Metody Systemowe i Decyzyjne

Laboratorium

Zofia Różańska

280526

Semestr letni 2024/2025

Raport - Analiza Zależności Między Stylami Życia a Parametrami Zdrowotnymi

Zofia Różańska

27.03.2025

1 Metodologia

1.1 Źródło Danych

Analiza oparta jest na zbiorze danych obejmującym kluczowe zmienne związane ze zdrowiem, aktywnością fizyczną, stylem życia oraz cechami demograficznymi.

1.2 Zmienne Kluczowe

- Wiek
- Wzrost
- Waga
- Płeć
- Częstotliwość aktywności fizycznej
- Czas użycia technologii
- Historia rodzinna
- Głowny środek transportu
- Spożycie wysokokalorycznych pokarmów
- Monitorowanie poziomu spożycia kalorii
- Poziom otyłości

2 Analiza Korelacji - Macierz Korelacji

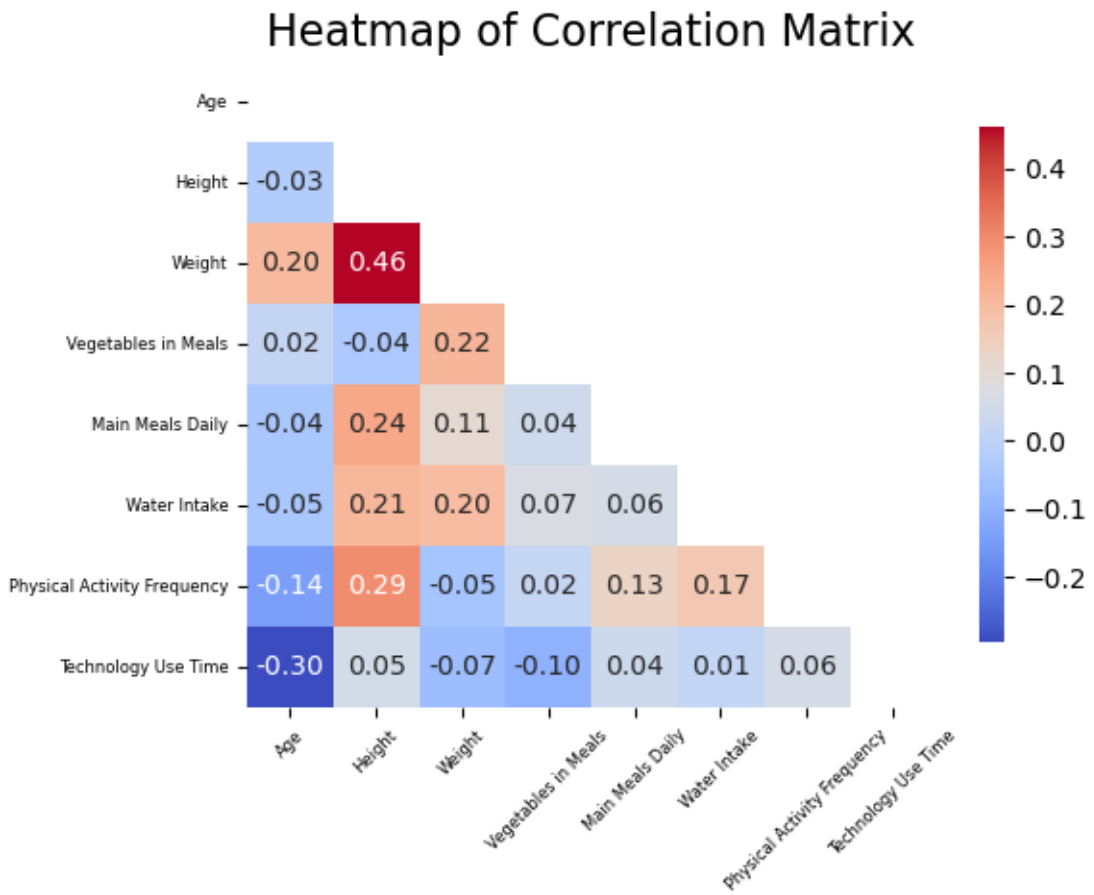


Figure 1: Heatmapa Koleracji

2.0.1 Kluczowe Obserwacje:

- Najsilniejsza dodatnia korelacja: Waga i Wzrost (0.46)
- Umiarkowane korelacje:
  - Aktywność fizyczna a wzrost (0.29)
  - Liczba posiłków a wzrost (0.24)
- Negatywne korelacje:
  - Czas użycia technologii a wiek (-0.30)
  - Wiek a aktywność fizyczna (-0.14)

2.0.2 Wnioski

- W badanej grupie istnieje silna dodatnia korelacja między wagą a wzrostem (0.46), co jest zjawiskiem naturalnym. Im większa waga, tym zazwyczaj większy wzrost, co może wskazywać na proporcjonalny rozwój organizmu.
- Co zaskakujące, zauważono umiarkowaną korelację między aktywnością fizyczną a wzrostem (0.29). Oznacza to, że osoby bardziej aktywne fizycznie mają tendencję do osiągania nieznacznie większego wzrostu.
- Liczba posiłków wykazuje umiarkowaną korelację ze wzrostem (0.24), sugerując, że osoby o większym wzroście częściej spożywają posiłki.
- Zaobserwowano negatywną korelację między czasem użycia technologii a wiekiem (-0.30). Oznacza to, że młodsze osoby spędzają więcej czasu korzystając z urządzeń elektronicznych, co wydaje się dosyć powszechne w dzisiejszych czasach.
- Wystąpiła słaba negatywna korelacja między wiekiem a aktywnością fizyczną (-0.14), co sugeruje tendencję do zmniejszania aktywności fizycznej wraz z wiekiem, choć zależność jest stosunkowo nieznaczna.

Warto podkreślić, że korelacja nie oznacza bezpośredniego związku przyczynowo-skutkowego, a jedynie wskazuje na współwystępowanie pewnych zjawisk w badanej grupie.

3 Analiza Demograficzna

3.1 Rozkład Wzrostu Z Podziałem Na Płeć

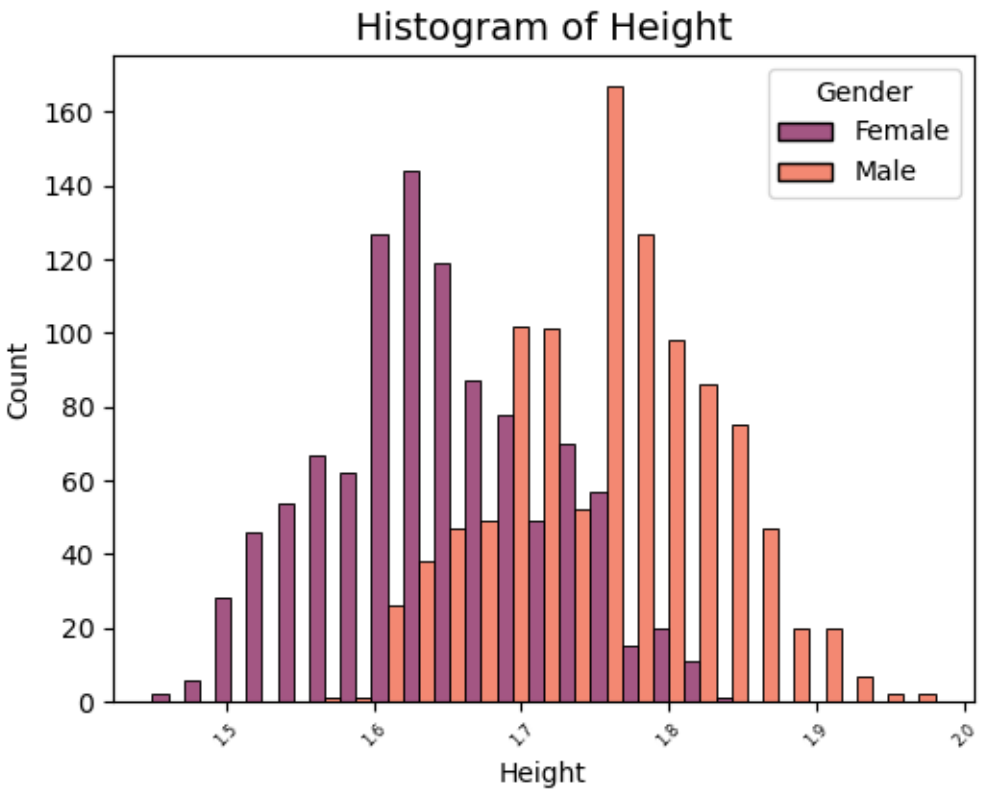


Figure 2: Histogram Wzrostu

3.1.1 Kluczowe Wnioski:

- Rozkład wzrostu zbliżony do jest normalnego
- Wyraźne różnice między płciami
- Mężczyźni mają tendecję do osiągania wyższych wartości niż kobiety
- Wśród kobiet większość wzrostów koncentruje się między 160-165 cm, a wsródmężnyzm między 175-180cm

3.2 Rozkład Wagi Z Podziałem Na Płeć

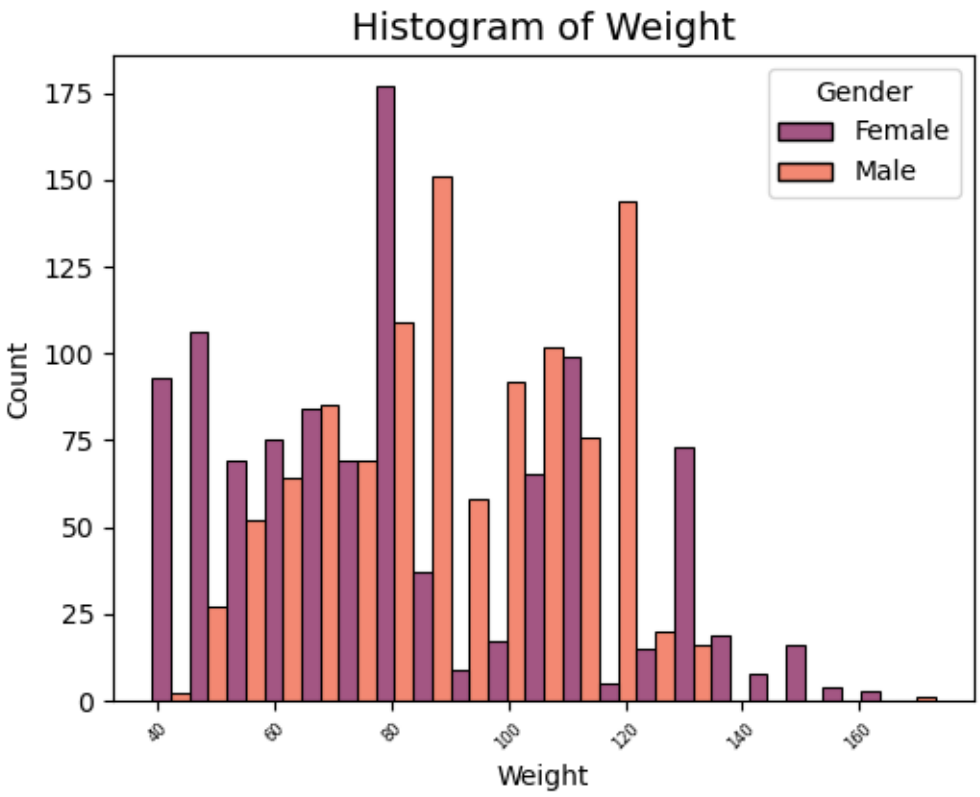


Figure 3: Histogram Wagi

3.2.1 Kluczowe Wnioski:

- Rozkład wagi zbliżony do normalnego, z lekką asymetrią w stronę wyższych wartości
- Nieznaczne różnice między płciami
- Mężczyźni mają tendecję do osiągania wyższych wartości niż kobiety
- Wśród kobiet większość wartości koncentruje się około 80kg, a wsródmężnyzm około 90kg

3.3 Wiek a Środek Transportu

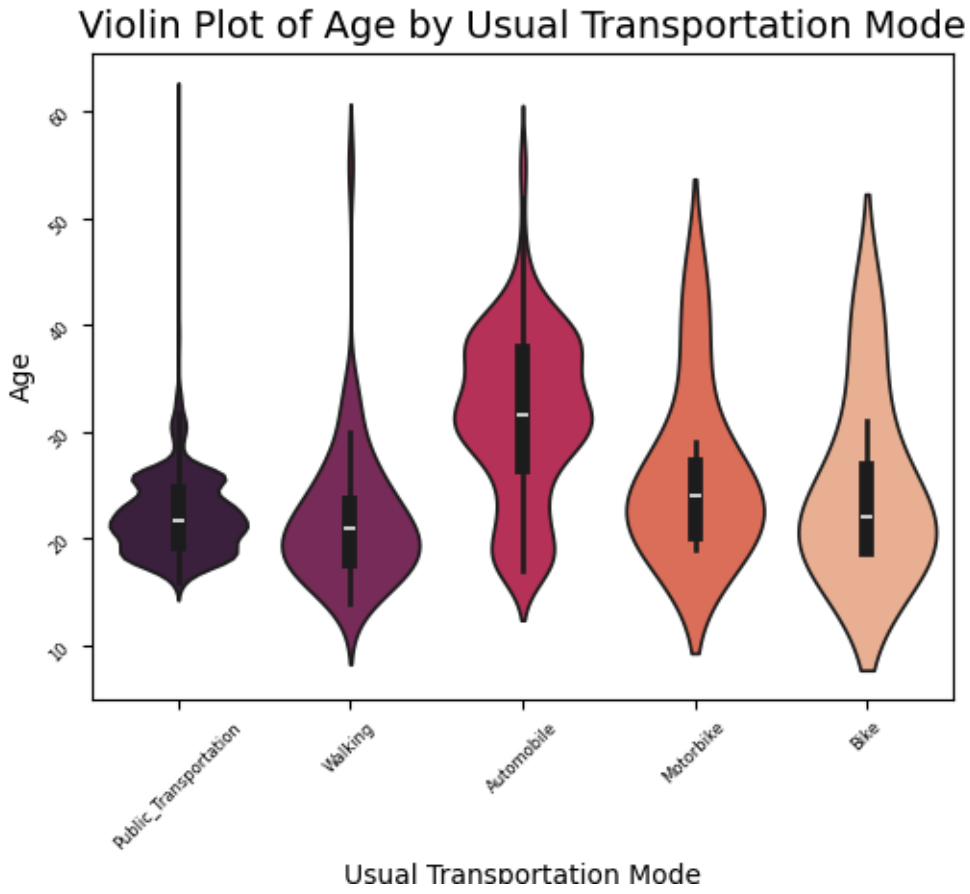


Figure 4: Violin Plot Wiek A Środek Transportu

3.3.1 Obserwacje:

- Transport publiczny: największe zróżnicowanie wiekowe
- Chodzenie: Stosunkowo równomierny rozkład, jednak najbardziej skoncentrowany około 20 lat
- Samochód: średni wiek użytkowników, skoncentrowany w wieku 30-40 lat
- Motocykl: koncentracja młodszych użytkowników, około 25 lat
- Rower: znaczne zróżnicowanie użytkowników

3.3.2 Wnioski:

Każdy środek transportu charakteryzuje się nieco innym rozkładem wiekowym użytkowników, co wskazuje na preferencje transportowe różnych grup wiekowych.

4 Analiza Technologiczna

4.1 Czas Użycia Technologii a Wiek

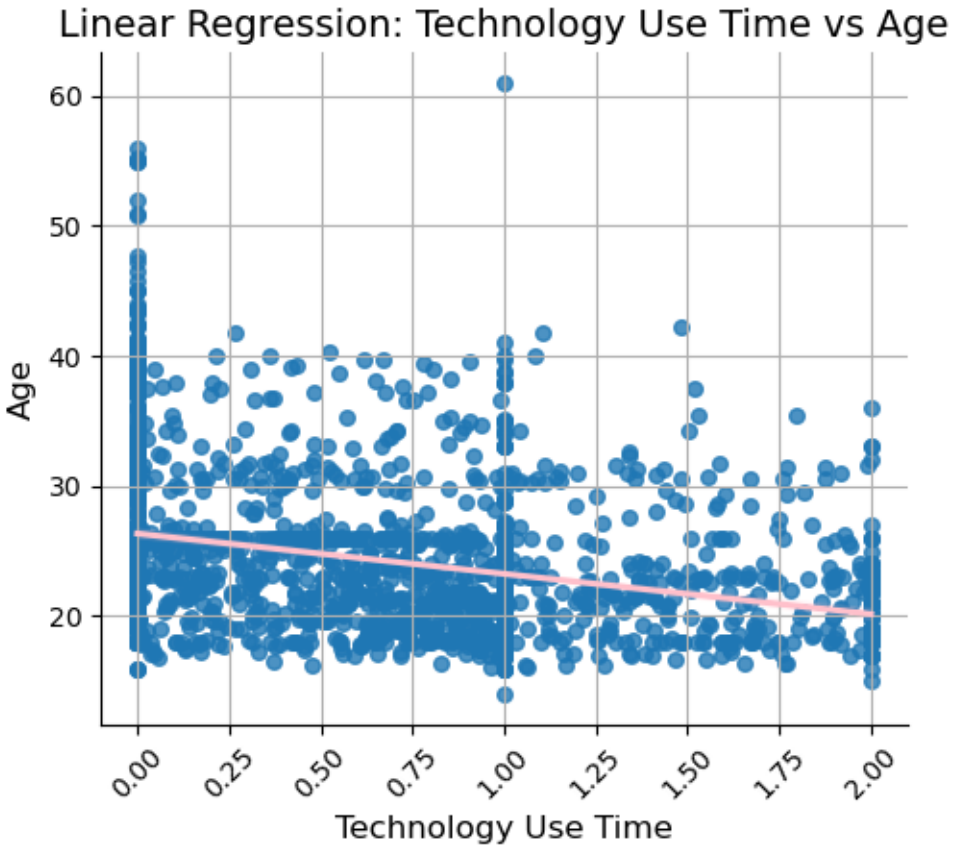


Figure 5: Regresja Technologia-Wiek

4.1.1 Kluczowe Wnioski:

- Umiarkowana negatywna korelacja (-0.3)
- Młodsi użytkownicy więcej czasu spędzają z technologią

5 Analiza Aktywności Fizycznej

5.1 Częstotliwość Aktywności Fizycznej

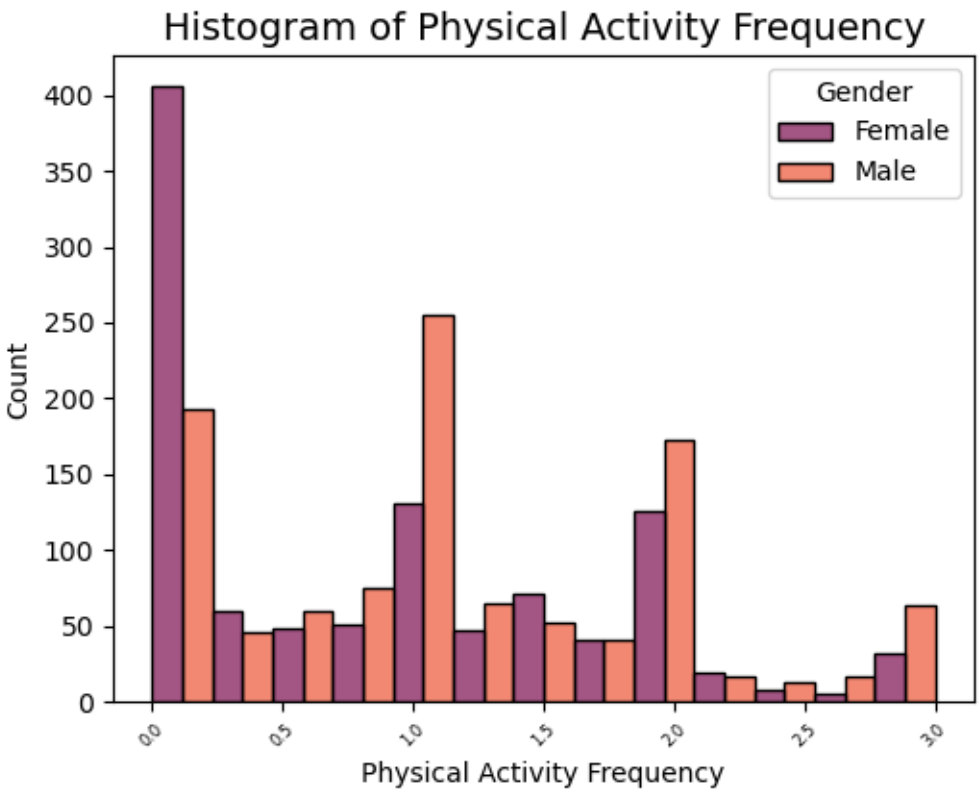


Figure 6: Histogram Aktywności Fizycznej

5.1.1 Kluczowe Obserwacje:

- Dominacja niskiej i umiarkowanej aktywności
- Nieznaczne różnice między płciami
- Mężczyźni wykazują nieznacznie wyższą aktywność

5.2 Aktywność a Wzrost

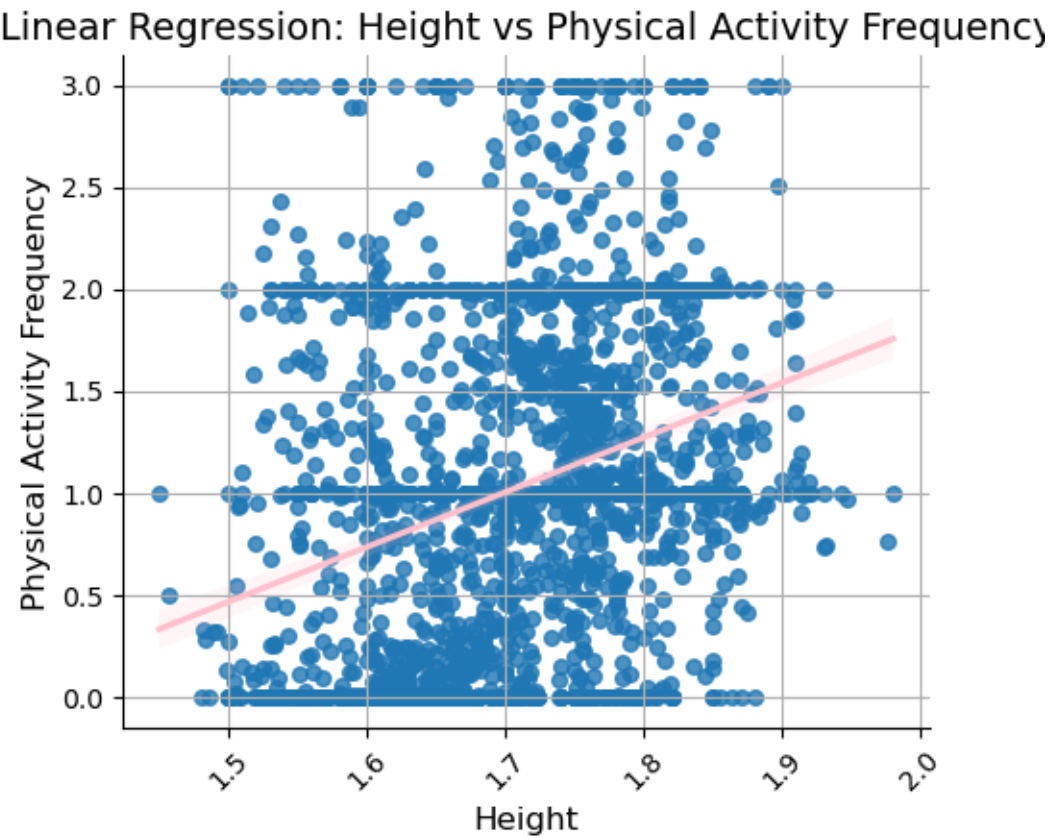


Figure 7: Regresja Liniowa Aktywność-Wzrost

5.2.1 Wnioski:

- Lekko pochylona linia regresji potwierdza słabą, ale pozytywną korelację między wzrostem a aktywnością fizycznąm która widoczna jest również na heatmapie korelacji.
- Istnieje delikatna, lecz zauważalna dodatnia zależność między wzrostem a częstotliwością aktywności fizycznej - wraz ze wzrostem wzrostu nieznacznie zwiększa się częstotliwość aktywności. Duży Rozrzut Danych
- Mimo trendu liniowego, widoczny jest znaczny rozrzut punktów, co oznacza, że wzrost nie jest jedynym ani decydującym czynnikiem wpływającym na aktywność fizyczną.

6 Analiza Wagowa

6.1 Waga a Wzrost

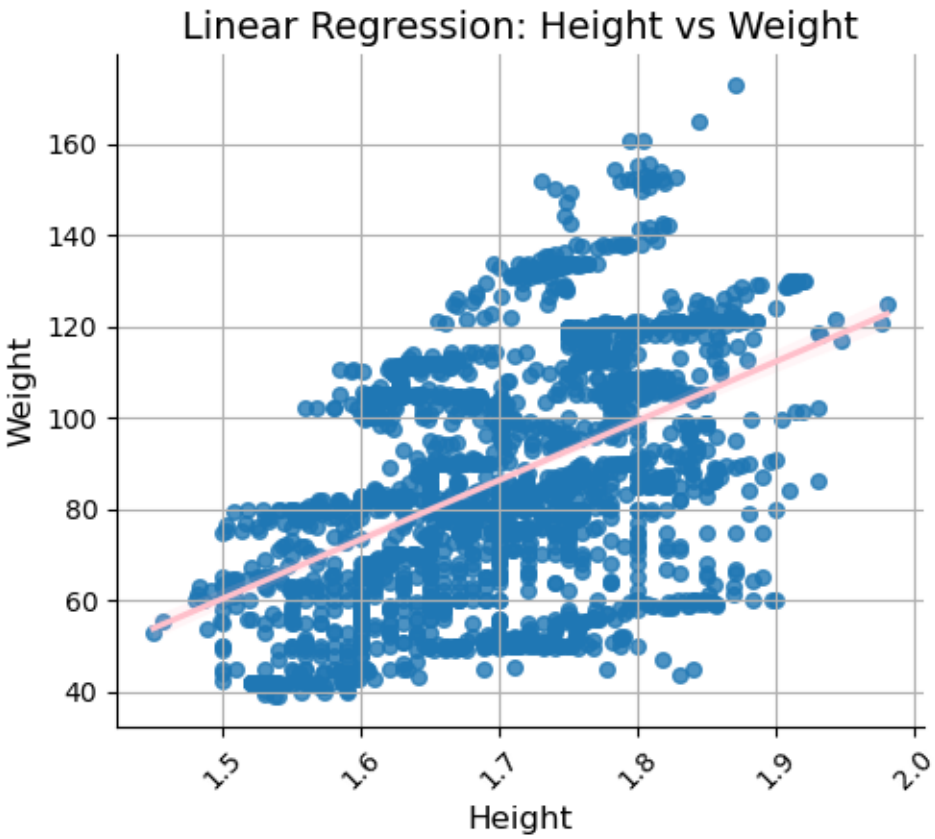


Figure 8: Regresja Waga-Wzrost

6.1.1 Kluczowe Wnioski:

- Różowa linia regresji potwierdza silną dodatnią korelację i pozytywny związek między wzrostem a wagą, pokazując systematyczny wzrost wagi wraz z wysokością.
- Mimo wyraźnego trendu, widoczny jest znaczny rozrzut punktów, co oznacza, że wzrost nie jest jedynym czynnikiem determinującym wagę.

6.2 Waga a Wiek

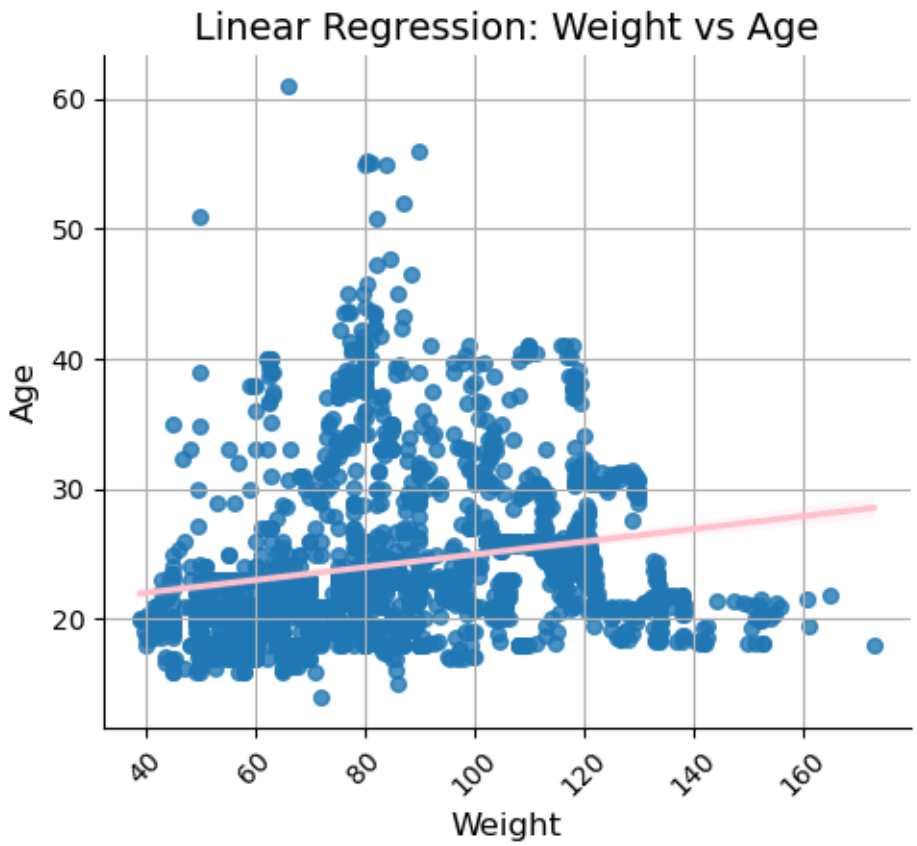


Figure 9: Regresja Waga-Wiek

6.2.1 Kluczowe Wnioski:

- Różowa linia regresji wskazuje na słaby, ale pozytywny związek między wiekiem a wagą.
- Istnieje delikatka tendencja do zwiększania masy ciała wraz z wiekiem

6.3 Waga a Monitorowanie Kalorii

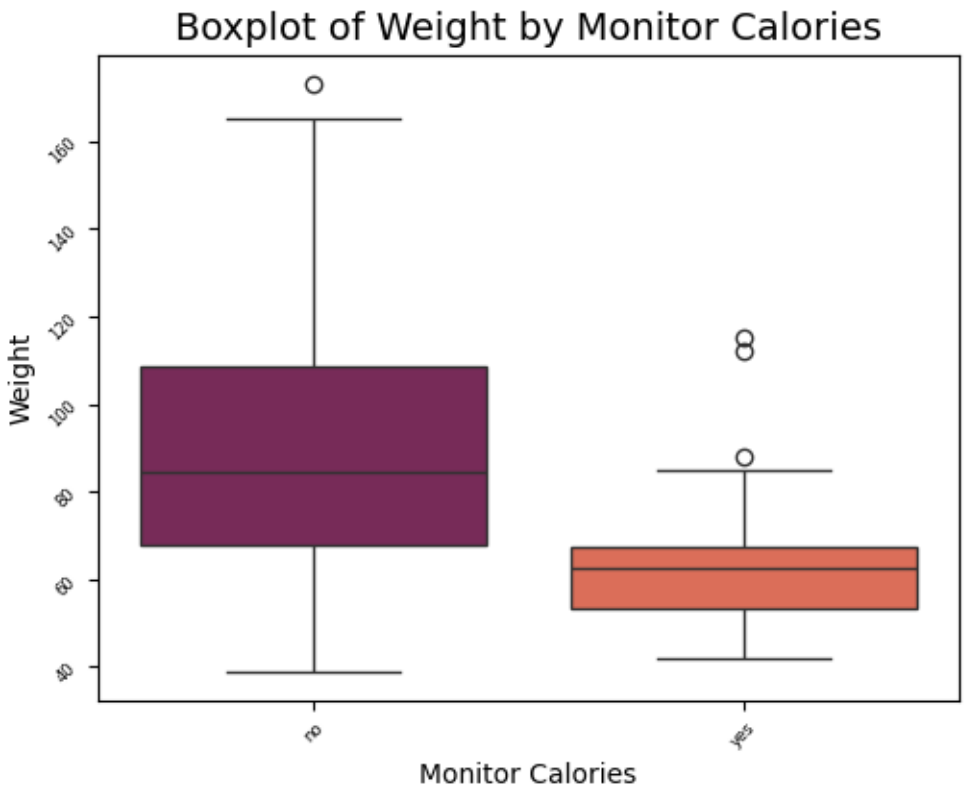


Figure 10: Boxplot Wagi wg Monitorowania Kalorii

6.3.1 Kluczowe Ustalenia:

- Wyraźny wpływ monitorowania kalorii na wagę
- Osoby niemonitorujące kalorii mają wyraźnie wyższą medianę wagi.
- Grupa niemonitorująca kalorie charakteryzuje się większym rozrzutem wagi.
- Osoby monitorujące kalorie mają bardziej zwartą dystrybucję wagi.

7 Analiza Otyłości

7.1 Otyłość a Historia Rodzinna

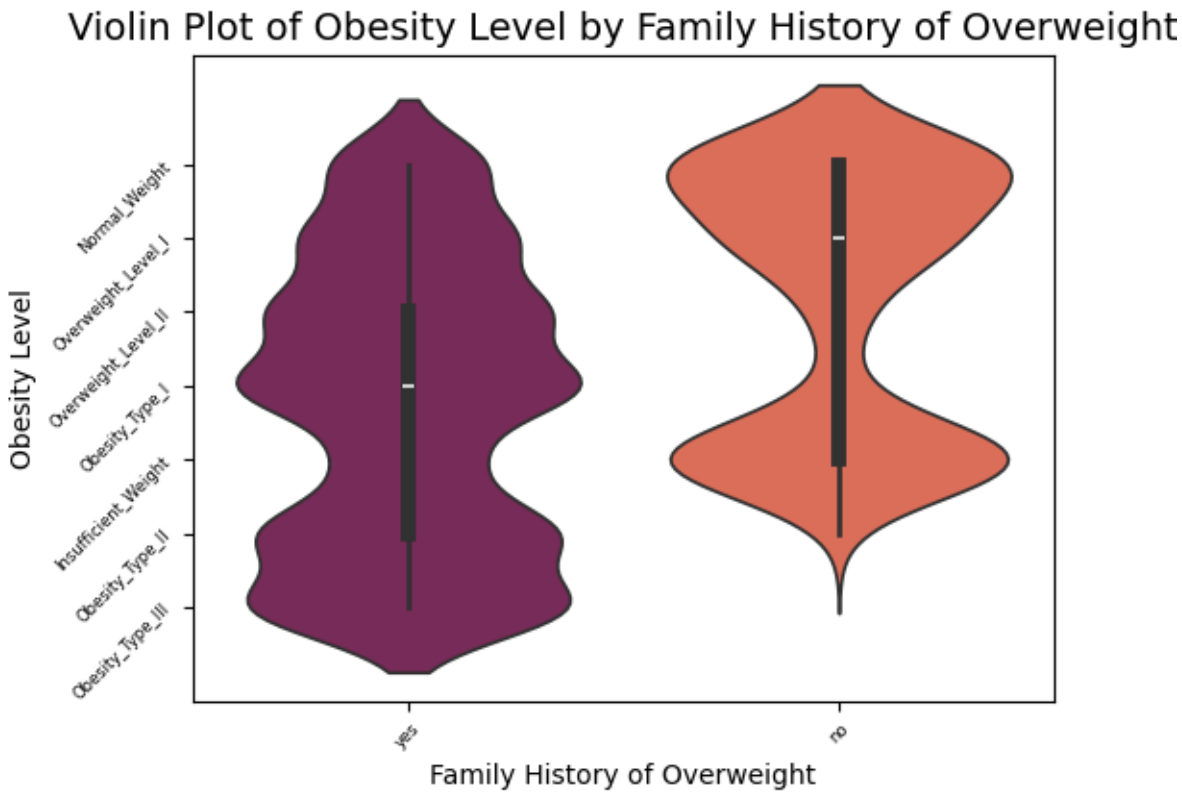


Figure 11: Violin Plot Otyłości wg Historii Rodzinnej

7.1.1 Kluczowe Obserwacje:

- Silny wpływ historii rodzinnej na poziom otyłości
- Szerszy zakres wariancji dla osób z obciążeniem genetycznym

7.2 Otyłość a Wysokokaloryczne Pokarmy

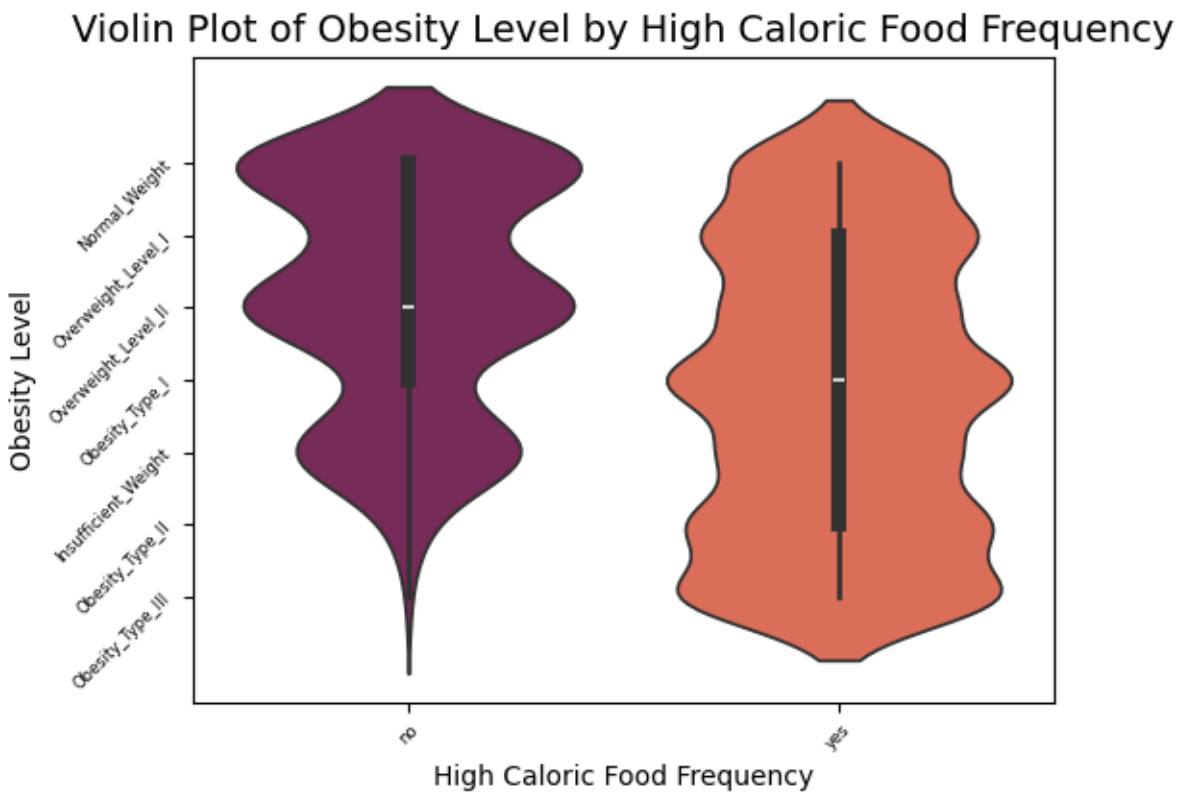


Figure 12: Violin Plot Otyłości wg Wysokokalorycznych Pokarmów

7.2.1 Kluczowe Obserwacje:

- Wyraźna korelacja spożycia wysokokalorycznych pokarmów z poziomem otyłości
- Większe zróżnicowanie poziomów otyłości przy wysokiej częstotliwości spożycia

7.3 Otyłość a Monitorowanie Kalorii

Figure 13: Violin Plot Otyłości wg Monitorowania Kalorii

7.3.1 Kluczowe Obserwacje:

- Osoby niemonitorujące kalorii wykazują szerszą rozpiętość poziomów otyłości.
- osoby monitorujące kalorie mają zdecydowanie niższą tendencję do otyłości

7.4 Otyłość a Środki Transportu

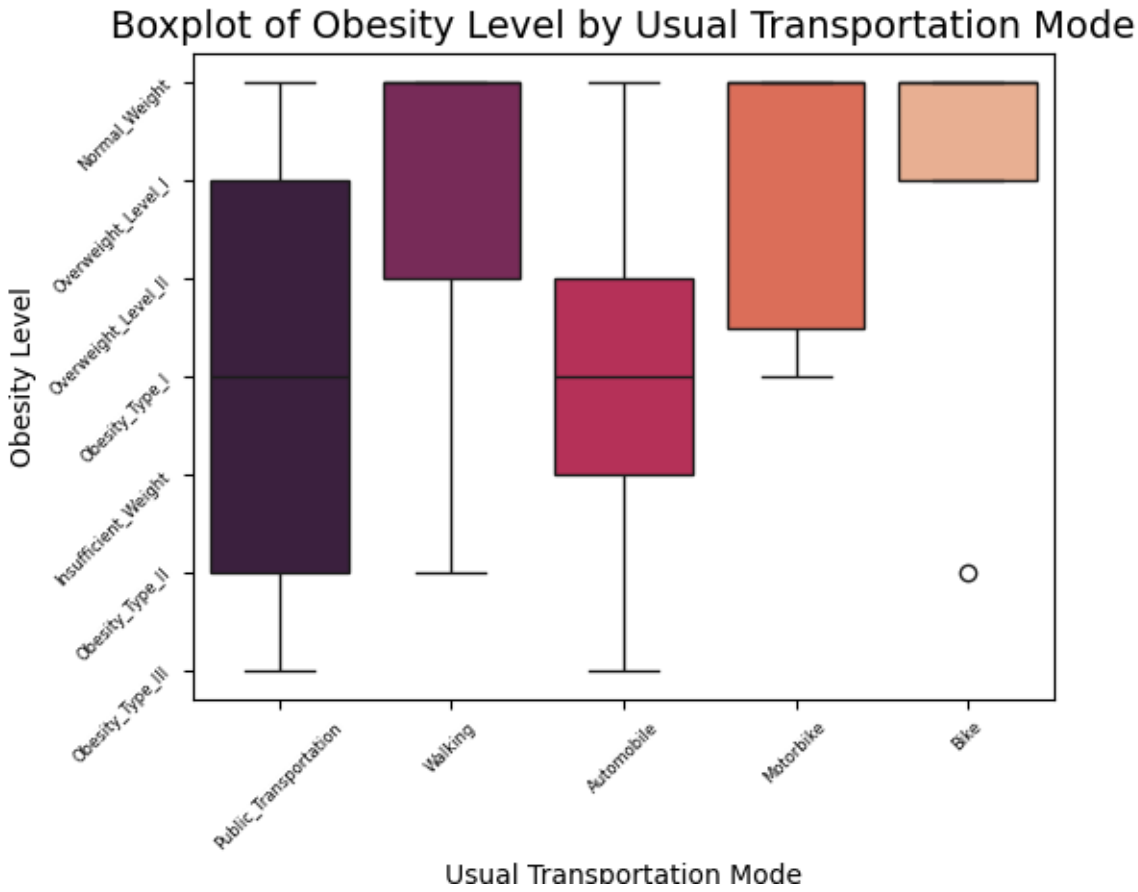


Figure 14: Boxplot Otyłości wg Głównego Środka Transportu

7.4.1 Kluczowe Obserwacje:

- Między grupami widoczne są istotne różnice w poziomie otyłości w zależności od preferowanego środka transportu.
- Rower
  - Najniższy poziom otyłości
  - Najbardziej zwarta dystrybucja
- Chodzenie
  - Niski poziom otyłości
  - Stosunkowo wąski zakres wartości
- Transport Publiczny
  - Średni poziom otyłości
  - Największe zróżnicowanie
- Mokocykl
  - Koncentracja wśród najniższych poziomów otyłości
- Samochód
  - Duże zróżnicowanie

8 Wnioski Końcowe

Analiza przeprowadzona na dostępnych danych pozwoliła na wyodrębnienie kluczowych czynników wpływających na zdrowie i otyłość. Wyniki podkreślają rolę zarówno stylu życia, jak i uwarunkowań genetycznych w kształtowaniu masy ciała i kondycji fizycznej.

8.1 Aktywność fizyczna jako kluczowy czynnik zdrowotny

- Osoby regularnie uprawiające aktywność fizyczną mają tendencję do niższego poziomu otyłości i lepszego stanu zdrowia.
- Zauważono pozytywną korelację między wzrostem a częstotliwością aktywności fizycznej, co może sugerować większą świadomość zdrowotną u osób wyższych.
- Spadek aktywności fizycznej wraz z wiekiem wskazuje na potrzebę promowania aktywnego trybu życia wśród starszych grup wiekowych.

8.2 Znaczenie diety i monitorowania kalorii

- Osoby monitorujące swoje spożycie kalorii mają mniejszą tendencję do otyłości i utrzymują stabilniejszą wagę.
- Spożywanie wysokokalorycznych pokarmów wykazuje silny wpływ na wzrost poziomu otyłości, a brak kontroli nad dietą prowadzi do większego rozrzutu wagi.
- Liczba posiłków wykazuje umiarkowaną korelację z wzrostem, co sugeruje, że osoby wyższe mogą mieć większe zapotrzebowanie energetyczne.

8.3 Genetyka i historia rodzinna

- Historia rodzinna otyłości istotnie wpływa na tendencję do zwiększonej masy ciała, co potwierdza rolę czynników genetycznych w predyspozycji do nadwagi.
- Osoby z obciążeniem genetycznym wykazują większy zakres poziomów otyłości, co podkreśla znaczenie świadomego podejścia do diety i aktywności fizycznej w tej grupie.

8.4 Technologia i jej wpływ na zdrowie

- Czas korzystania z urządzeń elektronicznych wykazuje negatywną korelację z wiekiem, czyli młodsze osoby spędzają więcej czasu przed ekranem.
- Nadmierne korzystanie z technologii może wpływać na zmniejszenie aktywności fizycznej i przyczyniać się do wzrostu masy ciała.

8.5 Środek transportu a zdrowie

- Osoby korzystające z aktywnych form transportu, takich jak rower czy chodzenie, mają niższy poziom otyłości i lepszą kondycję zdrowotną.
- Użytkownicy samochodów wykazują większą zmienność w poziomie masy ciała, co może być związane z siedzącym trybem życia.

9 Podsumowanie

Wyniki analizy jednoznacznie wskazują, że aktywność fizyczna, kontrola diety oraz wybór środków transportu mają istotny wpływ na zdrowie i poziom otyłości. Genetyka również odgrywa istotną rolę, jednak to świadome wybory związane ze stylem życia w największym stopniu determinują naszą kondycję fizyczną. Promowanie aktywnego trybu życia, ograniczenie spożycia wysokokalorycznych pokarmów oraz świadome monitorowanie diety mogą znacząco wpłynąć na poprawę zdrowia społeczeństwa.

Raport – ewaluacja modeli uczenia maszynowego

Zofia Różańska, 280526

01.05.2025

Wyniki ewaluacji modeli uczenia maszynowego

CLASSIFICATION ***Y = OBESITY LEVEL***			**TRAINING**	**VALIDATION**	**TESTING**	
model		**device**	**accuracy / cross-entropy loss**			**training time**
Scikit-learn	**Logistic Regression**	CPU	90.0% / –	–	90.5% / –	–
	Support Vector Classification	CPU	96.3% / –	–	92.2% / –	–
	Decision Tree Classifier	CPU	96.3% / –	–	92.6% / –	–
Numpy	**Logistic Regression**	CPU	77.0% / 0.63	78.1% / 0.11	78.5% / 0.74	1000 epochs lr = 0.01 3.7s
PyTorch	**Logistic Regression**	CPU	86.2% / 0.49	87.2% / 0.389	87.2% / 0.47	1000 epochs lr = 0.01 66.7s
		GPU	86.3% / 0.48	87.4% / 0.391	87.9% / 0.46	1000 epochs lr = 0.01 35.5s

LINEAR REGRESSION ***Y = WEIGHT***			**TRAINING**	**TESTING**
MODEL		**DEVICE**	**MSE**	
Numpy	Closed Form	CPU	25.670	24.829
Scikit-learn	Linear Regression	CPU	25.670	24.829

Raport – optymalizacja modeli uczenia maszynowego

Zofia Różańska, 280526

25.05.2025

1. Wyniki cząstkowe

a. Cross-validation i ewaluacja modelu

Model: regresja logistyczna
Cost function: gradient descent
Biblioteka: PyTorch
Learning rate: 0.5
Epochs: 1000
Metoda: StratifiedKFold

Fold	Accuracy
1	88.86%
2	93.60%
3	90.28%

Na podstawie wyników dokładności uzyskanych podczas 3-krotnej walidacji krzyżowej używając metody StratifiedKFold, można stwierdzić, że model cechuje się dobrą jakością predykcji i stosunkowo stabilnym działaniem. Niewielki rozrzut wyników między foldami sugeruje, że model nie jest nadmiernie wrażliwy na sposób podziału danych, co może świadczyć o jego ogólnej stabilności. Można jednak zauważyć, że wyniki różnią się o około 4.7 punktów procentowych, co sugeruje, że istnieje jeszcze przestrzeń do optymalizacji.

Model: LogisticRegression(max_iter=1000)
Biblioteka: Scikit-learn
Metoda: cross_val_score (KFold)

Fold	Accuracy
1	77.13%
2	91.19%
3	90.90%

Fold 1 osiągnął zauważalnie niższy wynik niż foldy 2 i 3 – różnica przekraczała 14 punktów procentowych, co jest istotne w kontekście oceny stabilności modelu. Taka rozbieżność może wynikać z faktu, że w tym przypadku zastosowano metodę KFold, która dzieli dane na równe części, ale

nie uwzględnia proporcji klas w poszczególnych foldach. Oznacza to, że rozkład klas w każdym foldzie może być nierównomierny, a to z kolei może prowadzić do sytuacji, w której jeden z foldów — jak np. fold 1 — zawiera więcej obserwacji z trudnych lub rzadkich klas, co znacząco utrudnia zadanie klasyfikatorowi. W rezultacie model może mieć mniejsze szanse na nauczenie się reprezentatywnych wzorców w takim podzbiorze danych, a jego skuteczność w predykcji spada.

b. Wykresy zbieżności i analiza błędów

Model: regresja logistyczna
Cost function: gradient descent
Biblioteka: PyTorch
Learning rate: 0.5
Epochs: 1000

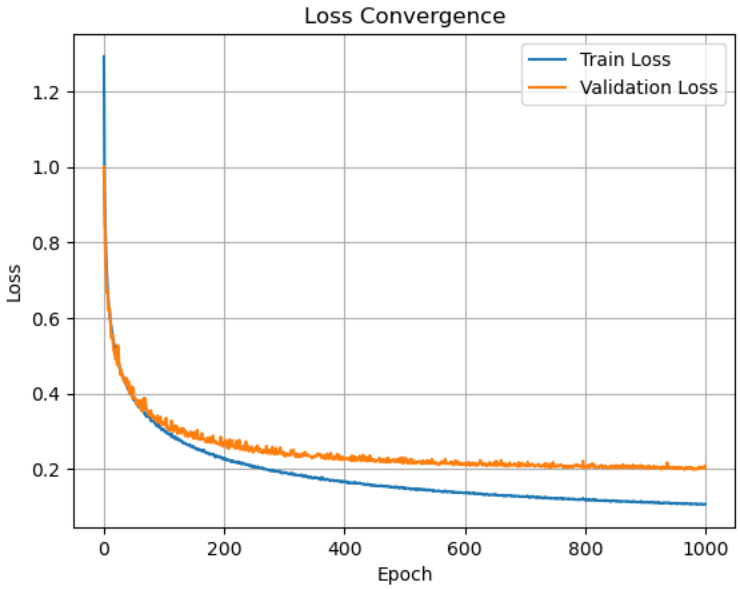
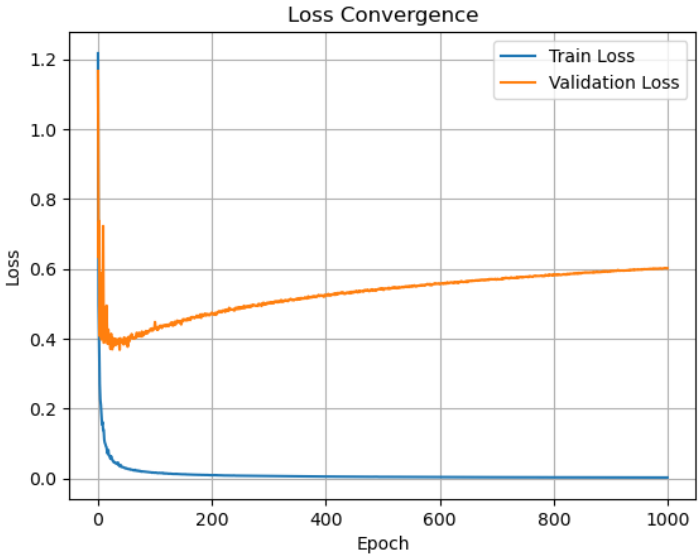


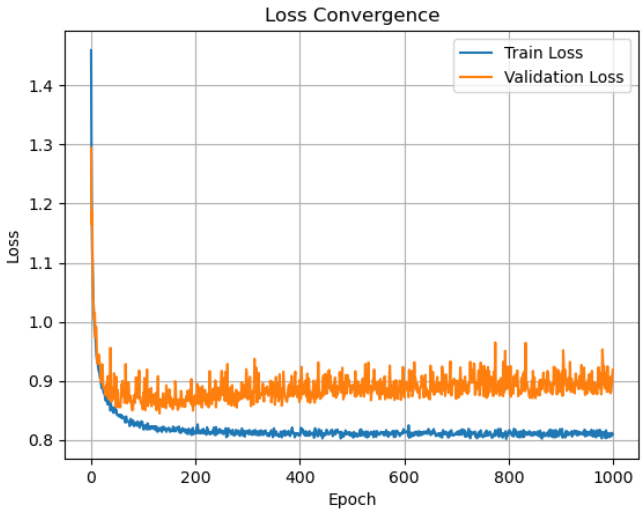
Figure 1 Wykres dla modelu o podstawowej złożoności

Dla modelu podstawowego wartość funkcji kosztu zarówno dla zbioru treningowego, jak i walidacyjnego systematycznie maleje wraz z liczbą epok, co świadczy o poprawnej zbieżności algorytmu gradientowego. Różnica między strata treningową, a walidacyjną jest stosunkowo niewielka i stabilna przez cały czas trwania uczenia, co oznacza, że model dobrze uczy się zależności danych. Nie zachodzi tutaj ani zjawisko nadmiernego dopasowania (overfitting), ani niedostatecznego dopasowania (underfitting)

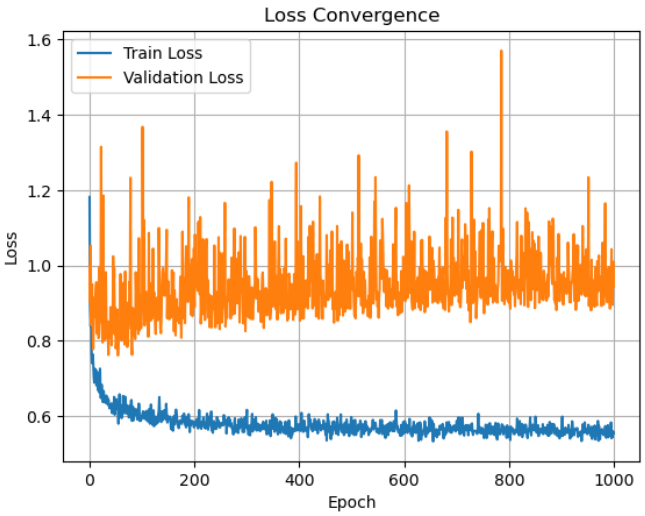


Wykres dla modelu ze zwiększoną złożonością (PolynomialFeatures degree = 2)

Model ze zwiększoną złożonością bardzo szybko minimalizuje stratę treningową, co świadczy o bardzo dobrym dopasowaniu danych treningowych. Tutaj jednak można zauważyć wyraźny rozdźwięk między stratą treningową a walidacyjną. Podczas gdy błąd treningowy szybko dąży do zera, strata walidacyjna po początkowym spadku zaczyna rosnąć co jest klasycznym objawem nadmiernego dopasowania. Model zapamiętuje dane treningowe i traci zdolność do generalizacji.



Wykres dla modelu o podstawowej złożoności uczonego na podstawie połowy (losowo) wybranych cech



Wykres dla modelu ze zwiększoną złożonością uczonego na podstawie połowy (losowo) wybranych cech

Jeśli zamiast wszystkich cech, weźmiemy tylko ich (losową) połowę, wyniki stają się bardziej chaotyczne i widać większe wahania strat zarówno treningowych i walidacyjnych względem poszczególnych epok. Dla obu modeli spada również wskaźnik dokładności – z około 90% do około 70%, co oznacza, że połowa analizowanych cech nie jest wystarczająca, by prawidłowo dokonać klasyfikacji.

Jednak dla wykresów obu modeli wydać, że krzywe straty, zarówno dla danych treningowych, jak i dla danych walidacyjnych, dążą do kształtów swoich odpowiedników przy modelach uczonych na podstawie wszystkich dostępnych cech. Tutaj także model o podstawowej złożoności dopasowuje dosyć dobrze, a model o zwiększonej złożoności doświadcza zjawiska overfittingu.

c. Regularyzacja L1 I L2

Model: regresja logistyczna
Cost function: gradient descent
Biblioteka: PyTorch
Learning rate: 0.5
Epochs: 1000

Regularyzacja	Train accuracy	Test accuracy	Wagi
brak	98.03%	95.56%	średnia = 0.367 odchylenie = 5.83
L1 (λ=0.1)	35.47%	35.93%	średnia = 0.0163 odchylenie = 0.0878
L2 (λ=0.1)	47.16%	44.68%	średnia = -0.001415 odchylenie = 0.10776

Model bez regularyzacji osiąga najwyższą dokładność zarówno na zbiorze treningowym, jak i testowym. Jednak bardzo wysokie odchylenie standardowe wag (5.83) sugeruje, że wagi mogą przyjmować duże wartości, co może prowadzić do overfittingu. Wysoka różnica między średnią a odchyleniem standardowym potwierdza brak kontroli nad rozkładem wag, co czyni model potencjalnie niestabilnym przy nowych danych.

Zastosowanie regularyzacji L1 skutecznie wymusiło wyzerowanie wielu wag, co przełożyło się na znacznie niższą średnią oraz odchylenie standardowe. Jednak tak silna penalizacja wag spowodowała niedouczenie modelu (underfitting) – dokładność zarówno na zbiorze treningowym, jak i testowym spadła do około 35%. Model nie zdołał uchwycić wystarczająco dużo informacji z danych, przez co jego zdolność generalizacji została drastycznie ograniczona.

L2 zmniejszyło wartości wag bez całkowitego ich zerowania, co pozwoliło modelowi zachować część zdolności do nauki. Wagi mają bardzo małą średnią i umiarkowane odchylenie, co świadczy o bardziej kontrolowanym rozkładzie parametrów w porównaniu do przypadku bez regularyzacji. Niemniej jednak, tutaj również doszło do niedouczenia – dokładność testowa i treningowa pozostały na niskim poziomie.

Im mniejsza wartość λ, tym model bardziej przypomina model bez regularyzacji – zarówno pod względem dokładności, jak i rozrzutu wag. Oznacza to, że małe wartości λ zapewniają subtelną kontrolę nad przeuczeniem bez drastycznej utraty jakości.

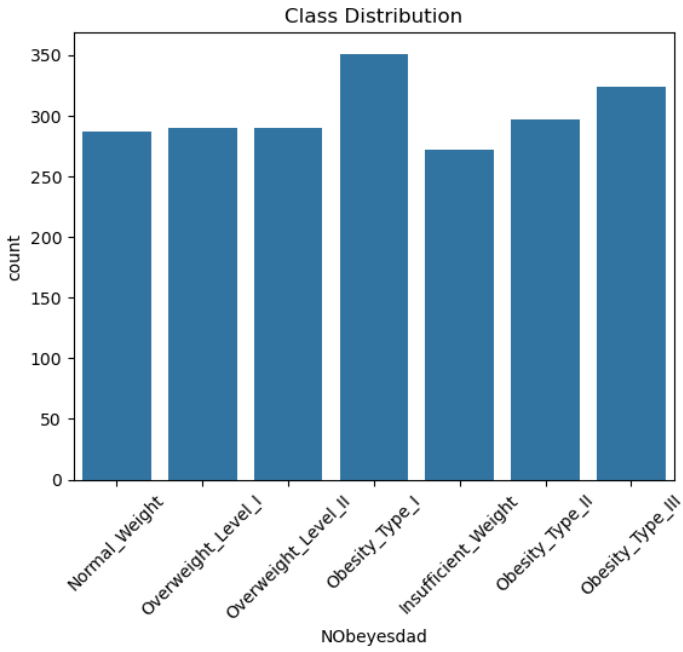
Regularyzacja najbardziej przydaje się w modelach, które cierpią na przeuczenie (overfitting). W analizowanym przypadku jednak model bez regularyzacji nie wykazywał istotnych oznak przeuczenia, co sugeruje, że dane są dobrze dopasowane, a sieć nie jest zbyt złożona. Wprowadzenie silnej regularyzacji w takiej sytuacji prowadzi jedynie do niedouczenia modelu

(underfitting). To właśnie obserwujemy przy zastosowaniu $\lambda = 0.1$ – znaczący spadek dokładności zarówno na zbiorze treningowym, jak i testowym.

alpha		Ridge Classifier	Logistic Regression penalty='l1'
1.0	**accuracy**	68.75%	81.63%
0.1		69.94%	81.06%
0.01		68.94%	80.30%
0.001		68.94%	79.92%

Dla metod Lasso i Ridge z biblioteki Scikit-learn, również otrzymujemy znacząco gorszą jakość predykcji, co znów doprowadza nas do wniosku, że te modele również nie cierpią na przeuczenie.

d. Usprawnienie danych – balansowanie zbiorów



Rozkład klas w zbiorze przed zbalansowaniem

Jak można zauważyć na powyższym wykresie, zbiór danych już jest dosyć dobrze zbalansowany, więc metody balansujące zbiór nie przyniosą tutaj wyraźnych efektów.

	Oryginalny zbiór			SMOTE			RandomUnderSampler		
accuracy	95.04%			94.33%			94.56%		
	precision	**recall**	**f1-score**	**precision**	**recall**	**f1-score**	**precision**	**recall**	**f1-score**
Insufficient_Weight	0.96	0.96	0.96	0.91	0.96	0.94	0.93	0.96	0.95
Normal_Weight	0.93	0.91	0.92	0.94	0.85	0.90	0.94	0.87	0.91
Obesity_Type_I	0.98	0.94	0.96	0.98	0.94	0.96	0.97	0.96	0.96
Obesity_Type_II	0.98	0.98	0.98	0.97	0.98	0.98	0.98	0.98	0.98
Obesity_Type_III	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Overweight_Level_I	0.91	0.89	0.91	0.88	0.93	0.90	0.88	0.93	0.90
Overweight_Level_II	0.87	0.95	0.87	0.89	0.91	0.90	0.89	0.89	0.89

Ogólna dokładność modelu po zbalansowaniu danych pozostała na takim samym poziomie. Obie metody balansowania danych nie zmieniły dużo w zbiorze danych, co nie może dziwić – dane były już niemalże idealnie zbalansowane.

Można jednak zauważyć delikatne zmiany w wynikach metryk. Przykładem jest klasa Overweight_Level_I, która jest jedną z mniej licznych zbiorze danych. Wskaźnik recall, który mówi, jaki procent rzeczywistych przypadków danej klasy model poprawnie rozpoznał, i który ważny jest, gdy nie chcemy przegapić przypadków wzrósł z 0.89 do 0.93.

e. Optymalizacja hiperparametrów

	Support Vector Classifier			
Top	Accuracy	**Hiperparameters**		
		C	gamma	kernel
1	96.53%	100	0.01	rbf
2	96.02%	1000	0.01	rbf
3	95.77%	1000	0.001	rbf
4	94.57%	10	0.1	rbf
5	94.06%	1000	0.1	rbf
BEZ OPTYMALIZACJI	96.34%	1	scale	rbf

Optymalizacja hiperparametrów modelu SVC metodą GridSearchCV pozwoliła na delikatną poprawę wydajności modelu (o 0.19 punktów procentowych).

Parametry:

- C
    - określa, jak bardzo model karze błędy klasyfikacji
    - Niska wartość – model dopuszcza więcej błędów – większa generalizacja

    - Wysoka wartość – model stara się dobrze dopasować do danych treningowych – ryzyko overfittingu
- Gamma
    - Określa zasięg wpływu pojedynczego punktu treningowego
    - Niska wartość – granice decyzyjne są gładkie – mniejsze ryzyko overfittingu
    - Wysoka wartość – granice decyzyjne ciasno otaczają dane – większe ryzyko overfittingu
- Kernel
    - Sposób transformacji danych do przestrzeni, w której klasy mogą być rozdzielone liniowo
    - Rbf – bardzo elastyczne, dobre dla danych nieliniowych
    - Poly – funkcja wielomianowa, mniej elastyczna, ale może lepiej dopasować do określonych typów danych

Z powyższej tabeli wynika, że dla tego zbioru danych najlepszy jest wysoki parametr C (ciężkie karanie za błędy), niski parametr gamma (większa generalizacja) i elastyczne jądro rbf.

	Decision Tree Classifier			
		Hiperparameters		
Top	Accuracy	criterion	max depth	min samples split
1	94.19%	log loss	20	5
2	94.13%	log loss	40	5
3	94.00%	entropy	None	2
4	93.94%	entropy	20	2
5	93.93%	log loss	10	10
BEZ OPTYMALIZACJI	92.62%	gini	None	2

Optymalizacja hiperparametrów modelu Decision Tree Classifier metodą GridSearchCV pozwoliła na poprawę wydajności modelu (o 1.57 punktów procentowych).
Parametry:

- Criterion
    - Określa funkcję, która używana jest do oceny jakości podziału
    - Gini – domyślna miara czystości – preferuje proste drzewa
    - Entropy – używa informacji entropicznej – nieco dokładniejsza, ale wolniejsza
    - log_loss – traktuje drzewo jako model probabilistyczny – może prowadzić do lepszego dopasowania, ale kosztowna obliczeniowo
- Max_depth
    - Określa maksymalna głębokość drzewa

    - Niska wartość – ogranicza złożoność drzewa – mniejsze ryzyko overfittingu
    - None – drzewo będzie rosło aż do pełnego podziału – może prowadzić do overfittingu
- Min_samples_split
    - Określa minimalną liczbę próbek potrzebną do podziału drzewa
    - Niska wartość – model może dzielić węzły bardzo agresywnie – ryzyko overfittingu
    - Wysoka wartość – model dzieli rzadziej – bardziej ogólne drzewo

Z powyższej tabeli wynika, że dla tego zbioru danych najlepsze jest kryterium log_loss i entropy (dokładniejsza ocena jakości), mała głębokość drzewa (zapobiegająca nadmiernemu dopasowaniu) i większe wartości min_samples_split (co zmniejsza szanse na overfitting).

Przeszukiwanie hiperparametrów jest ogólnie trudnym problemem, ze względu na kilka wad:

- **Wysoki koszt obliczeniowy** – każda kombinacja to osobny model do trenowania i walidacji, co może być bardzo czasochłonne przy większych zbiorach danych.
- **Nieintuicyjne zależności** – hiperparametry często oddziałują na siebie w nieliniowy sposób
- **Różna wrażliwość modeli** – niektóre modele są bardziej czułe na hiperparametry (np. drzewa decyzyjne), inne mniej (np. SVC) z domyślnymi ustawieniami często działają bardzo dobrze).
- **Overfitting na walidacji** – jeśli zestaw hiperparametrów pasuje zbyt dobrze do danych walidacyjnych, może nie generalizować na danych testowych.

h. Ensemble methods

	Wykorzystywane modele			**Model**
	Logistic Regression	**SVC**	**Decision Tree Classifier**	**Voting Classifier**
accuracy	90.53%	92.24%	92.61%	93.94%

Najlepszy wynik osiągnął klasyfikator zespołowy — Voting Classifier z głosowaniem miękkim (soft voting), który połączył predykcje kilku modeli, osiągając dokładność na poziomie 93.94%. Pokazuje to, że łączenie różnych modeli (nawet jeśli pojedynczo różnią się skutecznością) może prowadzić do uzyskania lepszej ogólnej wydajności. Głosowanie miękkie, które uwzględnia

prawdopodobieństwa klas, umożliwia bardziej zrównoważone podejmowanie decyzji i redukcję błędów pojedynczych klasyfikatorów.

	Base estimators			Final Estimator	Model
	Logistic Regression	**SVC**	**Decision Tree Classifier**	**Logistic Regression**	**Stacking Classifier**
accuracy	90.53%	92.24%	92.61%	90.53%	94.70%

Stacking Classifier osiągnął najwyższą skuteczność spośród wszystkich porównywanych tutaj modeli – 94.70%, co wskazuje na dużą efektywność tej metody w analizowanym zadaniu klasyfikacyjnym. Model ten łączy predykcje kilku bazowych klasyfikatorów, a następnie przekazuje je do modelu końcowego, który uczy się, jak najlepiej zestawić wyniki tych ekspertów w celu uzyskania ostatecznej prognozy.
Wysoka skuteczność Stackingu wynika z tego, że każdy bazowy klasyfikator może uczyć się innych wzorców w danych — dzięki temu model końcowy może nauczyć się, kiedy „ufać" danemu klasyfikatorowi najbardziej.

	Experts			Model
	Logistic Regression	**SVC**	**Decision Tree Classifier**	**Mixture of Experts**
accuracy	90.53%	92.24%	92.61%	95.70%

Mixture of Experts (MoE) osiągnął najwyższą skuteczność spośród wszystkich porównywanych tutaj modeli — 95.70%, co wskazuje na bardzo wysoką efektywność tej metody w zadaniu klasyfikacji. W porównaniu z wynikami poszczególnych ekspertów, MoE nie tylko dorównał najlepszemu z nich, ale jeszcze poprawił końcowy rezultat.
Mixture of Experts działa na zasadzie łączenia predykcji kilku modeli (ekspertów), przy czym każdemu z nich przypisywana jest dynamiczna waga obliczana przez tzw. gating model — osobny klasyfikator uczący się, który ekspert jest najbardziej wiarygodny dla danego przypadku. Dzięki temu, zamiast uśredniać decyzje, MoE uczy się kontekstowo przypisywać zaufanie do ekspertów w zależności od wejściowych cech.
Takie podejście sprawia, że MoE potrafi korzystać z mocnych stron poszczególnych klasyfikatorów w odpowiednich fragmentach przestrzeni cech, co znacząco zwiększa jego elastyczność i skuteczność.

Mixture of Experts to rozwiązanie szczególnie przydatne w przypadku dużych modeli z kilku kluczowych powodów:

- Efektywność obliczeniowa – zamiast uruchamiać cały, ogromny model, uruchamiamy tylko kilku mniejszych ekspertów
- Specjalizacja ekspertów – każdy ekspert może specjalizować się w konkretnym typie danych, co zwiększa ogólną precyzję modelu
- Możliwość rozbudowy – przy wprowadzaniu zmian nie trzeba przebudowywać całego modelu, wystarczy dodać kolejnego eksperta
- Lepsze wyniki – zastosowanie MoE pozwala uzyskać wyższą jakość predykcji

2. Studium ablacyjne

Aby znaleźć najbardziej optymalny model, wezmę pod uwagę tylko te metody, które faktycznie przyniosły poprawę wydajności wybranego modelu na wybranym zbiorze danych.

Opiszę krótko jak każda Regularyzacja wpłynęła na wydajność modelu:

1. Walidacja krzyżowa – delikatny spadek dokładności;
2. Zwiększenie złożoności modelu – overfitting i mocny spadek dokładności na zbiorze testowym;
3. Regularyzacja L1 i L1 – ze względu na niewystępowanie problemu overfittingu, Regularyzacja mocno negatywnie wpływa na dokładność modelu;
4. Balansowanie zbiorów – brak znaczącego wpływu – zbiór już jest dobrze zbalansowany;
5. Optymalizacja hiperparametrów – delikatna poprawa dokładności modeli;
6. Metody ensemble – tutaj można zauważyć największą poprawę wydajności.

Z powyższych informacji wynika, że przy tworzeniu studium ablacyjnego, rozważać będziemy tylko dwie ostatnie metody optymalizacji –optymalizację hiperparametrów oraz metody ensemble.

Ponieważ metody ensemble wymagają użycia kilku klasyfikatorów, wezmę pod uwagę modele: LogisticRegression, SVC oraz DecicionTreeClassifier z biblioteki Scikit-learn.

Metody optymalizacji	Accuracy score
Brak – LogisticRegression	90.53%
Brak – SVC	92.24%
Brak – DecisionTreeClassifier	93.18%
Optymalizacja Hiperparametrów	96.63% / 94.38%
Voting Classifier	93.94%
Stacking Classifier	94.70%
Mixture of Experts	95.70%
Optymalizacja Hiperparametrów + Voting Classifier	97.54%
Optymalizacja Hiperparametrów + Stacking Classifier	97.54%
Optymalizacja Hiperparametrów + Mixture of Experts	97.73%

Na podstawie poniższych wyników można wyciągnąć następujące wnioski:

- Optymalizacja hiperparametrów wyraźnie poprawiła skuteczność każdego z modeli. Różnice pomiędzy modelami bez optymalizacji a ich zoptymalizowanymi wersjami sięgają kilku punktów procentowych.
- Najlepsze rezultaty osiągnięto przy połączeniu optymalizacji hiperparametrów z technikami ensemble.

Najwyższą skuteczność osiągnął model **Mixture of Experts z optymalizacją hiperparametrów,** osiągając dokładność **97.73%**.

Ten model okazał się najlepszy z następujących powodów:

- Mixture od Experts łączy zalety wielu różnych klasyfikatorów (ekspertów) poprzez mechanizm wagujący, który dynamicznie decyduje, który ekspert lepiej poradzi sobie z danym przykładem.
- Optymalizacja hiperparametrów pozwoliła dobrać najlepsze możliwe konfiguracje zarówno dla poszczególnych ekspertów, jak i dla gating modelu, co zwiększyło elastyczność i dopasowanie modelu do danych.
- Dzięki takiemu podejściu MoE może:
    - Delegować trudniejsze przypadki do bardziej złożonych modeli,
    - Pozostawić proste obserwacje prostszym klasyfikatorom,
    - Dzięki temu lepiej radzić sobie z różnorodnością danych i klas,