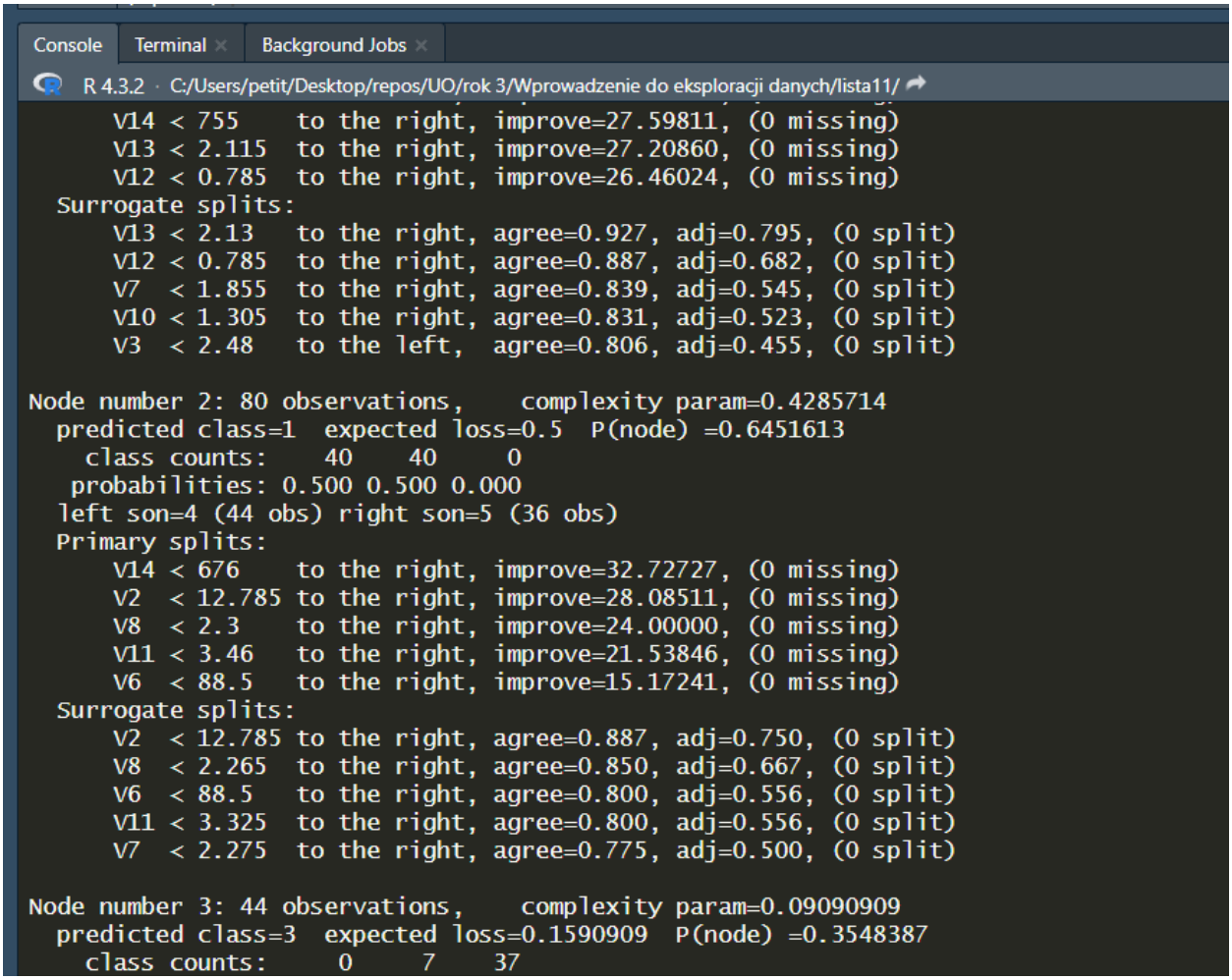
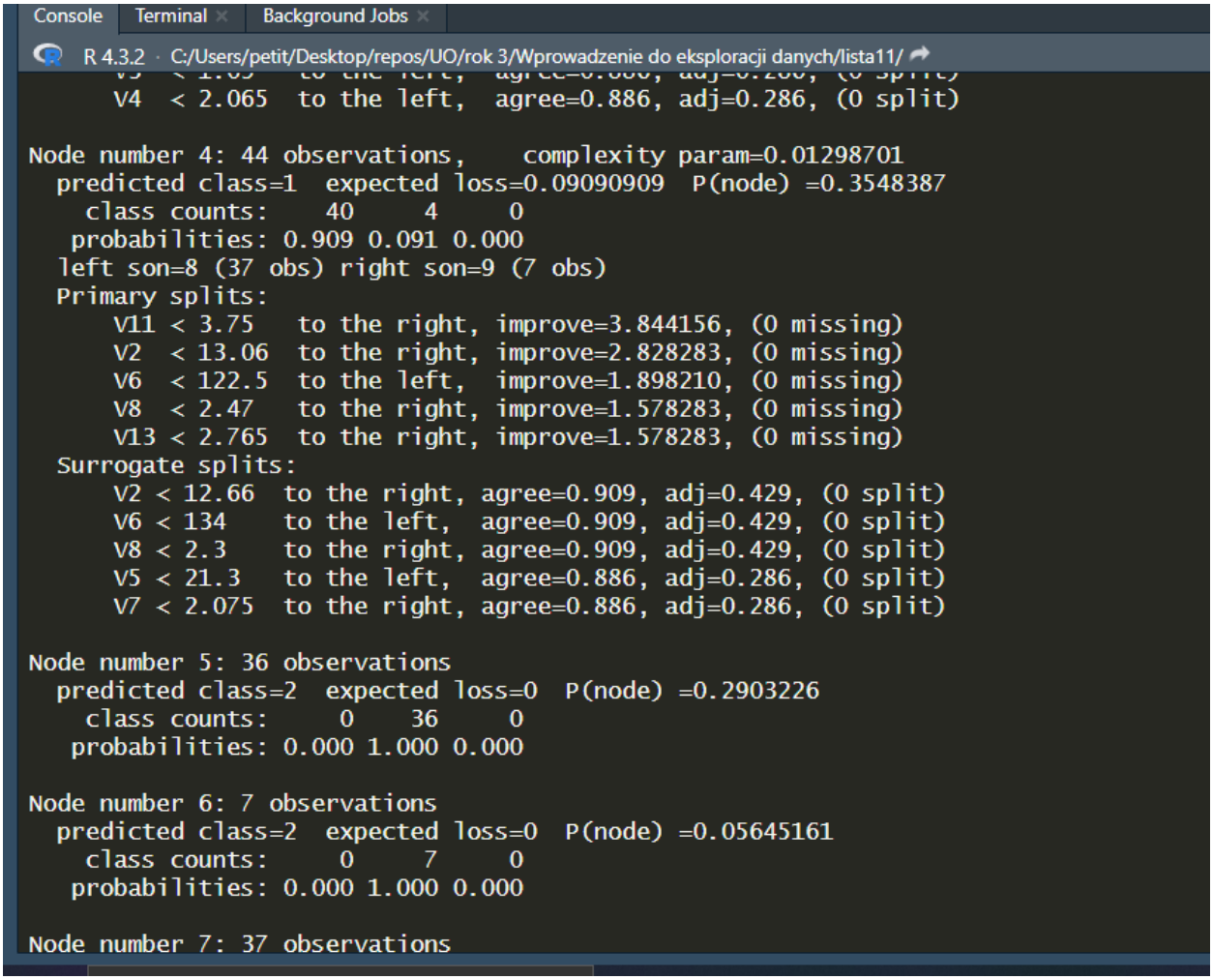
Lista 11

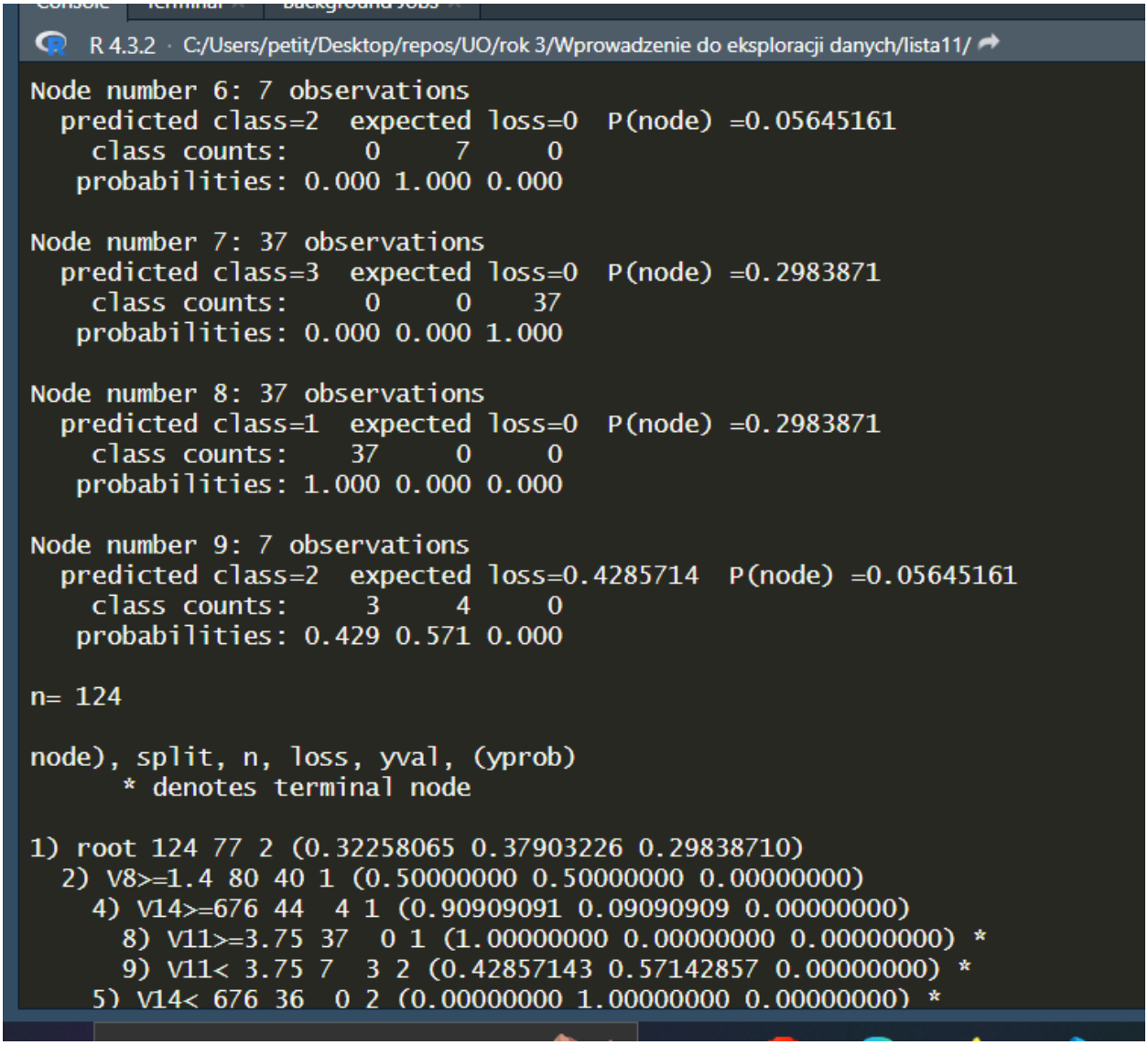
```
> library(rpart)
> library(rpart.plot)
>
> # Ustawienie ścieżki dostępu do danych
> path <- "C:/Users/petit/Desktop/repos/UO/rok 3/Wprowadzenie do eksploracji
danych/lista11"
> setwd(path)
>
> # Wczytanie danych
> wine <- read.csv('wine/wine.data', header = FALSE)
>
> # Podział danych na zbiór treningowy i testowy
> set.seed(123)  # Ustawienie ziarna dla reprodukowalności wyników
> indeksy <- sample(1:nrow(wine), nrow(wine) * 0.7)
> train_data <- wine[indeksy, ]
> test_data <- wine[-indeksy, ]
>
> # Budowa modelu drzewa decyzyjnego z ograniczeniem do 3 poziomów
> tree_model <- rpart(V1 ~ ., data = train_data, method = "class", cp = 0, ma
xdepth = 3)
>
> # Wyświetlenie podsumowania modelu
> print(summary(tree_model))
```

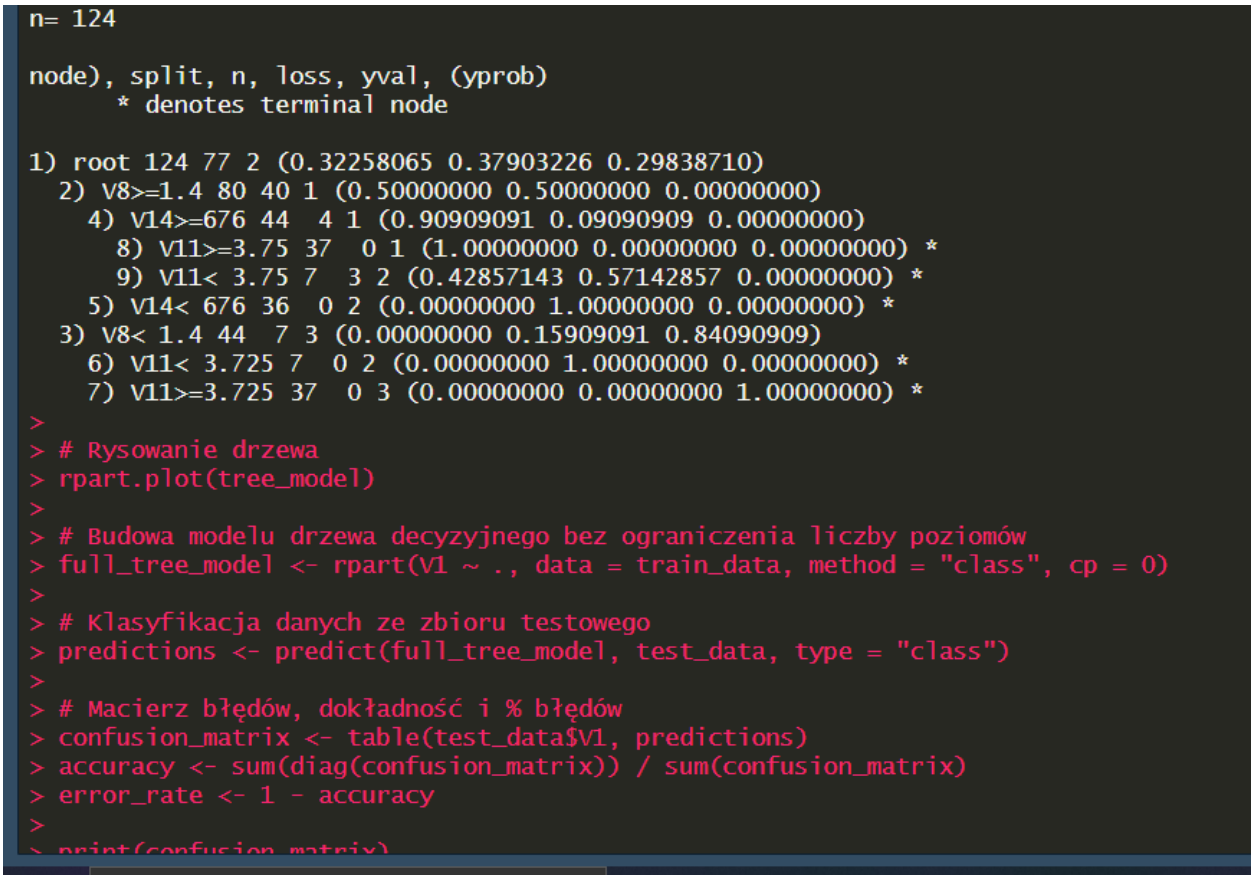
```
> print(summary(tree_model))
Call:
rpart(formula = V1 ~ ., data = train_data, method = "class",
    cp = 0, maxdepth = 3)
  n= 124

          CP nsplit  rel error    xerror       xstd
1 0.42857143      0 1.00000000 1.0000000 0.07016051
2 0.09090909      2 0.14285714 0.2727273 0.05424088
3 0.01298701      3 0.05194805 0.1428571 0.04111819
4 0.00000000      4 0.03896104 0.1688312 0.04430284

Variable importance
 V8  V7 V11 V14  V2 V12 V13  V6  V3 V10  V5  V4
 18  11  11  11  10  10   8   7   6   5   2   1

Node number 1: 124 observations,    complexity param=0.4285714
  predicted class=2  expected loss=0.6209677  P(node) =1
    class counts:    40    47    37
   probabilities: 0.323 0.379 0.298
  left son=2 (80 obs) right son=3 (44 obs)
  Primary splits:
      V8  < 1.4    to the right, improve=30.46921, (0 missing)
      V11 < 3.825  to the left,  improve=27.67748, (0 missing)
      V14 < 755    to the right, improve=27.59811, (0 missing)
      V13 < 2.115  to the right, improve=27.20860, (0 missing)
      V12 < 0.785  to the right, improve=26.46024, (0 missing)
  Surrogate splits:
      V13 < 2.13   to the right, agree=0.927, adj=0.795, (0 split)
      V12 < 0.785  to the right, agree=0.887, adj=0.682, (0 split)
```


Console
Terminal
Background Jobs
R 4.3.2 · C:/Users/petit/Desktop/repos/UO/rok 3/Wprowadzenie do eksploracji danych/lista11/
V14 < 755 to the right, improve=27.59811, (0 missing)
V13 < 2.115 to the right, improve=27.20860, (0 missing)
V12 < 0.785 to the right, improve=26.46024, (0 missing)
Surrogate splits:
V13 < 2.13 to the right, agree=0.927, adj=0.795, (0 split)
V12 < 0.785 to the right, agree=0.887, adj=0.682, (0 split)
V7 < 1.855 to the right, agree=0.839, adj=0.545, (0 split)
V10 < 1.305 to the right, agree=0.831, adj=0.523, (0 split)
V3 < 2.48 to the left, agree=0.806, adj=0.455, (0 split)
Node number 2: 80 observations, complexity param=0.4285714
predicted class=1 expected loss=0.5 P(node) =0.6451613
class counts: 40 40 0
probabilities: 0.500 0.500 0.000
left son=4 (44 obs) right son=5 (36 obs)
Primary splits:
V14 < 676 to the right, improve=32.72727, (0 missing)
V2 < 12.785 to the right, improve=28.08511, (0 missing)
V8 < 2.3 to the right, improve=24.00000, (0 missing)
V11 < 3.46 to the right, improve=21.53846, (0 missing)
V6 < 88.5 to the right, improve=15.17241, (0 missing)
Surrogate splits:
V2 < 12.785 to the right, agree=0.887, adj=0.750, (0 split)
V8 < 2.265 to the right, agree=0.850, adj=0.667, (0 split)
V6 < 88.5 to the right, agree=0.800, adj=0.556, (0 split)
V11 < 3.325 to the right, agree=0.800, adj=0.556, (0 split)
V7 < 2.275 to the right, agree=0.775, adj=0.500, (0 split)
Node number 3: 44 observations, complexity param=0.09090909
predicted class=3 expected loss=0.1590909 P(node) =0.3548387
class counts: 0 7 37

```
R 4.3.2 · C:/Users/petit/Desktop/repos/UO/rok 3/Wprowadzenie do eksploracji danych/lista11/
      V11 < 3.325  to the right, agree=0.800, adj=0.556, (0 split)
      V7  < 2.275  to the right, agree=0.775, adj=0.500, (0 split)

Node number 3: 44 observations,    complexity param=0.09090909
  predicted class=3  expected loss=0.1590909  P(node) =0.3548387
    class counts:     0     7    37
   probabilities: 0.000 0.159 0.841
  left son=6 (7 obs) right son=7 (37 obs)
  Primary splits:
      V11 < 3.725  to the left,  improve=11.772730, (0 missing)
      V12 < 0.898  to the right, improve= 8.112496, (0 missing)
      V8  < 0.975  to the right, improve= 3.835227, (0 missing)
      V2  < 12.41  to the left,  improve= 3.556854, (0 missing)
      V13 < 1.81   to the right, improve= 3.217172, (0 missing)
  Surrogate splits:
      V12 < 0.898  to the right, agree=0.955, adj=0.714, (0 split)
      V2  < 12.1   to the left,  agree=0.909, adj=0.429, (0 split)
      V5  < 17.25  to the left,  agree=0.909, adj=0.429, (0 split)
      V3  < 1.09   to the left,  agree=0.886, adj=0.286, (0 split)
      V4  < 2.065  to the left,  agree=0.886, adj=0.286, (0 split)

Node number 4: 44 observations,    complexity param=0.01298701
  predicted class=1  expected loss=0.09090909  P(node) =0.3548387
    class counts:    40     4     0
   probabilities: 0.909 0.091 0.000
  left son=8 (37 obs) right son=9 (7 obs)
  Primary splits:
      V11 < 3.75   to the right, improve=3.844156, (0 missing)
      V2  < 13.06  to the right, improve=2.828283, (0 missing)
      V6  < 122.5  to the left,  improve=1.898210, (0 missing)
      V8  < 2.47   to the right, improve=1.578283, (0 missing)
```


Console
Terminal
Background Jobs
R 4.3.2 · C:/Users/petit/Desktop/repos/UO/rok 3/Wprowadzenie do eksploracji danych/lista11/
V4 < 2.065 to the left, agree=0.886, adj=0.286, (0 split)
Node number 4: 44 observations, complexity param=0.01298701
predicted class=1 expected loss=0.09090909 P(node) =0.3548387
class counts: 40 4 0
probabilities: 0.909 0.091 0.000
left son=8 (37 obs) right son=9 (7 obs)
Primary splits:
V11 < 3.75 to the right, improve=3.844156, (0 missing)
V2 < 13.06 to the right, improve=2.828283, (0 missing)
V6 < 122.5 to the left, improve=1.898210, (0 missing)
V8 < 2.47 to the right, improve=1.578283, (0 missing)
V13 < 2.765 to the right, improve=1.578283, (0 missing)
Surrogate splits:
V2 < 12.66 to the right, agree=0.909, adj=0.429, (0 split)
V6 < 134 to the left, agree=0.909, adj=0.429, (0 split)
V8 < 2.3 to the right, agree=0.909, adj=0.429, (0 split)
V5 < 21.3 to the left, agree=0.886, adj=0.286, (0 split)
V7 < 2.075 to the right, agree=0.886, adj=0.286, (0 split)
Node number 5: 36 observations
predicted class=2 expected loss=0 P(node) =0.2903226
class counts: 0 36 0
probabilities: 0.000 1.000 0.000
Node number 6: 7 observations
predicted class=2 expected loss=0 P(node) =0.05645161
class counts: 0 7 0
probabilities: 0.000 1.000 0.000
Node number 7: 37 observations


R 4.3.2 · C:/Users/petit/Desktop/repos/UO/rok 3/Wprowadzenie do eksploracji danych/lista11/
Node number 6: 7 observations
predicted class=2 expected loss=0 P(node) =0.05645161
class counts: 0 7 0
probabilities: 0.000 1.000 0.000
Node number 7: 37 observations
predicted class=3 expected loss=0 P(node) =0.2983871
class counts: 0 0 37
probabilities: 0.000 0.000 1.000
Node number 8: 37 observations
predicted class=1 expected loss=0 P(node) =0.2983871
class counts: 37 0 0
probabilities: 1.000 0.000 0.000
Node number 9: 7 observations
predicted class=2 expected loss=0.4285714 P(node) =0.05645161
class counts: 3 4 0
probabilities: 0.429 0.571 0.000
n= 124
node), split, n, loss, yval, (yprob)
* denotes terminal node
1) root 124 77 2 (0.32258065 0.37903226 0.29838710)
2) V8>=1.4 80 40 1 (0.50000000 0.50000000 0.00000000)
4) V14>=676 44 4 1 (0.90909091 0.09090909 0.00000000)
8) V11>=3.75 37 0 1 (1.00000000 0.00000000 0.00000000) *
9) V11< 3.75 7 3 2 (0.42857143 0.57142857 0.00000000) *
5) V14< 676 36 0 2 (0.00000000 1.00000000 0.00000000) *

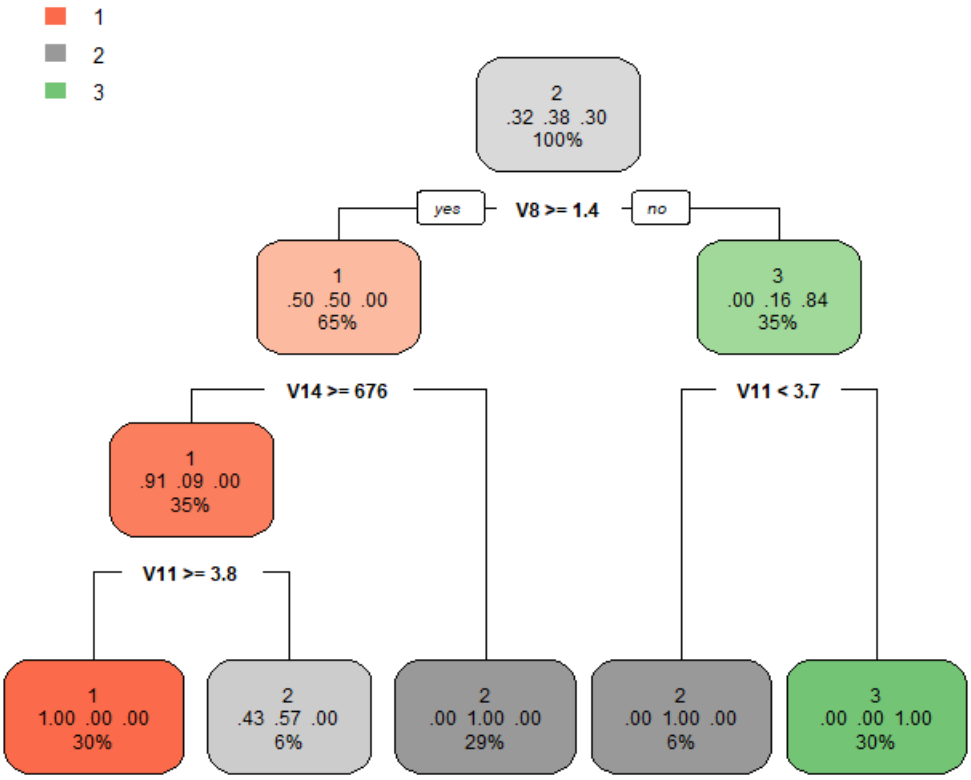


```
n= 124

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 124 77 2 (0.32258065 0.37903226 0.29838710)
  2) V8>=1.4 80 40 1 (0.50000000 0.50000000 0.00000000)
    4) V14>=676 44  4 1 (0.90909091 0.09090909 0.00000000)
      8) V11>=3.75 37  0 1 (1.00000000 0.00000000 0.00000000) *
      9) V11< 3.75 7  3 2 (0.42857143 0.57142857 0.00000000) *
    5) V14< 676 36  0 2 (0.00000000 1.00000000 0.00000000) *
  3) V8< 1.4 44  7 3 (0.00000000 0.15909091 0.84090909)
    6) V11< 3.725 7  0 2 (0.00000000 1.00000000 0.00000000) *
    7) V11>=3.725 37  0 3 (0.00000000 0.00000000 1.00000000) *
>
> # Rysowanie drzewa
> rpart.plot(tree_model)
>
> # Budowa modelu drzewa decyzyjnego bez ograniczenia liczby poziomów
> full_tree_model <- rpart(V1 ~ ., data = train_data, method = "class", cp = 0)
>
> # Klasyfikacja danych ze zbioru testowego
> predictions <- predict(full_tree_model, test_data, type = "class")
>
> # Macierz błędów, dokładność i % błędów
> confusion_matrix <- table(test_data$V1, predictions)
> accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
> error_rate <- 1 - accuracy
>
> print(confusion_matrix)
```

```
# Rysowanie drzewa
rpart.plot(tree_model)

# Budowa modelu drzewa decyzyjnego bez ograniczenia liczby
poziomów
full_tree_model <- rpart(V1 ~ ., data = train_data, method =
"class", cp = 0)

# Klasyfikacja danych ze zbioru testowego
predictions <- predict(full_tree_model, test_data, type =
"class")

# Macierz błędów, dokładność i % błędów
confusion_matrix <- table(test_data$V1, predictions)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
error_rate <- 1 - accuracy
```

```
print(confusion_matrix)
print(paste("Accuracy:", accuracy))
print(paste("Error rate:", error_rate))
```

```
> accuracy <- sum(diag(confusion_matrix)) / su
> error_rate <- 1 - accuracy
>
> print(confusion_matrix)
   predictions
     1  2  3
  1 19  0  0
  2  2 22  0
  3  0  1 10
> print(paste("Accuracy:", accuracy))
[1] "Accuracy: 0.944444444444444"
> print(paste("Error rate:", error_rate))
[1] "Error rate: 0.0555555555555556"
> |
```



Wnioski

- **Dokładność modelu** jest bardzo wysoka, co wskazuje na dobrą jakość klasyfikacji. Warto jednak zwrócić uwagę, czy nie ma ryzyka przeuczenia (overfitting), szczególnie w przypadku modelu bez ograniczeń głębokości.
- **Ważność zmiennych**: Zgodnie z podsumowaniem, najważniejszymi zmiennymi są V8, V7, V11, V14, co może być interesujące w kontekście analizy cech win.
- **Struktura drzewa**: Drzewo decyzyjne wydaje się logicznie podzielić dane, co można zaobserwować poprzez ścieżki decyzyjne i podział w węzłach.