



www.aileaders-project.eu

ĆWICZENIE SCENARIUSZOWE

Sprawiedliwość i stronniczość

- System wymiaru sprawiedliwości

w sprawach karnych



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

ĆWICZENIE SCENARIUSZOWE

Sprawiedliwość i stronniczość

- System wymiaru sprawiedliwości w sprawach karnych

01



Streszczenie 3

02



Wprowadzenie 3

03



Prezentacja narzędzi 4

04



Zajęcia praktyczne 6

05



Wnioski 8

06



Bibliografia 8

• 01 Streszczenie



Rodzaj OER

Ćwiczenie scenariuszowe

Cel/przeznaczenie

Analiza stronniczości algorytmu uczenia maszynowego stosowanego w systemie wymiaru sprawiedliwości w sprawach karnych

Oczekiwane efekty kształcenia

- Świadomość stronniczości w prognozach algorytmów uczenia maszynowego
- Wykrywanie stronniczości w prognozach

Słowa kluczowe

- Stronniczość
- Sprawiedliwość
- Uczenie maszynowe
- System wymiaru sprawiedliwości w sprawach karnych

Sugerowane Metodologiczne Podejście

Nauka oparta na przypadkach

• 02 Wprowadzenie



- 01** W szczególności dotyczy to stronniczości algorytmu uczenia maszynowego COMPAS stosowanego w amerykańskim systemie wymiaru sprawiedliwości w sprawach karnych.
- 02** Algorytm uczenia maszynowego COMPAS oblicza wynik ryzyka, który pomaga sędziemu podjąć decyzję o areszcie przedprocesowym i ustaleniu kaucji.
- 03** Sędzia podejmuje decyzję na podstawie oceny ryzyka, że zwolniony oskarżony nie stawi się na rozprawie lub wyrządzi szkodę społeczeństwu (recydywa).
- 04** Algorytm COMPAS oblicza wynik ryzyka, który mierzy ryzyko ponownego popełnienia przestępstwa przez oskarżonego (prognoza recydywy).
- 05** Wynik ryzyka jest wykorzystywany przez sędziego do podjęcia decyzji o areszcie przedprocesowym i kaucji.

ProPublica

- 01** ProPublica opublikowała [artykuł](#) na temat stronniczości algorytmu COMPAS.
- 02** Dane z COMPAS zostały również uzyskane i opublikowane przez ProPublica.

• 03 Prezentacja narzędzi



To ćwiczenie
scenariuszowe
wymaga znajomości:

- 01 Języka programowania Python;
- 02 Oprogramowania Jupyter notebook.

- Do wykonania tego scenariusza potrzebny jest również komputer z zainstalowanym językiem Python i oprogramowaniem Jupyter.
- Oprócz wbudowanych modułów języka Python wymagane są również biblioteki pandas, NumPy, Matplotlib i scikit-learn.

- 01 W tym scenariuszu wykorzystano treści opracowane przez [NOS](#) (portugalskiego operatora telekomunikacyjnego), które są publicznie dostępne na platformie programistycznej GitHub.
- 02 W tym scenariuszu ćwiczeniowym wykorzystamy treści [modułu SLU17 – Etyka i uczciwość](#).
- 03 Moduł ten jest częścią [kursu](#) NOS „Wprowadzenie do nauki o danych”.
- 04 Kurs ten jest z kolei częścią większej [ścieżki edukacyjnej](#) FAAST Advance Data Science.
- 05 Wszystkie te materiały są wykorzystywane przez NOS do szkolenia nowych pracowników, ale są publicznie dostępne na GitHub.
- 06 Ponadto wszystkie publiczne repozytoria NOS można obejrzeć na głównej [stronie](#) GitHub.
- 07 Wszystkie pliki wymagane do wykonania tego scenariusza znajdują się w folderze „oer_files”.
- 08 Nie ma więc potrzeby pobierania żadnych plików ze strony GitHub dla [modułu SLU17 – Etyka i uczciwość](#).
- 09 Wszystkie te pliki są już dostępne w folderze „oer_files”.
- 10 Oczywiście wszelkie zasługi za te pliki należą się NOS.
- 11 Zwracamy uwagę, że istnieją niewielkie różnice między kilkoma plikami w folderze „oer_files” a odpowiadającymi im plikami w serwisie GitHub.
- 12 W dostarczonym pliku „README.md” usunęliśmy link do pliku Google Docs, który jest dostępny tylko dla pracowników NOS. Jednak plik ten nie jest wymagany do wykonania scenariusza ćwiczenia.

W pliku „Exercise notebook.ipynb” wprowadziliśmy następujące zmiany

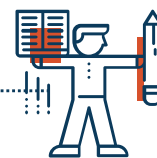
- 01** Pierwsza komórka kodu została udostępniona do edycji, a wywołanie przestarzałego stylu Matplotlib zostało usunięte;
- 02** Dodaliśmy komórkę kodu tuż poniżej obliczeń FPR dla czarnych oskarżonych, wyświetlającą „fpr_b”, aby była identyczna z tym, co zostało zrobione powyżej w przypadku białych oskarżonych;
- 03** Ostatnia komórka kodu była pusta i jako taka została usunięta.



Folder „oer_files” zawiera również plik, który nie jest obecny na stronie GitHub. Jest to plik „**Exercise notebook solved.ipynb**”.

Dla wygody i w celach informacyjnych plik ten zawiera kod / rozwiązania wszystkich ćwiczeń zawartych w pliku „Exercise notebook.ipynb”.

• 04 Ćwiczenia praktyczne



Pierwszym krokiem jest przeczytanie zawartości notatnika „Learning notebook.ipynb”.

Ten notatnik obejmuje podstawowe pojęcia dotyczące

01 elementy systemu uczenia się

02 domyślna prywatność oraz

03 stronniczość i sprawiedliwość.

- Następnie ćwiczenie scenariuszowe polega na wykonaniu ćwiczeń zawartych w notatniku „Notatnik ćwiczeń.ipynb”.
- Notatnik ten rozpoczyna się od podstawowego opisu sytuacji: stronniczości w ocenach ryzyka obliczanych przez algorytm COMPAS stosowany w amerykańskim systemie wymiaru sprawiedliwości w sprawach karnych.
- W celu uzyskania dalszych informacji zeszyt zawiera linki do [książki](#) na temat sprawiedliwości i uczenia maszynowego oraz do [artykułu](#) ProPublica na temat stronniczości maszyn w orzekaniu karnym.

ĆWICZENIE 01

- Ćwiczenie 1 polega na wykreśleniu rozkładów dla wyniku ryzyka obliczonego przez algorytm COMPAS.
- Celem jest wykreślenie ogólnego rozkładu, a także rozkładów według rasy.
- W szczególności należy wykreślić rozkłady wyników ryzyka zarówno dla oskarżonych rasy białej, jak i czarnej.

ĆWICZENIE 02

- Ćwiczenie 2 polega na wykreśleniu rozkładów wyników ryzyka uzyskanych przez klasę pozytywną (recydywistów).
- Celem jest wykreślenie ogólnego rozkładu, a także rozkładów według rasy.
- Ponownie należy przedstawić rozkłady wyników ryzyka zarówno dla białych, jak i czarnych (recydywistów) oskarżonych.

ĆWICZENIE 03

- Ćwiczenie 3 dotyczy oskarżonych, którzy zostali sklasyfikowani jako osoby o wysokim ryzyku recydywy (tj. mieli wysoką wartość wyniku ryzyka).
- Celem jest obliczenie wskaźnika fałszywych alarmów (FPR) dla tych oskarżonych wysokiego ryzyka.
- Wskaźnik FPR należy obliczyć zarówno dla oskarżonych rasy białej, jak i czarnej (wysokiego ryzyka).

Wskaźnik FPR jest również znany jako wskaźnik fałszywych alarmów lub prawdopodobieństwo fałszywego alarmu. Wskaźnik FPR został opisany w notatniku. Jednakże, ponieważ jest to kluczowy wskaźnik i klucz do zrozumienia tendencji w tym scenariuszu ćwiczeniowym, bardziej szczegółowe wyjaśnienie wskaźnika FPR znajduje się również tutaj.

Wskaznik FPR oblicza się jako

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\text{FP}}{N}$$

- 01 FP to liczba wyników fałszywie pozytywnych, czyli w naszym przypadku liczba rzeczywistych osób niebędących recydywistami, które zostały błędnie sklasyfikowane jako recydywiści;
- 02 TN to liczba prawdziwych wyników negatywnych, czyli w naszym przypadku liczba rzeczywistych osób niebędących recydywistami, które zostały prawidłowo sklasyfikowane jako osoby niebędące recydywistami;
- 03 N to całkowita liczba wyników negatywnych w rzeczywistości, czyli w naszym przypadku całkowita liczba osób, które faktycznie nie są recydywistami.

- 01 W związku z tym FPR jest wskaźnikiem rzeczywistych wyników negatywnych, które zostały błędnie sklasyfikowane jako pozytywne.
- 02 W naszym kontekście FPR to odsetek rzeczywistych osób, które nie popełniły ponownego przestępstwa, a zostały błędnie sklasyfikowane jako recydywiści.
- 03 Wyższe wartości są zatem mniej pożądane, ponieważ wyższy wskaźnik FPR oznacza, że większa część rzeczywistych osób niebędących recydywistami została błędnie sklasyfikowana jako recydywiści.
- 04 Jak wspomniano, celem jest napisanie kodu do wykonania wszystkich ćwiczeń zawartych w notatniku „**Exercise notebook.ipynb**”.
- 05 Jak wspomniano wcześniej, dla wygody i w celach informacyjnych udostępniamy również notatnik, który jest już wypełniony odpowiednim kodem („**Exercise notebook solved.ipynb**”).

• 05 Wnioski



To ćwiczenie scenariuszowe pokazuje, że prognozy algorytmów uczenia maszynowego mogą być stronnicze. Wyraźnie widać to na przykładzie wyraźnej różnicy w odsetku wyników fałszywie pozytywnych między oskarżonymi rasy białej i czarnej. Stronniczość algorytmów może więc mieć poważny wpływ na osoby lub grupy, których dotyczą decyzje oparte na tych algorytmach.

• 06 Referencje



- [Fairness and Machine Learning – Book](#)
- [ProPublica Article on Machine Bias in Criminal Sentencing](#)
- [NOS Ethics and Fairness Learning Unit on GitHub](#)
- [NOS Intro to Data Science Course on GitHub](#)
- [NOS FAAST Learning Path on GitHub](#)
- [Main NOS Page on GitHub](#)



Śledź naszą podróż



www.aileaders-project.eu



Co-funded by
the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.