

## DRZEWA DECYZYJNE – wstęp teoretyczny

**Modele drzew klasyfikacyjnych i regresyjnych** (CART, ang. *Classification and Regression Trees*), jak sama nazwa mówi, służą zarówno do rozwiązywania problemów regresyjnych (gdzie zmienną zależną jest cecha ilościowa – ciągła/liczbowa) jak i klasyfikacyjnych (zmienna zależna jakościowa – kategoriowa). Najogólniej, celem analizy z zastosowaniem algorytmu budowy drzew decyzyjnych jest znalezienie zbioru logicznych warunków podziału, typu *jeżeli, to*, prowadzących do jednoznacznego zaklasyfikowania obiektów.

**Drzewa decyzyjne** służą do wyboru deskryptorów o największym wpływie na modelowaną wielkość (najbardziej znaczących). Technika ta polega na „wzrastaniu drzewa” tj. dzielenia związków na wzajemnie wykluczające się grupy – węzły (ang. *nodes*). Linie łączące węzły nazywa się gałęziami (ang. *branches*). Algorytm rozpoczyna się od węzła głównego – korzenia (ang. *root*) – zawierającego wszystkie związki, które następnie dzielone są na węzły podrzędne. Końcowe węzły, które nie podlegają podziałom to liście (ang. *leaves*). Każdy podział określa reguła (próg) uwzględniająca wartości wybranego na danym etapie deskryptora.

Zarówno w przypadku klasycznych modeli jakościowych (SAR, ang. *Structure-Activity Relationships*), jak również modeli ilościowych (QSAR, ang. *Quantitative Structure-Activity Relationships*) związki dzielone są na dwa zbiory – uczący (wykorzystywany do opracowania drzewa decyzyjnego) oraz walidacyjny (służący do oceny zdolności predykcyjnych drzewa decyzyjnego).

W przypadku **drzew klasyfikacyjnych** deskryptory wybierane są pod kątem najmniejszego prawdopodobieństwa błędnej klasyfikacji, co oznacza, że binarny podział wykonywany z opracowaną regułą powinien prowadzić do maksymalnie dwóch jednorodnych grup związków. Prawdopodobieństwo błędnej klasyfikacji mierzy się za pomocą indexu Giniego, wyrażonego wzorem:

$$G = 1 - \sum_{j=1}^c \left(\frac{n_j}{n}\right)^2$$

gdzie  $n_j$  jest liczbą związków z klasy  $j$  zawartych w węźle.

**Do weryfikacji zdolności predykcyjnych modeli jakościowych służą miary statystyczne:**

$$\textbf{Sensitivity}(\textit{recall}, \textit{positiverate}) = TP/(TP + FN)$$

$$\textbf{Specificity} = TN/(FP + TN)$$

$$\textbf{Precision} = TP/(TP + FP)$$

$$\textbf{F1} \quad (\textit{harmonic mean of precision\&sensitivity}) = (2 \times TP)/(2 \times TP + FP + FN)$$

$$\textbf{Balanced accuracy} = (\textit{Sensitivity} + \textit{Specificity})/2$$

$$\text{Balanced error} = 1 - \text{Balanced accuracy}$$

		Predicted	
		Active	Inactive
Observed	Active	True positive (TP)	False positive (FP)
	Inactive	False negative (FN)	True negative (TN)

**Figure 1.** Confusion matrix describing the performance of a classification model (or 'classifier') on a set of test data for which the true values are known.

Wybór deskryptorów w przypadku **drzew regresyjnych** dokonywany jest przy pomocy metody najmniejszych kwadratów, czyli tak aby suma kwadratów różnic pomiędzy wartościami przewidywanymi przez model a zmierzonymi eksperymentalnie (tzw. rezyduałów) była jak najmniejsza.

**ZADANIE 1.** Dane wejściowe – **ftalany\_klasyfikacja.xlsx** zawiera dane dotyczące 32 ftalanów. Dane są już po autoskalowaniu. Podział na zbiór treningowy i walidacyjny znajdują się w poszczególnych arkuszach. Związki są podzielone na dwie kategorie: 1 (trwałość w przedziale dni-tygodnie) oraz 2 (tygodnie-miesiące).

1. Przygotuj macierz korelacji pomiędzy zmiennymi niezależnymi i zmienną zależną.
2. Sprawdź czy zbiór testowy i treningowy są zbalansowane.
3. Zbuduj model drzewa klasyfikacyjnego w celu przewidywania tego parametru (kategorii trwałości). Oceń zdolności prognostyczne modelu na podstawie macierzy pomyłek oraz statystyk: czułości, specyficzności, precyzji, współczynnika F1, dokładności oraz zbalansowanego błędu.
4. Dokonaj optymalizacji parametrów dwiema metodami (z uzasadnieniem ich wyboru)

**ZADANIE 2.** Dane wejściowe – **ftalany.xlsx** zawiera dane dotyczące 32 ftalanów. Dane są już po autoskalowaniu. Podział na zbiór treningowy i walidacyjny znajdują się w poszczególnych arkuszach.

1. Zbuduj model drzewa regresyjnego, aby przewidzieć stałą szybkości degradacji poszczególnych związków.
2. Oblicz statystyki  $R^2$ , RMSE,  $Q^2$  i RMSEex. Narysuj wykres zależności  $y_{pred}$  od  $y_{obs}$ , oraz wykres słupkowy zależności  $y_{pred}$  od  $y_{obs}$ .
3. Dokonaj optymalizacji parametrów dwiema metodami (z uzasadnieniem ich wyboru)
4. Dokonaj interpretacji uzyskanych wyników