🔎 Feature Selection – teoretyczne wprowadzenie

📌 1. Problem selekcji cech

W problemie selekcji cech (ang. *feature selection*) mamy dane:

- Macierz cech:

$$X = \begin{bmatrix} X_1^{(1)} & X_2^{(1)} & \cdots & X_p^{(1)} \\ X_1^{(2)} & X_2^{(2)} & \cdots & X_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ X_1^{(n)} & X_2^{(n)} & \cdots & X_p^{(n)} \end{bmatrix}$$

  gdzie n to liczba obserwacji, a p — liczba wszystkich cech.

- Wektor etykiet $Y \in \{0, 1\}^n$.

Zadaniem jest:
✅ Wskazanie (lub ranking) cech, które są **istotne** dla przewidywania Y.

📌 2. Generowanie danych (dataset 1 i dataset 2)

🔸 **Dataset 1:**

- $X_j \sim \mathcal{N}(0, 1)$, $j = 1, \ldots, p$.
- Oznaczmy przez k liczbę istotnych cech (np. $k = 10$).
- Dla każdej obserwacji:
  - Obliczamy:

$$S = \sum_{j=1}^{k} X_j^2$$

$$j=1$$

- Wyznaczamy wartość krytyczną:

$$c = \chi^2_k(0.5)$$

(czyli medianę rozkładu chi-kwadrat z k stopniami swobody).

- Etykieta:

$$Y = \begin{cases} 1 & \text{jeśli } S > c \\ 0 & \text{w przeciwnym razie} \end{cases}$$

🔹 **Dataset 2:**

- $X_j \sim \mathcal{N}(0,1), j = 1, \ldots, p.$
- Dla każdej obserwacji:
  - Obliczamy:

$$S = \sum_{j=1}^{k} |X_j|$$

  - Etykieta:

$$Y = \begin{cases} 1 & \text{jeśli } S > k \\ 0 & \text{w przeciwnym razie} \end{cases}$$

📌 3. Metody selekcji cech

🔹 **(a) Random Forest**

Random Forest (RF) dostarcza dwóch miar ważności cech:

- **Mean Decrease in Impurity (MDI)**:
  - Obliczana na podstawie spadku nieczystości (np. entropii) przy każdym podziale drzewa.
  - Dla cechy j:

$$\text{MDI}_j = \sum_{\downarrow} \left(\Delta i(t) \cdot \frac{N_t}{N} \right)$$

gdzie:

- $\Delta i(t)$ — zmniejszenie nieczystości przy podziale w węźle t.
- N_t — liczność próbek w węźle t.
- N — liczność wszystkich próbek.

- **Permutation Importance (Perm)**:
  - Szacowana poprzez permutowanie wartości cechy j i obserwowanie spadku dokładności modelu.

🔸 **(b) Boruta**

- Algorytm Boruta opiera się na Random Forest.
- Tworzy kopie (shadow features) — permutacje oryginalnych cech.
- Porównuje ważność oryginalnych cech z ważnością shadow features, testując istotność.
- Wynik: każda cecha oznaczana jako:
  - Istotna (ważna)
  - Nieistotna
  - Niepewna

📌 4. Prawidłowy ranking

Definiujemy **idealne uporządkowanie**:

- Cecha istotna powinna mieć ranking wyższy (większą ważność) niż dowolna cecha nieistotna.

Matematycznie:

- Dla k istotnych cech i $p - k$ szumowych:
  - Sortujemy cechy malejąco według ważności.
  - Sprawdzamy, czy wszystkie istotne cechy są na pierwszych k miejscach rankingu.

📌 5. Miary skuteczności

🔹 **Prawdopodobieństwo prawidłowego uporządkowania**

- Szacujemy:

$$\hat{P} = \frac{\text{liczba powtórzeń, w których istotne cechy są na topie}}{L}$$

gdzie L — liczba powtórzeń eksperymentu.

🔹 **Dokładność klasyfikacji**

- Trenujemy klasyfikator (np. Random Forest) na t najwyżej ocenionych cechach.
- Sprawdzamy dokładność:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- Rysujemy wykres: Accuracy vs. t.

📌 6. Eksperymenty

✅ Eksperymenty obejmują:

- Różne wartości:
  - n (np. 200, 500)
  - p (np. 50, 100, 500)
  - k (np. 5, 10, 20)
- Powtarzamy $L = 50$ razy, by oszacować stabilność wyników.