Praca domowa 1

Termin oddania: 28.10.2025

1 Wstęp

Celem pracy domowej jest sprawdzenie jak poszczególne hiperparametry modelu drzewa decyzyjnego wpływają na jego jakość predykcyjną.

2 Zbiór danych

W tym celu posłużymy się zbiorem danych $\mathcal{D} = (X, y)$, gdzie X = `X.csv`, y = `y.csv`. Aby przygotować dane do dalszej pracy należy podzielić zbiór $\mathcal{D}$ na treningowy i testowy w proporcji 8:2 ustawiając parametr `random_state = NUMER_INDEKSU`. Zbiór testowy należy wykorzystać do ostatecznej oceny wybranego modelu.

3 Oczekiwany wynik

Praca domowa składa się z czterech elementów. Pierwszym będzie przygotowanie zbioru danych do dalszej pracy. Drugim będzie przygotowanie eksperymentu pozwalającego odpowiedzieć na pytanie, które hiperparametry modelu drzewa decyzyjnego są najlepsze dla zadanych danych. Kolejnym elementem będzie ocena jakości modelu, a ostatnim sprawdzenie czy wielkość próbki treningowej wpływa na jakość predykcyjną modelu.

3.1 Przygotowanie zbioru danych *(2 punkty)*

Oceń na podstawie analizy czy otrzymany zbiór danych jest dobrej jakości. Jeżeli nie jest to zaproponuj w jaki sposób można zwiększyć jego jakość.

3.2 Eksperyment *(7 punktów)*

Przygotuj eksperyment badający miarę AUC drzewa decyzyjnego na zbiorze treningowym i testowym (wykorzystując kroswalidację, co najmniej 3-krotną na danych treningowych z ustawionym parametrem `random_state = NUMER_INDEKSU`) w zależności od hiperparametrów:

- kryterium podziału (*criterion*{*"gini", "entropy"*}),
- głębokość drzewa (*max_depth*),
- minimalna liczba obserwacji w liściu (*min_samples_leaf*).

Przejrzyj dokumentację dotyczącą budowy drzewa i spróbuj znaleźć inne hiperparametry, które poprawią jakość modelu.

3.3 Analiza jakości predykcyjnej modelu *(3 punkty)*

Na podstawie wyników z Sekcji 3.2 wybierz Twoim zdaniem najlepszy model oraz podaj uzasadnienie wyboru. Dla wybranego modelu i modelu z domyślnymi hiperparametrami na danych treningowych oraz testowych wyznacz:

- macierz pomyłek,
- dokładność (ang. *accuracy, ACC*), czułość (ang. *sensitivity, recall*), precyzja (*ang. precision*),
- krzywą ROC, wartość AUC.

3.4 Wpływ rozmiaru próbki treningowej na jakość predykcyjną modelu (*3 punkty*)

Dla wybranego modelu opisanego w Sekcji 3.3 oraz dla modelu z domyślnymi hiperparametrami przeprowadź eksperyment polegający na losowym wyborze `5%, 10%, 25%, 50%, 75%, 90%, 95%` danych treningowych. Wytrenuj oba modele — z wybranymi hiperparametrami i z domyślnymi ustawieniami — oraz oceń ich zdolność predykcyjną przy użyciu miary AUC, dla zbioru treningowego i testowego. Czy rozmiar próbki danych wpływa na jakość predykcyjną Twojego wybranego modelu lub modelu z domyślnymi hiperparametrami?

4 Szczegóły rozwiązania

Rozwiązanie powinno zawierać pliki:

- plik `NUMERINDEKSU_kody.ipynb` zawierający wszystkie potrzebne kody do odtworzenia rozwiązania zadania domowego,
- plik `NUMERINDEKSU_raport.pdf` opisujący analizę danych, przeprowadzony eksperyment, analizę wybranego modelu oraz badanie wpływu próbki treningowej na jakoś predykcyjną modelu (maksymalnie 4 strony).

5 Ocena

Łączna liczba punktów do zdobycia jest równa 15, w tym:

- 3.1 Analiza danych (*2 punkty)*
  - jakość kodu (porządek, czytelność) - 1 punkt,
  - raport - 1 punkty.
- 3.2 Eksperyment (*7 punktów)*
  - jakość kodu (porządek, czytelność) - 1 punkt,
  - jakość eksperymentu - 4 punkty,
  - raport - 2 punkty.
- 3.3 Analiza jakości predykcyjnej modelu (*3 punkty)*
  - jakość kodu (porządek, czytelność) - 1 punkt,
  - wnioski - 1 punkt,
  - raport - 1 punkt.
- 3.4 Wpływ rozmiaru próbki treningowej na jakość predykcyjną modelu (*3 punkty)*
  - jakość kodu (porządek, czytelność) - 1 punkt,
  - wnioski - 1 punkt,
  - raport - 1 punkt.

6 Oddanie pracy domowej

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NAZWISKO_IMIE_GR_PD1` (bez polskich znaków), gdzie

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

Tak przygotowany katalog należy przesłać na adres *anna.kozak@pw.edu.pl* do dnia 29.10.2025 do godziny 06:00. Prace przesłane po tym terminie będą miały odjęte 2 punkty za każdy rozpoczęty dzień spóźnienia. Tytuł wiadomości: *[WUM][PD1] Nazwisko Imię, Numer grupy: GR.*
Prace, które nie będą przygotowane w odpowiednim formacie nie będą oceniane.