Politechnika Wrocławska

Zadanie rekrutacyjne Solvro — analiza EDA dla zestawu cocktailDB

Jakub Jasiński

21.03.2025

Spis treści

1 Wprowadzenie

W niniejszym dokumencie przedstawiam analizę eksploracyjną (EDA), wykonaną w języku Java, zestawu danych pochodzących z bazy koktajli. Do większości operacji wykorzystano pakiet smile w wersji 4.3.0. Celem zadania rekrutacyjnego było:

- Przeprowadzenie wstępnej analizy danych (oczyszczanie, standaryzacja, usuwanie duplikatów, uzupełnianie brakujących wartości).
- Zastosowanie metod redukcji wymiarowości (PCA, UMAP) oraz klasteryzacji (K-Means, DBSCAN).
- Ewaluację jakości klasteryzacji przy użyciu metryk takich jak silhouette score oraz wizualizację wyników.

2 Opis zestawu danych

Zestaw danych zawiera informacje o 134 koktajlach, w tym nazwy, składniki, opisy oraz dane dotyczące zawartości alkoholu. Ze względu na charakter danych, przy przetwarzaniu zastosowano metody normalizacji nazw, uzupełniania brakujących pól oraz usuwania duplikatów składników.

3 Przetwarzanie danych

3.1 Czyszczenie i normalizacja

Przeprowadzono normalizację nazw koktajli oraz uzupełnianie brakujących wartości w składnikach koktalji, zastosowano m.in usuwanie zbednych znaków, wypełnianie pustych pól, stemming, usuwanie stop-words.

3.2 Tworzenie wektorów cech

Wykorzystano metodę TF-IDF do ekstrakcji cech na podstawie nazw składników. TF został zaimplementowany binarnie, a IDF klasyczną metodą. Warto w tym miejscu zaznaczyć, że inne konfiguracje tej metody, a także inne rozwiązania, nie wykazały znaczącej poprawy klasteryzacji danych. TF-IDF, jako jedna z nielicznych metod została zaimplementowana od zera, co może się przekładać na pogorszenie klasteryzacji, jednakże dalsze etapy ewaluacji nie wykazały takich prawidłowości.

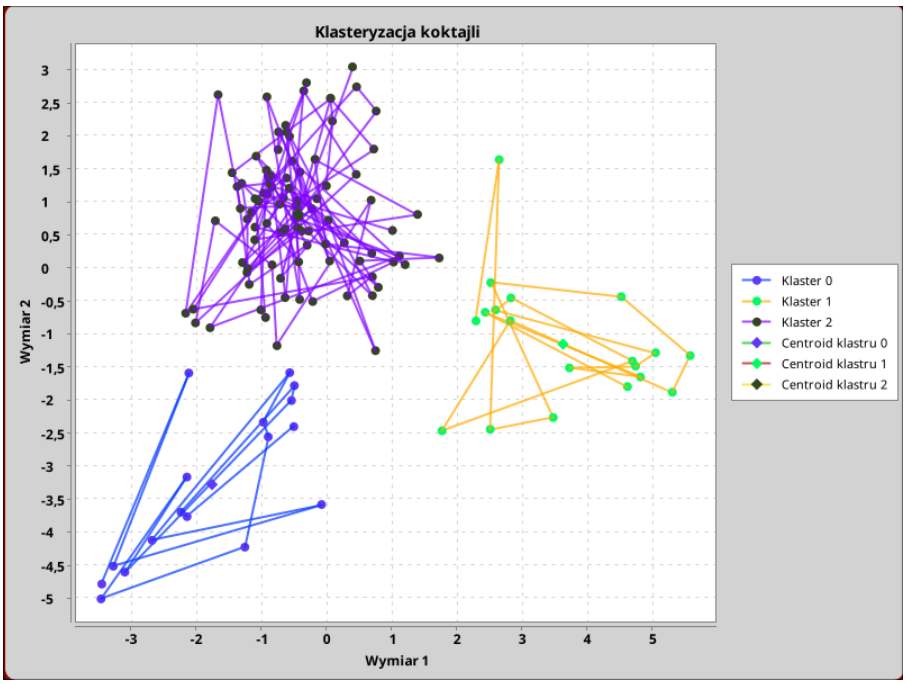
4 Redukcja wymiarowości i klasteryzacja

4.1 Redukcja wymiarowości

Do redukcji wymiarowości danych wykorzystano PCA (główne składowe). Na podstawie wyników grid searchu wybrano optymalną liczbę wymiarów, która maksymalizowała silhouette score. Ponadto zostało przeprowadzone badanie z wykorzystaniem metody UMAP, jednakże w każdym możliwym testowym przypadku (wykorzystanie klasteryzacji: K-means, DBSCAN, X-means), znaczenie pogorszenie wyniku klasteryzacji.

4.2 Klasteryzacja

Klasteryzacja została przeprowadzona głównie przy użyciu algorytmu K-Means. Dodatkowo eksperymentowano z DBSCAN, jednak wyniki nie przyniosły znaczącej poprawy, algorytm w większości wypadków niezależnie od użytej wartości eps czy minPts uznawał większość elementów zbioru jako szum. Metoda X-Means z pakietu smile. Grid search pozwolił na automatyczny dobór liczby klastrów i docelowej liczby wymiarów.



Rysunek 1: Wizualizacja klastrów uzyskanych metodą K-Means.

5 Ewaluacja klasteryzacji

W procesie oceny jakości klasteryzacji zastosowaliśmy szereg miar statystycznych, które pozwalają na ilościową ocenę spójności i separacji wyznaczonych klastrów. Oto, co wskazują poszczególne metryki:

- Silhouette Score (0.573) Ta miara mierzy, jak dobrze każdy punkt jest dopasowany do swojego klastra w porównaniu z sąsiadującymi klastrami. Wartość powyżej 0.5 sugeruje, że punkty są dobrze skwantyfikowane – mają one wysoką spójność wewnątrzklastrową oraz dobrą separację między klastrami. W naszym przypadku wynik 0.573 wskazuje, że klasteryzacja jest na ogół trafna, choć dla niektórych punktów wartość może być niższa, co wskazuje na pewne niejednorodności w strukturze danych.

- Davies-Bouldin Index (0.59) Indeks ten mierzy stosunek sumy wewnątrzklastrowych odległości (rozproszenia) do odległości między centroidami klastrów. Niższe wartości są pożądane, a wynik poniżej 1 (0.59) sugeruje, że klastery są dobrze oddzielone od siebie oraz wewnętrznie zwarte.

- Calinski-Harabasz Index (156.47) Ta miara ocenia stosunek międzyklastrowej wariancji do wariancji wewnątrzklastrowej. Wyższe wartości wskazują na silniejsze odróżnienie klastrów. Uzyskany wynik 156.47 świadczy o względnie dobrej strukturze klastrów, choć interpretacja tej metryki zależy także od liczby próbek i wymiarowości danych.

- Dunn Index (0.099) Dunn Index wskazuje stosunek najmniejszej międzyklastrowej odległości do największej wewnątrzklastrowej rozpiętości. Niska wartość (0.099) może sugerować, że minimalna odległość między niektórymi klastrami jest mała w porównaniu do rozpiętości wewnątrz klastrów. Mimo że ta miara bywa bardzo czuła na wartości odstające, wskazuje ona na obszary, w których separacja klastrów mogłaby być lepsza.

- Gap Statistic (0.0) Gap Statistic porównuje wyniki klasteryzacji z oczekiwaniami uzyskanymi przy losowym podziale danych. Wynik równy 0.0 może sugerować, że różnica między strukturą wyznaczoną przez algorytm a strukturą losową jest niewielka. Może to być wynikiem specyfiki zbioru danych – małej liczby próbek oraz dużej heterogeniczności.

- Dispersion (średnia odległość wewnątrz klastrów: 1.205) Ta wartość, wraz z medianą (1.175) i stosunkowo wysokim odchyleniem standardowym (0.626), sugeruje, że choć klastery są na ogół

zwarte, w niektórych przypadkach występuje większa zmienność odległości wewnątrzklastrowych. Może to oznaczać, że niektóre klastery mają mniej jednorodną strukturę.

6 Dyskusja i wnioski

- **Silhouette Score vs. Ocena wizualna:** Pomimo stosunkowo wysokiej wartości silhouette score (0.57 po usunięciu outlierów), wizualna ocena klastrów sugeruje, że dane mogą nie być idealnie podzielone. Może to wynikać z faktu, że zestaw danych zawiera tylko 132 próbki.
- **Ograniczenia zestawu danych:** Niewielka liczba próbek oraz zmienność cech (np. opisy składników, sposób zapisu nazw) mogą wpływać na trudność uzyskania lepszych wyników klasteryzacji.

7 Wnioski końcowe

Na podstawie analizy ilościowej (metryki ewaluacyjne) oraz wizualnej (wykresy centroidów, boxploty, silhouette ploty) można sformułować następujące wnioski:

- Jakość Klasteryzacji: Uzyskane metryki, takie jak Silhouette Score oraz Davies-Bouldin Index, wskazują, że struktura klastrów jest ogólnie dobra. Wynik Silhouette Score powyżej 0.5 sugeruje, że punkty są dobrze przypisane do klastrów, a klastery są dobrze oddzielone. Jednak niska wartość Dunn Index oraz zerowy Gap Statistic sugerują, że istnieją pewne problemy z minimalną separacją niektórych klastrów – co może być wynikiem naturalnej zmienności małego zbioru danych.
- Charakterystyka Zestawu Danych: Zestaw danych składający się z około 130 próbek (koktajli) jest niewielki, co utrudnia osiągnięcie bardzo wysokich wyników ewaluacyjnych. Dane te są także bardzo heterogeniczne – zawierają dużą zmienność składników oraz różnorodne kombinacje, co może wpływać na wyniki klasteryzacji.
- Wykorzystanie Wielu Metryk: Korzystanie z kilku miar ewaluacyjnych pozwala uzyskać bardziej wszechstronną ocenę jakości klasteryzacji. Choć Silhouette Score jest dobrym wskaźnikiem, warto także brać pod uwagę inne metryki (Calinski-Harabasz, Davies-Bouldin, Dunn, Gap Statistic), aby lepiej zrozumieć, jak klastery są rozproszone i jak silnie się od siebie różnią.
- Wnioski Wizualne: Wizualna analiza wykresów (np. mapy rozkładu klastrów, boxploty dla centroidów) dostarcza dodatkowego kontekstu – pokazuje, że mimo dobrych wyników metrycznych, dane wykazują duże zróżnicowanie. Może to wskazywać, że niektóre klastery są bardziej „rozciągnięte", a inne bardziej zwarte, co należy uwzględnić przy interpretacji wyników.

Podsumowując, choć uzyskane wyniki wskazują na dobrą jakość klasteryzacji w pewnych aspektach (np. Silhouette Score), istnieje kilka wskaźników, które sugerują potrzebę dalszej optymalizacji i uwzględnienia specyfiki danych. Niewielka liczba próbek oraz duża heterogeniczność zestawu danych stanowią wyzwanie

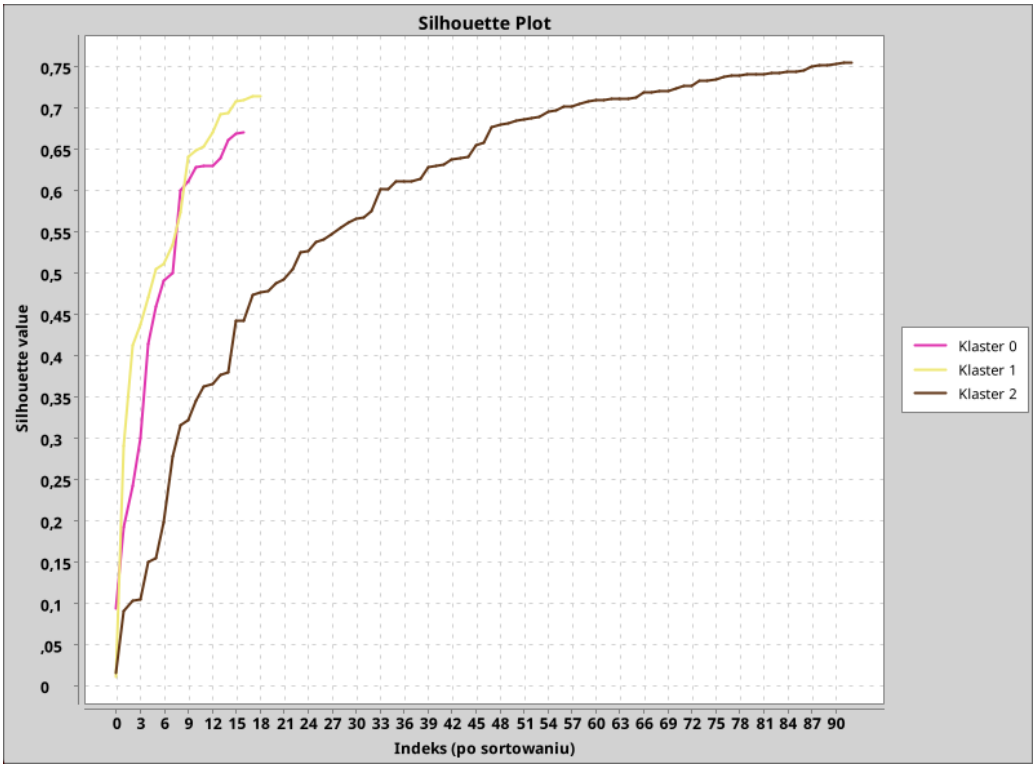
8 Refaktoryzacja

W ramach refaktoryzacji kodu zastosowano podejście modułowe, gdzie każdy komponent aplikacji (np. ładowanie danych, przetwarzanie, ekstrakcja cech, klasteryzacja, wizualizacja) jest testowany osobno przy użyciu narzędzi takich jak JUnit. Dzięki temu możliwe jest zapewnienie poprawności działania całej aplikacji oraz łatwiejsza konserwacja kodu.

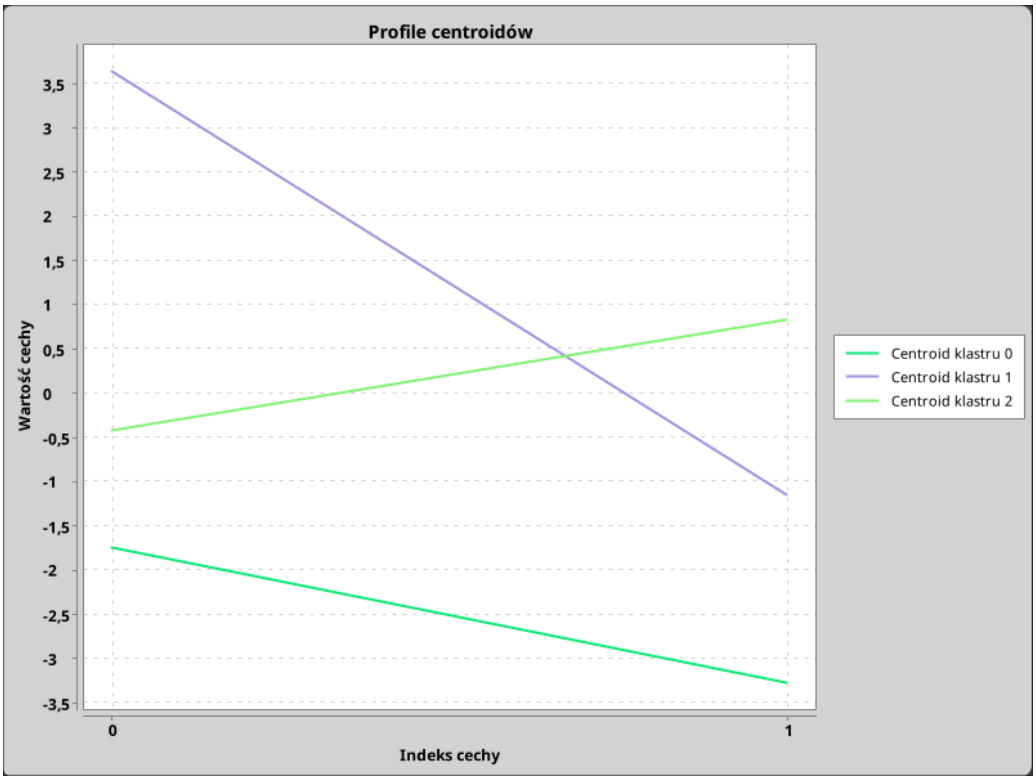
9 Podsumowanie

Przedstawiona analiza i wdrożenie metod klasteryzacji koktajli pozwoliło na identyfikację optymalnych parametrów klasteryzacji przy użyciu grid searchu. Mimo ograniczeń wynikających z niewielkiej liczby próbek, zastosowane metody umożliwiły uzyskanie spójnych wyników, które zostały poddane krytycznej analizie ilościowej i wizualnej.

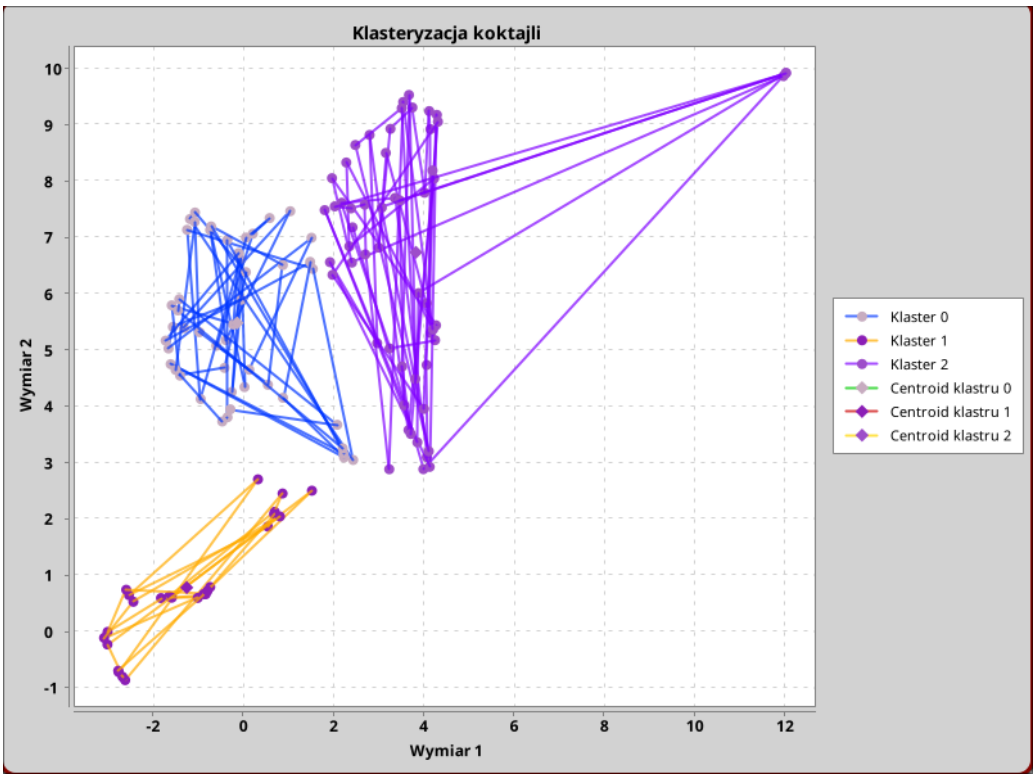
Załączniki



Rysunek 2: Wykres silhouette dla optymalnej konfiguracji.



Rysunek 3: Wykres centroidów klastrów.



Rysunek 4: Wizualizacja klastrów z użyciem UMAP.