

*Badanie czynników wpływających na
ilość spalanych kalorii podczas sesji
treningowej na siłowni*

Mateusz Dziendziel

129745

Warszawa, Maj 2025

Spis treści

Badany problem	2
Informacje o danych	2
Information Value oraz Macierz Korelacji	11
Drzewo klasyfikacyjne.....	13
Drzewo klasyfikacyjne (Głębokie)	14
Las losowy.....	15
Porównanie wyników	16
Krzywe ROC i Krzywe LIFT.....	16
Podsumowanie	17

Badany problem

Celem przeprowadzonego badania było ustalenie, które zmienne w istotny sposób wpływają na liczbę spalanych kalorii podczas jednej sesji treningowej na siłowni. Dodatkowo opracowano model predykcyjny, który – na podstawie dostępnych informacji o użytkowniku – pozwala oszacować, czy dana osoba znajdzie się w grupie osób spalających ponad 894 kalorie w trakcie pojedynczego treningu, czy też nie.

Informacje o danych

Dane: <https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset>

Dane wykorzystane w analizie pochodzą ze strony *kaggle.com*. Zbiór został sztucznie wygenerowany na podstawie rzeczywistych informacji. Zawiera 973 kompletne obserwacje – brak jest jakichkolwiek luk w danych. W zestawie znajduje się 15 zmiennych, z czego zmienną objaśnianą jest **Calories_Burned**.

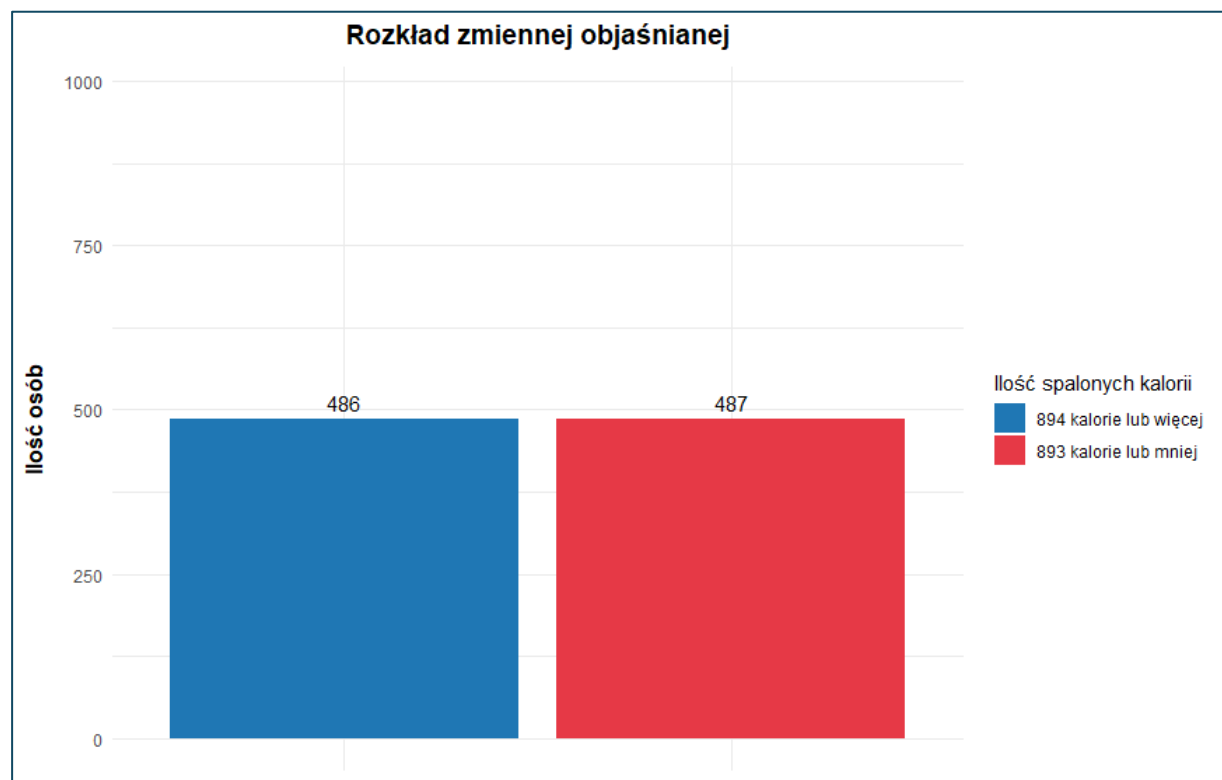
Wartości tej zmiennej zostały zaklasyfikowane do dwóch kategorii na podstawie mediany:

- **1** – osoby, które spaliły **894 kalorie lub więcej**,
- **0** – osoby, które spaliły **893 kalorie lub mniej** podczas sesji treningowej.

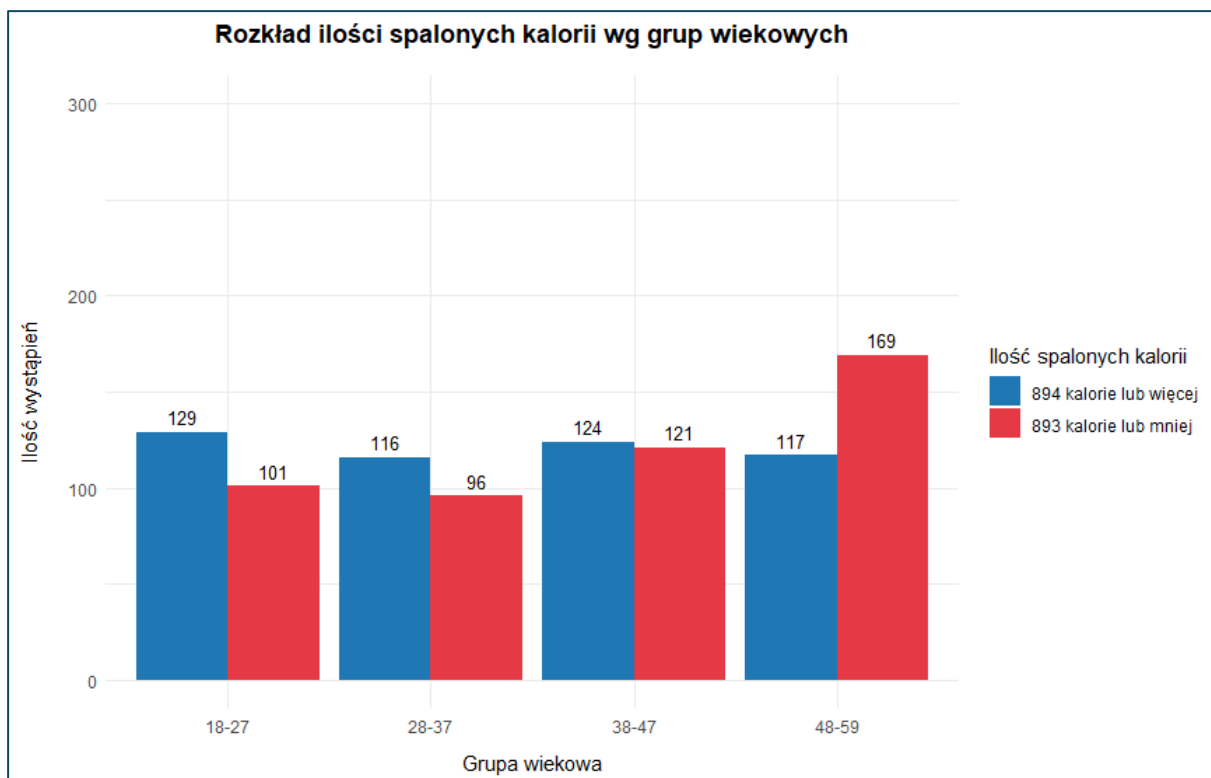
Pozostałe zmienne zostały przekształcone w formę kategoriczną i posłużyły jako predyktory w dalszej analizie. Szczegółowy opis zmiennych, liczba kategorii oraz ich znaczenie zostały przedstawione w tabeli oraz na poniższych wykresach.

Age	Wiek(lata)
Gender	pleć
Weight..kg.	Waga(kg)
Height..m.	Wzrost(m)
Max_BPM	Maksymalne tętno podczas treningu
Avg_BPM	Średnie tętno podczas treningu
Resting_BPM	Tętno mierzone przed rozpoczęciem treningu
Session_Duration_hours	Czas trwania sesjii(godziny)
Calories_Burned	Ilość spalonych kalorii podczas 1 sesji 1 – 894 kalorie lub więcej 0 – 893 kalorie lub mniej
Workout_type	Typ wykonywanych ćwiczeń
Fat_percentage	% tkanki tłuszczowej
Water_Intake..liters	Dzienna ilość spożywanej wody podczas treningów
Workout_Frequency..days.week.	Ilość sesji ćwiczeń w tygodniu(dni)
Experience_level	Doświadczenie na siłowni : 1 - Początkujący 2 - Średnio zaawansowany 3 - expert
BMI	Wskaźnik masy ciała

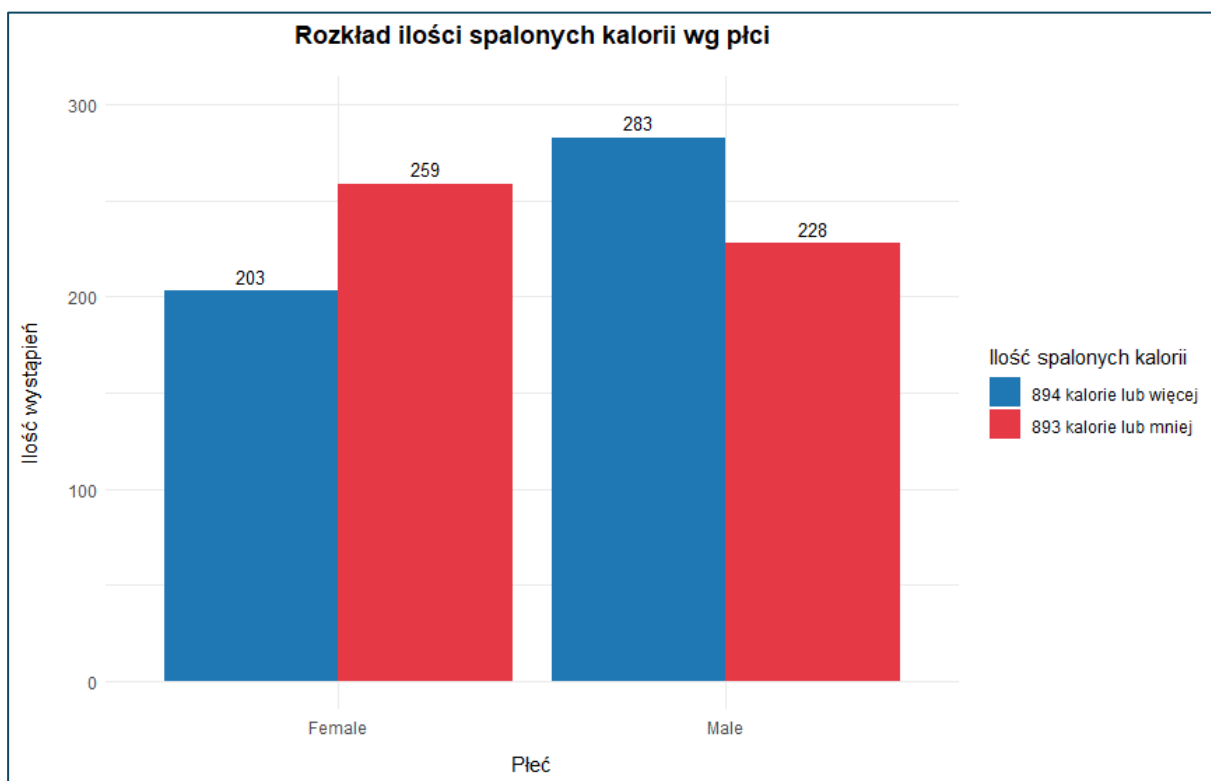
Tabela 1: Nazwa zmiennej w oryginalnym zbiorze (po lewej) i jej znaczenie (po prawej)



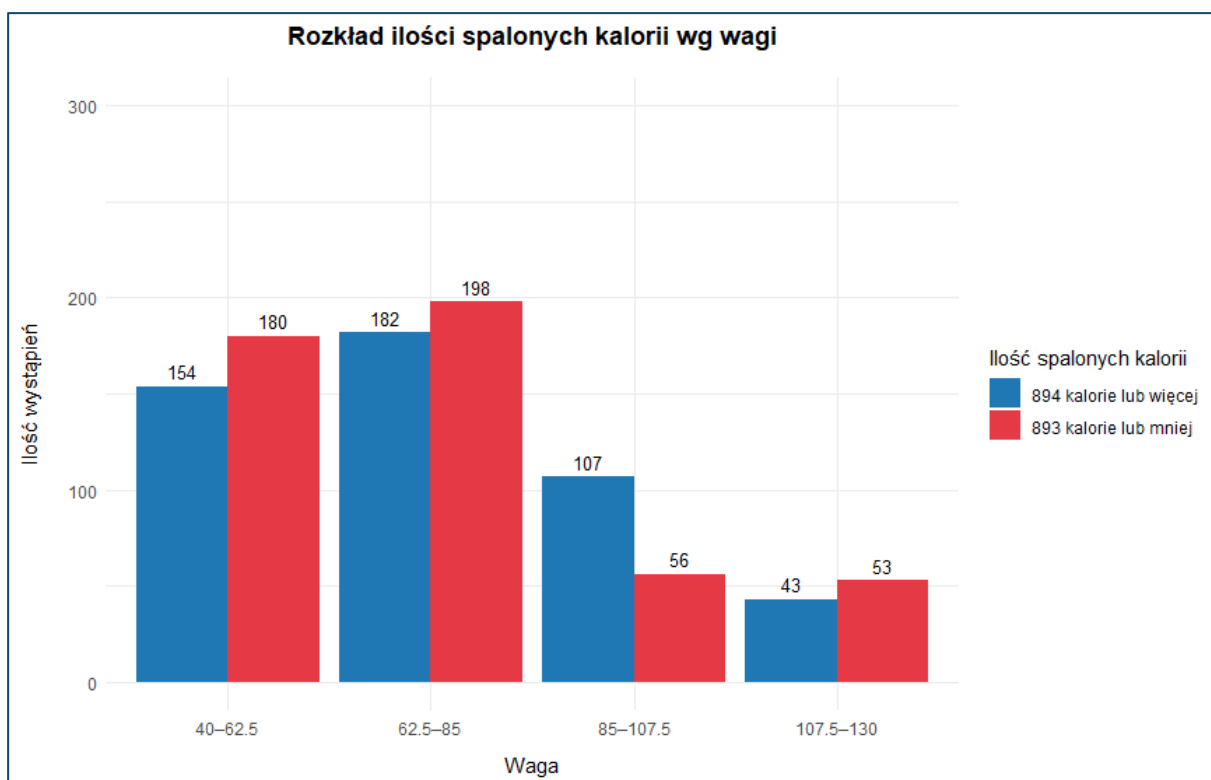
Wykres 1: Rozkład zmiennej objaśnianej i liczebność każdej z grup



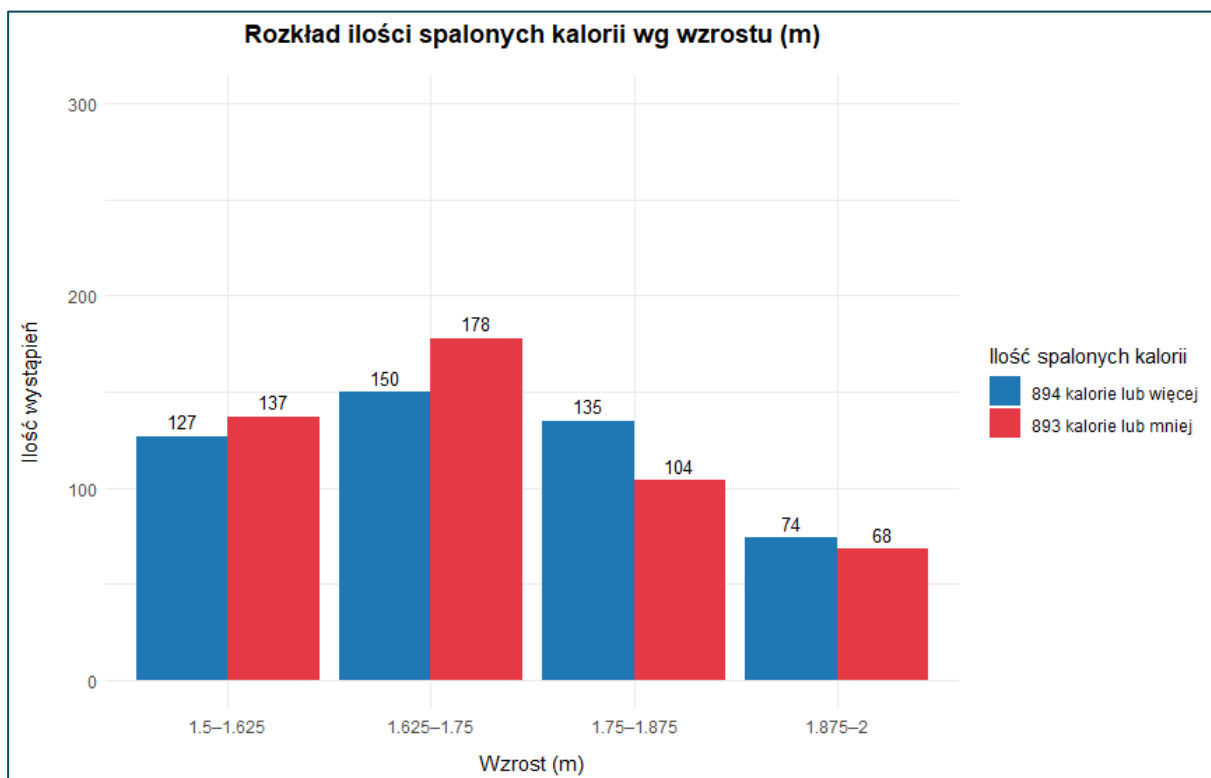
Wykres 2: Rozkład zmiennej „wiek” i liczebność w każdej z grup



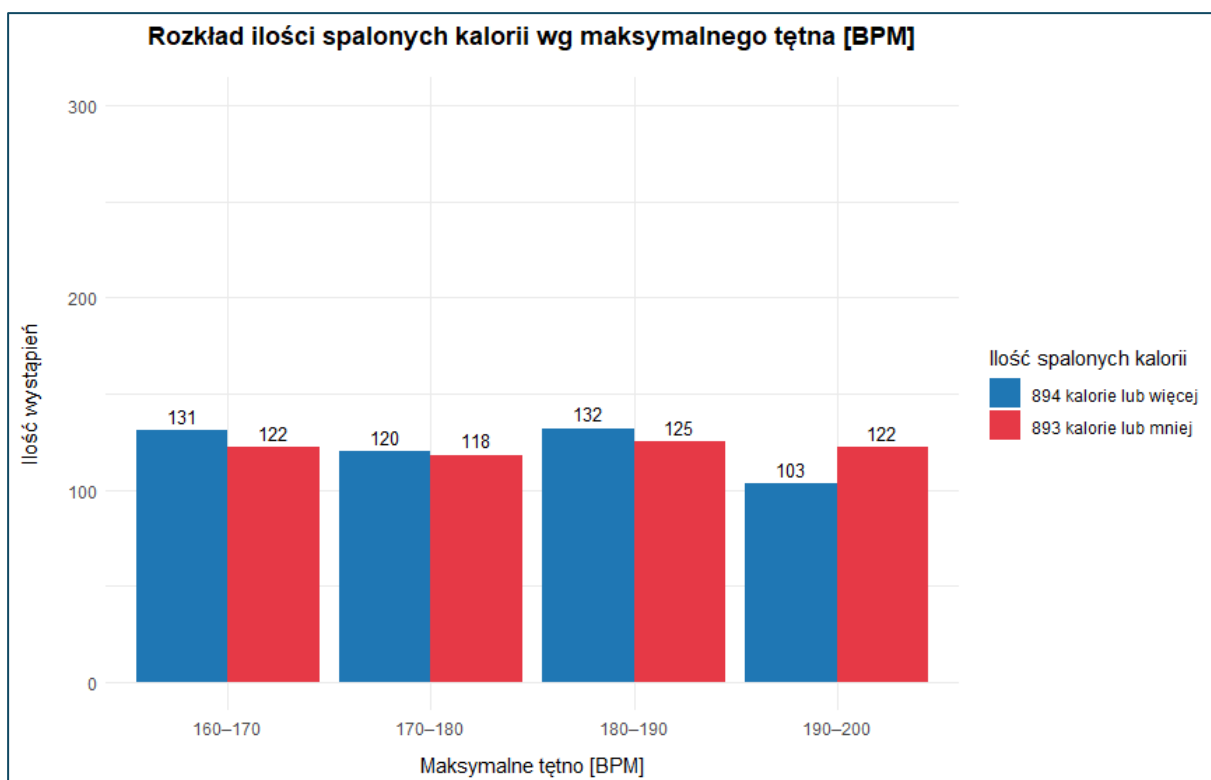
Wykres 3: Rozkład zmiennej „płeć” i liczebność w każdej z grup



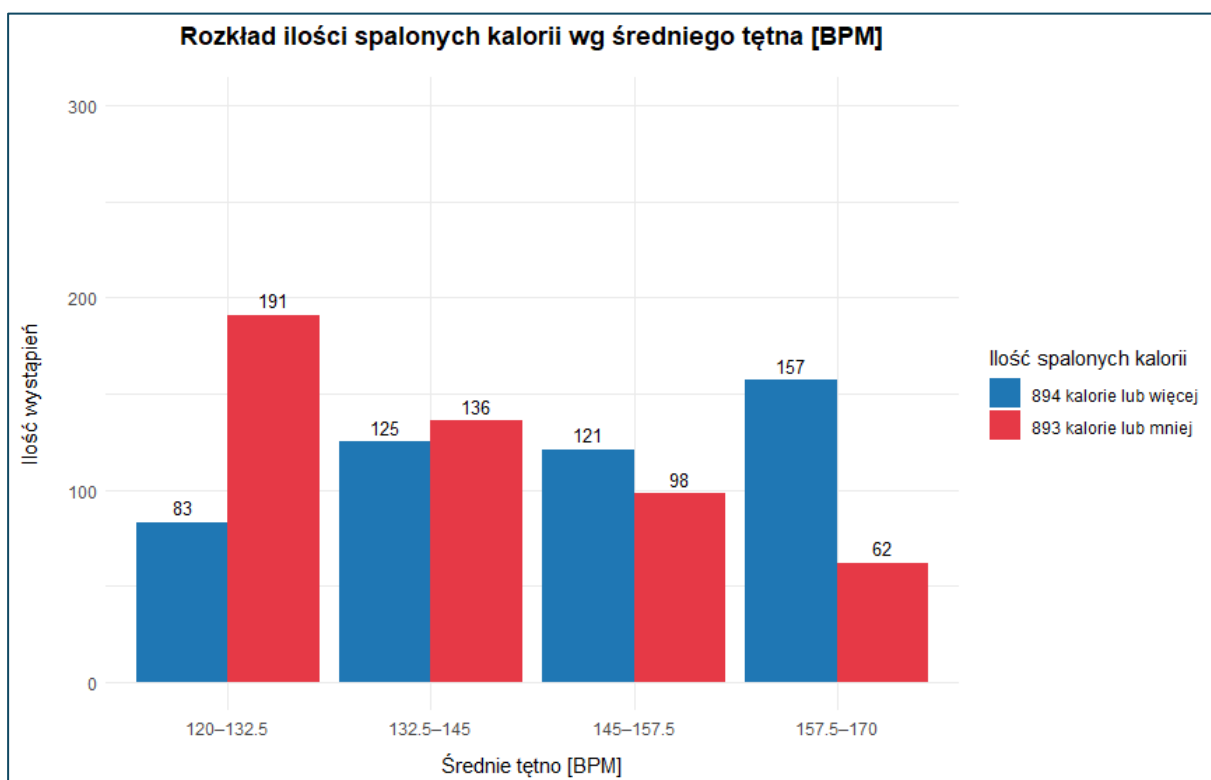
Wykres 4: Rozkład zmiennej „waga” i liczebność w każdej z grup



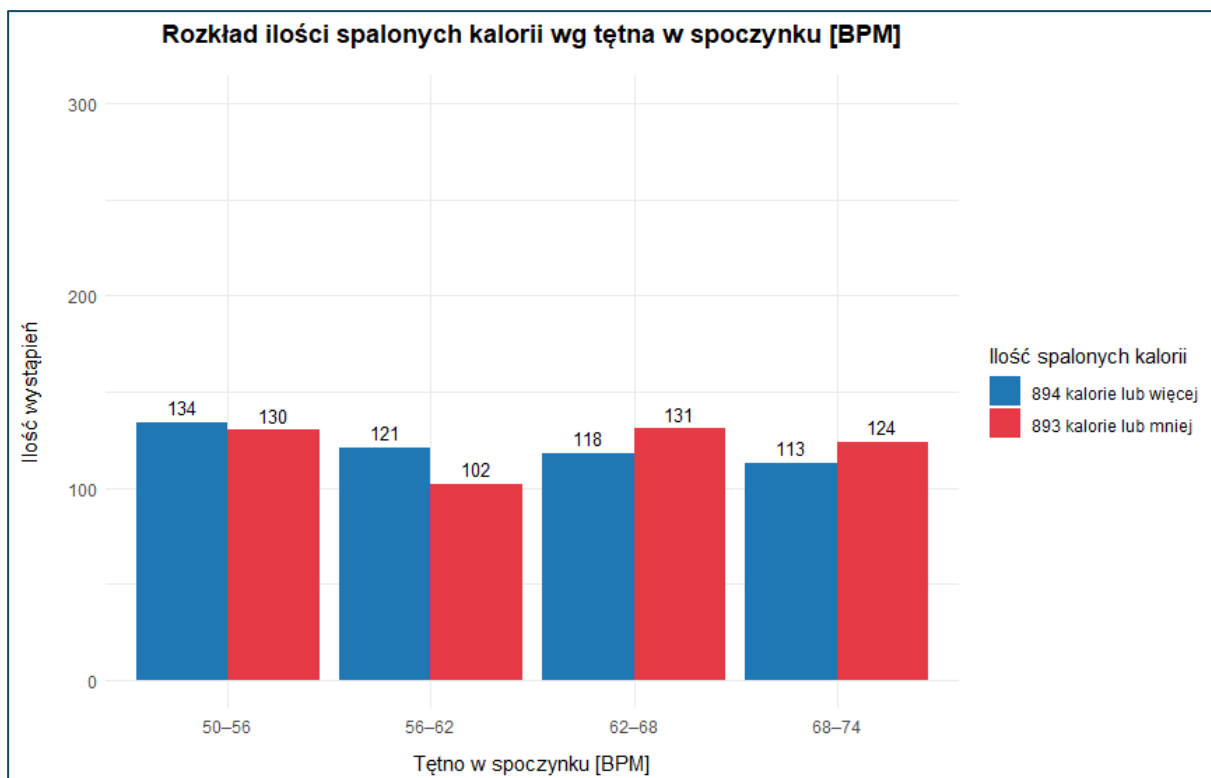
Wykres 5: Rozkład zmiennej „wzrost” i liczebność w każdej z grup



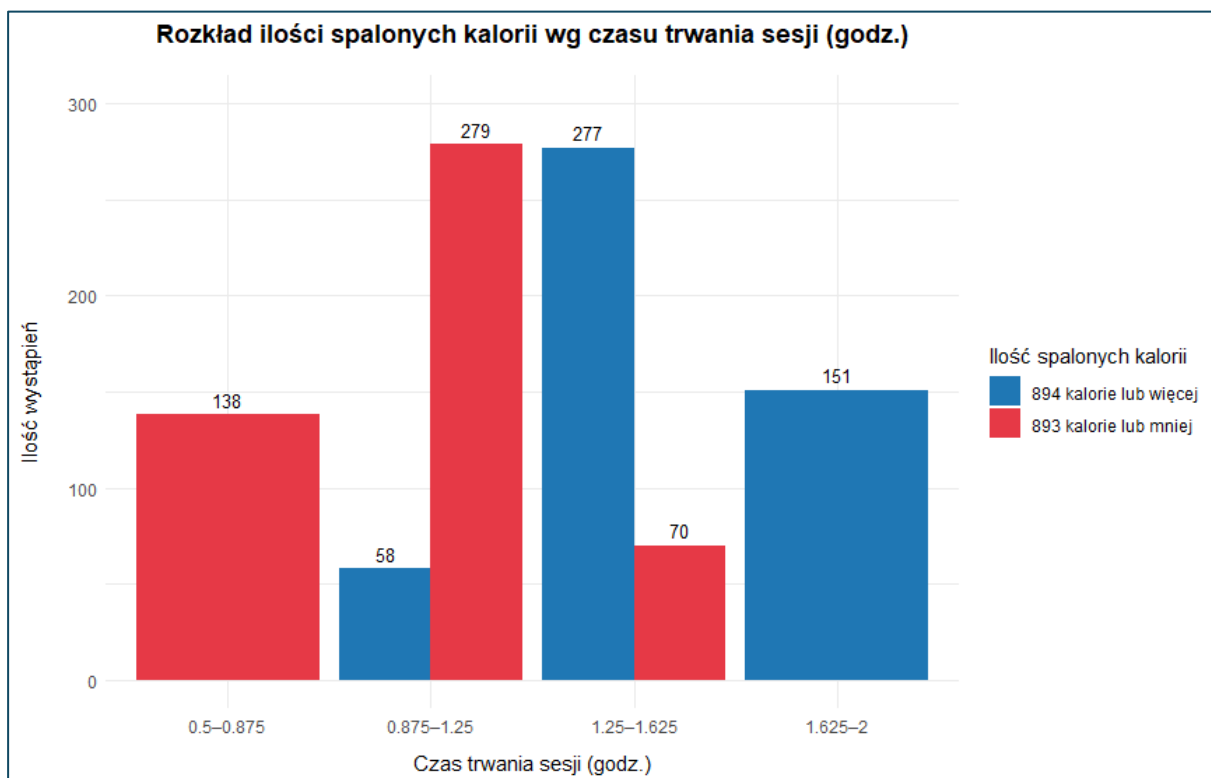
Wykres 6: Rozkład zmiennej „maksymalne tętno” i liczebność w każdej z grup



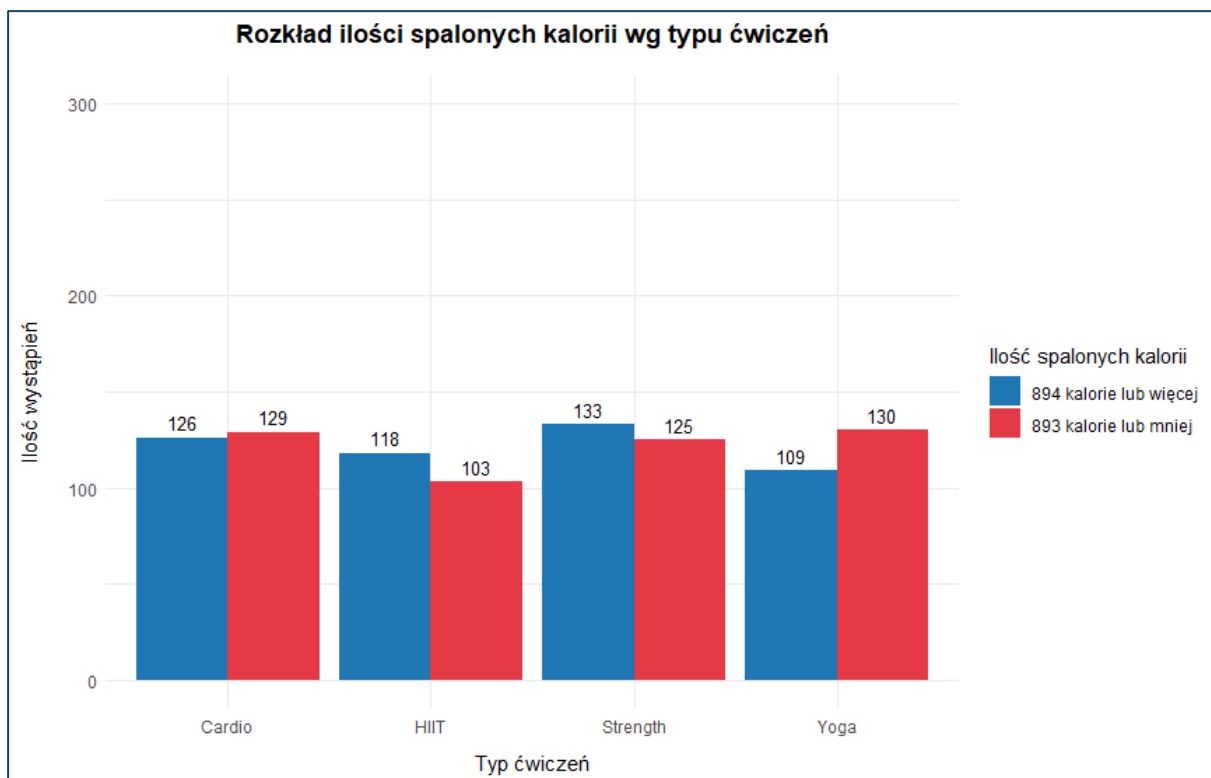
Wykres 7: Rozkład zmiennej „Średnie tętno” i liczebność w każdej z grup



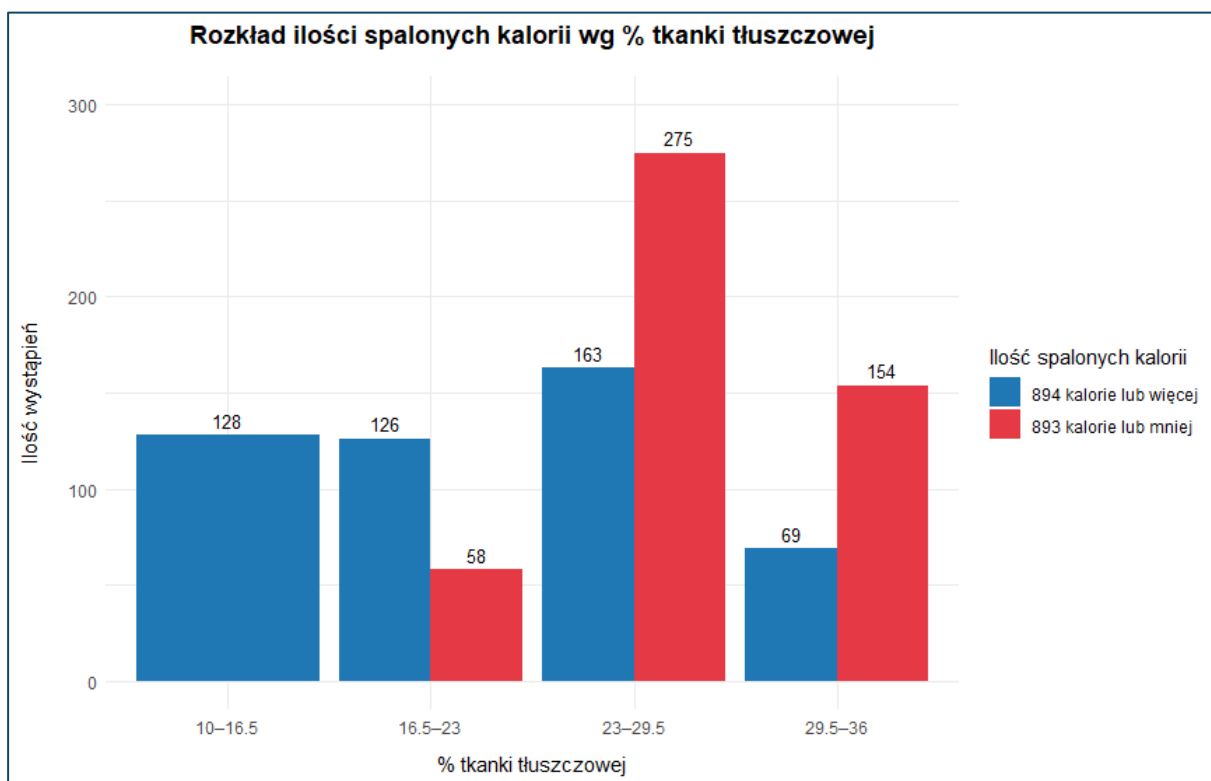
Wykres 8: Rozkład zmiennej „Tętno W spoczynku” i liczebność w każdej z grup



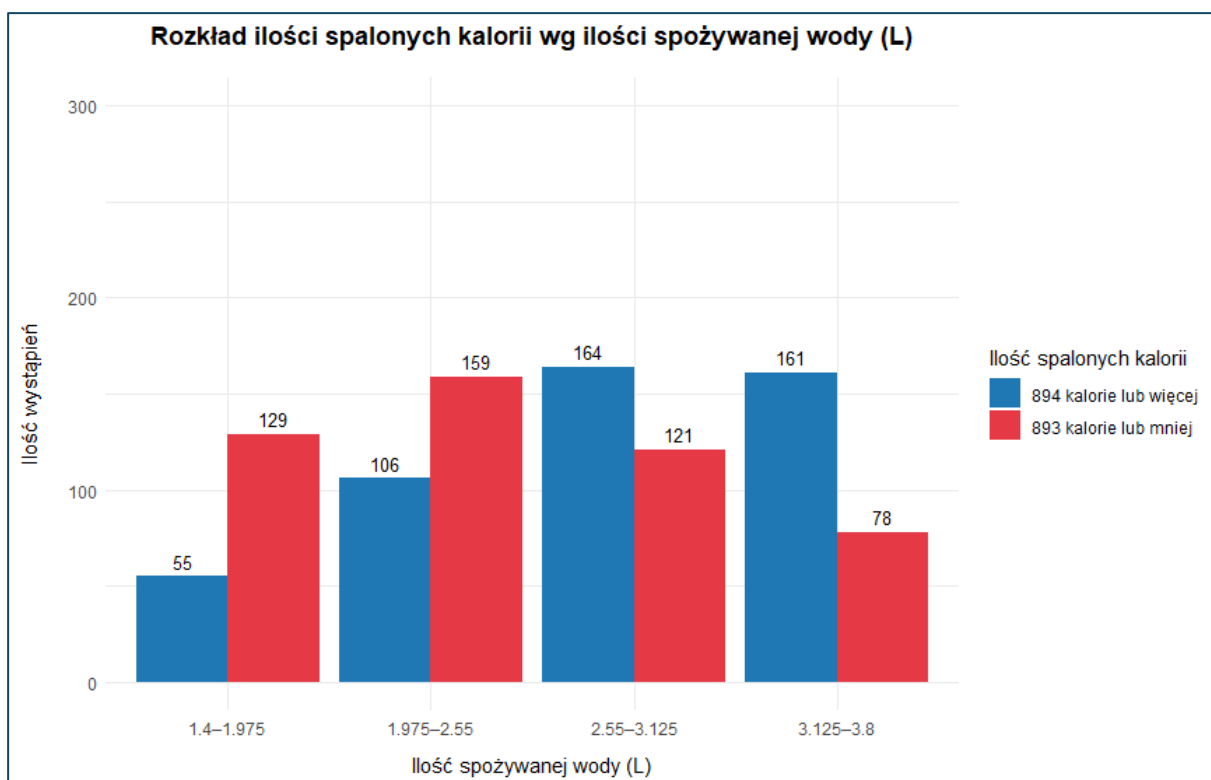
Wykres 9: Rozkład zmiennej „Czas trwania sesji” i liczebność w każdej z grup



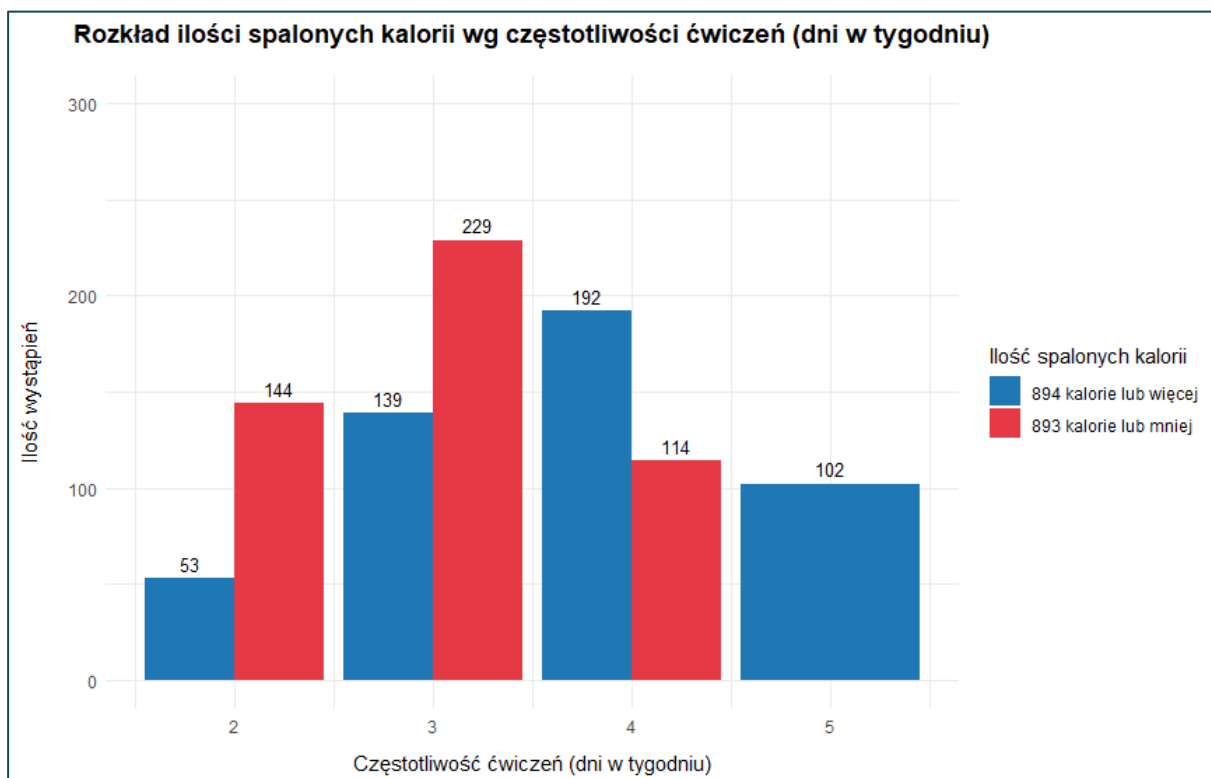
Wykres 10: Rozkład zmiennej „Typ ćwiczeń” i liczebność w każdej z grup



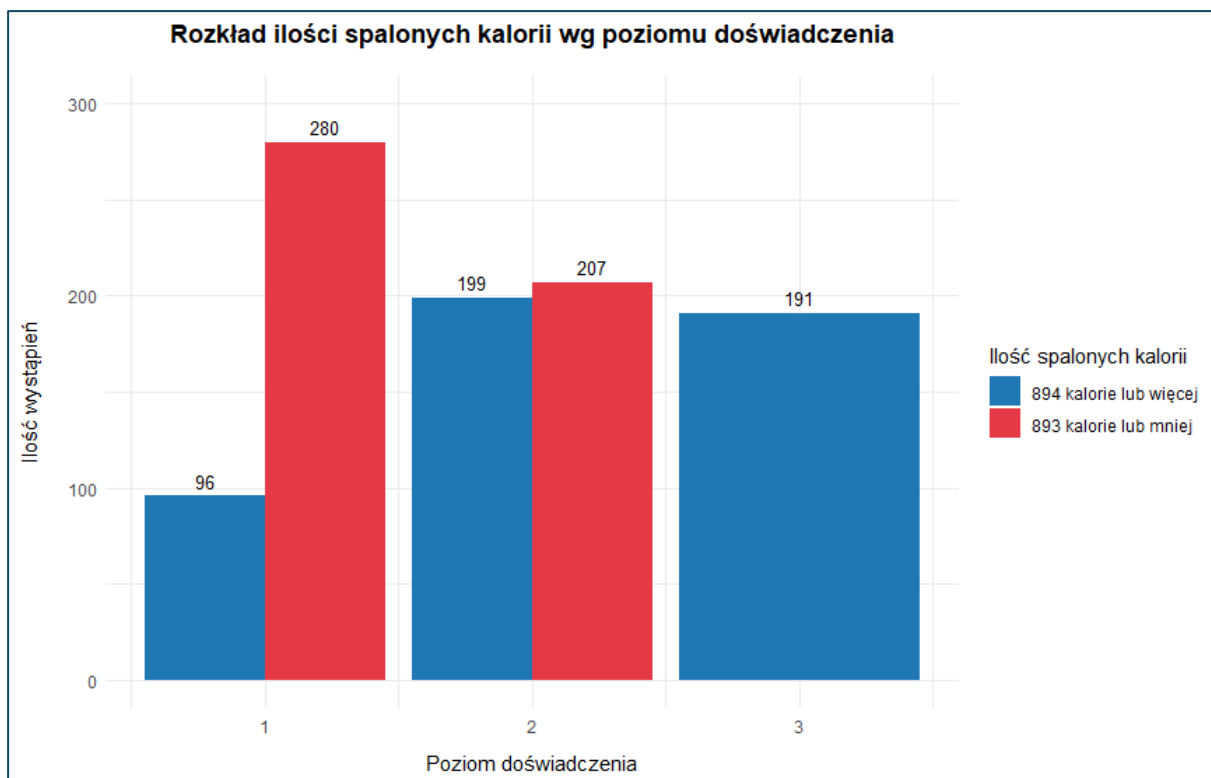
Wykres 11: Rozkład zmiennej „Procent tkanki tłuszczowej” i liczebność w każdej z grup



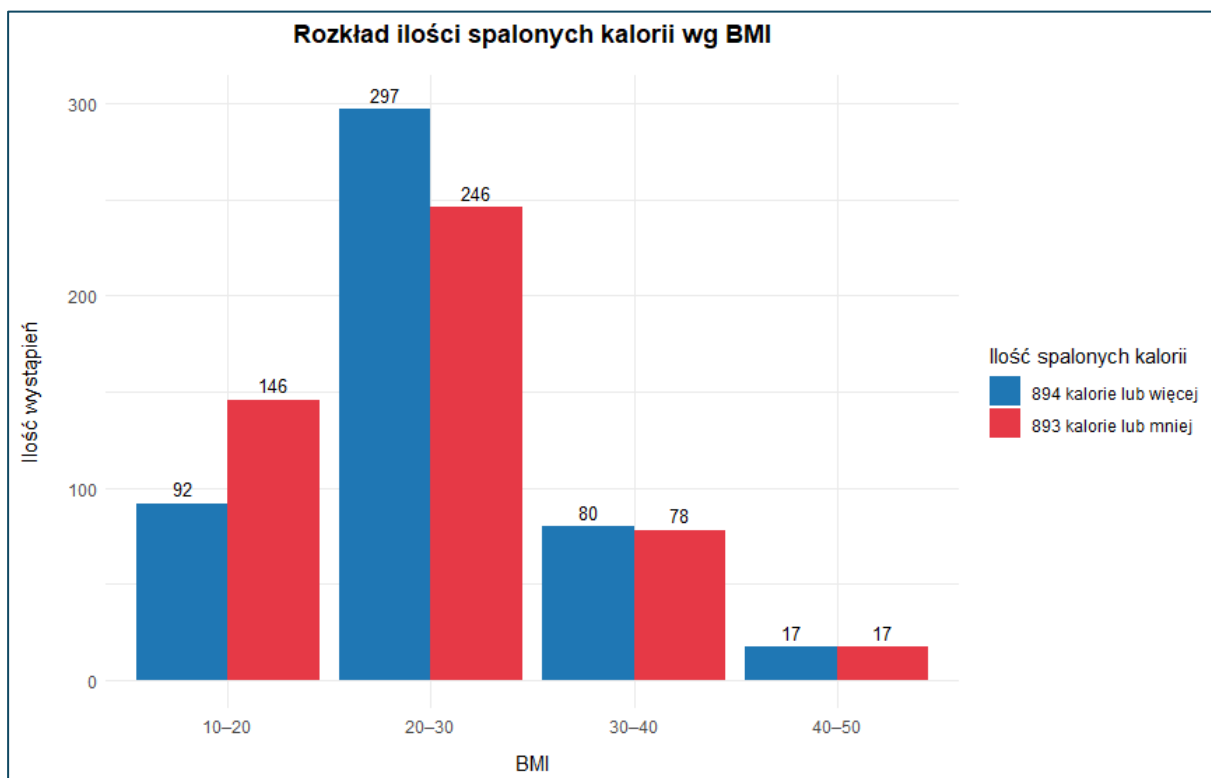
Wykres 12: Rozkład zmiennej „Ilość spożywanej wody” i liczebność w każdej z grup



Wykres 13: Rozkład zmiennej „Częstotliwość ćwiczeń” i liczebność w każdej z grup



Wykres 14: Rozkład zmiennej „poziom doświadczenia” i liczebność w każdej z grup



Wykres 15: Rozkład zmiennej „BMI” i liczebność w każdej z grup

Głównym wnioskiem wynikającym z analizy wykresów jest wyraźne zróżnicowanie liczebności w poszczególnych grupach. Przykładowo, zmienna „poziom doświadczenia” (wykres nr 14) pokazuje, że w grupie 1 dominuje niższy poziom spalonych kalorii, podczas gdy w grupie 3 przeważają osoby, które spaliły ponad 894 kalorie.

Tego rodzaju rozkłady sugerują, że zmienne takie jak „poziom doświadczenia” mogą odgrywać istotną rolę w dalszej analizie i wpływać na skuteczność modelu predykcyjnego.

Information Value oraz Macierz Korelacji

Obliczone wartości Information Value dostarczają kilku istotnych wniosków. Przede wszystkim widać wyraźnie, że jedynie **cztery zmienne bardzo dobrze tłumaczą zmienną objaśnianą, a dwie kolejne wykazują średnią siłę wyjaśniającą**. Pozostałe zmienne wydają się mieć niewielką wartość predykcyjną i prawdopodobnie nie będą przydatne w dalszym modelowaniu.

Kolejnym krokiem była analiza macierzy korelacji. Ze względu na występowanie istotnych współzależności pomiędzy niektórymi zmiennymi, zdecydowałem się na usunięcie części z nich. Do wykluczonych zmiennych należą: płeć, wzrost, waga oraz poziom doświadczenia.

Warto również zauważyć silną korelację pomiędzy czasem trwania sesji a liczbą spalonych kalorii. Ponieważ liczba spalonych kalorii stanowi zmienną objaśnianą, obecność tej zależności można uznać za pozytywny sygnał – wskazuje ona na potencjalną istotność zmiennej „czas trwania” w dalszej analizie.






Zmienna	Info_value
 Session_Duration (hours)	4.23
 Experience_Level	2.46
 Fat_Percentage	1.64
 Workout_Frequency (days/week)	1.32
 Avg_BPM	0.38
 Water_Intake (liters)	0.32
 Weight (kg)	0.08
 BMI	0.07
 Age	0.06
 Gender	0.05
 Height (m)	0.03
 Workout_Type	0.01
 Resting_BPM	0.01
 Max_BPM	0.01

Tabela 2: Information Value poszczególnych zmiennych

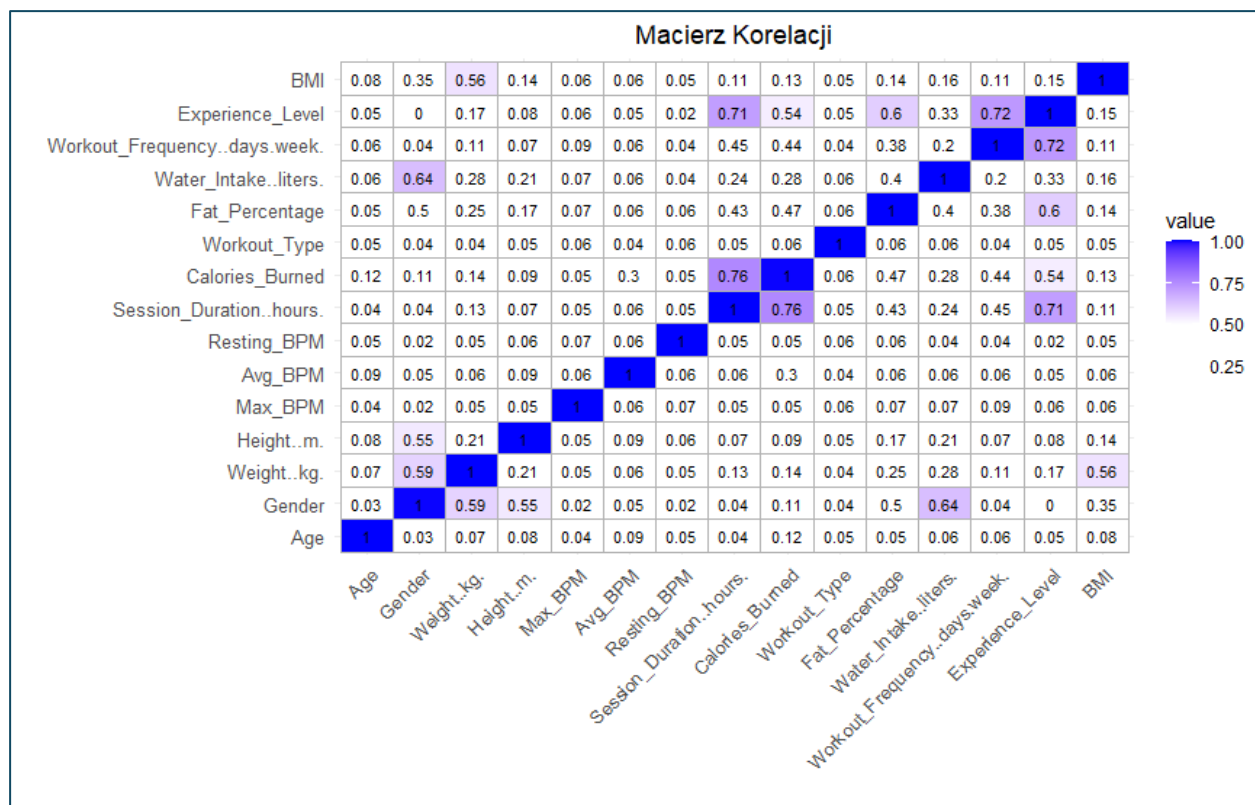


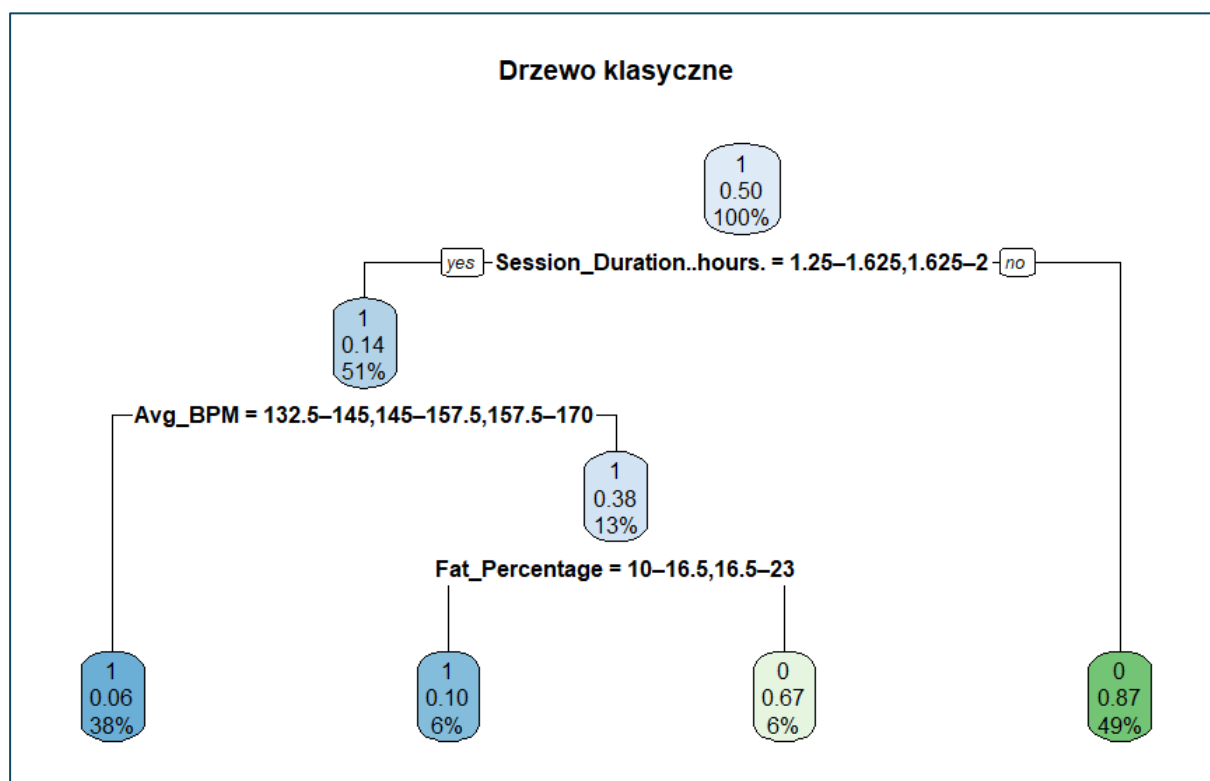
Tabela 3: Macierz korelacji pomiędzy zmiennymi

Drzewo klasyfikacyjne

Poniższe drzewo klasyfikacyjne zostało wygenerowane z wykorzystaniem parametru $cp = 0.02$, co miało na celu ograniczenie jego głębokości i zapobieżenie nadmiernemu dopasowaniu. Z analizy wynika, że pierwszą i zarazem najważniejszą zmienną decyzyjną jest **czas trwania sesji treningowej**.

Co ciekawe, drzewo wskazuje dwa istotne przedziały czasu: od **1,25 godziny do 1,625 godziny** oraz od **1,625 godziny do 2 godzin**. W praktyce granica między tymi przedziałami jest dość płynna, co sugeruje, że mogłyby zostać połączone w jedną kategorię.

Oprócz czasu trwania sesji, model wykorzystuje jeszcze dwie inne zmienne: **procent tkanki tłuszczowej** oraz **średnie tętno** osoby ćwiczącej.

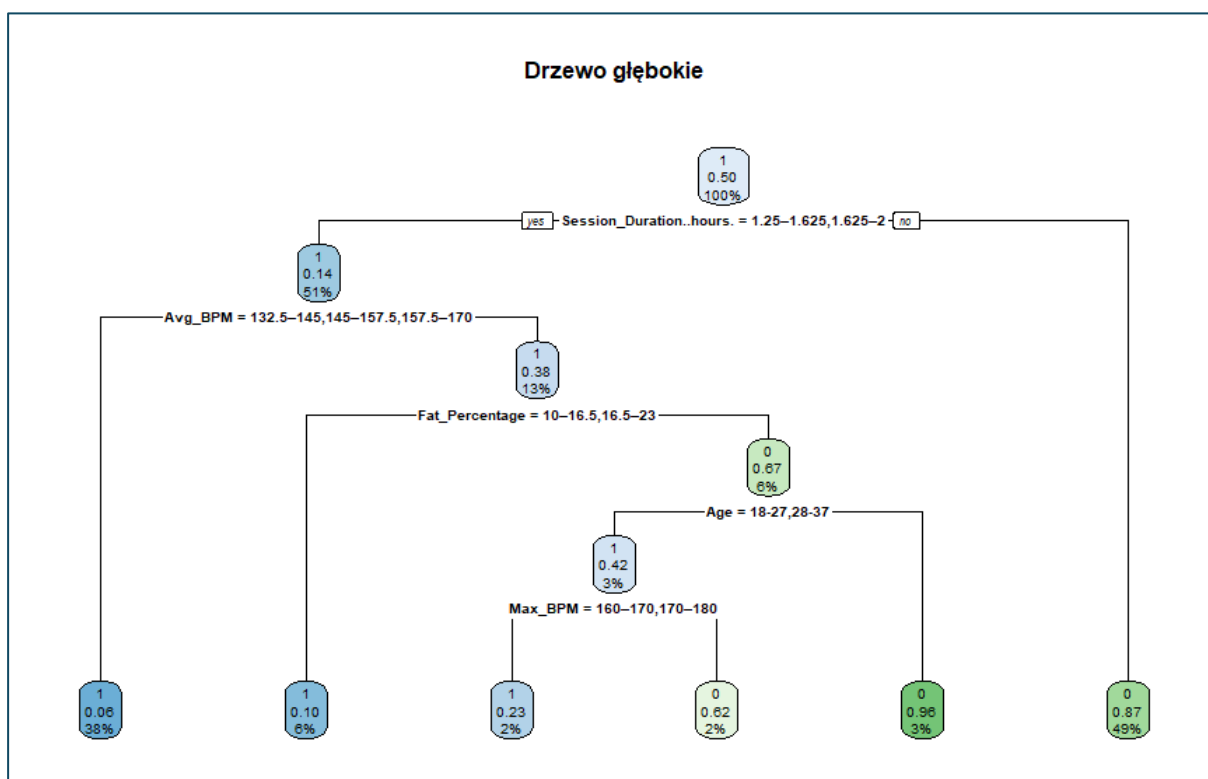


Wykres 16: Drzewo klasyfikacyjne z parametrem $cp = 0.02$

Drzewo klasyfikacyjne (Głębokie)

Głębsze drzewo klasyfikacyjne zostało wygenerowane z użyciem parametru $cp = 0.0075$. Umożliwiło to stworzenie bardziej złożonej struktury, która uwzględnia większą liczbę zmiennych, a jednocześnie zachowuje wymóg, aby końcowe liście zawierały co najmniej 1% wszystkich obserwacji w zbiorze treningowym.

W porównaniu do poprzedniego drzewa, model uwzględnił dodatkowo dwie nowe zmienne: **wiek osoby ćwiczącej** oraz **maksymalne tętno osiągnięte podczas treningu**.



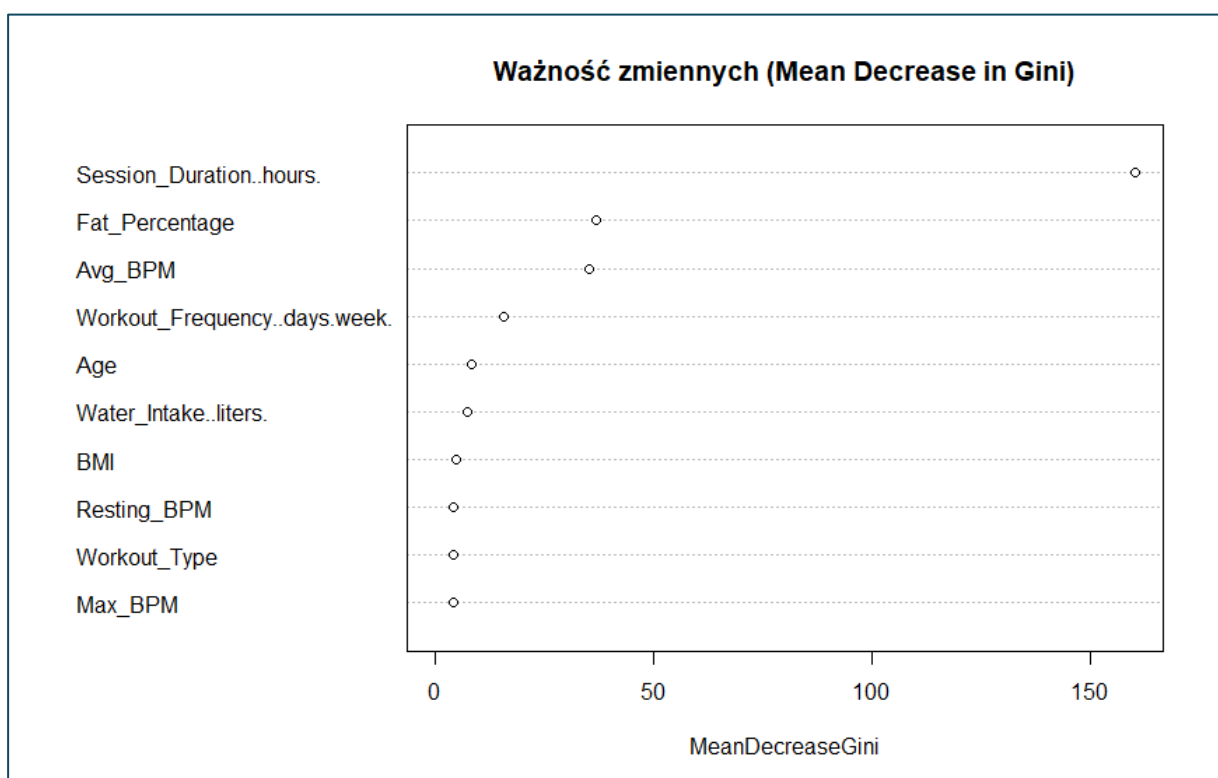
Wykres 17: Drzewo klasyfikacyjne z parametrem $cp = 0.0075$

Las losowy

Trzecim i ostatnim modelem klasyfikacyjnym był **las losowy**. Aby uniknąć problemu przeuczenia, zastosowano kilka ograniczeń, które miały na celu poprawę ogólnej jakości predykcji.

Po pierwsze, ustawiono parametr $mtry = 4$, co oznacza, że przy tworzeniu każdego drzewa, algorytm losował 4 zmienne spośród dostępnych predyktorów. Kolejnym ograniczeniem była **maksymalna głębokość drzewa**, ustawiona na 3, co pozwoliło zapobiec nadmiernemu rozrostowi struktury. Dodatkowo określono **minimalną liczbę obserwacji w liściu** na poziomie 20, co oznacza, że każdy liść musiał zawierać co najmniej 2,5% danych treningowych, aby mógł zostać utworzony.

Model wskazał, że najistotniejszą zmienną predykcyjną jest – podobnie jak wcześniej – **czas trwania sesji treningowej**, który zdecydowanie wyróżnia się pod względem wpływu. Oprócz niego, las losowy uwzględnił także: **procent tkanki tłuszczowej**, **średnie tętno** oraz **liczbę dni przeznaczonych na trening**.



Wykres 18: Las losowy z parametrami : $mtry = 4$, $max.depth = 3$ i $nodesize = 20$

Porównanie wyników

	Model	Set	accuracy	MER	precision	sensitivity	specificity	F1
1	tree	Train	0.8845	0.1155	0.9341	0.8295	0.9404	0.8787
2	tree_deep	Train	0.8935	0.1065	0.9282	0.8550	0.9326	0.8901
3	rf	Train	0.9230	0.0770	0.9280	0.9186	0.9275	0.9233
4	tree	Test	0.9072	0.0928	0.9518	0.8495	0.9604	0.8977
5	tree_deep	Test	0.9072	0.0928	0.9310	0.8710	0.9406	0.9000
6	rf	Test	0.9227	0.0773	0.9239	0.9140	0.9307	0.9189

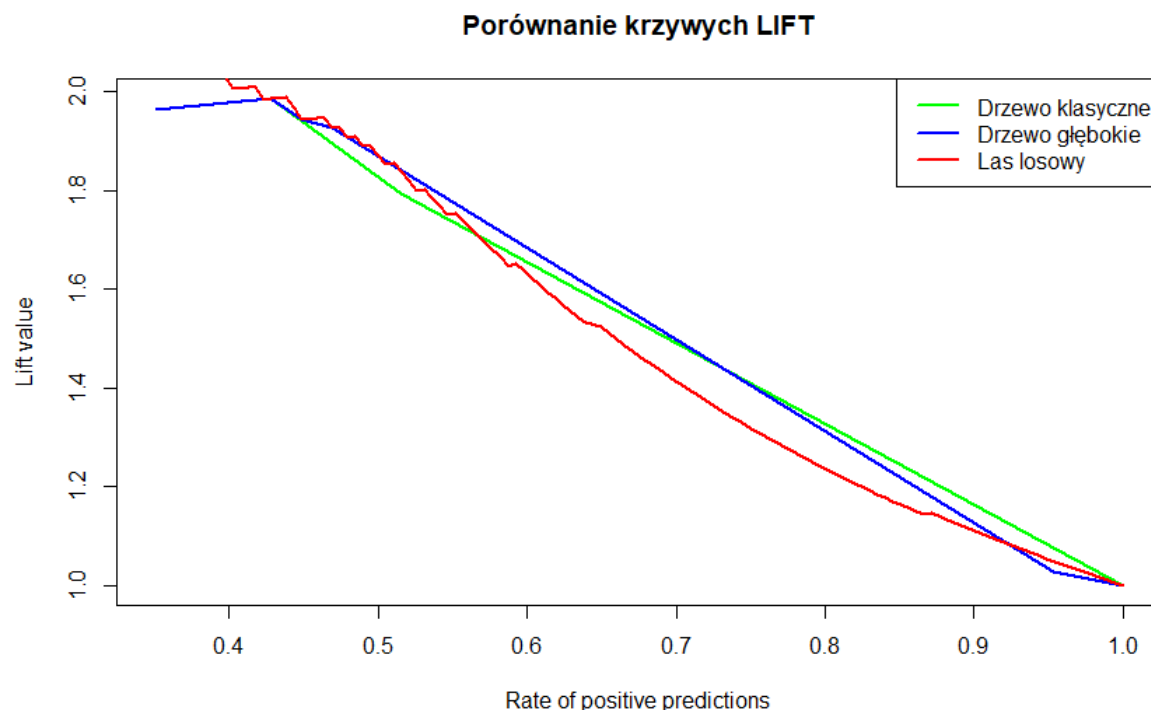
Tabela 4: Zestawienie wszystkich wyników razem

Porównując wyniki wszystkich modeli, można wyciągnąć kilka interesujących wniosków. Przede wszystkim zauważalne jest **podobieństwo wyników na zbiorze treningowym i testowym**, które zostały rozdzielone w proporcji 80:20. Taka zbieżność świadczy o **braku przeuczenia**, co jest istotnym aspektem przy ocenie jakości modeli predykcyjnych.

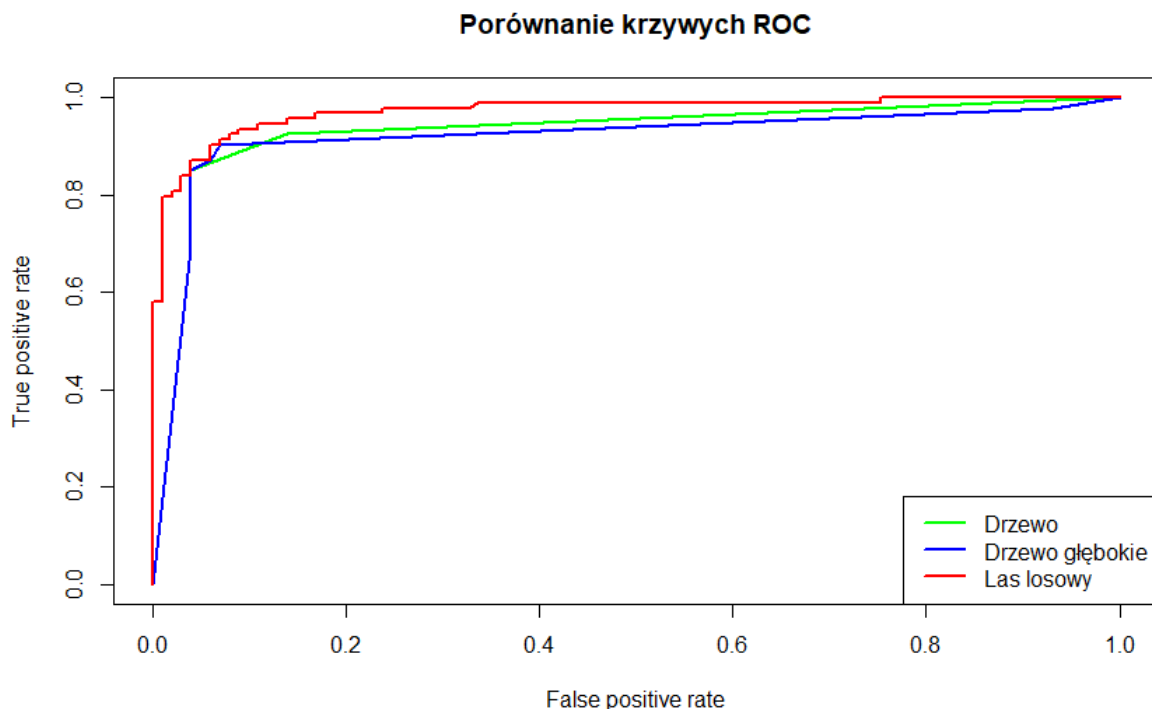
Głównym kryterium oceny skuteczności modelu w tym przypadku była **dokładność (accuracy)** – miara, która sprawdza się najlepiej w przypadku **zrównoważonych zbiorów danych**, takich jak nasz (gdzie zmienna objaśniana została utworzona na podstawie mediany).

Na podstawie tego kryterium można uznać, że **las losowy okazał się najskuteczniejszym modelem**. Co więcej – choć nie było to kluczowe w procesie oceny – osiągnął on również najwyższe wyniki pod względem **F1-score oraz czułości (sensitivity)**.

Krzywe ROC i Krzywe LIFT



Wykres 19 : Porównanie krzywych LIFT pomiędzy modelami



Wykres 20: Porównanie krzywych ROC pomiędzy modelami

W obu analizowanych wykresach las losowy zdecydowanie przewyższa pozostałe modele. Na wykresie **krzywej LIFT** znajduje się najwyżej, co oznacza, że osiąga **najwyższą skuteczność w identyfikacji pozytywnych przypadków**. Z kolei na wykresie **krzywej ROC** model ten plasuje się **najbliżej lewego górnego rogu**, co jest charakterystyczne dla klasyfikatora o najwyższej jakości.

Oba wykresy stanowią **dotychczasowe potwierdzenie**, że **las losowy jest najlepszym spośród zastosowanych modeli klasyfikacyjnych**.

Podsumowanie

Pomimo eliminacji kilku zmiennych w trakcie analizy, wszystkie finalne modele osiągnęły bardzo dobre wyniki – każdy z nich uzyskał **dokładność (accuracy) powyżej 0,9**. Spośród nich **najlepszym okazał się las losowy**, który nie tylko wyróżnił się najwyższymi miarami skuteczności, ale również potwierdził logiczne zależności między zmiennymi.

Na tej podstawie można uznać, że kluczowe czynniki wpływające na ilość spalonych kalorii podczas sesji treningowej to: **czas trwania sesji, procent tkanki tłuszczowej, średnie tętno** oraz **częstotliwość treningów**. Co istotne, wyniki te są **spójne z intuicją i wiedzą praktyczną** na temat aktywności fizycznej.