Wstęp

Wielokrotna regresja liniowa (MLR), znana również po prostu jako regresja wielokrotna, jest techniką statystyczną, która wykorzystuje kilka zmiennych objaśniających do przewidywania wyniku zmiennej odpowiedzi. Celem wielokrotnej regresji liniowej jest modelowanie liniowej zależności pomiędzy zmiennymi objaśniającymi (niezależnymi) a zmiennymi odpowiedzi (zależnymi). W istocie regresja wielokrotna jest rozszerzeniem zwykłej regresji najmniejszych kwadratów (OLS), ponieważ obejmuje więcej niż jedną zmienną objaśniającą. Modelowanie QSAR wymaga podzielenia związków zbadanych eksperymentalnie na zbiór uczący (kalibracyjny), który wykorzystuje się do zbudowania modelu i jego walidacji wewnętrznej, oraz zbir testowy (walidacyjny) służący do potwierdzenia zdolności prognostycznych modelu w procesie walidacji zewnętrznej.

Do stworzenia modelu wykorzystałam dane zawierające informacje nt. leków, które zostały uprzednio poddane procesowi autoskalowania.

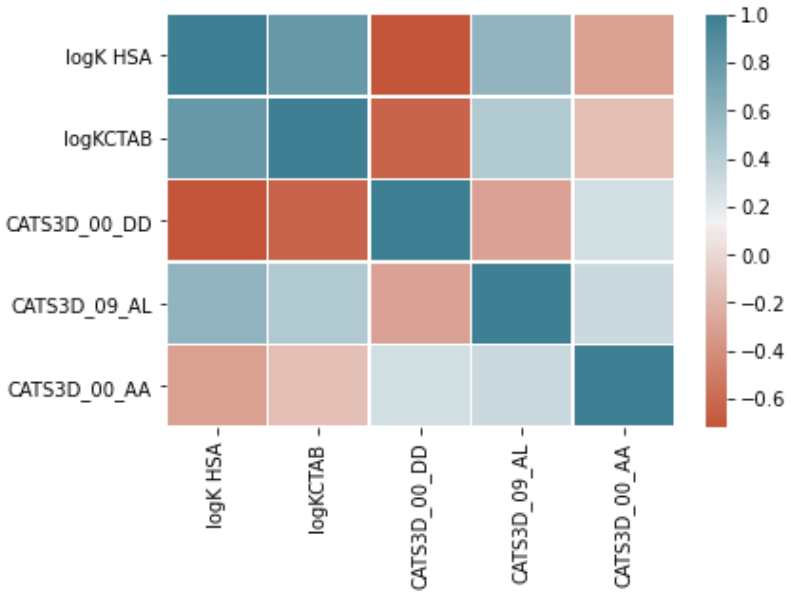
Zmienne uwzględnione w modelu:

1. Zmienna zależna
    a. parametr 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.

2. Zmienne niezależne:
    a. 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.
    b. 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
    c. CATS3D_09_AL – łączy informacje o lipofilowości i akceptorze wiązań wodorowych.
    d. 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.

1. Wykres korelacji.



Rysunek 1. Wykres korelacji między zmiennymi

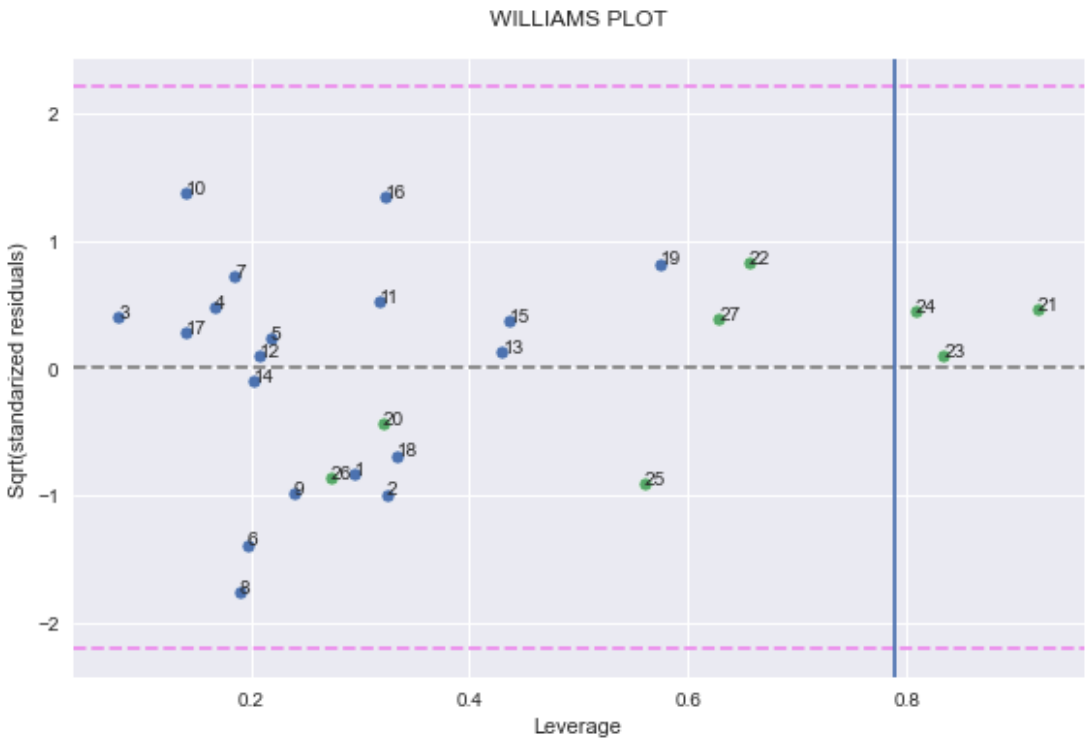
2. **Równanie modelu:**

y_pred = 0.8050 + 0.4722 * logKCTAB + (-0.2161) * CATS3D_00_DD + 0.3232 * CATS3D_09_AL + (0.2253) * CATS3D_00_AA

3. **Obliczone wartości statystyk:**

R^2 = 0.803
RMSEc = 0.427
F = 14.23
Q^2_{EX} = 0.949
$RMSE_{EX}$ = 0.301

4. **Wykres Williamsa z zaznaczoną wartością graniczną + interpretacja:**



Rysunek 2. Wykres Williamsa z zaznaczoną wartością dźwigni na poziomie 0.79.

1. acetaminophen
2. acetylsalicylic acid
3. bromazepam
4. carbamazepine
5. chlorpromazine
6. clonidine
7. diazepam
8. diltiazem
9. diphenhydramine
10. fluoxetine
11. flurbiprofen
12. imipramine
13. ketoconazole
14. ketoprofen
15. metronidazole
16. nizatidine
17. propranolol
18. ranitidine
19. trazodone
20. acyclovir
21. diclofenac
22. famotidine
23. ibuprofen
24. indomethacin
25. methylprednisolone
26. quinidine
27. zidovudine

1-19 – zestaw treningowy
20 – 27 – zestaw walidacyjny

Trzy z dwudziestu siedmiu leków zawartych w zestawie danych wykracza poza krytyczny współczynnik dźwigni „h". Wszystkie z nich należą do zestawu walidacyjnego (testowego). Ibuprofen, indomethacin i diclofenac to wartości odstające. Dla wszystkich związków w zbiorach treningowych i testowych ich standaryzowane reszty są mniejsze niż dwie i pół jednostki odchylenia standardowego.

5. Interpretacja:

R^2 (współczynnik determinacji) świadczy o zadowalającym dopasowaniu modelu do danych (im większy tym prosta regresji jest lepiej dopasowana do danych).

RMSEc (średni kwadratowy błąd kalibracji, miara jakości dopasowania modelu.) – im mniejsza wartość tym model lepiej dopasowany.

Q^2_{EX} (współczynnik walidacji zewnętrznej) – im wyższa wartość (bliższa jedności) tym lepszy model.

$RMSE_{EX}$ (średni kwadratowy błąd przewidywania) - lepszy model będzie miał mniejszy współczynnik $RMSE_{EX}$.

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.

RMSEc oraz $RMSE_{EX}$ nie różnią się zbytnio, podobieństwo związków należących do zbioru kalibracyjnego nie jest zbyt duże. R^2 oraz Q^2_{EX} są bliskie jedności oraz nie różnią się pomiędzy sobą o więcej niż 0.3, co także dobrze świadczy o modelu.