Zadanie rekrutacyjne do KN Solvro

Sekcja uczenia maszynowego

wiosna 2025

Krzysztof Głowacz

Spis treści

1 Wstęp

W niniejszym raporcie przedstawione zostało rozwiązanie zadania rekrutacyjnego do sekcji uczenia maszynowego Koła Naukowego Solvro (rekrutacja wiosenna 2025). Zadanie polegało na eksploracyjnej analizie danych oraz klasteryzacji podanego zbioru [1].

2 Analiza danych

Zbiór danych, pochodzący z bazy danych TheCocktailDB, zawierał listę koktajli wraz ze składnikami niezbędnymi do ich przyrządzenia. Link do zbioru danych został dołączony do sekcji Źródeł tego raportu [2].

Pierwsza analiza zbioru danych wykazała, że konieczna będzie osobna analiza składników koktalji, które początkowo były reprezentowane jako zagnieżdżone struktury danych wewnątrz rekordów opisujących koktajle. W celu przeprowadzenia poprawnej wstępnej analizy danych (*EDA*) napisany został skrypt `src/eda.py`, który miał za zadanie:

- wczytać dane z pliku w formacie JSON do formatu DataFrame z biblioteki Pandas,
- wydzielić dane dotyczące składników koktajli do osobnego DataFrame'a,
- wyświetlić podstawowe statystyki danych,
- wygenerować pełne raporty opisujący dane o koktajlach i składnikach koktajli korzystając z biblioteki ydata-profiling.

Na zdjęciu nr 1 przedstawiony został fragment wyjścia standardowego po uruchomieniu skryptu.

```
❯ uv run src/eda.py
Upgrade to ydata-sdk
Improve your data and profiling with ydata-sdk, featuring data quality scoring, redundancy detection, outlier identification,
Register at ttps://ydata.ai/register
2025-03-24 21:59:14 [info     ] Loading data from /home/kris/Studia/inne/kn-solvro/rekrutacja-2025/data/cocktail_dataset.json
2025-03-24 21:59:14 [info     ] Starting cocktail dataset exploration


=== Cocktail Dataset Exploration ===
Number of cocktails: 134

Columns info:
    Column        Data Type      Missing Values    Missing Percentage
--  ------------  -----------  ----------------  --------------------
 0  id            int64                       0                     0
 1  name          object                      0                     0
 2  category      object                      0                     0
 3  glass         object                      0                     0
 4  tags          object                     99                 73.88
 5  instructions  object                      0                     0
 6  imageUrl      object                      0                     0
 7  alcoholic     int64                       0                     0
 8  createdAt     object                      0                     0
 9  updatedAt     object                      0                     0
10  ingredients   object                      0                     0

Category distribution
---------------------
category
Ordinary Drink         127
Cocktail                 6
Punch / Party Drink      1
Name: count, dtype: int64
```

Zdjęcie 1: Uruchomienie skryptu src/eda.py

Informacje dostarczone przez skrypt i wygenerowany raport pozwoliły na wyciągnięcie nastepujących wniosków:

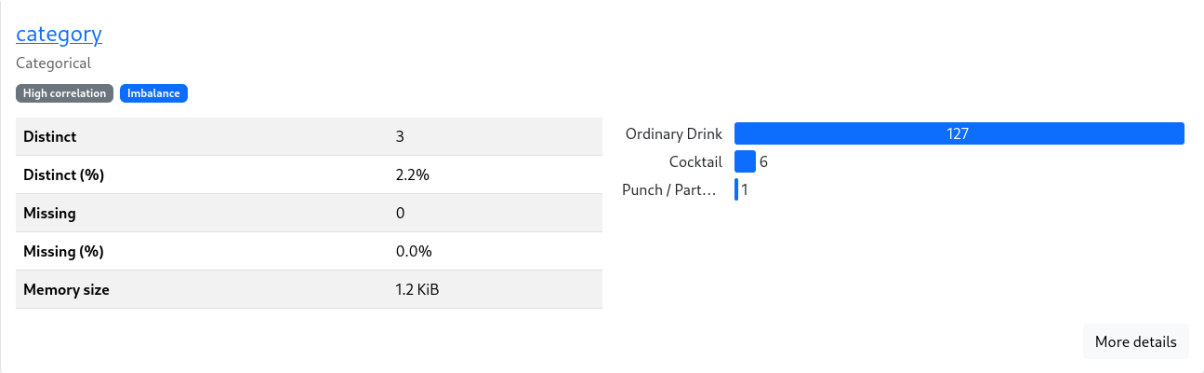
1. Każdy koktajl ze zbioru danych opisany jest przez 11 cech, przy czym 4 z nich są zupełnie nieistotne z punktu widzenia dalszej analizy, ponieważ albo zawierają informacje typowo bazodanowe (kolumny: *id*, *createdAt*, *updatedAt*), albo są linkiem URL do zdjęcia przedstawiającego dany koktajl (*imageUrl*). Cecha *tags* także została na tym etapie odrzucona - niespełna 74% koktajli nie

ma żadnego tagu przypisanego, a gdy koktajl ma przypisane tagi to sprawiają one wrażenie bardzo luźno powiązanych, niezbyt informatywnych. Cecha *(*instructions) także została odrzucona, ponieważ zawiera ona opis przygotowania koktajlu w języku naturalnym, co nie stanowiło dobrej podstawy do późniejszej klasteryzacji.

2. Cecha *category* jest mało wartościowa w kontekście późniejszej klasteryzacji ze względu na fakt, że zawiera jedynie 3 różne wartości, przy czym zdecydowana większość koktajli ma przypisaną taką samą kategorię (Ordinary Drink - zdjęcie 2).

3. Cecha *alcoholic* jest nieistotna ze względu na to, że przyjmuje stałą wartość równą 1 (tzn. wszystkie drinki są alkoholowe).

4. Jedyną cechą składników koktajli łatwą do uwzględnienia w późniejszej klasteryzacji była ich alkoholowość, tzn. flaga informująca o tym, czy dany składnik zawiera alkohol.



Zdjęcie 2: Analiza cechy *category*

3 Przygotowanie danych

Ciekawym pomysłem na poradzenie sobie z zagnieżdżoną strukturą składników koktajli była prosta ich agregacja - przypisanie do każdego koktajlu łącznej liczby składników, liczby składników zawierających alkohol i tych bez alkoholu. Następnie usunięto z DataFrame'u niepotrzebne kolumny (opisane w poprzedniej sekcji).

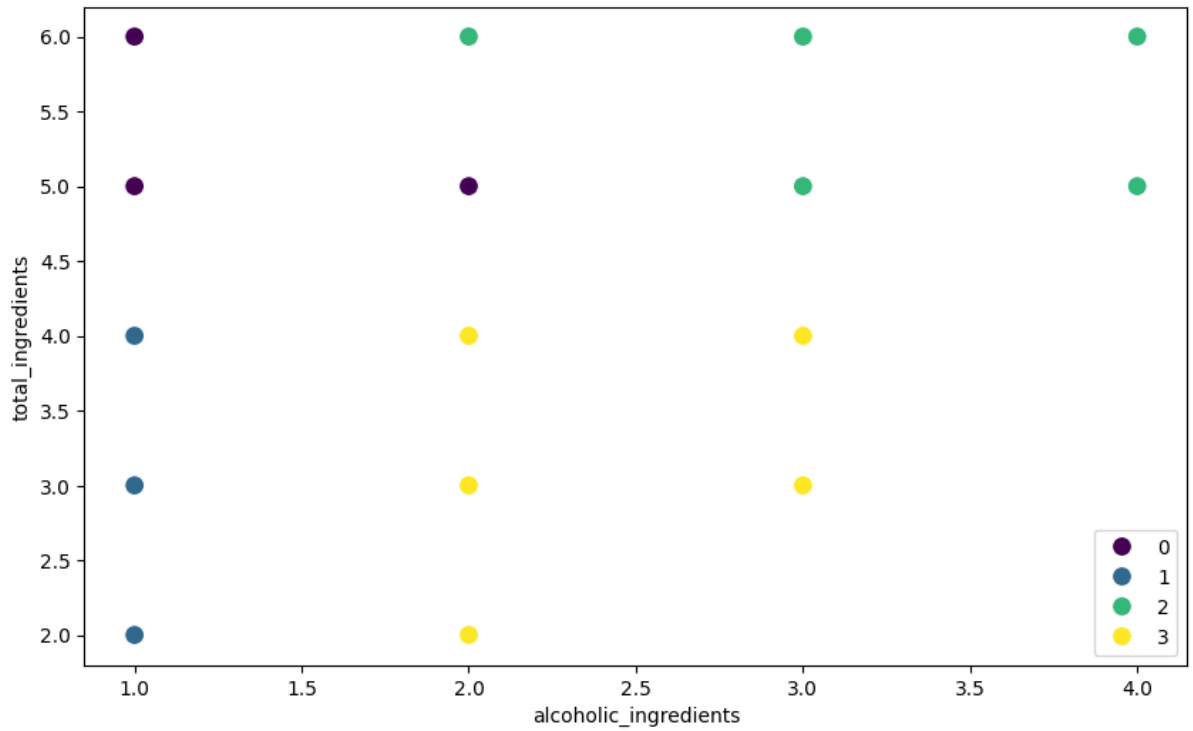
Kolejnym etapem przygotowania danych było ich enkodowanie - zastosowano `LabelEncoder` na kolumnie *glass*. Ostatnim krokiem tego procesu było przeskalowanie danych przy pomocy `StandardScalera`.

4 Klasteryzacja

Do wyznaczenia liczby klastrów użyto *Elbow Method*, która wskazywała na 4 lub 5 klastrów. Przyjęto wartość 4 ze względu na stosunkowo mały zbiór danych. Dane zostały sklasteryzowane metodą `KMeans` z biblioteki *scikit-learn*. Wykres ze zdjęcia 3 prezentuje efekty klasteryzacji.

Widoczny jest podział koktajli względem łącznej liczby potrzebnych składników, a także względem alkoholowych składników. Możemy wyróżnić 4 grupy koktajli:

- te, które wymagają tylko 1 składnika alkoholowego oraz niewielu składników sumarycznie (2/3/4).
- te, które wymagają niewielu składników alkoholowych (1/2) oraz wielu składników sumarycznie (5/6).
- te, które wymagają kilku składników alkoholowych (2/3) przy jednoczesnej niewielkiej liczbie sumarycznych składników (2/3/4).
- te, które wymagają wielu składników alkoholowych (2/3/4) i wielu składników sumarycznie (5/6).



Zdjęcie 3: Efekt klasteryzacji

Można się zgodzić, że dokonany podział jest zasadny i stosunkowo logiczny. Poza ewaluacją jakościową przeprowadzono także ewaluację ilościową i otrzymano następujące wartości metryk:

Metryka	wartość
Inertia	324.19
Silhouette Score	0.3119
Calinski-Harabasz Score	46.22
Davies-Bouldin Score	1.3301

Na ich podstawie możemy stwierdzić, że dokonana klasteryzacja, nawet bez odpowiedniego tuningowania hiperparametrów okazała się całkiem skuteczna, co potwierdziły przeprowadzone ewaluacje.

Źródła

[1] Pełna treść zadania rekrutacyjnego
`https://github.com/Solvro/rekrutacja/blob/main/machine_learning.md`.
Dostęp: 22.03.2025.

[2] Zbiór danych o koktajlach
`https://github.com/Solvro/rekrutacja/blob/main/data/cocktail_dataset.json`.
Dostęp: 22.03.2025.