Algorytmy analizy danych
Kierunek Informatyka, WE, sem. VII
Laboratorium nr 1

Grzegorz Cichosz

1. Wstępna ocena zbioru danych – plik iris.data.wn

a. Zapoznaj się z opisem zawartym w pliku iris.name.txt oraz z zawartością pliku iris.data.wn
Zbiór danych Iris Plants Database zawiera pomiary cech morfologicznych kwiatów irysa oraz przypisaną klasę określającą gatunek rośliny

b. Jaki jest ich charakter wczytanych danych (porządkowe, nominalne,…)?
Pierwsze cztery atrybuty mają charakter ilościowy ciągły, natomiast ostatni parametr ma charakter nominalny określający klasę (Gatunek Irysa)

c. Jaki jest wymiar danych (ile jest zmiennych zależnych i niezależnych)?
Zbiór danych zawiera 150 instancji
Występują 4 zmienne niezależne oraz 1 zmienna zależna
Zmienna zależna przyjmuje 3 wartości: Iris Setosa, Iris Versicolour, Iris Virginica.
Każda z klas zawiera po 50 instancji

d. osobny plik excel

e. Wylicz podstawowe wielkości statystyczne (suma, średnia, odchylenie standardowe, max, min, dominanta, liczebność) dla poszczególnych zmiennych niezależnych.
Co można stwierdzić w oparciu o te dane opisu statystycznego?

suma	876.5
średnia	5.843333333
odchylenie standardowe	0.828066128
min	4.3
max	7.9
dominanta	5
liczebność	150

Dla każdej ze zmiennych niezależnych obliczono podstawowe miary statystyczne
Na podstawie uzyskanych wyników można stwierdzić, że największym zróżnicowaniem charakteryzują się cechy związane z płatkiem kwiatu (petal length oraz petal width), co potwierdzają wyższe wartości odchylenia standardowego.

f. Dokonaj oceny danych pod kątem brakujących wartości.
Czy takie brakujące dane są w zestawie danych? Jeśli tak, to jak wyeliminujesz wartości brakujące? Opisz jakie rozwiązanie zastosowałeś celem wyeliminowania wartości brakujących.

sepal_length blank	0
sepal_width blank	0
petal_length blank	0
petal_width blank	0

Użyłem formuł
=COUNTBLANK(A2:A151)
=COUNTBLANK(B2:B151)
=COUNTBLANK(C2:C151)
=COUNTBLANK(D2:D151)

g. Wykonaj wykresy rozrzutu punktów „każdy z każdym z wymiarów” (zmiennej niezależnej) - pamiętaj o odpowiednim typie wykresu
Utwórz z otrzymanych wykresów tzw. macierz wykresów rozrzutu punktów.
Jaka jest twoja ocena tych danych (ich jakości) w oparciu o uzyskane wykresy?
Czy dane wykazują skorelowanie?
Na podstawie macierzy wykresów rozrzutu można stwierdzić, że dane charakteryzują się dobrą jakością. Nie występują wyraźne obserwacje odstające ani nielogiczne zależności pomiędzy zmiennymi
Widoczna jest silna zależność pomiędzy zmiennymi związanymi z płatkiem kwiatu, szczególnie pomiędzy petal length oraz petal width.

Wartości współczynników korelacji zostały obliczone przy użyciu funkcji CORREL w programie LibreOffice Calc
Najwyższe wartości korelacji występują pomiędzy zmiennymi petal length oraz petal width, co wskazuje na silną dodatnią korelację.

h. Wykonaj wykresy rozrzutu punktów „każdy z każdym z wymiarów”, ale dla każdej ze zmiennych zależnych osobno.
Utwórz macierze wykresów rozrzutu punktów.
Oblicz wartości współczynnika korelacji dla poszczególnych par zmiennych niezależnych, lecz niezależnie dla
poszczególnych kategorii.
Jak teraz oceniasz dane pod kątem ich jakości, ale również wpływu ich wartości na zmienną zależną?

Współczynniki korelacji zostały obliczone osobno dla każdej kategorii zmiennej zależnej z wykorzystaniem funkcji CORREL programu LibreOffice Calc
Analiza wykazała, że zależności pomiędzy zmiennymi różnią się w zależności od klasy, przy czym najsilniejsze korelacje dotyczą cech płatków kwiatu

i. Wykonaj histogramy (wykresy) każdej ze zmiennych niezależnych z osobna, biorąc wszystkie wartości.
Skorzystaj ze standardowych ustawień dotyczących wskazania przedziałów dla częstości (czyli nie wskazuj
zakresu zbioru) w narzędziu Dane>Analiza danych>Histogram
Wykonaj też histogramy (wykresy) każdej ze zmiennych niezależnych z osobna, lecz również niezależnie dla
każdej z kategorii
Co możesz powiedzieć o tych danych w oparciu o zbudowane histogramy?
Czy nie dostrzegasz czegoś nietypowego, np. danych odstających?
Wykonaj też wykresy liniowe dla każdej ze zmiennych niezależnych oraz każdej z kategorii z osobna.
Czy dane Twoim zdaniem wymagają dodatkowej interwencji pod kątem jakościowym? Jeśli tak, to jakiej?

j. Dokonaj teraz skalowania cech
Skalowanie wykonaj dla dwóch przypadków: normalizując poszczególne zmienne niezależne oraz w oparciu
o standaryzację zmiennych niezależnych
Wykonaj wykresy po znormalizowaniu oraz po standaryzacji każdej ze zmiennych niezależnych z osobna