Akademia Ekonomiczno-Humanistyczna w Warszawie

SPRAWOZDANIE

INTELIGENTNA ANALIZA DANYCH

**LAB1**

**KWANTYFIKATORY**

23.10.2021

JOANNA PRAJZENDANC

36358

MIŁOSZ SAKOWSKI

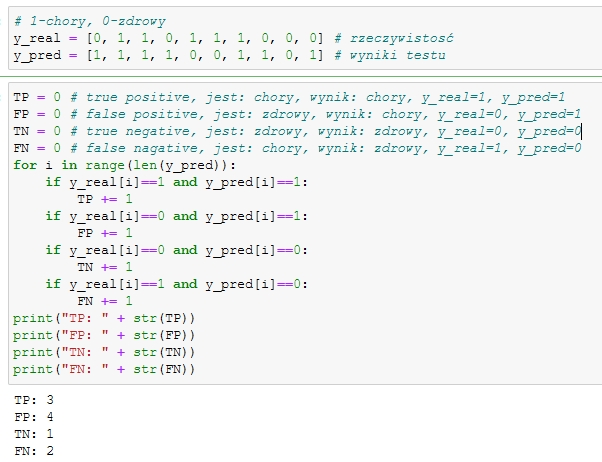
36381

# Cel ćwiczenia

Celem ćwiczenia było zapoznanie się z nowymi pojęciami używanymi podczas analizy danych: klasyfikatory, metryki, macierz pomyłek (tablica pomyłek, ang. *confusion matrix*), czułość (ang. *sensitivity, recall*), swoistość (ang. *specificity, SPC*), precyzja (ang. *precision*) i dokładność (ang. *accuracy, ACC*) oraz F1-score.

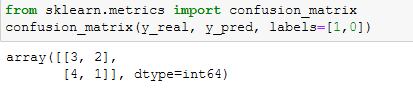
W pierwszej kolejności dokonano analizy kodu, w którym zostały użyte te pojęcia, następnie należało wykonać dwa zadania w celu utrwalenia nowych definicji.

# Analiza kodu i wyjaśnienie pojęć



Obraz 1: Fragment kodu z wynikami TP, FP, TN, FN

Powyższy zrzut kodu przedstawia zaproponowane 10 obserwacji oraz 10 predykcji na podstawie których możemy stwierdzić wyniki TP, FP, TN, FN i utworzyć macierz pomyłek.



Obraz 2: Fragment kodu z macierzą pomyłek

Przedstawiony zrzut ekranu przedstawia **macierz pomyłek**, która pozwala dokonać oceny jakości klasyfikacji binarnej. W takiej tablicy dane są oznaczone etykietami: „pozytywną”   
(TP - *true positive*, TN - *true negative*) oraz „negatywną” (FP - *false positive*, FN - *false negative*).



Obraz 3: Schemat tworzenia macierzy pomyłek. Źródło: Wikipedia

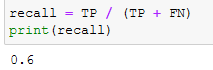
Obliczając sumę elementów tego samego typu można zauważyć, że wartości „negatywne” (FP + FN) mają wyższą sumę

2 + 4 = 6

niż wartości „pozytywne” (TP + TN)

3 + 1 = 4

Licząc teraz **czułość** (odsetek prawdziwie pozytywnych):



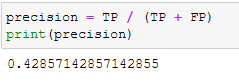
Obraz 4: Fragment kodu z wynikiem czułości

Czułość określa, jaki jest udział prawidłowo prognozowanych przypadków pozytywnych (TP) ze wszystkich przypadków pozytywnych. Powinna dążyć do 1.

Oraz **swoistość** (odsetek prawdziwie negatywnych):

Swoistość bada, jaki jest udział prawidłowo przewidzianych przypadków negatywnych (TN) ze wszystkich przypadków negatywnych. Powinna dążyć do 1.

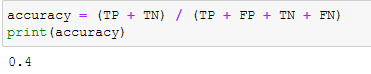
**Precyzję**:



Obraz 5: Fragment kodu z wynikiem precyzji

Precyzja wskazuje, że wartość oznaczona jako dodatnia jest rzeczywiście dodatnia. Im wyższa precyzja tym mniejsza ilość FP.

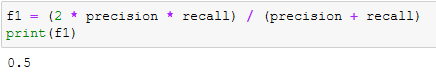
**Dokładność**:



Obraz 6: Fragment kodu z wynikiem obliczeń dokładności

Widzimy, że predykcja jest niewystarczająca, ma zbyt małą dokładność. Dokładność, to wskaźnik który pozwala ocenić jakość klasyfikacji testu i powinna dążyć do 1. Z obliczeń możemy dowiedzieć się jaka część testów ze wszystkich zaklasyfikowanych, została zaklasyfikowana poprawnie.

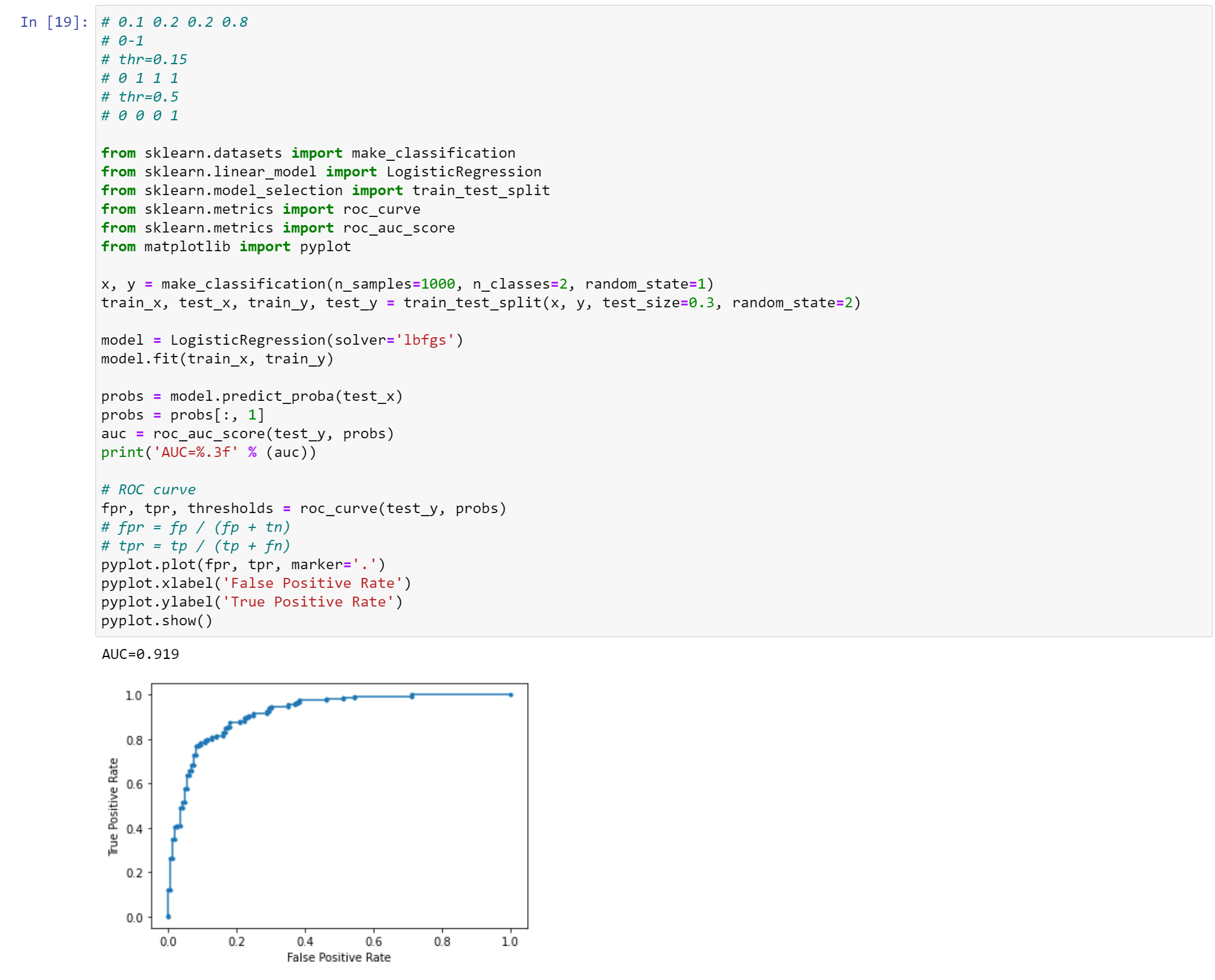
Możemy również policzyć średnią skuteczność modelu w tym zbiorze danych,   
tzw. **F1-score**:



Obraz 7: Fragment kodu z wynikiem wskaźnika F1-score

Im bliżej wartości 1, tym lepiej to świadczy o algorytmie klasyfikującym. Wskaźnik F1-score przyjmuje wartość 1, kiedy mamy do czynienia z idealną czułością i precyzją.

Na koniec zaprezentowany został przykład sztucznej inteligencji, która na podstawie modelu rozkładu normalnego uzyskała dokładność 0.919.



Obraz 8: Fragment kodu z przykładem sztucznej inteligencji rozwiązującej ten sam problem

# Zadanie nr 1

**Zadanie:**

Prosze wygenerowac 10 obserwacji i 10 predykcji (y\_real i y\_pred), dla danych obliczyć:

a) macierz pomyłek

b) accuracy

c) recall

d) precision

e) f1

Dla każdej z pięciu powyższych metryk proszę wytłumaczyć jej znaczenie.

**Rozwiązanie:**

# Zadanie nr 2

**Zadanie:**

Jesteśmy programistami sterowników do samochodów z systemem automatycznego sterowania. Nasz system ma na celu jak najbezpieczniejszą jazdę. Po otrzymaniu sygnału z kamery musimy sklasyfikować, czy obraz przedstawia obiekt który należy ominąć (np. pies) czy inny, niegroźny obiekt (np. cień). Przyjmijmy, że obiekt do ominięcia reprezentuje klasę pozytywną. Proszę opisać w 5 zdaniach czym powinien charakteryzować się nasz system (np. wysokie/niskie TP,FP,TN,FN) i dlaczego?

**Odpowiedź:**

# Zadanie nr 3

**Zadanie:**

Mamy dwie macierze pomyłek:

a) 9 1

0 4

b) 4 6

3 1

Która macierz pomyłek reprezentuje lepszy model i dlaczego?

**Odpowiedź:**

Macierz a) reprezentuje lepszy model, ponieważ suma TP i TN jest większa od sumy FP + FN, czyli suma jest wyższa dla wartości sklasyfikowanych jako przypadki pozytywne niż dla przypadków sklasyfikowanych jako negatywne.