1 Opis problemu:

Wybrany temat klasteryzacja, problem klasteryzacji polega na grupowaniu zbioru danych na podstawie ich podobieństwa. Głównym celem jest, aby obiekty w jednym klastrze były do siebie bardziej podobne niż do obiektów z innych klastrów.

2 Opis algorytmów

K-średnich: Proste i powszechne, grupuje dane w predefiniowaną liczbę (k) klastrów w oparciu o centra (środki).

Klastrowanie hierarchiczne: Tworzy hierarchię klastrów, łącząc najbliższe punkty danych (aglomeracyjne) lub dzieląc większe klastry (podziałowe).

DBSCAN: Znajduje klastry na podstawie gęstości punktów danych, dobrze radzi sobie z szumem.

Gaussian Mixture Models (GMM): Zakłada, że dane pochodzą z wielu rozkładów gaussowskich (krzywe dzwonowe), przydatne dla niesferycznych klastrów.

Klastrowanie aglomeracyjne (wspomniane dwukrotnie): Patrz Hierarchical Clustering (podejście łączące).

Spectral Clustering: Wykorzystuje właściwości macierzy podobieństwa do identyfikacji klastrów, działa dobrze na złożonych kształtach.

Mean Shift Clustering: Przesuwa punkty danych, aż osiągną obszary o dużej gęstości, przydatne w przypadku klastrów podobnych do kropli.

Propagacja pokrewieństwa: Wysyła wiadomości między punktami danych w celu znalezienia klastrów na podstawie wzajemnych podobieństw.

OPTICS: Porządkuje punkty danych na podstawie gęstości, przydatne do znajdowania klastrów o różnej gęstości.

BIRCH: Tworzy wielopoziomowe podsumowanie danych w celu wydajnej obsługi dużych zbiorów danych.

3 Opis algorytmu/metody, która została wybrana do rozwiązania problemu wraz z uzasadnieniem i wybranego algorytmu do nauki.

Opis Algorytmu

K-średnich to popularna metoda klasteryzacji, która jest stosowana do grupowania zestawu danych na k klastrów, gdzie k jest liczbą z góry określoną przez użytkownika.

Algorytm działa w następujących krokach: 1. Inicjalizacja: Wybierz kkk początkowych centroid, które mogą być losowo wybrane z danych wejściowych

2. Przypisanie punktów do klastrów: Każdy punkt danych jest przypisany do najbliższego centroidu na podstawie odległości euklidesowej.

3. Aktualizacja centroidów: Przelicz centroidy jako średnie z punktów przypisanych do każdego klastra.

4. Iteracja: Powtarzaj kroki 2 i 3 aż do zbieżności, czyli kiedy przypisania punktów do klastrów przestaną się zmieniać lub zmiany będą minimalne.

Uzasadnienie Wyboru Algorytmu

1. Intuicyjność i Prostota: Algorytm K-średnich jest prosty do zrozumienia i implementacji. Jego podstawowe operacje, takie jak obliczanie odległości i średnich, są łatwe do zaimplementowania i zoptymalizowania.

2. Efektywność: K-średnich jest stosunkowo szybki i efektywny obliczeniowo, szczególnie przy użyciu dużych zbiorów danych.

3. Szerokie Zastosowanie: Algorytm K-średnich jest używany w różnych dziedzinach, takich jak rozpoznawanie wzorców, przetwarzanie obrazów, segmentacja rynku, i wiele innych. Jest to uniwersalny algorytm, który można dostosować do wielu problemów.

4. Opis wybranych zbiorów danych

Wybrany zbiór zawiera imiona, ilośc i plec dzieci urodzonych pomiędzy 1880 a 2015 r.
I 1858692 rekordy
link do zbioru: https://www.kaggle.com/datasets/thedevastator/us-baby-names-by-year-of-birth

5. Proces przetwarzania danych dostosowany do wybranej metody.

1. Zbieranie Danych: pierwszym krokiem jest zgromadzenie danych, które będą podlegać klasteryzacji. Dane mogą pochodzić z różnych źródeł, w moim wypadku używany jest plik CSV

2. Czyszczenie Danych: Dane często zawierają błędy, braki, duplikaty i inne niedoskonałości. Przed przystąpieniem do klasteryzacji należy przeprowadzić:
usuwanie brakujących wartości, usuwanie duplikatów, obsługę wartości odstających

3. Standaryzacja Danych: jest kluczową, ponieważ algorytm K-średnich opiera się na odległości euklidesowej, która jest wrażliwa na skalę cech.
Normalizacja: Skaluje cechy do przedziału-100, 100]. co również ułatwia wizualizacje danych

6. Przygotowanie zbiorów uczących i testujących.

1. Podział Danych

Dane są podzielone na zbiór uczący i testowy, w stosunku 80:20.
Podział ten jest losowy, aby zapewnić reprezentatywność obu zbiorów

2. Wizualizacja Klastrow

Wizualizacja na wykresie dwuwymiarowym (np. scatter plot) może pomóc w ocenie jakości klasteryzacji. Dobrze oddzielone klastry będą widoczne jako oddzielne grupy punktów.