UCZENIE MASZYNOWE – wstęp teoretyczny

Uczenie maszynowe (ML, ang. *Machine Learning*) jest gałęzią sztucznej inteligencji, której ideą jest umożliwienie maszynom wykonywania swoich zadań przy użyciu inteligentnych programów. Istnieje wiele definicji wyjaśniających czym jest uczenie maszynowe. Jedna z nich mówi, że: „Program komputerowy (maszyna) uczy się na podstawie doświadczenia E w odniesieniu do pewnej klasy zadań T i miary efektywności P, jeśli jego efektywność wykonywania zadania T (mierzona za pomocą P) poprawia się wraz z doświadczeniem T". Wyróżnia się trzy główne kategorie algorytmów uczenia maszynowego: przez wzmacnianie (ang. *Reinforcement Learning/Deep Learning)*, nienadzorowane (ang. *Unsupervised Learning*) oraz nadzorowane (ang. *Supervised Learning*).

Metoda **uczenia przez wzmocnienie** ma zastosowanie w sytuacjach, gdzie brak jest zarówno określonych parametrów wejściowych, jak i wyjściowych. Jej działanie opiera się na wykorzystywaniu obserwacji zbieranych na podstawie interakcji ze środowiskiem. Po każdej akcji, jedyną informacją zwrotną otrzymywaną przez maszynę jest tzw. sygnał wzmocnienia (pozytywny lub negatywny). Celem algorytmu jest podejmowanie działań w taki sposób, aby zmaksymalizować poprawność rezultatów. Do przykładów tej techniki należą Q-Learning oraz Monte Carlo.

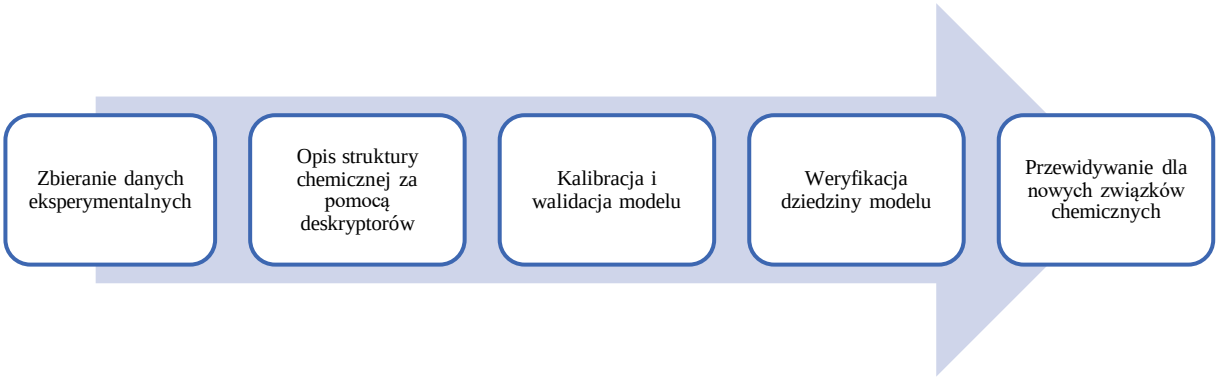
Uczenie nienadzorowane polega na dostarczeniu maszynie do nauki nieoznaczonego zbioru uczącego składającego się jedynie z cech/atrybutów wejściowych (wektor $\vec{x}$) i niezawierającego etykiet. Algorytm samodoskonali się, wykorzystując duże, zróżnicowane zbiory danych. System uczy się modelu (funkcji h), którego zadaniem jest opisanie danych wyjściowych (np. wyszukanie prototypów, rozpoznawanie skupień). Przykładami tej kategorii są grupowanie metodą k-średnich i sieci neuronowe.

W **uczeniu nadzorowanym** maszyna otrzymuje instrukcję, czego ma się nauczyć i w jaki sposób. Zestaw danych uczących zawiera zbiór cech/atrybutów wejściowych (wektor $\vec{x}$) oraz poprawnie zaetykietowaną odpowiedź – wartość wyjściową ($f(\vec{x})$). System uczy się modelu (funkcji h) mającego jak najlepiej aproksymować funkcję f w celu poprawnej predykcji etykiet przykładów, z którymi nie miał jeszcze styczności. Algorytmy uczenia nadzorowanego rozwiązują problemy regresyjne i klasyfikacyjne. Należą do nich m.in. regresja liniowa, regresja głównych składowych, regresja częściowych najmniejszych kwadratów, maszyna wektorów nośnych oraz drzewa decyzyjne. W uczeniu nienadzorowanym system szuka wzorców z bieżącego przykładu, natomiast w nadzorowanym korzysta z przykładów podanych wcześniej.

Przykładowym obszarem chemii komputerowej, w którym wykorzystywane są algorytmy uczenia nadzorowanego jest **QSAR/QSPR** (ang. *Quantitive Activity/Structure-Property Relationships*) – ilościowe modelowanie zależności pomiędzy strukturą chemiczną ($\boldsymbol{X}$) a aktywnością biologiczną/właściwością fizykochemiczną ($\boldsymbol{y}$). Podstawą QSAR/QSPR jest założenie, że różnice w aktywności/właściwościach substancji chemicznych wynikają z różnic w ich budowie. W oparciu o zbiór obliczonych deskryptorów struktury, numerycznie

wyrażających zmienność budowy chemicznej, oraz odpowiedni model matematyczny można interpolować brakujące informacje na temat związków, dzięki wartościom eksperymentalnym dostępnym dla podobnych do nich strukturalnie cząsteczek. Ogólne równanie QSAR/QSPR wyrażone jest poniższym wzorem. Podstawowe kroki modelowania QSAR/QSPR przedstawia **Rysunek 1.**

$$\boldsymbol{y} = f(\boldsymbol{X})$$



Rysunek 1. Prosty schemat modelowania QSAR/QSPR.

Dopełnieniem modelowania QSAR/QSPR są różnorodne metody chemoinformatycznej, wielowymiarowej analizy danych (analiza głównych składowych, metoda najmniejszych kwadratów) oraz inne algorytmy sztucznej inteligencji np. algorytm genetyczny. Połączenie wymienionych technik umożliwia modelowanie skomplikowanych zależności, tym samym stając się przyszłością badań naukowych nad powiązaniem aktywności/właściwości ze strukturą związków chemicznych oraz nadzieją na rozszerzenie obecnego stanu wiedzy.

Przygotowanie zbioru danych

Zgodnie z dobrą praktyką chemometryczną wyniki pomiarów do analizy problemów wieloparametrycznych (zależnych od dużej liczy zmiennych) powinny być gromadzone w odpowiednio zaplanowanych tabelach arkusza kalkulacyjnego. W przypadku modelowania QSAR należy skonstruować wektor $\boldsymbol{y}$ (modelowania wielkość) oraz macierz $\boldsymbol{X}$ (deskryptory). Dane zostały zorganizowane w oparciu o wzór macierzy przedstawiony na **Rysunku 2.**

Lek	Modelowana wielkość	Deskryptor 1	Deskryptor 2	Deskryptor 3	…	Deskryptor m
Lek *i*	y_i	X_{1i}	X_{2i}	X_{3i}	…	X_{mi}
Lek *ii*	y_{ii}	X_{1ii}	X_{2ii}	X_{3ii}	…	X_{mii}
Lek *iii*	y_{iii}	X_{1iii}	X_{2iii}	X	…	X_{miii}
…	…	…	…	…	…	…
Lek *n*	y_n	X_{1n}	X_{2n}	X_{3n}	…	X_{mn}

Rysunek 2. Wzór macierzy do modelowania QSAR.

Modelowanie QSAR wymaga podzielenia związków zbadanych eksperymentalnie na zbiór uczący (kalibracyjny), który wykorzystuje się do zbudowania modelu i jego walidacji wewnętrznej, oraz zbiór testowy (walidacyjny) służący do potwierdzenia zdolności prognostycznych modelu w procesie walidacji zewnętrznej. Najczęściej stosowanym algorytmem podziału jest 1:X polegający na posortowaniu wartości odpowiedzi (malejąco/rosnąco) i przypisaniu co x-tego obiektu do zbioru testowego, co zapewnia równomierny rozkład związków w zakresie modelowanej wielkości.

Autoskalowanie (standaryzacja) polega na transformacji zmiennych, aby wartość średnia każdej z nich była równa 0, a odchylenie standardowe równe jedności. W taki sposób wszystkie wymiary wielowymiarowej przestrzeni cech stają się współmierne. Autoskalowanie łączy w sobie centrowanie zmiennych względem początku układu współrzędnych oraz skalowanie przedziałowe, na skutek którego wyeliminowany zostaje wpływ jednostek użytych do pomiaru zmiennych. Wykonuje się je według wzoru:

$$z_{ij} = \frac{x_{ij} - x_j}{s_j}$$

gdzie: z_{ij} – standaryzowana wartość cechy,
x_{ij} – początkowa wartość zmiennej,
x_j – wartość średnia j-tej zmiennej,
s_j – odchylenie standardowe j-tej zmiennej.

METODA REGRESJI WIELOKROTNEJ (MLR)

Regresja wielokrotna (MLR, ang. *Multiple Linear Regression*) to jedno z najprostszych, a zarazem najważniejszych narzędzi statystycznych, gdzie zmienna zależna (modelowana wielkość) wyrażana jest jako kombinacja liniowa zmiennych niezależnych (deskryptorów). W metodzie tej zakłada się model w postaci:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

gdzie: y – zmienna zależna (modelowana wielkość),
b_0 – wyraz wolny,
b_n – współczynniki modelu,
x_n – zmienne objaśniające (deskryptory).

Współczynniki dobierane są za pomocą metody najmniejszych kwadratów, czyli tak, aby suma kwadratów różnić pomiędzy wartościami przewidywanymi przez model a zmierzonymi eksperymentalnie (tzw. rezyduałów) była jak najmniejsza. Wzór na współczynniki modelu regresyjnego przedstawia się następująco:

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Po obliczeniu współczynników możliwe jest otrzymanie wektora $\boldsymbol{y}$ zgodnie z równaniem:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

gdzie $\boldsymbol{H}$ jest macierzą dźwigni wiążącą obliczone oraz eksperymentalne odpowiedzi modelu. Współczynniki dźwigni h są przydatne do sprawdzenia strukturalnej dziedziny modelu.

Istotnym ograniczeniem MLR w modelowaniu QSAR jest to, że gdy pomiędzy zmiennymi występują silne korelacje nie jest możliwe poprawne odwrócenie macierzy $(\boldsymbol{X}^T\boldsymbol{X})$, a więc wzór nie może zostać użyty do obliczenia współczynników $\boldsymbol{b}$. Występowanie takich korelacji w przypadku wielowymiarowych metod QSAR jest bardzo powszechne (niektóre deskryptory WHIM wyjaśniają tę samą cechę według różnych schematów ważenia, a więc są prawie całkowicie skorelowane), dlatego konieczne jest skorzystanie z innej metody np. PCR (ang. *Principal Component Regression*) lub dokonanie eliminacji jednego z pary skorelowanych zmiennych – zwykle o wyższej korelacji z innymi deskryptorami.

Miarę jakości dopasowania modelu (ang. *goodness-of-fit*) stanowią współczynnik determinacji (R^2, ang. *determination coefficient*) oraz średni kwadratowy błąd kalibracji ($RMSE$, $RMSE_C$, ang. *Root Mean Sqare error of calibration*) obliczane według wzorów:

$$R^2 = 1 - \frac{\sum\left(y_{pred} - y_{obs}\right)^2}{\sum(y_{obs} - \bar{y}_{obs})^2}$$

$$RMSE_C = \sqrt{\frac{\sum\left(y_{pred} - y_{obs}\right)^2}{n_C}}$$

gdzie:

y_{pred} – przewidywana wartość zmiennej zależnej,

y_{obs} – obserwowana (eksperymentalna) wartość zmiennej zależnej,

n_C – liczebność zbioru kalibracyjnego.

Model QSAR uznaje się za dobrze dopasowany, gdy $R^2 > 0{,}65$ ($R^2 > 0{,}9$ w przypadku modelowania QSPR). R^2 zwiększa się wraz z dodawaniem kolejnych deskryptorów niezależnie od tego, czy redukują one niewyjaśnioną wariancję zmiennej zależnej.

Oceny statystycznej istotności przyjętego modelu dokonuje się z wykorzystaniem testu F-Snedecora – istotny, gdy informacja wyjaśniana przez model jest większa niż zawarta w jego błędach (im większa wartość tej statystyki, tym większa istotność modelu).

Dziedzina modelu

Dziedzina modelu (AD, ang. *Applicability Domain*) opisuje granice teoretycznego obszaru w przestrzeni cech związków, w którym przewidywania są wiarygodne. Przestrzeń tę wyznacza się poprzez zestawienie na osi odciętych współczynników dźwigni h, a na osi rzędnych wartości standaryzowanych rezyduałów (wykres Williamsa). Współczynniki dźwigni określają strukturalne podobieństwo między związkiem a zbiorem uczącym. Wartość graniczna dźwigni jest wyrażana wzorem:

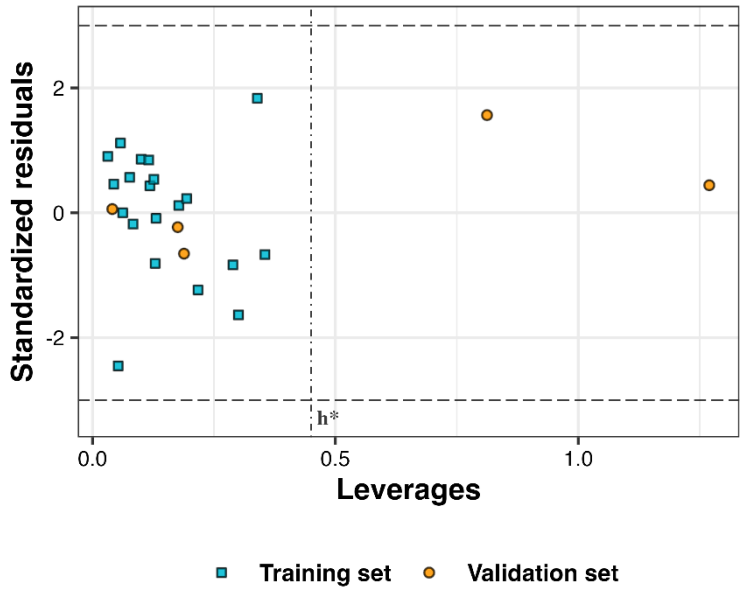
$$h^* = \frac{3p}{n}$$

gdzie: p – liczba deskryptorów w modelu,

n – liczba związków w zbiorze uczącym.

Związki o wartościach $h > h*$ rozpatrywane są jako leżące poza AD. Oś rzędnych przedstawia precyzję przewidywania przez model, w związku z czym wartości odchylenia

standardowe. Zdefiniowanie dziedziny modelu pozwala na sporządzenie mapy powierzchni błędu modelu w zależności od użytych deskryptorów oraz ocenę jego jakości.



Rysunek 3. Przykładowy wykres Williamsa.

Walidacja modelu

Elastyczność modelu QSAR (ang. *robustness*) to jego odporność na usunięcie jednego lub większej liczby związków ze zbioru uczącego. Weryfikuje się ją za pomocą walidacji krzyżowej (CV, ang. *cross-validation*). Zdolności prognostyczne modelu (ang. *predictive capability*) ocenia się przeprowadzając walidację zewnętrzną (ang. *external validation*).

Walidację krzyżową można określić mianem „walidacji wewnętrznej" (ang. *internal validation*), ponieważ wykorzystuje zbiór kalibracyjny również jako walidacyjny (niejednocześnie). Najprostszym typem walidacji krzyżowej jest „wyrzuć jeden" (LOO, ang. *leave-one-out*), który polega na wyłączaniu pojedynczych obiektów ze zbioru uczącego – stanowią one wówczas jednoelementowe zbiory walidacyjne. Na podstawie modelu skalibrowanego dla pozostałych $n-1$ elementów przewiduje się wartość modelowanej wielkości dla wyrzuconego związku. Procedurę powtarza się n razy (kolejno dla każdego elementu), czego wynikiem jest n-elementowy zbiór wartości wyjść z modelu. W efekcie otrzymuje się zestaw n różnic pomiędzy prognozami modelu i odpowiedziami obiektów, które wykorzystywane są do wyznaczania współczynnika walidacji krzyżowej (Q^2_{CV}, ang. *cross-validation determination coefficient*) oraz średniego kwadratowego błędu walidacji krzyżowej ($RMSE_{CV}$, ang. *root mean square error of cross-validation*) zgodnie z równaniami:

$$Q^2_{CV} = 1 - \frac{\sum(y_{cv} - y_{obs})^2}{\sum(y_{obs} - \bar{y}_{obs})^2}$$

$$RMSE_{CV} = \sqrt{\frac{\sum(y_{cv} - y_{obs})^2}{n_C}}$$

gdzie:

y_{cv} – wartość zmiennej zależnej otrzymana z zastosowaniem walidacji krzyżowej,
y_{obs} – obserwowana (eksperymentalna) wartość zmiennej zależnej,
n_C – liczebność zbioru kalibracyjnego.

Istnieją również inne odmiany walidacji krzyżowej takie jak „wyrzuć więcej" (LMO, ang. *leave-more-out*), gdzie kolejno wyłącza się po kilka obiektów ze zbioru uczącego, a także metoda randomizowanej walidacji krzyżowej, w której elementy wyłączanych jednorazowo bloków wybierane są losowo.

W przeciwieństwie do walidacji krzyżowej, walidacja zewnętrzna dostarcza informacji dotyczących jakości przewidywania modelu posługując się zbiorem testowym. Miarą zdolności prognostycznych modelu QSAR są wartości współczynnika walidacji zewnętrznej (Q^2_{Ex}, *external validation coefficient*) oraz średniego kwadratowego błędu przewidywania ($RMSE_{Ex}$, ang. *root mean square error of prediction*).

$$Q^2_{Ex} = 1 - \frac{\sum\left(y_{pred} - y_{obs}\right)^2}{\sum\left(y_{pred} - \bar{y}_{pred}\right)^2}$$

$$RMSE_{Ex} = \sqrt{\frac{\sum\left(y_{pred} - y_{obs}\right)^2}{n_T}}$$

gdzie:

y_{pred} – przewidywana wartość zmiennej zależnej,
y_{obs} – obserwowana wartość zmiennej zależnej,
n_T – liczebność zbioru walidacyjnego.

Dobry model QSAR powinien charakteryzować się możliwie bliskimi jedności wartościami R^2, Q^2_{CV}, Q^2_{Ex} oraz porównywalnymi i możliwie małymi wartościami średnich błędów kwadratowych. Występowanie znacznych różnic pomiędzy wartościami błędów $RMSE_C$, $RMSE_{CV}$ i $RMSE_{Ex}$ wskazuje na zbyt duże podobieństwo związków należących do zbioru kalibracyjnego (ang. *overfiting*), a tym samym na małą zdolność modelu do generalizowania informacji.

ZADANIE 1. Cytotoksyczność nanocząstek tlenków metali względem komórek ludzkiej linii komórkowej keratynocytów (zmienna zależna) została wyrażona jako liniowa kombinacja dwóch deskryptorów (zmiennych niezależnych): (i) entalpii tworzenia nanoklastra MeOx (ΔHfc), reprezentującego fragment powierzchni, oraz (ii) elektroujemności Mullikena (X^C) obliczonej dla całego klastra. Na podstawie poniższego równania proszę wyznaczyć wartości przewidywanej cytotoksyczności, następnie obliczyć współczynnik determinacji, średni kwadratowy błąd kalibracji oraz sprawdzić istotność statystyczną modelu. Proszę pamiętać o przeprowadzeniu autoskalowania danych.

$$log(EC_{50})^{-1} = 2{,}466 + 0{,}244\,\Delta Hfc + 0{,}394\,X^C$$

Metal oxide	**ΔHfc** [kcal/mol]	**X^C** [eV]	**Obs. log(EC_{50})$^{-1}$** [molar]
TiO_2	-1492.00	4.9	**1.76**
ZnO_2	-638.10	4.95	**2.02**
SiO_2	-618.30	3.81	**2.12**
V_2O_3	-139.50	3.24	**2.24**
Sb_2O_3	-206.70	4.46	**2.31**
Bi_2O_3	-148.50	5.34	**2.50**
Mn_2O_3	-96.30	5.00	**2.64**
CoO	-786.80	7.44	**2.83**
In_2O_3	-52.10	6.78	**2.92**
ZnO	-449.40	8.33	**3.32**

ZADANIE 2. Na podstawie wstępu teoretycznego, poniższych informacji oraz wiedzy własnej proszę zbudować model „krok po kroku” z uwzględnieniem:

(i) sprawdzenia korelacji pomiędzy zmiennymi objaśniającymi (macierz korelacji),
(ii) równania modelu,
(iii) wykreślenia ypred od yobs z podziałem na zbiór uczący i walidacyjny (legenda),
(iv) sprawdzenia dziedziny modelu (wykres Williamsa z zaznaczoną wartością graniczną, podział na zbiór uczący i walidacyjny),
(v) obliczenia statystyk: $R^2, RMSE_C, Q^2_{CVloo}, RMSE_{CVloo}, Q^2_{Ex}, RMSE_{Ex}, F$.

Powyższe podpunkty należy zawrzeć w sprawozdaniu wraz z kodem. Proszę krótko zinterpretować uzyskane wyniki.

„dane_leki.xlsx” – dane wejściowe (autoskalowane!!!):

WEKTOR ***y***.

W wektorze ***y*** umieściłam zaczerpnięte z literatury naukowej wyniki pomiarów laboratoryjnych parametru **logK HSA** – stała równowagowa tworzenia się kompleksu w roztworze; **miara siły interakcji między reagentami**; wyraża powinowactwo leku do albuminy surowicy człowieka.

Albumina (HSA, ang. *human serum albumin*) to jedno z dwóch głównych białek osocza odpowiedzialnych za wiązanie leków. Białka osocza są odpowiedzialne za utrzymanie równowagi kwasowo-zasadowej, prawidłowego ciśnienia osmotycznego oraz transport substancji nierozpuszczalnych w wodzie (takich jak endogenne hormony sterydowe lub kwasy tłuszczowe). Czynniki wpływające na stopień wiązania leku z białkami: stężenie leku, powinowactwo leku do białek, oddziaływania cząsteczki leku z „kieszeniami białek". Lek związany z białkami jest nieaktywny farmakologiczne, nie przenika przez błony biologiczne i nie ulega metabolizmowi, więc zmniejszenie stopnia wiązania leku z białkami osocza skutkuje wzrostem siły działania i skróceniem czasu działania leku. Tylko wolna frakcja leku może przenikać przez błony biologiczne. Przykłady leków o dużym stopniu wiązania z białkami: pochodne kumaryny, fenylobutazon, salicylany, sulfonamidy, NLPZ, penicyliny. Powinowactwo do białka osocza jest jedną z najważniejszych właściwości biologicznych, które należy wziąć pod uwagę podczas projektowania i oceny przyszłych potencjalnych leków.

MACIERZ ***X***.

logK CTAB – retencja (czas retencji = czas uwalniania) w fazie pseudostacjonarnej CTAB (bromek heksadecylotrimetyloamoniowy) z wykorzystaniem metody micelarnej chromatografii elektrokinetycznej (MEKC); micele utworzone w CTAB mają strukturę podobną do HSA.

Deskryptory CATS – dostarczają dodatkowych informacji o strukturze cząsteczki oraz mogą dostarczyć użytecznych informacji odzwierciedlających zachowanie leku w regionie wiążącym HSA; kodują informację o częstościach par atomów, które mogą być potencjalnymi miejscami wiązania leku; kryterium wyboru: wartość statystyki F.

CATS3D_09_AL – łączy informacje o lipofilowości i akceptorze wiązań wodorowych.

CATS3D_00_AA i CATS 3D_00_DD – ważone tylko przez dawcę wiązania wodorowego (D), siłę akceptora (A). Wpływ wiązania wodorowego jest dostrzegany jako jeden z krytycznych czynników determinujących interakcję między miejscem II HSA, a niektórymi typami ligandów – małymi, zwykle aromatycznymi kwasami karboksylowymi.

Zbiór uczący – 19 związków, zbiór walidacyjny – 8 związków (70:30 %, wybrane losowo).

WZORY.

Autoskalowanie

$$z_{ij} = \frac{x_{ij} - x_j}{s_j}$$

gdzie: z_{ij} – standaryzowana wartość cechy,

x_{ij} – początkowa wartość zmiennej,

x_j – wartość średnia j-tej zmiennej,

s_j – odchylenie standardowe j-tej zmiennej ($s_j = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$).

Jakość dopasowania modelu

$$R^2 = 1 - \frac{\sum\left(y_{pred} - y_{obs}\right)^2}{\sum(y_{obs} - \bar{y}_{obs})^2}$$

$$RMSE_C = \sqrt{\frac{\sum\left(y_{pred} - y_{obs}\right)^2}{n_C}}$$

gdzie:

y_{pred} – przewidywana wartość zmiennej zależnej,

y_{obs} – obserwowana (eksperymentalna) wartość zmiennej zależnej,

n_C – liczebność zbioru kalibracyjnego.

Elastyczność modelu

$$Q^2_{CV} = 1 - \frac{\sum(y_{cv} - y_{obs})^2}{\sum(y_{obs} - \bar{y}_{obs})^2}$$

$$RMSE_{CV} = \sqrt{\frac{\sum(y_{cv} - y_{obs})^2}{n_C}}$$

gdzie:

y_{cv} – wartość zmiennej zależnej otrzymana z zastosowaniem walidacji krzyżowej,

y_{obs} – obserwowana (eksperymentalna) wartość zmiennej zależnej,

n_C – liczebność zbioru kalibracyjnego.

Jakość przewidywania modelu

$$Q^2_{Ex} = 1 - \frac{\sum\left(y_{pred} - y_{obs}\right)^2}{\sum\left(y_{pred} - \bar{y}_{pred}\right)^2}$$

$$RMSE_{Ex} = \sqrt{\frac{\sum\left(y_{pred} - y_{obs}\right)^2}{n_T}}$$

gdzie:

y_{pred} – przewidywana wartość zmiennej zależnej,
y_{obs} – obserwowana wartość zmiennej zależnej,
n_T – liczebność zbioru walidacyjnego.

Ocena statystycznej istotności modelu – test F-Snedecora

$$F = \frac{S_M^2}{S_E^2}$$

gdzie:
S_M^2– wariancja modelu,
S_E^2 – wariancja resztowa.

Wariancja modelu

$$S_M^2 = \frac{\sum\left(y_{pred} - \bar{y}_{pred}\right)^2}{n-1}$$

Wariancja resztowa

$$S_E^2 = \frac{\sum\left(y_{obs} - y_{pred}\right)^2}{n-p-1}$$

gdzie:
n– liczebność całego zbioru obiektów,
p – liczba deskryptorów.