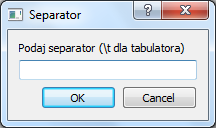
|  |  |
| --- | --- |
| Wydział Informatyki  Systemy Wspomagania Decyzji | Data: 02.04.2012 |
| **Klasyfikacja obiektów za pomocą grupowania.**  **Warpechowska Anna**  **Nowak Krzysztof** | Prowadzący:  Mgr inż. Paweł Zabielski |

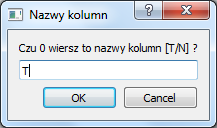
Zadanie 1

Napisać w dowolnym języku i środowisku programowania narzędzie pozwalające na wczytanie danych z pliku z uwzględnieniem

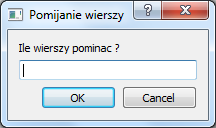
* różnych separatorów (kolejne wartości dzielą: spacje, tabulatory, ‘;’, itd.)



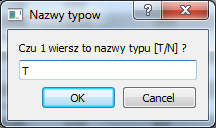
* nazw zmiennych w pierwszym wierszu (tak/nie);



* pominięcie pierwszych n linii



* typów kolejnych zmiennych (liczba, ciąg znaków, data itd.



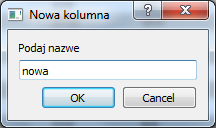
Dodatkowo:

Tworzenie zbioru danych z możliwością:

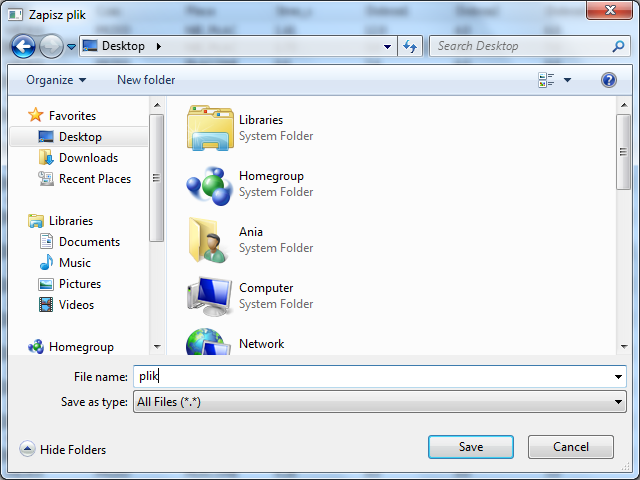
* określenia liczby zmiennych
* określenia liczby obserwacji

Program umożliwia utworzenie nowego, pustego zbioru oraz dynamicznie dodawanie zmiennych / obserwacji.

* dodawania/usuwania zmiennych/obserwacji



* zapisu zbioru danych

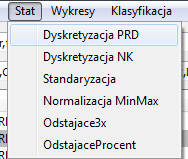


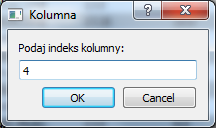
Zadanie2

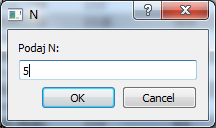
Umożliwienie wykonania na wczytanym/utworzonym zbiorze danych następujących operacji:

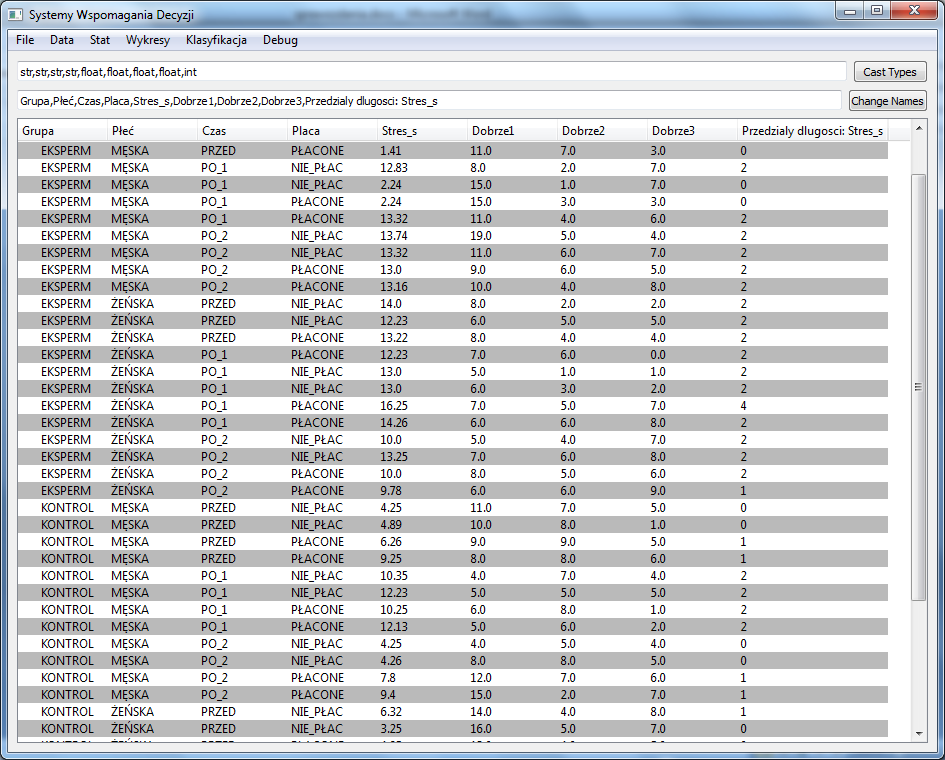
a) Dyskretyzacja

* przedziały równej długości - dodanie do zbioru danych nowego atrybutu o wartościach nominalnych, przyjmującego dla poszczególnych obserwacji wartość odpowiadającą numerowi przedziału przypisanego wartości wybranego atrybutu a, przy podziale wartości atrybutu a na zadaną liczbę n przedziałów o równej długości.

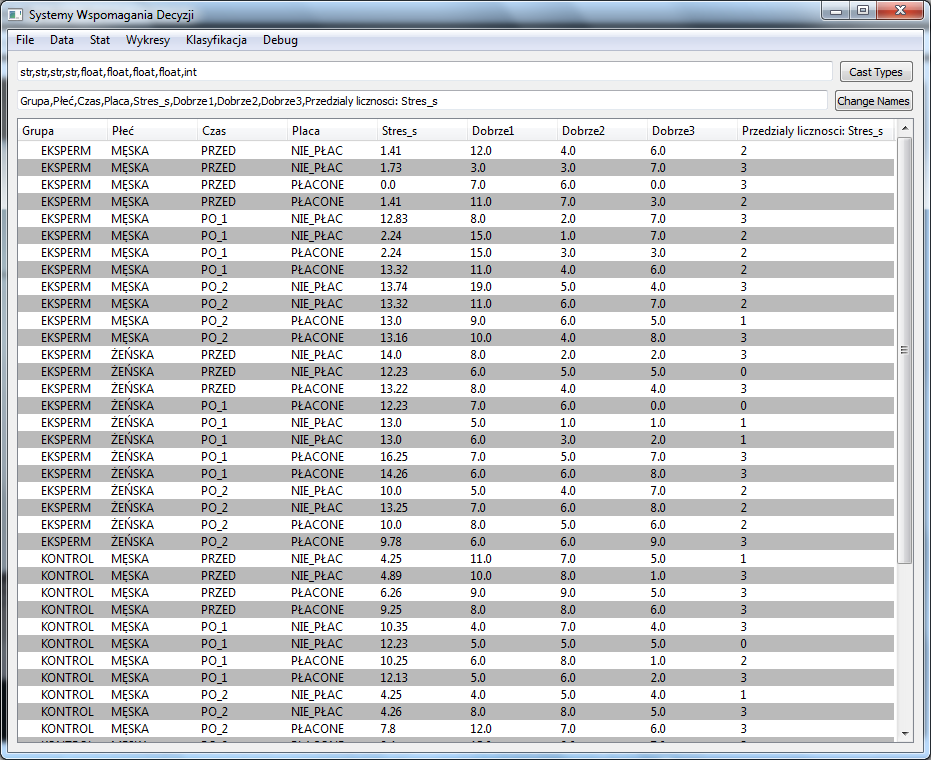






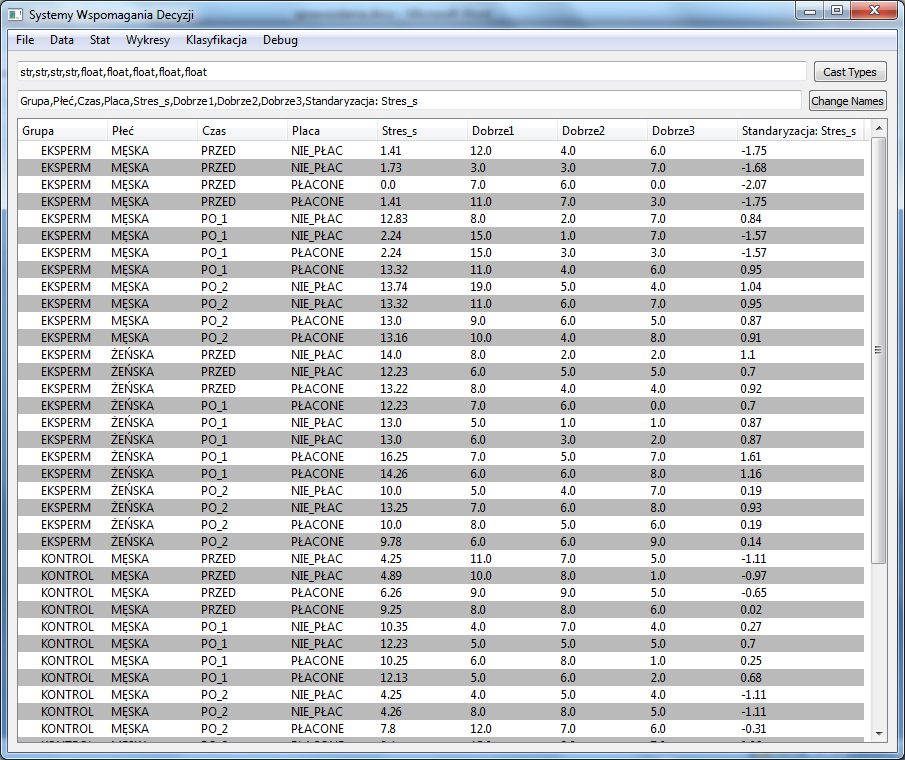


* preferowanie najliczniejszych klas - stosowane do atrybutów o wartościach nominalnych, dodanie do zbioru danych nowego atrybutu o wartościach nominalnych przypisującego najliczniej reprezentowanej wartości - wartość 1, drugiej pod względem liczności wartości - wartość 2, itd. aż do zadanej liczby n, pozostałym wartościom - wartość n+1



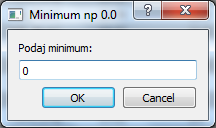
b) Standaryzacja

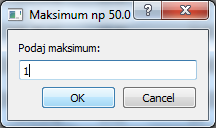
* dodanie nowego atrybutu przyjmującego dla poszczególnych obserwacji znormalizowaną wartość wybranego atrybutu

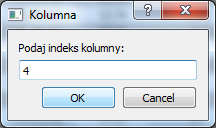


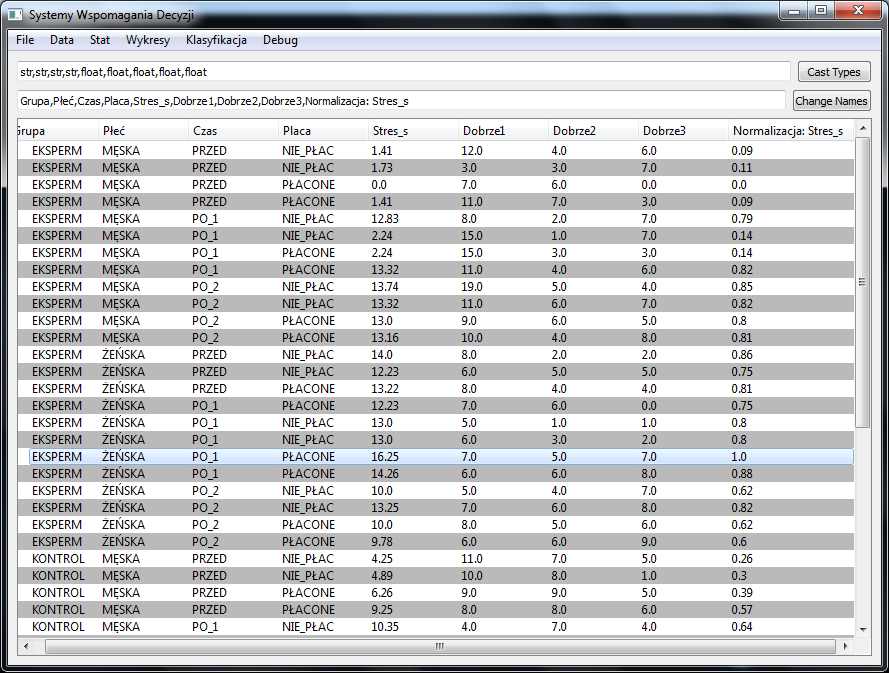
c) Normalizacja min-max

* dodanie nowego atrybutu przyjmującego dla poszczególnych obserwacji wartość wybranego atrybutu zrzutowaną do przedziału o zadanych wartościach minimum i maksimum



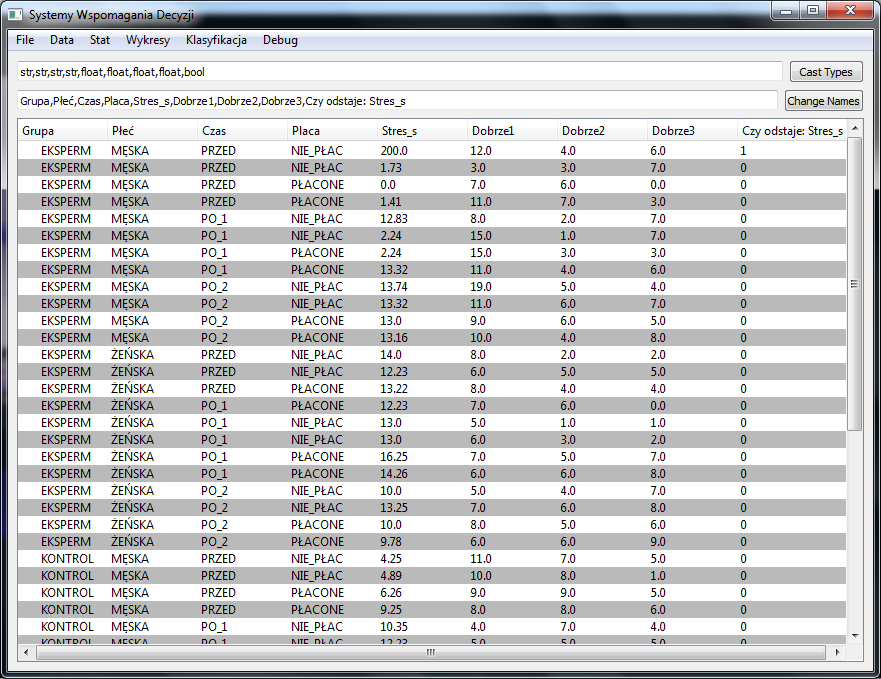




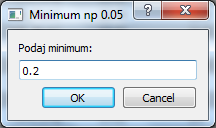
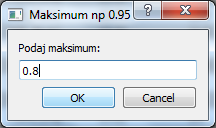


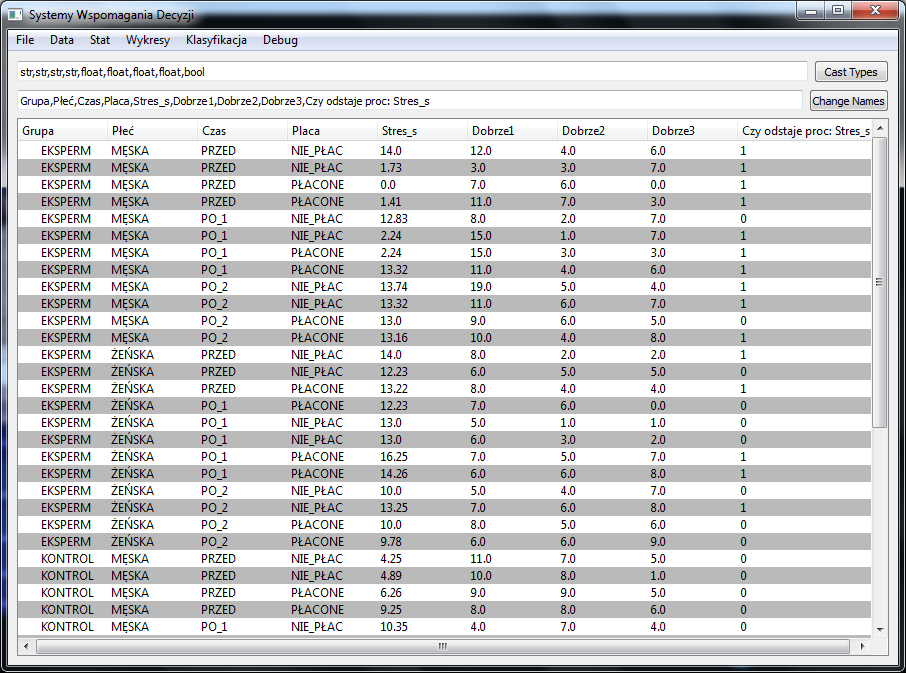
d) Wykrywanie obserwacji odstających

* 3 x odchylenie standardowe - oznaczenie jako obserwacji odstającej tej, której odległość od średniej jest większa niż trzykrotność odchylenia standardowego (wyliczenia na podstawie wybranego atrybutu)



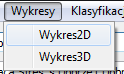
* określony procent najmniejszych i największych wartości - oznaczenie jako obserwacji odstającej tej, której wartość wybranego atrybutu należy do ustalonego procenta (np. 5%) najmniejszych lub największych wartości tego atrybutu dla całego zbioru danych

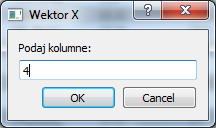
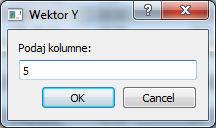
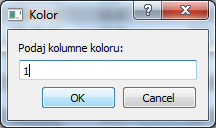
 

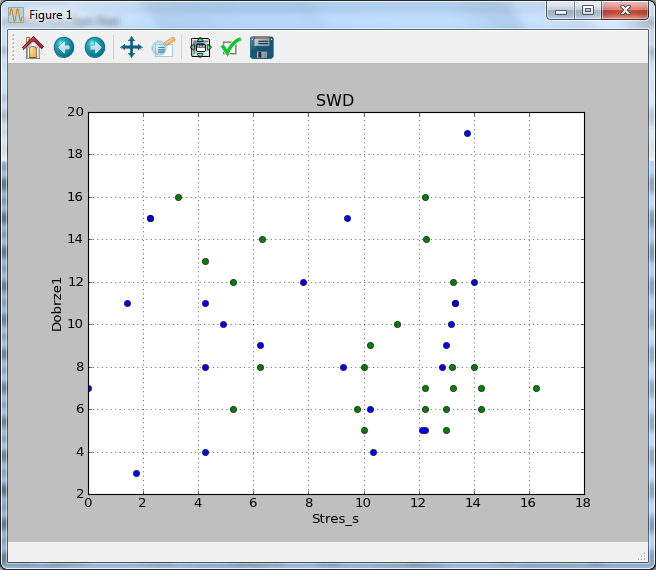


e) Wykresy rozproszenia

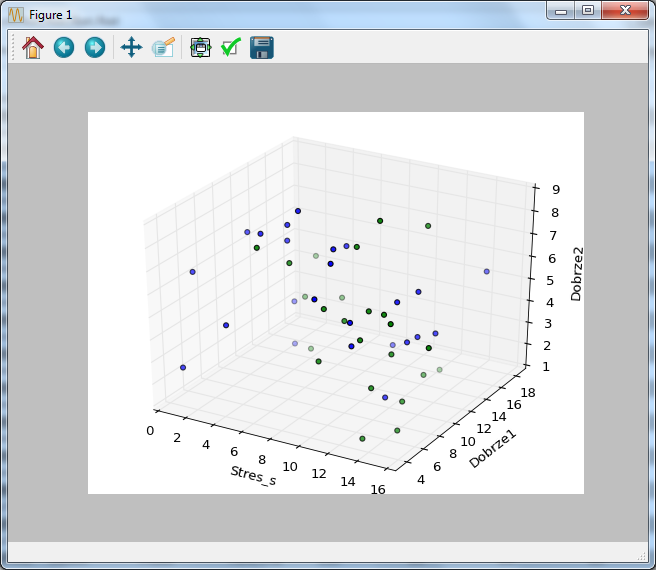
* 2D z możliwością wyświetlania klas w kolorach





* 3D (dla chętnych)

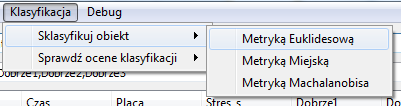


Zadanie3

Klasyfikacja metodą k-nn (k najbliższych sąsiadów – K Nearest Neighbours). Napisać narzędzie pozwalające na ocenę jakości klasyfikacji metodą k-nn oraz klasyfikację nowego obiektu tą metodą.

a) Klasyfikacja

Zadajemy nowy obiekt bez przypisanej klasy decyzyjnej, określamy ilość sąsiadów k i metrykę w ocenie odległości obiektów. Zadaniem narzędzia powinno być przydzielenie nowego obiektu do klasy decyzyjnej, do której należy najwięcej spośród jego k najbliższych sąsiadów wyznaczonych zgodnie z określoną metryką.



b) Ocena jakości klasyfikacji

Określamy jedynie ilość sąsiadów k i metrykę w ocenie odległości obiektów. Oceny jakości klasyfikacji dokonujemy metodą leave-one-out. Jako jakość klasyfikacji rozumiemy iloraz poprawnie zaklasyfikowanych obiektów do ilości wszystkich obiektów w zbiorze wejściowym.

Metryki oceny odległości obiektów:

* odległość Euklidesowa
* odległość Mahalanobisa
* metryka Manhattan