Kolejny rzut oka na pandemię

Dane

	Timestamp	Country	STATE	CITY	Are you above 18 Years of Age.	EYE PAIN	CHEST PAIN	SOAR
0	4/6/2021 0:48:07	INDIA	TELANGANA	HYDERABAD	Yes	YES	Yes	No
1	4/6/2021 15:31:21	INDIA	TELANGANA	HYDERABAD	Yes	YES	Yes	No
2	4/19/2021 12:06:03	INDIA	TELANGANA	HYDERABAD	Yes	NO	No	No
3	4/19/2021 12:29:41	INDIA	TELANGANA	HYDERABAD	Yes	NO	No	No
4	4/19/2021 12:34:24	INDIA	TELANGANA	OTHERS	Yes	NO	No	No

W ramach przygotowania danych usunąłem kolumny: - 'Timestamp',

- 'CITY',
- 'Country',
- 'STATE',
- 'Are you above 18 Years of Age.'
- 'Would you be more likely or less likely to have a COVID-19 vaccination if it was recommended to you by each' 'of the following: [WHO]'
- 'Would you be more likely or less likely to have a COVID-19 vaccination if it was recommended to you by each' 'of the following: [Politicians]'
- 'Would you be more likely or less likely to have a COVID-19 vaccination if it was recommended to you by each' 'of the following: [Government Health Officials]'
- 'Would you be more likely or less likely to have a COVID-19 vaccination if it was recommended to you by each' 'of the following: [Doctors & Healthcare Staff]'
- 'Would you be more likely or less likely to have a COVID-19 vaccination if it was recommended to you by each' 'of the following: [Friends and Fa,ily]'
- 'How concerned are you that you would experience a side effect from a COVID-19 vaccination?'

Dane z kolumny "If a vaccine to prevent COVID-19 was offered to you today, would you choose to be vaccinated?" zmpaowałem do zbioru {"Yes", "No"} - "Yes, Probably" i "Yes, Definitely" do "Yes" pozostałe do "No".

Otrzymane wartości zmapowałem do wartości numerycznych: - dla kolumn które miały "Yes" albo "No" do 1.0, 0.0

- dla kolumn, które miały kilka wartości do liniowej skali, na przykład:

```
freq_dict = {
    'Always': 4.0,
    'Often': 3.0,
    'Sometimes': 2.0,
    'Rarely': 1.0,
    'Never': 0.0
}
```

- została jeszcze jedna kolumna która zawierała kilka wymienionych chorób po przecinku, tutaj zastosowałem `one-hot encoding`

you access to Sanitizer/Hand wash at Home	AGE BAND	ASTHAMA	HIGH BLOOD PRESSURE	OBESITY	WEEKEND IMMUNE SYSTEM
1	1	0	1	1	1
1	1	1	1	0	0
1	2	0	0	0	0
0	0	0	0	0	0

Klasyfikatory

Wykorzystałem następujące klasyfikatory:

- kNN - biorący pod uwagę wszystkie cechy, k = 5
- kNN Subset - biorący pod uwagę tylko pewien podzbiór cech - wszelakie symptompy.
- RF - Random Forest - domyślny ze sklearna.
- RF All Traits - Random Forest - bez bootstrapowania i bez losowania cech na którym uczą się drzewa.

Można pomyśleć, że wszystkie drzewa będą takie same ale jest jeszcze jedno źródło wariancji, czyli permutacja features przy każdym splicie.

Diagnostyka

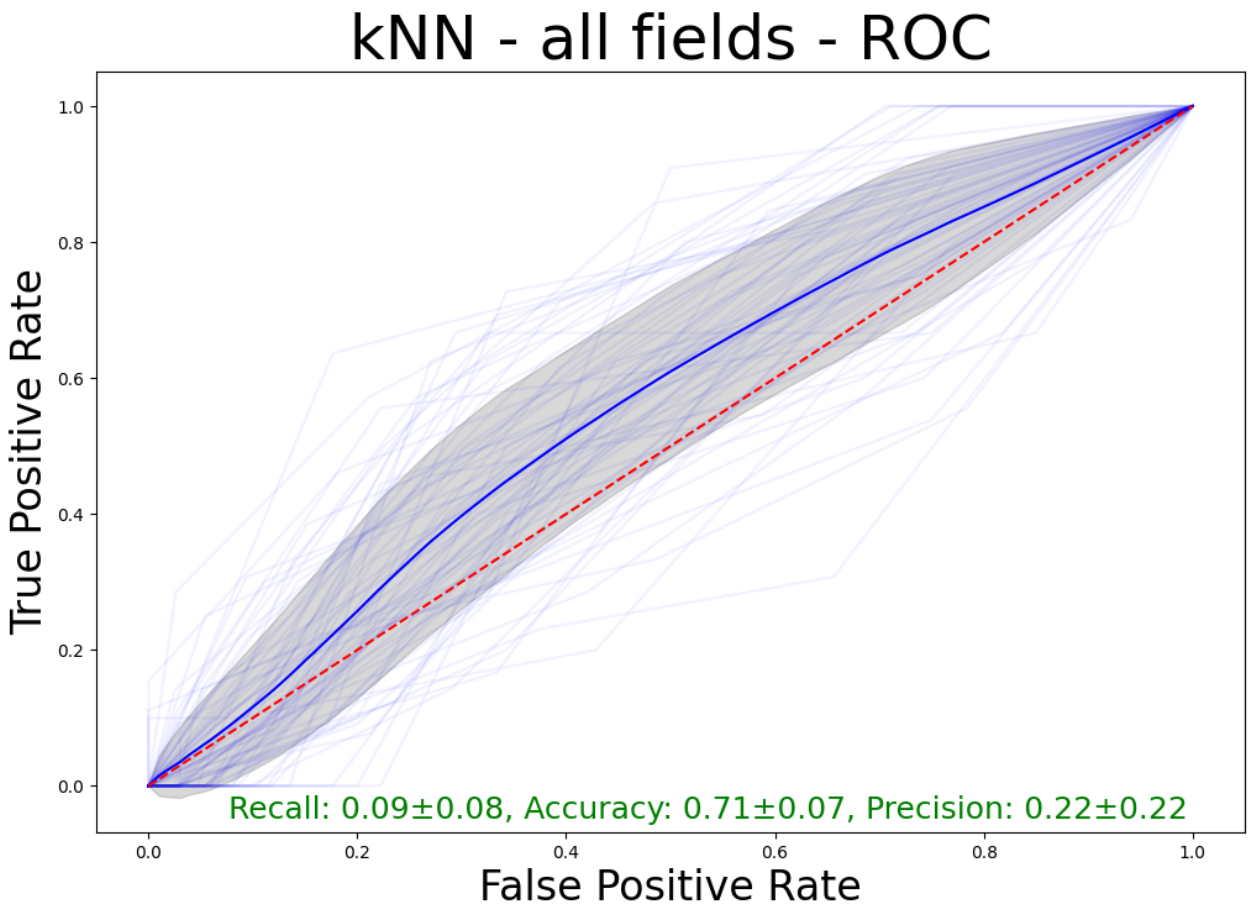
Do podziału danych na część treningową i testową wykorzystałem 5-fold cross validation, i powtarzałem 10 razy.

Jako positive uznaję brak chęci na przyjęcie szczepienia.

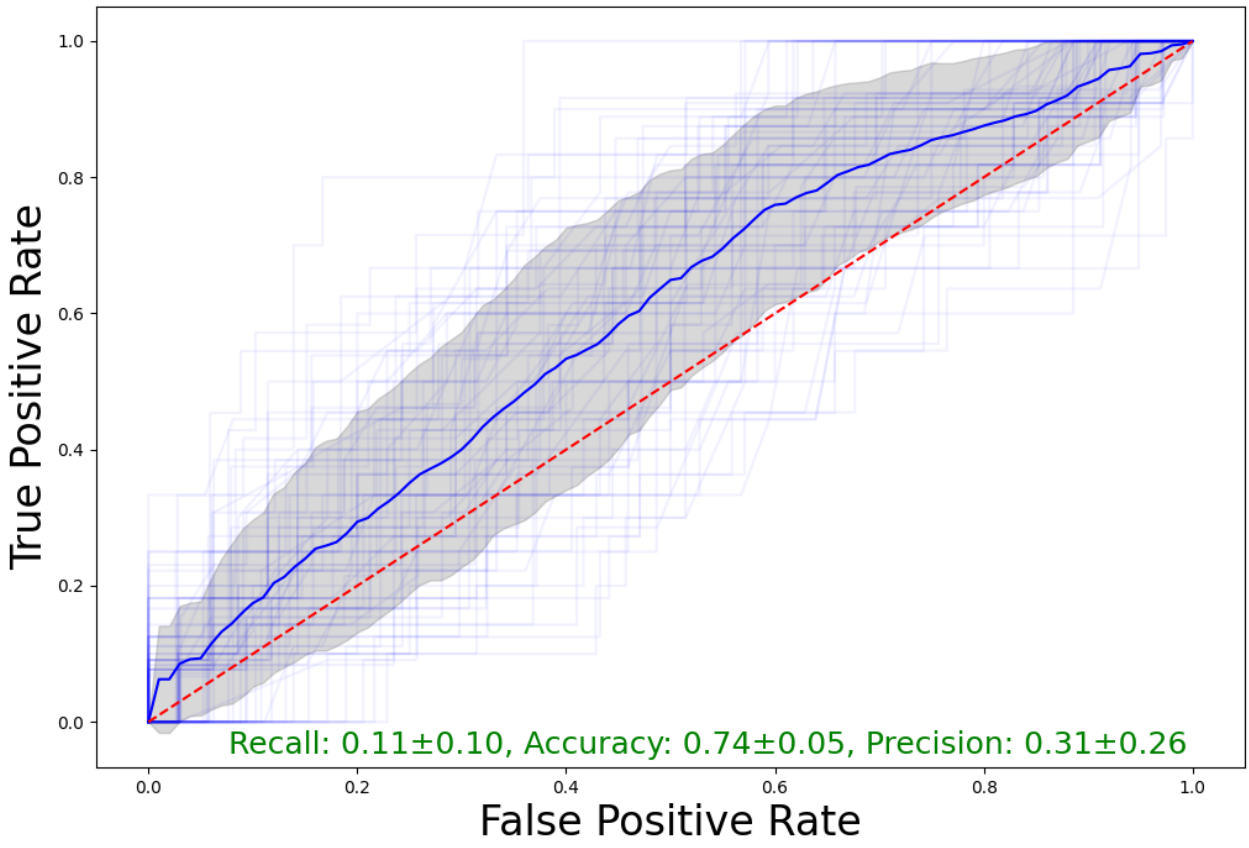
Recall, accuracy, precision policzyłem dla progu czułości = 50% (default).

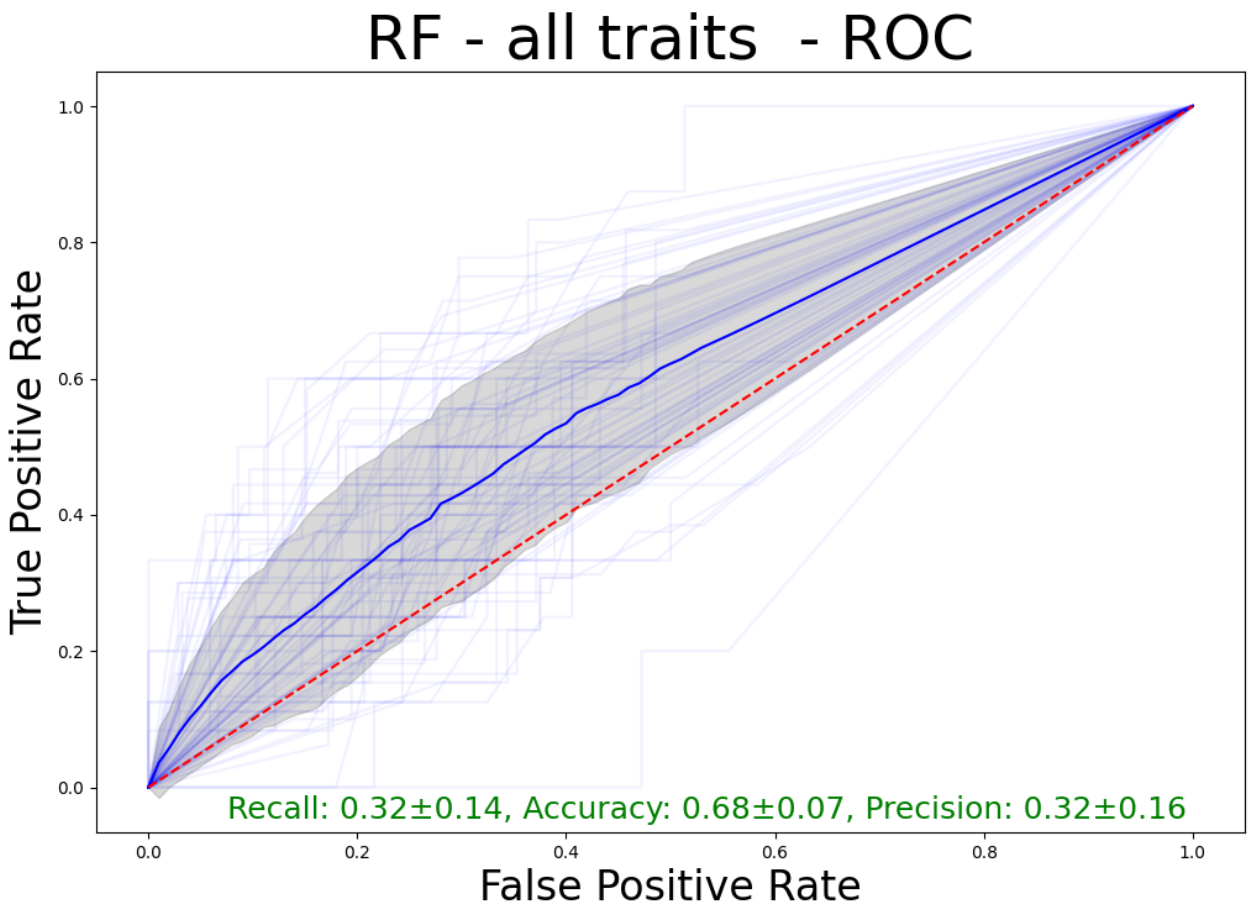
Klasyfikatory radziły sobie raczej przeciętnie z problemem, uzyskanie accuracy na poziomie 70% to był dobry wynik.

Ich krzywe ROC są bliższe tej krzywej losowego klasyfikatora niż idealnego. RF miał większą wariancję niż kNN ale miał też lepsze wyniki - choć i tak przeciętne bo precision (czyli jaka część egzemplarzy zaklasyfikowanych jako pozytywne faktycznie były pozytywne) na poziomie .3 to słaby wynik.




kNN - subset - ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Recall: 0.13±0.15, Accuracy: 0.73±0.08, Precision: 0.27±0.25


RF - default - ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Recall: 0.11±0.10, Accuracy: 0.74±0.05, Precision: 0.31±0.26



Ważność Cech

5 najważniejszych cech dla klasyfikatora Random Forest. Właściwie to 3 są stabilne, pozostałe zmieniają się.

Cechy, które mają największe znaczenie wydają się być rozsądne:

- unikanie kontaku z ludźmi,
- czy w ciągu 24h ktoś robił coś w co zamieszani byli inni ludzie,
- czy ktoś używa maski na zewnątrz.