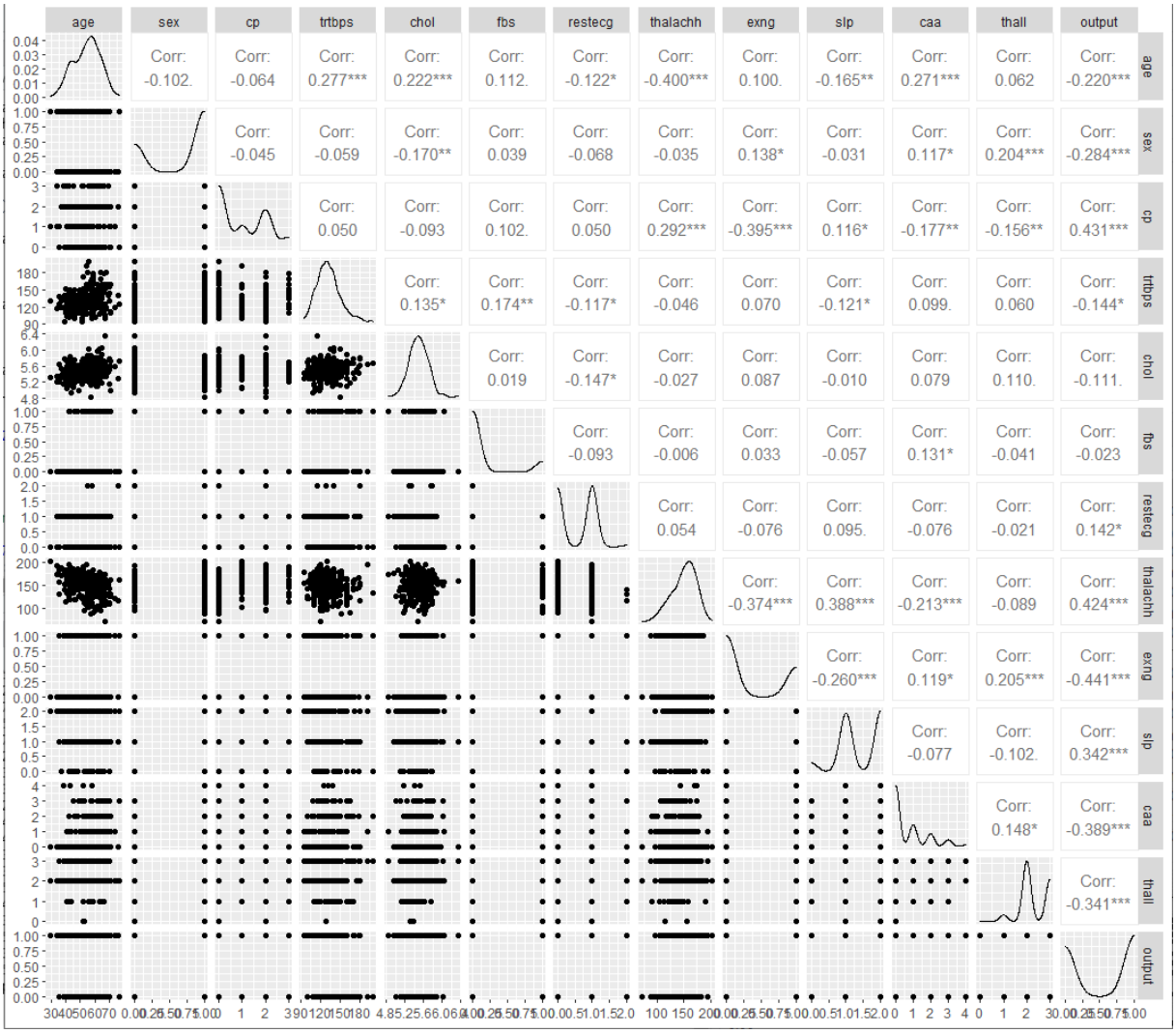
Budowa modeli predykcji wystąpienia ataku serca

Dane, na których zostały zbudowane dwa modele predykcji, składają się z informacji o 300 osobach płci męskiej i żeńskiej. Dane osób, na których będziemy bazować to między innym:

- Wiek
- Płeć
- Ból w klatce piersiowej
- Ciśnienie krwi
- Cholesterol
- Maksymalne tętno

Dane zostały podzielone na dwa zbiory – uczący i testowy – zawierające po 150 osób oraz zostały sprawdzone pod kątem normalizacji, zależności zmiennych i odstałych danych, które mogłyby zaburzyć analizę .



Modele, które zostały zbudowane to:

- Model regresji logistycznej

  Model regresji logistycznej jest szczególnym przypadkiem uogólnionego modelu liniowego. Znajduje zastosowanie, gdy zmienna zależna jest dychotomiczna, to znaczy przyjmuje tylko dwie wartości takie jak na przykład sukces lub porażka, wystąpienie lub brak pewnej jednostki chorobowej. W zapisie matematycznym wartości te reprezentowane są jako 1 i 0.

- Drzewo decyzyjne

  Korzystają one z graficznego przedstawienia decyzji i ich możliwych konsekwencji. Drzewa decyzyjne są tak skonstruowane aby pomóc w podejmowaniu decyzji.

- Model kNN

  Algorytm klasyfikacji opierający swoją predykcję na k najbliższych obserwacjach względem parametrów obserwacji, którą chcemy odpowiednio przyporządkować.

Pierwszy model to model regresji logistycznej oparty na wszystkich zmiennych, do których mamy dostęp. Szczegółowe informacje modelu wskazują, które zmienne są mniej, a które bardziej istotne.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 15.890296  10.453621   1.520 0.128492
age          0.003564   0.041586   0.086 0.931712
sex         -1.848839   0.827165  -2.235 0.025407 *
cp           1.029712   0.310648   3.315 0.000917 ***
trtbps      -0.008887   0.019436  -0.457 0.647479
chol        -3.009413   1.824420  -1.650 0.099042 .
fbs         -0.828193   0.931068  -0.890 0.373730
restecg      1.432288   0.616161   2.325 0.020097 *
thalachh     0.034682   0.018967   1.829 0.067465 .
exng        -0.727678   0.698688  -1.041 0.297647
slp          0.849990   0.443745   1.915 0.055430 .
caa         -0.845364   0.319425  -2.647 0.008133 **
thall       -1.645138   0.502856  -3.272 0.001069 **
```

Na tej podstawie został zbudowany drugi model regresji logistycznej, składający się tylko z najbardziej istotnych zmiennych. Model ten został poddany próbie predykcji ataku serca na zbiorze testowym. Skuteczność wyniosła 79%. Szczegółowe informacje tego modelu wskazały jedną, mniej istotną zmienną.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   19.8811     9.2064   2.159 0.030812 *
dane$sex      -1.5330     0.7259  -2.112 0.034697 *
dane$cp        1.0544     0.2792   3.776 0.000159 ***
dane$chol     -3.0569     1.5950  -1.917 0.055294 .
dane$restecg   1.3601     0.5654   2.406 0.016141 *
dane$thall    -1.7723     0.4684  -3.784 0.000154 ***
dane$slp       1.2180     0.4128   2.951 0.003169 **
dane$caa      -0.9007     0.2985  -3.017 0.002550 **
```

Na tej podstawie został zbudowany kolejny model, aby możliwie jak najbardziej poprawić skuteczność wcześniejszego modelu. Wynik nie zmienił się – 79%. W opisie modelu znalazła się jeszcze jedna zmienna, która ma znacznie mniejszy wpływ niż reszta – płeć.

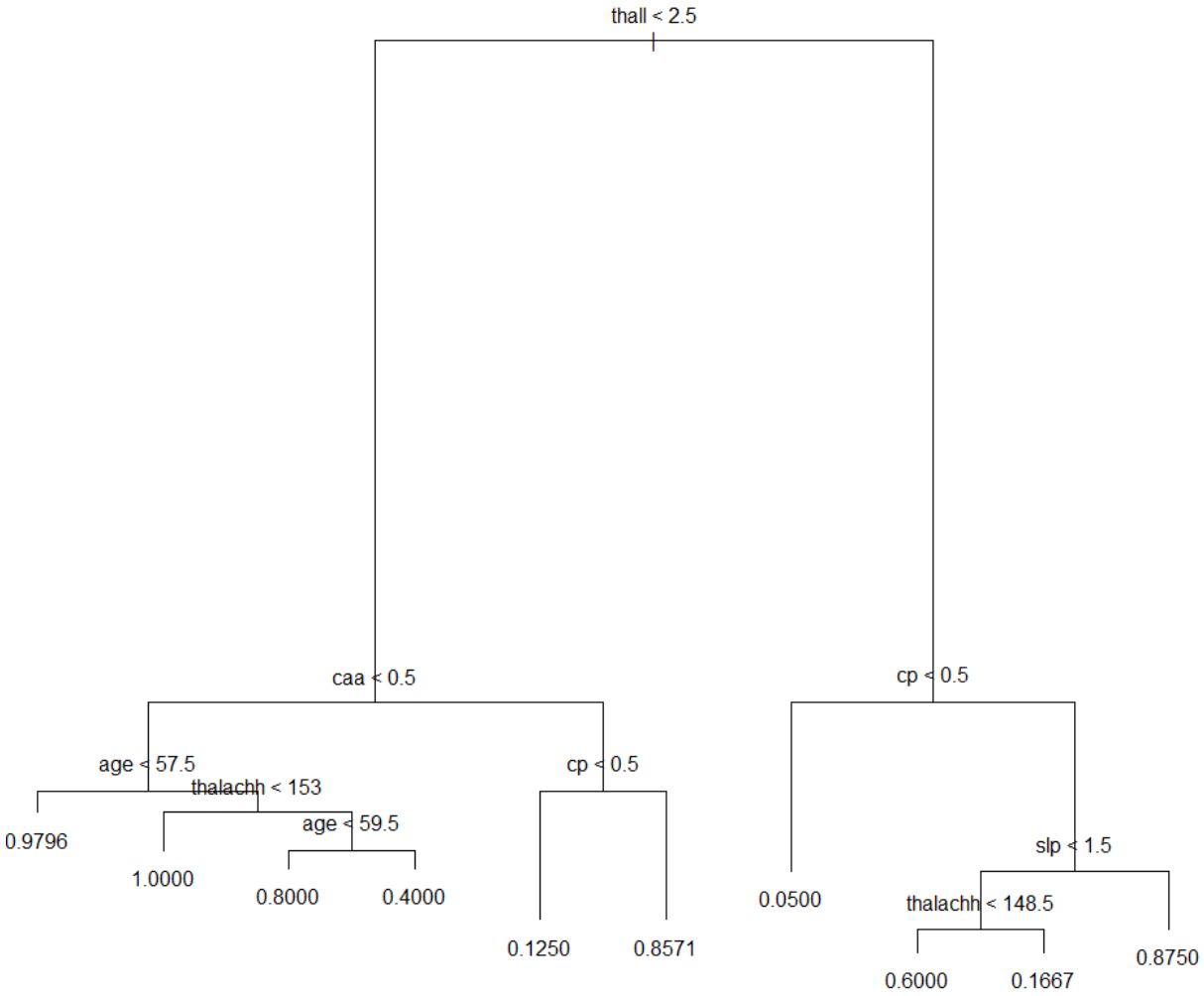
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.6321     1.2872   2.045 0.040872 *
dane$cp        1.0695     0.2740   3.903  9.5e-05 ***
dane$sex      -1.0831     0.6413  -1.689 0.091205 .
dane$restecg   1.5081     0.5547   2.719 0.006556 **
dane$thall    -1.7502     0.4554  -3.843 0.000121 ***
dane$slp       1.1973     0.4058   2.950 0.003173 **
dane$caa      -0.8380     0.2812  -2.981 0.002876 **
```

Ostatecznie, model został zbudowany tylko z najistotniejszych zmiennych.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.2421     1.2735   1.761 0.078312 .
dane$cp        1.0512     0.2709   3.880 0.000104 ***
dane$restecg   1.5210     0.5485   2.773 0.005550 **
dane$thall    -1.8987     0.4627  -4.103 4.08e-05 ***
dane$slp       1.1675     0.4008   2.913 0.003585 **
dane$caa      -0.8954     0.2852  -3.140 0.001691 **
```

Jego skuteczność nie zmieniła się względem wcześniejszych modeli. Otrzymana dokładność to 79%.

Do porównania wyników został stworzony inny model – drzewo decyzyjne. Inicjalizacja drzewa została oparta na wszystkich zmiennych. Ostatecznie algorytm wybrał tylko 6 z 13. Zmienne, które zostały wzięte pod uwagę trochę się różnią od tych, które zostały użyte w modelu regresji logistycznej. Wspólne zmienne obu modeli to między innymi: ból w klatce piersiowej, caa oraz slp.



Ten model również został poddany testowi predykcji ataku serca u pewnych osób. Jego skuteczność wynosiła 75%.

Do uzyskania jeszcze lepszej skuteczności w predykcji ataku serca, został stworzony trzeci model – kNN. Pierwszy model predykcji korzystał ze wszystkich zmiennych, które w żaden sposób nie zostały przeskalowane. Dokładność tego modelu wyniosła zaledwie 67%. Druga próba predykcji opierała się na zmiennych, które były bardzo istotne w modelu regresji logistycznej. Skuteczność tego modelu

została sprawdzona dla różnych k i maksymalna wartość była równa 93%. Może to wynikać z faktu, że wykorzystane dane w tym modelu były z przedziału [0,3], przez co były w pewnym sensie znormalizowane.

```
  [1] 0.9200000 0.9333333 0.9066667 0.9000000 0.9066667 0.9000000 0.8866667 0.8866667 0.8866667
 [10] 0.8866667 0.8800000 0.8733333 0.8866667 0.8866667 0.8866667 0.8666667 0.8800000 0.8666667
 [19] 0.8600000 0.8600000 0.8666667 0.8600000 0.8600000 0.8600000 0.8600000 0.8600000 0.8533333
 [28] 0.8466667 0.8466667 0.8400000 0.8466667 0.8466667 0.8400000 0.8466667 0.8466667 0.8466667
 [37] 0.8400000 0.8533333 0.8400000 0.8533333 0.8400000 0.8400000 0.8466667 0.8400000 0.8400000
 [46] 0.8466667 0.8333333 0.8466667 0.8266667 0.8266667 0.8333333 0.8333333 0.8333333 0.8333333
 [55] 0.8400000 0.8400000 0.8400000 0.8400000 0.8400000 0.8400000 0.8466667 0.8200000 0.8200000
 [64] 0.8200000 0.8200000 0.8200000 0.8200000 0.8200000 0.8200000 0.8200000 0.8200000 0.8066667
 [73] 0.8000000 0.8000000 0.8000000 0.8000000 0.8000000 0.8000000 0.8000000 0.8000000 0.7933333
 [82] 0.8000000 0.7933333 0.7800000 0.7733333 0.7800000 0.7800000 0.7666667 0.7666667 0.7466667
 [91] 0.7400000 0.7466667 0.7400000 0.7466667 0.7466667 0.7266667 0.7266667 0.7200000 0.7066667
[100] 0.7066667 0.7000000 0.7000000 0.6666667 0.6666667 0.6600000 0.6600000 0.6600000 0.6200000
[109] 0.5933333 0.5866667 0.5866667 0.5866667 0.5600000 0.5333333 0.5333333 0.5133333 0.5133333
[118] 0.5066667 0.5066667 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
[127] 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
[136] 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
[145] 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
```

Dla porównania został stworzony ten sam model, ale opierający się na wszystkich dostępnych danych, które zostały znormalizowane do przedziału [0,1]. Skuteczność tego modelu wyniosła 100%.

```
  [1] 0.9933333 0.9933333 0.9933333 0.9933333 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [10] 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333
 [19] 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333
 [28] 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333 0.9933333
 [37] 0.9933333 0.9800000 0.9800000 0.9666667 0.9666667 0.9666667 0.9666667 0.9666667 0.9666667
 [46] 0.9600000 0.9600000 0.9666667 0.9600000 0.9600000 0.9600000 0.9600000 0.9600000 0.9600000
 [55] 0.9600000 0.9600000 0.9666667 0.9666667 0.9666667 0.9600000 0.9533333 0.9533333 0.9533333
 [64] 0.9600000 0.9600000 0.9600000 0.9600000 0.9600000 0.9600000 0.9600000 0.9600000 0.9533333
 [73] 0.9533333 0.9533333 0.9600000 0.9600000 0.9600000 0.9533333 0.9600000 0.9600000 0.9666667
 [82] 0.9666667 0.9533333 0.9466667 0.9466667 0.9466667 0.9466667 0.9466667 0.9466667 0.9466667
 [91] 0.9466667 0.9333333 0.9266667 0.9266667 0.9266667 0.9200000 0.9200000 0.9200000 0.9266667
[100] 0.9200000 0.9200000 0.9333333 0.9266667 0.9266667 0.9133333 0.9000000 0.9000000 0.8933333
[109] 0.8800000 0.8666667 0.8533333 0.8333333 0.8200000 0.8266667 0.8066667 0.8000000 0.7933333
[118] 0.7533333 0.7266667 0.6266667 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
[127] 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
[136] 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
[145] 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000 0.5000000
```

Oba modele – model regresji logistycznej i drzewo decyzyjne – sprawdziły się tak samo dobrze, osiągając bardzo podobną skuteczność. Chcąc przewidzieć atak serca pacjenta, mając taki sam zestaw informacji, który został wykorzystany przy budowie obu modeli, mamy odpowiednio 79% lub 75% szans na to, że u pacjenta faktycznie wystąpi atak serca. Różnice w skuteczności modeli mogą wynikać z doboru różnych zmiennych do utworzenia predykcji. Najlepiej natomiast sprawdził się model kNN, osiągając 100% dokładności w dopasowaniu. Może to wynikać z faktu, iż algorytm ten jest odporny na odstające dane, w przeciwieństwie do modelu regresji logistycznej.