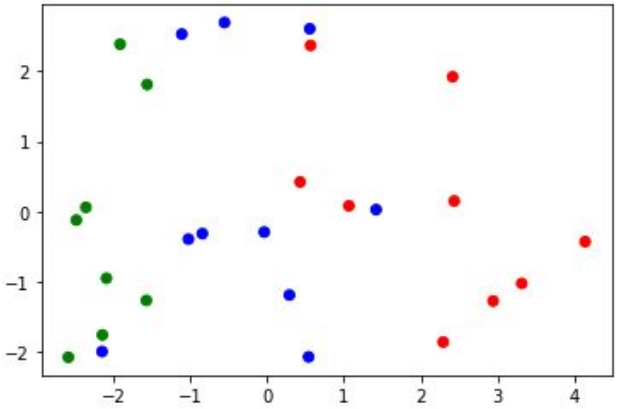
Wizualizacja metodami MDS

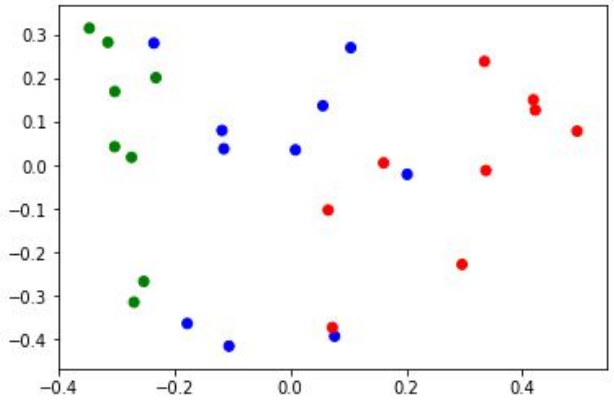
Józef Jasek

Zadanie 1

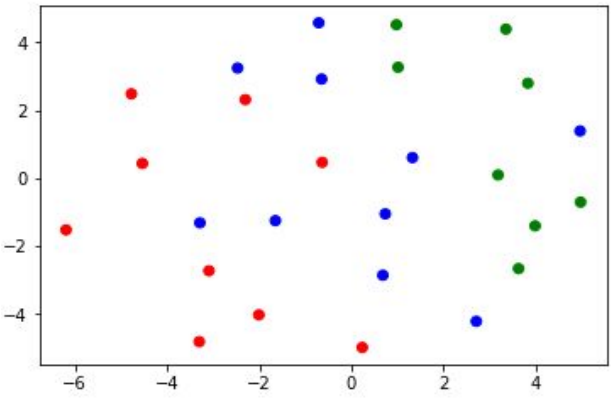
Wizualizowany zbiór danych to lung-cancer rozmiaru 27x56 po wyczyszczeniu danych.
Klasy od 1 do 3 reprezentują różne stany pacjentów.



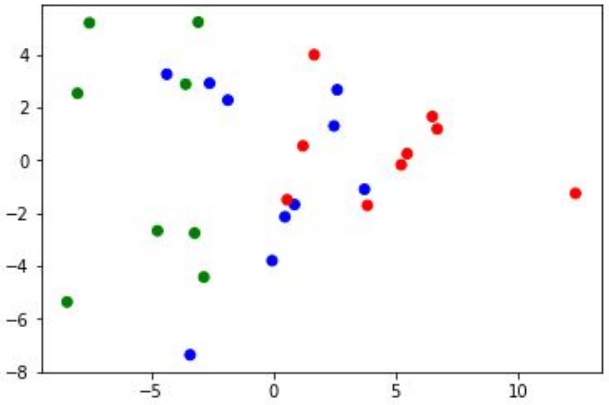
PCA



KPCA



MDS



ISOMAP

Porównajmy dwie wartości, by wyznaczyć najlepszy algorytm dla tego zbioru.
Pierwsza wartość: ile średnio sąsiadów po przeprowadzeniu danej metody jest tymi samymi sąsiadami co po wyliczeniu k-NN na pierwotnym zbiorze.
Druga wartość: ile średnio najbliższych sąsiadów należy do tej samej klasy (w obu przypadkach k=15).

Dla dwóch wymiarów:

pca:	11.925925925925926/15.	0.48148148148148157
kpca:	11.814814814814815/15.	0.49135802469135803
mds:	11.74074074074074/15.	0.44691358024691363
isomap:	11.851851851851851/15.	0.471604938271605
lle:	10.703703703703704/15.	0.44444444444444453

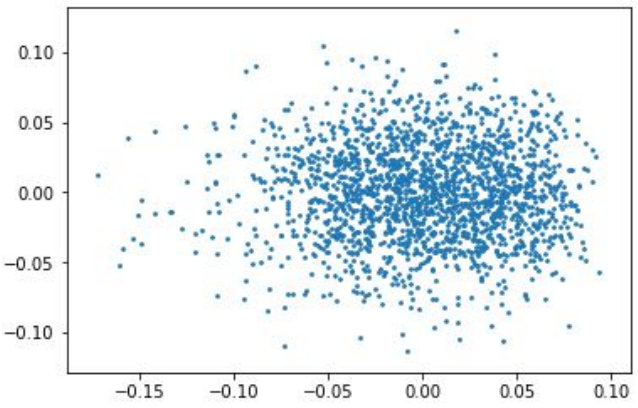
Dla trzech wymiarów:

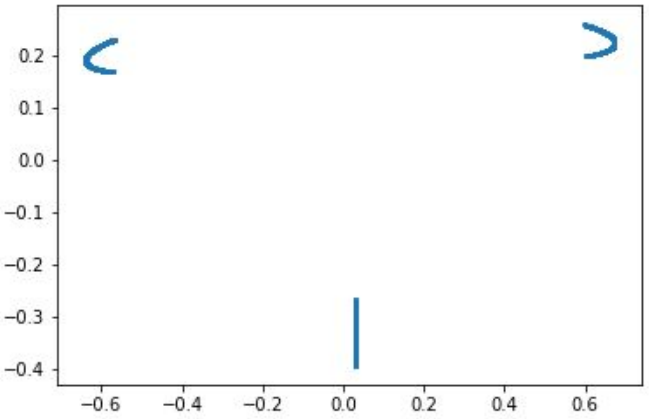
pca:	12.296296296296296/15	0.4740740740740741
kpca:	11.925925925925926/15	0.471604938271605
mds:	12.37037037037037/15	0.45679012345679015
isomap:	12.074074074074074/15	0.4740740740740741
lle:	9.666666666666666/15	0.4320987654320988

Widzimy, że w zależności od wybranej metryki i liczby wymiarów najlepiej wypada PCA, KPCA lub MDS. Nie ma jednak bardzo dużej różnicy w jakości wizualizacji.

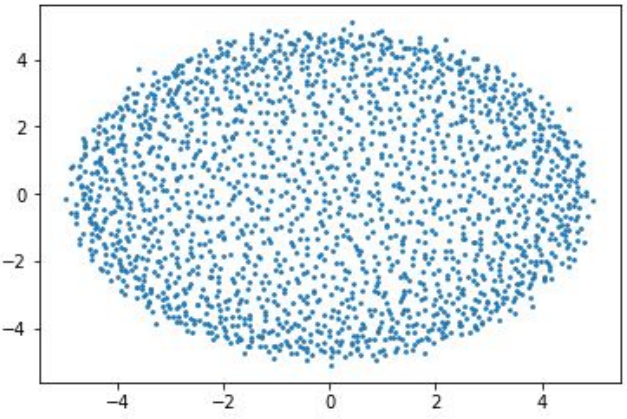
Zadanie 2

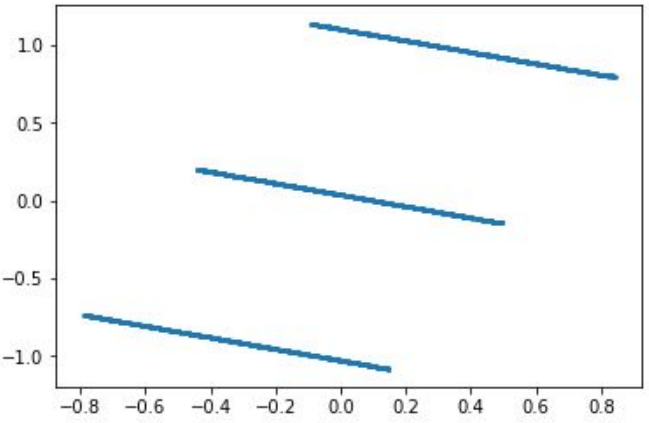
Mamy zbiór 1891 na 208. Niemal wszystkie kolumny są losowo wygenerowane. Przeprowadzenie PCA i KPCA na całym zbiorze pozwala nam bardzo dobrze wydzielić 3 zbiory danych. Pozostałe metody nie zwracają wartościowych informacji. Wykorzystujemy metody PCA i KCPA na połowie zbioru i analogicznie do przeszukiwania binarnego ograniczamy zbiór przeszukiwanych kolumn o połowę. Jeżeli jedna z połów generuje podobny wykres do całego zbioru, to tam znajdują się dane, które nas interesują. W ten sposób otrzymujemy, że tylko ostatnia kolumna tego zbioru zawiera dane wartościowe. Poniżej porównanie wizualizacji dla pierwszych 207 kolumn i obok dla ostatniej kolumny:





KPCA





MDS

Dla porównania wizualizacja pełnego zbioru metodą PCA i KPCA: