Dokumentacja projektowa

Inżynieria Biomedyczna w Praktyce

Michał Pawełek

Sekcja 3, Kierunek: Informatyka

Wydział Matematyki Stosowanej

Rok akademicki 2021/2022

Tytuł: *Ocena poziomu tłuszczu w organizmie na podstawie danych antropometrycznych*

Spis treści

1 Wstęp

W dzisiejszych czasach problemy spowodowane nadwagą przybierają poważniejszego tempa i nie zapowiada się, żeby szybko zniknęły z naszego życia. Najczęściej dotyka to dzieci i nastolatków nie biorący pod uwagę konsekwencji nadmiernego pochłaniania niezdrowego jedzenia. Najczęstszym współczynnikiem nadwagi jest m.in. poziom tłuszczu w organizmie. Oczywiście, jest to bardzo zależna oraz indywidualna skala, na podstawie której nie można wywnioskować precyzyjnych wniosków, lecz dobrze pokazuje w jakim stopniu otyłości może znajdować się pojedyncza jednostka.
Celem projektu jest oszacowanie na podstawie wcześniej zmierzonych danych antropometrycznych poziomu tłuszczu w ciele. Granicami wskazującymi czy dana wartość pojedynczej osoby będzie należała do normy, będą dane pochodzące z testów analizatora składu ciała znanego pod nazwą InBody. Obiektami przeprowadzonych testów były dzieci oraz nastolatkowie w wieku szkolnym.
Po uprzednim uporządkowaniu bazy pod kątem potrzebnych nam danych do właściwej analizy, owe dane będą przekazywane do odpowiednich algorytmów uczenia maszynowego w celu ustalenia własnego poziomu tłuszczu i porównania ich z pierwotnymi wynikami. Dzięki temu będzie można sprawdzić, czy za pomocą otrzymanych wyników antropometrycznych jest możliwość zautomatyzować wskazanie poziomu tłuszczu danej osoby.

2 Analiza zagadnienia

Globalny wzrost częstości występowania nadwagi i otyłości we wszystkich kategoriach wiekowych, związanych z brakiem aktywności fizycznej i niewłaściwymi nawykami żywieniowymi, motywuje do badań i uzasadnia skuteczną edukację oraz edukację społeczną jak również zdrowotną, mające na celu złagodzenie tych negatywnych tendencji. Częstość występowania otyłości coraz częściej występuje w niższych kategoriach wiekowych. Otyłość to choroba związana ze zwiększonym poziomem rozwoju gospodarczego kraju oraz charakteryzująca się nadmiernym nagromadzeniem tłuszczu w organizmie człowieka, jest czynnikiem ryzyka wielu chorób niezakaźnych w populacjach dorosłych i dzieci. Ponadto otyłość dziecięca ma tendencję do utrzymywania się w okresie dojrzewania, a nawet w wieku dorosłym, z poważnymi konsekwencjami zdrowotnymi.
W związku z narastającą nadwagą, otyłość oraz stan niedowagi w dzieciństwie negatywnie wpływają na sprawność motoryczną, która zależy od koordynacji nerwowo-mięśniowej, sprawności i rozwoju mięśni. Z nadwagą i otyłością wiąże

się nadmiar tkanki tłuszczowej, której ilość mierzy się w celu oceny składu ciała. Tkanka tłuszczowa jest największym narządem dokrewnym w organizmie - jego wzrost prowadzi do poważnych zmian metabolicznych, a w konsekwencji do przemian somatycznych w organizmie. Mięśnie szkieletowe są również bardzo ważne, ponieważ stosunek masy mięśniowej do masy tłuszczowej jest często używany do oceny zdrowia metabolicznego.
Najskuteczniejszym sposobem leczenia otyłości u dzieci jest profilaktyka pierwotna. Centralną rolę w profilaktyce odgrywa diagnostyka somatyczna. Poprzez diagnozę somatyczną jesteśmy w stanie określić, a następnie, w trakcie rozwoju, porównać ukierunkowanie wzrostu wybranych parametrów somatycznych dziecka na wykresach centylowych. Jedną z podstawowych metod oceny wzrostu jest porównanie zmierzonych danych antropometrycznych dziecka ze średnią populacyjną za pomocą wykresów wzrostu. Karty wzrostu są nieodzowną częścią praktyki pediatrycznej ze względu na łatwość ich użycia w ocenie wzrostu i rozwoju. Badania antropometryczne przeprowadzane w regularnych odstępach czasu są ważnymi wskaźnikami stanu zdrowia dzieci i najwłaściwszym sposobem oceny stanu odżywienia i stanu zdrowia populacji pediatrycznej.

Projekt został wykonany w środowisku programistycznym *Jupyter Notebook* przy użyciu języka *Python*. Bazy danych ('dane_inbody' oraz 'dane_antropo') zostały zapożyczone z kursu "Inżynieria Biomedyczna w Praktyce"dla kierunku Informatyka, umiejscowione na Platformie Zdalnej Edukacji należącej do Politechniki Śląskiej. W załączniku do dokumentacji projektu będzie znajdywać się kod projektu w dwóch wariantach: w notebooku (dla odczytania w Jupiterze) oraz w czystym Pythonie (dla jakiegokolwiek środowiska programistycznego umożliwiającym odczyt owego języka)

3 Specyfikacja wewnętrzna

3.1 Przystosowanie bazy danych pod temat projektu

Pierwszym jak i najważniejszym krokiem rozpoczynającym każdy projekt jest uprzednie zaimportowanie potrzebnych do pracy bibliotek. W tym projekcie w skład wchodziły biblioteki takie jak:

- **numpy,**
- **pandas,**
- **seaborn,**
- **math,**
- **collections,**
- **sklearn**

Ostatnia z wymienionych jest szczególnie ważna, gdyż dostarczyła najwięcej gotowych funkcji w obszarze klasyfikacji m.in.:

(a) różne klasyfikatory (KNN, Naiwny Bayes itp.),

(b) macierz błędów

(c) metryki

(d) normalizację zbioru

Po zaimportowaniu bibliotek jak również samych baz danych, potrzebne było wyodrębnić z bazy 'dane_inbody' kolumny zawierającą *BodyFatMass* wraz z jego granicami górnymi oraz dolnymi (wygenerowanymi na podstawie innych danych takich jak chociażby wiek) jak i *id* danego pomiaru.

	id	BFM	BFM_LowLim	BFM_UpLim
0	101	12.5	17.7	8.9
1	102	5.8	14.5	7.2
2	104	5.6	18.7	9.3
3	105	5.1	16.9	8.4
4	107	27.0	13.3	6.6
...	...	...	...	...
867	509	0.0	21.0	20.9
868	510	0.0	9.4	10.9
869	511	0.0	16.5	19.3
870	601	0.0	17.1	19.8
871	999	0.0	22.4	22.2

Rysunek 1: Tablica zawierającą BFM

Tak stworzona tabela (Rysunek 1) posiada pewne nieprawidłowości - pewne dane nie zgadzają się z nazwą kolumny, dane większe są w kolumnie tam gdzie powinny być mniejsze i na odwrót. Do naprawienia jak również do stworzenia odpowiedniej klasy do klasyfikacji (o czym w następnych krokach) została stworzona autorska klasa (Rysunek 2).

```
class tworzenie:
    @staticmethod
    def zamiany(x):
        for i in range(len(x)):
            low=x.iloc[i,2]
            up=x.iloc[i,3]
            if low>up:
                x.iloc[i,2]=up
                x.iloc[i,3]=low
        return x

    @staticmethod
    def klasa(x):
        for i in range(len(x)):
            low=x.iloc[i,2]
            up=x.iloc[i,3]
            bfm=x.iloc[i,1]
            if bfm<low:
                x.iloc[i,1]=0
            elif bfm>up:
                x.iloc[i,1]=2
            else:
                x.iloc[i,1]=1
        return x
```

Rysunek 2: Klasa do zamiany komórek oraz tworzenia klasy

Można łatwo zauważyć, że funkcja **'zamiania'** to po prostu zmiana kolumny z jednej na drugą w razie złego ustawienia.
Po wykonaniu pierwszej funkcji, w celu ułatwienia oraz ograniczenia się do prawidłowych danych, już od tego momentu należało zacząć redukcję bazy o błędne dane. Z załączonych w *Literaturze* źródeł, zostało określone, że dane o ilości tłuszczu w organizmie dla danej grupy badawczej będą oscylować w granicach nie większych niż 30 oraz nie mniejszych niż 2.
Następnie użyto drugiej z wcześniej wspomnianych funkcji jakim jest **'klasa'**. Służy ona do ustalenia i nadania odpowiedniej numeracji poziomu tłuszczu. Szacuje ona czy dana wartość mieści się w przedziale poprzednio naprawianym: jeśli wartość jest poniżej dolnego limitu to otrzymuje wartość 0, jeśli jest ona powyżej górnego limitu to 2 a gdy mieści się pomiędzy nimi to dostaje 1. Na podstawie tego sposobu można wstępnie założyć czy dana osoba ma dobry poziom tłuszczu czy odbiega on normy w zależnym od tego kierunku.
Gdy wszystkie rekordy zostały odpowiednio zaklasyfikowane potrzebne było usunąć już niepotrzebne kolumny z granicami jak również przejść do kolejnego kroku jakim było złączenie przygotowanej tabelki wraz z tabelą zawierającą dane antropometryczne. Tabele są łączone poprzez wspólne numery id.

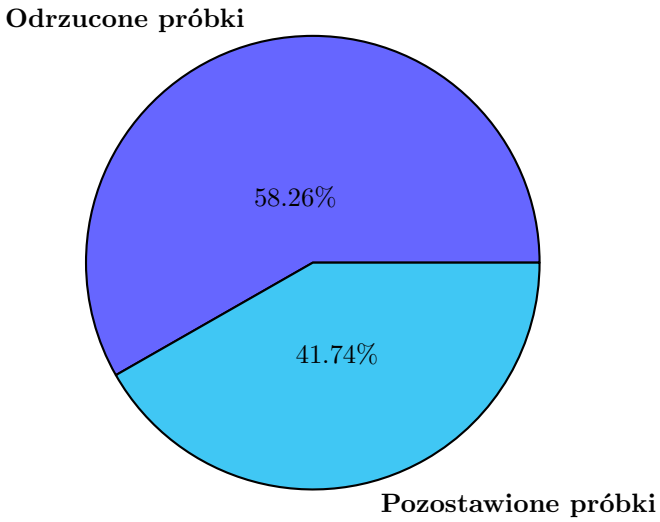
	id_x	Wysokosc	WysokoscSiedzeniowa	TetnoKrwi	id_y	BFM
0	101	182.0	96.0	66.0	101	1.0
1	102	167.0	84.4	75.0	102	0.0
2	104	187.4	93.5	0.0	104	0.0
3	105	177.9	90.0	92.0	105	0.0
4	107	161.7	84.8	88.0	107	2.0
5	108	160.1	86.4	74.0	108	1.0

Rysunek 3: Widok połączonych tabel z wybranymi kolumnami i wierszami

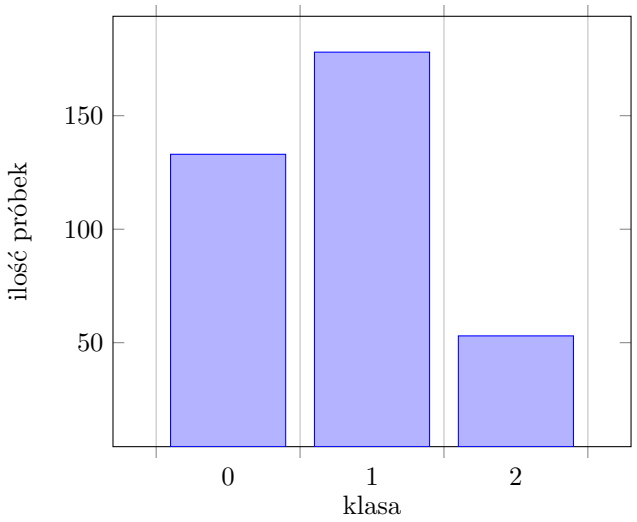
Posiadając pełną tabelę z danymi (Rysunek 3), należało raz jeszcze zredukować bazę o kolejne niepotrzebne do analizy dane. W tym celu zostały postawione 3 warunki (wszystkie dane zostały dobrane na podstawie źródeł uwzględnionych na samym dole projektu):

1. pomiar miał posiadać ciśnienie skurczowe w granicach od 60 do 181
2. pomiar miał posiadać ciśnienie rozkurczowe w granicach 30 do 111
3. pomiar miał posiadać tętno krwi od 50 do 110

Przy użyciu wcześniej stworzonej zmiennej możemy się dowiedzieć, iż dzięki filtrowaniu usunięto 508 rekordów z 872 a pozostało 364 poprawnych rekordów do analizy, co pokazuje poniższy diagram.



Za pomocą funkcji z biblioteki 'Counter', dowiemy się także o ważnej informacji, jaką jest ilość przypisanych pomiarów do odpowiednich klas. Według otrzymanych wniosków (oraz poniższego diagramu), najwięcej pomiarów jest w klasie 1, druga w wielkości jest klasa 0 a najmniej znajduje się w klasie 2.



3.2 Normalizacja i podział danych

Po usunięciu niepotrzebnych do analizy kolumn takich jak *'id_x'* czy *'data badania'*, baza danych była prawie gotowa do dalszego działania. Jedyną rzeczą jaką można byłoby jeszcze wykonać przed oddaniem danych pod uczenie maszynowe to normalizacja zbioru oraz jej podział.
Do tej pierwszej posłużyła funkcja z biblioteki **'sklearn.preprocessing'** pod nazwą *MinMaxScaler*. Warto ówcześnie pamiętać, że normalizacji nie powinna podlegać sama klasa do klasyfikacji, w tym przypadku to 'BFM'. Aby rozwiązać owy problem wystarczyło zwyczajnie zapisać kolumnę z wcześniej wymienionymi wartościami w pewnej zmiennej, usunąć tę kolumnę z bazy, następnie poddać bazę pod normalizację a na koniec przywrócić zapisaną wcześniej kolumnę z zapisanej zmiennej do już przygotowanej bazy. Tak przygotowane dane (Rysunek 4) będą poddawane dalszym procedurom.

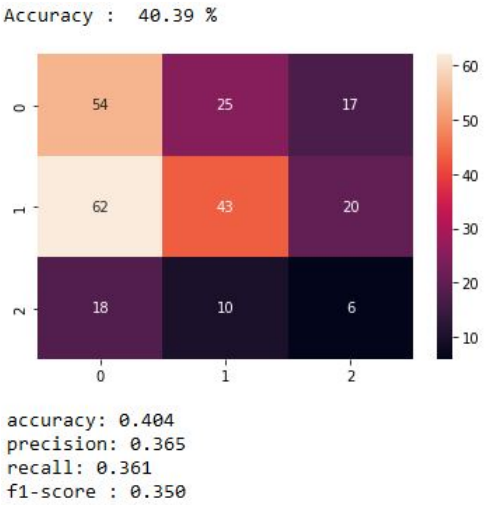
	Wysokosc	WysokoscSiedzeniowa	TetnoKrwi	BFM
0	0.786260	0.849398	0.245614	1.0
1	0.557252	0.500000	0.403509	0.0
2	0.723664	0.668675	0.701754	0.0
3	0.476336	0.512048	0.631579	2.0
4	0.451908	0.560241	0.385965	1.0
5	0.322137	0.439759	0.561404	1.0

Rysunek 4: Znormalizowana tabela z wybranymi kolumnami i wierszami

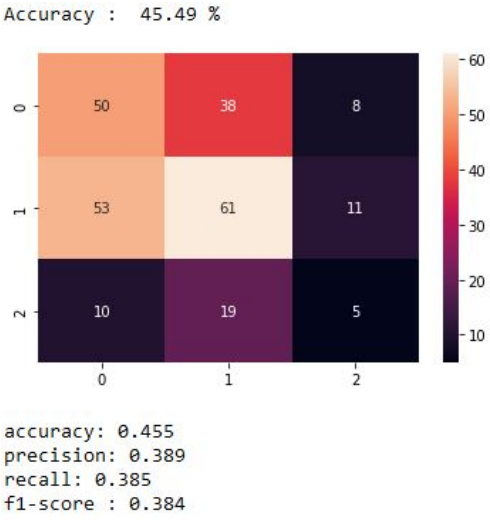
Przed rozpoczęciem klasyfikacji zbiór danych należało podzielić na zbiór treningowy i testowy. Pierwszy z nich, traktowany jako znane dane, służył jako punkt odniesienia dla algorytmów uczenia ich. Drugi zbiór traktowany był jako dane wprowadzane do programu w celu ich sklasyfikowania. Dzięki posiadaniu wiedzy o klasach, jakie przypisane były do tych próbek można było porównać je do wyników klasyfikacji i przeprowadzić analizę skuteczności algorytmów. Proporcje podziału zbioru to 7:3.

4 Wyniki

Algorytmami do klasyfikowania danych były kolejno: Naiwny Bayes, regresja logistyczna, KNN, drzewo decyzyjne oraz lasy losowe. Wszystkie algorytmy pochodzą z biblioteki *'sklearn'*.

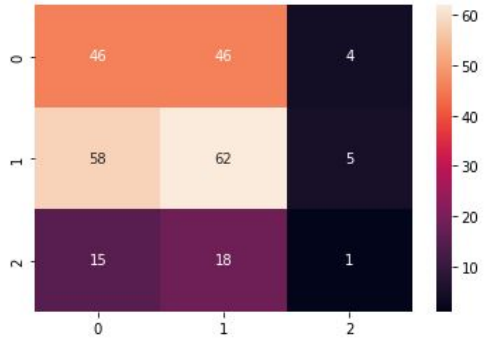


Rysunek 5: Naiwny Bayes z użyciem wzoru Gaussa



Rysunek 6: Regresja logistyczna

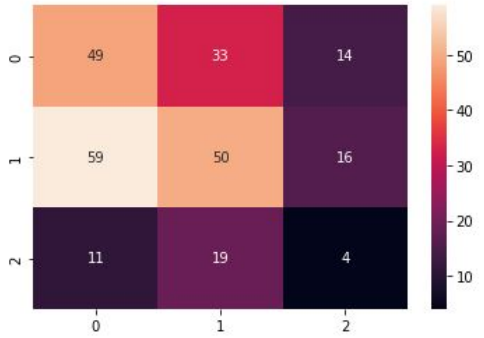
Accuracy : 42.75 %



accuracy: 0.427
precision: 0.326
recall: 0.335
f1-score : 0.322

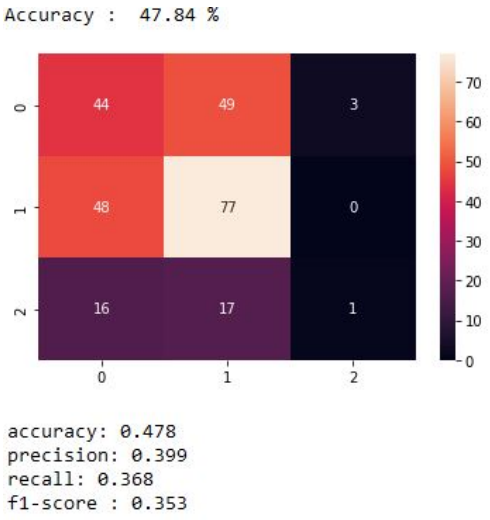
Rysunek 7: KNN

Accuracy : 40.39 %



accuracy: 0.404
precision: 0.340
recall: 0.343
f1-score : 0.338

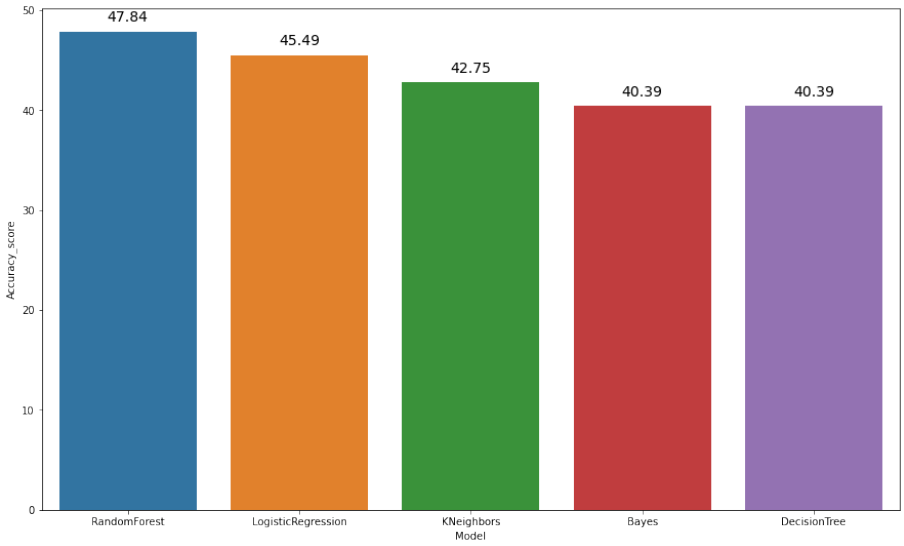
Rysunek 8: Drzewo decyzyjne



Rysunek 9: Lasy losowe

Jak można zauważyć na powyższych macierzach błędów, klasyfikatory miały problem z klasą 2 gdzie ogółem próbek było najmniej. Może to dać do zrozumienia, że dane w tychże próbkach były zbyt bardzo zbliżone np. do klasy 1 i tym sposobem algorytmom mogło obliczać się inny wynik.
Dobrym spostrzeżeniem jest też to, że wszystkie metryki są do siebie bardzo zbliżone - zwłaszcza *'precyzja', 'czułość' i 'f1'*, gdzie wartości oscylują pomiędzy 32% a 39%. Świadczy to o specyficzności bazy jak również o braku różnicy skuteczności danej metody.
Żeby uzyskać prawidłowe wyniki, na macierzy musiałyby być liczby tylko na przekątnej zaczynając od lewego górnego rogu, lecz jak widać ani jednemu klasyfikatorowi się to nie udało poprawnie w całości zrobić.

Poniższy diagram (Rysunek 10) przedstawia porównanie wszystkich użytych w tym projekcie algorytmów klasyfikacyjnych w formie diagramu słupkowego.



Rysunek 10: Porównanie klasyfikatorów

5 Wnioski

Analizując powyższy diagram mogą się nasunąć dwa wnioski. Pierwszym z nich jest to, że zależność skuteczności pomiędzy danymi algorytmami klasyfikującymi dane na tejże bazie nie jest duża, albowiem nie ma większej różnicy którego algorytmu użyjemy. Może się wydawać przy największym i najmniejszym słupku (*RandomForest i DecisionTree*), iż różnica jest dość znacząca i przekracza około 7%, ale nie jest to całkowicie prawidłowe spostrzeżenie, gdyż ponownie uruchamiając podział danych wraz z klasyfikatorami będzie można dostrzec znaczącą różnice otrzymanych wyników. Dzieje się tak przez losowy podział zbioru, który za każdym razem inaczej porcjuje dane dla zbioru treningowe jak i testowego.
Drugi wniosek dotyczy samej bazy. Posiada ona trudne do uniwersalnego pogrupowania dane. Owe dane są tak mało znacząco zróżnicowane między sobą, że powoduje to trudności podczas klasyfikowania. Nawet po dogłębnym zredukowaniu danych, nie spowodowało to powstaniu idealnej bazy. Jest to oczywiście praktycznie niemożliwe, żeby wszystkie dane (zwłaszcza będące na podstawie prawdziwych badań przeprowadzonych na ludziach) mogły dawać jednoznaczny rezultat.
Z obu tych wniosków możemy więc wyciągnąć tezę, która mówi o możliwym, lecz niedokładnym procesie klasyfikowania poziomu tłuszczu przez uczenie maszynowe z danych antropometrycznych.

6 Podsumowanie

Projekt został zrealizowany całkowicie z ustalonymi wcześniej celami. Głównym celem projektu było oszacowanie poziomu tłuszczu danego uczestnika badania co udało się z powodzeniem. Niestety nie możemy tego powiedzieć o próbie zautomatyzowania procesu przez algorytmy klasyfikujące, ponieważ jak efekt swoich działań był widoczny, tak owe wyniki nie pokrywały się z realnym stanem rzeczy. Dzięki temu można wyciągnąć wnioski, że na tych zbiorach danych nie opłaca się wykonywać uczenia maszynowego. Nie jest to zła wiadomość, lecz blokuje to możliwość dalszego rozwoju w tymże kierunku (ale w innych nadal jak najbardziej pozwala).
Wyniki na poziomie 40%/50% nie są zadowalające, zwłaszcza gdy chodzi o klasyfikatory gdzie powinny one być 'górnolotne' - oscylować w okolicach 80%/100%. Pomimo tych niedogodności, oficjalnie można oświadczyć że wszystkie cele jakie ten projekt miał dostarczyć zostały spełnione.
Oczywiście dalszy rozwój projektu nie kończy się z oddaniem go do oceny - jest w nim wiele rzeczy, które mogłyby być lepiej zrealizowane oraz bardziej rozszerzone. Jednymi z takich rzeczy są:

- Urozmaicenie zakresu danych o inne zbiory np. poprzez przeprowadzenie badań metodą DXA,
- Zebranie danych od ludzi z różnych części świata lub różnych sfer społecznych,
- Spróbowanie przełożenia projektu na inny język programowania i porównanie wyników,
- Lepsze i bardziej precyzyjne redukowanie danych (potrzebne większe doświadczenie w zakresie biologi),
- Wykonanie obliczeń klasyfikacyjnych oraz powtórzenie ich np. 100 razy i dopiero wtedy spróbowanie wyciągnięcia wniosków

Literatura

[1] *https://www.builtlean.com/body-fat-percentage-men-women/*

[2] *https://www.mayoclinic.org/diseases-conditions/low-blood-pressure/symptoms-causes/syc-20355465*

[3] *https://apteline.pl/artykuly/prawidlowy-puls-za-niski-puls-tetno-prawidlowe-tetno-spoczynkowe-za-wysoki-puls*

[4] *https://www.mdpi.com/2227-9067/8/5/366/html*

[5] *https://academic.oup.com/advances/article/5/3/320S/4562745*

[6] Jonathan C. K. Wells *Advances in Nutrition,* Volume 5, Issue 3, May 2014, Pages 320S–329S

[7] Bogin.B. *Patterns of human growth,* Cambridge (UK): Cambridge University Press; 1988

[8] Nasreddine, L.; Naja, F.; Chamieh, M.; Adra, N.; Sibai, A.M.; Hwalla, N. Trends in overweight and obesity in Lebanon: Evidence from two national cross-sectional surveys (1997 and 2009). *BMC Public Health* 2012.