Kolokwium zaliczeniowe

Metody walidacji modeli statystycznych

Zadanie 1

Wczytaj zbiór danych **DIA.csv** zawierający dane dotyczące związków chemicznych, scharakteryzowanych za pomocą deskryptorów molekularnych wyznaczonych przy użyciu biblioteki RDKit. Głównym celem zadania jest klasyfikacja binarna związków chemicznych pod kątem zdolności do wywoływania polekowej choroby autoimmunologicznej (*Drug-Induced Autoimmunity*, DIA). Szczegółowy opis wszystkich zmiennych znajduje się w pliku `DIA.pdf`.
Wykonaj następujące zadania:

a) **(3 pkt)** Wykorzystując dwa wybrane algorytmy uczenia maszynowego (nie korzystaj z siatek parametrów, dobierz parametry ręcznie), przedstaw oraz zinterpretuj wyniki predykcyjne uzyskane na zbiorze treningowym oraz testowym. Zastosuj co najmniej dwie metryki klasyfikacji i uzasadnij ich wybór.

b) **(5 pkt)** Przeprowadź selekcję cech z wykorzystaniem poznanych na zajęciach metod, uzasadnij swój wybór oraz zastosowaną metodologię. Zastosuj wcześniej wybrane algorytmy uczenia maszynowego z tymi samymi parametrami i porównaj uzyskane wyniki przed oraz po selekcji cech. Następnie dla algorytmu charakteryzującego się wyższą skutecznością predykcyjną przeprowadź dostrajanie hiperparametrów z wykorzystaniem siatki parametrów na zbiorze po selekcji cech. Porównaj wszystkie uzyskane wyniki i sformułuj ostateczne wnioski.

Zadanie 2

Wczytaj zbiór danych **Boston.csv** zawierający informacje dotyczące nieruchomości na przedmieściach Bostonu. Głównym celem zadania jest regresja prognozująca cenę mieszkania. Szczegółowy opis wszystkich zmiennych znajduje się w pliku `Boston.pdf`.
Wykonaj następujące zadania:

a) **(3 pkt)** Wykorzystując dwa wybrane algorytmy uczenia maszynowego (nie korzystaj z siatek parametrów, dobierz parametry ręcznie), przedstaw oraz zinterpretuj wyniki predykcyjne na zbiorze treningowym oraz testowym. Zastosuj co najmniej dwie metryki regresji i uzasadnij ich wybór.

b) **(5 pkt)** Dla wybranych modeli przeprowadź dostrajanie hiperparametrów z wykorzystaniem dwóch rodzajów siatek parametrów. Porównaj uzyskane wyniki przed oraz po dostrajaniu, wybierz najlepszy model ze względu na wybraną metrykę, uzasadnij swój wybór i sformułuj ostateczne wnioski.

Zadanie 3

Wczytaj zbiór danych **Diabetes.csv** zawierający dane kliniczne pacjentów. Głównym celem zadania jest klasyfikacja binarna mająca na celu przewidywanie wystąpienia cukrzycy. Szczegółowy opis wszystkich zmiennych znajduje się w pliku `Diabetes.pdf`.
Wykonaj następujące zadania:

a) **(3 pkt)** Wykorzystując dwa wybrane algorytmy uczenia maszynowego (nie korzystaj z siatek parametrów, dobierz parametry ręcznie), przedstaw oraz zinterpretuj wyniki predykcyjne uzyskane na zbiorze treningowym oraz testowym. Zastosuj co najmniej dwie metryki klasyfikacji i uzasadnij ich wybór. Dodatkowo przeprowadź 5-krotną walidację krzyżową bez powtórzeń.

b) **(6 pkt)** Przeprowadź balansowanie danych z wykorzystaniem 3 poznanych na zajęciach technik, krótko opisz zasady ich działania i uzasadnij swój wybór. Dla algorytmów z poprzedniego podpunktu ponownie przeprowadź ewaluację i porównaj uzyskane wyniki przed oraz po balansowaniu. Dla algorytmu charakteryzującego się wyższą skutecznością predykcyjną

wyznacz optymalny próg odcięcia prawdopodobieństwa klasyfikacji, porównaj macierz klasyfikacji z domyślnym progiem (0,5) i skomentuj wpływ zmiany progu na wyniki predykcyjne. Następnie przeprowadź dostrajanie hiperparametrów z wykorzystaniem siatki parametrów i dla uzyskanego modelu ponownie wyznacz optymalny próg odcięcia. Porównaj wszystkie uzyskane wyniki i sformułuj ostateczne wnioski dotyczące całego zadania.

Uwaga: Ocenie podlegać będzie zarówno poprawność merytoryczna uzyskanych wyników, jak i estetyka oraz czytelność ich prezentacji.

Oddawanie pracy: Cały folder roboczy należy spakować do pliku `.zip` i przesłać za pośrednictwem platformy Microsoft Teams.

Życzę Państwu Powodzenia!!!