

Ćwiczenie laboratoryjne

12. Badania przypadków klinicznych: analiza danych pacjentów i budowa modeli predykcyjnych w oparciu o rzeczywiste przypadki

1 Badania przypadków klinicznych: analiza danych pacjentów i budowa modeli predykcyjnych w oparciu o rzeczywiste przypadki

1.1 Cel ćwiczenia

Celem ćwiczenia jest zapoznanie studentów z procesem **analizy rzeczywistych (lub realistycznych) przypadków klinicznych** oraz budowy **modeli predykcyjnych wspomagających decyzje medyczne**. Ćwiczenie integruje wiedzę z zakresu:

- analizy danych pacjentów,
- eksploracji danych klinicznych,
- uczenia maszynowego,
- interpretacji wyników modeli w kontekście klinicznym.

Ćwiczenie ma charakter **dydaktyczny i symulacyjny** i nie stanowi narzędzia diagnostycznego.

1.2 Wprowadzenie teoretyczne

Badania przypadków klinicznych (ang. *case studies*) polegają na analizie indywidualnych lub grupowych danych pacjentów w celu:

- identyfikacji czynników ryzyka,
- wykrywania wzorców chorobowych,
- przewidywania zdarzeń klinicznych (np. hospitalizacji, powikłań).

Formalnie problem predykcyjny można zapisać jako:

$$\hat{y} = f(\mathbf{x}),$$

gdzie:

- $\mathbf{x} = (x_1, x_2, \dots, x_n)$ – wektor cech klinicznych pacjenta,
- y – zmienna docelowa (np. rozpoznanie, ryzyko),
- f – model uczenia maszynowego.

1.3 Dane kliniczne

W ćwiczeniu wykorzystywany jest zbiór danych pacjentów zawierający m.in.:

- dane demograficzne (wiek, płeć),
- parametry antropometryczne (BMI),
- parametry fizjologiczne (ciśnienie, glukoza),
- zmienną docelową (np. ryzyko choroby lub diagnoza).

Dane mają charakter tabelaryczny i są zapisane w formacie CSV.

1.4 Technologie i narzędzia

W ćwiczeniu wykorzystywane są następujące technologie:

- język **Python**,
- środowisko **Jupyter Notebook**,
- biblioteki:
 - `pandas` – analiza danych,
 - `numpy` – obliczenia numeryczne,
 - `matplotlib / seaborn` – wizualizacja,
 - `scikit-learn` – modele ML i ewaluacja.

1.5 Przebieg ćwiczenia laboratoryjnego

Etap 1: Analiza przypadku klinicznego

- Zapoznaj się z opisem przypadku (profil pacjenta).
- Określ problem predykcyjny (np. przewidywanie choroby, ryzyka).
- Zidentyfikuj istotne cechy kliniczne.

Etap 2: Przygotowanie danych

- Załaduj dane pacjentów z pliku CSV.
- Sprawdź braki danych i zastosuj imputację.
- Zakoduj zmienne kategoryczne.
- Dokonaj standaryzacji wybranych cech.

Etap 3: Analiza eksploracyjna

- Oblicz statystyki opisowe.
- Wykonaj wizualizacje (histogramy, wykresy pudełkowe).
- Przeanalizuj zależności między cechami a zmienną docelową.

Etap 4: Budowa modelu predykcyjnego

- Podziel dane na zbiór treningowy i testowy.
- Zbuduj co najmniej jeden model ML (np. regresja logistyczna).
- Przeprowadź trening modelu.

Etap 5: Ewaluacja i interpretacja

- Oblicz metryki jakości (Accuracy, Precision, Recall, ROC AUC).
- Zinterpretuj wyniki w kontekście klinicznym.
- Wskaż ograniczenia modelu.

1.6 Przykładowy kod (fragment)

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

X = data.drop("diagnosis", axis=1)
y = data["diagnosis"]

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42
)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
```

1.7 Warianty zadań dla studentów

1. Analiza pojedynczego przypadku klinicznego i jego klasyfikacja.
2. Porównanie dwóch modeli (np. logistyczny vs. drzewo decyzyjne).
3. Wpływ wybranej cechy na predykcję modelu.
4. Analiza błędnych predykcji (false positives / false negatives).
5. Ocena stabilności modelu przy różnych podziałach danych.

1.8 Aspekty etyczne i praktyczne

- Dane pacjentów muszą być anonimizowane.
- Modele predykcyjne nie zastępują decyzji lekarza.
- Należy unikać nadmiernego dopasowania (overfitting).

1.9 Podsumowanie

Ćwiczenie pokazuje pełny proces pracy z przypadkami klinicznymi: od analizy danych pacjentów, przez eksplorację, aż po budowę i ocenę modeli predykcyjnych. Umiejętności te stanowią podstawę praktycznego wykorzystania sztucznej inteligencji w medycynie.

2 Warianty zadań (15) — modele, progi decyzyjne i analiza przypadków klinicznych

Każdy wariant dotyczy pracy na notebooku oraz zbiorze `cases_clinical_for_lab12.csv` (lub własnym CSV o tej samej strukturze). W każdym wariantie student ma wykonać: (i) preprocessing i EDA, (ii) trening co najmniej jednego modelu, (iii) ocenę metryk, (iv) analizę błędów FP/FN jako **badania przypadków**.

Wspólna modyfikacja kodu (próg decyzyjny). W funkcji ewaluacji należy wprowadzić parametr `threshold`:

$$\hat{y} = \mathbb{1}(p(y=1 | \mathbf{x}) \geq \tau),$$

gdzie τ to próg decyzyjny. Studenci mają raportować wpływ τ na FP i FN.

Wariant 1: Baseline + próg 0.4 vs 0.6 (wrażliwość na FN/FP)

Modele: Regresja logistyczna (LR).

Progi: $\tau \in \{0.4, 0.5, 0.6\}$.

Zadania:

- Porównaj macierze pomyłek dla trzech progów.
- Wybierz próg minimalizujący FN (uzasadnij klinicznie).
- Opisz 3 przypadki FN i 3 przypadki FP (case studies).

Wariant 2: Random Forest + strojenie progu dla minimalizacji FN

Modele: Random Forest (RF).

Progi: przeszukanie $\tau \in [0.20, 0.80]$ co 0.05.

Zadania:

- Znajdź próg zapewniający $\text{Recall} \geq 0.90$ (jeśli możliwe).
- Podaj koszt uboczny w postaci wzrostu FP.
- Przedstaw 5 najbardziej „niepewnych” przypadków (p w okolicy progu).

Wariant 3: Porównanie dwóch rodzin modeli (liniowy vs nieliiniowy)

Modele: LR vs RF.

Progi: stały $\tau = 0.5$ oraz próg dobrany pod $\text{Recall}=0.85$.

Zadania:

- Porównaj ROC AUC i PR AUC (dla obu modeli).
- Dobierz próg tak, aby osiągnąć zadaną czułość i porównaj FP.
- Opisz, które cechy dominują w Permutation Importance dla RF.

Wariant 4: Model SVM z prawdopodobieństwami

Modele: SVM (`SVC(probability=True)`).

Progi: $\tau \in \{0.4, 0.5, 0.6\}$.

Zadania:

- Dodaj SVM do notebooka i porównaj z LR.
- Oceń wpływ progów na FN/FP.
- Zidentyfikuj 3 przypadki, gdzie SVM i LR się nie zgadzają (i opisz je).

Wariant 5: Gradient Boosting (model trzeciej kategorii)

Modele: Gradient Boosting (`GradientBoostingClassifier`) lub HistGB.

Progi: $\tau \in \{0.35, 0.5, 0.65\}$.

Zadania:

- Porównaj z RF pod kątem PR AUC.
- Ustal próg minimalizujący FN przy ograniczeniu $\text{FP} \leq$ określonej wartości (np. 25%).
- Przedstaw analizę 5 błędów FN i możliwe przyczyny.

Wariant 6: Skrajna nierównowaga klas (symulacja) + metryki

Modele: LR i RF.

Modyfikacja danych: undersampling klasy 1 do 10–15% udziału.

Progi: τ dobierany pod F1 lub Recall.

Zadania:

- Pokaż, jak zmienia się Precision/Recall.
- Uzasadnij, czemu Accuracy jest mylące przy nierównowadze.
- Dobierz próg w zależności od celu klinicznego.

Wariant 7: Wpływ skalowania cech na modele liniowe

Modele: LR bez skalowania vs LR ze skalowaniem (w pipeline).

Progi: $\tau = 0.5$ oraz τ dobrany pod najlepsze F1.

Zadania:

- Porównaj współczynniki LR (po standaryzacji) i ich znaki.
- Oceń zmianę FN/FP po zmianie progu.
- Opisz 3 przypadki, w których model zmienił decyzję po skalowaniu.

Wariant 8: Kalibracja prawdopodobieństw

Modele: RF + kalibracja (CalibratedClassifierCV).

Progi: $\tau \in \{0.4, 0.5, 0.6\}$.

Zadania:

- Porównaj rozkład predykowanych prawdopodobieństw przed i po kalibracji.
- Sprawdź, czy dla tego samego progu zmieniają się FN/FP.
- Uzasadnij, kiedy kalibracja jest ważna klinicznie.

Wariant 9: Reguła decyzyjna typu „strefa niepewności”

Modele: dowolny probabilistyczny (LR lub RF).

Progi: dwuprogiowe: $\tau_L = 0.4$, $\tau_H = 0.6$.

Zadania:

- Zdefiniuj 3 decyzje: NEG jeśli $p < \tau_L$, POS jeśli $p > \tau_H$, REVIEW w przeciwnym razie.
- Zmierz odsetek przypadków REVIEW.
- Opisz 5 przypadków REVIEW i zaproponuj dodatkowe badania/cechy.

Wariant 10: Analiza podgrup pacjentów (fairness / segmenty)

Modele: LR lub RF.

Progi: $\tau = 0.5$ oraz progi osobne dla podgrup (np. płeć).

Zadania:

- Oceń metryki osobno dla podgrup (np. F vs M, przedziały wieku).
- Sprawdź, czy próg 0.5 powoduje istotnie więcej FN w którejś grupie.
- Zaproponuj strategię doboru progów (jednolity vs per-grupa) i omów konsekwencje etyczne.

Wariant 11: Analiza cech i redukcja wymiaru (feature ablation)

Modele: RF.

Progi: $\tau = 0.5$.

Zadania:

- Usuń po jednej cechę (ablation) i oceń spadek ROC AUC.
- Porównaj z Permutation Importance.
- Wskaż minimalny zestaw 3–4 cech dający akceptowalne wyniki.

Wariant 12: Interpretacja lokalna przez „najbardziej podobne przypadki”

Modele: LR lub RF.

Progi: $\tau = 0.5$ i $\tau = 0.4$.

Zadania:

- Dla wybranego pacjenta (np. FP lub FN) znajdź 5 najbardziej podobnych przypadków (np. odległość euklidesowa po skalowaniu).
- Porównaj ich wyniki i cechy.
- Uzasadnij, czy decyzja modelu jest spójna z „sąsiadami”.

Wariant 13: Optymalizacja progu względem kosztu klinicznego

Modele: RF.

Progi: przeszukanie $\tau \in [0.1, 0.9]$.

Założenie kosztów: $C_{FN} \gg C_{FP}$ (np. $C_{FN} = 10$, $C_{FP} = 1$).

Zadania:

- Zdefiniuj koszt całkowity: $C(\tau) = 10 \cdot FN(\tau) + 1 \cdot FP(\tau)$.
- Znajdź próg minimalizujący $C(\tau)$.
- Zinterpretuj wynik w kontekście screeningu.

Wariant 14: Porównanie metryk i wybór modelu pod zastosowanie

Modele: LR, RF, Boosting (3 modele).

Progi: $\tau = 0.5$ oraz próg dopasowany do Precision=0.80 (jeśli możliwe).

Zadania:

- Wybierz model do: (a) screeningu, (b) diagnostyki pogłębionej.
- Uzasadnij wybór metrykami (Recall/Precision/PR AUC).
- Przedstaw analizę 5 błędów krytycznych (FN) dla wybranego scenariusza.

Wariant 15: Stabilność modelu i wpływ losowości podziału

Modele: RF.

Progi: $\tau = 0.5$ i $\tau = 0.4$.

Zadania:

- Powtórz podział train/test dla 5 różnych `random_state`.
- Raportuj rozrzut ROC AUC oraz FN/FP przy różnych progach.
- Opisz, które przypadki najczęściej stają się FN (czy są „trudne” systematycznie?).

Wymagania do raportu (wspólne dla wszystkich wariantów)

W raporcie należy uwzględnić:

- opis problemu i danych (kolumny, braki, preprocessing),
- porównanie modeli i metryk (co najmniej: Confusion Matrix, ROC AUC lub PR AUC),
- wpływ progów decyzyjnych na FN i FP (co najmniej 2 wartości τ),
- analizę przypadków klinicznych: min. 3 FP i 3 FN (opis cech i możliwych przyczyn),
- wnioski i ograniczenia (np. generalizacja, bias, brak walidacji klinicznej).