Assignment IV

Multiple Logistic Regression and Decision Tree

Silpa Soni Nallacheruvu (19980824-5287) Hernan Aldana (20000526-4999)

2025-01-03

Overview:

Analysis of The ICU Study data (Hosmer & Lemeshow (1989): Applied Logistic Regression), with 200 patient records admitted to an Intensive Care Unit (ICU) using multiple logistic regression and decision tree models to identify factors that affect the survival of such patients.

Exercise 4:1 (Multiple Logistic Regression)

Question 1

Report the model selection process briefly. Based on your chosen model, which factors affect the probability of not surviving? Report odds ratios with confidence intervals for the most important variables/factors, and interpret them. Use the variable names from the table (not V3, V4, etc.).

Approach:

- **Data Preparation**:
  1. Load the dataset and rename variables for clarity.
  2. Simplify categorical variables for improved analysis: Categories for Ethnicity and ConsciousnessLevel were combined to reduce their three-level factors into binary factors, enhancing interpretability and resolving the issue of their cell counts being low.

     Ethnicity:

     - The existing categories: White, Black and Other were re-categorized to ensure sufficient representation, with White and Others grouped together due to the majority cell count being White.
     - Category 1 = White was retained as is.
     - Categories 2 = Black and 3 = Other were merged into a single category labeled as 0.

     ConsciousnessLevel:

     - The existing categories: Awake, Unconscious, and Coma were re-categorized to ensure sufficient representation, with Awake and Others grouped together due to the majority cell count being Awake.
       * Category 0 = Awake remained unchanged.
       * Categories 1 = Unconscious and 2 = Coma were combined into a single category labeled as 1.

3. Remove unnecessary variables: Variables that do not contribute meaningfully to the analysis or prediction, such as PatientID (which lacks clinical significance), were excluded to streamline the dataset and focus on relevant factors.

Rationale: To ensure sufficient representation in each category for statistical analysis.

- **Model Fitting**:
  1. Fit an empty logistic regression model and a full logistic regression model to be used in the stepwise selection.
  2. We chose AIC with a bidirectional stepwise method for variable selection due to its balance between model fit and complexity, which is suitable for predictive modeling. AIC prioritizes minimizing prediction error over identifying the true model, making it ideal for datasets with a moderate sample size (200 records) and a relatively large number of predictors (20 variables). Bidirectional selection combines forward and backward approaches, ensuring a comprehensive search for significant variables while avoiding overfitting. The stepwise selection process helps identify the most relevant predictors of survival.
- **Analysis of the Final Model**:
  1. Extract coefficients, odds ratios, and their 95% confidence intervals for significant variables.
  2. Significant variables were identified as those with p-values less than 0.05.
- **Interpretation**:
  1. Interpret the results from the AIC, odds ratios, and confidence intervals to determine the most impactful predictors.

Results:

Summary of the final model after performing stepwise selection using AIC:

```
##
## Call:
## glm(formula = Survival ~ ConsciousnessLevel + TypeOfAdmission +
##     Age + Cancer + BloodPressure + BloodCarbonDioxide + BloodPH,
##     family = binomial, data = data_ca4)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -5.326431   1.607500  -3.313 0.000921 ***
## ConsciousnessLevel  4.496229   1.014153   4.433 9.27e-06 ***
## TypeOfAdmission     3.027648   0.955149   3.170 0.001525 **
## Age                 0.042684   0.013707   3.114 0.001845 **
## Cancer              2.402019   0.888072   2.705 0.006835 **
## BloodPressure      -0.014538   0.007083  -2.053 0.040101 *
## BloodCarbonDioxide -2.253218   0.994392  -2.266 0.023456 *
## BloodPH             1.854290   0.874168   2.121 0.033904 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 128.44  on 192  degrees of freedom
## AIC: 144.44
##
```

```
## Number of Fisher Scoring iterations: 6
```

Model selection:

- The final logistic regression model includes the following variables: ConsciousnessLevel, TypeOfAdmission, Age, Cancer, BloodCarbonDioxide, BloodPH, and BloodPressure.
- These variables were selected using a stepwise AIC, which ensures a balance between model complexity and goodness of fit.

Significant Variables:

- Variables with a p-value < 0.05 are considered significant predictors of survival:
  - ConsciousnessLevel
  - TypeOfAdmission
  - Age
  - Cancer
  - BloodPressure
  - BloodCarbonDioxide
  - BloodPH

Odds Ratios and Confidence Intervals:

Here's the final report after extracting odds ratios and confidence intervals for significant variables:

```
## Waiting for profiling to be done...
```

Table 1: Odds Ratios and Confidence Intervals for Significant Variables

	Variable	OddsRatio	CI.Lower	CI.Upper
ConsciousnessLevel	ConsciousnessLevel	89.6783468	15.4206181	894.7216188
TypeOfAdmission	TypeOfAdmission	20.6486198	4.0327671	186.1696351
Age	Age	1.0436083	1.0178804	1.0746098
Cancer	Cancer	11.0454497	1.9976034	73.2056473
BloodPressure	BloodPressure	0.9855667	0.9712796	0.9989335
BloodCarbonDioxide	BloodCarbonDioxide	0.1050606	0.0124716	0.6405209
BloodPH	BloodPH	6.3871615	1.1433812	38.4807236

1. ConsciousnessLevel:
   - Odds Ratio: 89.68 (CI: 15.42–894.72)
   - ConsciousnessLevel: 1 = Unconscious or Coma, 0 = Awake
   - Patients who are unconscious or in a coma are over 89.68 times more likely to not survive compared to those who are conscious. This indicates that ConsciousnessLevel is a critical predictor of survival.
2. TypeOfAdmission:
   - Odds Ratio: 20.65 (CI: 4.03–186.17)
   - TypeOfAdmission: 1 = Acute, 0 = Non-acute
   - Acute admissions are associated with approximately 20.65 times higher odds of not surviving compared to non-acute admissions. This suggests that the type of admission significantly impacts survival outcomes.

3. Age:
    - Odds Ratio: 1.04 (CI: 1.01–1.07)
    - For every additional year of age, the odds of not surviving increase by 4%. This suggests that older patients are at a higher risk of not surviving.
4. Cancer:
    - Odds Ratio: 11.05 (CI: 1.99–73.2)
    - Cancer: 1 = Yes, 0 = No
    - Patients with Cancer have 11.05 times higher odds of not surviving compared to those without Cancer. This highlights the significant impact of Cancer on survival outcomes.
5. BloodPressure:
    - Odds Ratio: 0.986 (CI: 0.97–0.998)
    - For each mm Hg increase in blood pressure, the odds of not surviving decrease by 1.4%. This indicates that higher blood pressure, within the normal range, reflecting a healthier cardiovascular system, is associated with better survival outcomes.
6. BloodCarbonDioxide:
    - Odds Ratio: 0.105 (CI: 0.01–0.64)
    - BloodCarbonDioxide Levels: 1 = above 45, 0 = below 45
    - Patients with blood carbon dioxide levels above 45 have a 89.5% lower odds of not surviving compared to those with below 45. This suggests that maintaining normal blood carbon dioxide levels is crucial for survival.
7. BloodPH:
    - Odds Ratio: 6.38 (CI: 1.14–38.48)
    - BloodPH Levels: 1 = below 7.25, 0 = above 7.25
    - Patients with blood pH levels below 7.25 have 6.38 times higher odds of not surviving compared to those with above 7.25. This indicates that abnormal blood pH levels are associated with a higher risk of not surviving.

Conclusion:

The model indicates that factors such as consciousness level, type of admission, age, cancer, blood carbon dioxide, blood pressure and blood pH are significant predictors of survival. Patients who are unconscious, have acute admissions, are older, have cancer, exhibit very low blood pressure, higher levels of blood carbon dioxide, and lower levels of blood pH are at a higher risk of not surviving. This analysis helps identify high-risk patients and can guide clinical interventions to improve survival rates by addressing these critical factors.

Based on the prior knowledge of p-values from the model summary, **ConsciousnessLevel** appears to be the **most significant predictor** of survival, with an odds ratio of 89.68, indicating a strong impact on the odds of survival. Conversely, **BloodPressure** is the **least significant** predictor, with an odds ratio of 0.986, suggesting that higher blood pressure reduces the odds of non-survival by only 1.4%. This interpretation will be further validated in the following questions by testing the significance of these predictors through ROC curve analysis and LOOCV.

Question 2

How well does your chosen model fit the data? In assignment 3, deviance was used to assess model fit. However, for individual-level data, deviance is unsuitable. Instead, perform the Hosmer-Lemeshow goodness-of-fit test using the recommended R code.

Approach:

- Understand the Hosmer-Lemeshow Test:
  - The Hosmer-Lemeshow test evaluates whether the observed event rates matches the expected probabilities predicted by the model.
  - The null hypothesis is that the model fits the data well (a high p-value suggests no evidence of poor fit).
- Implementation:
  - Use the function for the test ResourceSelection::hoslem.test()
  - Calculate the predicted probabilities from the final model.
  - Specify the predicted probabilities from the final model and the actual outcomes while performing Hosmer-Lemeshow test (m_step and Survival respectively).
  - 10 is selected as the number of groups for the test because dividing by deciles is a common choice and it was sufficient for this dataset of 200 observations.

Hosmer-Lemeshow Test Results:

Here are the results of the Hosmer-Lemeshow goodness-of-fit test:

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  data_ca4$Survival, predicted_probs
## X-squared = 7.2202, df = 8, p-value = 0.5131
```

Interpretation:

1. The p-value of 0.5131 tells us to not reject the null hypothesis that the model fits the data well at a 5% significance level.

Conclusion:

The Hosmer-Lemeshow test with 10 groups (df = 8) yielded a chi-squared statistic of 7.2202 and a p-value of 0.5131. Since the p-value is much greater than 0.05, we fail to reject the null hypothesis that the model fits the data well. This indicates that the predicted probabilities align closely with the observed survival outcomes, confirming the adequacy of the logistic regression model for this dataset.

Question 3

Create a confusion matrix for the chosen model. Calculate the accuracy, sensitivity, specificity, and positive and negative predictive values for three values of threshold. Describe and explain the result.

Approach:

- Understand the Confusion Matrix:
  - A confusion matrix is a table that summarizes the performance of a classification model.
  - It shows the number of true positives, true negatives, false positives, and false negatives.
- Implementation:
  - Threshold of 0.3, 0.5 and 0.7 are used to classify the predicted probabilities into binary outcomes to view the range of sensitivity and specificity.

- Calculate the accuracy, sensitivity, specificity, and positive and negative predictive values from the confusion matrix for each threshold.
- Here, the positive of the model is "Not Survived" and the negative is "Survived".

Accuracy: Accuracy measures the proportion of correct predictions made by the model.

$$\text{Accuracy} = \frac{\text{True Positives (TP)+True Negatives (TN)}}{\text{Total Predictions (TP + TN + FP + FN)}}$$

Sensitivity: Sensitivity (True Positive Rate) measures the proportion of actual positive cases that are correctly identified by the model.

$$\text{Sensitivity (True Positive Rate)} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)+False Negatives (FN)}}$$

Specificity: Specificity (True Negative Rate) measures the proportion of actual negative cases that are correctly identified by the model.

$$\text{Specificity (True Negative Rate)} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)+False Positives (FP)}}$$

Threshold = 0.5: The model classifies predictions with probabilities greater than or equal to 0.5 as positive (predicts "not survive") and those below 0.5 as negative (predicts "survive").

Results:

Table 2: Confusion Matrix at threshold= 0.3

	Predicted Not Survived	Predicted Survived
Actual Not Survived	26	14
Actual Survived	19	141

Table 3: Confusion Matrix at threshold= 0.5

	Predicted Not Survived	Predicted Survived
Actual Not Survived	17	23
Actual Survived	3	157

Table 4: Confusion Matrix at threshold= 0.7

	Predicted Not Survived	Predicted Survived
Actual Not Survived	14	26
Actual Survived	2	158

Table 5: Performance Metrics at threshold= 0.3

Metric	Value
Accuracy	0.83500
Sensitivity	0.65000
Specificity	0.88125

Table 6: Performance Metrics at threshold= 0.5

Metric	Value
Accuracy	0.87000
Sensitivity	0.42500
Specificity	0.98125

Table 7: Performance Metrics at threshold= 0.7

Metric	Value
Accuracy	0.8600
Sensitivity	0.3500
Specificity	0.9875

Interpretation:

Threshold = 0.3:

Accuracy: The model has an accuracy of 0.835, meaning that it correctly predicted 83.5% of the cases.

Sensitivity: The sensitivity of 0.65 indicates that the model correctly identified 65% of the actual non-survivors.

Specificity: The specificity of 0.88125 suggests that the model correctly identified 88.125% of the actual survivors.

Threshold = 0.5:

Accuracy: The model has an accuracy of 0.87, meaning that it correctly predicted 87% of the cases.

Sensitivity: The sensitivity of 0.425 indicates that the model correctly identified 42.5% of the actual non-survivors.

Specificity: The specificity of 0.98125 suggests that the model correctly identified 98.1% of the actual survivors.

Threshold = 0.7:

Accuracy: The model has an accuracy of 0.86, meaning that it correctly predicted 86% of the cases.

Sensitivity: The sensitivity of 0.35 indicates that the model correctly identified 35% of the actual non-survivors.

Specificity: The specificity of 0.9875 suggests that the model correctly identified 98.75% of the actual survivors.

Impact of Threshold:

- Thresholds control the trade-off between sensitivity and specificity:
  - A lower threshold (e.g., 0.3) typically increases sensitivity because more cases are classified as "not survive," but it may decrease specificity.
  - A higher threshold (e.g., 0.7) typically increases specificity because fewer cases are classified as "not survive," but sensitivity may decrease.
  - Compared to 0.3 and 0.7 thresholds, the accuracy is highest at the threshold of 0.5, which is the default threshold for binary classification.

Conclusion:

The confusion matrix and performance metrics provide insights into the model's predictive accuracy. In general over the three thresholds, the model has a high specificity, indicating that it is highly effective at identifying survivors. However, the sensitivity is relatively low, suggesting that the model has difficulty identifying actual non-survivors.

Question 4

Create plots of ROC curves for the chosen model. Calculate the AUC for the full model and two more models. Choose the best model based on the AUC.

Approach:

- Understand ROC Curves and AUC:
  - ROC curves are used to evaluate the performance of classification models by plotting the true positive rate against the false positive rate.
  - The AUC (Area Under the Curve) summarizes the ROC curve, with higher values indicating better model performance.
- Implementation:
  - Compare the AUC values for the full model and two additional models and the model with the highest AUC is considered the best model for (in-sample) predicting survival probabilities.
  - The two additional models used for comparison are derived by removing the highest significant variable (ConsciousnessLevel) and the least significant variable (Blood Pressure) from the full model that includes all the significant predictors available in the dataset as mentioned in Q1.
  - Thereby, we can compare the significance of these variables in predicting survival probabilities by observing the change in AUC values using ROC curves.
    - Full Model with only significant predictors from the dataset:

    $$Survival \sim ConsciousnessLevel + TypeOfAdmission + Age + Cancer + BloodCarbonDioxide + BloodPH + Patient + BloodPressure$$

    - Model A: Exclude Blood Pressure from the full model.

    $$Survival \sim ConsciousnessLevel + TypeOfAdmission + Age + Cancer + BloodCarbonDioxide + BloodPH + Patient$$

    - Model B : Exclude ConsciousnessLevel from the full model.

    $$Survival \sim TypeOfAdmission + Age + Cancer + BloodCarbonDioxide + BloodPH + Patient + BloodPressure$$

Results:

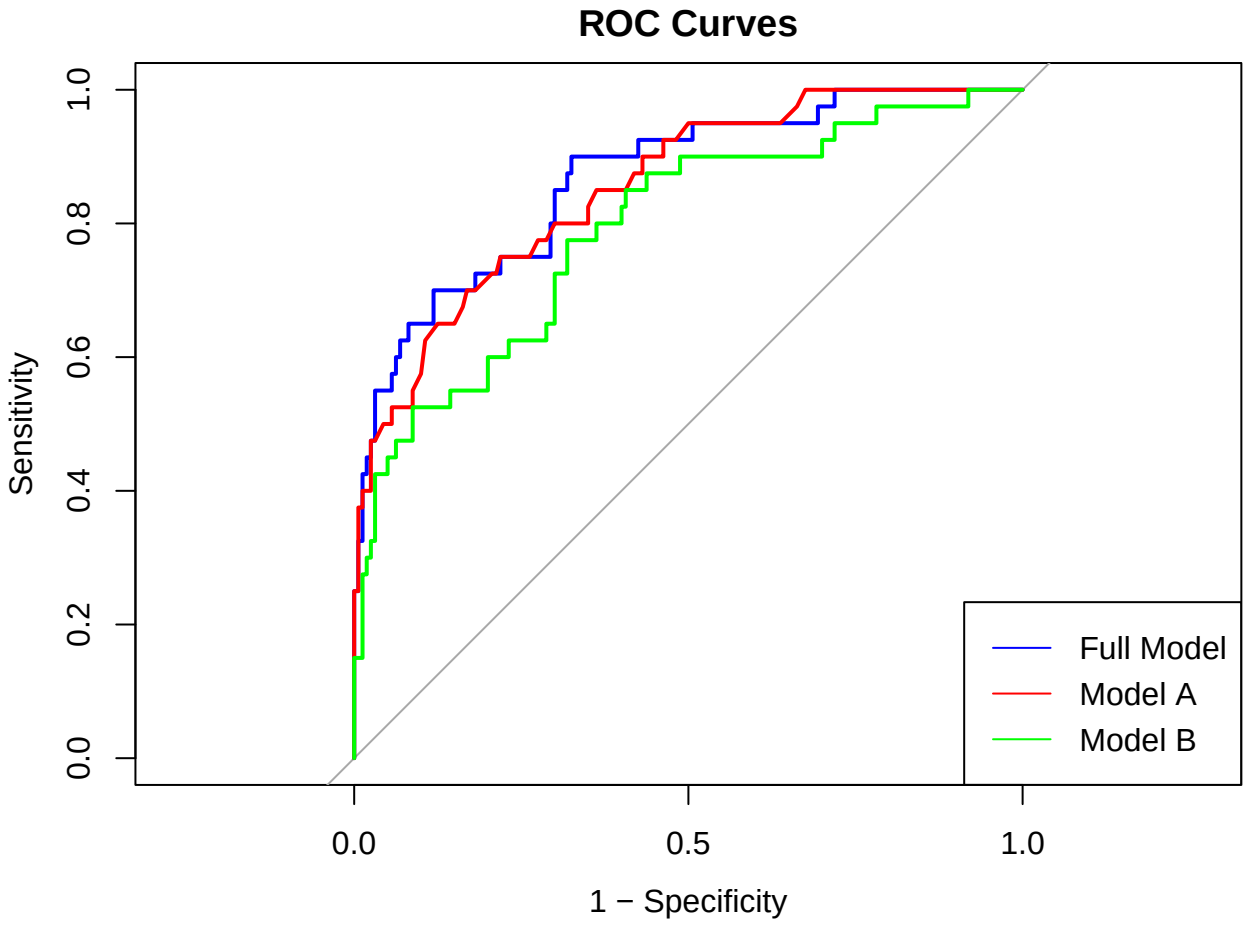


Table 8: AUC Values

Model	AUC
Full Model	0.8678125
Model A	0.8542188
Model B	0.7917188

Model Selection:

- The full model has the highest AUC of 0.8678, indicating that it has a better predictive performance for survival probabilities compared to the other models.
- Model A, which excludes Blood Pressure, has an AUC of 0.854 is quite close to the full model, whereas Model B, which excludes ConsciousnessLevel, has an AUC of 0.792, which has a bigger difference with full model compared to the Model A.
- This confirms our assumption that Blood Pressure does not significantly contribute to the patient survival prediction, as its exclusion does not significantly affect the AUC of the model, whereas ConsciousnessLevel is a significant predictor of survival, and its exclusion has a more significant impact on the model's predictive performance.
- The AUC of all three models are much higher than 0.5 (random guessing) and can be considered effective for predicting survival probabilities, with the full model being the best among them.

Question 5

Perform Leave One Out Cross Validation (LOOCV) for the above full model and the additional models. Calculate the LOOCV-adjusted AUC for the three models and compare with the results from question 4. Which model indicates the best predictive performance?

Approach:

- Understand Leave One Out Cross Validation (LOOCV):
  - LOOCV is a technique for assessing the predictive performance of a model by training on all but one observation and testing on the left-out observation.
  - The AUC values from LOOCV provide an estimate of the model's performance on unseen data.
  - The three models used in question 4 are evaluated using LOOCV to determine the best predictive performance.

Results:

Table 9: LOOCV-Adjusted AUC Values

Model	AUC
Full Model	0.8203125
Model A	0.8153125
Model B	0.7503125

Interpretation:

- According the LOOCV-adjusted AUC values, Full Model has the highest AUC of 0.82, followed by the Model A with an AUC of 0.815, and Model B with an AUC of 0.75.
- LOOCV-adjusted AUC is a more reliable AUC as it is calculated by leaving out one observation at a time and predicted using the model trained on the remaining data. The decrease in the AUC values compared to the AUC values from the previous analysis is expected as LOOCV provides a more realistic estimate of the model's performance on unseen data and helps to avoid overfitting.
- Hence, we can conclude that the Full Model has the best predictive performance among the three models based on the LOOCV-adjusted AUC values.
- Model A can be considered as the second-best model, as it has a slightly lower AUC compared to the Full Model. This means that even though Blood Pressure may not be a highly significant variable, it is still needed to improve the predictive performance of the model.
- The AUC of Model B has further decreased compared to the previous analysis, suggesting that there is a bigger impact of the ConsciousnessLevel of the patient in predicting their survival than previously thought. This signifies that ConsciousnessLevel of the patient is very essential to predict the survival probability.

Exercise 4:2 (Decision Tree)

Question 1

Fit a decision tree model to the ICU data using the rpart package. Use the same predictors as in the multiple logistic regression model. Plot the decision tree and interpret the results.

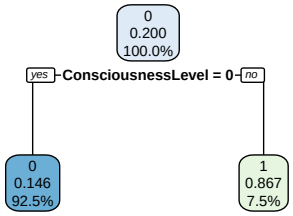
Approach:

To classify the survival status of patients admitted to the ICU, a decision tree model is applied. Decision trees offer an interpretative way to identify key predictors and their thresholds that affect survival. For this task, the following steps were followed:

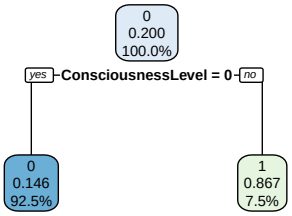
1. **Model Fitting**:
   - A decision tree model was fit using the ICU dataset.
   - The response variable (`v2`) indicates survival (0 = survived, 1 = not survived).
   - The algorithm was allowed to automatically determine the most suitable variables for splitting at each node based on the Gini Index or Information Gain criteria.
2. **Parameters Adjusted**:
   - **Splitting criterion**: Both "information" and "gini" criteria were tested to evaluate their effects on splits.
   - **Complexity parameter (`cp`)**: Different values were used to control the depth of the tree and prevent overfitting.
3. **Visualization**:
   - Tree diagrams were created using the `rpart.plot` package to assess the structure and interpretation of the models.
4. **Tree Selection**:
   - The selection criteria for a good decision tree includes:
     - **Interpretation**: The tree should be easy to interpret and explain. Trees with fewer splits are preferred.
     - **Relevance of Splits**: Identify important predictors of survival.
     - **Clinical Relevance**: For ICU patients, identifying key factors affecting survival is crucial.

Results:

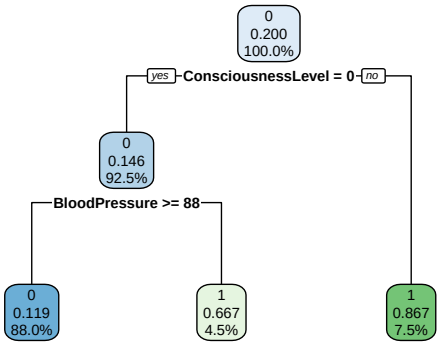
Gini, cp = 0.1



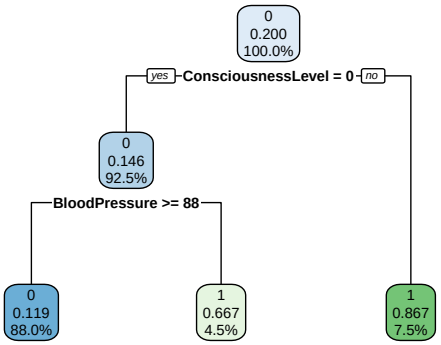
Info Gain, cp = 0.1



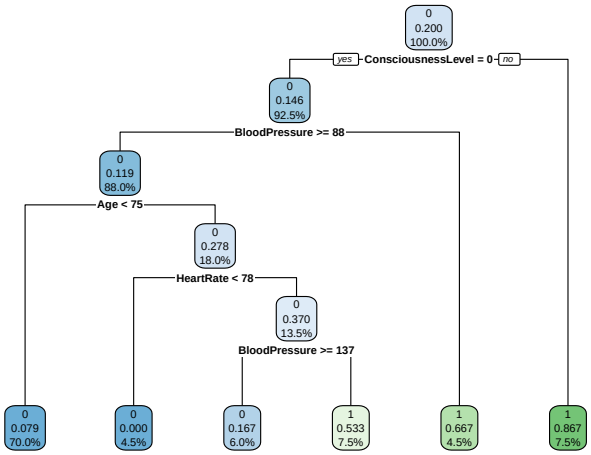
Gini, cp = 0.01



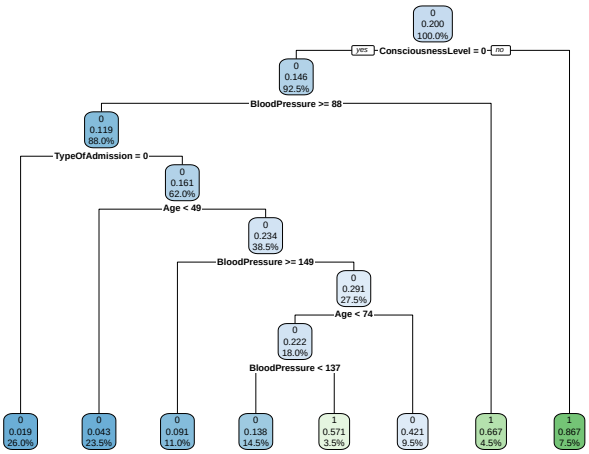
Info Gain, cp = 0.01



Gini, cp = 0.001



Info Gain, cp = 0.001



Interpretation and selection:

Splitting Criteria:

- **Gini index** tends to favor splits with balanced distributions of the target classes, which may better handle uncertainty in clinical settings.
- **Information gain** focuses on maximizing the reduction in entropy, which could lead to splits that are sightly more biased but more efficient for some datasets.

Simplicity and Predictors:

- All trees highlight Consciousness Level as the primary predictor of survival.
- Trees with cp = 0.1 are the simplest, making them easy to interpret, but they do not refine predictions beyond the primary split.
- Trees with cp = 0.01 added a split on Blood Pressure, providing additional insights into survival probabilities. The trees are identical for both Gini and Information criteria. This suggests that the model is not overfitting and the relationships in the data are consistent.
- Trees with cp = 0.001 are the most complex tested trees, with multiple splits that may lead to overfitting. The additional splits vary for Gini and Information criteria, indicating that the choice of splitting method affects the tree structure. While Gini chose Age and HeartRate for further splits, Information chose TypeOfAdmission and Age.

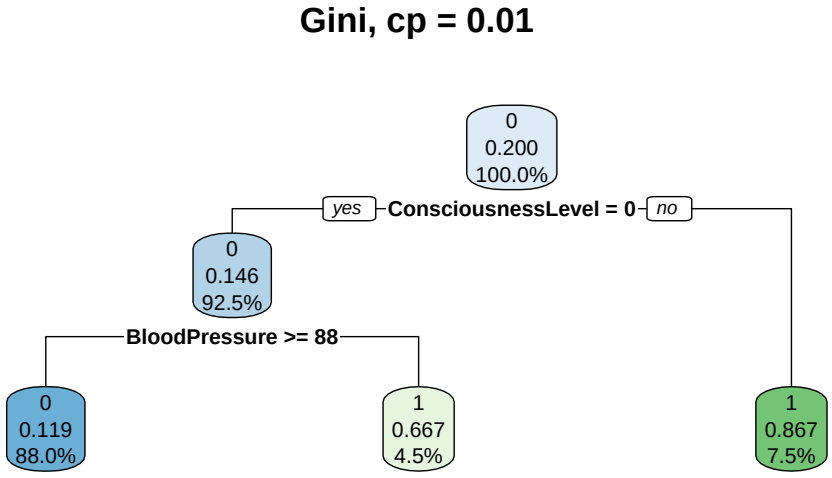
Simplicity vs refinement:

- a tree with cp = 0.01 is a good compromise, as it adds refinement without sacrificing interpretation, and without risking overfitting.
- Among the splitting methods with same cp, the preference depends on the dataset:
  - If the data has imbalanced classes, gini might be better.
  - If entropy reduction is preferred, Information could perform better.

Since the trees with cp = 0.01 are the most balanced in terms of simplicity and refinement, the tree with Gini Index and cp = 0.01 is selected as a good decision tree for predicting survival probabilities in the ICU dataset as it offers a good trade-off between interpretation and predictive performance and is identical to the tree with Information Gain and cp = 0.01, indicating robustness in the model selection. This tree is simple enough for clinical use while providing actionable insights.

Conclusion:

Good Decision tree: Gini, cp=0.01.



Variables of Interest:

- **Consciousness level**: The primary variable splitting the data in both trees, indicating its importance for predicting survival.
- **Blood pressure**: A secondary variable used in the second tree, capturing further distinctions in survival likelihood.

Interpretation:

At the root node, the proportion of patients who do not survive (class = 1) is 0.800 (80%), while the proportion who survive (class = 0) is 0.200 (20%).

- **Consciousness Level > 0:** Among patients who are not conscious (Consciousness Level > 0), the probability of not surviving (class = 1) is 0.867 (86.7%), and the probability of surviving (class = 0) is 0.133 (13.3%).
- **Consciousness Level = 0:** Among conscious patients (Consciousness Level = 0), the probability of surviving (class = 0) is 0.146 (14.6%), and the probability of not surviving (class = 1) is 0.854 (85.4%).
- **Blood Pressure >= 88:** Among conscious patients with blood pressure greater than or equal to 88, the probability of surviving (class = 0) is 0.119 (11.9%), and the probability of not surviving (class = 1) is 0.881 (88.1%).
- **Blood Pressure < 88:** Among conscious patients with blood pressure less than 88, the probability of surviving (class = 0) is 0.333 (33.3%), and the probability of not surviving (class = 1) is 0.667 (66.7%).

Question 2

In this question we have to assess how well the chosen tree predicts for Survival, using the AUC metric, to measure predictive performance. Finally we will compare the AUC of the decision tree with the AUC of the logistic regression model.

Approach:

- **Calculate Predicted Probabilities**: Use the `predict()` function to calculate the predicted probabilities for the chosen decision tree model.
- **Calculate AUC**: Use the `roc()` function from the `pROC` package to calculate the AUC for the decision tree model.
- **Logistic Regression Model**: Compare the AUC of the decision tree with the AUC of the logistic regression model to assess predictive performance.

Results:

Model	AUC
Decision Tree (Gini, cp = 0.01)	0.7239844
Logistic Regression	0.8678125

1. **Decision Tree AUC**: The decision tree model with Gini Index and cp = 0.01 has an AUC of 0.724, indicating a good predictive performance.
2. **Logistic Regression AUC**: The logistic regression model has an AUC of 0.8678, which is higher than the decision tree model.

Interpretation:

1. **Decision Tree vs. Logistic Regression**: The decision tree is slightly less accurate but offers better interpretation making it a useful option when simplicity is a priority. (For real-time clinical applications)
2. **Logistic Regression**: Logistic regression achieves higher predictive accuracy, which may be preferable if accuracy outweighs interpretation.

In conclusion based on the AUC values, the logistic regression model is the better-performing model for predicting survival. However, the decision tree remains as a valuable option for its interpretation and simplicity, especially in clinical settings where understanding the decision-making process is crucial.

Question 3

Calculate a LOOCV-corrected AUC for the decision tree model and comment on the result.

Approach:

- **LOOCV for Decision Tree**:

  Perform Leave One Out Cross Validation (LOOCV) for the decision tree model to calculate the LOOCV-adjusted AUC.

- We have chosen the above mentioned decision tree model with Gini Index and cp = 0.01 for this analysis.

Results:

```
##
## Call:
## roc.default(response = data_ca4$Survival, predictor = predprob_tm_LOOCV)
##
## Data: predprob_tm_LOOCV in 160 controls (data_ca4$Survival 0) > 40 cases (data_ca4$Survival 1).
## Area under the curve: 0.5595
```

Interpretation:

- The LOOCV-adjusted AUC is 0.5595 which is a significant decrease from the unadjusted AUC of 0.724. This indicates that the decision tree model may not generalize well to new data and could be overfitting the training data.
- The decrease in AUC suggests that the decision tree model may not be as effective in predicting survival probabilities on unseen data as initially thought. This highlights the importance of assessing model performance using cross-validation techniques to avoid overfitting and ensure robust predictions.
- The LOOCV-adjusted AUC is just above 0.5, indicating that the model's predictive performance is only slightly better than random guessing. This suggests that the decision tree model may not be suitable for predicting survival probabilities in this dataset and may require further refinement or a different modeling approach. This can be linked to overfitting, as the model may have learned the noise in the training data rather than the underlying patterns. Further refinements may be necessary to improve predictive performance.
- The decision tree with Gini Index and cp = 0.001 or Information Gain and cp = 0.001 could be explored instead to improve predictive performance and generalization to new data.
