Assignment IV

Multiple Logistic Regression and Decision Tree

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

2024-12-16

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:
  - Load the dataset and rename variables for clarity.
  - Combine categories for categorical variables if necessary.
- Model Fitting:
  - Fit an empty logistic regression model and a full logistic regression model to be used in the stepwise selection.
  - Here, Logistic Regression Model is used to predict the binary outcome of probability of not surviving based on multiple predictors.
- Model Selection Process:
  - Use a stepwise selection with AIC to identify a parsimonious model.
- Analysis of the Final Model:
  - Extract coefficients, odds ratios, and their 95% confidence intervals for significant variables.
- Interpretation:
  - Interpret the results from the AIC, odds ratios, and confidence intervals, to determine the best model.

Results:

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

```
##
## Call:
## glm(formula = Survival ~ ConsciousnessLevel + TypeOfAdmission +
##     Age + Cancer + Patient + BloodCarbonDioxide + BloodPH + BloodPressure,
##     family = binomial, data = data_ca4)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.7353420  1.6104573  -2.940 0.003278 **
## ConsciousnessLevel  2.6208042  0.6859650   3.821 0.000133 ***
## TypeOfAdmission     3.0547147  0.9339217   3.271 0.001072 **
## Age                 0.0385864  0.0133655   2.887 0.003889 **
## Cancer              2.3388380  0.8671971   2.697 0.006997 **
## Patient            -0.0020714  0.0008783  -2.359 0.018345 *
## BloodCarbonDioxide -2.4646334  1.0619854  -2.321 0.020299 *
## BloodPH             2.0884994  0.9031831   2.312 0.020757 *
## BloodPressure      -0.0099893  0.0070360  -1.420 0.155682
## ---
## 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: 130.19  on 191  degrees of freedom
## AIC: 148.19
##
## Number of Fisher Scoring iterations: 6
```

Model selection:

- The final logistic regression model includes the following variables: ConsciousnessLevel, TypeOfAdmission, Age, Cancer, Patient, 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
  - BloodCarbonDioxide,
  - BloodPH
- Patient variable was exlcluded from the final model as the ID code was not significant.
- BloodPressure was also excluded from the final model as it was not significant according to the p-value.

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...
```

```
##                                Variable    OddsRatio     CI.Lower    CI.Upper
## (Intercept)                (Intercept)  0.008779445 0.0002996746   0.1795376
## ConsciousnessLevel ConsciousnessLevel 13.746774142 4.3144886742  65.2810381
## TypeOfAdmission       TypeOfAdmission 21.215131928 4.3561881520 189.1540175
## Age                               Age  1.039340550 1.0141827727   1.0692793
## Cancer                         Cancer 10.369181076 1.9513659807  66.5483329
## Patient                       Patient  0.997930702 0.9961324780   0.9995944
## BloodCarbonDioxide BloodCarbonDioxide  0.085040009 0.0080634712   0.5539141
## BloodPH                       BloodPH  8.072791946 1.4001269889  53.6032965
```

1. ConsciousnessLevel:
    - Odds Ratio: 13.75 (CI: 4.31-65.28)
    - Patients who are unconscious or in a coma have a significantly higher probability of not surviving compared to those who are conscious.
2. TypeOfAdmission:
    - Odds Ratio: 21.25 (CI: 4.36-189.15)
    - Acute admissions are associated with a significantly higher probability of not surviving compared to non-acute admissions.
3. Age:
    - Odds Ratio: 1.04 (CI: 1.01-1.07)
    - For each additional year of age, the odds of not surviving increase by 4%.
4. Cancer:
    - Odds Ratio: 10.37 (CI: 1.95-66.54)
    - Patients with cancer have over 10 times higher odds of not surviving compared to those without cancer.
5. BloodCarbonDioxide:
    - Odds Ratio: 0.085 (CI: 0.008-0.55)
    - Lower blood carbon dioxide levels significantly reduce the odds of not surviving.
6. BloodPH:
    - Odds Ratio: 8.07 (CI: 1.4-53.60)
    - Lower blood pH levels significantly increase the odds of not surviving.

Conclusion:

The selected model indicates that factors such as consciousness level, type of admission, age, cancer, blood carbon dioxide, and blood pH are significant predictors of survival. Patients who are unconscious, have acute admissions, are older, have cancer, and have abnormal blood gas levels are at higher risk of not surviving. These results can help identify high-risk patients and improve treatment strategies to increase survival rates.

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 = 2.2064, df = 6, p-value = 0.8998
```

Interpretation:

1. The p-value of 0.8998 tells us not reject the null hypothesis that the model fits the data well.

Conclusion:

The Hosmer-Lemeshow test indicates that the model fits the data well. The observed event rates are consistent with the expected probabilities predicted by the model. This suggests that the model is a good fit for the data and can be used to make accurate predictions about survival probabilities.

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 1: Confusion Matrix at threshold= 0.3

	Predicted Not Survived	Predicted Survived
Actual Not Survived	25	15
Actual Survived	18	142

Table 2: Confusion Matrix at threshold= 0.5

	Predicted Not Survived	Predicted Survived
Actual Not Survived	18	22
Actual Survived	5	155

Table 3: Confusion Matrix at threshold= 0.7

	Predicted Not Survived	Predicted Survived
Actual Not Survived	10	30
Actual Survived	2	158

Table 4: Performance Metrics at threshold= 0.3

Metric	Value
Accuracy	0.8350
Sensitivity	0.6250
Specificity	0.8875

Table 5: Performance Metrics at threshold= 0.5

Metric	Value
Accuracy	0.86500
Sensitivity	0.45000
Specificity	0.96875

Table 6: Performance Metrics at threshold= 0.7

Metric	Value
Accuracy	0.8400
Sensitivity	0.2500
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 0.625 indicates that the model correctly identified 62.5% of the actual non-survivors.

Specificity: The specificity of 0.8875 suggests that the model correctly identified 88.75% of the actual survivors.

Threshold = 0.5:

Accuracy: The model has an accuracy of 0.865, meaning that it correctly predicted 86.5% of the cases.

Sensitivity: The sensitivity of 0.45 indicates that the model correctly identified 45% of the actual non-survivors.

Specificity: The specificity of 0.96875 suggests that the model correctly identified 96.9% of the actual survivors.

Threshold = 0.7:

Accuracy: The model has an accuracy of 0.84, meaning that it correctly predicted 84% of the cases.

Sensitivity: The sensitivity of 0.25 indicates that the model correctly identified 25% 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 predicting survival probabilities.
  - The two additional models used for comparison with the full model are derived by removing a significant variable and a non-significant variable from the full model (such as ConsciousnessLevel and Blood Pressure 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:

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

    - 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:

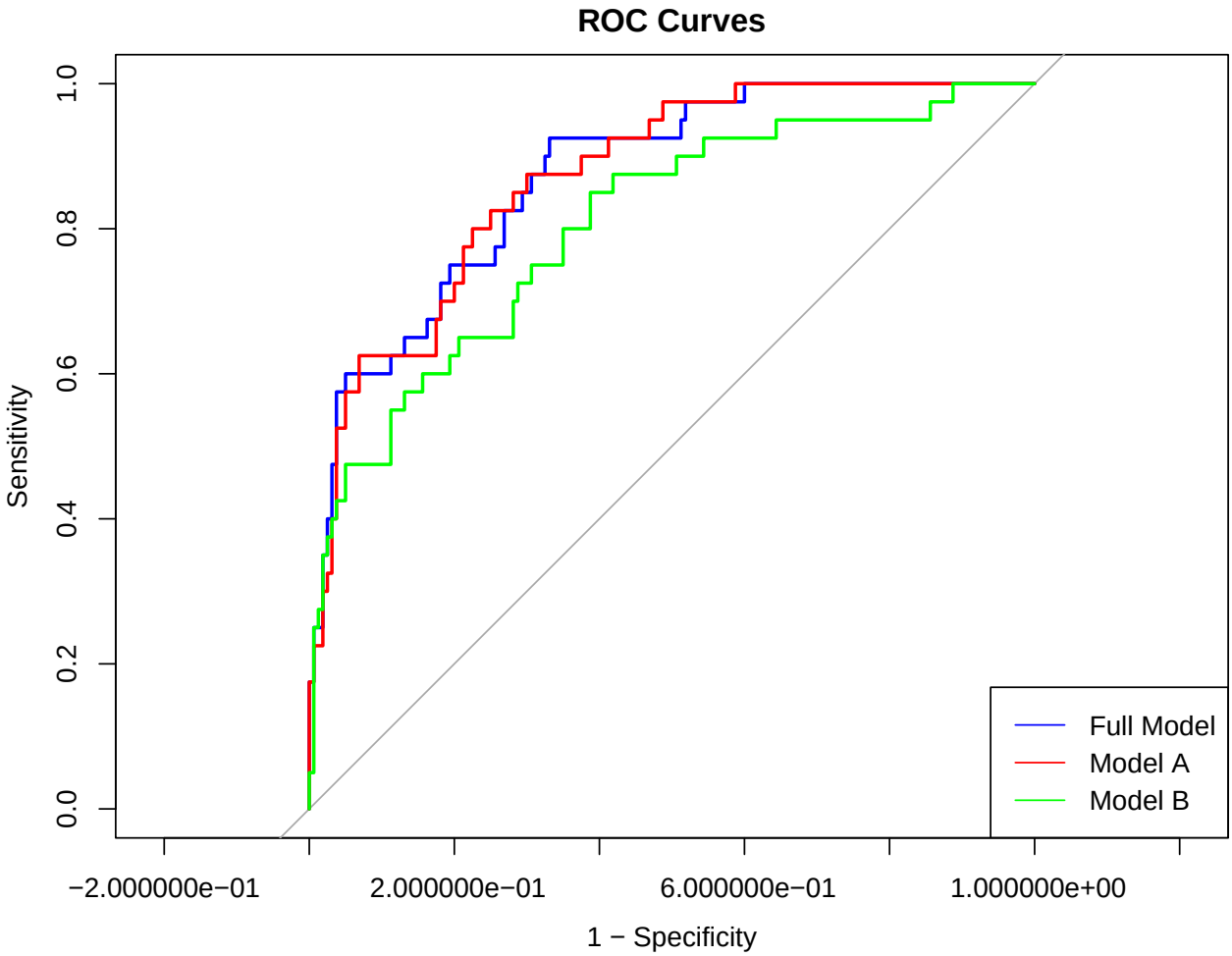


Table 7: AUC Values

Model	AUC
Full Model	0.8729687
Model A	0.8710938
Model B	0.8043750

Model Selection:

- The full model has the highest AUC of 0.873, 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 0.871 is quite close to the full model, whereas Model B, which excludes ConsciousnessLevel, has an AUC of 0.804, which has a bigger difference with full model compared to the Model B.
- This can be interpreted as 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 8: LOOCV-Adjusted AUC Values

Model	AUC
Full Model	0.8193750
Model A	0.8242187
Model B	0.7565625

Interpretation:

- According the LOOCV-adjusted AUC values, Model A has the highest AUC of 0.824, followed by the Full Model with an AUC of 0.819, and Model B with an AUC of 0.756.
- 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.
- Model A having a higher AUC than full model indicates that the full model was overfitted in the previous analysis, and Model A is the best model for predicting survival probabilities.
- This means that Blood Pressure is not required to be added as a explanatory variable while predicting the survival probabilities of the patients and removing it from the model improves the predictive performance.
- The AUC of Model B has further decreased compared to the previous analysis, suggesting that the previous Model B was also overfitted and 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.

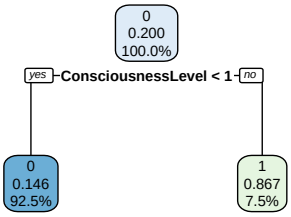
Approach:

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:

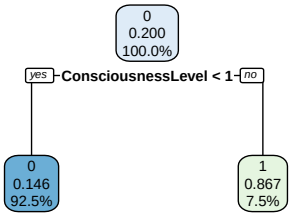
1. **Model Fitting**:
   - A decision tree model was fit using the ICU dataset.
   - The response variable (`v2`) indicates survival (0 = survived, 1 = not survived).
   - Predictor variables include age, blood pressure, consciousness level, and other clinical factors.
2. **Parameters Adjusted**:
   - **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 interpretability of the models.
4. **Tree Selection**:
   - The selecion criteria for the best tree includes:
     - **Interpretability**: 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.
5. **Variables of Interest**:
   - **`v21` (Consciousness level)**: The primary variable splitting the data in both trees, indicating its importance for predicting survival.
   - **`v11` (Blood pressure)**: A secondary variable used in the second tree, capturing further distinctions in survival likelihood.

Results:

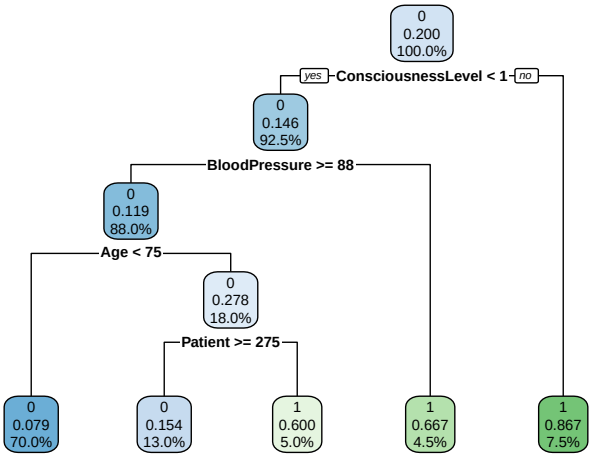
Gini, cp = 0.1



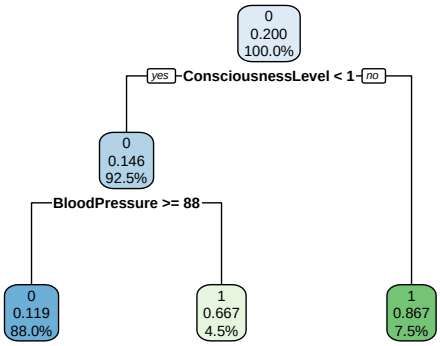
Info Gain, cp = 0.1



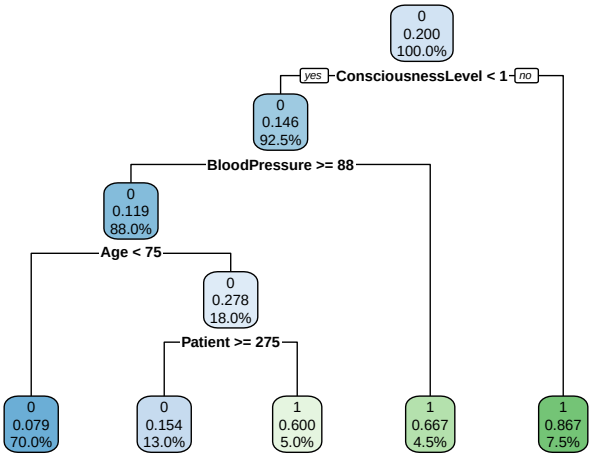
Gini, cp = 0.01



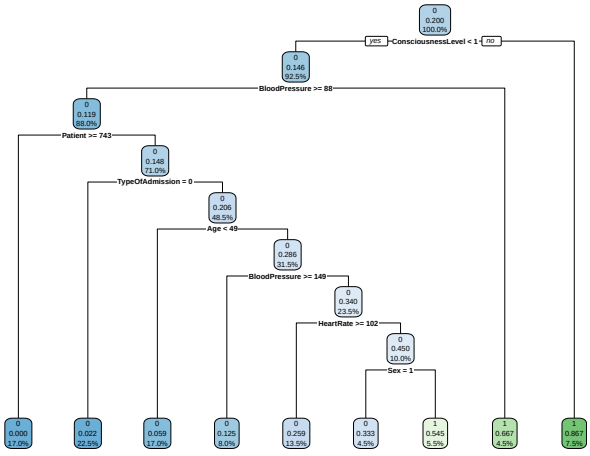
Info Gain, cp = 0.01



Gini, cp = 0.001



Info Gain, cp = 0.001



Interpretation and selection:

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 add an extra split, offering more detailed classification without adding excessive complexity.
- Trees with cp = 0.001 are the most complex, with multiple splits that may lead to overfitting.

Compare 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 vs refinement

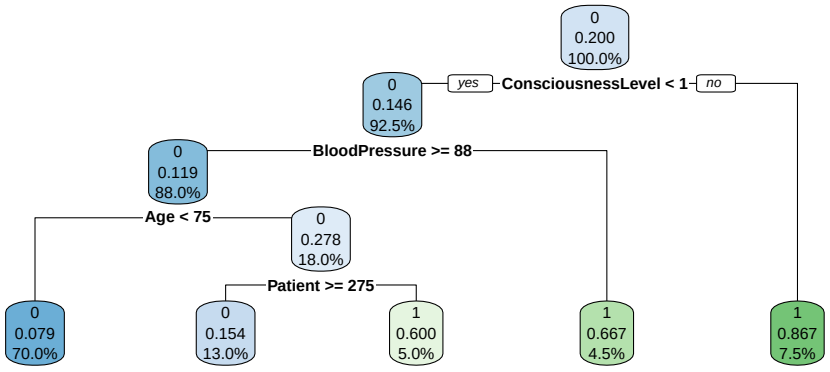
- a tree with cp = 0.01 is a good compromise, as it adds refinement without sacrificing interpretation, and without risking overfitting.
- Among the splitting methods with same cp, the preference depends on the dataset:
  - If the data has imbalanced classes, gini might be better.
  - If entropy reduction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

Conclusion:

Based on the uploading plots and logic:

- **Best tree:** Gini, cp=0.01. The tree with cp = 0.01 using Gini Index was selected because it offers a good trade-off between interpretation and predictive refinement. Consciousness Level remains the primary driver of survival predictions, while the additional split on Blood Pressure refines subgroups with varying survival probabilities. This tree is simple enough for clinical use while providing actionable insights.

Gini, cp = 0.01



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)	0.8114844
Logistic Regression	0.8729687

1. **Decision Tree AUC**: The decision tree model with Gini Index and cp = 0.01 has an AUC of 0.8118, indicating a good predictive performance.
2. **Logistic Regression AUC**: The logistic regression model has an AUC of 0.8729, 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 best 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: 0.6284
```

Interpretation:

- The LOOCV-adjusted AUC is 0.6284 which is a significant decrease from the unadjusted AUC of 0.8115. This indicates that the decision tree model may be overfitting the data, leading to a lower predictive performance on unseen data.
- The decision tree model may benefit from additional tuning or simplification to enhance its predictive performance and avoid overfitting.
- This could be due to the presence of the parameters BloodPressure and Patient ID in the model, which may not be significant predictors of survival and could be contributing to overfitting.