TECHNIKI EKSPLORACJI DANYCH

Projekt

Metoda: Drzewo Decyzyjne

Prowadzący: prof. Zbigniew Michna

Grupa: nr 3, środa 11:15 TN

Autorzy: Kaja Kuchnia (273947)

Karolina Kluz (273924)

Spis treści

1. Opis danych: 3
Korekta danych 3
Struktura danych....4
Statystyki opisowe....5
Wykresy....6
Wykres pudełkowy 6
Histogram....7
2. Cel projektu:....8
3. Wybór metody: 9
4. Zastosowanie danej metody do danych:....9
Analiza danych 10
Trafność zbioru testowego i treningowego....11
Prognoza dla nowych danych....11
5. Podsumowanie, wnioski:....12

1. Opis danych:

Wybranym przez nas zbiorem danych używanym do analizy jest zbiór "*tips.csv*". Dane te zostały pozyskane ze strony eio.upc.edu.

Zbiór danych składa się z rejestracji napiwków (w dolarach amerykańskich) otrzymywanych przez kelnera pracującego w jednej restauracji przez okres kilku miesięcy.

Korekta danych

Dane zostały poddane korekcie w postaci zmian nazw kolumn oraz wartości zmiennych kategorialnych na polskie. Te zmiany pozwalają na bardziej intuicyjne zrozumienie danych oraz ułatwiają interpretację wyników analizy. Dodatkowo, eliminują możliwość wystąpienia mieszanki językowej na wykresach i drzewie decyzyjnym, co prowadzi do spójniejszej prezentacji danych.

Wprowadzone zmiany w danych obejmują konwersję zmiennych na czynniki oraz przekształcenie etykiet na polskie, co zwiększa czytelność analizy:

1. Usunięcie kolumny indeksowej: Pierwsza kolumna z indeksem została usunięta, ponieważ R tworzy własny indeks dla danych. Ta modyfikacja pozwala na zachowanie spójności struktury danych i eliminuje niepotrzebne powtórzenie.
2. Konwersja na czynniki z polskimi etykietami: Dane zostały przekształcone na dane typu czynnik, co umożliwia łatwiejszą analizę i interpretację wyników ze względu na dodanie poziomów. Etykiety zostały zmienione na polskie dla lepszej zrozumiałości.
2. Zmiana nazw zmiennych: zamiana nazw wartości w kolumnach danych. Oryginalne nazwy ("*total_bill*","*tip*", "*sex*", "*smoker*", "*day*", "*time*", "*size*") zostały zastąpione przez: "*Rachunek*", "*Napiwek*", "*Płeć*", "*Palacz*", "*Dzień*", "*Pora*", "*Rozmiar*". Ta modyfikacja ułatwia identyfikację i zrozumienie poszczególnych zmiennych.
3. Zmiana etykiet w zmiennych kategorialnych:
    - Zmienna "*Płeć*": Oryginalne etykiety "*Female*" i "*Male*" zostały zmienione odpowiednio na "*Kobieta*" i "*Mężczyzna*".

- Zmienna "*Dzień*": Oryginalne etykiety "*Thur*", "*Fri*", "*Sat*", "*Sun*" zostały zmienione na polskie nazwy dni tygodnia: "*Czwartek*", "*Piątek*", "*Sobota*", "*Niedziela*".
- Zmienna "*Pora*": Oryginalne etykiety "*Lunch*" i "*Dinner*" zostały zmienione na polskie odpowiedniki: "*Lunch*" i "*Obiad*".
- Zmienna "*Palacz*": Oryginalne etykiety "*Yes*" i "*No*" zostały zmienione na polskie odpowiedniki: "*Tak*" i "*Nie*".

Struktura danych

Dane składają się z 244 wierszy, gdzie każdy wiersz stanowi opis jednego uzyskanego napiwku, składającego się z 7 zmiennych:

- *Rachunek* - Kwota rachunku w dolarach, typ danych: liczbowy zmiennoprzecinkowy o podwójnej dokładności, przyjmuje wartości w zakresie 3.07-50.81
- *Napiwek* - Kwota napiwku w dolarach, typ danych: liczbowy zmiennoprzecinkowy o podwójnej dokładności, przyjmuje wartości w zakresie 1-10
- *Płeć* - Płeć osoby płacącej rachunek, typ danych: znakowy, przyjmuje wartości „Kobieta” i „Mężczyzna”
- *Palacz* - Czy w grupie były osoby palące, typ danych: znakowy, przyjmuje wartości „Tak” i „Nie”
- *Dzień* - Dzień tygodnia, typ danych: znakowy, przyjmuje wartości „Niedziela”, „Sobota”, „Czwartek” i „Piątek”
- *Pora* - Pora dnia w trybie działania restauracji, typ danych: znakowy, przyjmuje wartości „Obiad” i „Lunch”
- *Rozmiar* - Liczba osób w grupie, typ danych: liczbowy całkowity, przyjmuje wartości w zakresie od 1 do 6

```
> str(dane)
'data.frame':   244 obs. of  7 variables:
 $ Rachunek: num  17 10.3 21 23.7 24.6 ...
 $ Napiwek : num  1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
 $ Płeć    : Factor w/ 2 levels "Kobieta","Mężczyzna": 1 2 2 2 1 2 2 2 2 2 ...
 $ Palacz  : Factor w/ 2 levels "Tak","Nie": 2 2 2 2 2 2 2 2 2 2 ...
 $ Dzień   : Factor w/ 4 levels "Czwartek","Piątek",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Pora    : Factor w/ 2 levels "Lunch","Obiad": 2 2 2 2 2 2 2 2 2 2 ...
 $ Rozmiar : int  2 3 3 2 4 4 2 4 2 2 ...
```

Zdjęcie 1 Użycie funkcji str()

Strukturę zbioru danych uzyskałyśmy poprzez użycie funkcji *str*(), która zwraca takie informacje jak: typ obiektu, liczba elementów w obiekcie, lista nazwanych składowych obiektu, wraz z ich typami danych oraz funkcji *typeof*() zwracającej typ danych z podziałem typu numerycznego na całkowity oraz zmiennoprzecinkowy. Dodatkowo użyto funkcji *range*() i *unique*() do uzyskania informacji o zakresie danych liczbowych oraz wartościach jakie przyjmują zmienne typu znakowego.

Poniżej przedstawiłyśmy pierwsze 6 wierszy tego zbioru danych.

```
> head(dane)
  Rachunek Napiwek      Płeć Palacz     Dzień  Pora Rozmiar
1    16.99    1.01   Kobieta    Nie Niedziela Obiad       2
2    10.34    1.66 Mężczyzna    Nie Niedziela Obiad       3
3    21.01    3.50 Mężczyzna    Nie Niedziela Obiad       3
4    23.68    3.31 Mężczyzna    Nie Niedziela Obiad       2
5    24.59    3.61   Kobieta    Nie Niedziela Obiad       4
6    25.29    4.71 Mężczyzna    Nie Niedziela Obiad       4
```

Zdjęcie 2 Użycie funkcji head()

Statystyki opisowe

Dokonałyśmy także opisu danych poprzez statystyki opisowe. Aby tego dokonać użyłyśmy funkcji *summary*(). Funkcja ta zwraca: wartość minimalną, pierwszy kwantyl, medianę, średnią, trzeci kwantyl oraz wartość maksymalną. Funkcji tej użyłyśmy tylko dla zmiennych numerycznych.

```
> summary(dane[,sapply(dane, is.numeric)])
    Rachunek        Napiwek          Rozmiar
 Min.   : 3.07   Min.   : 1.000   Min.   :1.00
 1st Qu.:13.35   1st Qu.: 2.000   1st Qu.:2.00
 Median :17.80   Median : 2.900   Median :2.00
 Mean   :19.79   Mean   : 2.998   Mean   :2.57
 3rd Qu.:24.13   3rd Qu.: 3.562   3rd Qu.:3.00
 Max.   :50.81   Max.   :10.000   Max.   :6.00
```

Zdjęcie 3 Użycie funkcji summary() dla zmiennych numerycznych

Z powyższego opisu można uzyskać następujące informacje dla poszczególnych zmiennych:

1. **Rachunek** (zakres wartości od 3.07 do 50.81)
   - 25% obserwacji znajduje się poniżej wartości 13.35, a poniżej wartości 24.13 znajduje się 75% obserwacji
   - Średnia wartość rachunku wynosi 19.79, a mediana wynosi 17.80 – to oznacza, że zmienna ta ma rozkład skośny prawostronny

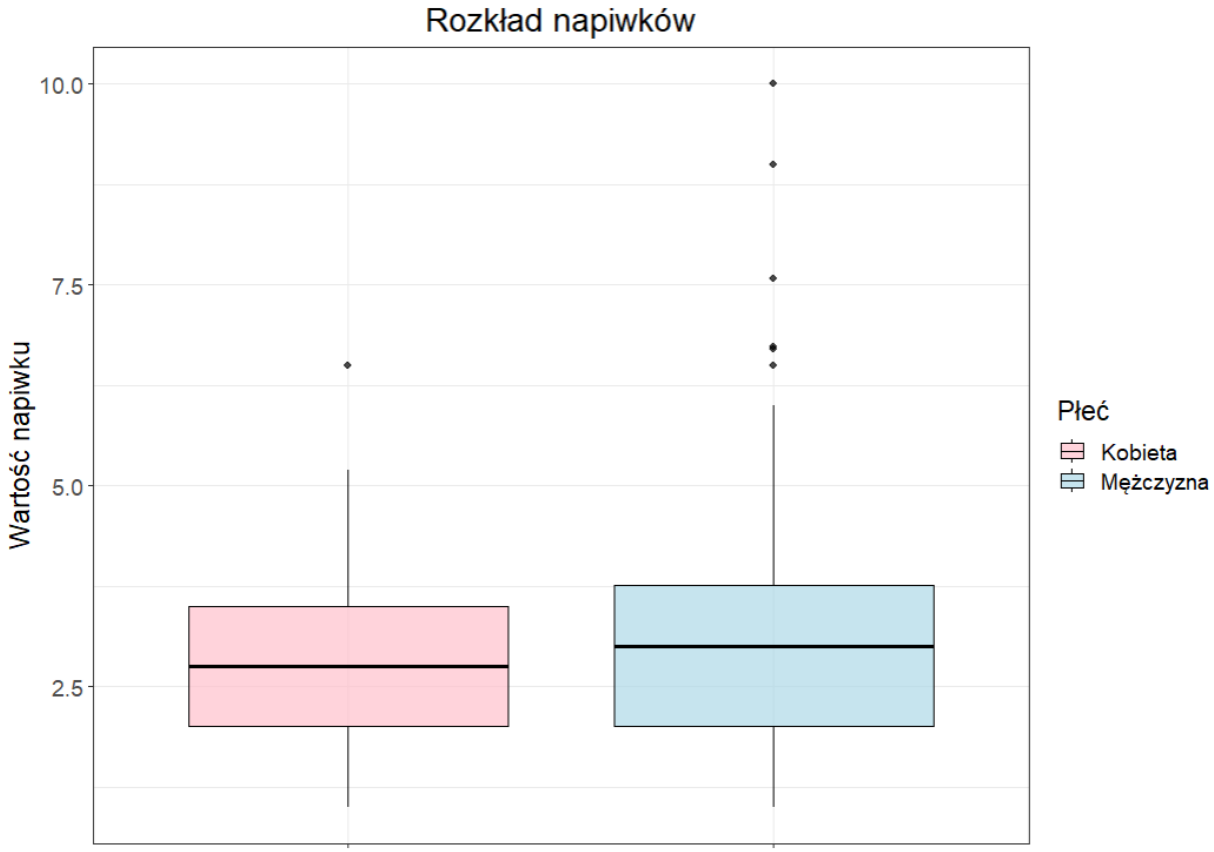
2. **Napiwek** (zakres wartości od 1 do 10)
    - 25% obserwacji znajduje się poniżej wartości 2, a poniżej wartości 3.562 znajduje się 75% obserwacji
    - Średnia wartość napiwku wynosi 2.90, a mediana wynosi 2.998 – to oznacza, że zmienna ta ma rozkład lekko skośny prawostronny
3. **Rozmiar** (zakres wartości od 1 do 6)
    - 25% obserwacji znajduje się poniżej wartości 2, a poniżej wartości 3 znajduje się 75% obserwacji
    - Średnia liczba osób w grupie wynosi 2.57, a mediana wynosi 2 – to oznacza, że zmienna ta ma rozkład skośny prawostronny

Wykresy

Następnie stworzyłyśmy wykresy: pudełkowy i histogram, które ukarzą nam rozkład wartości napiwków w zależności od płci, umożliwiając szybką analizę różnic oraz trendów w zachowaniach konsumenckich.

Wykres pudełkowy

Jako pierwszego stworzyłyśmy wykres pudełkowy za pomocą funkcji z pakietu *ggplot2* w języku R. Wykorzystując dane dotyczące napiwków podzielonych na płcie, zdefiniowałyśmy estetykę graficzną, gdzie oś X reprezentuje płcie, a oś Y przedstawia wartości napiwków. Dodatkowo, znajdują się na nim elementy, takie jak linia środkowa oznaczająca medianę oraz prostokątna skrzynia reprezentująca rozstęp międzykwartylowy.

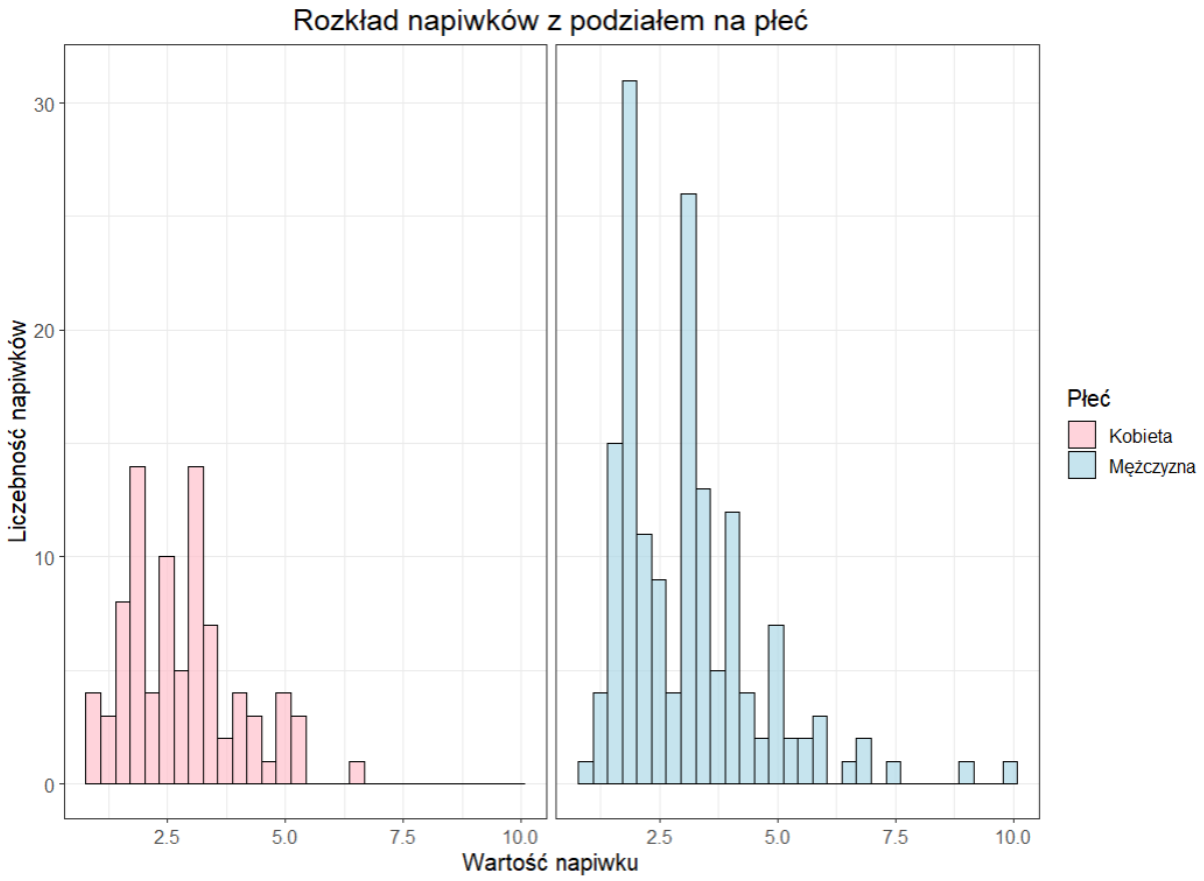


Wykres 1 Wykres pudełkowy dla wartości napiwku z podziałem ze względu na płeć

Na wykresie pudełkowym centralna linia w obu przypadkach reprezentuje medianę, która w obu przypadkach wynosi ponad 2.5 jednak dla mężczyzn jest trochę wyższa niż dla kobiet. Dodatkowo, warto zauważyć, że istnieje tylko jedna wartość odstająca dla kobiet oraz kilka dla mężczyzn. Pudełko przedstawiające próbę mężczyzn charakteryzuje się skośnością lewostronną, co wskazuje na przewagę wartości wyższych niż średnia, a lewy ogon rozkładu jest wydłużony. Wygląd pudełka przedstawiającego próbę kobiet może sugerować, że rozkład wartości napiwku jest zbliżony do symetrycznego.

Histogram

Następnie przeprowadziłyśmy analizę danych napiwków, tworząc histogram za pomocą biblioteki *ggplot2* w języku R. Dane te zostały podzielone ze względu na płeć. Na osi X przedstawione są przedziały wartości napiwków, natomiast na osi Y znajduje się liczba wystąpień w danym przedziale. Ten graficzny zapis danych umożliwia szybką analizę rozkładu wartości napiwków w obu grupach płciowych.



Wykres 2 Histogram dla wartości napiwku z podziałem na płeć

Histogram napiwków ukazuje, że dominująca liczba napiwków koncentruje się w przedziale od około 1 do 4 dolarów, bez względu na płeć. Jednakże, w przypadku mężczyzn, ze względu na większą liczbę obserwacji – co może wskazywać na pewną tendencję w płaceniu za posiłek przez mężczyzn – możemy zaobserwować większe rozproszenie danych i więcej obserwacji reprezentujących wyższe kwoty niż przeciętna.

2. Cel projektu:

Celem analizy jest wykorzystanie drzewa decyzyjnego w celu przewidzenia płci osób na podstawie wartości napiwków.

Jest to przydatne narzędzie dla restauracji, które może automatycznie przypisać płeć klientom na podstawie wartości ich napiwków, co ułatwi gromadzenie danych osobowych potencjalnych konsumentów. Dodatkowo po odnalezieniu pewnych prawidłowości w wielkości napiwków

obsługa restauracji może zwiększyć swoje zarobki poprzez wykrycie i lepszą obsługę tych klientów, od których potencjalnie mogą uzyskać wyższe napiwki.

3. Wybór metody:

W analizie danych zastosowano metodę drzewa decyzyjnego, popularną technikę uczenia maszynowego, wykorzystując pakiety *rpart*, party oraz *rpart.plot* w języku R. Drzewa decyzyjne dzielą dane na podgrupy, korzystając z różnych cech, co umożliwia modelowanie zależności między nimi a przewidywaną zmienną.

4. Zastosowanie danej metody do danych:

W pierwszej kolejności zbiór danych został podzielony na część uczącą (treningową) i testową. Część treningowa zawiera 70% rekordów, natomiast część testowa będzie stanowić 30% zbioru danych. Aby zapewnić losowość podziału, użyte zostało ziarno generatora liczb pseudolosowych. Po wyświetleniu fragmentów każdego ze zbiorów można zaobserwować poprawność losowego doboru danych do poszczególnych części.

```
> head(dane_testowe)
   Rachunek Napiwek      Płeć Palacz     Dzień  Pora Rozmiar
1     16.99    1.01   Kobieta    Nie Niedziela Obiad       2
3     21.01    3.50 Mężczyzna    Nie Niedziela Obiad       3
5     24.59    3.61   Kobieta    Nie Niedziela Obiad       4
7      8.77    2.00 Mężczyzna    Nie Niedziela Obiad       2
12    35.26    5.00   Kobieta    Nie Niedziela Obiad       4
13    15.42    1.57 Mężczyzna    Nie Niedziela Obiad       2
```

Zdjęcie 4 Użycie funkcji head() dla danych testowych

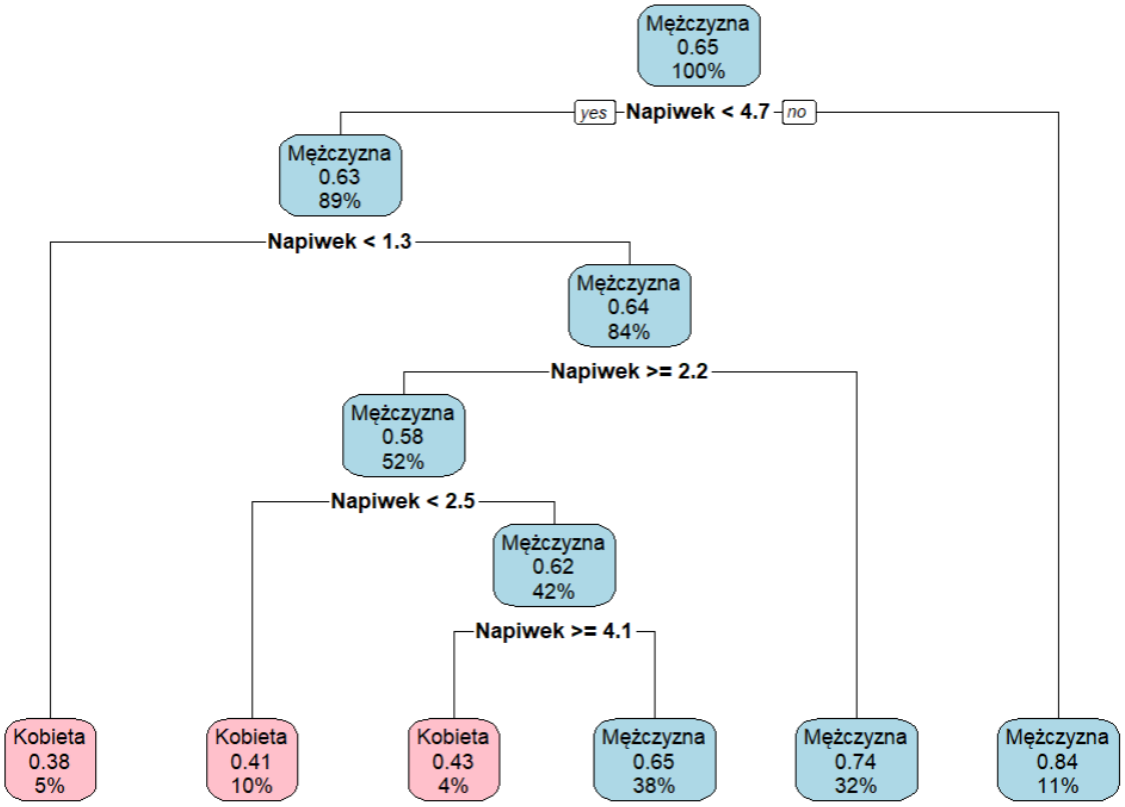
```
> head(dane_treningowe)
    Rachunek Napiwek      Płeć Palacz    Dzień  Pora Rozmiar
28     12.69    2.00 Mężczyzna    Nie   Sobota Obiad       2
80     17.29    2.71 Mężczyzna    Nie Czwartek Lunch       2
150     7.51    2.00 Mężczyzna    Nie Czwartek Lunch       2
101    11.35    2.50   Kobieta    Tak   Piątek Obiad       2
236    10.07    1.25 Mężczyzna    Nie   Sobota Obiad       2
111    14.00    3.00 Mężczyzna    Nie   Sobota Obiad       2
```

Zdjęcie 5 Użycie funkcji head() dla danych treningowych

Atrybutem decyzyjnym jest wartość napiwku która przyjmuje wartości z zakresu od 1 do 10. Atrybutem warunkowym jest natomiast "*Płeć"*, która przyjmuje wartości "Tak", "Nie". Na ich

podstawie zbudowałyśmy i podzieliłyśmy drzewo decyzyjne. Oznacza to, że możemy przypisać płeć osoby płacącej, poprzez wartość przekazanego napiwku.

Analiza danych



Wykres 3 Drzewo decyzyjne z atrybutem decyzyjnym „Wartość napiwku" i z atrybutem warunkowym „Płeć"

W wyniku zastosowania funkcji *rpart*() i funkcji estetycznych z nią związanych, uzyskałyśmy model drzewa decyzyjnego. Przeprowadzając szybką analizę, zauważyłyśmy, że przypadki, w których napiwek przekracza wartość 4.7, z dużym prawdopodobieństwem pochodzą od mężczyzn. Z kolei sytuacje, w których napiwek jest znacznie niższy niż 1.3, najprawdopodobniej dotyczą klientek. Te założenia pozwalają na lepsze zrozumienie zależności między płcią a wysokością napiwków w analizowanym zbiorze danych.

Trafność zbioru testowego i treningowego

```
> błąd_treningowy
[1] 0.3117647
```

Zdjęcie 6 Błąd treningowy - wartość

```
> błąd_testowy
[1] 0.4324324
```

Zdjęcie 7 Błąd testowy - wartość

Błąd treningowy wynoszący 0.3117647 oznacza, że model nieprawidłowo sklasyfikował około 31% próbek ze zbioru treningowego. Natomiast błąd testowy wynoszący 0.4324324 wskazuje, że model popełnił błędnie sklasyfikowane przypadki w około 43% próbek ze zbioru testowego. Różnica między błędem treningowym a testowym jest niewielka, co sugeruje, że model może nieco nadmiernie dopasowywać się do danych treningowych, ale nie występuje znaczące zjawisko przeuczenia.

Prognoza dla nowych danych

Na podstawie przeprowadzonej analizy modelu drzewa decyzyjnego możemy stwierdzić, że dla nowych danych dotyczących wartości napiwków, model jest w stanie przewidzieć płeć osób płacących rachunek.

```
> prognoza
        1         2
  Kobieta Mężczyzna
Levels: Kobieta Mężczyzna
```

Zdjęcie 8 Prognoza płci osób płacących napiwek

Dla nowych danych, gdzie wartości napiwków wynoszą odpowiednio 4.6 i 3.5 dolarów, model przewiduje, że pierwsza osoba jest kobietą, a druga mężczyzną. Te same wyniki prognoz można również odczytać z gotowego drzewa decyzyjnego.

5. Podsumowanie, wnioski:

Po przeprowadzonej analizie danych oraz budowie drzewa decyzyjnego można wysnuć następujące wnioski:

- Analiza danych rachunku, napiwków i rozmiaru grupy sugeruje, że istnieją różnice między płciami w kontekście wysokości napiwków oraz rozmiaru grupy.
- Średnia wartość napiwków dla mężczyzn jest wyższa niż dla kobiet, co sugeruje, że mężczyźni mogą być skłonni do zostawienia większych napiwków.
- Rozkłady wartości rachunku, napiwków i rozmiaru grupy są skośne prawostronnie, co oznacza, że wartości odstające występują w górnej części rozkładu.
- Model drzewa decyzyjnego wykazał zdolność do przewidywania płci klientów na podstawie wysokości napiwków. Jednak błędy treningowy i testowy wskazują na to, że model może nieco nadmiernie dopasowywać się do danych treningowych.
- Mimo niewielkiej różnicy między błędem treningowym a testowym, brak znaczącego zjawiska przeuczenia sugeruje, że model może być użyteczny w przewidywaniu płci na podstawie wysokości napiwków.
- Dla nowych danych dotyczących wysokości napiwków, model jest w stanie dokonywać prognoz dotyczących płci klientów. Na przykład, dla wartości napiwków wynoszących odpowiednio 4.6 i 3.5 dolarów, model przewiduje, że pierwsza osoba jest kobietą, a druga mężczyzną.