🔹 1. Generowanie zbiorów danych

📌 (a) Zbiór danych rzeczywistych

Wybieramy rzeczywisty zbiór danych do klasyfikacji binarnej — np. Breast Cancer Wisconsin z `scikit-learn`.

Dane:

- $y \in \{0, 1\}$ — etykieta klasy
- $x_1, x_2, \ldots, x_p$ — zmienne wejściowe.

📌 (b) Zbiór danych syntetycznych

Dane generowane według modelu regresji logistycznej:

$$y_i \sim \text{Bernoulli}(p_i)$$

gdzie:

$$p_i = \frac{1}{1 + \exp[-(\alpha + \beta^T x_i)]}$$

oraz:

- $x_i \sim \mathcal{N}(0, I_p)$
- $\beta = (b, b, b, b, b, 0, \ldots, 0)$
- α, b, k, n — parametry zadania:
  - α — intercept,
  - b — współczynnik istotnych zmiennych,
  - k — liczba nieistotnych zmiennych (szum),
  - n — liczba obserwacji.

🔹 2. Dopasowanie modeli klasyfikacyjnych

📌 **Modele:**

- **Regresja logistyczna:**

$$P(y = 1|x) = \frac{1}{1 + \exp[-(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)]}$$

- **Drzewo decyzyjne:**
  Algorytm CART — dzieli przestrzeń cech na prostokątne regiony w celu klasyfikacji.

🔹 3. Schematy oceny błędu klasyfikacji

Dla ustalonych parametrów (np. $n = 1000, b = 1, k = 20$) oceniamy błąd klasyfikacji na różne sposoby:

📌 **(a) Refitting (apparent error)**

Uczenie i testowanie na tym samym zbiorze:

$$\text{Error} = 1 - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_i = \hat{y}_i)$$

📌 **(b) 10-krotna walidacja krzyżowa**

Podział zbioru na 10 części:

- W każdej iteracji model jest trenowany na 9 częściach i testowany na 1.
- Błąd to średnia z 10 powtórzeń.

📌 (c) Bootstrap

Losujemy próbki z powtórzeniami (bootstrap sample), model uczymy na próbce bootstrapowej, a oceniamy na pozostałych obserwacjach (out-of-bag).

📌 (d) Bootstrap 0.632

Kombinacja:

$$\text{Error}_{.632} = 0.368 \cdot \text{Error}_{app} + 0.632 \cdot \text{Error}_{OOB}$$

gdzie:

- Error_{app} — błąd dopasowania (na wszystkich danych)
- Error_{OOB} — błąd bootstrapowy na out-of-bag.

🔹 4. Krzywe ROC i Precision-Recall

📌 Krzywa ROC

- Oś x: False Positive Rate
- Oś y: True Positive Rate

📌 Krzywa Precision-Recall

- Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall:

$$\text{Recall} \downarrow \frac{\text{TP}}{\text{TP} + \text{FN}}$$

W analizie sprawdzamy różne parametry ($n = 100, 1000; b = 0.5, 1; k = 5, 50$).

🔹 5. Analiza progu decyzyjnego

📌 Predykcja

Próg decyzyjny t:

$$\hat{y} = \begin{cases} 1, & \text{jeśli } P(y = 1|x) > t \\ 0, & \text{w przeciwnym razie} \end{cases}$$

📌 Accuracy

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

📌 Balanced Accuracy

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right)$$

📌 Uwagi:

- $t = 0.5$ — optymalny dla klasycznej precyzji.
- $t = p(y = 1)$ — często optymalny dla balanced accuracy (w przypadku niezrównoważonych klas).