Politechnika Wrocławska
Wydział Elektroniki

Hurtownie i eksploracja danych
Projekt

Sklep monopolowy

Autorzy:
Karol Jędrusik nr 226794
Adam Łaput nr 241350

Prowadzący zajęcia projektowe:
Dr hab. inż. Robert Burduk

Termin zajęć: **Piątek TN 8:15-11:00**

26 listopada 2020

Spis treści

1 Cel projektu

Celem projektu jest stworzenie schematu hurtowni danych sklepu monopolowego znajdującego się w bliskiej i dozwolonej prawnie odległości od stadionu miejskiego na którym odbywają się różnego rodzaju wydarzenia (sportowe oraz kulturalne). W zależności od czasu odbywania się wydarzenia na stadionie (oraz jego typu), pogody w danym terminie oraz typu płatności przeprowadzane zostaną badania na temat sprzedaży artykułów alkoholowych.

2 Fakt

Analizowanym **faktem**, czyli pojedynczym zdarzeniem będącym podstawą analiz [1], **jest sprzedaż** artykułów monopolowych.

3 Miary

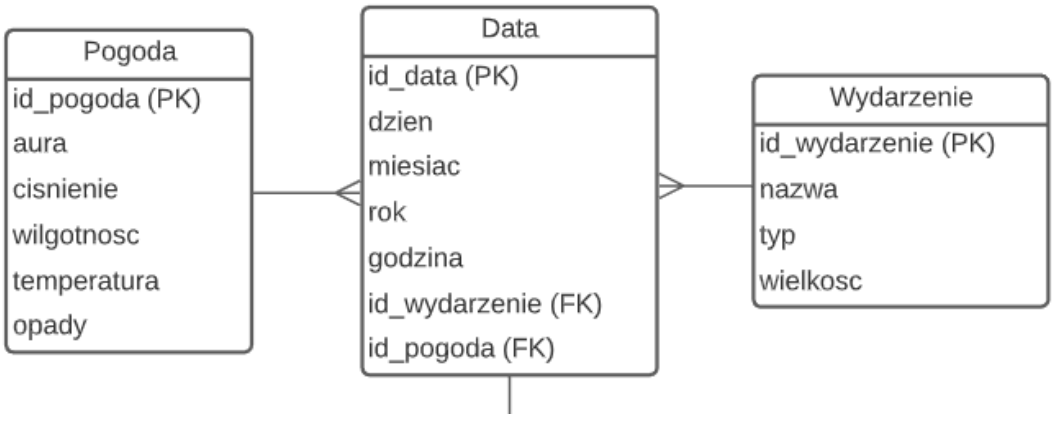
Do określonego **faktu sprzedaży** można określić następujące **miary** (wartości liczbowe przyporządkowane do danego faktu [1]):

- **wartość przychodu z transakcji** - miara opisuje, ile pieniędzy udało się uzyskać;
- **liczba sprzedanego typu alkoholu** - miara opisuje, ile sprzedano sztuk artykułu;

4 Wymiary

Do opisu **wymiarów** (cech opisujących dany fakt) służą **atrybuty**, które opisują wymiary i przechowują dodatkowe informacje na temat faktu[1]. W Naszym projekcie wyszczególniamy następujące wymiary i ich atrybuty:

- **Data** - wymiar dotyczący czasu wykonania transakcji, odbywającego się wydarzenia oraz stanu pogodowego;
  - Tabela Data
    - Dzień - od 1 do 31;
    - Miesiąc - od 1 do 12;
    - Rok - od 2020 do 2100;
    - Godzina - od 0 do 23;
  - Tabela Wydarzenie
    - Nazwa wydarzenia - dokładna nazwa wydarzenia;
    - Typ wydarzenia - m.in. kulturalne, sportowe;
    - Wielkość wydarzenia - szacowana liczba uczestników wydarzenia;
  - Tabela Pogoda
    - Aura - m.in. pochmurna, słoneczna, mglista;
    - Ciśnienie - wyrażone w hPa;
    - Wilgotność - wyrażona w %;
    - Temperatura - wyrażona w Celcjuszach;
    - Opady - m.in. brak, małe, średnie, duże;

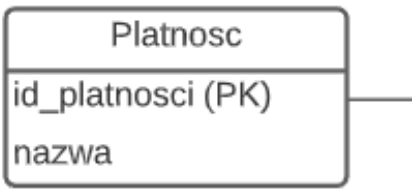


Rysunek 1: Źródło danych - Data - Model logiczny

- **Produkt** - wymiar dotyczący sprzedawanego produktu;
  - Tabela Produkt
    - Nazwa - dokładna nazwa produktu;
    - Marka - marka produktu;
    - Pojemność - wyrażona w mililitrach;
    - Zawartość alkoholu - wyrażona w %;
    - Typ opakowania - m.in. butelka szklana, butelka plastikowa, puszka;
    - Cena produktu - wyrażona w PLN;

Rysunek 2: Źródło danych - Produkt - Model logiczny

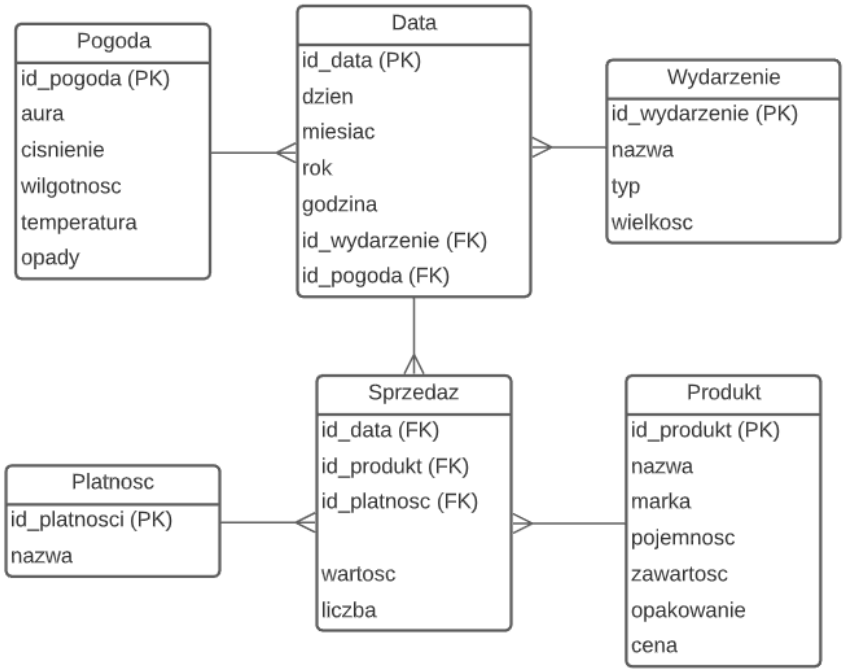
- **Płatność** - wymiar dotyczący typu płatności.
  - Tabela Płatność
    - Nazwa - nazwa typu płatności;



Rysunek 3: Źródło danych - Płatność - Model logiczny

5 Schemat logiczny hurtowni danych

Schemat logiczny hurtowni danych przedstawia się w następujący sposób:



Rysunek 4: Schemat logiczny hurtowni danych

Schemat logiczny zawiera w sobie tablicę faktów (*Sprzedaz*) oraz tablice wymiarów (*Wydarzenie, Data, Pogoda, Produkt, Płatność*).

6 Pytania analityczne

- Jak wpływa rodzaj opadów lub ich brak na średnią wartość pojedynczych transakcji?
- Czy istnieje zależność pomiędzy wartością wszystkich transakcji podczas trwania wydarzeń, a ich typem oraz wielkością?
- Czy wartość transakcji determinuje użycie wybranego typu płatności?
- Czy wielkość sumy wartości pojedynczych transakcji zależy od miesiąca dokonania transakcji?
- Jakie marki produktów są najczęściej kupowane w danym miesiącu?

7 Eksploracja danych

7.1 Cel eksploracji

Celem eksploracji danych jest analiza informacji oraz próba znalezienia użytecznych reguł i zależności wśród wybranego zestawu danych[2].

7.2 Zestaw danych

Wybrany zestaw danych opisuje cechy różnych wariantów białego portugalskiego wina *'Vinho Verde'* oraz ich ocenę w oczach ekspertów [3]. Zawiera w sobie 4898 instancji oraz 12 atrybutów (11 cech ocenianego wina oraz 1 końcowa ocena eksperta):

Tablica 1: Opis atrybutów

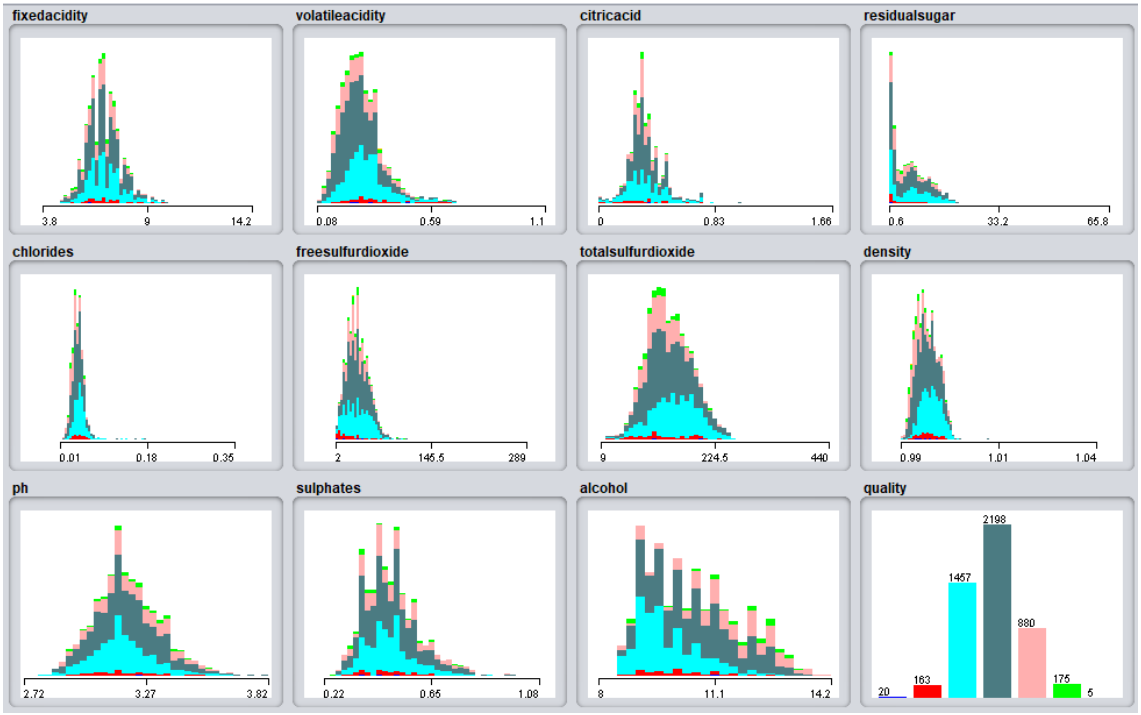
Lp.	*Atrybut*	*Opis*
1	Fixed acidity	stała kwasowość (g/l wina)
2	Volatile acidity	lotna kwasowość (g/l wina)
3	Citric acid	kwas cytrynowy (g/l wina)
4	Residual sugar	cukier resztkowy (g/l wina)
5	Chlorides	chlorki (g/l wina)
6	Free sulfur dioxide	dwutlenek siarki (mg/l wina)
7	Total sulfur dioxide	dwutlenek siarki i pochodne (mg/l wina)
8	Density	gęstość (g/cm^3 wina)
9	pH	pH wina (skala pH)
10	Sulphates	siarczany (g/l wina)
11	Alcohol	zawartość alkoholu (%)
12	Quality	ocena eksperta (w skali od 0 do 10)

7.3 Eksploracja zestawu danych

Jako narzędzie pomocne przy eksploracji danych wykorzystano oprogramowanie Weka 3.8.4.[4]. Oryginalny zestaw danych dostępny do pobrania ze strony internetowej istnieje jednak jako plik „.csv”. W celu operowania na danych w zakresie klasyfikacji przy wykorzystaniu oprogramowania Weka jesteśmy zmuszeni przekonwertować zestaw danych na plik „.arff”.

Wykresy cech

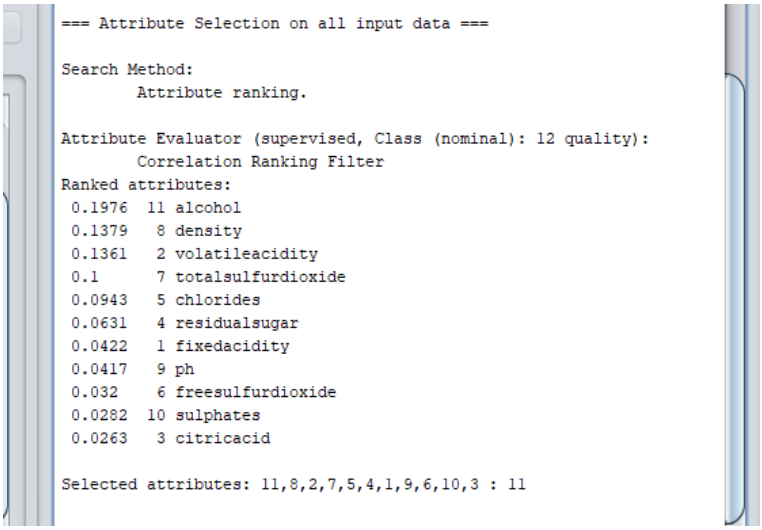
Po otworzeniu zestawu danych jesteśmy w stanie sprawdzić informacje na temat każdej cechy (wartość minimalna, wartość maksymalna, wartość średnia) oraz wykresy powiązań pomiędzy wartościami cech rekordów oraz przyporządkowaną oceną końcową (klasą).



Rysunek 5: Wykresy cech - zależnie od oceny końcowej

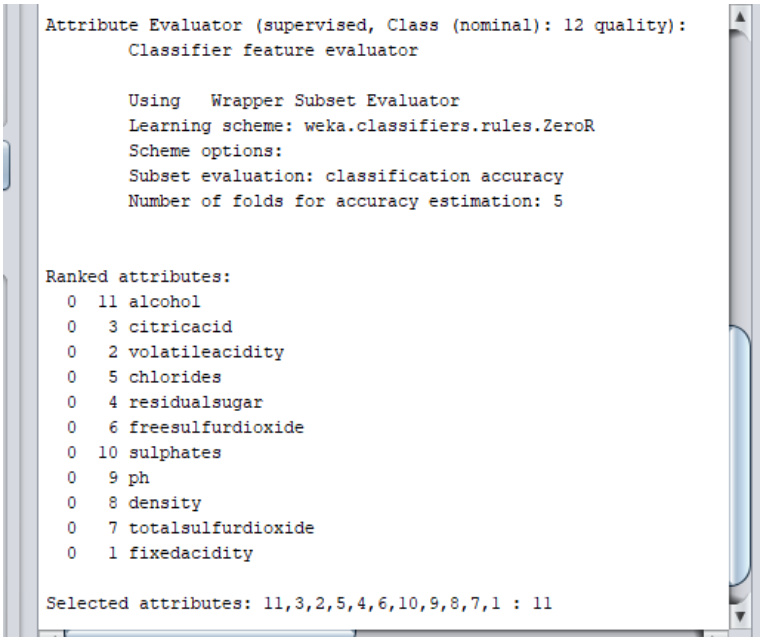
Ranking cech

Przy użyciu wbudowanych metod obliczamy ranking cech dla podanego zestawu danych. Wynikiem działania funkcji *CorrelationAttributeEval,* której zadaniem jest obliczenie korelacji pomiędzy daną cechą, a klasą wyjściową, jest poniższy ranking cech:



Rysunek 6: Ranking cech - CorrelationAttributeEval

Dodatkowo, w celu wyznaczenia rankingu cech użyliśmy funkcji *ClassifierAttributeEval.* Poniżej prezentujemy wyniki działania funkcji:



Rysunek 7: Ranking cech - ClassifierAttributeEval

Jak możemy zauważyć, w obu rankingach cech na pierwszym miejscu znajduje się atrybut „alcohol” wskazujący na wartość zawartości alkoholu w danej próbce wina. W obu rankingach na wysokim miejscu znajduje się także atrybut „volatile-acidity” (lotna kwasowość) oraz „chlorides” (chlorki). Podane rankingi mogą wskazywać więc, że są to jedne z najważniejszych elementów, przeważających o ocenie końcowej wystawianej przez *sommeliera*.

Klasyfikacja

Następnie przy pomocy różnych modeli klasyfikacji oraz zastosowania 2-krotnej walidacji krzyżowej przeprowadzamy eksperymenty klasyfikacji zestawu danych dla różnej liczby cech zgodnie z wyznaczonym wcześniej rankingiem cech (rozpoczynając od pierwszej cechy, aż do wzięcia pod uwagę wszystkich cech) obliczonym za pomocą funkcji *CorrelationAttributeEval*. Jako wskaźnik poprawności działania klasyfikatora uznajemy procent poprawnie rozpoznanych instancji. W problemie klasyfikacji wykorzystamy wbudowane funkcje z ich domyślnymi wartościami:

- HoeffdingTree
- RandomTree
- RandomForest

Tablica 2: Klasyfikacja - HoeffdingTree

Liczba atrybutow	*Wskaznik*
1	49.5917
2	47.3663
3	48.8975
4	48.4688
5	46.5904
6	44.2221
7	44.5284
8	44.3855
9	43.7117
10	44.1609
11	43.7730

Tablica 3: Klasyfikacja - RandomTree

Liczba atrybutow	*Wskaznik*
1	49.2650
2	50.8371
3	52.7154
4	53.4504
5	53.6954
6	54.5325
7	54.3283
8	53.9200
9	54.9816
10	54.9816
11	55.2266

Tablica 4: Klasyfikacja - RandomForest

Liczba atrybutow	*Wskaznik*
1	49.2854
2	47.3663
3	57.5949
4	59.6162
5	61.3516
6	62.5766
7	62.4336
8	62.5153
9	63.3932
10	63.8016
11	55.2266

Wnioski

Na podstawie otrzymanych wyników jesteśmy w stanie stwierdzić, iż najwyższą wartość wskaźnika otrzymaliśmy dla klasyfikatora „RandomForest” oraz liczby cech równej 10, a procent poprawnie sklasyfikowanych instancji wynosi 63,8016%.

Wyniki dla klasyfikatora „HoeffdingTree” obniżają się wraz ze wzrostem liczby cech. Możemy nawet stwierdzić, iż klasyfikator radzi sobie najlepiej wykorzystując wartości tylko jednej cechy - „alcohol”.

Wyniki dla klasyfikatora „RandomTree” wzrastają wraz ze wzrostem liczby cech, aż do osiągnięcia liczby cech równej 4. Kolejne wyniki są do siebie bardzo zbliżone.

Wyniki dla klasyfikatora „RandomForest” wzrastają wraz ze wzrostem liczby cech. Co ciekawe, wykorzystanie maksymalnej liczby cech poskutkowało otrzymaniem jednego z najniższych wyników.

Porównując klasyfikatory między sobą, klasyfikator „RandomForest” okazał się być najlepszym z nich. Drugim najlepszym klasyfikatorem jest „RandomTree”. W związku z tym dla liczby cech równej 10 przeprowadziliśmy badania eksperymentalne dla 8 różnych wartości *liczby drzew*. Wyniki zamieściliśmy w poniższej tabeli:

Tablica 5: RandomForest - liczba drzew, a wartość wskaźnika

Liczba drzew	*Wskaznik*
100	63.8016
150	63.5770
200	63.5361
300	63.9241
500	63.6995

W problemie klasyfikacji dla wybranego zestawu danych zwiększenie liczby drzew nie skutkuje zwiększeniem wartości wskaźnika, tzn. zwiększeniem liczby poprawnie sklasyfikowanych instancji.

typ płatności - miara opisuje, czy sprzedaż została wykonana za pomocą gotówki, karty płatniczej lub kodu rabatowego;

Bibliografia

[1] Polsko-japońska akademia technik komputerowych - hurtownie danych - wykład 3 - modele logiczne hurtowni danych. https://edu.pjwstk.edu.pl/wyklady/hur/scb/rW3.htm.

[2] wazniak.mimuw.edu.pl - eksploracja danych. http://wazniak.mimuw.edu.pl/images/3/3d/ED-4.2-m01-1.0.pdf.

[3] Uci machine learning repository - wine quality data set. https://archive.ics.uci.edu/ml/datasets/wine+quality.

[4] Weka 3.8.4. https://www.cs.waikato.ac.nz/ml/weka/.