PPPD - Lab. 08

Załaduj jeszcze raz zbiór danych `iris` i dokonaj jego transpozycji. Zakładamy od tej pory, że macierz `iris` jest listą 4 list, każda o długości 150.

Rysunek 1: Gatunki kosaćców (irysów)

Okazuje się, że ww. zbiór danych reprezentuje pomiary dla 4 gatunków kwiatów:

- 50 pierwszych wierszy to *Iris setosa* (kosaciec szczecinkowy),
- 50 drugich – *Iris versicolor* (kosaciec różnobarwny),
- 50 trzecich – *Iris virginica* (kosaciec wirginijski).

---

1. Narysuj jeszcze raz wykresy rozproszenia dla każdej pary zmiennych (zob. rysunek poniżej). Tym kolor i-tego punktu powinien odzwierciedlać gatunek kosaćca. Wywołaj `plt.scatter(zmienna_x, zmienna_y, c=col, alpha=0.3)`, gdzie `col` jest listą napisów taką, że `col[k]` to angielska nazwa koloru, przy użyciu której będziemy reprezentować gatunek k-tego irysa (`alpha` to stopień przezroczystości dla poprawy estetyki).

   Uwaga: aby zwiększyć czytelność wykresu, możesz zaburzyć współrzędne każdego punktu małą wartością losową (np. z przedziału $[-0{,}05, 0{,}05]$*).*

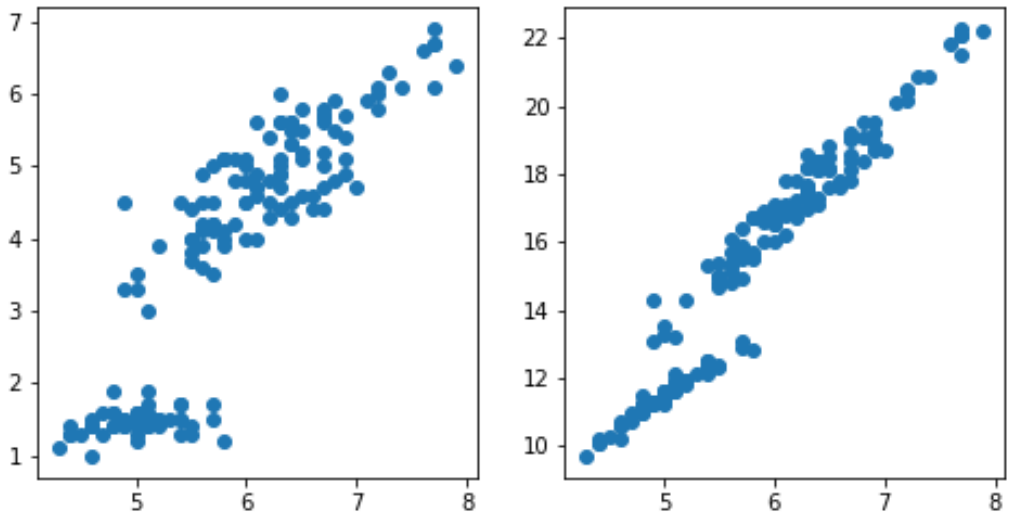
2. Wyznacz tzw. centroid każdej klasy (dla każdego gatunku kosaćca), tj. oblicz średnią arytmetyczną dla każdej z 4 zmiennych w obrębie każdej z 3 klas. Zapisz je w postaci macierzy `C` typu 4×3.

3. Dodaj do ww. wykresu punkty oznaczające centroidy: `plt.plot(wsp_x, wsp_y, markersize=25, color="#00000055", marker='o', linestyle='')`.

4. Dla każdego ze 150 punktów oblicz jego odległość euklidesową do każdego z centroidów (w przestrzeni 4-wymiarowej). Wskaż, który środek (pierwszy, drugi, czy trzeci) jest jemu bliższy. Oblicz proporcję liczby punktów takich, że najbliższy mu centroid jest faktycznie centroidem klasy, z której on pochodzi.

5. Powtórz ćwiczenie z p. 4 dla każdej zmiennej, każdej pary i każdej trójki zmiennych (odległość euklidesowa w przestrzeni 1-, 2- lub 3-wymiarowej). Wskaż przypadek, który prowadzi do największej proporcji poprawnie zaklasyfikowanych punktów.

6. Powtórz ćwiczenie z p. 3 i 4, ale tym razem na wersji `iris`, w której zmienne zostały wystandaryzowane, tj. od obserwacji w każdym wierszu macierzy `iris` (dla przypomnienia: macierz jest typu 4×150) została odjęta jego średnia arytmetyczna i następnie wynik został podzielony przez jego odchylenie standardowe (upewnij się, że po standaryzacji średnia arytmetyczna w każdym wierszu wynosi 0, a odchylenie standardowe – 1).



Rysunek 2: Wykresy rozproszenia dla każdej pary zmiennych (współrzędne niezaburzone szumem losowym), kolor oznacza gatunek kosaćca