**Projekt z R**

Rozważ dane: **train\_d.csv** oraz **test\_d.csv**

Dane dotyczą różnych cząsteczek chemicznych. Pierwsze dwie kolumny to odpowiednio nazwa oraz informacja o strukturze (w formacie SMILES). Pozostałe: MolWt to masa cząsteczkowa, LogP opisuje hydrofobowość (logarytm współczynnika podziału oktanol/woda), HDonors to liczba donorów wiązań wodorowych, HAcceptors – liczba akceptorów wiązań wodorowych, TPSA to powierzchnia polarna cząsteczki, RotBonds – liczba rotowalnych wiązań, AromRings – liczba pierścieni aromatycznych, HeteroAtoms – liczba heteroatomów (innych niż C i H), HeavyAtoms – liczba atomów ciężkich (wszystkich poza wodorem), a FracCSP3 to ułamek atomów węgla w hybrydyzacji sp³ (miara nasycenia cząsteczki). Ostatnia kolumna mpC przedstawia temperaturę topnienia (w stopniach Celcjusza).

Celem projektu jest zamodelowanie zmiennej mpC w oparciu o pozostałe/dostępne informację zawarte w pliku **train\_d.csv.** Projekt powinien mieć formę kilku stronnicowego sprawozdania (pdf) z załączonym kodem do wykonania kolejnych kroków.

**Co powinno się znaleźć?**

1. Przedstawienie danych i problemu.

2. Charakterystyka numeryczna i graficzna danych (np. wykresy rozkładu zmiennych, podstawowe statystyki).

3. Badanie korelacji pomiędzy zmiennymi (uwaga: nie wszystkie korelacje mogą być sensowne).

4. Budowa różnych modeli (w tym ewentualne stosowanie transformacji nieliniowych). Wskazanie zmiennych istotnych i nieistotnych w rozważanych modelach.

5. Diagnostyka modeli (wizualna i/lub za pomocą testów).

6. Odkrywanie obserwacji odstających i wpływowych.

7. Droga dojścia do najlepszego modelu (jakie stosujemy kryterium? R2, AIC, F test lub inne).

8. Interpretacja modeli i wyników (także pośrednich).

9. Krótkie podsumowanie.

Dla najlepszego modelu narysuj wykres (wartości mpC) vs (przewidywania modelu). Dodatkowo, proszę o wyznaczenie Adjusted R2 oraz RMSD na zbiorze treningowym (**train\_d.csv)** oraz RMSD na zbiorze testowym (**test\_d.csv**) dla każdego ze zbudowanych modeli. Wyniki należy zebrać w tabeli.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model1 | … | ModelN |
| Train R2 | ?? | ?? | ?? |
| Train RMSD | ?? | ?? | ?? |
| Test RMSD | ?? | ?? | ?? |

**Uwagi**

1. Można pracować w parach.

2. Proszę szczególną uwagę zwrócić na przejrzystość prezentowanych danych i wyników oraz ich interpretację.

3. Mile widziane jest korzystane z ggplot2 oraz dplyr.

4. Zwróć uwagę, że niektóre zmienne mogą być traktowane zarówno jako numeryczne jak i kategoryczne.

5. R2 dla najlepszego modelu powinno wynieść co najmniej 65%.