DATA MINING 7

Zadanie 1 (grupowanie)

Dla danych z pliku **penguins.csv** wykonaj polecenia:

1. Wyświetl podstawowe informacje o zbiorze.
2. Sprawdź czy w zbiorze nie brakuje danych. Jeżeli są – usuń je ().
3. Ogranicz się do atrybutów **'bill_length_mm'** i **'flipper_length_mm'**.
4. Narysuj dendogram, zinterpretuj go i wyznacz ilość klastrów.
5. Zastosuj algorytm hierarchiczny aglomeracyjny do zbioru i wyznacz klastry.
6. Stwórz wykres rozrzutu z zaznaczonymi klastrami.

UWAGA: w punktach 3-5 przetestuj różne sposoby obliczenia odległości między klastrami (linkage - *complete, average, single, ward*).

Zadanie 2 (klasyfikacja)

A. Przetestuj algorytm kNN. Wykorzystaj metodę `KNeighborsClassifier` z `sklearn.neighbors`. Wykorzystaj notatnik **ED_kNN_przykład.ipynb.**

B. Napisz samodzielnie program wykorzystujący algorytm kNN do klasyfikacji punktu na płaszczyźnie (x,y) w przypadku zbioru danych wykorzystanego w punkcie A.

Zadanie 3 (klasyfikacja)

Rozważmy następujący zbiór danych:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

A. Rozważ dwie kolumny: **Outlook** i **Temperature**. Zaklasyfikuj przypadek:

(Rain, Hot)

Zobacz: **ED_kNN_zadanie_3_TODO.ipynb**

B. Rozważ wszystkie kolumny i zaklasyfikuj przypadek:

(Sunny, Cool, High, Strong).

Zadanie 4 (klasyfikacja)

Zastosuj algorytm kNN do zbioru: https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

1. Przygotuj zbiór danych – dodaj nazwy atrybutów.
2. Przekonwertuj dane kategoryczne do numerycznych.
3. Jako etykiety przyjmij wartości atrybutu 'class'.
4. Wydziel ze zbioru danych zbiór treningowy i zbiór testowy `(sklearn.model_selection.train_test_split)`
5. Znajdź precyzję nauczonego modelu.

Zadanie 5 (klasyfikacja)

Zapoznaj się z przykładem implementacji algorytmu kNN do zbioru **Titanic dataset** (https://www.kaggle.com/c/titanic).

Przykład jest dostępny online: https://www.kaggle.com/gaurav9297/titanic-using-knn

W przykładzie uwzględnione są 3 kolumny. Spróbuj uwzględnić 2 kolumny więcej i przedyskutuj uzyskane wyniki.