Politechnika Śląska
Wydział Matematyki Stosowanej

Algorytmy Eksploracji Danych

Dokumentacja Projektowa

Analiza zachowań klientów sklepu internetowego przy użyciu metod redukcji wymiarów, klasteryzacji i reguł asocjacyjnych

Autorzy: Jakub Darul, Mateusz Lamla
Grupa: 1
Semestr: V
Stopień: I

24 stycznia 2026

Spis treści

1 Wstęp

1.1 Cel projektu

Celem niniejszego projektu jest przeprowadzenie kompleksowej analizy danych transakcyjnych pochodzących ze sklepu internetowego. Analiza ma na celu wyodrębnienie grup klientów o podobnych profilach zakupowych oraz odkrycie ukrytych wzorców (reguł asocjacyjnych) rządzących doborem produktów do koszyka. Wiedza ta w warunkach rzeczywistych pozwoliłaby na optymalizację strategii marketingowej oraz układu sklepu.

1.2 Zakres prac

Projekt obejmuje trzy główne etapy analizy:

1. Przygotowanie danych i inżynierię cech (stworzenie modelu RFM - Recency, Frequency, Monetary).
2. Redukcję wymiarowości danych w celu wizualizacji struktury zbioru.
3. Segmentację klientów (klasteryzację).
4. Wykrywanie reguł asocjacyjnych (analiza koszykowa).

2 Opis wykorzystanego zbioru danych

Do analizy wykorzystano zbiór danych **"Online Retail Data Set"**, dostępny publicznie w repozytorium UCI Machine Learning Repository.

Zbiór zawiera transakcje z brytyjskiego sklepu internetowego (e-commerce) sprzedającego upominki, z okresu od 01.12.2010 do 09.12.2011.

Główne atrybuty zbioru:

- `InvoiceNo`: Unikalny numer transakcji.
- `StockCode`: Kod produktu.
- `Description`: Nazwa produktu.
- `Quantity`: Liczba sztuk produktu w transakcji.
- `InvoiceDate`: Data i czas transakcji.
- `UnitPrice`: Cena jednostkowa (w funtach szterlingach).
- `CustomerID`: Unikalny identyfikator klienta.
- `Country`: Kraj zamieszkania klienta.

Przed analizą dane zostały oczyszczone z brakujących identyfikatorów klientów oraz zwrotów (ujemne wartości w polu `Quantity`).

3 Opis zastosowanych metod

W projekcie wykorzystano trzy grupy metod eksploracji danych.

3.1 Redukcja wymiarów: PCA

Cel: Zmniejszenie liczby zmiennych opisujących klienta przy zachowaniu jak największej ilości informacji (wariancji), co umożliwia wizualizację wielowymiarowych danych na płaszczyźnie 2D.
Charakterystyka: PCA (Principal Component Analysis) to technika statystyczna przekształcająca zbiór skorelowanych zmiennych w mniejszy zbiór nieskorelowanych zmiennych zwanych głównymi składowymi.

3.2 Klasteryzacja: k-Means

Cel: Podział bazy klientów na rozłączne grupy (segmenty), wewnątrz których klienci są do siebie podobni.
Charakterystyka: Algorytm k-średnich (k-Means) iteracyjnie przypisuje punkty danych do jednego z k skupień, dążąc do minimalizacji wariancji wewnątrz klastrów. Wymaga wcześniejszego określenia liczby grup.

3.3 Reguły asocjacyjne: Algorytm Apriori

Cel: Znalezienie powiązań między produktami, tzn. określenie, jakie produkty są często kupowane razem.
Charakterystyka: Algorytm Apriori przeszukuje bazę transakcji w celu znalezienia częstych zbiorów produktów, a następnie generuje reguły typu "Jeżeli klient kupił A, to kupi B"na podstawie miar takich jak *Support*, *Confidence* i *Lift*.

4 Opis implementacji

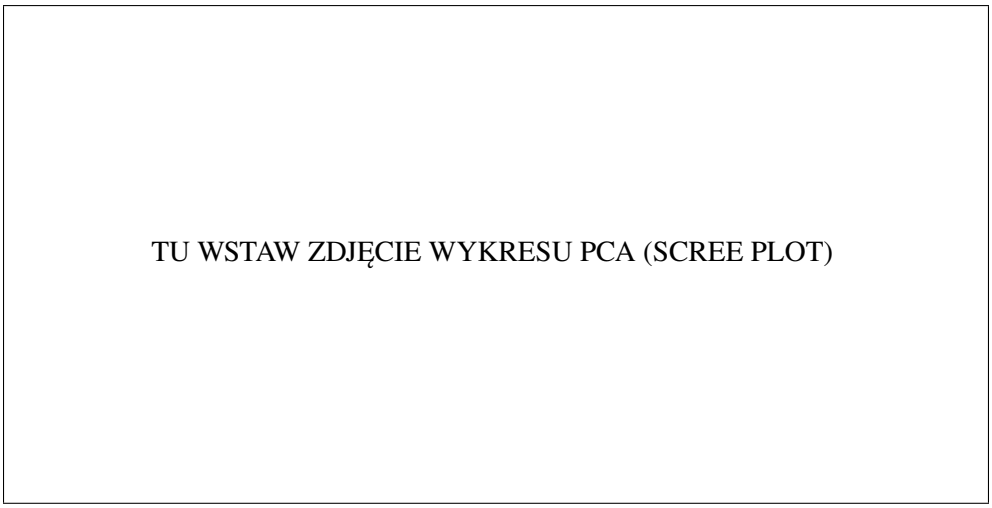
Projekt został zrealizowany w języku **Python**. Do analizy wykorzystano następujące biblioteki:

- **Pandas & NumPy:** Przetwarzanie i agregacja danych (stworzenie tabeli z cechami: Frequency, Monetary, Variety).
- **Scikit-learn:** Standaryzacja danych (`StandardScaler`), implementacja PCA oraz algorytmu k-Means.
- **Mlxtend:** Implementacja algorytmu Apriori oraz generowanie reguł asocjacyjnych.
- **Matplotlib & Seaborn:** Wizualizacja wyników.

5 Wyniki i interpretacja

5.1 Wyniki redukcji wymiarów (PCA)

Zastosowanie PCA pozwoliło zredukować trzy wymiary opisujące klienta (częstotliwość zakupów, wartość koszyka, różnorodność) do dwóch głównych składowych.

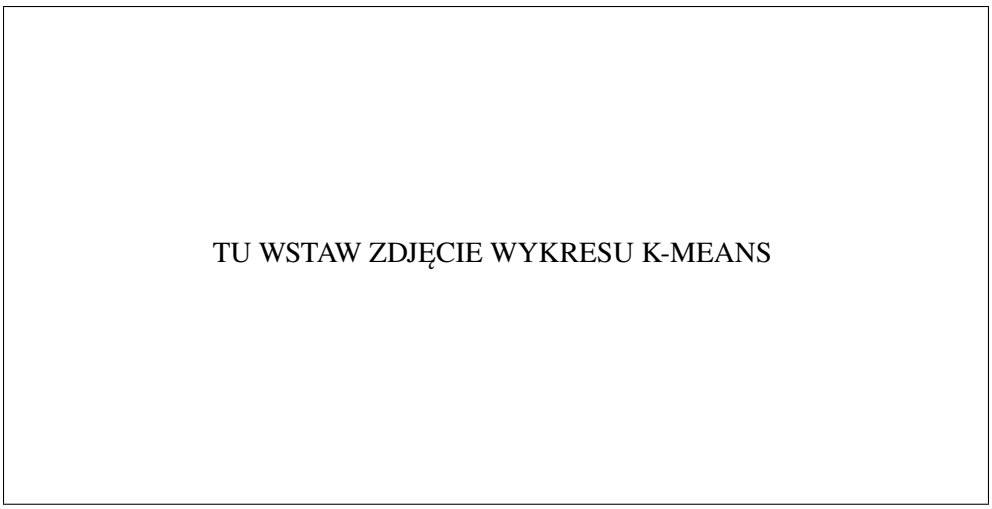


Rysunek 1: Wykres osypiska (Scree Plot) pokazujący wariancję wyjaśnioną przez składowe.

Wynik liczbowy: Pierwsze dwie składowe (PC1 i PC2) wyjaśniają łącznie około 90% wariancji zbioru. Oznacza to, że reprezentacja 2D jest wiarygodnym przybliżeniem rzeczywistej struktury danych.

5.2 Wyniki klasteryzacji (k-Means)

Algorytm k-Means podzielił klientów na 3 klastry. Poniższy wykres przedstawia rozmieszczenie klientów w przestrzeni wyznaczonej przez PCA.



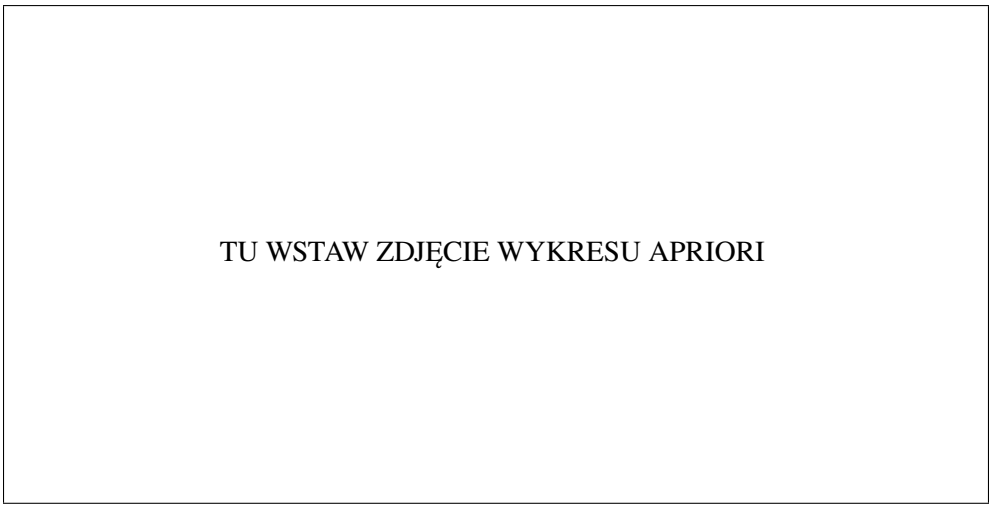
Rysunek 2: Segmentacja klientów - wizualizacja klastrów na płaszczyźnie PCA.

Interpretacja klastrów (na podstawie średnich wartości):

- **Klaster 0 (Niebieski):** Klienci okazjonalni. Niska częstotliwość zakupów i niska wartość koszyka. Stanowią najliczniejszą grupę.
- **Klaster 1 (Zielony):** Klienci regularni. Średnie wydatki, częstsze powroty do sklepu.
- **Klaster 2 (Żółty):** Klienci VIP / Hurtownicy. Bardzo wysoka wartość `Monetary` i `Frequency`. Jest to grupa nieliczna, ale kluczowa dla przychodów sklepu.

5.3 Wyniki reguł asocjacyjnych (Apriori)

Analiza koszykowa (dla transakcji z Francji) pozwoliła wykryć silne zależności między produktami. Wykres przedstawia zależność między wsparciem (Support) a pewnością (Confidence) reguł.



Rysunek 3: Rozkład reguł asocjacyjnych: Support vs Confidence (Kolor = Lift).

Przykładowa znaleziona reguła:

```
Antecedents: {SET/6 RED SPOTTY PAPER PLATES}
Consequents: {SET/6 RED SPOTTY PAPER CUPS}
Lift: > 1, Confidence: ~0.8-0.9
```

Interpretacja: Klienci kupujący papierowe talerzyki w czerwone kropki z bardzo dużym prawdopodobieństwem (bliskim 90%) kupują również pasujące do zestawu kubeczki. Wysoki wskaźnik *Lift* potwierdza, że nie jest to zbieg okoliczności, lecz silna korelacja produktowa.

6 Podsumowanie

Przeprowadzona analiza pozwoliła na skuteczne przetworzenie surowych danych transakcyjnych w użyteczną wiedzę biznesową. Dzięki metodzie PCA możliwa była wizualizacja wielowymiarowych danych. Klasteryzacja k-Means pozwoliła wyodrębnić grupę najbardziej dochodowych klientów (VIP), do których można skierować dedykowane kampanie marketingowe. Z kolei reguły asocjacyjne wskazały konkretne pary produktów (np. zestawy imprezowe), które powinny być oferowane razem (cross-selling) w celu zwiększenia sprzedaży.

Wybór metod okazał się trafny dla specyfiki danych e-commerce, łącząc analizę behawioralną (segmentacja) z analizą produktową (koszykową).