1. Treść zadania

✓ Zadanie1

a) 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.)

b) <u>Dodatkowo:</u>

Tworzenie zbioru danych z możliwością:
• określenia liczby zmiennych
• określenia liczby 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)