# Podstawy uczenia maszynowego
## Kolejny rzut oka na pandemię – raport

Mateusz Kocot

22 maja 2021

## Spis treści



# 1 Przygotowanie danych

Przygotowanie danych podzielono na kilka kroków.

1. Zmodyfikowano wartości w kolumnie „If a vaccine to prevent COVID-19 was offered to you today, would you choose to be vaccinated?". Wybrano wariant, w którym tylko „Yes, Defintely" zmieniane jest na „Yes". Pozostałe wartości, tj. „Yes, Probably", „No, Probably Not" oraz „No, Definitely Not" zmienione zostały na „No".
2. Usunięto kolumny dotyczące rekomendacji oraz kolumnę z odpowiedziami na pytanie o obawy dotyczące skutków ubocznych.
3. Usunięto kolumny zawierające tylko jedną wartość. Usunięto także kolumnę „Timestamp".
4. Zauważono, że dane dotyczące wieku nie są spójne. Dwie osoby udzieliły odpowiedzi „No" na pytanie o pełnoletność oraz zaznaczyły „18-25" jako „AGE BAND". Dodano więc nową kategorię w kolumnie „AGE BAND" – „<18" i przypisano ją dla tych dwóch wierszy.
5. Dla ułatwienia przetransformowano zbiór na małe litery.
6. W zbiorze danych kilka razy pojawiła się wartość „nan". Dla ułatwienia zastąpiono ją najczęstszymi odpowiedziami z odpowiednich kolumn.

7. By skorzystać z funkcjonalności biblioteki *scikit-learn*, kategorie zastąpiono liczbami całkowitymi. Odpowiednie przekształcenia zapisano w słowniku.
8. Finalnie rozdzielono zbiór na zbiór właściwy oraz etykiety.

# 2 Metryka

Na potrzeby klasyfikatora k-NN stworzono metrykę, która uwzględnia każdą cechę z osobna. Wartości są odpowiednio rzutowane na przedział $[0, 1]$, a następnie obliczana jest wartość absolutna różnicy. Wartości dla każdej kolumny są sumowane, a suma zwracana jest jako odległość dwóch obserwacji. Domyślnie, waga dla wyniku z każdej kolumny wynosi 1. Przygotowano także drugi zestaw wag. Premiuje on symptomy, np. „Eye pain" (waga: 4) czy „Difficulty in Breathing" (waga: 6). Większe wagi dostały także kolumny dotyczące kontaktów z innymi ludźmi lub wychodzenia z domu w ciągu ostatnich 24 godzin. Zmniejszono wagę kolumny z miastami. Największą wagę, tj. 10, dostały przedziały wiekowe.

# 3 Eksperymenty

Przyjęto, że *positive* oznacza brak chęci szczepienia. Modele k-NN będą działać dla $k = 5$, a modele Random Forest wyposażone zostaną w 500 drzew. Wszystkie eksperymenty zostaną przeprowadzone 20 razy.

W sumie zostaną wykorzystane cztery modele – po dwa warianty k-NN i losowych lasów:

1. k-NN z metryką z domyślnymi wagami (k-NN (1)),
2. k-NN z metryką z drugim zestawem wag (k-NN (2)),
3. Random Forest – domyślny (RF (1)),
4. Random Forest, gdzie wszystkie drzewa uczą się na wszystkich cechach oraz wyłączony jest bootstrap (RF (2)).

## 3.1 Pierwsze wyniki

Na początku przetestowano klasyfikatory w domyślnej konfiguracji głosowania. Testy wykonano z użyciem *5-fold cross-validation*. Wyniki po przeprowadzeniu walidacji krzyżowej 20 razy zaprezentowano w tab. 1.

| **Model** | **accuracy** [%] | **precision** [%] | **recall** [%] |
|---|---|---|---|
| **K-NN (1)** | $(51.2 \pm 6.1) \pm 2.2$ | $(57.1 \pm 8.8) \pm 1.8$ | $(64.4 \pm 8.9) \pm 2.6$ |
| **K-NN (2)** | $(51.8 \pm 6.4) \pm 2.1$ | $(58.1 \pm 9.1) \pm 1.9$ | $(62.4 \pm 8.8) \pm 3.2$ |
| **RF (1)** | $(56.1 \pm 6.2) \pm 2.0$ | $(61.0 \pm 8.0) \pm 1.8$ | $(68.6 \pm 8.6) \pm 2.8$ |
| **RF (2)** | $(52.6 \pm 6.6) \pm 2.2$ | $(59.4 \pm 9.2) \pm 2.1$ | $(58.1 \pm 8.6) \pm 2.9$ |

**Tab. 1:** Wyniki uzyskane przez klasyfikatory w domyślnej konfiguracji głosowania

Wyniki nie są zbyt satysfakcjonujące. Szczególnie martwią duże odchylenia, niemniej przy tak małej liczbie obserwacji, można się było tego spodziewać.

Najlepiej sprawuje się las losowy w domyślnej konfiguracji. Widać także, że dzięki zmianie wag uzyskano nieco lepsze średnie wartości. Jednakże, duże niepewności sprawiają, że ciężko wyciągnąć racjonalne wnioski.

## 3.2 ROC

Najpierw ustalono sposób wyznaczania progów czułości dla obu metod.

1. K-NN – progi definiowane są przez minimalną liczbę sąsiadów *positive* potrzebnych do zakwalifikowania obserwacji jako *positive*.
2. Random Forest – metoda ta zaimplementowana jest w pakiecie *scikit-learn* w taki sposób, że każde drzewo zwraca prawdopodobieństwo przynależności obserwacji do klas. Brane są średnie prawdopodobieństw ze wszystkich drzew i wybierana jest klasa, której średnia jest prawdopodobieństw jest najwyższa. W związku z tym, progi, które zostały wykorzystane w tym zadaniu, bazują na tymże prawdopodobieństwie. Progi definiowane są przez minimalną wartość prawdopodobieństwa klasy *positive*. Wzięte zostały wartości od 0, do 1, postępujące co 0.1

Wizualizację krzywych ROC przedstawiono na rys. 1.

Krzywe dla K-NN prawie pokrywają się z dorysowaną na pomarańczowo prostą. Oznacza to, że równie dobrze można by wziąć klasyfikator losowy. Nieco lepiej sytuacja wygląda w przypadku RF. Oba warianty znajdują się nieco nad pomarańczową prostą. Więcej kontroli nad *precision* i *recall* daje wariant pierwszy. Dla wszystkich wariantów, ponownie można zaobserwować duże odchylenia, co jest spowodowane małą liczbą danych i dużą zależnością jakości klasyfikatora od permutacji danych przy podziale na zbiór treningowy i testowy.

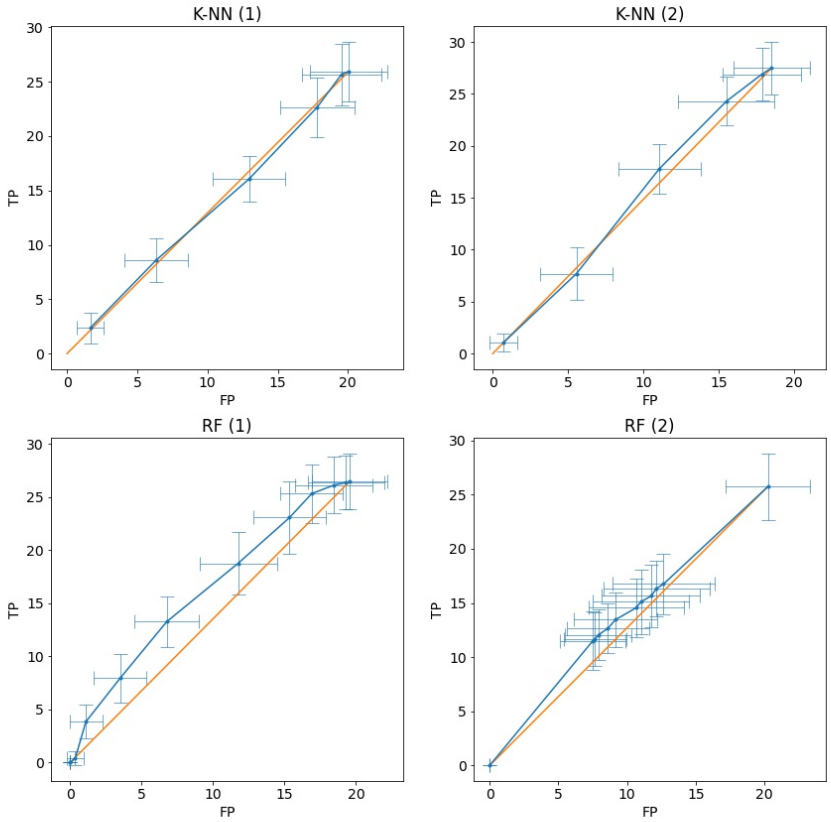
Wybrano najbardziej obiecujące punkty:

- K-NN (1): punkt 5. od prawej strony,
- K-NN (2): punkt 4. od prawej strony,
- RF (1): punkt 7. od prawej strony,
- RF (2): punkt 7. od prawej strony,

Współczynniki *accuracy*, *precision* i *recall* przedstawiono w tab. 2

| **Model** | **accuracy** [%] | **precision** [%] | **recall** [%] |
|---|---|---|---|
| **K-NN (1)** | $48.5 \pm 5.6$ | $58.2 \pm 8.6$ | $33.6 \pm 8.7$ |
| **K-NN (2)** | $54.9 \pm 5.7$ | $62.0 \pm 5.9$ | $65.4 \pm 10.8$ |
| **RF (1)** | $56.6 \pm 5.7$ | $66.6 \pm 8.5$ | $50.4 \pm 9.2$ |
| **RF (2)** | $53.4 \pm 7.5$ | $59.9 \pm 11.2$ | $52.3 \pm 8.0$ |

**Tab. 2:** Wyniki uzyskane przez klasyfikatory dla konfiguracji odpowiadających wybranym punktom z krzywych ROC




**Rys. 1:** Na niebiesko: Krzywe ROC dla poszczególnych metod, na pomarańczowo: prosta przechodząca przez punkty: $(0, 0)$ oraz $(max(FP), max(TP))$ (dodana dla porównania jakości poszczególnych metod)

Tym razem, pierwszy wariant RF uzyskał najlepszy dotąd wynik *precision* (oczywiście ze sporym odchyleniem, co sprawia, że wyciągnięte wnioski mogą minąć się z prawdą). Jednakże, wariant ten uzyskał także bardzo niską wartość współczynnika *recall*. Jest to spowodowane tym, że odpowiadający temu wariantowi punkt znajduje się blisko lewej strony w porównaniu do punktów pozostałych (oprócz K-NN (1), który znajduje się jeszcze bliżej lewej strony).

## 3.3 Najbardziej znaczące cechy w losowym lesie

Z wykorzystaniem pola *feature_importances_* znaleziono znaczenia poszczególnych cech w pierwszym wariancie RF (*Gini importance*).

1. Najważniejszą cechą (15.4%) okazała się być kolumna „For how many days have you had at least one of these symptoms?". Wydaje się to być racjonalne – w końcu ludzie

ciężej przechodzący choroby zapewne bardziej boją się zachorowania na COVID, co sprawia, że są bardziej chętni na szczepienia. W tym przypadku jednak tak wysoka pozycja tej kolumny najprawdopodobniej spowodowana jest nierzetelnością ankieterów. Kolumna ta zawiera kilkadziesiąt unikalnych wartości, które bardzo często się pokrywają. Przez to, często jedna wartość występuje w kolumnie tylko raz albo kilka razy, a w takim przypadku łatwo stworzyć drzewo odpowiednio klasyfikujące ten przypadek.

2. Na drugim miejscu znajduje się kolumna „CITY". Jest to nieco zaskakujące. Przy tak małej liczbie obserwacji ciężko jednak stwierdzić, czy podobnie jak w poprzednim przypadku jest to spowodowane 10 różnymi wartościami, czy rzeczywiście istnieje jakaś korelacja.

3. Na następnych pozycjach znajdują się już pozycje, które da się uzasadnić, np. kolumna z odpowiedziami na pytanie o znajomości osób mających objawy wskazujące na COVID albo unikanie kontaktu z innymi.

4. Spory wpływ na wybór ma też płeć. Co ciekawe, wszystkie osoby, które nie chciały wskazać swojej płci („prefer not to say") nie chcą się szczepić. Najprawdopodobniej ta zależność jest powodem sporego znaczenia tej cechy.

5. Co ciekawe, pytania dotyczące konkretnych objawów (katar, ból, katar, itd.) nie mają dużego znaczenia. Niepotrzebnie więc przyznano im wcześniej większe wagi (wykorzystane w metryce z drugiego wariantu K-NN).