# Data mining - zadanie 44

Gabriel Budziński

January 20, 2024

## 1 Treść

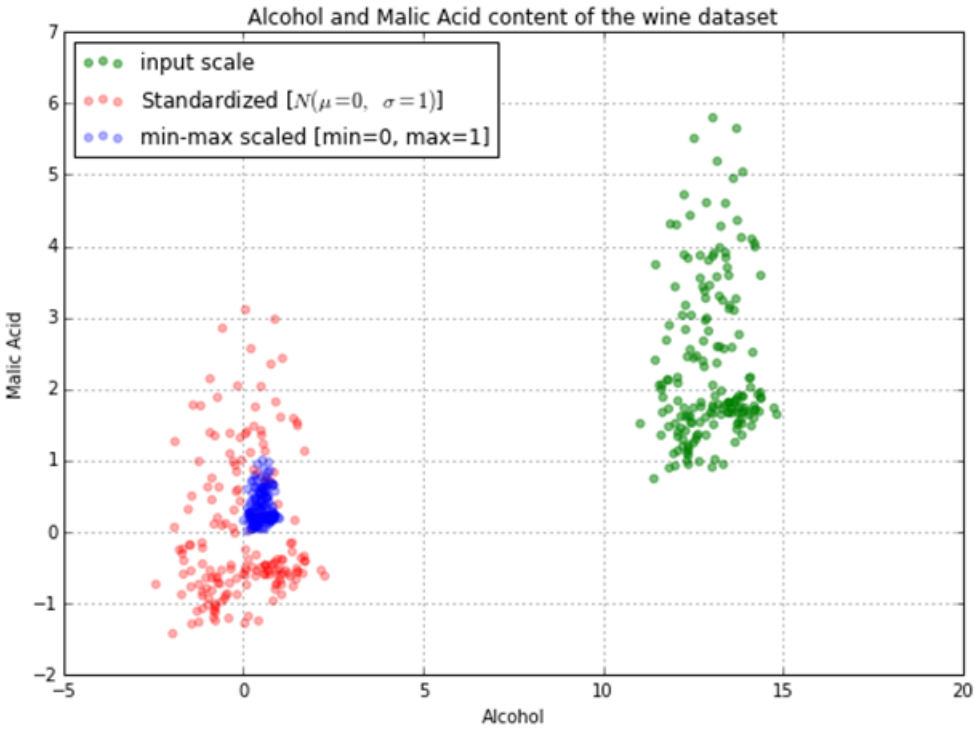
Explain why feature scaling of the input features is important.

## 2 Skalowanie

Cechy w danych które dostarczamy do naszej sieci mogą mieć różne zakresy, jednostki czy rzędy wielkości. W takich wypadkach model niektóre z cech mogą zdominować inne, które zostaną pominiętem, co negatywnie wpłynie na dokładność naszego modelu.

## 3 Metody skalowania

Najczęstszymi metodami skalowania są normalizacja, standaryzacja i min-max scaling.




### 3.1 Standaryzacja

$$x' = \frac{x - \bar{x}}{\sigma}$$

W ten sposób otrzymujemy zbiór ze średnią $\mu = 0$ oraz odchyleniem standardowym $\sigma = 1$.

### 3.2 Normalizacja

$$x' = \frac{x - \bar{x}}{\max(x) - \min(x)}$$

W ten sposób otrzymujemy wartości $x' \in [-1, 1]$ ze średnią $\mu = 0$.

### 3.3 Min-max

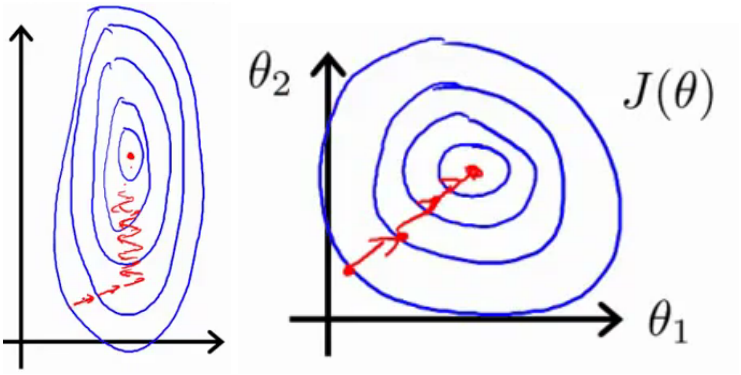
$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

W ten sposób otrzymujemy wartości $x' \in [0, 1]$.

## 4 Skalowanie a trening

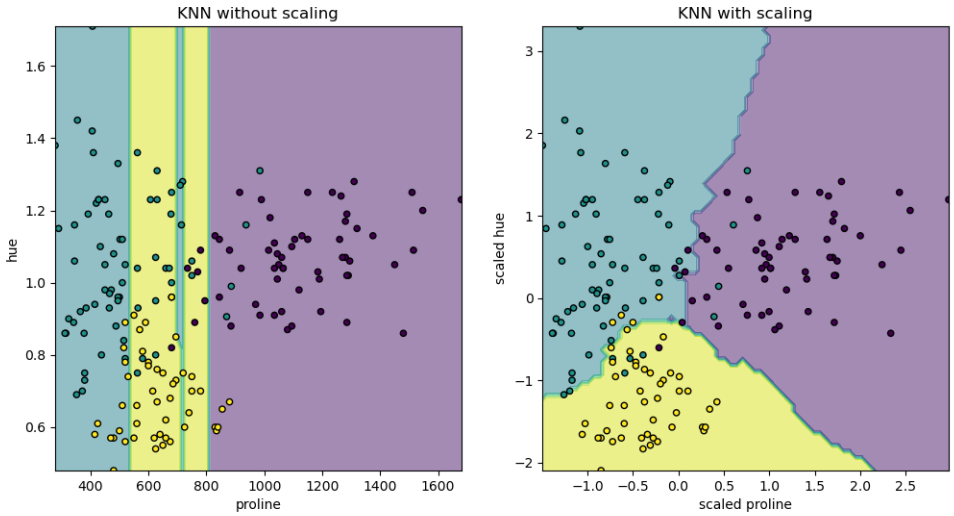
Skalowanie zbioru danych ma też pozytywny wpływ na szybkość uczenia lub dokładność modelu w zależności od metody.

### 4.1 Gradient decent




Jeśli dane są rozciągniete w jednym z wymiarów (ma duże odchylenie standardowe), gradient może długo skakać na drodze do optimum. Kiedy przeskalujemy te dane droga do rozwiązania optymalnego jest łatwiejsza.

### 4.2 KNN



Jak widzimy dane oryginalne oraz przeskalowane prowadzą do powstania zupełnie różnych modeli.

## 4.3 Modele oparte o drzewa

W modelach opartych o drzewa wpływ sklaowania jest znikomy.