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](#_Toc166685119)

[1.1 Korekta danych 3](#_Toc166685120)

[1.2 Struktura danych 4](#_Toc166685121)

[1.3 Statystyki opisowe 5](#_Toc166685122)

[1.4 Wykresy 6](#_Toc166685123)

[1.4.1 Wykres pudełkowy 6](#_Toc166685124)

[1.4.2 Histogram 7](#_Toc166685125)

[2. Cel projektu 8](#_Toc166685126)

[3. Wybór i opis metody 9](#_Toc166685127)

[4. Zastosowanie danej metody do danych 9](#_Toc166685128)

[4.1 Analiza danych 11](#_Toc166685129)

[4.2 Obliczenie średniego błędu kwadratowego na zbiorze treningowym i testowym 11](#_Toc166685130)

[4.3 Prognoza dla nowych danych 12](#_Toc166685131)

[5. Podsumowanie, wnioski 13](#_Toc166685132)

# Opis danych

Wybranym przez nas zbiorem danych używanym do analizy jest zbiór "*tips.csv*". Dane te zostały pozyskane ze strony [eio.upc.edu](http://www-eio.upc.edu/~pau/cms/rdata/datasets.html).

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

## 1.1 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.
3. 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.
4. 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*".

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

Obraz zawierający tekst, Czcionka, zrzut ekranu

Opis wygenerowany automatycznie

Rysunek 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.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

Rysunek 2 Użycie funkcji head()

## 1.3 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.

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

Rysunek 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

1. **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

1. **Rozmiar** (zakres wartości od 1 do 6)

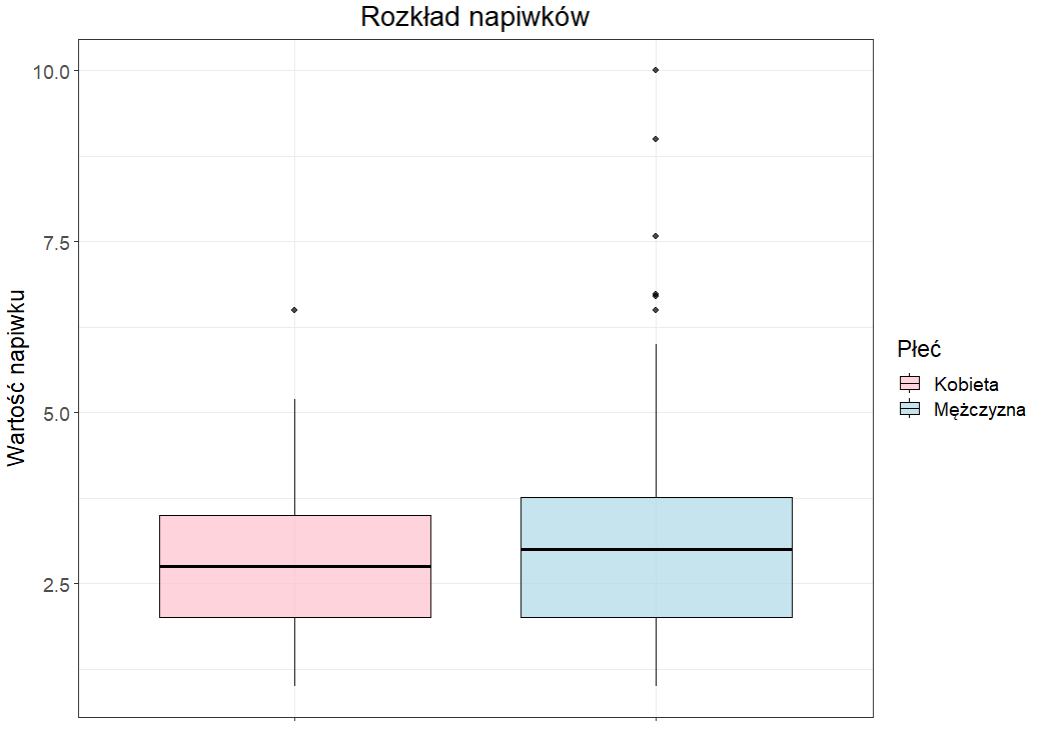
* 25% obserwacji znajduje się poniżej wartości 2, a poniżej wartości 3 znajduje się 75% obserwacji
* Średnia liczba osób w grupie wynosi 2.57, a mediana wynosi 2 – to oznacza, że zmienna ta ma rozkład skośny prawostronny

## 1.4 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.

### 1.4.1 Wykres pudełkowy

Jako pierwszy 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.

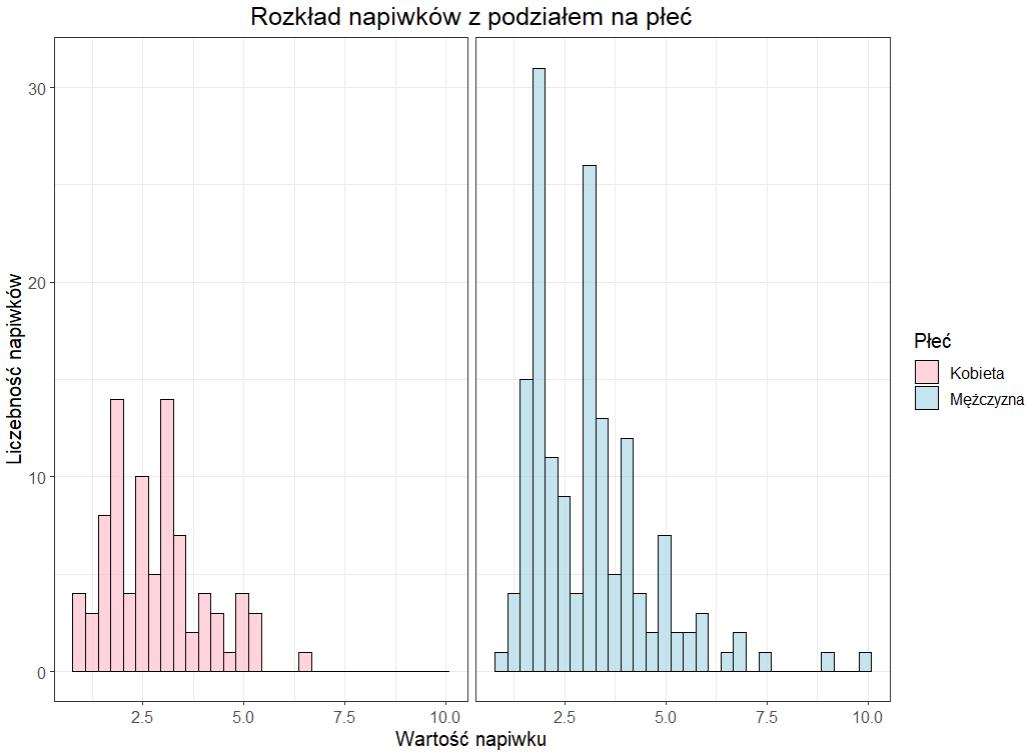


Wykres 1 Wykres pudełkowy dla wartości napiwku z podziałem ze względu na płeć

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.

### 1.4.2 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.

# Cel projektu

Celem analizy jest przewidzenie wartości napiwku w zależności od następujących czynników: płeć płacącego, rozmiar obsługiwanej grupy klientów oraz czy osoby są palaczami.

Wynik tej analizy będzie przydatny dla kelnera, który kolekcjonował te dane. Będzie on wiedział, które stoliki (grupy klientów) do obsługi wybrać, tak żeby zebrać jak największą kwotę w napiwkach, czyli zmaksymalizować swój zarobek.

# Wybór i opis metody

Do przeprowadzenia naszej analizy zdecydowałyśmy się wybrać drzewo regresyjne, ponieważ jest ono jedną z najbardziej intuicyjnych i łatwych w interpretacji metod modelowania. Jego struktura pozwala na przejrzystą wizualizację. Dodatkowo celem naszej analizy jest przewidzenie wartości zmiennej ciągłej („*Napiwek*”) na podstawie zmiennych kategorycznych, co sprawia, że drzewo decyzyjne jest najodpowiedniejszą metodą eksploracji danych.

Drzewo regresyjne to jeden z typów drzewa decyzyjnego. W porównaniu do drzew klasyfikacyjnych, w których atrybutem decyzyjnym jest cecha dyskretna (kategoryczna), w drzewach regresyjnych atrybut decyzyjny jest ciągły, najczęściej są to liczby rzeczywiste. Jest ono wykorzystywane w uczeniu maszynowym, a głównym jego celem jest przewidywanie wartości zmiennej docelowej na podstawie danych wejściowych. Drzewo regresyjne składa się z węzłów, gałęzi i liści. Węzły reprezentują testy na atrybutach danych, każda gałąź odpowiada jednemu z wyników tego testu, a każdy liść reprezentuje wartość prognozowaną przez model. Proces tworzenia drzewa regresyjnego zaczyna się od węzła głównego i odbywa się iteracyjnie – w każdej iteracji dokonywany jest podział zbioru danych na dwa podzbiory, tak aby minimalizować całkowitą sumę kwadratów odchyleń. Proces kończy się w momencie osiągnięcia danego kryterium zatrzymania.

# Zastosowanie danej metody do danych

W pierwszej kolejności zbiór danych został podzielony na część uczącą (treningową) i testową. Dane zostały losowo przyporządkowane do obydwu części. Część treningowa zawiera 70% rekordów, natomiast część testowa stanowi 30% zbioru danych. Po wyświetleniu fragmentów każdego ze zbiorów można zaobserwować poprawność losowego doboru danych do poszczególnych części.

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

Rysunek 4 Użycie funkcji head() dla danych testowych

Obraz zawierający tekst, Czcionka, zrzut ekranu, numer

Opis wygenerowany automatycznie

Rysunek 5 Użycie funkcji head() dla danych treningowych

Atrybutem decyzyjnym w naszej analizie jest „*Napiwek*”, który przyjmuje wartości z zakresu od 1 do 10. Natomiast atrybutami warunkowymi są: „*Płeć*”, „*Rozmiar*”, „*Palacz”*. Na ich podstawie utworzyłyśmy drzewo regresyjne.

## 4.1 Analiza danych

Obraz zawierający tekst, diagram, linia, mapa

Opis wygenerowany automatycznie

Wykres 3 Drzewo decyzyjne z atrybutem decyzyjnym „Wartość napiwku” i z atrybutami warunkowymi „Płeć”, „Rozmiar” oraz „Palacz”

W wyniku zastosowania funkcji *rpart*() i funkcji estetycznych z nią związanych, uzyskałyśmy model drzewa regresyjnego. Okazuje się, że osoby niepalące płci męskiej, które uczestniczą w posiłku w towarzystwie minimum 4 osób zostawiają największy napiwek kelnerowi. Natomiast najmniejszy napiwek zostawiają kobiety, także niepalące, jedzące posiłek w towarzystwie jednej osoby lub samotnie. Możemy także zauważyć ogólną zależność wartości napiwku od rozmiaru grupy – im więcej osób uczestniczy w posiłku, tym większy jest napiwek.

## 4.2 Obliczenie średniego błędu kwadratowego na zbiorze treningowym i testowym

Dla naszego modelu drzewa regresyjnego postanowiłyśmy obliczyć średni błąd kwadratowy (MSE) na zbiorze treningowym oraz testowym. Wartości średniego błędu kwadratowego na obu tych zbiorach przedstawiono poniżej.



Rysunek 6 Błąd treningowy - wartość



Rysunek 7 Błąd testowy - wartość

Średni błąd kwadratowy na zbiorze treningowym wyniósł około 1,40. Oznacza to, że średni kwadrat różnicy między wartościami przewidywanymi przez model a rzeczywistymi wartościami na danych treningowych wynosi około 1.40. Natomiast na zbiorze testowym wartość dla tego błędu wyniosła około 1,46. Oznacza to, że średni kwadrat różnicy między wartościami przewidywanymi przez model a rzeczywistymi wartościami na danych testowych wynosi około 1.46. Zakres wartości napiwków wynosi od 1 do 10, więc te błędy są miarą tego, jak dobrze model radzi sobie z przewidywaniem wartości napiwków w tym zakresie. Im mniejszy błąd, tym lepiej model jest dopasowany do danych. Wartości błędów na poziomie około 1.4 sugerują, że model może nie być idealny, ale radzi sobie stosunkowo dobrze z przewidywaniem wartości napiwków w tym zakresie.

## 4.3 Prognoza dla nowych danych

Na podstawie przeprowadzonej analizy modelu drzewa regresyjnego możemy określić przewidywaną wartość napiwku na podstawie informacji jakiej płci była osoba płacąca, w jak licznym towarzystwie jadła oraz czy osoby te były palaczami dla nowych danych.

Obraz zawierający tekst, Czcionka, biały

Opis wygenerowany automatycznie

Rysunek 8 Prognoza wartości napiwku dla nowych danych

Dla nowych danych mających postać („Kobieta”, 2, „Tak”) oraz („Mężczyzna”, 4, „Nie”), model przewiduje, że pierwsza osoba (kobieta) da napiwek kelnerowi o wartości 2,72$, natomiast druga osoba (mężczyzna) da napiwek o wartości 4,58$. Te same wyniki prognoz można również odczytać z gotowego drzewa decyzyjnego.

# Podsumowanie, wnioski

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

* Prognozy wartości napiwków na podstawie modelu mogą pomóc kelnerom w lepszym zarządzaniu swoim czasem i zasobami, poprzez skoncentrowanie się na obsłudze tych grup klientów, które mają tendencję do zostawiania wyższych napiwków.
* Osoba, która zgromadziła te dane (kelner w restauracji) będzie wiedziała, które stoliki (grupy klientów) wybrać do obsługi, aby zmaksymalizować swój zarobek (niepalący mężczyźni w dużym towarzystwie).
* Model drzewa regresyjnego okazał się przydatnym narzędziem do przewidywania wysokości napiwków na podstawie różnych czynników, takich jak płeć klienta, czy grupa jest paląca oraz rozmiar grupy.
* Wartości średniego błędu kwadratowego obliczone na zbiorze treningowym oraz testowym wskazują na stosunkowo dobrą jakość dopasowania modelu do danych.
* Drzewo regresyjne jest łatwe do zrozumienia, co czyni go przydatnym narzędziem dla personelu restauracyjnego, którzy mogą wykorzystać te informacje do lepszego zrozumienia preferencji klientów i dostosowywania obsługi.