W08. Uczenie maszynowe w psychologii

Psychoinformatyka | Wykład 08

2026-11-04

Cele i zakres wykładu

- Odróżnić wyjaśnianie mechanizmu od predykcji nowych obserwacji.
- Rozróżnić regresję, klasyfikację i klasteryzację.
- Zbudować procedurę oceny bez przecieku danych: rozdzielenie osób, pipeline, metryki.
- Zinterpretować wynik modelu w kontekście psychologicznym.

Studium przypadku: syntetyczne dane 120 osób × 4 pomiary; przewidywanie etykiety podwyższonego stresu.

Wyjaśnianie a przewidywanie

- Wyjaśnianie: jakie procesy i zależności mogą tłumaczyć obserwowane zjawisko?
- Predykcja: jak trafnie oszacujemy wynik dla nieznanej obserwacji lub nowej osoby?
- Model o wysokiej skuteczności predykcyjnej nie dowodzi przyczynowego wpływu cech.
- Oba cele wymagają odmiennego planu badania i walidacji.

Yarkoni i Westfall (2017): DOI.

Uczenie nadzorowane: X → model → y

- Cechy X: sen [h], czas korzystania z telefonu [min], liczba kroków.
- Etykieta y: umowna klasa 0/1 w zbiorze syntetycznym.
- Uczenie: oszacowanie parametrów z danych treningowych.



Rysunek 1: Schemat modelu nadzorowanego: cechy, model i etykieta

Uczenie nadzorowane i nienadzorowane

- Klasyfikacja: przewidujemy kategorię, np. wynik 0/1.
- Regresja: przewidujemy wartość ciągłą, np. liczbę punktów.
- Klasteryzacja: poszukujemy podobnych profili bez gotowej etykiety.
- Klaster nie jest automatycznie typem osobowości ani kategorią diagnostyczną.

Pytanie badawcze przed wyborem algorytmu

1. Kogo chcemy przewidywać: kolejną próbę tej osoby czy nowego uczestnika?
2. Kiedy predykcja ma powstać: tego dnia czy przed kolejnym pomiarem?
3. Jakim kosztem: które błędy są bardziej dotkliwe?
4. Na jakiej podstawie zdefiniowano etykietę i zgodę na użycie danych?

Cechy, etykieta i przeciek informacji

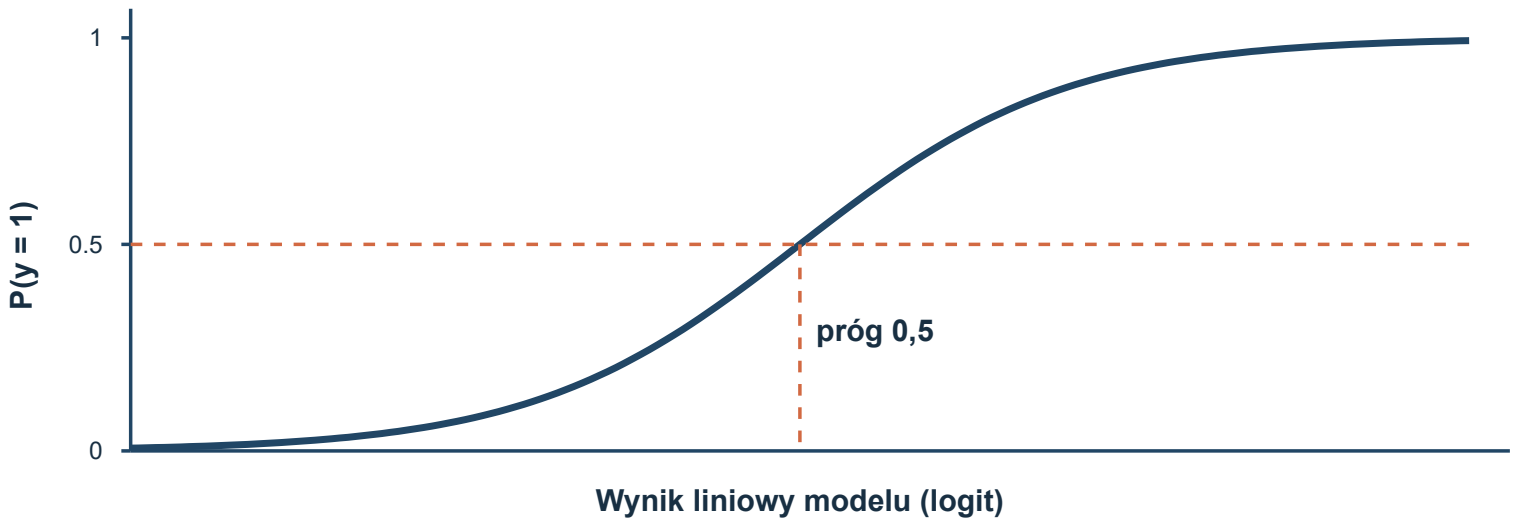
- Cecha: wielkość dostępna w czasie predykcji.
- Etykieta: wynik, którego model ma się nauczyć.
- Przeciek celu: cecha zawiera fragment odpowiedzi, np. wynik kwestionariusza użytego do utworzenia etykiety.
- Przeciek podziału: ten sam uczestnik trafia do treningu i testu.

Regresja, klasyfikacja i prawdopodobieństwo

- W regresji wynik ma skalę liczbową, np. przewidywany czas reakcji [ms].
- W klasyfikacji wynik to etykieta klasy oraz często jej oszacowane prawdopodobieństwo.
- Próg decyzyjny zamienia prawdopodobieństwo na kategorię, np. (p 5).
- Wartość (0{,}8) jest predykcją modelu, nie pewnością kliniczną.

Regresja logistyczna

- Modeluje (P(y=1X)=(eta_0+eta^ op X)), gdzie ((z)=1/(1+e^{-z})).
- Odpowiedni punkt odniesienia dla klasyfikacji binarnej.
- Regularizacja ogranicza nadmierne dopasowanie współczynników.



Rysunek 2: Sigmoida regresji logistycznej z progiem 0,5

Drzewo decyzyjne: reguły podziału

- Węzeł dzieli dane na podstawie progu wybranej cechy.
- Kolejne podziały tworzą lokalne reguły przypisywania klas.
- Głębokie drzewa mogą dopasowywać się do szumu pomiarowego.
- Interpretacja ścieżki nie oznacza identyfikacji mechanizmu psychologicznego.

Pytanie: co się stanie, gdy zmienimy próg dla snu z 6,5 h na 7 h?

Las losowy: agregacja drzew

- Random Forest uczy wiele drzew na różnych próbkach i podzbiorach cech.
- Głosowanie lub uśrednianie stabilizuje predykcje względem pojedynczego drzewa.
- Wciąż wymaga walidacji poza próbką treningową.
- Większa złożoność nie gwarantuje lepszej generalizacji w małym badaniu.

Klasteryzacja profili zachowania

- Przykładowe wektory: średnia długość snu, zmienność aktywności, regularność rytmu.
- K-means minimalizuje odległości wewnątrz klastrów dla zadanej liczby grup.
- Skalowanie cech i metryka odległości wpływają na przypisania.
- Stabilność klastrów i ich trafność konstruktu trzeba oceniać osobno.

Braki, skale i kodowanie cech

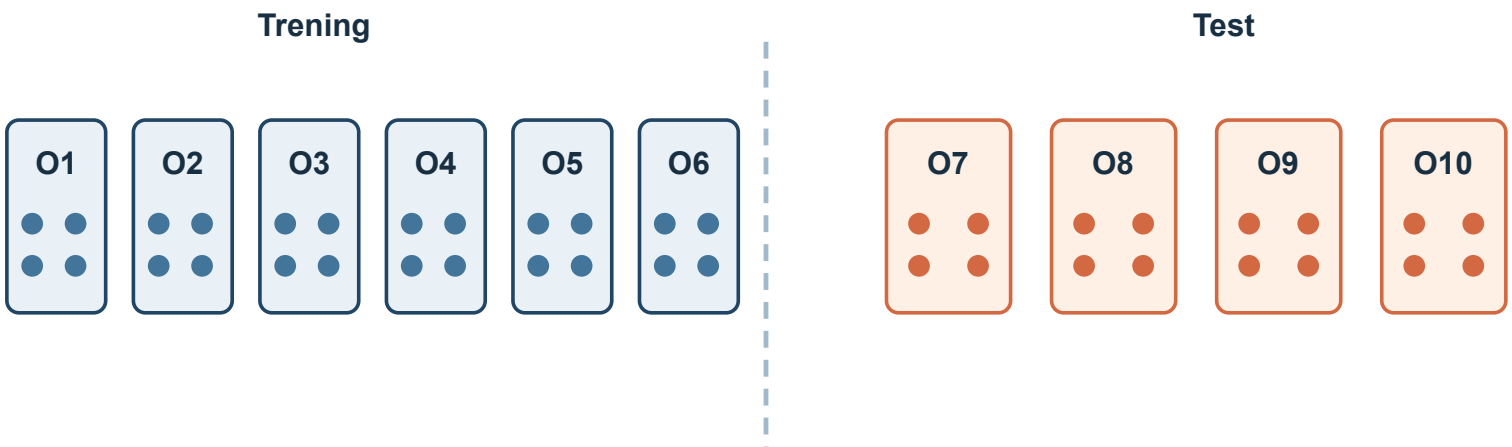
- Braki pomiaru mogą wynikać z niedziałającego czujnika lub braku aktywności.
- Imputacja medianą i standaryzacja mają być dopasowane wyłącznie na treningu.
- `Pipeline` przenosi te kroki na walidację i test bez ponownego uczenia transformacji.
- Zmienne kategoryczne wymagają zdefiniowanego kodowania.

Niezrównoważenie klas

- Jeśli 90% przypadków należy do klasy 0, stała predykcja 0 daje 90% accuracy.
- Takie rozwiązanie ma recall = 0 dla klasy pozytywnej.
- Używaj miar zależnych od celu: balanced accuracy, precision, recall, F1.
- Porównuj każdy model z prostym punktem odniesienia.

Trening i test: podział według osób

- Obserwacje tego samego uczestnika są statystycznie zależne.
- Dla celu „nowa osoba„ grupą podziału jest identyfikator uczestnika.
- `GroupShuffleSplit` zapewnia rozłączność osób w treningu i teście.



Rysunek 3: Przydział wszystkich pomiarów jednego uczestnika do jednego zbioru

Projekt badania: syntetyczne dane stresu

- 120 osób × 4 pomiary = 480 wierszy.
- Trzy cechy: `sen_h`, `telefon_min`, `kroki`.
- 20 braków snu; etykieta generowana ze znanego modelu probabilistycznego.
- Podział 75/25 według osób: 90 treningowych, 30 testowych.

Punkt odniesienia: `DummyClassifier`, który przewiduje najczęstszą klasę treningową.

5 minut przerwy

Pytania do przemyślenia:

1. Dlaczego wysoka accuracy nie musi oznaczać dobrej klasyfikacji?
2. Co oznacza przeciek informacji między treningiem i testem?
3. Po co rozdzielać pomiary tej samej osoby?
4. Czym różni się ocena predykcji od wyjaśniania przyczyn?

Etapy budowania modelu

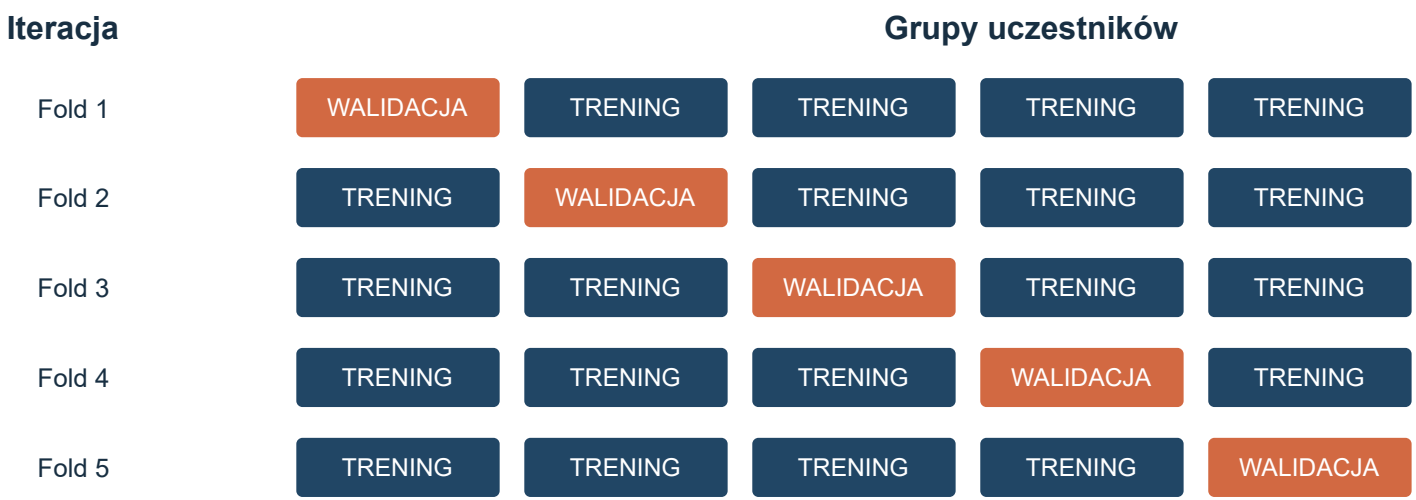
1. Zdefiniuj cel predykcji, populację i punkt czasowy.
2. Zamroź zestaw testowy według uczestników.
3. Wewnątrz treningu: imputacja, skalowanie, wybór modelu, walidacja grupowa.
4. Raz oceń model na teście; zapisz macierz pomyłek i metryki.
5. Zbadaj stabilność, kalibrację i ograniczenia przenoszenia.

Niedouczenie i przeuczenie

- Niedouczenie: model jest zbyt prosty lub źle dopasowany do wzorca.
- Przeuczenie: model odtwarza osobliwości treningu, tracąc jakość poza nim.
- Porównanie błędów treningowych i walidacyjnych pomaga wykrywać problemy.
- Złożoność, liczebność osób i regularizacja wpływają na generalizację.

Walidacja krzyżowa według grup

- W GroupKFold każdy uczestnik należy tylko do jednego foldu walidacyjnego.
- Każdy fold uczy wszystkie kroki `Pipeline` na własnym treningu.
- Ocena opiera się na wynikach foldów, ich rozrzucie i liczbie niezależnych osób.



Rysunek 4: Pięć iteracji walidacji krzyżowej według grup uczestników

LOSO i walidacja w przyszłości

- LOSO (leave-one-subject-out): każdorazowo testujemy model na jednej pominiętej osobie.
- GroupKFold: mniej dopasowań, większe grupy w walidacji.
- Walidacja czasowa: wymaga respektowania kolejności zdarzeń i chwili predykcji.
- Walidacja zewnętrzna: nowa kohorta, urządzenie lub ośrodek.

Macierz pomyłek

- TN: poprawnie rozpoznany brak podwyższonego stresu.
- FP: fałszywy alarm; FN: pominięta klasa pozytywna.
- TP: poprawnie rozpoznana klasa dodatnia.



Rysunek 5: Macierz pomyłek na syntetycznym zbiorze testowym

Dane syntetyczne, próg 0,5, test: 120 pomiarów od 30 nowych osób.

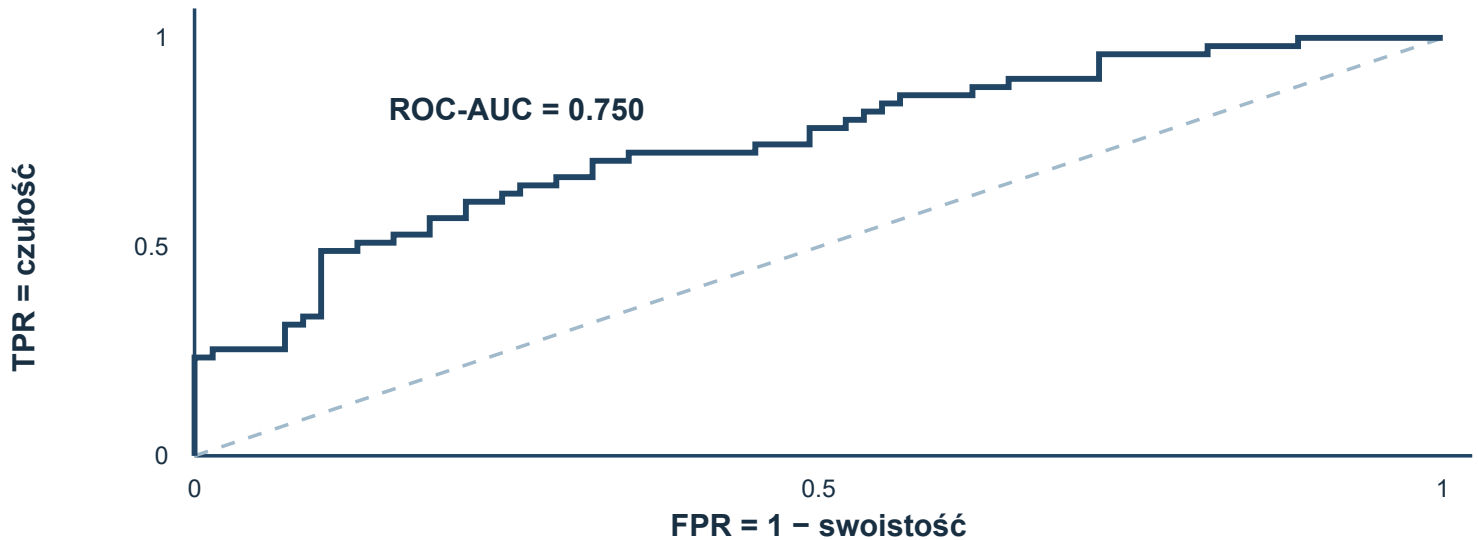
Miary jakości predykcji

- Accuracy: odsetek poprawnych predykcji; wrażliwa na niezrównoważenie klas.
- Balanced accuracy: średnia czułości i swoistości.
- Precision: odsetek prawdziwie dodatnich wśród prognoz dodatnich.
- Recall: czułość wykrywania klasy pozytywnej; F1: średnia harmoniczna precision i recall.

Na syntetycznym teście: balanced accuracy 0.623, F1 0.459.

ROC, PR-AUC i progi

- Krzywa ROC zestawia czułość (TPR) i odsetek fałszywie dodatnich (FPR).
- ROC-AUC ocenia zdolność do rangowania klas w wielu progach.
- PR-AUC pomaga analizować klasy rzadkie; kalibracja bada zgodność prognozowanych prawdopodobieństw z częstościami obserwowanymi.



Rysunek 6: Krzywa ROC dla syntetycznego zbioru testowego

Python: generowanie danych syntetycznych

- Wytwarzamy trzy cechy i losową składową uczestnika.
- Etykieta powstaje probabilistycznie; nie jest pomiarem psychologicznym.

```
120 uczestników × 4 pomiary.
Cechy snu, telefonu i aktywności.
Syntetyczne etykiety oraz brakujące wartości.
```

Python: rozdzielenie osób i pipeline

- Rozdzielamy 90 osób do treningu i 30 do testu.
- Imputacja i skalowanie są uczone w każdym `fit` modelu.

```
GroupShuffleSplit: rozdział całych osób.
Imputacja → skalowanie → regresja logistyczna.
Każdy krok dopasowany tylko do treningu.
```

Python: walidacja grupowa i model bazowy

- GroupKFold(5) porównuje działanie modelu na niewidzianych osobach w zbiorze treningowym.
- W końcowym teście porównujemy `DummyClassifier` z regresją logistyczną.

```
GroupKFold(5): walidacja na nowych osobach.
DummyClassifier: punkt odniesienia.
Regresja logistyczna: niezależny test.
```

Python: macierz pomyłek i ROC-AUC

- Z prawdopodobieństw tworzymy predykcje przy ustalonym progu 0,5.
- Raportujemy macierz pomyłek, F1, ROC-AUC i PR-AUC.

```
from sklearn.metrics import roc_curve
proba = model.predict_proba(X.iloc[test])[:, 1]
pred = (proba >= .5).astype(int)  # próg ustalony przed oceną testową
tn, fp, fn, tp = confusion_matrix(y.iloc[test], pred,
    labels=[0, 1]).ravel()
print('TN, FP, FN, TP:', tn, fp, fn, tp)
print('Precision:', round(precision_score(y.iloc[test], pred), 3),
      'Recall:', round(recall_score(y.iloc[test], pred), 3),
      'F1:', round(f1_score(y.iloc[test], pred), 3))
print('ROC-AUC:', round(roc_auc_score(y.iloc[test], proba), 3),
      'PR-AUC:', round(average_precision_score(y.iloc[test], proba), 3))
fpr, tpr, _ = roc_curve(y.iloc[test], proba)
```

Interpretacja cech i ograniczenia wyjaśnienia

- Współczynniki modelu logistycznego opisują zmiany log-odds przy kontroli pozostałych cech.
- Standaryzacja pozwala porównywać skale, ale nie identyfikuje przyczyn.
- Ważność cech i SHAP objaśniają sposób działania modelu, nie mechanizm psychiczny.
- Zmienne współliniowe oraz zależność między pomiarami zmieniają interpretację.

Trafność, równość i odpowiedzialne wdrożenie

- Trafność konstruktu: czy etykieta i predyktory mierzą zamierzone zjawisko?
- Przenoszenie: sprawdzaj wyniki dla nowych osób, urządzeń, okresów i populacji.
- Nierówności: oceniaj różnice błędów w podgrupach przy dostatecznej liczebności.
- Prywatność: ogranicz cechy wrażliwe, retencję, zakres zgody i wtórne wykorzystanie.

Reguła: skuteczność predykcyjna nie jest równoznaczna z przydatnością kliniczną.

Literatura i dokumentacja

- Yarkoni, T., Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science, 12(6), 1100–1122. DOI.
- Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. NeuroImage, 180, 68–77. DOI.
- Roberts, D. R., i in. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929. DOI.

Dokumentacja narzędzi: scikit-learn: przeciek danych, GroupKFold, metryki klasyfikacji.

Wnioski i samodzielna praca

- Zdefiniuj pytanie predykcyjne i jednostkę generalizacji przed modelowaniem.
- Rozdziel osoby, a następnie dopasuj transformacje i model w pipeline.
- Raportuj macierz pomyłek, metryki, niestabilność, wynik bazowy i kalibrację prawdopodobieństw.
- Sprawdź trafność konstruktu, kalibrację, prywatność i walidację zewnętrzną.

Ćwiczenie: uruchom `pi_w08.ipynb`, zmień odsetek braków lub próg i opisz wpływ na precision oraz recall. Powiązanie z laboratorium L05.