**Zadanie**

**Zebrano pomiary 250 mężczyzn w różnym wieku (źródło: http://www.byu.edu/chhp).**

**Zebrane dane zawierają poprawnie zmierzony % tłuszczu „Pct.BF” (%bodyfat) oraz inne**

**zweryfikowane wartości zmiennych chrakteryzujących sylwetkę. Tab. 1 zawiera tylko**

**fragment danych i stanowi tylko ich ilustrację. Pełny zestaw danych znajduje się w pliku o**

**nazwie „MED-Lab-1-Zad-Dobor zmiennych-bodyfat-dane-i-opis.txt”**

**Funkcje zawarte w programie zad2.py:**

* r\_limit\_func – funkcja służy do obliczenia współczynnika granicznego korelacji według wzoru . Na wejściu funkcji należy podać liczbę parametrów.
* check\_y\_xn\_correlation(correlation\_matrix, r\_limit, data)

Funkcja bada korelację pomiędzy y oraz . Parametry posiadające korelację z y mniejszą niż współczynnik graniczny korelacji zostają odrzucone. Na wejście funkcji należy podać macierz korelacji, współczynnik graniczny korelacji oraz macierz z danymi wejściowymi. Po zakończeniu działania funkcja zwraca zredukowaną macierz danych.

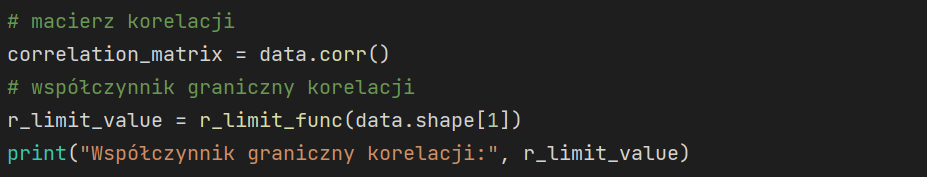
* check\_xn\_xn\_correlation(data)

Funkcja służy do odrzucenia parametrów silnie skorelowany ze sobą. Wyszukuje największą korelację pomiędzy , a następnie odrzuca ten parametr, który jest słabiej skorelowany z y. Kroki te są powtarzane, aż do momentu uzyskania wszystkich korelacji mniejszych od współczynnika granicznego korelacji. Z każdą iteracją obliczana jest ponownie macierz korelacji oraz współczynnik graniczny korelacji. Funkcja pobiera i zwraca macierz danych.

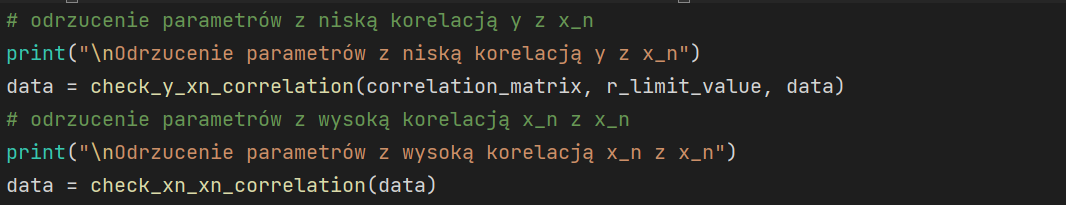
* Pozostałe funkcje użyte w programie zostały opisane w pierwszym zadaniu.

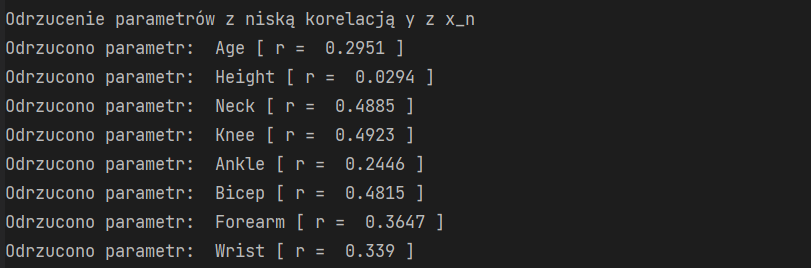
**MODEL LINIOWY DLA ZMIENNEJ Pct\_BF**

Eliminacja zmiennych została przeprowadzona na podstawie macierzy korelacji. Macierz korelacji zawiera kolumny y (Pct\_BF) oraz , , …, . W pierwszej kolejności został obliczony współczynnik graniczny korelacji.



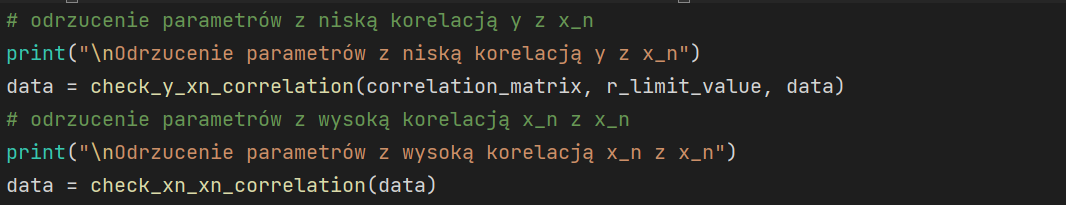


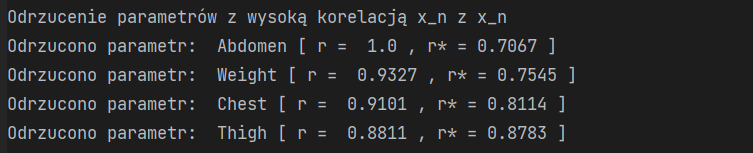
Wszystkie niebędące w silnej korelacji z y zostały odrzucone. 



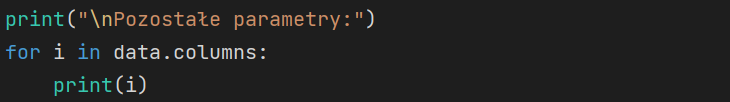
W powyższym kroku została odrzucona aż połowa parametrów.

Następnie została sprawdzona korelacja między x-ami. Jeżeli miedzy x-ami występuje silna korelacja, należy odrzucić słabiej skorelowanego z y.



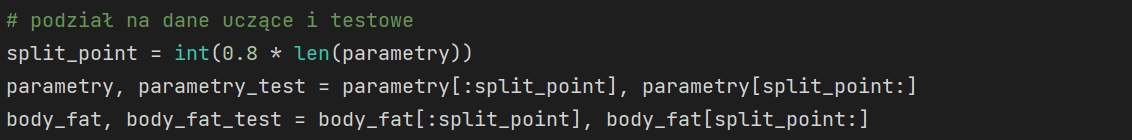


Pozostałe parametry:

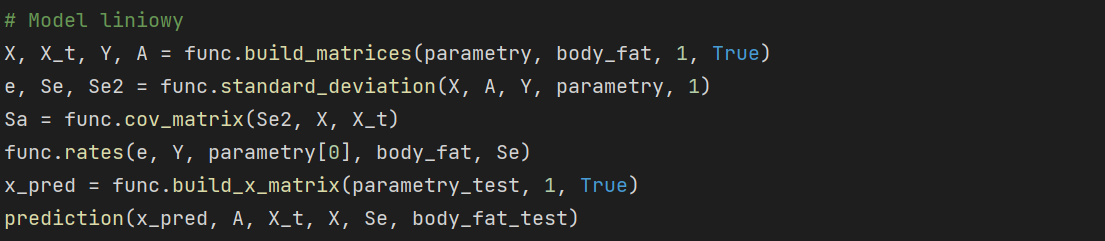




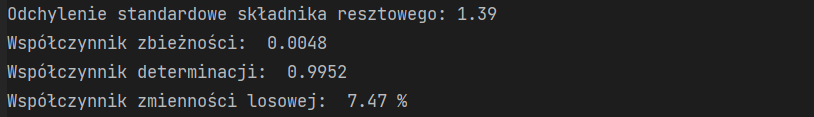
Przygotowane dane zostały podzielone na zbiór uczący oraz zbiór testowy w proporcji 8:2.



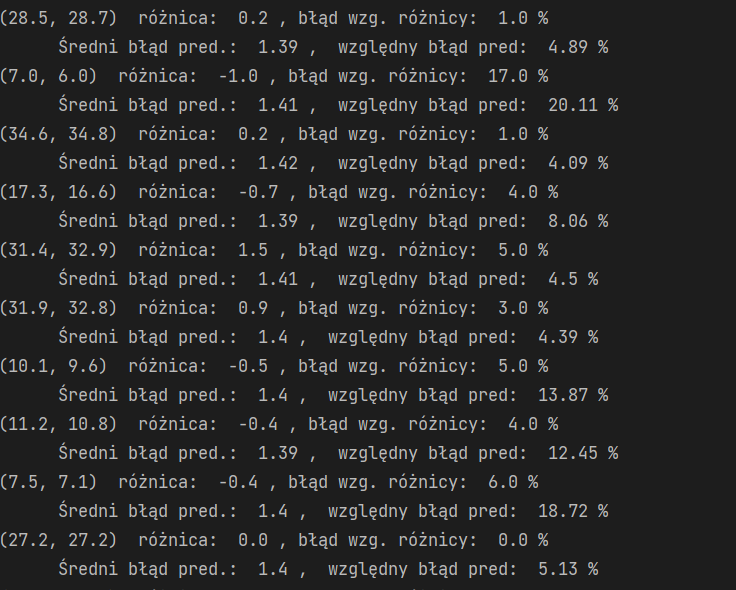
Dla danych uczących został utworzony model liniowy.

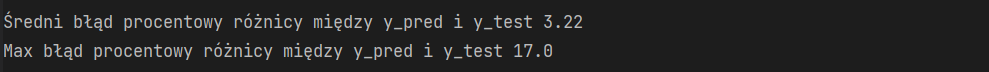


Poniżej zaprezentowano wskaźniki opisujące model.



Dla danych testowych sprawdzono poprawność działania modelu. Poniżej podano fragment zestawienia zmiennej objaśnianej wraz z y z danych testowych, ich różnicę oraz błąd procentowy tej różnicy. Dodatkowo podano średni błąd predykcji ex ante oraz względny błąd predykcji.





Model w stopniu bardzo dobrym wykonuje predykcję. Średni błąd procentowy różnicy między i wynosi 3.22%. Najwyższy błąd procentowy wynosi 17%.

Współczynnik determinacji dla modelu wynosi 0.995.

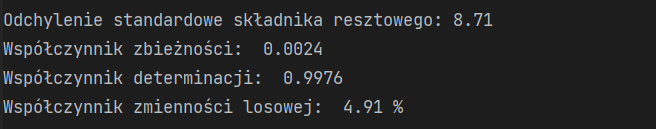
**MODEL LINIOWY DLA ZMIENNEJ WEIGHT**

Dodatkowo utworzono model liniowy dla zmiennej objaśnianej „Weight”. Przeprowadzono analogiczne kroki jak dla poprzedniego modelu.

Zmienne objaśniające które pozostały po redukcji to Pct\_BF, Waist oraz Hip.

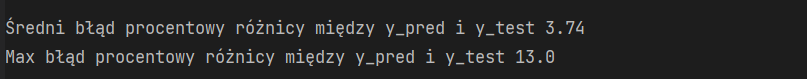


Współczynnik determinacji ponownie prezentuje się na bardzo wysokim poziomie wynoszącym 0.9976.



Fragment zestawienia zmiennej objaśnianej wraz z y z danych testowych.





Model dla zmiennej objaśnianej „Weight” również wykonuje predykcję na bardzo dobrym poziomie. Średni błąd procentowy różnicy między y\_pred i y\_test wynosi 3.74%. Najwyższy błąd procentowy wynosi 13%.