Wstęp do Analizy Danych

Przewidywanie jakości wody

Projekt zaliczeniowy

Szymon [redacted], Wojciech [redacted]

15.06.2022r.

1. Sformułowanie tematu oraz celu analizy danych

Dostęp do bezpiecznej wody pitnej jest niezbędny dla zdrowia, jest podstawowym prawem człowieka i elementem skutecznej polityki ochrony zdrowia. Jest to ważne jako kwestia zdrowia i rozwoju na poziomie krajowym, regionalnym i lokalnym. W niektórych regionach wykazano, że inwestycje w zaopatrzenie w wodę i urządzenia sanitarne mogą przynieść korzyści ekonomiczne netto, ponieważ zmniejszenie negatywnych skutków zdrowotnych i kosztów opieki zdrowotnej przewyższają koszty podjęcia interwencji.

Celem naszej analizy będzie skonstruowanie algorytmu, który na podstawie przedstawionych danych będzie potrafił przewidywać czy woda jest zdatna do picia czy też nie.

2. Opis zbioru danych

Dane, z których skorzystaliśmy pochodzą ze strony kaggle.com
Źródło: https://www.kaggle.com/datasets/adityakadiwal/water-potability

Dane zawierają wskaźniki jakości wody dla 3276 różnych zbiorników wodnych. Cechy wody, które pomogą nam stworzyć algorytm to:

1. Wartość pH:

PH jest ważnym parametrem w ocenie równowagi kwasowo-zasadowej wody. Jest także wskaźnikiem kwaśnego lub zasadowego stanu wód. WHO zaleciła maksymalny dopuszczalny limit pH od 6,5 do 8,5. Obecne zakresy badań wynosiły 6,52–6,83, które mieszczą się w zakresie standardów WHO.

2. Twardość:

Twardość powodują głównie sole wapnia i magnezu. Sole te są rozpuszczane ze złóż geologicznych, przez które przepływa woda. Czas kontaktu wody z materiałem powodującym twardość pomaga określić, jaka jest

twardość wody surowej. Twardość była pierwotnie definiowana jako zdolność wody do wytrącania mydła powodowana przez wapń i magnez.

3. Ciała stałe (Całkowite rozpuszczone substancje stałe (TDS)):

Woda ma zdolność rozpuszczania szerokiej gamy nieorganicznych i niektórych organicznych minerałów lub soli, takich jak potas, wapń, sód, wodorowęglany, chlorki, magnez, siarczany itp. Minerały te nadają wodzie niepożądany smak i osłabiają kolor. Jest to ważny parametr do wykorzystania wody. Woda o wysokiej wartości TDS wskazuje, że woda jest silnie zmineralizowana. Pożądany limit dla TDS to 500 mg/l, a maksymalny limit to 1000 mg/l, który jest zalecany do picia.

4. Chloraminy:

Chlor i chloramina to główne środki dezynfekujące stosowane w publicznych systemach wodociągowych. Chloraminy najczęściej powstają, gdy amoniak dodaje się do chloru w celu uzdatniania wody pitnej. Poziomy chloru do 4 miligramów na litr (mg/l lub 4 części na milion (ppm)) są uważane za bezpieczne w wodzie pitnej.

5. Siarczan:

Siarczany to naturalnie występujące substancje, które znajdują się w minerałach, glebie i skałach. Są obecne w otaczającym powietrzu, wodach gruntowych, roślinach i żywności. Główne komercyjne zastosowanie siarczanu to przemysł chemiczny. Stężenie siarczanów w wodzie morskiej wynosi około 2700 miligramów na litr (mg/l). W większości zasobów słodkiej wody waha się od 3 do 30 mg/l, chociaż w niektórych lokalizacjach geograficznych stwierdza się znacznie wyższe stężenia (1000 mg/l).

6. Przewodność:

Czysta woda nie jest dobrym przewodnikiem prądu elektrycznego, jest raczej dobrym izolatorem. Wzrost stężenia jonów poprawia przewodnictwo elektryczne wody. Ogólnie rzecz biorąc, ilość rozpuszczonych ciał stałych w wodzie określa przewodność elektryczną. Przewodność elektryczna (EC) faktycznie mierzy proces jonowy roztworu, który umożliwia mu przesyłanie prądu. Zgodnie ze standardami WHO wartość EC nie powinna przekraczać 400 µS/cm.

7. Węgiel organiczny:

Całkowity węgiel organiczny (TOC) w wodach źródłowych pochodzi z rozkładającej się naturalnej materii organicznej (NOM), a także ze źródeł syntetycznych. TOC jest miarą całkowitej ilości węgla w związkach organicznych w czystej wodzie. Według US EPA < 2 mg/L jako TOC w uzdatnionej / pitnej wodzie i < 4 mg/Lt w wodzie źródłowej używanej do leczenia.

8. Trihalometany:

THM to substancje chemiczne, które można znaleźć w wodzie uzdatnionej chlorem. Stężenie THM w wodzie pitnej zmienia się w zależności od poziomu materiału organicznego w wodzie, ilości chloru wymaganej do uzdatniania wody oraz temperatury uzdatnianej wody. Poziom THM do 80 ppm jest uważany za bezpieczny w wodzie pitnej.

9. Mętność:

Zmętnienie wody zależy od ilości ciał stałych obecnych w stanie zawieszonym. Jest miarą właściwości emitujących światło wody, a test służy do określenia jakości odprowadzanych ścieków w odniesieniu do materii koloidalnej. Średnia wartość zmętnienia uzyskana dla Wondo Genet Campus (0,98 NTU) jest niższa niż zalecana przez WHO wartość 5,00 NTU.

10. Zdatność do picia:

Wskazuje, czy woda jest bezpieczna do spożycia przez ludzi, gdzie 1 oznacza zdatną do picia, a 0 oznacza niezdatną do picia.

3. Wybór narzędzi do analizy danych

W naszym zbiorze danych nasz problem jest problemem binarnym – będziemy starać się przewidzieć wartość kolumny „zdatność do picia” - dlatego też użyjemy do niego metody najbliższych sąsiadów (KNN) oraz będziemy chcieli jego skuteczność porównać z modelem drzew decyzyjnych. Do dopracowania naszych wyników będziemy próbować kilku metod:

- obliczenia modelu KNN dla różnych wartości liczby ***k,***
- normalizowania oraz standaryzowania danych,
- zastosowania AdaBoost do modelu drzew decyzyjnych.

4. Opis jakości zbudowanych modeli/ opis projektu

Nasze dane we wszystkich cechach są ilościowe lecz wartości pH, siarczany oraz trihalometany w niektórych zbiornikach wodnych były niezarejestrowane co sprawiało że w naszym zbiorze brakowało niektórych danych. Zastąpiliśmy te braki średnimi z całej kolumny odpowiednio: średnia wartości pH, siarczany oraz trihalometany.

Dzielimy zbiór danych na zbiór uczący oraz zbiór testowy w proporcjach 80% do 20% **w sposób losowy**.

Na początek sprawdziliśmy algorytm KNN dla znormalizowanych danych. Wyniki przedstawiliśmy w macierzy krzyżowej.

Następnie użyliśmy algorytmu KNN dla standaryzowanych danych. Wyniki również zostały przedstawione w tabeli krzyżowej.

Kolejnym etapem jest zbudowanie drzewa decyzyjnego i przedstawienie wygenerowanych wyników. Algorytm uznał, że najistotniejszą cechą, która decyduje czy woda jest zdatna do picia jest ilość obecnych siarczanów w zbiorniku.

Ostatnim etapem było zbudowanie Adaboost w celu poprawy drzewa decyzyjnego.

5. Wyniki

Przedstawione zostaną tutaj wyniki użytych algorytmów. Dla algorytmu KNN ze standaryzowanymi danymi wynik to:

	water_s_test_pred		
water_test_labels	0	1	Row Total
0	387	22	409
	0.946	0.054	0.623
	0.625	0.595	
	0.590	0.034	
1	232	15	247
	0.939	0.061	0.377
	0.375	0.405	
	0.354	0.023	
Column Total	619	37	656
	0.944	0.056	

Oznacza to skuteczność na poziomie 61.3%. Wynik ten jest osiągnięty dla liczby sąsiadów k=56 i jest to najlepszy wynik jaki udało nam się osiągnąć dla tego modelu przy sprawdzeniu innych wartości liczby ***k***. Jak widzimy wartość naszego „najlepszego” k jest w przybliżeniu równa pierwiastkowi liczby danych w zbiorze ($\sqrt{3276} \approx 57$).

Wynik dla drzewa decyzyjnego:

	predicted_potability		
actual potability	0	1	Row Total
0	393	13	406
	0.599	0.020	
1	236	14	250
	0.360	0.021	
Column Total	629	27	656

Oznacza to skuteczność na poziomie 62%.

Po sprawdzeniu modelu Adaboost otrzymaliśmy:

	predicted_potability		
actual potability	0	1	Row Total
0	397	9	406
	0.605	0.014	
1	240	10	250
	0.366	0.015	
Column Total	637	19	656

Zatem otrzymaliśmy skuteczność na tym samym poziomie, czyli model Adaboost nic nie poprawił.

Najlepszym modelem okazał się algorytm KNN dla znormalizowanych danych:

	water_n_test_pred		
water_test_labels	0	1	Row Total
0	404	5	409
	0.988	0.012	0.623
	0.643	0.179	
	0.616	0.008	
1	224	23	247
	0.907	0.093	0.377
	0.357	0.821	
	0.341	0.035	
Column Total	628	28	656
	0.957	0.043	

Skuteczność w tym modelu to ok. 65.1%. Wynik ten jest osiągnięty dla liczby sąsiadów k=57 i jest to najlepszy wynik jaki udało nam się osiągnąć dla tego modelu przy sprawdzeniu innych wartości liczby ***k***. Jak widzimy wartość naszego „najlepszego” k jest w przybliżeniu równa pierwiastkowi liczby danych w zbiorze ($\sqrt{3276} \approx 57$).

6. Podsumowanie

Podsumowując, najskuteczniejszym z rozważanych algorytmów okazał się algorytm KNN zastosowany dla znormalizowanych danych dla liczby k = 57 – skuteczność, którą uzyskaliśmy jest na niezbyt satysfakcjonującym poziomie 65% - być może użycie innego modelu do klasyfikacji binarnej, którego nie poznaliśmy na zajęciach pomoże nam uzyskać większą skuteczność.