1. Wstęp.

Cukrzyca jest przewlekłą chorobą metaboliczną, wynikającą z zaburzonego wydzielania lub działania insuliny. Przyczyn cukrzycy jest wiele, nie wynika ona bezpośrednio z jednego czynnika. Wśród nich znajdują się m.in. obciążenie genetyczne, zła dieta i otyłość, małe ilości snu, zespół Crushinga czy stres. Osoby dotknięte cukrzycą mają bardzo wysoki poziom glukozy we krwi, nie jest on widoczny od razu gołym okiem, lecz często towarzyszy mu duże pragnienie, spory apetyt połączony z traceniem wagi, senność i osłabienie oraz nieostre bądź podwójne widzenie.

Choroba ta dzieli się na cukrzycę typu 1, cukrzycę typu 2, cukrzycę ciążową, cukrzycę wtórną (typu 3). Różnią się one zachowaniem insuliny (w typie 1. jest jej brak, za to w typie 2. ona występuje, lecz wydziela się w nieregularny, nieprawidłowy sposób), oraz przebiegiem, np. cukrzyca ciążowa ustaje po urodzeniu dziecka. Cukrzyca typu 3. występuje u osób starszych, w wieku emerytalnym.

Leczenie cukrzycy polega na normowaniu poziomu cukru we krwi, oraz zapobieganiu dalszym powikłaniom.

Dane, które wykorzystujemy podczas naszego projektu pochodzą z badania na kobietach, które są rdzennymi amerykankami wywodzącymi się z kultury "Pima". Wszystkie badane miały co najmniej 21 lat, żyły w pobliżu miasta Phoenix (stolicy Arizony). Dane zostały zebrane przez "US National Institute of Diabetes and Digestive and Kidney Diseases" (Amerykański Narodowy Instytut Badań nad Cukrzycą, Chorobami nerek oraz Chorobami Układu Trawiennego (tłum. red.))

2. Analiza danych.

2.1 Dostępne dane.

Do dyspozycji mamy szereg cech, które są dostępne w plikach z danymi:

npreg – liczba ciąż

glu – stężenie glukozy w osoczu na podstawie doustnego testu tolerancji glukozy

bp – ciśnienie rozkurczowe krwi (mm Hg)

skin – grubość fałdu skóry tricepsu (mm)

bmi – wskaźnik masy ciała (masa wyrażona w $\frac{kg}{wzrost\ w\ m^2}$)

ped – funkcja rodowodu cukrzycy

age – wiek wyrażony w latach

type – no / yes (chory/zdrowy)

2.2 Wektor wejściowy

Wydaje się nam, że w wektorze wejściowym zostaną użyte wszystkie znalezione cechy w folderze z danymi(oprócz TYPE). Mamy niepewności co do ciśnienia rozkurczowego, ale pewnie i ta cecha jest powiązana z cukrzycą. Cecha TYPE nie znajdzie się w wektorze wejściowym – po to budujemy naszą sieć, żeby znaleźć, czy pacjentka jest chora czy zdrowa.

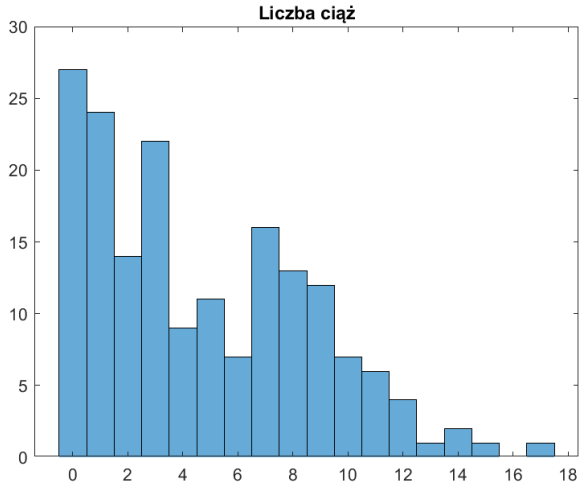
2.3 Podział zbioru danych na dane treningowe i testowe.

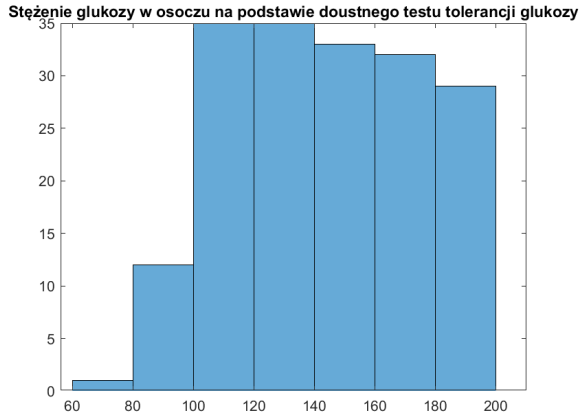
Ten podział został narzucony z góry. Plik pima.tr zawiera w sobie 200 linijek danych treningowych, natomiast w pliku prima.te można znaleźć 332 linijki danych testowych.

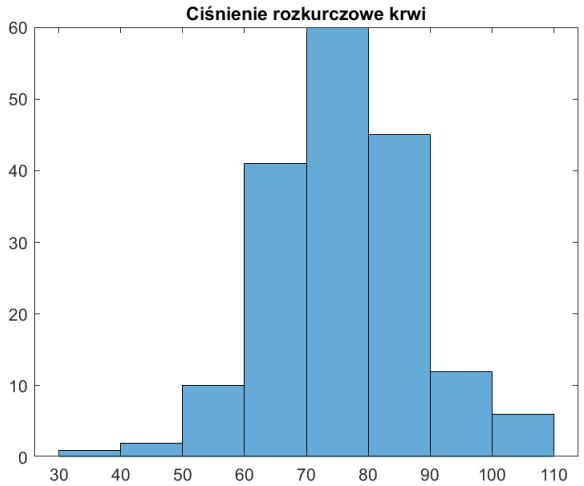
2.4 Zakres zmienności cech

Nazwa	Zakres zmienności dla zdrowych	Zakres zmienności dla chorych	Średnia dla zdrowych	Średnia dla chorych	Odchylenie standardowe dla zdrowych	Odchylenie standardowe dla chorych
Liczba ciąż	0-13	0-17	2.9268	4.7005	2.7872	3.919
Stężenie glukozy	56-197	78-199	110.0169	143.1186	24.2869	31.2650
Ciśnienie rozkurczowe krwi	24-110	30-110	69.9127	74.7006	11.9031	12.5239
Grubość fałdu skóry tricepsu	7-60	7-99	27.2901	32.9774	10.0803	10.3950
Wskaźnik masy ciała	18.20 - 57.30	22.90-67.10	31.4295	35.8198	6.5468	6.6116
Funkcja rodowodu cukrzycy	0.0850-2.3290	0.1270-2.4200	0.4463	0.6166	0.2988	0.3989
Wiek	21-81	21-70	29.2225	36.4124	9.9034	10.8374
Klasa (zdrowy/chory)	–	–	–	–	–	–

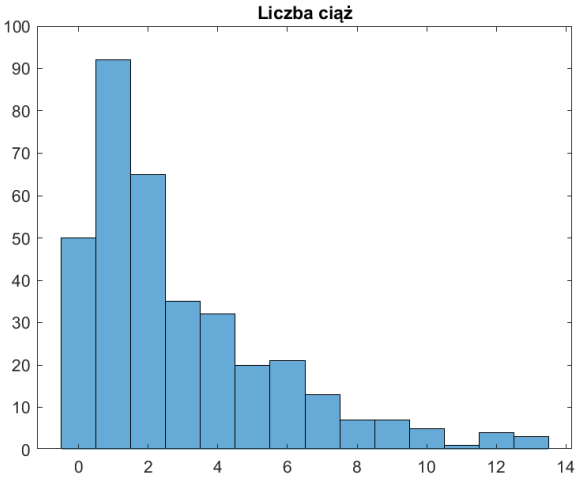
Kobiety chore

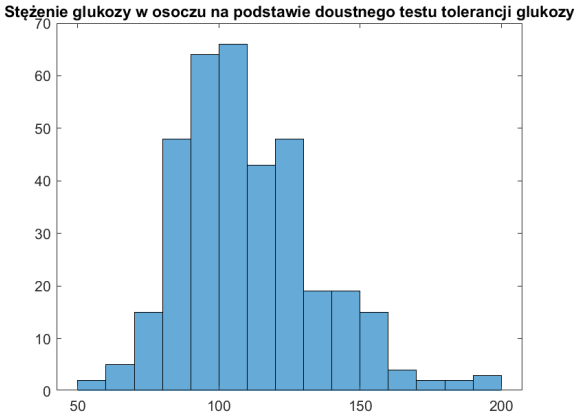


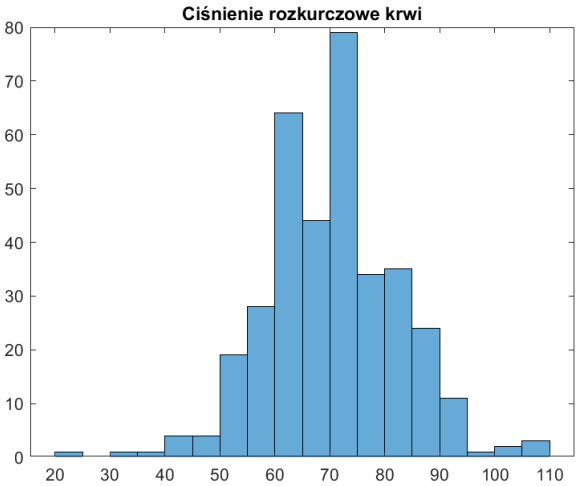


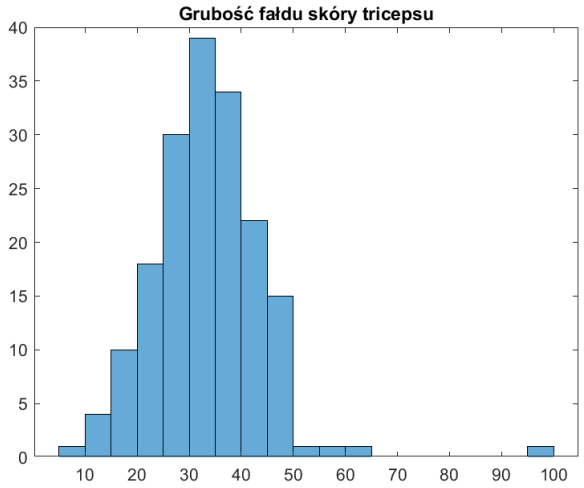


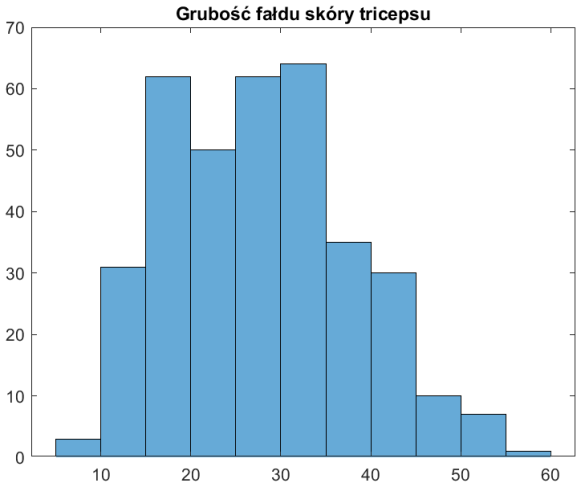
Kobiety zdrowe

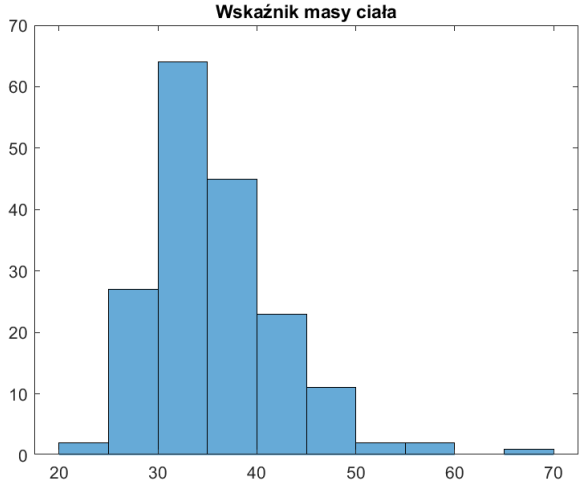


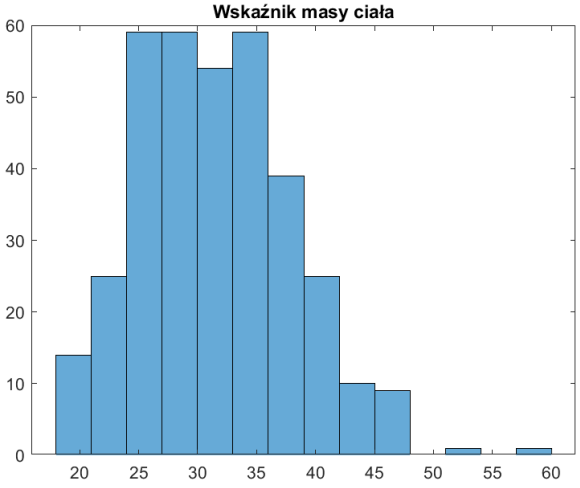


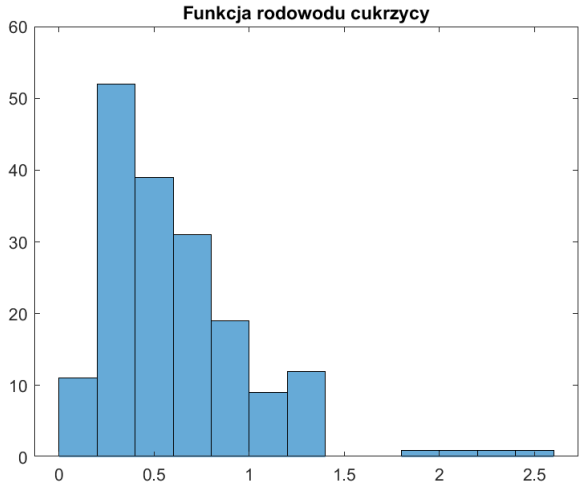


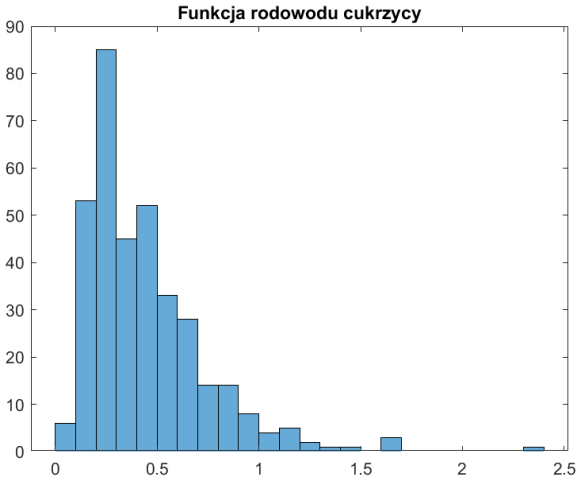

Grubość fałdu skóry tricepsu
0 5 10 15 20 25 30 35 40
10 20 30 40 50 60 70 80 90 100

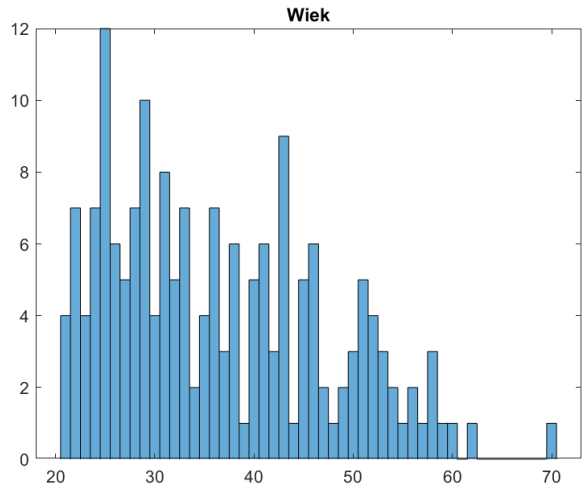

Grubość fałdu skóry tricepsu
0 10 20 30 40 50 60 70
10 20 30 40 50 60

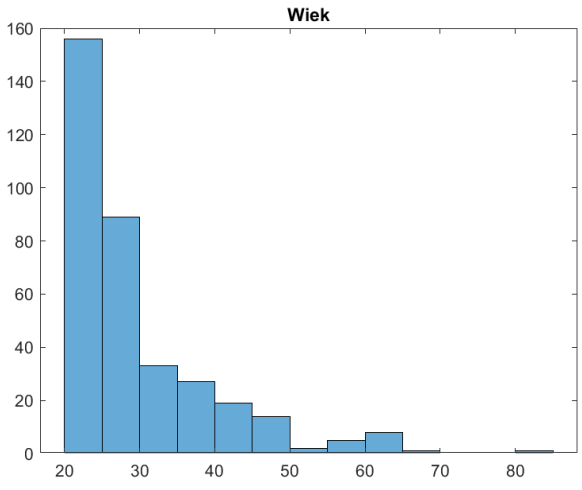

Wskaźnik masy ciała
0 10 20 30 40 50 60 70
20 30 40 50 60 70


Wskaźnik masy ciała
0 10 20 30 40 50 60
20 25 30 35 40 45 50 55 60


Funkcja rodowodu cukrzycy
0 10 20 30 40 50 60
0 0.5 1 1.5 2 2.5


Funkcja rodowodu cukrzycy
0 10 20 30 40 50 60 70 80 90
0 0.5 1 1.5 2 2.5





2.5 Proponowany sposób kodowania danych nienumerycznych

W naszym wypadku jedyna cecha która zawiera w sobie dane nienumeryczne to cecha TYPE. Proponowany sposób kodowania:

Czy jest chory? (zmienna type) - yes → 1,

no → 0.

Warto też podkreślić, że ta cecha nie jest używana w warstwie wejściowej, jest nam potrzebna tylko po to żeby zrozumieć i dobrze umieć oszacować zaimplementowaną sieć. Użyliśmy kodowania, bo nam jest tak po prostu wygodniej, żeby później podzielić dane na osoby chore i zdrowe i wyznaczyć wszystkie wymagane wartości: wartość średnia, odchylenie standardowe itd.

3. Koncepcja realizacji

3.1 Metoda wstępnego przetwarzania danych

Liczbę neuronów w warstwie ukrytej można próbować oszacować wzorem:

$$N_u = \sqrt{N_{we} * N_{wy}}$$

Co w naszym przypadku oznacza $N_u \approx 3$

Warstwa	**Liczba neuronów**
Wejściowa	7
Ukryta	3
Wyjściowa	1

Zwykle uczenie przebiega w ten sposób, że zaczynamy od mniejszej liczby neuronów w warstwie ukrytej, stopniowo ją zwiększając. Wadą zbyt małej liczby neuronów jest to, że sieć nie potrafi poprawnie odwzorować funkcji. Z drugiej strony - zbyt wiele elementów warstwy ukrytej prowadzi do uczenia “na pamięć” i wydłużenia procesu uczenia.

Inną “szkołą” wyznaczania warstwy ukrytej, jest zastosowanie wzoru $N_u = \frac{N_{we}}{2} + N_{wy}$, co w naszym przypadku da liczbę 4.5 (zaokrągloną do 5.).
Podczas implementacji naszej sieci, będziemy próbowały dobrać optymalną liczbę neuronów w warstwie ukrytej, zgodną z naszymi założeniami.

Źródło danych:

Metody Heurystyczne[pdf].
https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjF3M_j8Yz3AhXxo4sKHc77C4IQFnoECAQQAQ&url=http%3A%2F%2Fwww.imio.polsl.pl%2FDopobrania%2FMH%2520ME3%2520wyklad%25204%2520do%2520druku.pdf&usg=AOvVaw216lJLcqW2tlU_6FNuGtUx

3.2 Funkcja aktywacji, algorytm uczenia oraz model sieci

Funkcję aktywacji, na którą się zdecydowałyśmy jest funkcja tangensu hiperbolicznego.

$$y_m = tanh(\beta \cdot v_m)$$

Gdzie z reguły $\beta \in (0, 1]$. Pozwala ona na przyzwoitą szybkość efektywnego wyjścia neuronu.

Do uczenia sieci zostanie zastosowany **algorytm** wstecznej propagacji błędów.
Wybór padł na ten algorytm, ponieważ wykorzystujemy perceptrony wielowarstwowe.
Jego (poglądowy) schemat krokowy wygląda następująco:

1. Inicjalizacja sieci i algorytmu
2. Obliczanie wartości wyjściowej sieci na podstawie danych
3. Obliczanie błędu sieci
4. Korekcja wag
5. Czy sieć nauczona?

Jeśli tak - przechodzimy dalej.
Jeśli nie - wracamy do punktu 2.

6. Koniec

Poglądowy **model sieci**: