# 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, maxdepth = 3)

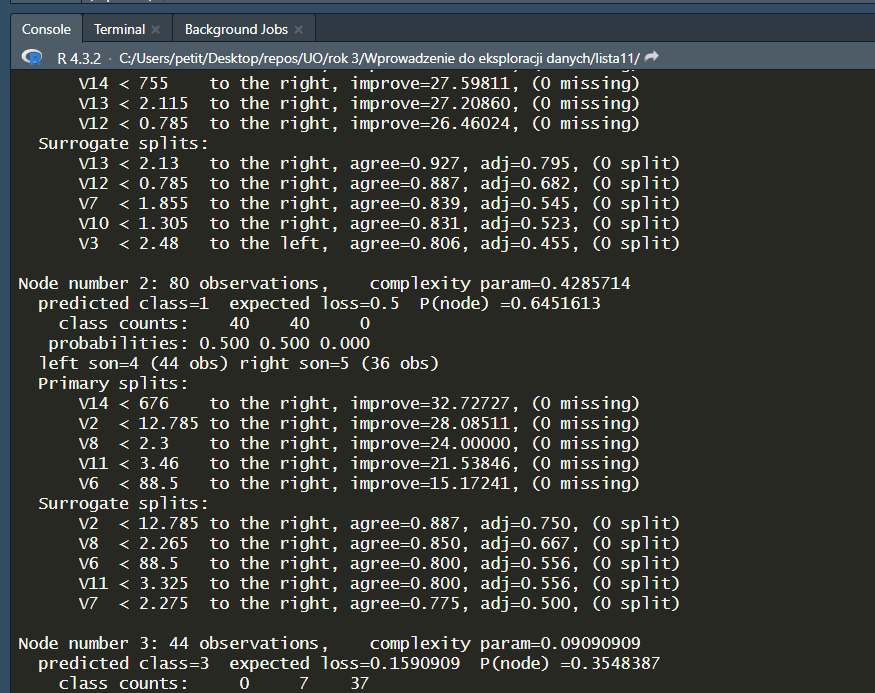
>

> # Wyświetlenie podsumowania modelu

> print(summary(tree\_model))

Obraz zawierający tekst, zrzut ekranu, oprogramowanie, wyświetlacz

Opis wygenerowany automatycznie



Obraz zawierający tekst, elektronika, zrzut ekranu, wyświetlacz

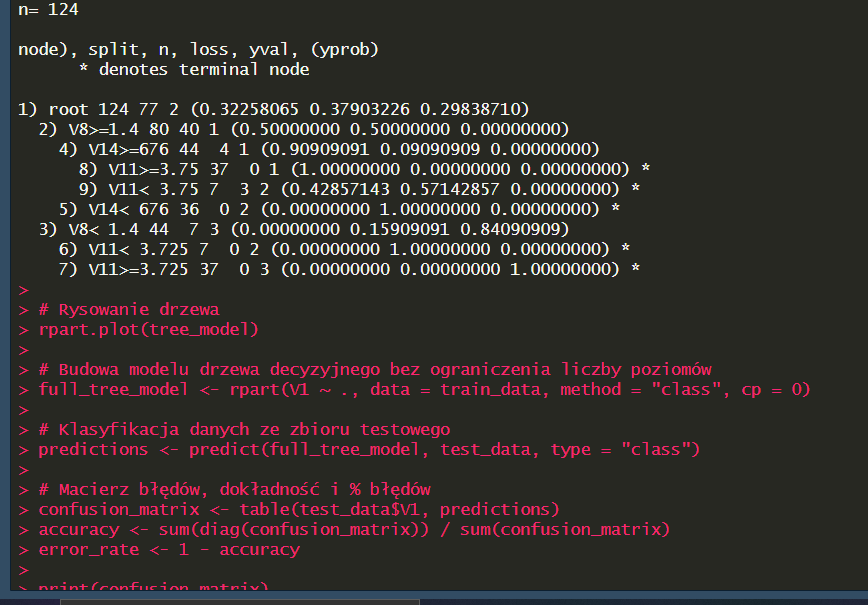
Opis wygenerowany automatycznie

Obraz zawierający tekst, elektronika, zrzut ekranu, wyświetlacz

Opis wygenerowany automatycznie

Obraz zawierający tekst, elektronika, zrzut ekranu, komputer

Opis wygenerowany automatycznie



*# 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

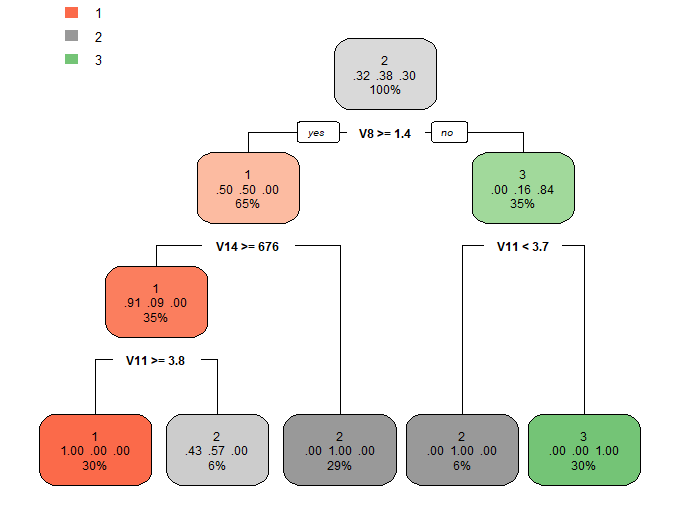
print(confusion\_matrix)

print(paste("Accuracy:", accuracy))

print(paste("Error rate:", error\_rate))

Obraz zawierający tekst, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie



## 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.