Computer Assignment 4 Part 2

Gabriel Lindqvist, Jiahui Li

2024-12-13

```
##
##   1   2   3
## 175  15  10
```

```
##
##   0   1
##  25 175
```

```
##
##   0   1   2
## 185   5  10
```

```
##
##   0   1
## 185  15
```

Exercise 4:2 Decision Tree

Task 1

We start off by reporting the changes in the decision trees while adjusting the `cp`, for the two splitting methods. This was done in the following way:

1. Set `cp = 0.1` as our starting value.
2. Generate and plot the two decision trees for each splitting method.
3. Write down the observations.
4. Lower the `cp` value and go to step 2.

Observations and Choosing Tree:

- `cp = 0.1`: Both splitting methods generate the same tree and the height of the tree is 1, with `v21` (Consciousness level) as the root.
- `cp = 0.05`: Both splitting methods generate the same tree and the height of the tree is now 2, where the left node has two children `v11 >= 88` (blood pressure).

- `cp = 0.005`: Information splitting method still has the same tree as previously. Gini splitting method has grown larger, including the predictors: `v3 < 75` (age), `v12 < 78` (heart rate) and `v11 >= 137` as a leaf.
- `cp = 0.001`: Information splitting method has generated the largest tree and has now included the predictor `v14` (Type of Admission). The tree generated by using gini splitting method remains the same.
- `cp = 0.0005`: No notable changes.
- `cp = 0.0001`: Again, no notable changes, indicating that no further reduction in complexity parameter will let the tree grow larger.

```
## Warning: package 'rpart' was built under R version 4.2.3
```

```
## Warning: package 'rpart.plot' was built under R version 4.2.3
```

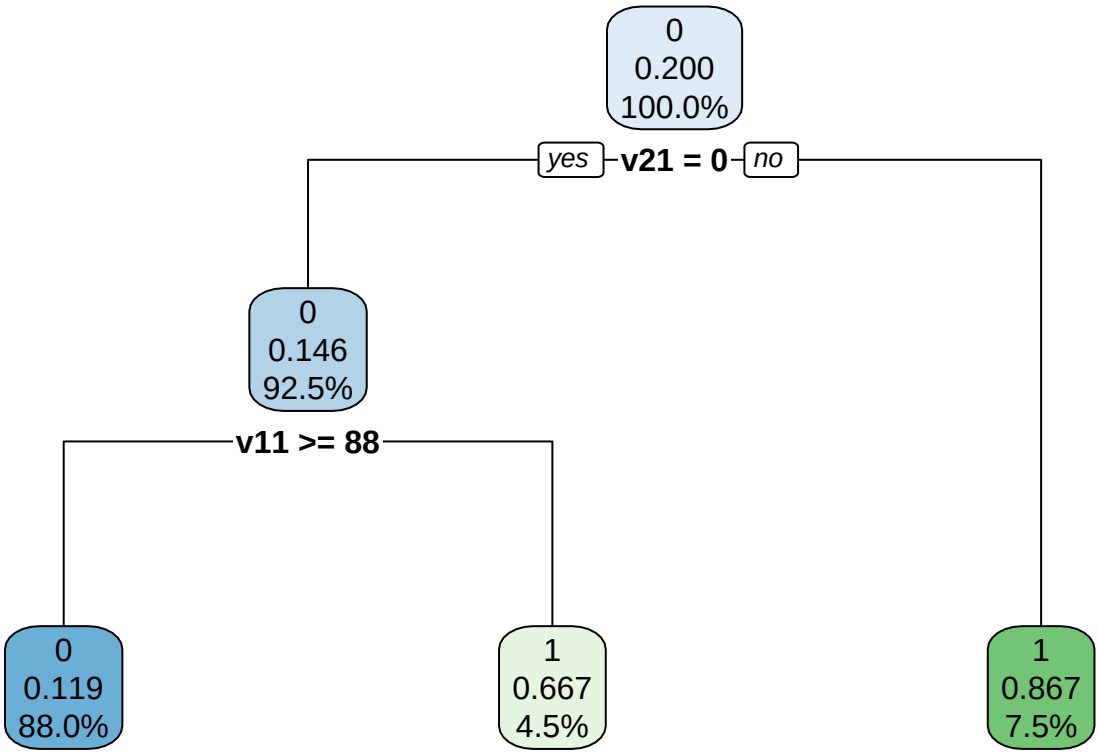
To decide on a "good" decision tree (without computing AUC or LOOCV), intuitively, we compare the predictors selected in the decision tree with the logistic regression chosen in the previous exercise. We also weigh in interpretation, with simpler decision trees are in favor.

A key difference from the observations is that the gini splitting method has included **Heart rate**, while excluded in the chosen logistic regression model. This does not necessarily imply that the predictor has no influence in patient survival, due to AIC-based model selection. Furthermore, **Heart rate** is treated as binary, while its continuous in the logistic regression model.

In contrast, the information splitting method is consistent with its inclusion of predictors (relative to the chosen logistic regression model). That is, all predictors found in the largest tree are also found in the chosen logistic regression. However, this tree is rather complex with a height of 7. Thus, we decided to go with the simpler tree that only includes **Conciousness level** and **Blood pressure**.

Interpretation of Chosen Tree Model

First step is to plot the chosen decision tree.



In order to interpret the illustrated decision tree, we need to understand the structure of each node.

1. **Top Value (Binary Code)**: This is the so called *majority class*. That is, the decision tree assigns a class to each node based on the most frequent (majority) class among the observations in that node. If the majority of observations in a node have `v2 = 0` (survived), then the top value will be `0`. Conversely, if the majority of observations in a node have `v2 = 1` (did not survive), then the top value will be `1`.
2. **Middle Value (Proportion)**: Reflects the proportion of observations in the node that belong to class `1` (did not survive). The majority class is determined based on the proportion. For example, the right child of the root has a middle value of `0.867`, meaning 86.7% did not survive. Thus the top value is `1`
3. **Bottom Percentage**: Represents the proportion of the entire dataset that ends up at that specific node. This percentage is distributed between siblings. For example, `92.5%` of the total observations fall into the left child of the root, while `7.5%` fall into the right child of the root.

With that in mind, we can deduce the following:

- **Consciousness level**: This factor serves as the root in the decision tree, meaning it is the most important factor. Being unconscious (represented by the right child of the root), is strongly associated with non-survival.
- **Blood Pressure**: Represented as the left child of the root. For patients that are conscious (`v21 = 0`), blood pressure is a key predictor for survival outcome. High blood pressure ($\geq 88\ mm\ Hg$) is associated with better outcomes, while low blood pressure ($< 88\ mm\ Hg$) indicates a higher risk of non-survival.

Task 2

In this task, we want to investigate the predictive power of our chosen tree model (AUC without cross validation) and compare the results with that for the multiple logistic regression models in the previous exercise.

For curiosity, we calculate the AUC for the largest decision tree model that information splitting method generated

```
## Warning: package 'pROC' was built under R version 4.2.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## [1] "AUC for Chosen Tree Model: 0.724"
```

```
## [1] "AUC for Largest Tree Model: 0.8891"
```

Comparing AUC between the chosen tree model and the logistic regression models, the chosen tree model ranks at the bottom regarding predictive power. However, the largest tree model edges all three multiple logistic regression models, with the highest calculated AUC. Thus, the largest tree model generated with information splitting method performs best based on AUC.

Thoughts Regarding The Results

Based on intuition, the results are not surprising. The chosen tree model is simplistic, including only two predictors, likely limits its ability to capture the full complexity of the relationship found in the data. The largest tree model includes all key predictors identified as significant in the AIC-based logistic regression model. In that sense, the decision tree has access to the same information as the logistic regression.

Task 3

In this task, we want to calculate the LOOCV-corrected AUC for our chosen tree model (Again, we include the other generated tree model).

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls > cases
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
## [1] "AUC for Chosen Tree Model: 0.5356"
```

```
## [1] "AUC for Largest Tree Model: 0.8891"
```

The results indicates that our chosen tree model is much worse at predicting new data, compared to the results in Exercise 4:1 Task 5. However, the other tree model outperforms all considered model, emphasizing that a decision tree model should be considered if the goal is to model the best predictive model.