Podstawy uczenia maszynowego

14.03.2024

Laboratorium 2

Analiza głównych składowych

Łukasz Stępień, Kacper Fus

1. Cel zadania

Celem zadania jest zapoznanie się z metodą analizy głównych składowych (ang. Principal Component Analysis, PCA). W trakcie nauki wykorzystano dane „Plantdoc dataset” . Jest to zbiór zdjęć przedstawiający choroby popularnych roślin uprawnych.

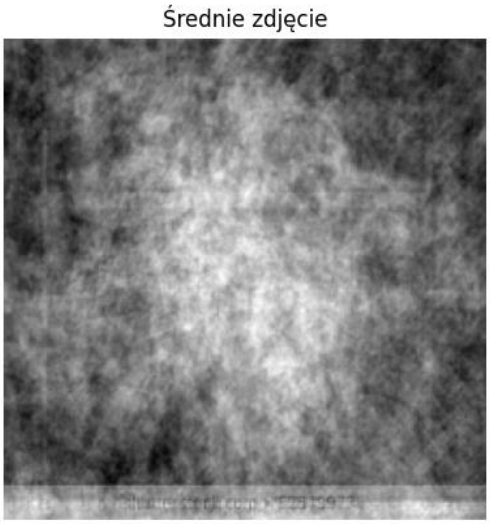
2. Implementacja

2.1 Preprocesing danych

- Ze zbioru danych wybrano podzbiór 60 zdjęć przedstawiających trzy choroby roślin jednego gatunku ("mold", "mosaic", "septoria"), po 20 zdjęć dla każdej choroby.
- Przeskalowano wszystkie zdjęcia do rozdzielczości 224×224, tak aby wszystkie obrazy miały ten sam rozmiar, równy 224 × 224 × 3.
- Skonwertowano obrazy do skali szarości, tak aby z trójwymiarowego tensora reprezentującego dane zdjęcie otrzymać tablicę dwuwymiarową.
- Skonwertowano obrazy, będące teraz tablicami dwuwymiarowymi (macierzami) na wektory, dzięki czemu każdy obraz jest reprezentowany przez wektor o rozmiarze 50 176.
- Przeprowadzono centrowanie zbioru,

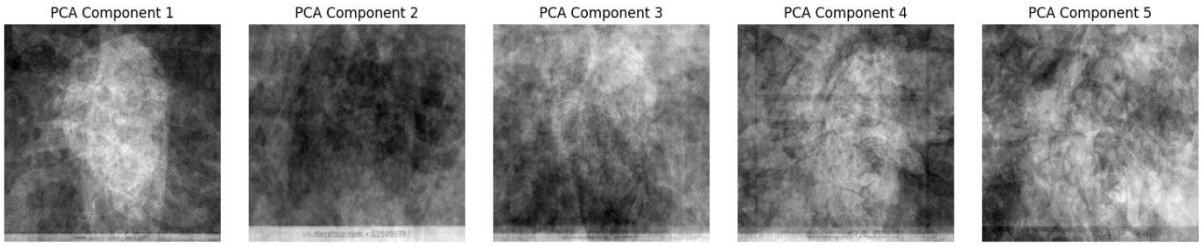
2.2 Analiza głównych składowych

- Wykonano transformację PCA, poprzez użycie funkcji sklearn.decomposition.PCA.
- Porównanie macierzy kowariancji przed i po transformacji PCA. Przed transformacją macierz ta posiadała rozmiar 50176x50176, zaś po tylko 60x60.
- Średnie zdjęcie, które wykorzystano w trakcie centrowania zbioru wyglądało następująco.



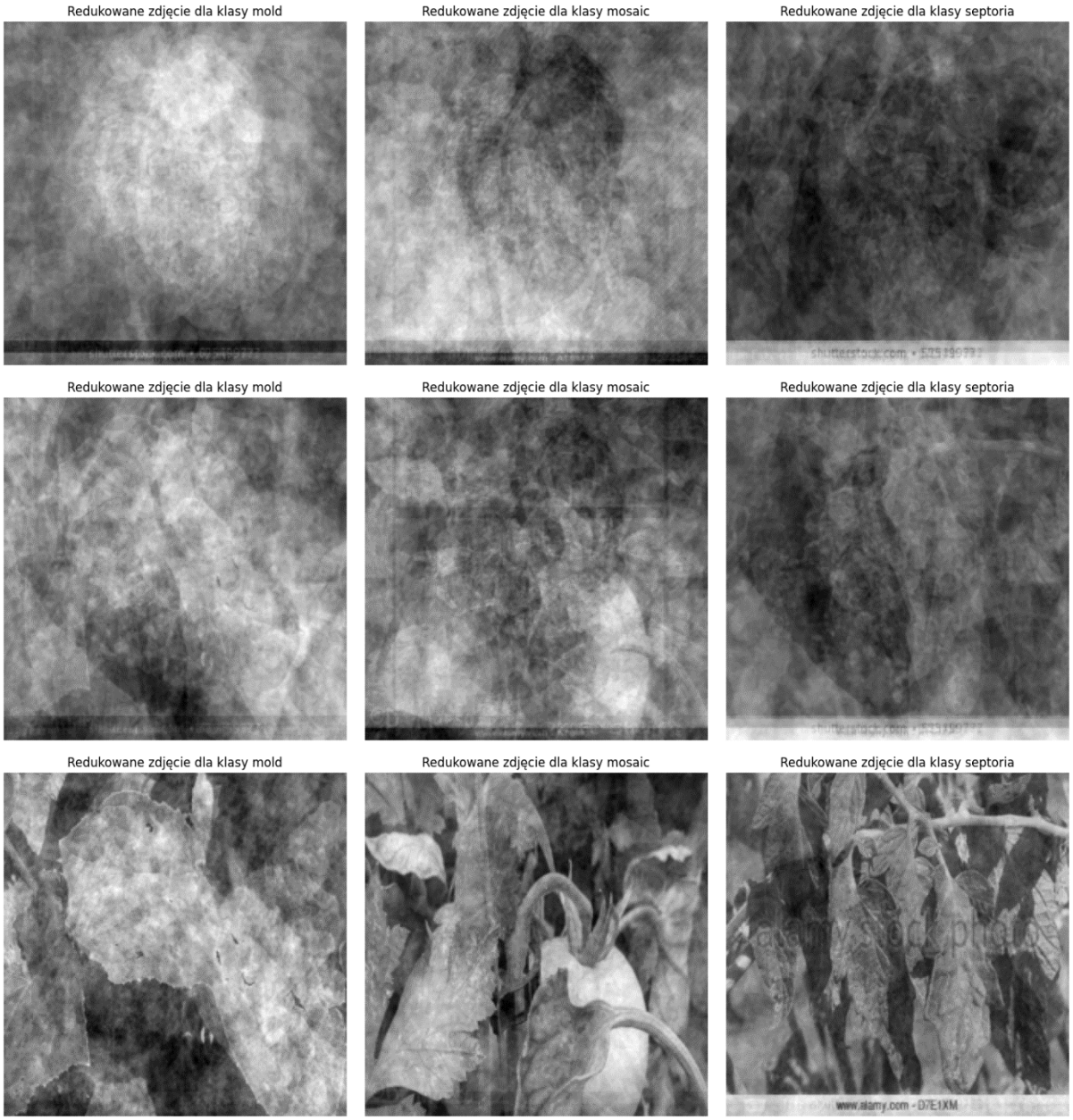
Ryc 3.

- Zwizualizowano nowe wektory bazowe, które posortowano według powiązanej wariacji

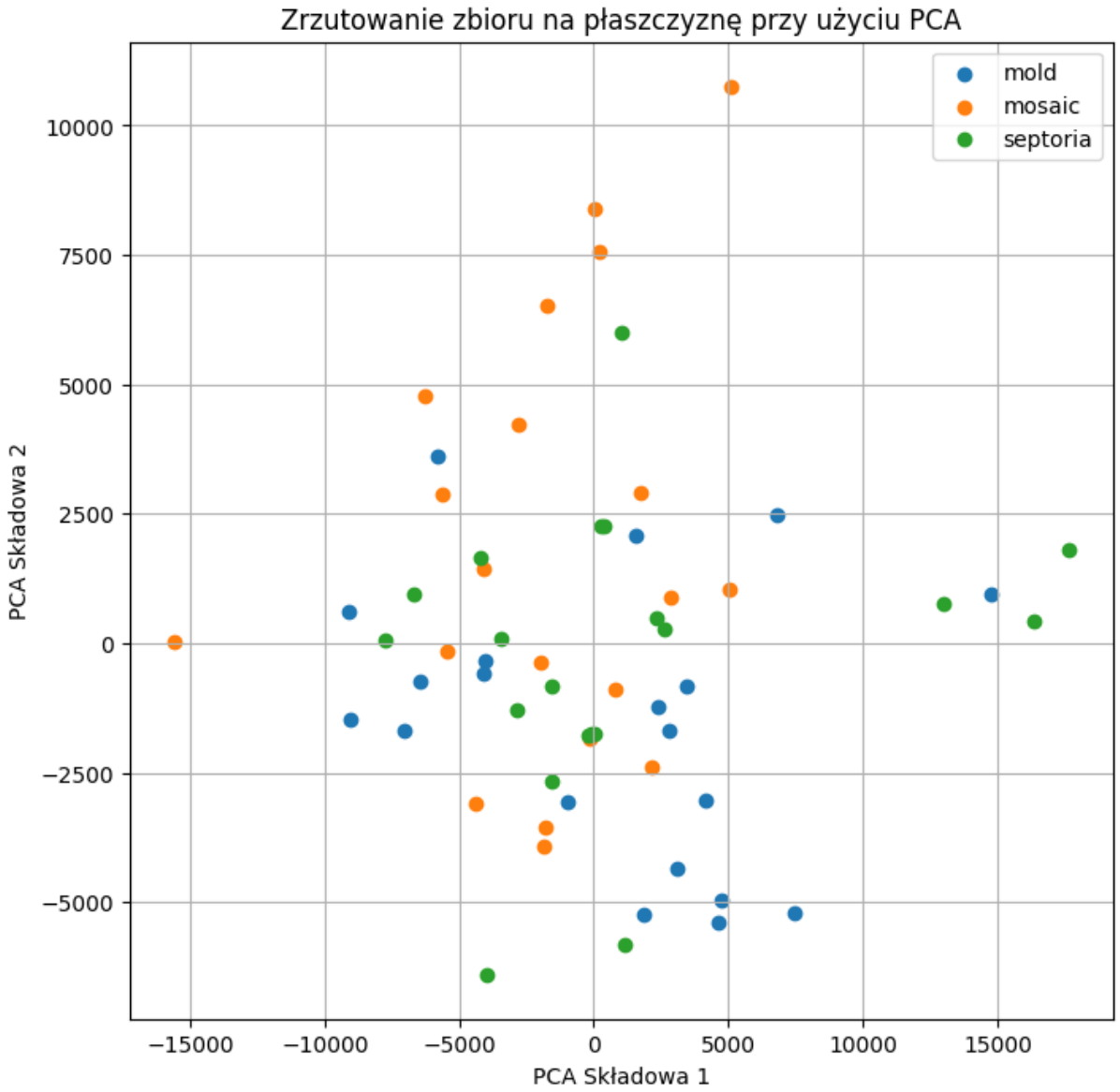


Ryc. 4

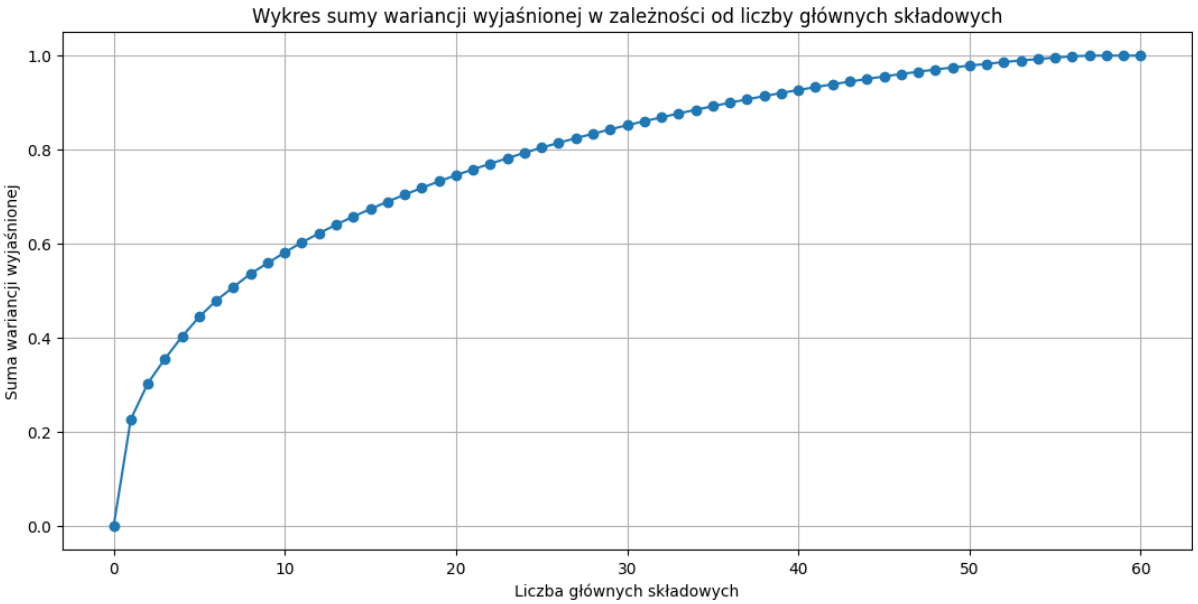
- Zredukowano wymiarowość obserwacji do odpowiednio 3, 9 i 27 najważniejszych składowych, czyli cech w nowej bazie. Wyniki przedstawiono poniżej.



- Zredukowano wymiarowość do 2 najważniejszych aspektów danych. Powstałe wektory 2D użyto jako wektory na płaszczyźnie, aby wykorzystać PCA do zrzutowania zbioru na płaszczyznę. Wyniki zaprezentowano na poniższych wykresach.



- Dzięki redukcji wymiarowości danych, został usunięty tzw. „szum” co ułatwia nam wizualizacje rozkładu danych. Pomaga to zauważyć jak dane są uzależnione od składowych PCA.



- Wykres sumy wariancji wyjaśnionej w zależności od liczby głównych składowych pozwala zrozumieć, jak wiele informacji przechowuje się w kolejnych składowych głównych w PCA. Pokazuje, ile wariancji w danych jest wyjaśniane przez każdą kolejną składową główną dodaną do analizy. Oś x na wykresie reprezentuje liczbę głównych składowych (PC1, PC2, PC3, ..., PCk), które są uwzględniane w analizie. Początkowo, gdy dodajemy tylko pierwszą składową główną (PC1), suma wariancji wyjaśnionej zaczyna od wartości odpowiadającej wariancji tej pierwszej składowej. W miarę dodawania kolejnych składowych głównych, suma wariancji wyjaśnionej rośnie. Początkowo wzrost ten jest dosyć szybki, zwłaszcza gdy pierwsze składowe główne przechowują dużą ilość informacji. W pewnym momencie, wzrost ten zaczyna maleć, ponieważ kolejne składowe główne przechowują coraz mniej informacji o wariancji w danych.