Sprawozdanie – Zastosowanie metody składowych głównych (PCA)

Jakub Zięba

# Cel pracy

Celem pracy jest przedstawienie metody składowych głównych i opisanie jej założeń, następnie zastosowanie metody na odpowiednio dużym zbiorze danych z uprzednią weryfikacją ww. założeń, interpretacja uzyskanych wyników oraz porównanie ich z wynikami uzyskanymi przy zaniedbaniu części z nich.

# Opis zbioru danych i założeń

## Opis datasetu

Wybrany zbiór danych składa się z informacji o słuchotkach, ślimakach morskich, wyglądem przypominających małże czy ostrygi. Uwzględnione cechy zwierząt to:

* Sex – płeć zwierzęcia, męska lub żeńska;
* Length – pomiar skorupy w najdłuższym jej miejscu, długość wyrażona w mm;
* Diameter – pomiar skorupy w miejscu prostopadłym do miejsca wymienionego wyżej, tj. ze zmiennej Length. Długość wyrażona w mm.
* Height – wysokość zwierzęcia, mierzona od miejsca styku z podłożem do najwyższego miejsca skorupy, wyrażona w mm;
* Whole Weight – całkowita waga ślimaka, łącznie skorupy oraz jej wnętrzności, wyrażona w gramach;
* Shucked Weight – waga „mięsa” ślimaka, wyrażona w mm;
* Viscera Weight – waga wnętrzności po ich wykrwawieniu, wyrażona w mm;
* Shell Weight – waga muszli po wysuszeniu, wyrażona w mm;
* Rings – liczba pierścieni na muszli, liczba całkowita. Pierścieni można użyć, podobnie jak słojów w przypadku drzew, do określania wieku zwierząt. Można to osiągnąć dodając 1.5 do liczby pierścieni.

## Opis założeń

Przy stosowaniu metody PCA należy spełnić szereg założeń, które pozwolą na poprawne jej wykonanie i w efekcie otrzymanie poprawnych wyników:

* Należy pozbyć się outlierów, ponieważ będą zakłócać przeprowadzanie badania, które może w efekcie jedynie podkreślić to że występują outlierzy;
* Dane należy zestandaryzować. Jest to wymagane szczególnie w przypadku danych mocno się od siebie różniących jeśli chodzi o rząd wielkości. W naszym przykładzie można spojrzeć na zmienne Diameter oraz Rings. Średnia wartość tej pierwszej to 0.4 mm, największa to 0.65 mm, natomiast dla zmiennej Rings już najniższą wartością jest 1, średnia to prawie 10 a maksimum 29. Takich zmiennych nie można badać przed ich standaryzacją, gołym okiem widać jak bardzo od siebie odstają.
* Poszczególne zmienne powinny mieć również rozkłady przypominające normalne.

# Analiza i obróbka danych

Przed przystąpieniem do obróbki i stosowania PCA, najpierw zajmiemy się podstawową analizą zbioru danych.

Obraz zawierający tekst, zrzut ekranu, czarne i białe, czarne

Opis wygenerowany automatycznie

Analizując statystyki związane z wymienionymi wyżej cechami, możemy zobaczyć że dane zawierają informacje o bardzo różnorodnych osobnikach. Najmłodszy z nich miał niespełna 2.5 roku, najstarszy natomiast ponad 30 lat. Zbadano tutaj zarówno osobniki maleńkie, o wadze tak małej jak 0.002 grama, jak i nieporównywalnie większe względem tych poprzednich, ważące niemal 3kg.

Obraz zawierający tekst, diagram, Wykres, linia

Opis wygenerowany automatycznie

Rozkłady poszczególnych zmiennych wskazują na występowanie wartości odstających, których należy się pozbyć.

Obraz zawierający tekst, diagram, Prostokąt, Równolegle

Opis wygenerowany automatycznie

Wykresy pudełkowe potwierdzają to przypuszczenie, wartości odstających jest całkiem sporo. Aby się ich pozbyć można m.in. użyć metody wykorzystującej rozstęp ćwiartkowy.

Obraz zawierający diagram, origami

Opis wygenerowany automatycznie

Obraz zawierający diagram, origami

Opis wygenerowany automatycznie

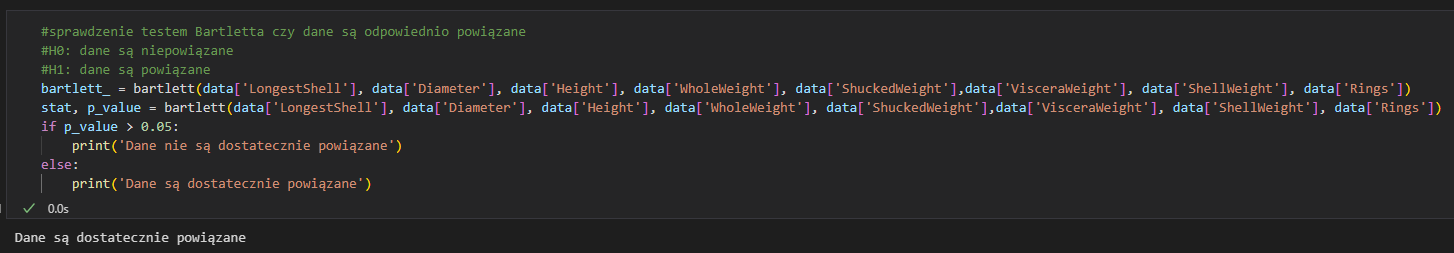
Po wyrzuceniu outlierów można porównać zbiór danych na wykresach skrzypcowych – widać znaczną poprawę jeśli chodzi o wartości odstające i długość „ogonów” wykresów. Wykresy te pokazują również, że większość rozkładów zmiennych przypomina w pewnym stopniu rozkłady normalne, choć na pewno żaden z nich nie jest dokładnie taki.

Kolejnym elementem który należy poddać analizie są współczynniki korelacji między zmiennymi.

Obraz zawierający tekst, zrzut ekranu, kwadrat, Prostokąt

Opis wygenerowany automatycznie

Współczynniki korelacji wydają się być wystarczająco wysokie, jednak sprawdzimy to jeszcze z pomocą formalnego testu – testu Bartletta.



Według testu Bartletta dane są dostatecznie powiązane. Pozostało więc jedynie poddać dane standaryzacji i możemy przejść do stosowania metody PCA.

# Opis i zastosowanie metody PCA

# Interpretacja wyników

# Porównanie z metodą stosowaną na danych niestandaryzowanych

# Podsumowanie