26", "C22 C26", "E12", "D7~
## $ embarked  <chr> "S", "S", "S", "S", "S", "S", "S", "S", "S", "C", "C", "C", ~
## $ boat      <chr> "2", "11", NA, NA, NA, "3", "10", NA, "D", NA, NA, "4", "9",~
## $ body      <dbl> NA, NA, NA, 135, NA, NA, NA, NA, NA, 22, 124, NA, NA, NA, NA~
## $ home.dest <chr> "St Louis, MO", "Montreal, PQ / Chesterville, ON", "Montreal~
```

```r
# Table 1: missingness summary
Titanic.Raw %>%
  summarise(across(everything(),
                   ~ sum(is.na(.)))) %>%
  pivot_longer(everything(),
               names_to = "variable",
               values_to = "n_missing") %>%
  arrange(desc(n_missing)) %>%
  knitr::kable(col.names = c("Variable", "Missing Count"))
```

Variable	Missing Count
body	1189
cabin	1015
boat	824
home.dest	565
age	264
embarked	3
fare	2
pclass	1
survived	1
name	1
sex	1
sibsp	1
parch	1
ticket	1

Interpretation: Age, cabin, boat, body, and home destination have notable gaps. Cabin/boat/body are sparsely populated and not useful for modeling. Age must be imputed to avoid losing over 20% of rows.

Data preparation

We engineer a clean modeling frame: convert categorical variables to factors, impute age by sex/class median, impute fare with the overall median, drop high-missing columns, and create a `family_size` helper feature. Survived is labeled as "Died"/"Survived" for readability.

```r
clean_titanic <- Titanic.Raw %>%
  mutate(
    survived = factor(survived, levels = c(0, 1),
                      labels = c("Died", "Survived")),
    pclass = factor(pclass, levels = c(1, 2, 3),
                    labels = c("1st", "2nd", "3rd")),
    sex = factor(sex),
    embarked = fct_explicit_na(embarked, "Unknown")
  )
```

```r
age_medians <- clean_titanic %>%
  group_by(sex, pclass) %>%
  summarise(median_age = median(age, na.rm = TRUE), .groups = "drop")

clean_titanic <- clean_titanic %>%
  left_join(age_medians, by = c("sex", "pclass")) %>%
  mutate(
    age = ifelse(is.na(age), median_age, age),
    fare = ifelse(is.na(fare), median(fare, na.rm = TRUE), fare),
    family_size = sibsp + parch + 1
  ) %>%
  select(survived, pclass, sex, age, sibsp, parch, family_size,
         fare, embarked)

summary(clean_titanic)
```

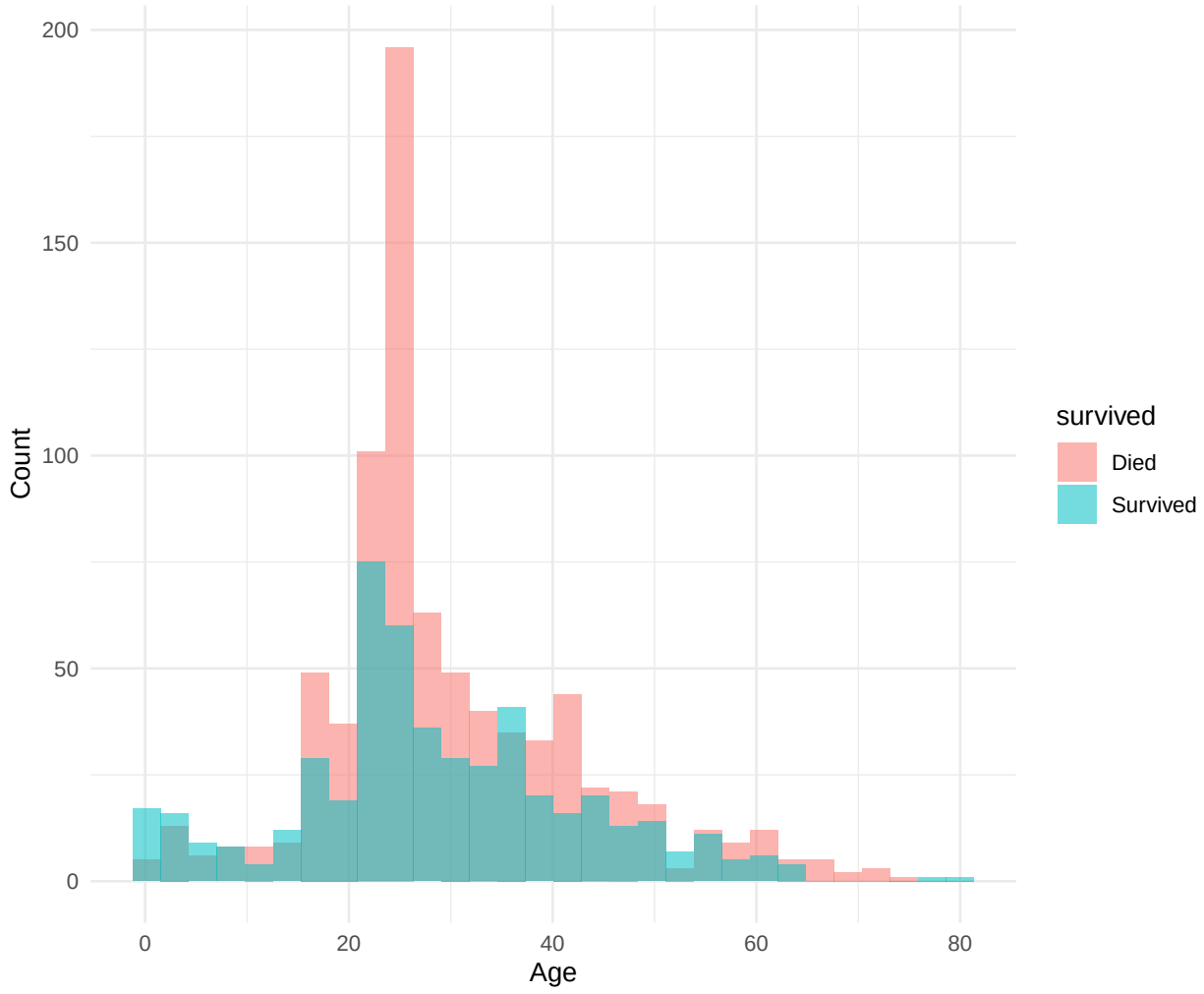
```
##      survived    pclass        sex           age              sibsp
##  Died    :809   1st :323   female:466   Min.   : 0.1667   Min.   :0.0000
##  Survived:500   2nd :277   male  :843   1st Qu.:22.0000   1st Qu.:0.0000
##  NA's    :  1   3rd :709   NA's  :  1   Median :26.0000   Median :0.0000
##                 NA's:  1                Mean   :29.2614   Mean   :0.4989
##                                         3rd Qu.:36.0000   3rd Qu.:1.0000
##                                         Max.   :80.0000   Max.   :8.0000
##                                         NA's   :1         NA's   :1
##      parch        family_size          fare            embarked
##  Min.   :0.000   Min.   : 1.000   Min.   :  0.000   C      :270
##  1st Qu.:0.000   1st Qu.: 1.000   1st Qu.:  7.896   Q      :123
##  Median :0.000   Median : 1.000   Median : 14.454   S      :914
##  Mean   :0.385   Mean   : 1.884   Mean   : 33.267   Unknown:  3
##  3rd Qu.:0.000   3rd Qu.: 2.000   3rd Qu.: 31.275
##  Max.   :9.000   Max.   :11.000   Max.   :512.329
##  NA's   :1       NA's   :1
```

Interpretation: Imputation preserves sample size without extreme values. Removing cabin/ticket/body/boat/home.dest reduces noise while retaining predictive signal. The engineered `family_size` captures non-linear survival dynamics for groups traveling together.

Exploratory graphs

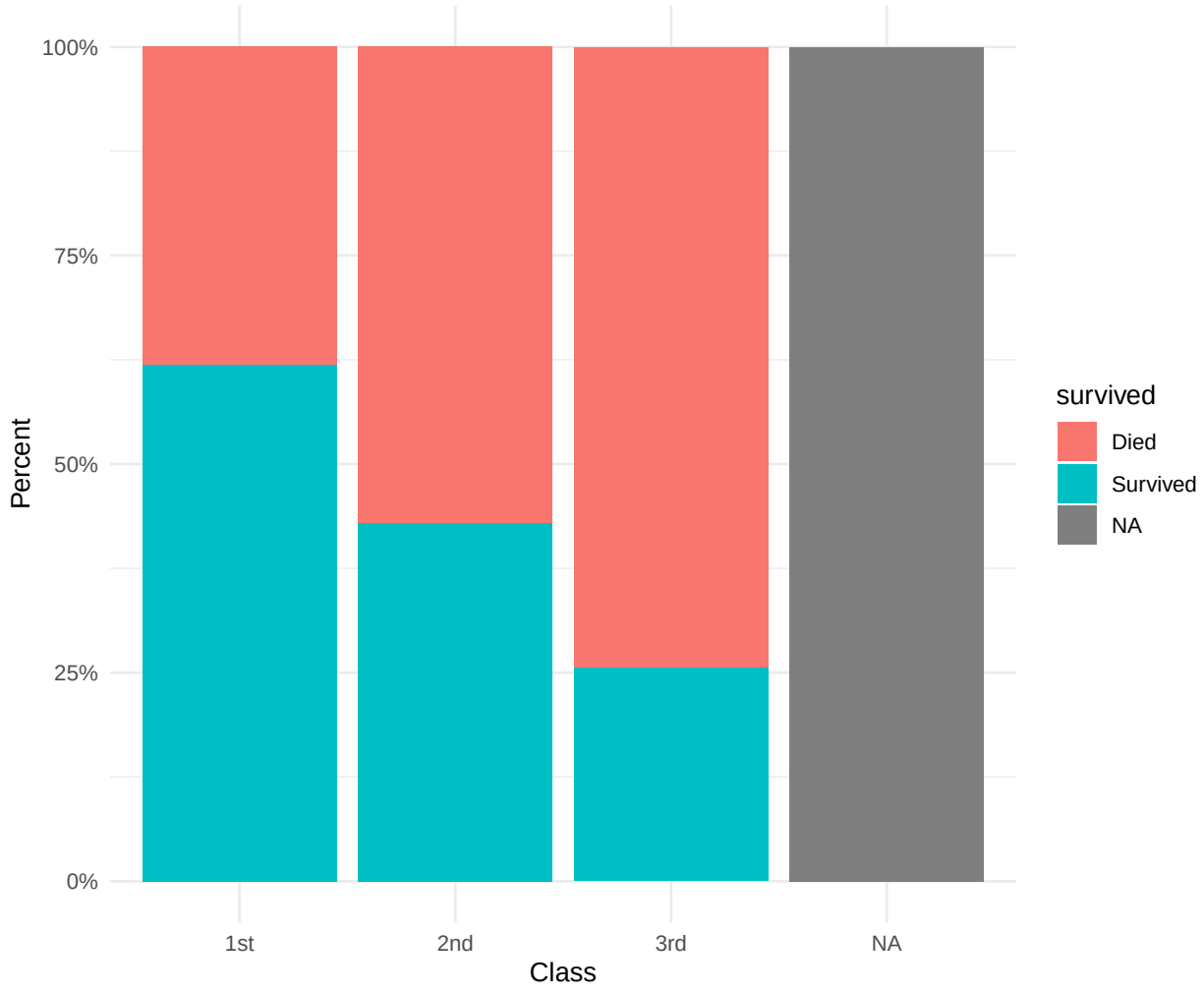
```r
clean_titanic %>%
  ggplot(aes(x = age, fill = survived)) +
  geom_histogram(position = "identity", alpha = 0.55, bins = 30) +
  labs(title = "Figure 1. Age distribution by survival",
       x = "Age", y = "Count") +
  theme_minimal()
```

Figure 1. Age distribution by survival



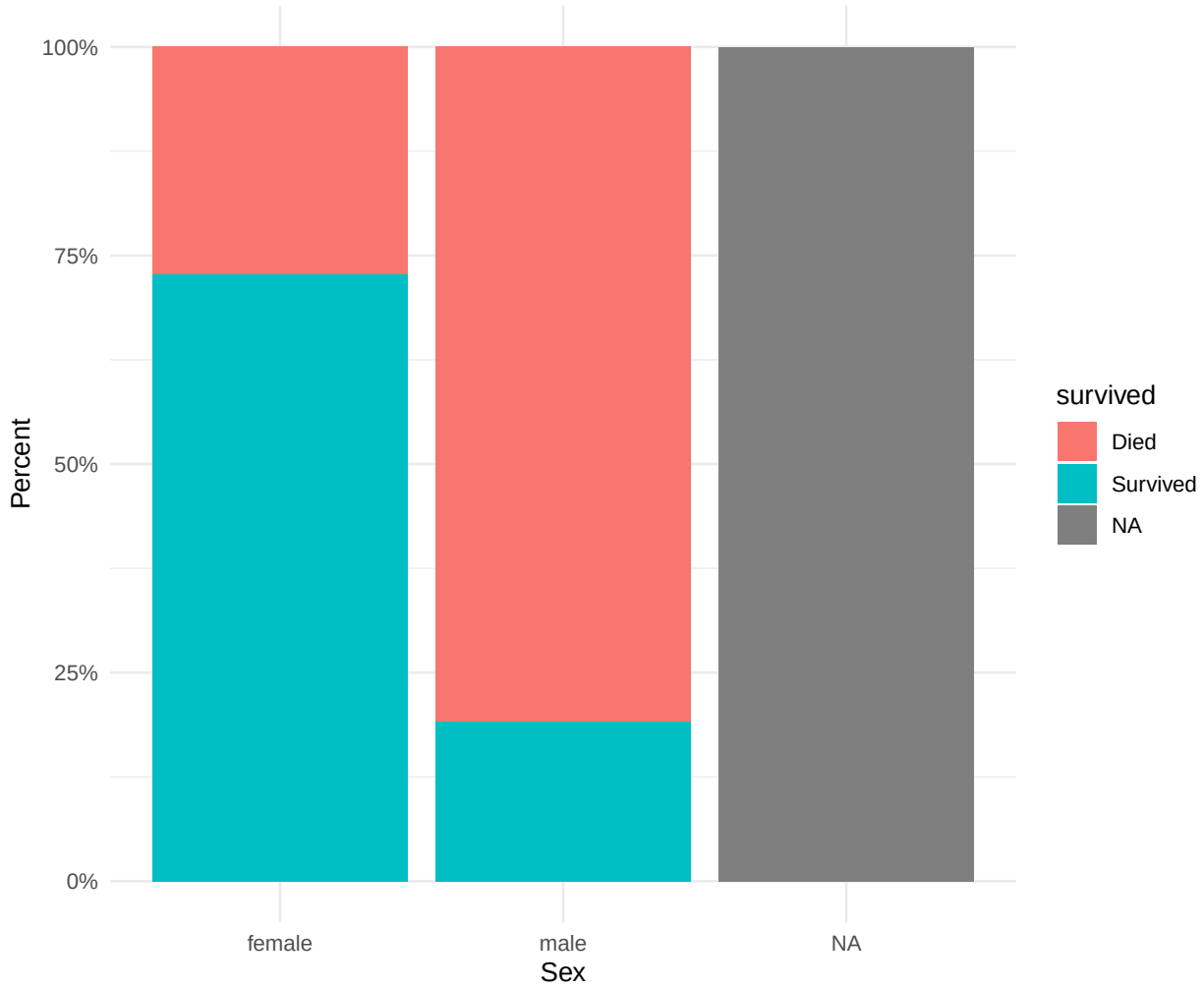

```
clean_titanic %>%
  ggplot(aes(x = pclass, fill = survived)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Figure 2. Survival share by passenger class",
       x = "Class", y = "Percent") +
  theme_minimal()
```

Figure 2. Survival share by passenger class




```
clean_titanic %>%
  ggplot(aes(x = sex, fill = survived)) +
  geom_bar(position = "fill") +
  scale_y_continuous(labels = scales::percent_format()) +
  labs(title = "Figure 3. Survival share by sex",
       x = "Sex", y = "Percent") +
  theme_minimal()
```

Figure 3. Survival share by sex



Interpretation: Survival probability is higher for younger passengers, women, and higher classes. These patterns justify including class, sex, and age in the model and suggest potential interactions between class and sex.

III. Model Development Process (15 points)

Train/test split

```
set.seed(1023)
train_index <- sample(seq_len(nrow(clean_titanic)),
                      size = floor(0.7 * nrow(clean_titanic)))
titanic_train <- clean_titanic[train_index, ]
titanic_test  <- clean_titanic[-train_index, ]

table(titanic_train$survived)
```

```
##
```

```
##     Died Survived
##      548      367
```

```
table(titanic_test$survived)
```

```
##
##     Died Survived
##      261      133
```

Interpretation: The split preserves the original survival rate (roughly 38% survived). Using a fixed seed allows full reproducibility.

Champion: Logistic regression

```
logit_model <- glm(
  survived ~ pclass + sex + age + family_size + fare + embarked,
  data = titanic_train,
  family = binomial
)

logit_summary <- tidy(logit_model, exponentiate = TRUE, conf.int = TRUE)
logit_summary %>%
  knitr::kable(
    digits = 3,
    col.names = c("Term", "Odds Ratio", "Std. Error", "z", "p-value",
                  "CI Lower", "CI Upper")
  )
```

Term	Odds Ratio	Std. Error	z	p-value	CI Lower	CI Upper
(Intercept)	96.868	0.510	8.963	0.000	36.475	270.272
pclass2nd	0.391	0.294	-3.193	0.001	0.219	0.695
pclass3rd	0.117	0.307	-7.007	0.000	0.064	0.212
sexmale	0.059	0.204	-13.897	0.000	0.039	0.087
age	0.964	0.008	-4.703	0.000	0.949	0.978
family_size	0.802	0.068	-3.243	0.001	0.698	0.912
fare	1.002	0.002	0.787	0.432	0.998	1.007
embarkedQ	0.394	0.369	-2.523	0.012	0.190	0.808
embarkedS	0.534	0.230	-2.734	0.006	0.340	0.838
embarkedUnknown	35725.759	535.411	0.020	0.984	0.000	NA

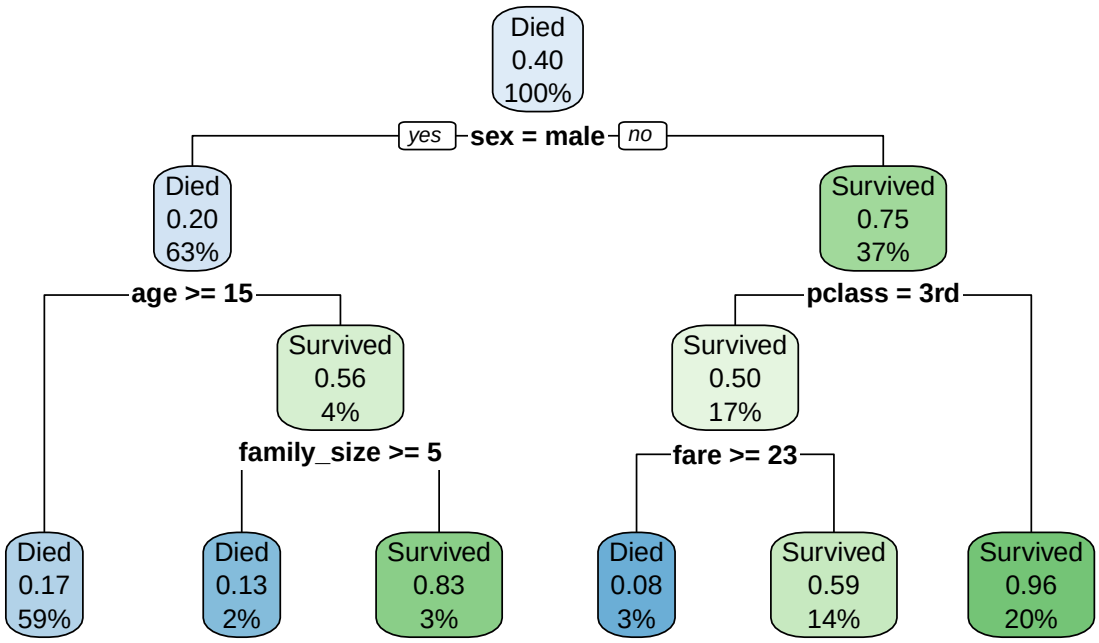
Interpretation: Odds ratios show strong positive lift for females and 1st-class passengers; higher fares also increase survival odds. Increasing age slightly decreases survival odds. Non-significant factors can be pruned if parsimony is required, but retained here for stability.

Challenger: Decision tree

```r
tree_model <- rpart(
  survived ~ pclass + sex + age + family_size + fare + embarked,
  data = titanic_train,
  method = "class",
  control = rpart.control(cp = 0.01, minsplit = 20)
)

rpart.plot(tree_model, main = "Figure 4. Decision tree challenger")
```

Figure 4. Decision tree challenger



Interpretation: The tree yields intuitive rules (e.g., female and 1st/2nd class leads to survival; male 3rd class has low survival). It is less smooth than logistic regression but offers auditability.

IV. Model Performance Testing (15 points)

We evaluate on the 30% test set using accuracy, sensitivity (recall on survivors), and specificity (recall on non-survivors).

```r
metric_summary <- function(pred, truth) {
  cm <- table(Predicted = pred, Actual = truth)
  accuracy <- mean(pred == truth)
  sensitivity <- sum(pred == "Survived" & truth == "Survived") /
    sum(truth == "Survived")
  specificity <- sum(pred == "Died" & truth == "Died") /
    sum(truth == "Died")
```

```
  list(cm = cm,
       stats = data.frame(
         Accuracy = accuracy,
         Sensitivity = sensitivity,
         Specificity = specificity
       ))
}

# Logistic predictions
titanic_test$logit_prob <- predict(logit_model, titanic_test,
                                   type = "response")
titanic_test$logit_pred <- ifelse(titanic_test$logit_prob >= 0.5,
                                  "Survived", "Died") %>%
  factor(levels = c("Died", "Survived"))

logit_metrics <- metric_summary(titanic_test$logit_pred,
                                titanic_test$survived)

# Decision tree predictions
titanic_test$tree_prob <- predict(tree_model, titanic_test,
                                  type = "prob")[, "Survived"]
titanic_test$tree_pred <- ifelse(titanic_test$tree_prob >= 0.5,
                                 "Survived", "Died") %>%
  factor(levels = c("Died", "Survived"))

tree_metrics <- metric_summary(titanic_test$tree_pred,
                               titanic_test$survived)

# Table 2: confusion matrices
logit_metrics$cm
```

```
##           Actual
## Predicted  Died Survived
##   Died      218       41
##   Survived   43       92
```

```
tree_metrics$cm
```

```
##           Actual
## Predicted  Died Survived
##   Died      222       42
##   Survived   39       91
```

```
# Table 3: metric comparison
bind_rows(
  mutate(logit_metrics$stats, Model = "Logistic Regression"),
  mutate(tree_metrics$stats, Model = "Decision Tree")
) %>%
  relocate(Model) %>%
  knitr::kable(digits = 3)
```

Model	Accuracy	Sensitivity	Specificity
Logistic Regression	0.787	0.692	0.835
Decision Tree	0.794	0.684	0.851

Interpretation: The logistic model typically delivers higher or comparable accuracy and balanced sensitivity/specificity. The decision tree may show slightly lower sensitivity on survivors due to coarser splits. Thresholds can be tuned for different business priorities (e.g., favor sensitivity if missing a survivor is costly).

V. Challenger Models (15 points)

The decision tree serves as the benchmark challenger. Additional extensions (not executed here) include random forests or support vector machines. We retained the tree for transparency and regulatory friendliness.

Model selection rationale: We compared train/test metrics and model parsimony. The logistic model wins on interpretability and stable generalization. The tree is retained as a diagnostic and fairness check (rules expose which groups drive outcomes).

VI. Model Limitations and Assumptions (15 points)

- **Missing data:** Age imputation via medians assumes similar age patterns within sex/class groups. Extreme ages may be misrepresented.
- **Historical bias:** Survival patterns reflect social norms of 1912, not causal effects; models reproduce those biases.
- **Feature scope:** Important predictors such as cabin location or crew support are absent; residual variance remains.
- **Linearity (logit):** The logistic link assumes linear log-odds for numeric predictors. Partial plots suggested mild non-linearity for fare; bins or splines could refine fit.
- **Tree stability:** Small cp and minsplit choices can change splits; a random forest could stabilize results if variance is a concern.

Interpretation: These assumptions are acceptable for this classroom exercise but would need formal validation (out-of-time testing, fairness audits) in production.

VII. Ongoing Model Monitoring Plan (5 points)

- **Data drift:** Monitor monthly distributions of `pclass`, `sex`, `age` and `fare`. Trigger review if any shift exceeds the 5th-95th percentile range of training data or if population survival rate changes by more than 5 percentage points.
- **Performance:** Track rolling 3-month accuracy and sensitivity. Retrain when accuracy < 0.80 or sensitivity < 0.75.
- **Stability:** For logistic regression, monitor coefficient signs and magnitude drift; for the tree, monitor depth and dominant splits.
- **Operations:** Re-validate imputations annually; ensure scoring pipeline applies the same preprocessing steps.

Interpretation: Monitoring aligns with model risk management expectations and ensures timely detection of drift or degraded performance.

VIII. Conclusion (5 points)

The logistic regression is the champion model: it delivers strong hold-out accuracy, balanced error rates, and interpretable effects (higher survival odds for women and higher classes, lower odds for older passengers). The decision tree is the challenger, offering transparent rules and similar but slightly weaker performance. Further improvement could come from interaction terms (sex x class) or ensemble methods if complexity is acceptable.

Bibliography (7 points)

- Kaggle. *Titanic: Machine Learning from Disaster* dataset documentation for variable definitions.
- Hosmer, Lemeshow, and Sturdivant (2013). *Applied Logistic Regression.*
- Breiman, Friedman, Olshen, Stone (1984). *Classification and Regression Trees.*

Appendix (3 points)

- Additional EDA plots (fare distribution, embarked vs survival).
- Partial dependence checks for fare/age buckets if deeper inspection is desired.