Analiza czy po rozmiarze odcisku palca można określić parametry fizyczne osoby

Konrad Kiełczyński
Informatyka Stosowana
Politechnika Wrocławska
MSiD Lab 17:05 TP
260409@student.pwr.edu.pl

I. Wstęp

Poszukując zbioru danych do analizy, zainteresował mnie ten dotyczący pomiarów odcisków palców oraz parametrów fizycznych osób. Celem analizy jest zbadanie czy wielkość odcisku ma jakiś związek z innymi parametrami dotyczącymi badającej osoby takimi jak wzrost i waga.

II. Zbiór danych i jego przetwarzanie

A. Zbiór danych

Zbiór danych dotyczy zebranych danych m.in. na temat pomiaru wagi, wzrostu oraz wielkości służących do określenia rozmiaru odcisku palca wśród 200 uczestników badania. Dane zostały przygotowane przez uczonych z Loughborough University [1].

B. Przetwarzanie wstępne

1) Zmiana formatu pliku: Plik z danymi na stronie jest plikiem w formacie excel-a, zmieniam go na bardziej mi odpowiadający do analizy format CSV (comma-separated values).

2) Kolumny mają długie nie użyteczne nazwy: Dla prostszej analizy danych dodaję ręcznie krótsze skrótowe nazwy kolumn.

3) Ujednolicenie wartości w tabelach: Wartości w kolumnie określającej płeć powinny być wartością Femeale/Male dlatego zmieniamy wszystkie inne wpisy jak np. male na Male. Następnie wartość Male zamieniamy na wartość numeryczną 1 a Female 0.

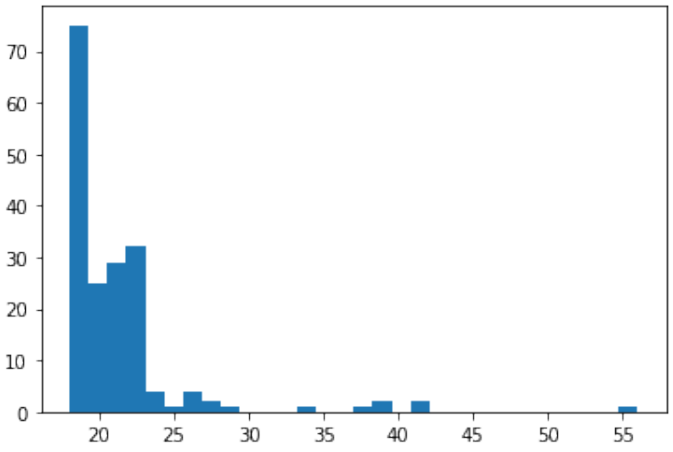
C. Analiza eksploracyjna

Zbiór danych zawiera 200 instancji na temat pomiarów poszczególnych uczestników badania. Zawarte jest w nim 11 atrybutów:

- Numer będący identyfikatorem danej osoby w zbiorze,
- Płeć, wiek, dominująca ręka badanej osoby,
- Wzrost i waga będące wartością średnią z 3 pomiarów,
- Temperaturę, długość, szerokość, powierzchnie i obwód odcisku palca.

1) Pozbycie się zbędnych kolumn:

- Ze względu na mało zróżnicowane dane odrzucam kolumnę dotyczącą dominującej ręki. Blisko 90% badanych ma prawą rękę jako tą dominującą.
- Podobnie decyduje o odrzuceniu kolumny dotyczącej wieku ze względu na to że dane skupiają wśród ludzi w wieku pomiędzy 18-22 rokiem życia Rys. 1.



Rysunek 1. Rozkład wieku badanych osób

- Odrzucam także kolumnę z wartościami dotyczącymi temperatury palca jak i numerze uczestnika badania ze względu na to że nie będą potrzebne przy poniższej analizie.

2) Reprezentacyjny przykład danych: Mała tabelka przykładowych, wykorzystywanych danych dotyczących badanych uczestników Tab. I.

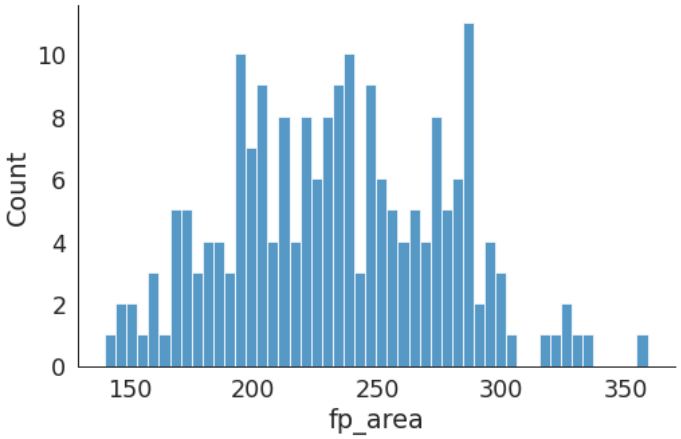
W danych występuje 117 mężczyzn oraz 83 kobiety. Wartości średnie analizowanych danych wyglądają następująco Rys. 2.

Tabela I
Przykładowe wykorzystane dane

gender	height [cm]	weight [kg]	fp_height [mm]	fp_width [mm]	fp_area [mm^2]	fp_circ [mm]
1	174.0	70.0	19.8	13.7	240.6	57.7
1	202.0	99.0	24.0	14.1	278.8	62.7
0	164.0	66.3	19.1	12.5	202.5	52.9
1	182.3	82.0	20.0	13.7	223.8	55.5

```
height       173.1700
weight        72.3900
fp_height     20.3635
fp_width      13.4870
fp_area      233.5530
fp_circ       56.5745
dtype: float64
```

Rysunek 2. Średnie wartości analizowanych danych



Rysunek 3. Rozkład powierzchni odcisku

III. Eksperymenty

1) Celem analizy jest próba:
   a) Dopasowania rozkładu do wartości powierzchni odcisku palca.
   b) Dopasowania modeli regresji: liniowej, GLM i SVR modelującego wzrost człowieka na podstawie jednego z parametrów odcisku palca,
   c) Dopasowania modeli regresji: liniowej, GLM i SVR modelującego wagę człowieka na podstawie jednego z parametrów odcisku palca,
   d) Klasyfikacji płci na podstawie kilku pomiarów odcisku palca.

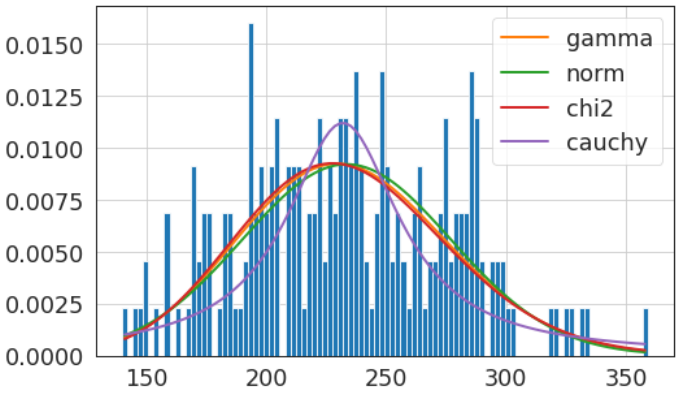
2) Podział danych na trenujące i testowe

   Wszystkie dane użyte do wyznaczenia regresji jak i klasyfikatora w kolejnych etapach, dziele na dane trenujące jaki i testowe w proporcji 75:25 z ustawionym ziarnem losowości na 143 aby dane były losowane zawsze w ten sam sposób. Analizować będę modele stworzone na podstawie estymatorów dostępnych w sckit-learn.

3) Wykonywane eksperymenty z modelami / rozkładami

   A. Dopasowanie rozkładu do wartości powierzchni odcisku palca.

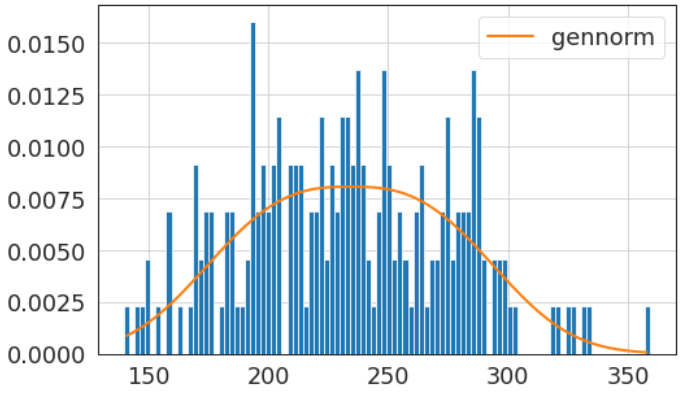
   Pierwszym etapem będzie wizualizacja danych za pomocą histogramu Rys. 3. Następnie za pomocą modułu fitter opartym na module SciPy określimy wartości błędu kwadratowego dla kilku popularnych rozkładów tj. Normalnego, Gamma, Chi2 oraz Cauchy'ego Rys. 4. Jakoś każdego dopasowania określamy błędem średnio-kwadratowym Tab. II. Następnie dzięki metodzie get_distributions() sprawdzam który z wszystkich dostępnych rozkładów posiada najmniejszy błąd średnio-kwadratowy oraz jednocześnie najlepiej wpasowuje się w nasze dane. Tym rozkładem jest UOGÓLNIONY ROZKŁAD NORMALNY (ang. generalized normal distribution or generalized Gaussian distribution [GGD]) którego błąd wynosi 0.000847 Rys. 5.



Rysunek 4. Próba dopasowania kilku popularnych rozkładów do danych

Tabela II
Rozkład i jego wartość błędu średnio-kwadratowego

Rozkład	Błąd
GGD	0.000847
Gamma	0.000883
Normalny	0.000883
Chi2	0.000885
Cauchy'ego	0.001148



Rysunek 5. Najlepszy dopasowany rozkład tj. GGD

B. Model opisujący wzrost człowieka względem obwodu odcisku palca

Po analizie współczynnika korelacji Tab. III możemy wywnioskować że atrybutem najlepiej odwzorowującym wartość wzrostu jest pomiar obwodu odcisku palca. Dlatego też wybieramy ją na daną naszego modelu.

Tabela III
Współczynnik korelacji wzrostu względem różnych danych pomiarowych odcisku palca

Atrybut	Współczynnik
fp_height	0.570573
fp_width	0.500564
fp_area	0.593244
fp_circ	0.609828

Za pomocą danych testujących zostały stworzone 3 modele regresji a konkretniej: regresji liniowej, GLM jak i SVR Rys. 6. Parametry, błąd średniokwadratowy (Mean Square Error (MSE)) jak i uśredniony błąd bezwględny (Mean Absolute Error (MAE)) tych modeli został ukazany w Tabeli Tab. IV.

Tabela IV
Parametry i błąd poszczególnych modeli dotyczących 2 eksperymentu

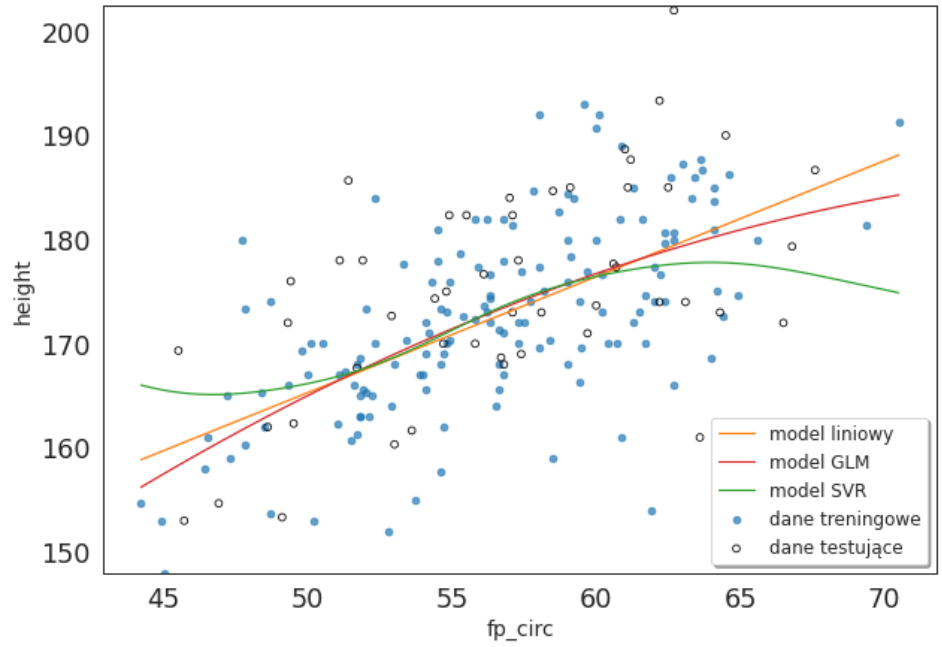
Model	Parametry	MSE	MAE
Model liniowy	[1.11257], 109.68061	76.7	∓7.22
GLM	[0; 3.5954; -0.022], 40.36998	74.7	∓7.12
SVR	———	78.1	∓7.3

Najbardziej dokładnym modelem regresji opisującym wzrost człowieka jest GLM czyli generalized linear model. Model linowy i SVR dają większe wartości błędu więc gorszej modelują nasze dane. Estymatory w środkowej części przypominają znaną nam świetnie regresje liniową. Natomiast na dwóch krańcach widać pewne odchylenia. Jak jednak widzimy, najlepiej odwzorowujący model ma ciągle dość dużą wartość błędu. Dlatego sprawdzimy czy wartość błędu zmieni się gdy uwzględnimy dane mężczyzn jak i kobiety osobno Tab V.

Tabela V
Parametry i błąd poszczególnych modeli dotyczących 2 eksperymentu

Gender	Model	MSE	MAE
Mężczyzna	Model liniowy	29.7	∓4.23
	GLM	29.6	∓4.24
	SVR	29.7	∓4.06
Kobieta	Model liniowy	54.5	∓5.11
	GLM	60.2	∓5.92
	SVR	56.6	∓5.73

Jak możemy zauważyć, gdy rozważymy dane osobno płciami dostaniemy o wiele lepsze, dokładniejsze modele regresji. Predykcja jest znacznie dokładniejsza od tej dotyczącej całego zbioru. Wzrost mężczyzn jak i kobiet można całkiem trafnie określić powyższymi estymatorami.



Rysunek 6. Modele regresji próbujące opisać wzrost człowieka względem obwodu odcisku palca

C. Model opisujący wagę człowieka względem szerokości odcisku palca

Po analizie współczynnika korelacji Tab. VI możemy wywnioskować że atrybutem najlepiej odwzorowującym wartość wagi jest pomiar szerokości odcisku palca. Dlatego też wybieramy ją na daną naszego modelu.

Tabela VI
Współczynnik korelacji wagi względem różnych danych pomiarowych odcisku palca

Atrybut	Współczynnik
fp_weight	0.360701
fp_width	0.500167
fp_area	0.455277
fp_circ	0.425761

Za pomocą danych testujących zostały stworzone 3 modele regresji a konkretnie regresji liniowej, GLM jak i SVR Rys. 7. Parametry, błąd średniokwadratowy (Mean Square Error (MSE)) jak i uśredniony błąd bezwzględny (Mean Absolute Error (MAE)) tych modeli został ukazany w Tabeli Tab. VII.Patrząc na wykres widać że dane całkiem ładnie układają się w linię, szczególnie w centrum wykresu.

Najbardziej dokładnym modelem opisującym wagę człowieka jest SVR. Model linowy i GLM dają większe wartości błędu więc gorszej modelują nasze dane. Błąd wszystkich estymatorów jest duży, dlatego ponownie sprawdzamy co się stanie w przypadku uwzględnienia płci osobno Tab VIII.Czy nasze modele ponownie poprawią swoją dokładność?

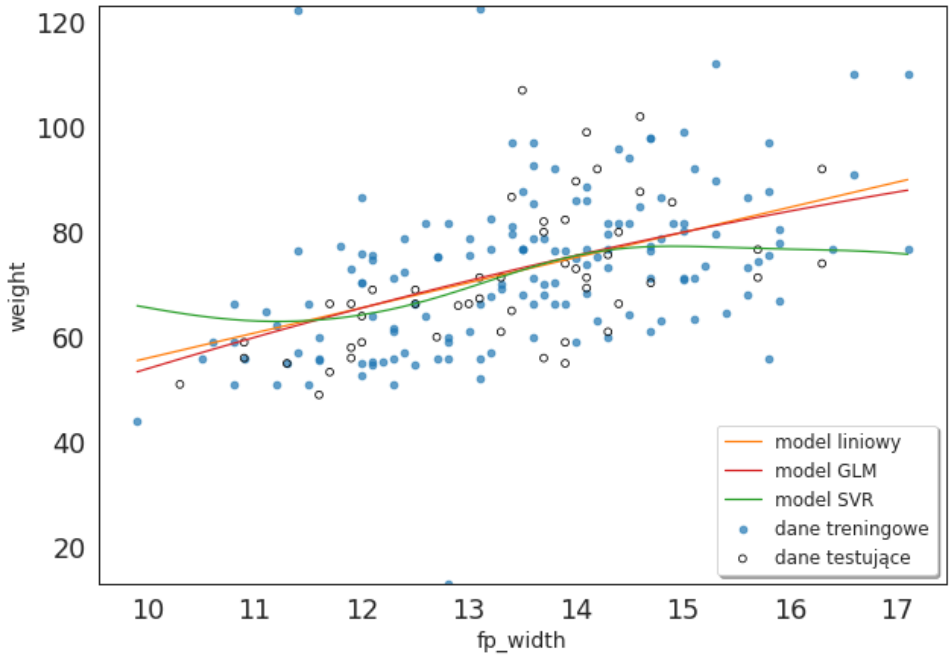
Tabela VII
Parametry i błąd poszczególnych modeli dotyczących 2 eksperymentu

Model	Parametry	MSE	MAE
Model liniowy	[4.78581], 8.14961	120	∓8.47
GLM	[0; 10.0501; -0.1943], -27.07357	118	∓8.4
SVR	———	122	∓8.41

Tabela VIII
Parametry i błąd poszczególnych modeli dotyczących 2 eksperymentu

Gender	Model	MSE	MAE
Mężczyzna	Model liniowy	99.5	∓8.14
	GLM	102	∓8.19
	SVR	108	∓8.15
Kobieta	Model liniowy	221	∓9.88
	GLM	221	∓9.54
	SVR	230	∓10.5

Tym razem podział danych według płci nie skutkował stworzeniu lepszego modelu a w przypadku danych dotyczących kobiet wręcz gorszego. U mężczyzn model polepszył się lecz wartość błędu pozostaje ciągle duża. Podsumowując wagę człowieka jest możliwa do zamodelowania względem odcisku palca. Niestety wartości błędu są na tyle duże, że określenie wagi nie aż tak dokładne.



Rysunek 7. Modele regresji próbujące określić wagę człowieka względem szerokości odcisku palca

D. Dopasowanie płci na podstawie pomiarów wartości odcisku palca

Jak mogliśmy zauważyć dzięki znajomości płci jesteśmy wstanie lepiej określić pewne wartości fizyczne osób. Więc postaramy się ją wyznaczyć, za pomocą danych opisujących wartości parametrów odcisku palca jak i płci uczestników. Na podstawie tych danych tworzymy nasz klasyfikator, będzie nim KNN (K-Nearest Neighbours) z biblioteki scikit-learn. Dzięki niemu określamy klasę/wartość(u nas płeć), robimy to na podstawie odległości od K - sąsiadów w naszym przypadku K jest równe 7. Nasze dane na tych samych zasadach dzielimy na testowe i trenujące co w poprzednich eksperymentach.

Dokładność rozwiązania tego klasyfikatora wynosi 86%. Czyli z całkiem dużą dokładnością możemy określić płeć osoby na podstawie jej odcisku palca a co za tym idzie dokładniej określić profil fizyczny poszukiwanej osoby.

IV. Wnioski

Wnioskując z wykonanych eksperymentów można dowiedzieć się, że można przewidzieć wzrost jak i wagę człowieka za pomocą odcisku palca. Określamy to dzięki modelom regresji takim jak SVR, GLM oraz zwykłej regresji liniowej. Całkiem ważnym aspektem jest zaznaczenie tego że dokładniej można określić wzrost, gdy posiadamy informację jaka jest płeć poszukiwanej przez nas osoby Tab. VIII. Natomiast gdy nie mamy takiej informacji, możemy ją łatwo uzyskać. Dostaniemy ją za pomocą klasyfikatora KNN (Eksperyment III-D) którego danymi są pomiary odcisków. Jesteśmy w stanie z 86% poprawnością określić czy dana osoba jest kobietą czy mężczyzną, tylko za pomocą kilku parametrów dotyczących odcisku osoby. Natomiast określenie wagi osoby nie jest aż tak dokładne za pomocą modeli regresji ponieważ jest obarczone dość dużym błędem.
Podsumowując wartości pomiarów odcisku palca poszukiwanej osoby, mogą służyć do całkiem trafnego określeniu profilu fizycznego poszukiwanej osoby (wzrostu oraz wagi). Jednak musimy się liczyć z tym że nasze modele są obarczone pewnym błędem, w przypadku wzrostu mniejszym a wagi większym.

Literatura

[1] "Height, weight and fingerprint measurements collected from 200 participants," https://repository.lboro.ac.uk/articles/dataset/Height_weight_and_fingerprint_measurements_collected_from_200_participants/7539206/1, dostęp: 03.06.2022.