

Uczenie maszynowe w Python

Członkowie grupy:

1. Yahor Koval, ykoval@edu.cdv.pl
2. Tymur Popovych, t.popovych@edu.cdv.pl
3. Anton Talmachou, a.talmachou@edu.cdv.pl

Uczenie maszynowe jako narzędzie wczesnego wykrywania chorób wątroby

Wybranym problemem badawczym jest automatyczna klasyfikacja pacjentów pod kątem chorób wątroby na podstawie parametrów biochemicznych krwi. Celem projektu było stworzenie modelu uczenia maszynowego, który na podstawie danych takich jak *wiek(age)*, *płeć(gender)* czy stężenie enzymów wątrobowych (*SGOT*, *SGPT*) i *bilirubiny* (*tot_bilirubin*,*direct_bilirubin*), potrafi odróżnić osoby zdrowe od chorych (*is_patient*).

Przeprowadzone operacje i ich uzasadnienie:

1. Wstępne przygotowanie danych:

Dane załadowano ze zbioru *Indian Liver Patient Dataset (ILPD)* i przypisano im czytelne nazwy kolumn. Braki w danych, szczególnie w kolumnie *ag_ratio*, zostały usunięte lub uzupełnione wartością średnią(mean), co zapobiegło utracie istotnych informacji podczas uczenia modelu.

2. Kodowanie i transformacja:

Zmienną kategoryczną „płeć” zamieniono na wartości numeryczne (*Mężczyzna*: 1, *Kobieta*: 0), a etykietę celu *is_patient* zmapowano tak, aby wartość 1 oznaczała chorobę, a 0 jej brak. W celu zredukowania skośności rozkładu i poprawy jakości predykcji, dla cech takich jak *bilirubina* czy *enzymy wątrobowe*, zastosowano *logarytmizację* (log1p).

3. Analiza korelacji:

Wygenerowano mapę ciepła korelacji cech, co pozwoliło na identyfikację silnych zależności między parametrami, takimi jak *bilirubina całkowita(tot_bilirubin)* i bezpośrednia.

4. Budowa modelu:

Do klasyfikacji wybrano *Random Forest Classifier* z liczbą 300 drzew decyzyjnych. Zastosowanie parametru *class_weight='balanced'* było kluczowe dla zrównoważenia wpływu klas, gdyż w zbiorze danych występuje więcej przypadków osób chorych niż zdrowych.

Analiza wyników i poprawność modelu:

Analiza poprawności modelu opiera się na zestawie metryk uzyskanych na zbiorze testowym:

1. **Dokładność (Accuracy):** Model poprawnie zaklasyfikował **70,69%** wszystkich przypadków.
2. **Czułość (Recall):** Osiągnięto bardzo wysoki wynik **92,77%**, co jest najważniejszym parametrem w diagnostyce medycznej, ponieważ minimalizuje ryzyko przeoczenia osoby chorej.
3. **F1-Score:** Wynik na poziomie **81,91%** potwierdza wysoką stabilność modelu w odniesieniu do harmonii między precyzją a czułością.

Macierz błędów: Wizualizacja wyników pokazuje, że model bardzo skutecznie identyfikuje osoby chore, przy stosunkowo niskiej liczbie błędów typu „fałszywy negatyw”.

Wyniki projektu:

Finalny model został zapisany w formacie binarnym jako plik *liver_model.pickle*, co umożliwia jego ponowne wykorzystanie bez konieczności powtarzania procesu uczenia. Testy predykcyjne na profilach pacjentów wykazały, że model poprawnie ocenia ryzyko – dla pacjenta z zaburzonymi wynikami badań wskazał **95%** prawdopodobieństwo choroby, natomiast dla pacjenta o typowych parametrach wykazał status osoby zdrowej z prawdopodobieństwem **72,3%**. Projekt zakończył się stworzeniem gotowego narzędzia do wsparcia wstępnej diagnozy chorób wątroby.