**Statystyczne metody rozpoznawania obrazu**

**Zadanie 4**

**Bartłomiej Gorzela**

1. **Zidentyfikować wartości odstające w wybranym zbiorze danych.**

W celu wykonania zadania posłużyłem się przykładowym zbiorem danych ‘California Housing Dataset’ możliwym do zaimportowania bezpośrednio z biblioteki scikit-learn dostępnej w Python. Do identyfikacji wartości odstających zdecydowałem się na wybór dwóch atrybutów:

- *MedInc* (Median Income): przedstawia medianę dochodu gospodarstwa domowego.  
- *HouseAge* (Median Age of Housing Units): średnia wieku wszystkich budynków mieszkalnych.

Wartości wybranych atrybutów zwizualizowałem za pomocą histogramów (argument ‘bins’ ustawiłem na wartość 100 w celu poprawy czytelności wizualizacji):

A comparison of a graph

AI-generated content may be incorrect.

Oryginalne dane zwizualizowałem za pomocą wykresu 2D:  
A chart of green dots

AI-generated content may be incorrect.

Kolejnym krokiem jaki wyknałem było zidentyfikowanie wartości odstających. Posłużyłem się w tym celu algorytmem DBSCAN (Density-Based Spatial Clustering of Applications with Noise), który umożliwia wykrywanie skupisk danych oraz jednoczesne oznaczanie punktów, które można uznać za wartości odstające. Algorytm jest dostępny w bibliotece scikit-learn.

Algorytm wymagał przeprowadzenia standaryzacji danych, czego dokonałem przy użyciu obiektu StandardScaler(). Po przeskalowaniu danych, zastosowałem algorytm DBSCAN z parametrami: **eps=0.3** oraz **min\_samples=10**. Oznacza to, że punkt jest uznany za rdzeniowy jeśli ma co najmniej 10 sąsiadów w promieniu 0.3 jednostki (w przestrzeni standaryzowanej).

W wyniku działania DBSCAN każdemu punktowi została przypisana etykieta klastra (dbscan\_label), przy czym wartość -1 oznacza punkt uznany za odstający.

Poddane działaniu algorytmu atrybuty ponownie zwizualizowałem za pomocą wykresu 2D, zaznaczając kolorem czerwonym punkty uznane za wartości odstające.

A graph of red and blue dots

AI-generated content may be incorrect.

Ten sam krok powtórzyłem dla modyfikując wartości ‘eps’ oraz ‘min\_samples’. Wyniki znajdują się poniżej:

A diagram of red and blue dots

AI-generated content may be incorrect.

A diagram of a red and blue graph

AI-generated content may be incorrect.

A red and blue dotted chart

AI-generated content may be incorrect.

Kod źródłowy:

A screen shot of a computer program

AI-generated content may be incorrect.  
A computer screen shot of a program code

AI-generated content may be incorrect.