Lista 2 - Konwersja danych

Podstawowym narzędziem, dla którego wykonana zostanie analiza danych jest środowisko *Weka* do pobrania ze strony `www.cs.waikato.ac.nz/ml/weka/`. Narzędzie to operuje na plikach danych o rozszerzeniu *.arff* (zobacz również *Weka-wprowadzenie*). Przykładowe zestawy danych znajdują się w katalogu *data* w miejscu w którym została zainstalowana *Weka*. Należy otworzyć zbiór danych *weather.arff* wykorzystując tryb edycji pliku w narzędziu *Notatnik*, bądź z wykorzystaniem pochodnego edytora tekstu (*Notepad++*, *TextMate*).

W ramach pliku z danymi wyróżnia się sekcję *attributes* gdzie definiowane są atrybuty (cechy) występujących obiektów, oraz sekcję *data*, gdzie w każdym wierszu pojawiają się rekordy (obiekty) opisane konkretnymi wartościami cech. Atrybuty mogą przyjmować wartości różne rodzaje wartości, przy czym najczęściej rozpatruje się wartości rzeczywiste (*real*), bądź też nominalne (*nominal*). Przykładowo, w zbiorze *weather.arff* atrybutami nominalnymi są *outlook*, *windy*, oraz *play*, natomiast atrybutami rzeczywistymi *temperature*, oraz *humidity*. Wartości atrybutów w sekcji *data* są podawane zgodnie z kolejnością atrybutów zadaną w sekcji *attributes*.

Do przeglądania i analizy danych wykorzystuje się GUI jakie dostarcza *Weka* w wersji aplikacyjnej (zobacz również *Weka-wprowadzenie*). Poprzez Wybranie modułu *Explorer*, a następne opcji *Open File*. W przypadku, gdy zbiór danych jest poprawnie zdefiniowany w formacie *.arff* dane zostaną poprawnie wczytane do narzędzia. Interfejs graficzny daje możliwość przeglądania statystyk dotyczących każdego z atrybutów. W ramach zakładki *Preprocess* możliwe jest wykorzystanie filtrów służących do obróbki danych.

Zadania

1. Z wykorzystaniem narzędzia *Notatnik* (bądź pochodnego edytora tekstu: *Notepad++*, *TextMate*) należy skonstruować zbiór danych składający się z 4 atrybutów, oraz 5 rekordów. Należy zdefiniować 2 atrybuty rzeczywiste (kwotę kredytu oraz wiek) oraz dwa atrybuty nominalne(płeć, przyjmującą wartości K, oraz M, oraz status decyzji kredytowej, przyjmujący wartości pozytywny, oraz negatywny). Średnia kwota kredytu dla 5 klientów powinna wynosić 500 zł, średni wiek klientów o negatywnym statusie powinien wynosić 35 lat, natomiast liczba klientów o pozytywnym statusie powinna wynosić 3. Skonstruowany plik z danymi zapisz pod nazwą *XXXXXXL2_1.arff*. Zbadaj poprawność zbudowanego zbioru danych poprzez wczytanie go do interfejsu graficznego *Weki* - *2 pkt*.

2. Wykorzystując plik *XXXXXXL1_2.xls* skonstruowany podczas prac nad poprzednią listą zbuduj odpowiadający mu plik danych w formacie *.arff*. Skonstruowany plik z danymi zapisz pod nazwą *XXXXXXL2_2.arff*. Zbadaj poprawność zbudowanego zbioru danych poprzez wczytanie go do interfejsu graficznego *Weki*. Opisz w w punktach sposób przejścia z jednego formatu danych do drugiego. - *3 pkt* (zgodność utworzonego pliku z wersją *xls*), *3 pkt* (subiektywna ocena sposobu przejścia z jednego formatu danych do drugiego).