Robert Kozakiewicz Gdańsk, 30.05.2015

Dawid Piasecki

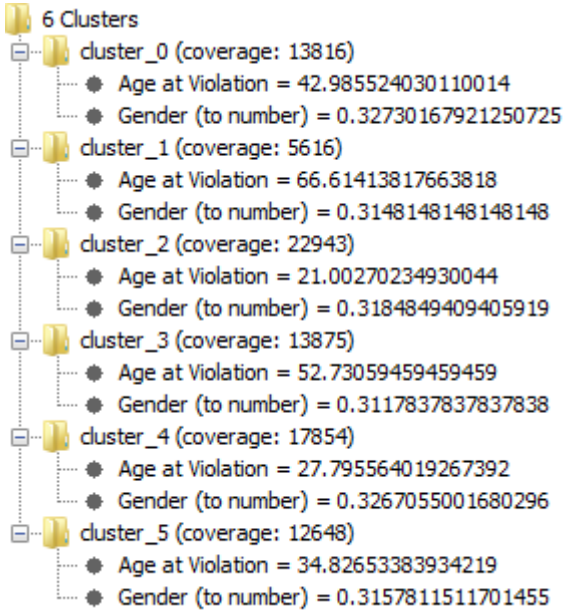
Eksploracja danych

Zrozumienie warunków biznesowych i identyfikacja źródeł danych

Cele biznesowe eksploracji danych.

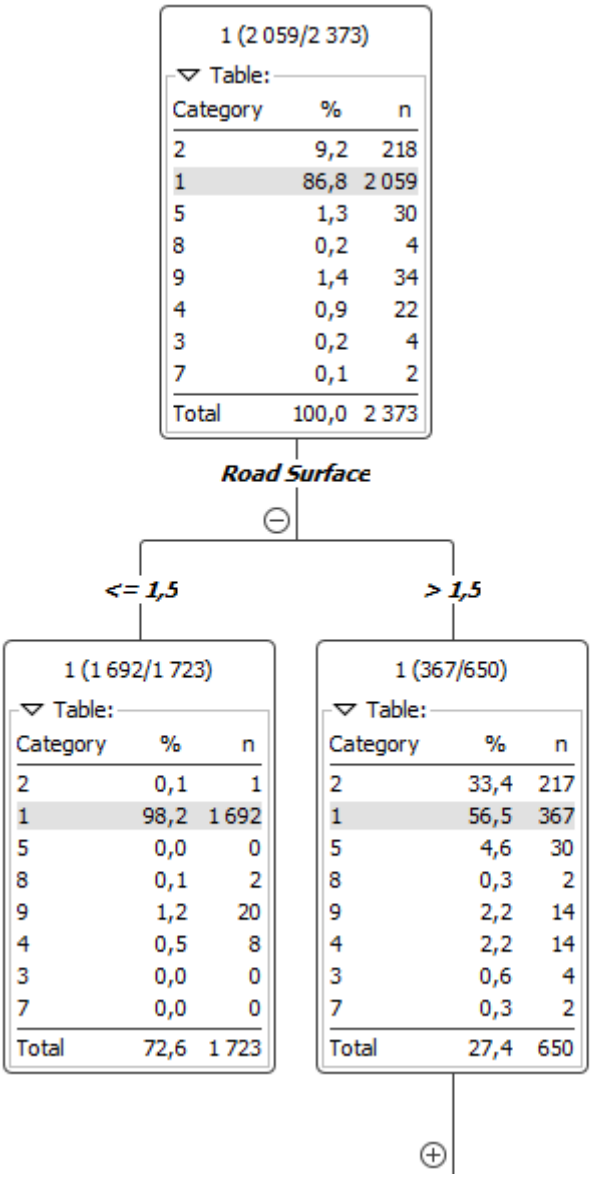
a) Jaka jest charakterystyka kierowcy, który uczestniczy w wypadku?

Do tego celu użyliśmy danych o mandatach, interesujące nas kolumny to 'Gender' oraz 'Age at Violation'. Wartość 0 oznacza płeć męską, natomiast 1 płeć żeńską. Użyliśmy algorytmów do klasteryzacji k-Means.



b) Jaka jest charakterystyka warunków pogodowych podczas wypadków?

W rozwiązaniu tego celu wykorzystaliśmy dane ze źródła o wypadkach. Kolumny "Road Surface" i "Weather Conditions". Wybrany model to drzewo decyzyjne zależności warunków pogodowych od powierzchni. Atrybutem nominalnym uczyniliśmy tylko jeden z dwóch a mianowicie "Weather Conditions", po to by można było uzależnić warunki pogodowe od kilku rodzajów dróg.



Road Surface:
1 - Dry
2 - Wet / Damp
3 - Snow
4 - Frost / Ice
5 - Flood (surface water over 3cm deep)

Weather Conditions:
1 - Fine without high winds
2 - Raining without high winds
3 - Snowing without high winds
4 - Fine with high winds
5 - Raining with high winds
6 - Snowing with high winds
7 - Fog or mist – if hazard
8 - Other

9 - Unknown

c) Jaka jest charakterystyka pojazdu, który uczestniczy w wypadku?

W tym problemie posłużyliśmy się źródłem danych o wypadkach. Interesująca nas kolumna to 'Type of Vehicle'. Przyjmuje ona następujące wartości:

1 - Pedal cycle

2 - M/cycle 50cc and under

3 - Motorcycle over 50cc and up to 125cc

4 - Motorcycle over 125cc and up to 500cc

5 - Motorcycle over 500cc

8 - Taxi/Private hire car

9 - Car

10 - Minibus (8 – 16 passenger seats)

11 - Bus or coach (17 or more passenger seats)

14 - Other motor vehicle

15 - Other non-motor vehicle

16 - Ridden horse

17 - Agricultural vehicle (includes diggers etc.)

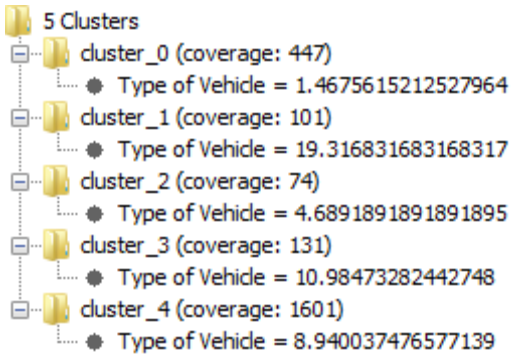
18 - Tram / Light rail

19 - Goods vehicle 3.5 tonnes mgw and under

20 - Goods vehicle over 3.5 tonnes and under 7.5 tonnes mgw

21 - Goods vehicle 7.5 tonnes mgw and over

Do wyznaczenia charakterystyk użyliśmy algorytmu do klasteryzacji k-Means.



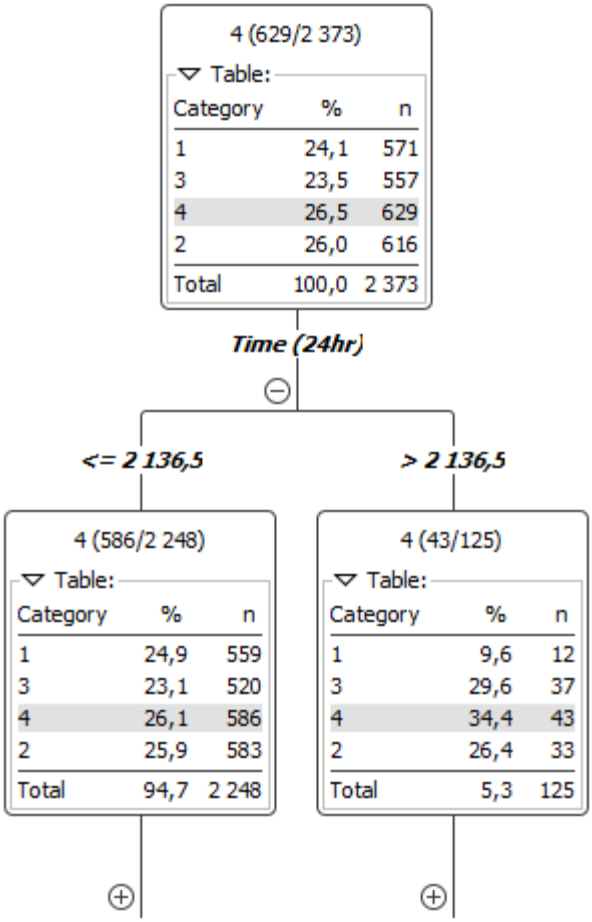
Utworzone klastry wyznaczają charakterystyki pojazdów uczestniczących w wypadkach.

d) Jak pora dnia i data w roku wpływa na ilość wypadków?

Danę, które przyczyniły się do rozwiązania tego problemu to: kwartał roku, otrzymany za pomocą "Date Field Extractor" z kolumny "Accident Date" oraz godzina w której miał miejsce wypadek z kolumny "Time (24hr)" ze źródła o wypadkach. Wybrany model to drzewo decyzyjne zależności kwartału od godziny.

Godzina w formacie hhmm, gdzie minuty to koniecznie 2 cyfry np. 06, a godzina może być tylko jedną np. 152 oznacza 1:52.

Kwartały od 1-4.



e) Jak charakterystyka drogi(rodzaj drogi i oświetlenie) wpływa na ilość wypadków?

Interesujące nas dane znajdują się w źródle z wypadkami, a konkretnie w kolumnach '1st Road Class' oraz 'Lighting Conditions'. Możliwe wartości atrybutów to:

1st Road Class:

1 - Motorway

2 - A(M)

3 - A

4 - B

5 - C

6 - Unclassified

Lighting Conditions:

1 - Daylight: street lights present

2 - Daylight: no street lighting

3 - Daylight: street lighting unknown

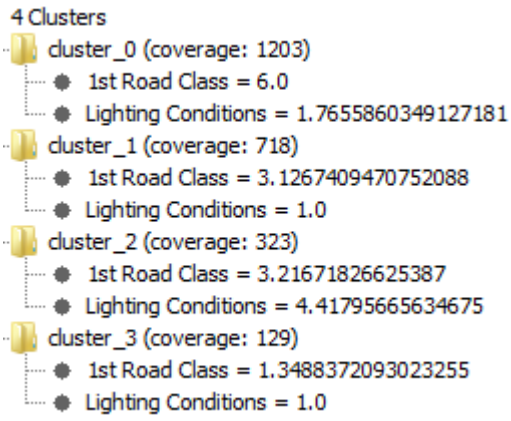
4 - Darkness: street lights present and lit

5 - Darkness: street lights present but unlit

6 - Darkness: no street lighting

7 - Darkness: street lighting unknown

Użyliśmy algorytmów do klasteryzacji k-Means oraz EM.




```
EM
==

Number of clusters selected by cross validation: 4
Number of iterations performed: 3


                    Cluster
Attribute                0      1      2      3
                     (0.14) (0.36) (0.38) (0.12)
=================================================
1stRoadClass
  mean                5.9976 2.8557 5.9995 2.8984
  std. dev.           0.0698 0.7324 0.0307 0.6152

LightingConditions
  mean                4.2262      1      1 4.2003
  std. dev.           0.7142 1.4621 1.4621 0.6339

Clustered Instances

0       327 ( 14%)
1       852 ( 36%)
2       909 ( 38%)
3       285 ( 12%)


Log likelihood: -1.71862
```