# Określenie problemu

Celem projektu jest prognoza dekady wiekowej pacjentów wykorzystując dostępne cechy~~.~~

Problem rozwiąże za pomocą uczenie nadzorowanego, ponieważ każdy przykład uczący ma zdefiniowany oczekiwany wynik. Problem rozwiąże przy pomocy regresji – będę przewidywać wartość liczbową. Możliwa regresja wieloraka, gdyż mogę potrzebować wiele cech do prognozowania. Jest to również regresja jednoczynnikowa – dla każdego pacjenta będę chciała przewidzieć jedną wartość.

Jako miarę wydajności stosować będę RMSE.

# Dane

Dane, które będę używać wygenerowałam z wyników mojego poprawionego, pierwszego projektu. Wgrywam 3 tabele danych: pierwsza tabela zawiera wszystkie dane o każdym pacjencie, druga informacje o oknie z maksymalnym średnim RR a trzecia o oknie z maksymalnym odchyleniem. Zmienne dotyczące pierwszej tabeli będę oznaczać poprzez o, dotyczące drugiej – m (mean) a trzeciej s (std).

We wszystkich przypadkach dane zawierają informacje o 181 pacjentach. Wszystkie atrybuty oprócz płci są numeryczne. Płeć jest atrybutem kategorialnym. Nie ma brakujących danych.

Aby przyjrzeć się lepiej danym stworzyłam tabele zawierające podsumowanie atrybutów numerycznych oraz histogramy tych atrybutów. Zaobserwowałam z nich, że atrybuty ukazane są w różnej skali: w przypadku prawdopodobieństw jest to skala od 0 do 1 a na przykład atrybut RMSSD oraz średnia mają już dużo większą skalę. Niektóre atrybuty prawdopodobieństw mają rozkład długoogonowy.

Następnie dziele dane na zbiory uczące i testowe. Zdecydowałam się zrobić to metodą losowania warstwowego, aby dobrze odzwierciedlić każdą grupę wiekową. Poprzez funkcje „porównanie” sprawdzam, czy losowanie warstwowe przebiegło prawidłowo.

Od tego momentu będę zajmować się po kolei zbiorami danych, zaczynając od zbioru ogólnego.

# Model 1 – zbiór ogólny:

## Analiza danych

Tworzę kopię danych uczących i na nim będę pracować.

Sprawdzam korelację każdego atrybutu z wiekiem. Największą korelację dodatnią wykazują atrybuty: p(0da) 0.616358

p(da0) 0.588166

natomiast ujemną:

p(add) -0.510118

p(dd) -0.513811

p(aad) -0.535412

Następnie przedstawiłam tą korelację na wykresie.

Z macierzy korelacji mogę zauważyć także, że niektóre cechy wykazują względem siebie nawzajem dużą korelację, np. kwartyle. W późniejszych krokach wykonam analizę składowych głównych.

## Przygotowanie danych

Rozdzielam czynniki prognostyczne od etykiety (atrybutu ‘Wiek’). Następnie oddzieliłam dane numeryczne od kategorycznych.

Stworzyłam transformator, który skaluje dane numeryczne oraz zamienia wartości kategoryczne na 0 lub 1 w zależności od płci pacjenta. Do skalowania wybrałam metodę normalizacji, ponieważ spora część danych jest już w skali 0-1 (prawdopodobieństwa) oraz sieci neuronowe często oczekują na wejściu wartości 0-1. Przed użyciem transformatora tworzę 2 zmienne: pierwszą z listą atrybutów numerycznych a drugą z listą atrybutów kategorialnych. Te zmienne będę używać również przy kolejnych modelach, ponieważ wszystkie tabele zawierają te same atrybuty. Transformer zwraca array, więc aby mieć wygodniejszy wgląd w dane przekształciłam je z powrotem do dataframe z nazwanymi kolumnami.

Następnie wykonuje analizę głównych składowych. Wybieram taką liczbę komponentów, aby objaśniały 95% wariancji. W przypadku tego modelu 95% wariancji opisuje 9 komponentów. Następnie przedstawiłam jaki % wariancji jest zachowany wzdłuż kolejnych składowych głównych. Przedstawiłam również te dane na wykresie.

## Wybór i uczenie modelu

Będę korzystać z modelu SGDRegressor poznanego na zajęciach.

Tworzę funkcję do porównywania błędów.

Następnie metodą przeszukiwania siatki wyszukuję najlepsze hiper parametry. Ponieważ po pierwszym przeszukiwaniu najlepszymi hiper parametrami okazały się być największe lub najmniejsze z podanych wartości numerycznych, ponownie przeszukuję siatkę.

Dokonuje sprawdzianu krzyżowego dla ustalonych hiperparametrów.

Gdy uzyskałam już finalny model, oceniam go za pomocą zbioru testowego. W tym celu oddzielam etykiety od czynników prognostycznych, za pomocą wcześniej utworzonego transformera normalizuje dane oraz dokonuje redukcji wymiarów do wymiaru, który wyszedł przy zbiorze trenującym.

Otrzymałam błąd 14.5 oraz przedział ufności: [11.15255456 , 17.25405972]

# Model 2 – dane dla okien o max średnim RR

Będę pracować analogicznie do modelu 1.

## 3. Analiza danych

Tworzę kopię danych uczących i na nim będę pracować.

Sprawdzam korelację każdego atrybutu z wiekiem. Największą korelację dodatnią wykazują atrybuty: p(0d0) 0.420036

natomiast ujemną:

std -0.444582

RMSSD -0.469231

pNN20 -0.476106

pNN50 -0.520601

Następnie przedstawiłam tą korelację na wykresie.

W tym przypadku jest mniej zmiennych o dużej korelacji, więc prawdopodobnie będę potrzebować więcej składowych głównych.

## Przygotowanie danych

Rozdzielam czynniki prognostyczne od etykiety (atrybutu ‘Wiek’). Następnie oddzieliłam dane numeryczne od kategorycznych.

Korzystam z wcześniej utworzonego transformatora.

Następnie wykonuje analizę głównych składowych. Wybieram taką liczbę komponentów, aby objaśniały 95% wariancji. W przypadku tego modelu 95% ~~zmienności~~ wariancji opisuje 16 komponentów. Można zaobserwować, że w tym przypadku kolejne składowe główne wyjaśniają mniejszy % wariancji niż w przypadku 1 modelu, dlatego potrzebujemy więcej składowych głównych.

## Wybór i uczenie modelu

Ponownie będę korzystać z modelu SGDRegressor.

Następnie metodą przeszukiwania siatki wyszukuję najlepsze hiper parametry. Po pierwszym przeszukaniu najlepszą wartością eta jest 0.001. Dokonuje ponownego przeszukiwania siatki.

Dokonuje sprawdzianu krzyżowego.

Gdy uzyskałam już finalny model, oceniam go za pomocą zbioru testowego. W tym celu oddzielam etykiety od czynników prognostycznych, za pomocą wcześniej utworzonego transformera normalizuje dane oraz dokonuje redukcji wymiarów do wymiaru, który wyszedł przy zbiorze trenującym.

Otrzymałam błąd 13.96 oraz przedział ufności: [10.71835672, 16.58126633]

# Model 3 – dane dla okien o max odchyleniu standardowym

## 3. Analiza danych

Tworzę kopię danych uczących i na nim będę pracować.

Sprawdzam korelację każdego atrybutu z wiekiem. Największą korelację dodatnią wykazują atrybuty: p(0da) 0.338157

natomiast ujemną:

max -0.339862

pNN50 -0.367990

std -0.589627

SDNN -0.589636

Następnie przedstawiłam tą korelację na wykresie.

Ponownie mamy mało zmiennych o dużej korelacji, więc będziemy potrzebować wielu składowych głównych, aby opisać 95% wariancji.

## Przygotowanie danych

Rozdzielam czynniki prognostyczne od etykiety (atrybutu ‘Wiek’). Następnie oddzieliłam dane numeryczne od kategorycznych.

Korzystam z wcześniej utworzonego transformatora.

Następnie wykonuje analizę głównych składowych. Wybieram taką liczbę komponentów, aby objaśniały 95% wariancji. W przypadku tego modelu 95% wariancji opisuje 18 komponentów.Można zaobserwować, że w tym przypadku kolejne składowe główne wyjaśniają mniejszy % wariancji niż w przypadku 1 modelu, dlatego potrzebujemy więcej składowych głównych.Tworzę wykresy tak samo jak w przypadku poprzednich modeli.

## Wybór i uczenie modelu

Ponownie będę korzystać z modelu SGDRegressor.

Następnie metodą przeszukiwania siatki wyszukuję najlepsze hiper parametry. Po pierwszym przeszukaniu najlepszą wartością ety jest 0.001, dokonuje ponownego przeszukania.

Dokonuje sprawdzianu krzyżowego.

Gdy uzyskałam już finalny model, oceniam go za pomocą zbioru testowego. W tym celu oddzielam etykiety od czynników prognostycznych, za pomocą wcześniej utworzonego transformera normalizuje dane oraz dokonuje redukcji wymiarów do wymiaru, który wyszedł przy zbiorze trenującym.

Otrzymałam błąd 11.95 oraz przedział ufności [ 9.59738631, 13.91414359]

Na sam koniec tworzę tabelkę porównującą błędy. Najmniejszy błąd otrzymałam dla modelu 3 a największy dla modelu 1.