## **Klasteryzacja danych EMNIST MNIST metodą k-średnich**

### **1. Opis zadania**

Celem było przeprowadzenie klasteryzacji danych EMNIST MNIST metodą k-średnich, z wykorzystaniem poprawionej inicjalizacji centroidów. Dla różnych wartości liczby klastrów (10, 15, 20, 30) wybrano najlepszy wynik na podstawie minimalnej inercji (suma kwadratów odległości punktów od ich centroidów).

### **2. Klasteryzacja (k = 10)**

Dla każdej ilości klastrów wykonano 10 prób inicjalizacji, wybierając wynik z najmniejszą inercją. Uzyskana macierz procentowego przydziału cyfr do klastrów pozwoliła ocenić jakość dopasowania.

Wizualizacja centroidów pokazała, że większość z nich chociażby przypomina odpowiednie cyfry, choć niektóre klastry zawierały dane niejednoznaczne (np. zniekształcone cyfry).

### **3. Klasteryzacja dla k = 15, 20, 30**

Dla większych wartości k analogicznie przeprowadzono klasteryzację i wizualizację wyników:

* Przy **k = 15** i **k = 20** zauważono, że niektóre klastry reprezentują warianty tej samej cyfry (np. pionowe i ukośne „1”).
* Przy **k = 30** klastry stawały się bardziej szczegółowe — można było rozróżnić różne style zapisu tej samej cyfry.

W niektórych przypadkach (szczególnie przy k=20 i 30) możliwe byłoby połączenie kilku klastrów w jedną klasę cyfry — co może być użyteczne w konstrukcji klasyfikatora z klasteryzacją jako etapem wstępnego grupowania.

### **4. Wnioski**

* Centroidy dla k=10 są zbliżone do średnich obrazów cyfr, co świadczy o poprawnym działaniu algorytmu.
* Zwiększanie liczby klastrów poprawia rozróżnialność wariantów cyfr, ale utrudnia bezpośrednią interpretację.
* Dla klasyfikatora cyfr najtrafniejsze wydaje się użycie 10 lub 15 klastrów, z ewentualnym scalaniem podobnych w wyższych wartościach k.

# **Klasteryzacja zbioru EMNIST MNIST za pomocą algorytmu DBSCAN**

## **1. Opis zadania**

Celem było zastosowanie algorytmu DBSCAN do klasteryzacji danych obrazowych ze zbioru EMNIST (cyfry), tak aby uzyskać możliwie najniższąliczbę punktów szumu.

## **2. Metodologia**

Dane wejściowe: zbiór EMNIST z cyframi 0–9 (zredukowany do 2D przy użyciu PCA).  
 Algorytm: DBSCAN z różnymi parametrami eps (promień sąsiedztwa) oraz min\_samples (min. liczba sąsiadów).  
 Dobór parametrów był przeprowadzony eksperymentalnie, w celu maksymalizacji dokładności klasyfikacji przy minimalnym szumie i sensownej liczbie klastrów( w zakresie od 10 do 30).

## **3. Najlepsze uzyskane wyniki**

Wyniki uzyskane dla:

Epsylon = 11.5

Min\_samples = 4

|  |  |
| --- | --- |
| **Metryka** | **Wartość** |
| Liczba wyznaczonych klastrów | **19** |
| Dokładność klasyfikacji (bez szumu) | **0,1176** |
| Odsetek błędów w klastrach | **0.8824** |
| Procent punktów uznanych za szum | **0,0360** |