Trenowanie i ewaluacja modeli – część 3

1. Utwórz model średniej rocznej ilości opadów używając wysokości, logarytmu z wysokości oraz wartości wysokości podniesionej do kwadratu na zbiorze treningowym. Oblicz błąd średniokwadratowy RMSE.
2. Zastosuj funkcję predict() na testowym zbiorze danych dla modelu z ćw. 1. Oblicz błąd średniokwadratowy (RMSE) pomiędzy wartościami obserwowanymi a przewidywanymi.
3. Utwórz model średniej rocznej ilości opadów używając max ilości opadów, logarytmu z max ilości opadów oraz wartości max opadów podniesionej do kwadratu na zbiorze treningowym. Oblicz metrykę RMSE.
4. Zastosuj funkcję predict() na testowym zbiorze danych dla modelu z ćw. 3. Oblicz RMSE pomiędzy wartościami obserwowanymi a przewidywanymi.
5. Rozbuduj model tak, aby średnia ilość opadów była wyjaśniana przez wysokość, logarytm z wysokości, wartość wysokości podniesionej do kwadratu, max wielkość opadów, logarytm z max wysokości opadów, wartość max wysokosci opadów podniesionej do kwadratu oraz interakcję pomiędzy wysokością a maksymalną wielkością opadów. Oblicz RMSE.
6. Zastosuj funkcję predict() na testowym zbiorze danych dla modelu z ćw. 5. Oblicz RMSE pomiędzy wartościami obserwowanymi a przewidywanymi.

Kryteria informacji są miarą jakości dopasowania modelu statystycznego. Są różne kryteria wyboru zmiennych objaśniających do modelu. Najczęściej stosowanym jest kryterium **Akaike (AIC – Akaike's Information Criterion)**

$$\boldsymbol{AIC = 2k - 2\ln(L)}$$

k- liczba parametrów
L- funkcja wiarygodności

Aby nie sprawdzać wszystkich możliwych kombinacji- każdego możliwego podzbioru cech w celu wybrania modelu z najniższą wartością AIC, często stosowane są metody krokowe. Wyróżniamy 2 metody:

- **Forward selection** – polega na stopniowym dołączaniu do modelu kolejnych zmiennych zaczynając od modelu bez żadnej zmiennej objaśniającej. W kolejnych krokach dołączana jest po jednej zmiennej niezależnej, która minimalizuje kryterium AIC. Jeśli dodanie kolejnej dodatkowej cechy nie zmniejszy kryterium AIC, algorytm zatrzymuje się podając najlepsze specyfikacje modelu. W języku R wykonujemy **forward selection** używając funkcji step() definiując argument **direction='forward'**. Funkcja **step()** oblicza kryterium **AIC** i wybiera tą cechę która najbardziej minimalizuje AIC.

- **Backward selection** – polega na stopniowym usuwaniu z modelu kolejnych zmiennych zaczynając od modelu ze wszystkimi zmiennymi objaśniającymi. W kolejnych krokach usuwane są po jednej zmiennej, które najmniej wnoszą do modelu. Liczba obserwacji musi być większa od liczby parametrów. W języku R wykonujemy **backward selection** używając funkcji step() definiując argument **direction='backward'**.

7. Wykonaj selekcję modeli używając metody **forward selection** używając funkcji **step()** z argumentem **direction='forward'**.
   a) Utwórz model bazowy (bez żadnych zmiennych wyjaśniających).
   b) Zdefiniuj zakres zmiennych które powinny zostać uwzględnione (**Altitude, Max Rainfall, Mean cloud cover oraz Mean annual air temp**). Argument **scope()** funkcji step() przyjmuje jako dane wejściowe notację formuły. Uzyj funkcji **as.formula()**
   c) Wyświetl i zinterpretuj wynik.
8. Utwórz model na podstawie wybranych zmiennych z poprzedniego ćwiczenia . Oblicz RMSE dla zbioru treningowego.
9. Zastosuj funkcję predict() na testowym zbiorze danych dla modelu z ćw. 8. Oblicz RMSE pomiędzy wartościami obserwowanymi a przewidywanymi.
10. Wykonaj selekcję modeli analogicznie do ćw. 7. Tym razem zdefiniuj zakres w którym będą wszystkie zmienne objaśniające.
11. Wykonaj selekcję modeli używając metody **backward selection** używając funkcji step() z argumentem **direction='backward'**. Jako model początkowy utwórz model zmiennej objaśnianej przez następujace zmienne objaśniające: Altitude, Max Rainfall, Mean Cloud Cover oraz Mean Annual Air Temp.
12. Utwórz model na podstawie wybranych zmiennych z poprzedniego ćwiczenia . Oblicz RMSE dla zbioru treningowego.
13. Zastosuj funkcję predict() na testowym zbiorze danych dla modelu z ćw. 11. Oblicz RMSE pomiędzy wartościami obserwowanymi a przewidywanymi.
14. Porównaj wynik modelu z ćw. 7 i 11.
15. Wykonaj selekcję modelu analogicznie do ćw. 11 używając metody backward selection. Tym razem zacznij od modelu startowego w którym są wszystkie zmienne objaśniające.
16. Utwórz model na podstawie wybranych zmiennych z poprzedniego ćwiczenia . Oblicz RMSE dla zbioru treningowego.
17. Zastosuj funkcję predict() na testowym zbiorze danych dla modelu z ćw. 16. Oblicz RMSE pomiędzy wartościami obserwowanymi a przewidywanymi.
18. Porównaj wyniki wszystkich utworzonych modeli (z części 1-3). Przedstaw wartości RMSE dla zbioru testowego i treningowego dla każdego z nich.