

Wstęp do Uczenia Maszynowego

Projekt

Termin oddania projektu: 06.01.2026

Termin odpowiedzi: 28-29.01.2026

1 Wstęp

Celem jest zaproponowanie metody klasyfikacji, która pozwoli zbudować model o jak największej mocy predykcyjnej. Dysponujemy sztucznie wygenerowanym zbiorem danych *artificial*, w którym zostały ukryte istotne zmienne. Należy dokonać klasyfikacji do dwóch klas. Dokładność modelu będzie mierzona za pomocą miary zrównoważonej dokładności (*balanced accuracy*).

Projekt jest wykonywany samodzielnie!

2 Zbiór danych

Dane do projektu to sztucznie wygenerowany zbiór, który zawiera 100 zmiennych objaśniających (część z tych kolumn może być zbędna). Zbiór treningowy zawiera 1500 obserwacji, natomiast zbiór testowy 500.

Dostępne są następujące pliki:

- zbiór treningowy: `artificial_train_data.csv`,
- etykiety zbioru treningowego: `artificial_train_labels.csv`,
- zbiór testowy: `artificial_test_data.csv`.

Aby wczytać zbiór danych w języku Python wystarczy użyć funkcji `read_csv` z pakietu `pandas`.

3 Oczekiwany wynik

Na przygotowanie rozwiązania projektu będą składały się następujące elementy:

- jakość predykcji na zbiorze testowym mierzona przez *balanced accuracy*

$$BA = \frac{1}{2} \left(\frac{TP}{P} + \frac{TN}{N} \right),$$

- raport opisujący wykorzystane metody i wyniki eksperymentów (maksymalnie 4 strony A4),
- odpowiedź ustna podsumowująca rozwiązanie.

4 Szczegóły rozwiązania

Zbiór treningowy oraz etykiety do zbioru treningowego należy wykorzystać do przygotowania modelu. Oczekiwany wynik to wektor prawdopodobieństw przynależności do klasy 1 dla obserwacji ze zbioru testowego.

Rozwiązanie powinno zawierać pliki:

- `NUMERINDEKSU_artificial_prediction.txt` - prawdopodobieństwo przynależności do klasy 1 dla danych testowych, gdzie `NUMERINDEKSU` zastępujemy swoim numerem indeksu (przykładowy plik `example_artificial_prediction.txt`),
- folder `Kody` zawierający wszystkie potrzebne kody do przygotowania rozwiązania projektu,
- plik `NUMERINDEKSU_raport.pdf` opisujący wykorzystane metody i wyniki eksperymentów (maksymalnie 4 strony).

5 Ocena

Łączna liczba punktów do zdobycia jest równa 35, w tym:

- jakość kodu (porządek, czytelność, obszerność eksperymentów) - 10 punktów,
- jakość predykcji rozwiązania* - 4 punkty,
- raport - 17 punktów,
- odpowiedź ustna - 4 punkty.

* - jakość predykcji jest oceniana miarą *balanced accuracy* na zbiorze testowym. Wyniki zostaną ustawione w ranking (od najlepszego do najgorszego). Osoba z najlepszym wynikiem (najbliższym wartości 1) zyskuje 4 punktów. Osoba z najgorszym wynikiem (najbliższym wartości 0) zyskuje 2 punkty. Pozostałe wyniki zostaną przeskalowane i zaokrąglone do wartości 0.1.

6 Oddanie projektu

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NAZWISKO.IMIE.GR.projekt` (bez polskich znaków), gdzie

$$GR = \begin{cases} 1 & \text{dla środy, 12:15 (AK),} \\ 2 & \text{dla środy, 12:15 (KW),} \\ 3 & \text{dla środy, 14:15 (AK),} \\ 4 & \text{dla czwartek, 14:15 (AK).} \end{cases}$$

Tak przygotowany katalog należy przesłać na adres anna.kozak@pw.edu.pl do dnia 06.01.2026 do godziny 23:59. Prace przesłane po tym terminie będą miały odjęte 2 punkty za każdy rozpoczęty dzień spóźnienia. Tytuł wiadomości: *[WUM]/[P] Nazwisko Imię, Numer grupy: GR*.

Prace, które nie będą przygotowane w odpowiednim formacie nie będą oceniane.