Mikołaj Błaszczyk

Klasyfikacja na zbiorach nieregularnych na przykładzie statystyk infekcji wirusem CoViD19

Spis treści

Co to są zbiory nieregularne?

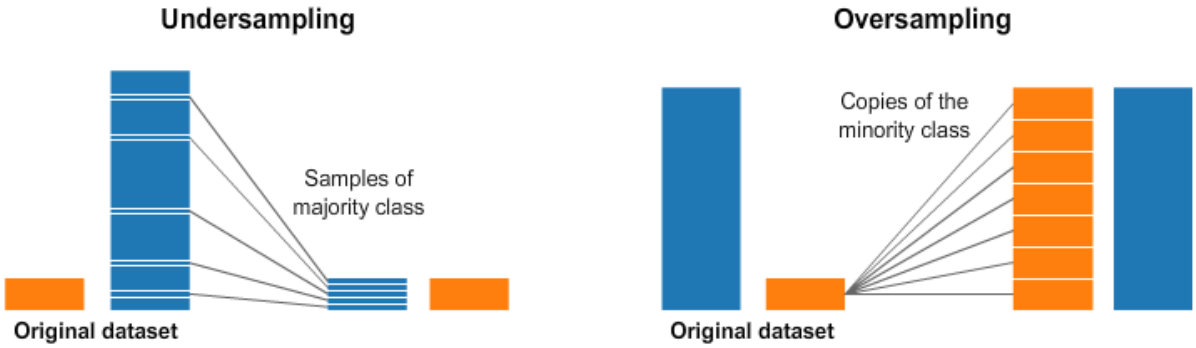
Często gdy chcemy sklasyfikować pewien zbiór danych spotykamy się z sytuacją gdzie komórki nie są klasyfikowane po równo w stosunku 1:1. Występuje to w szczególności w sytuacji gdzie próbujemy zobaczyć jakąś anomalię np. przekręty bankowe w tabeli miliona transakcji, czy ilość ludzi, u których wykryto raka spośród wszystkich badanych. Są to sytuacje gdzie często poszukiwana przez nas klasa (w powyższych przykładach kolejno: przekręt bakowy i wystąpienie raka) pojawia się w mniej niż 10% danych.

W takich sytuacjach zwykłe sposoby klasyfikacji zwyczajnie nie przynoszą wymiernych rezultatów. Najczęściej wynik będzie taki, że dominująca klasa ma prawie 100% wykrywalności, a nasza anomalia mniej niż 40%. Jest to szczególnie ważne chociażby w przytoczonym powyżej przykładzie wykrywania raka. Przecież nie chodzi nam o to by zobaczyć czy ktoś nie posiada komórek raka tylko o to by wykryć je u jak największej liczbie badanych ludzi.

Sposoby radzenia sobie z nieregularnymi danymi

Cały czas możemy usłyszeć o coraz to nowych algorytmach radzących sobie z tym problemem, jednak chciałbym tutaj opisać te najczęściej używane, a co za tym idzie takie, które są już stosowane[i] i przynoszą bardzo dobre rezultaty.

Jednym z najbardziej podstawowych podejść do problemu jest tzw. ***preprocessing*** danych. Polega on na tworzeniu takiego środowiska w którym obie klasy występują w proporcji 1:1, a działa on poprzez zwiększanie liczby danych rzadkich, zmniejszanie liczby danych często występujących lub kombinacje obu tych czynności.



Rysunek 1 - wyjaśnienie dwóch metod preprocessingu danych: undersamplingu i oversamplingu

W moich badaniach użyłem zarówno ***undersamplingu*** jak i ***oversamplingu***.

Po takim procesie najczęściej wykorzystujemy metodę klasyfikacji danych według naszego uznania, czyli taką, która będzie dla naszych danych działać w najlepszy i najbardziej optymalny sposób. W moich badaniach po oversamplingu danych użyłem znanej w statystyce metody **regresji logistycznej**. Regresja opiera się na liczeniu prawdopodobieństwa występowania którejś z dwóch klas za pomocą sposobu *szansy,* jest to stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki.

Innym podejściem są tzw. **metody zespołowe (eng. *ensemble methods*)**. Techniki tę najczęściej łączą wiele podstawowych modeli w celu stworzenia jak

najbardziej optymalnego modelu predykacyjnego. Najczęściej metody te wybierają jak najlepsze drzewo decyzyjne spośród puli wcześniej testowanych modeli.

Wybrane metody zespołowe użyte w moich badaniach:

- **Bagging** – (skrót od ang. **b**ootstrap **agg**regat**ing**) tworzy klasyfikatory na podstawie losowo wybranych zbiorów danych, następnie poszczególne klasyfikatory są łączone i podejmują decyzję na podstawie demokratycznej większości głosów.
- **Las losowy** – ang. random forest, jest ulepszeniem metody baggingu. Jego główną zaletą jest to, że utworzone pod-drzewa mają ze sobą mniej wspólnych cech. Działa to dzięki małej zmianie, która narzuca algorytmowi określoną limitowaną próbkę cech do przeszukiwania (w przeciwieństwie do baggingu gdzie algorytm przeszukuje przez wszystkie cechy).
- **Boosting** – w przeciwieństwie do baggingu korzysta z całego zbioru danych do trenowania każdego klasyfikatora, jednocześnie dając odpowiednią wagę przypadkom, w których nastąpiła błędna klasyfikacja, aby kolejne klasyfikatory zwracały bardziej na nie uwagę.

Środowisko badań

Wszystkie testy zostały przeprowadzone wykorzystując język programowania *python*, natomiast normalizacja tabeli danych nastąpiła za pomocą skryptu napisanego w języku *R*.

Dwie najistotniejsze biblioteki, które użyłem w pythonie to:

- ***Imbalanced-learn*** [ii]– biblioteka stworzona w celu ułatwienia klasyfikowania zbiorów nieregularnych
- ***Sklearn*** [iii]– biblioteka zawierająca wiele metod potrzebnych do klasyfikowania danych oraz ich przygotowywania pod różne algorytmy

Kolejne pomocnicze biblioteki to:

- ***Matplot***- biblioteka do rysowania wykresów matematycznych
- ***Pandas*** – biblioteka do analizy danych
- ***Numpy*** – biblioteka do obliczeń matematycznych

Środowisko programistyczne: **Visual Studio 2019** z rozszerzeniem do programowania w pythonie.

Baza danych

W celach badawczych postanowiłem wykorzystać współczesną statystykę zachorowań i śmierci z powodu nowego wirusa CoViD19.

Zbiór danych zawiera informacje o ludziach zarażonych tym wirusem, a także informacje o ich stanie. Niestety sama baza danych nie jest idealna – informacje nie są aktualizowane biorąc pod uwagę stan współczesny i nie posiada wielu szczegółowych informacji medycznych, które zwyczajnie ciężko otrzymać od tak wielu placówek hospitalizujących na świecie. Niemniej jednak myślę, że nawet używając tak szczątkowych informacji możemy zobaczyć ciekawe rezultaty.

Dane o pacjentach zawierają kilka kluczowych informacji:

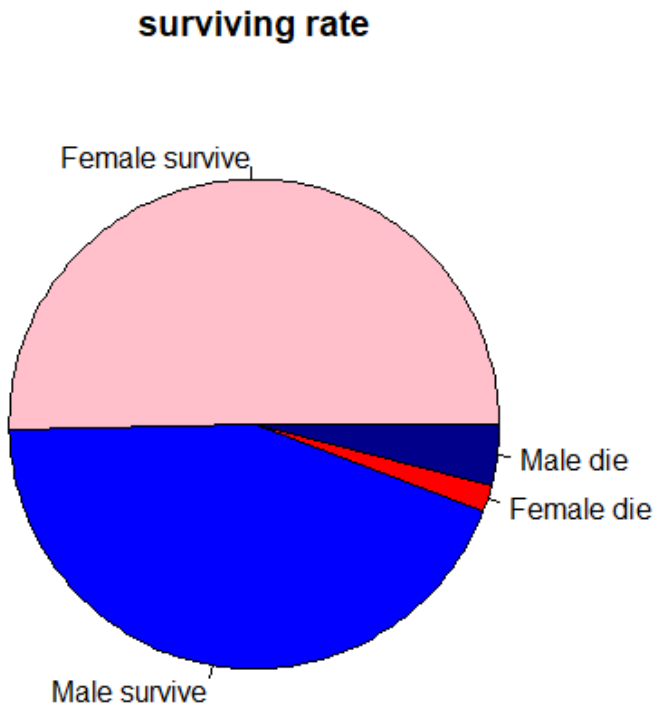
- Kraj
- Płeć
- Wiek
- Czy dana osoba odwiedzała centrum epidemii
- Czy dana osoba pochodziła z centrum epidemii
- Stan danej osoby (śmierć, uzdrowienie)

Źródło badanej bazy danych: https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

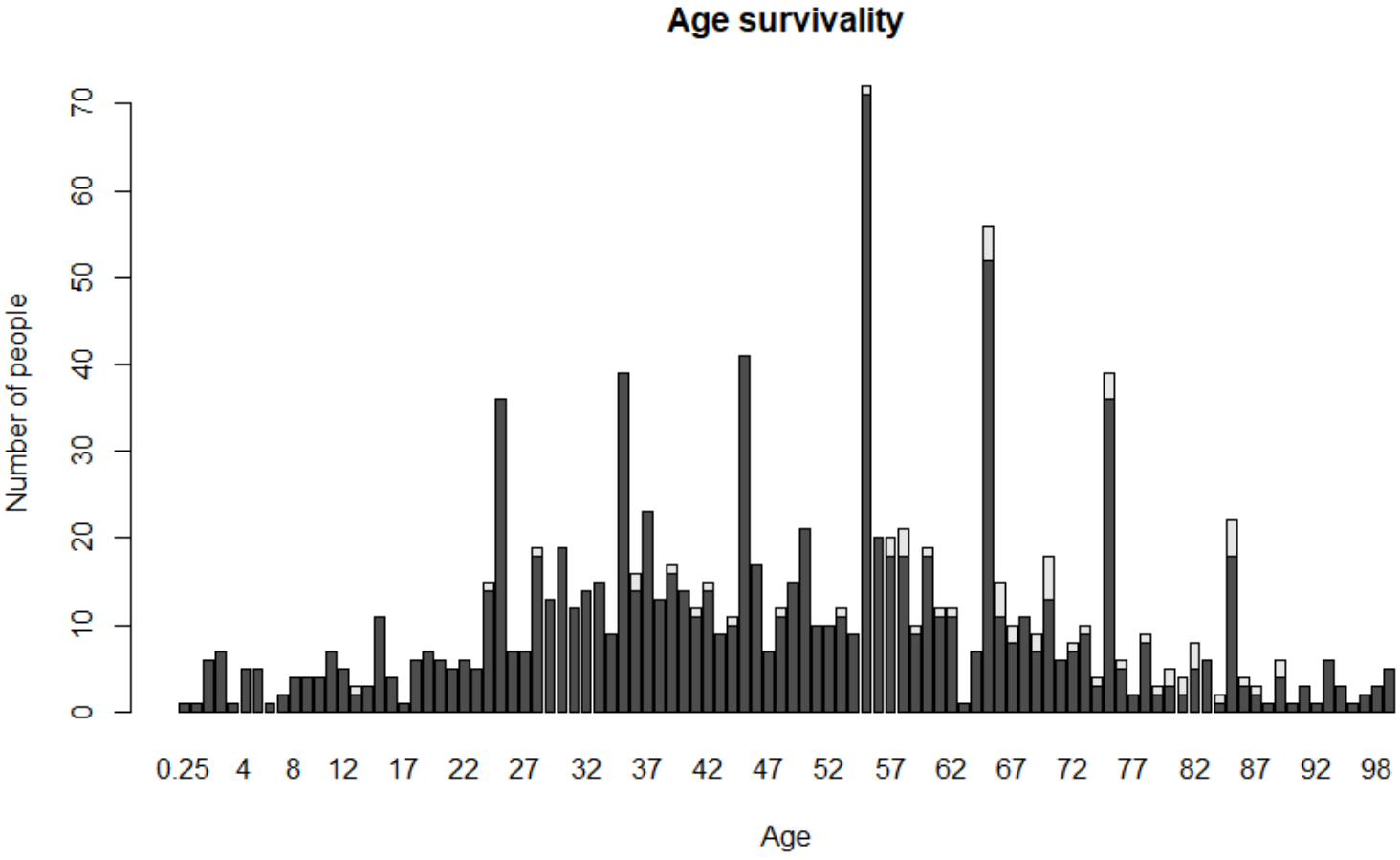
Biorąc pod uwagę stan bieżącej bazy danych, która na moment robienia badań była jedyną sensowną, możliwe jest, że kiedy ta praca jest prezentowana, istnieje już pełniejsza i bardziej szczegółowa baza. Jeżeli tak się okaże to ten dokument zostanie stosownie zaktualizowany.

Badania

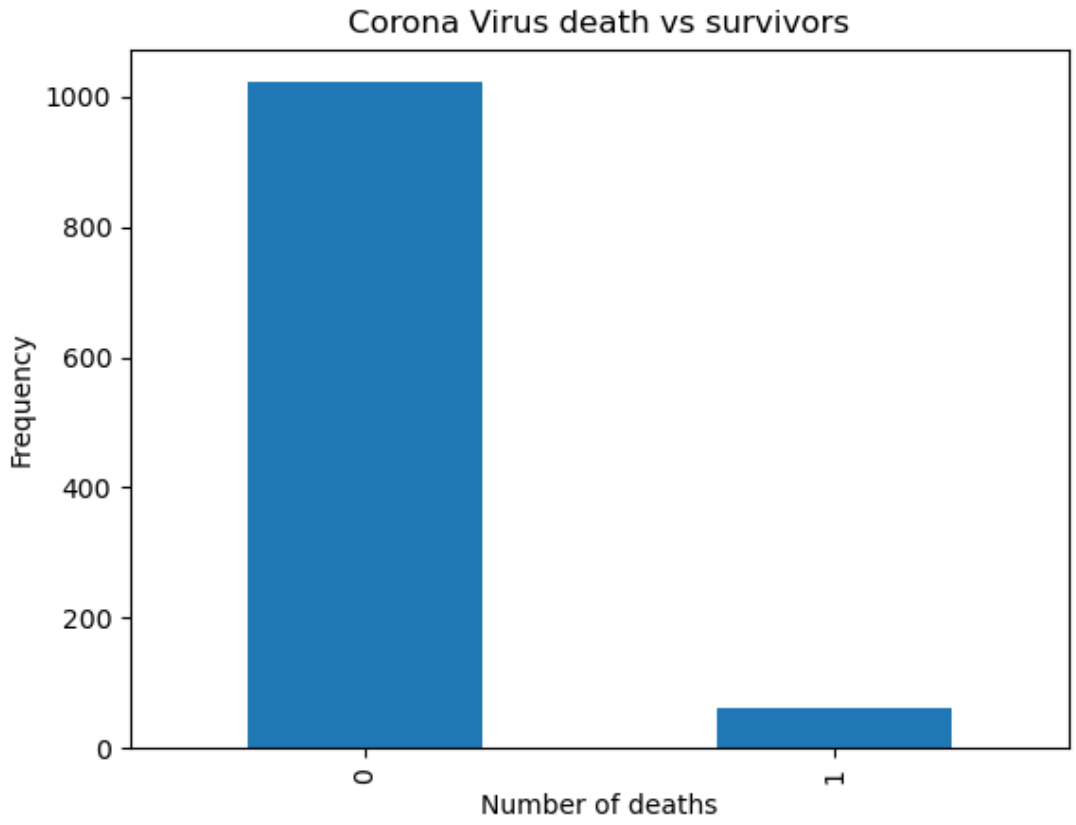
Zacznijmy od statystyk:



Rysunek 2 - wykres przeżywalności wirusa według płci



Rysunek 3 - wykres przeżywalności wirusa według wieku



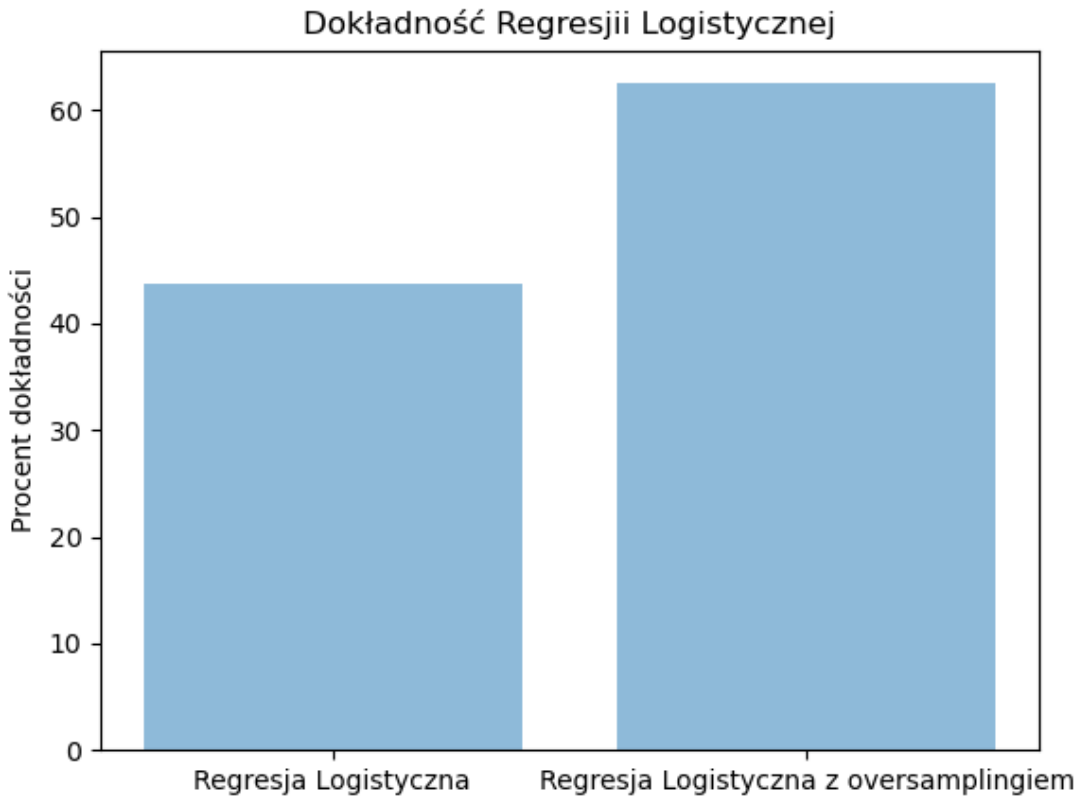
Rysunek 4 - przedstawia stosunek pacjentów, którzy umarli do tych, którzy przeżyli infekcję (0 - pacjent wyzdrowiał, 1 - pacjent poniósł śmierć)

Jak widać po wstępnych statystykach najbardziej zagrożoną grupą są ludzie po 50 roku życia oraz mężczyźni względem kobiet dużo częściej nie zdrowieją po zarażeniu się wirusem.

Wykorzystajmy teraz wymienione powyżej metody do klasyfikowania zbiorów nieregularnych. Jednak przed przedstawieniem wyników trzeba omówić w jaki sposób liczyłem prawdopodobieństwo wystąpienia klasy śmierci.

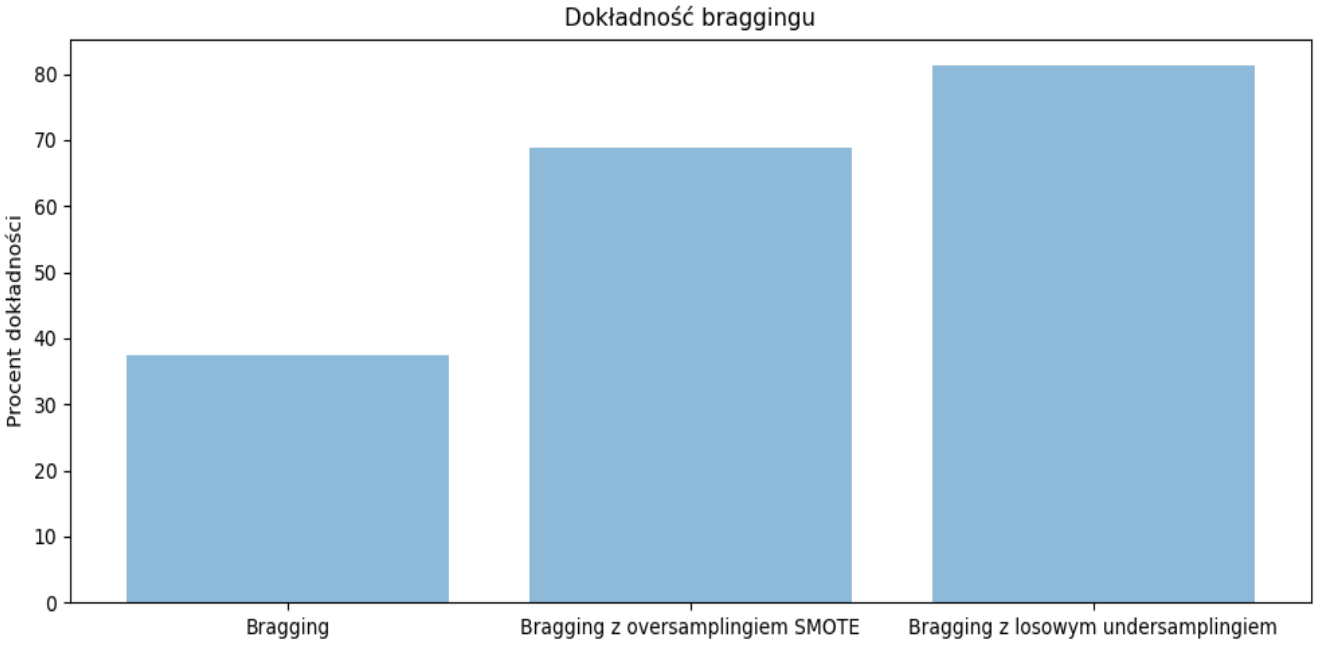
Gdybyśmy liczyli prawdopodobieństwo w sposób normalny, a więc stosunek dobrych odpowiedzi do złych to z powodu małej ilości klasy odpowiedzialnej za śmierć pacjenta moglibyśmy postawić tezę, że wszyscy ludzie przeżyją i mieć dokładność ponad 90%. Dlaczego? Śmiertelność wirusa wynosi około 5%, a więc teoretycznie ze 100 osób chorych tylko 5 umrze co oznacza, że stawiając, że wszyscy przeżyją otrzymamy dokładność 95% bo pomylimy się tylko w 5 przypadkach.

W tym celu zmieniłem sposób liczenia prawdopodobieństwa i skupiłem się na liczbie poprawnie wywnioskowanych śmierci z powodu korona wirusa. Dzięki takiemu sposobowi liczenia możemy określić jak ryzykowna jest infekcja dla danego pacjenta i dla których z nich należy podjąć niezwłoczne działania medyczne w celu uratowania życia.



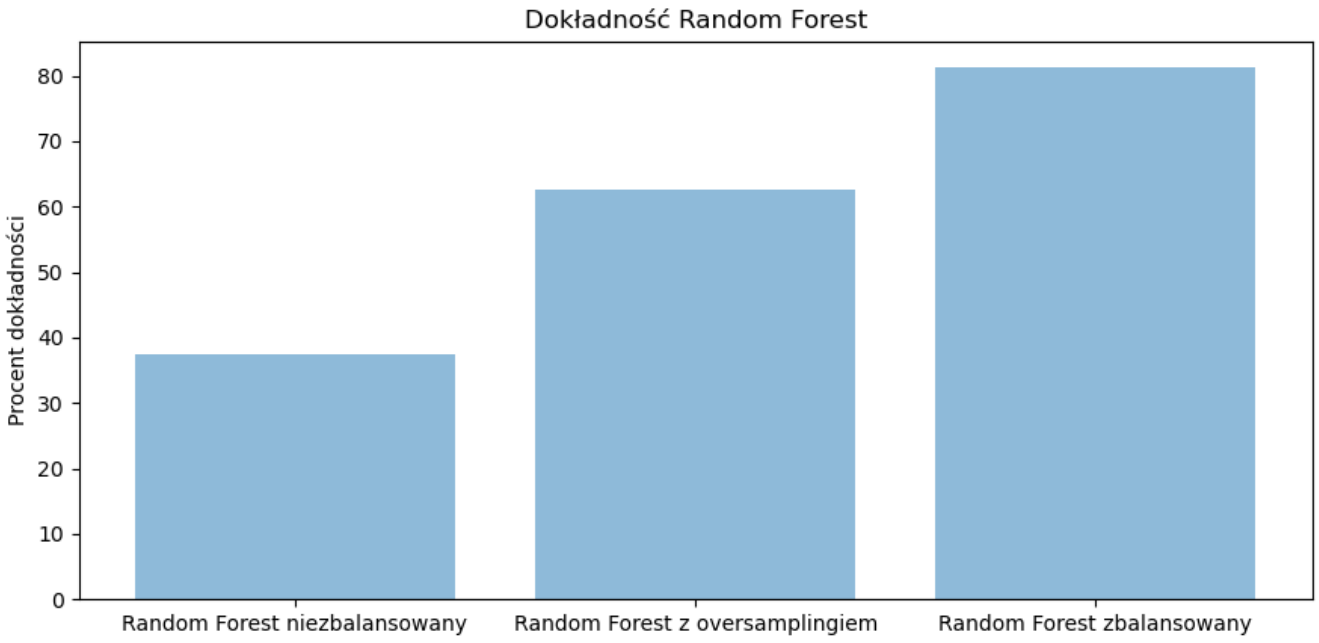
Rysunek 5 - regresja logistyczna bez oversamplingu i z oversamplingiem

Jak widać na wykresie po wykorzystaniu metody oversamplingu, prawdopodobieństwo wykrycia szukanej przez nas klasy wzrosło o prawie 20% niemniej jednak wzrost z 42% do 62% nie jest do końca satysfakcjonującym wynikiem. Dlatego zobaczmy jak z zadaniem poradzą sobie metody zespołowe.



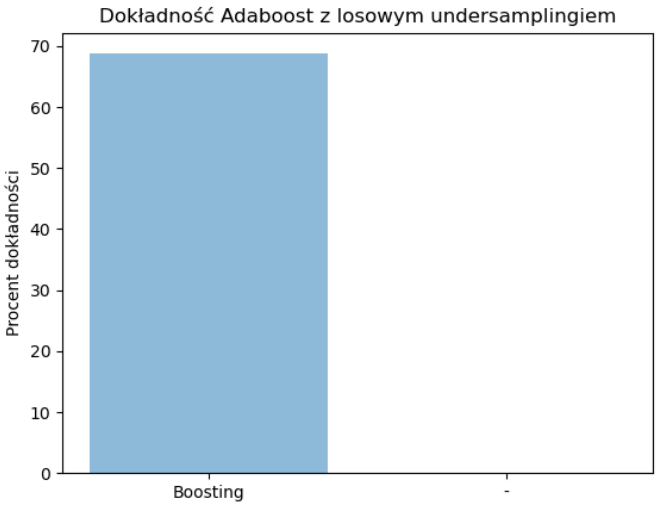
Rysunek 6 - bragging bez undersamplingu i z undersamplingiem

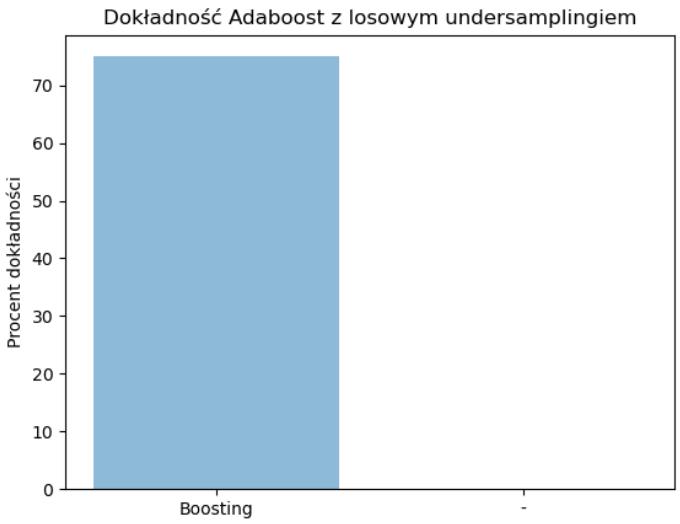
Przy algorytmie braggingu można zauważyć już widoczną poprawę nie dość, że dokładność procentowa wzrosła o ponad 40% to sposób ten jest o 20% lepszy od zwykłej regresji liniowej z undersamplingiem porównując 82% do 62%.

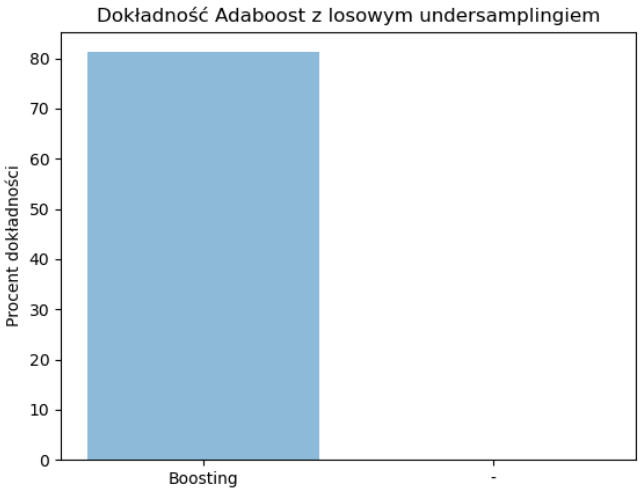


Rysunek 7 - Random Forest bez undersamplingu i z undersamplingiem

Ciekawie wygląda algorytm Random Forest. Teoretycznie powinien poradzić on sobie lepiej od metody braggingu, jednak w praktyce okazało się, że pokazał dokładnie takie same wyniki. Najbardziej prawdopodobną przyczyną jest niska liczba atrybutów w bazie danych.



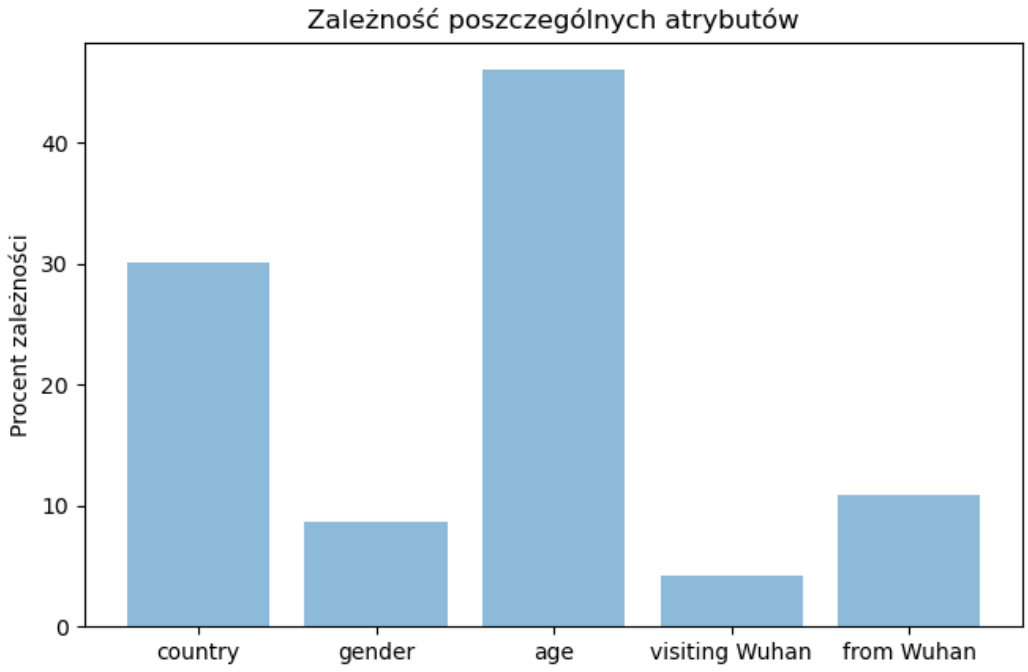




Rysunek 8 - przedstawienie 3 wykresów wahań dokładności algorytmu ADABOOST

Bardzo interesująco wyglądała sytuacja z boostingiem. Algorytm ADABOOST miał inną dokładność z każdą kompilacją programu, gdzie w najlepszym przypadku tworzył on model, którego poprawność dorównywała tym z braggingu i losowego lasu, ale w gorszych wypadkach była o 14% gorsza. Najczęściej jego dokładność otrzymywała się w przedziale od 70% do 80%.

Po przeanalizowaniu wyników używanych algorytmów warto także zobaczyć, które atrybuty najbardziej wpłynęły na klasyfikacje według stworzonego przez nas modelu losowego lasu.



Rysunek 9 - procentu wpływ każdego z atrybutów na klasyfikacje możliwości śmierci pacjenta

Widać dosyć wyraźnie na przedstawionym wykresie, że najważniejszą zmienną przy klasyfikowaniu pacjenta jest jego wiek. Pokrywa się to z informacjami, które już znamy o wirusie CoViD19 – najbardziej zagrożonymi osobami są osoby starsze. Ciekawa jest też statystyka mówiąca o tym, że kraj pochodzenia również wpływa w sposób znaczny na klasyfikacje. Jak wiemy bardzo dużo zgonów następuje w krajach ze słabą opieką medyczną np. Iranie oraz krajach, które bardzo lekceważąco podeszły do zagrożenia – Włochy. Co ciekawe wizyta w strefie Wuhan, gdzie wirus zaczął się rozprzestrzeniać i zaraził największą liczbę osób miała najmniejszy procent zależności.

Wnioski

Powyższe badania przyniosły bardzo ciekawe rezultaty. Udało się mi między innymi osiągnąć bardzo duże prawdopodobieństwo wykrycia czy dana osoba jest w tzw. grupie ryzyka, a co za tym idzie czy powinniśmy rozpocząć niezwłoczne leczenie u takiego pacjenta. Prawdopodobieństwo wykrywalności osiągnęło tutaj poziom 81.25%.

Niestety może to nie być faktyczne prawdopodobieństwo, ponieważ jest ono oparte o małą statystykę ponad 1000 osób, które przeszły infekcję wywołaną nowym korona wirusem. Nie tylko liczba danych jest tutaj problemem ale także ilość zawartych atrybutów. Bardzo pomogłyby nam dane, które zawierają informacje o wcześniej przebytych chorobach, posiadanych schorzeniach, czy o systemie immunologicznym pacjenta. Niestety ale z wielu powodów takich informacji nie znajdziemy w internecie.

To co należałoby zrobić w przyszłości to przede wszystkim poszerzyć zebrane dane o nowych pacjentów i zwiększyć liczbę atrybutów, tylko w ten sposób możemy polepszyć dokładność wykrywania pacjentów w tzw. grupie ryzyka.

Bibliografia

[i] Guo Haixiang, L. Y. (2017). **Learning from class-imbalanced data: Review of methods and applications.** *Expert Systems With Applications*, pp. 220-239.

[ii] Guillaume Lema, Fernando Nogueira, Christos K. Aridas (2017). https://imbalanced-learn.readthedocs.io/en/stable/

[iii] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., and Blondel M., and Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.. https://scikit-learn.org/stable/index.html