Analiza danych jakościowych

Agnieszka Tracz

1. Cel analizy

Celem mojej analizy jest znalezienie kluczowych czynników wpływających na śmierć z powodu niewydolność serca oraz stworzenie modelu predykcyjnego, który pozwoli na skuteczną identyfikację pacjentów zwiększonego ryzyka śmierci.

2. Baza danych

Wykorzystuję do tego celu bazę danych, którą pobrałam ze strony:

https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

Dane zostały zebrane od 299 pacjentów w okresie od kwietnia do grudnia 2015 roku w Faisalabad Institute of Cardiology oraz w Allied Hospital w Faisalabad, w stanie Punjab, w Pakistanie.

Zbiór ten zawiera 299 obserwacji i cechy takie jak:

- Age: Wiek pacjentów.

- Anaemia: Wskazuje obecność (1) lub brak (0) anemii u pacjentów.

- Creatinine Phosphokinase: Poziom enzymu kreatynokinazy (CK) w organizmie pacjentów.

- Diabetes: Wskazuje obecność (1) lub brak (0) cukrzycy u pacjentów.

- Ejection Fraction: Frakcja wyrzutowa, czyli procent krwi wyrzucanej przez lewą komorę serca podczas skurczu.

- High Blood Pressure: Wskazuje obecność (1) lub brak (0) nadciśnienia u pacjentów.

- Platelets: Liczba płytek krwi w organizmie pacjentów.

- Serum Creatinine: Poziom kreatyniny we krwi.

- Serum Sodium: Poziom sodu we krwi.

- Sex: Płeć pacjentów, gdzie 0 oznacza kobietę, a 1 mężczyznę.

- Smoking: Wskazuje obecność (1) lub brak (0) palenia tytoniu u pacjentów.

- Time: Czas obserwacji pacjentów od momentu rozpoczęcia badania.

- Death Event: Wskazuje, czy pacjent zmarł (1) czy przeżył (0) do zakończenia badania.

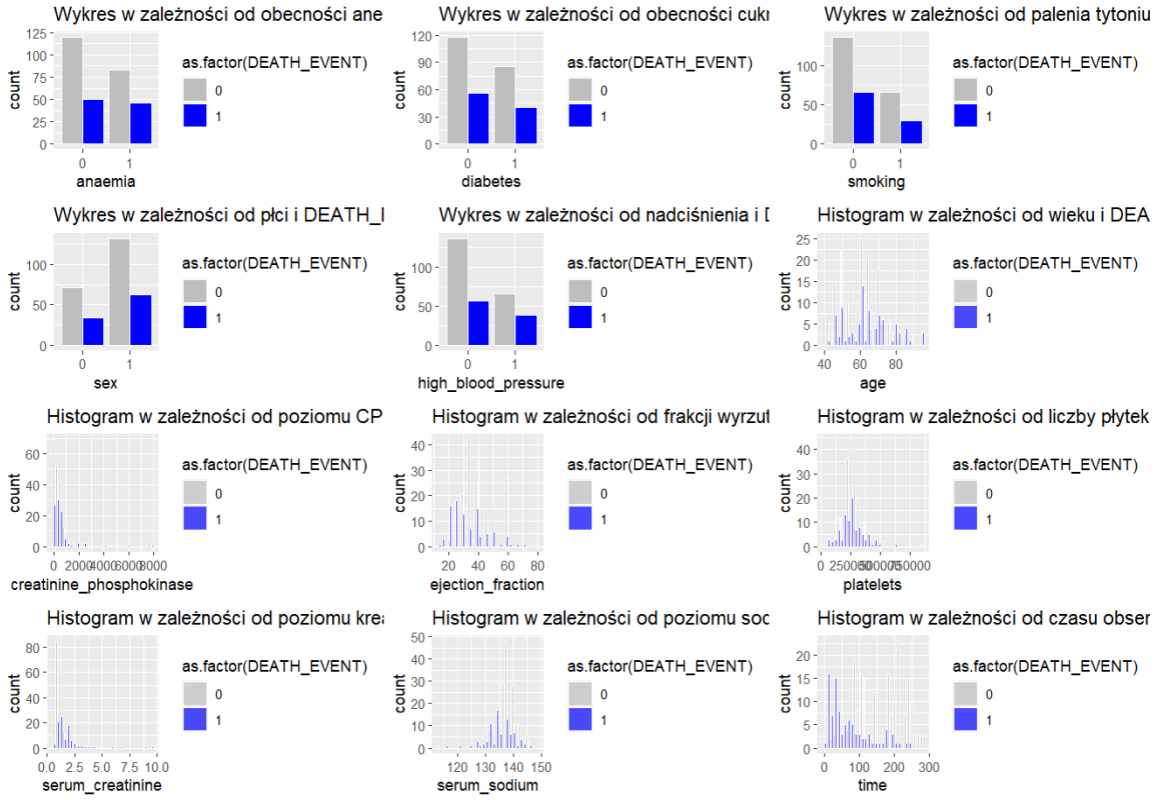
3. Przygotowanie danych do analizy

W celu usprawnienia analizy, dokonałam przekształceń niektórych zmiennych numerycznych na jakościowe w pierwotnej bazie danych. Zmiany wprowadziłam w przypadku zmiennych takich jak: obecność anemii, obecność cukrzycy oraz zmiennej objaśnianej DEATH_EVENT (śmierć pacjenta), która została przekształcona w kategoryczną zmienną binarną.

Te modyfikacje mają na celu poprawę interpretowalności danych i zwiększenie skuteczności analizy.

4. Wstępna analiza danych

4.1 Podstawowe wykresy:



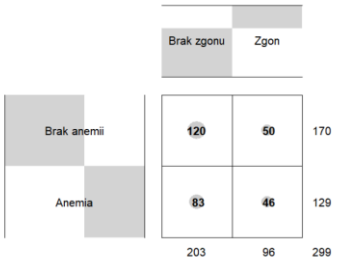
Analiza powyższych wykresów wskazuje, że obecność anemii, występowanie nadciśnienia, wiek oraz czas obserwacji pacjenta mogą być istotnymi czynnikami wpływającymi na ryzyko zgonu pacjentów. Dodatkowo, z powyższych wykresów można wysunąć wnioski, że zmienne takie jak płeć, cukrzyca i status palenia raczej nie powinny mieć istotnego wpływu na zmienną objaśnianą. Co do reszty zmiennych, na tym etapie analizy nie jestem w stanie wyciągnąć dodatkowych wniosków.

4.2 Analiza tablic kontyngencji

Analiza tablic kontyngencji częściowo potwierdziła moje początkowe założenia. Zmienne takie jak wiek i czas obserwowania pacjenta okazały się być istotnymi czynnikami wpływającymi na zmienną objaśnianą. Natomiast zmienne takie jak płeć, cukrzyca i status palenia nie są istotnymi czynnikami w tej analizie.

Podczas analizy wykresów popełniłam błąd co do zmiennych anemia i nadciśnienie, które okazały się nie być istotnymi zmiennymi.

Wyniki testów dla obydwóch zmiennych były bardzo podobne dlatego pokaże tylko wyniki dla pierwszej z nich, czyli obecności anemii u pacjenta.



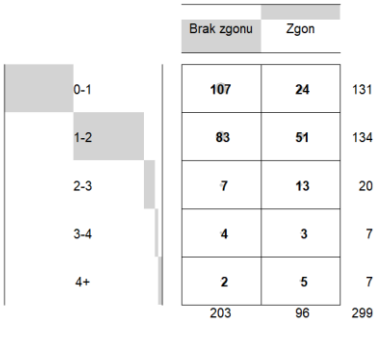

```
                          X^2 df P(> X^2)
Likelihood Ratio 1.3086  1  0.25265
Pearson          1.3131  1  0.25183

Phi-Coefficient   : 0.066
Contingency Coeff.: 0.066
Cramer's V        : 0.066
```

Na podstawie wyników testu ilorazu wiarygodności oraz testu chi-kwadrat Pearsona mogę stwierdzić brak istotnego związku między występowaniem anemii, a ryzykiem zgonu pacjenta.

Analiza tablic kontyngencji wykazała, że istnieją także inne istotne zmienne związane z ryzykiem zgonu pacjenta. Tymi czynnikami są: frakcja wyrzutowa, poziom kreatyniny we krwi oraz poziom sodu we krwi.

W celu ułatwienia interpretacji wyników, zmienną „poziom kreatyniny we krwi” podzieliłam na kategorie, gdzie przedstawiłam liczbę przypadków braku zgonu i zgonu w poszczególnych grupach stężeń: 0-1, 1-2, 2-3, 3-4, 4+.




```
                          X^2 df   P(> X^2)
Likelihood Ratio 28.697  4 9.0057e-06
Pearson          28.864  4 8.3325e-06

Phi-Coefficient   : NA
Contingency Coeff.: 0.297
Cramer's V        : 0.311
```

Wyniki testów potwierdziły istotną statystyczną zależność pomiędzy analizowanymi zmiennymi, co wskazuje na istotny związek pomiędzy poziomem kreatyniny we krwi, a ryzykiem zgonu pacjenta.

4.3 Interpretacja ilorazów szans dla istotnych zmiennych w analizie:

4.3.1 Poziom kreatyniny we krwi.

Po analizie wartości procentowych zauważam, że istnieje pewna zależność: pacjenci z wyższym poziomem kreatyniny mieli większą szansę na zgon w porównaniu do tych z niższym poziomem.

	0-1	1-2	2-3	3-4	4+
Brak zgonu	0.8167939	0.6194030	0.3500000	0.5714286	0.2857143
Zgon	0.1832061	0.3805970	0.6500000	0.4285714	0.7142857

Potwierdza to również iloraz szans.

Wynik jest fascynujący, zwłaszcza gdy porównujemy pierwszą i ostatnią grupę. Pacjenci z poziomem kreatyniny 0-1 mają około 11 razy większe szanse na przeżycie niż ci z poziomem 4+.

Brak zgonu	Zgon
11.14583333	0.08971963

4.3.2 Poziom sodu we krwi.

Analiza proporcji poziomu sodu we krwi wskazuje, że pacjenci z zakresem 136-145 mają znacznie więcej procent szans na przeżycie w porównaniu do pozostałych grup. Iloraz szans pomiędzy grupą o poziomie sodu 0-135 a 136-140 wskazuje, że pacjenci z poziomem sodu 0-130 mają około 4 razy mniejszą szansę na przeżycie w porównaniu do tych z poziomem sodu 136-140.

4.3.3 Czas obserwacji pacjentów.

Analiza tablicy kontyngencji dla zmiennej czas obserwacji sugeruje, że ryzyko zgonu zależy od czasu obserwacji. Ilorazy szans wykazują istotne różnice między grupami, szczególnie w przypadku porównania grupy obserwowanej od 40 do 50 dni z grupą obserwowaną od 80 do 90 dni, gdzie szansa na przeżycie jest ponad pięciokrotnie większa dla pierwszej grupy

4.3.4 Wiek pacjentów.

Analiza tablicy kontyngencji dla zmiennej wiek wskazuje, że ryzyko zgonu rośnie wraz z wiekiem pacjentów. Ilorazy szans ukazują istotną różnicę między grupami wiekowymi, zwłaszcza pomiędzy osobami poniżej 70 roku życia a tymi powyżej.

4.3.4 Frakcja wyrzutowa.

Analiza tablicy kontyngencji dla zmiennej frakcja wyrzutowa wskazuje na istnienie zależności między tą zmienną, a ryzykiem zgonu pacjenta. Szczególnie istotną różnicę obserwujemy między grupą 0-20 a 21-40, gdzie szansa na przeżycie dla drugiej grupy jest około 16 razy większa niż dla pierwszej.

5. Model GLM()

W tym etapie przeprowadziłam analizę kilku modeli regresji logistycznej. Wybrałam te modele ponieważ zmienna objaśniana jest zmienną binarną. Rozpoczęłam od bazowych modeli, czyli takich które uwzględniały wszystkie dostępne zmienne lub tylko część z nich. Następnie rozszerzyłam modele o interakcje rzędu 2 i 3 między czynnikami.

Ostateczny wybór modelu oparłam na kryterium informacyjnym Akaike (AIC), które sugeruje, że model logit2, uwzględniający zmienne takie jak: wiek, czas obserwacji pacjenta, poziom kreatyniny i sodu we krwi oraz frakcja wyrzutowa jest najlepszym modelem ze względu na uzyskanie najniższego poziomu AIC.

```
Call:
glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
    time + serum_sodium, family = binomial, data = dane)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1590  -0.5888  -0.2281   0.5144   2.7959

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        9.493034   5.405768   1.756  0.07907 .
age                0.042466   0.015030   2.825  0.00472 **
ejection_fraction -0.073430   0.015785  -4.652 3.29e-06 ***
serum_creatinine   0.685990   0.174044   3.941 8.10e-05 ***
time              -0.020895   0.002916  -7.166 7.74e-13 ***
serum_sodium      -0.064557   0.038377  -1.682  0.09254 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 375.35  on 298  degrees of freedom
Residual deviance: 223.49  on 293  degrees of freedom
AIC: 235.49

Number of Fisher Scoring iterations: 6
```

W kolejnym kroku sprawdzam dopasowanie modelu, wykorzystując dwa różne testy: test wiarygodności (Likelihood Ratio Test) i test chi-kwadrat.

```
> lrtest(logitNULL,logit2)
Likelihood ratio test

Model 1: DEATH_EVENT ~ 1
Model 2: DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + time +
    serum_sodium
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   1 -187.67
2   6 -111.74  5 151.86  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(logitNULL,logit2,test="Chisq")
Analysis of Deviance Table

Model 1: DEATH_EVENT ~ 1
Model 2: DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + time +
    serum_sodium
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       298     375.35
2       293     223.49  5   151.86 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Oba testy potwierdzają, że uwzględnienie zmiennych w modelu przynosi statystycznie istotne ulepszenie dopasowania.

Sprawdzę teraz jak zachowuje się uproszczony model (zawierający mniej zmiennych) w porównaniu do modelu zawierającego wszystkie zmienne.

```
Analysis of Deviance Table

Model 1: DEATH_EVENT ~ age + ejection_fraction + serum_creatinine + time +
    serum_sodium
Model 2: DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase + diabetes +
    ejection_fraction + high_blood_pressure + platelets + serum_creatinine +
    serum_sodium + sex + smoking + time
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       293     223.49
2       286     219.55  7   3.9321   0.7876
```

Dodanie dodatkowych zmiennych do Modelu 2 nie przyniosło istotnej poprawy w dostosowywaniu modelu do danych w porównaniu do Modelu 1.

Podobne wyniki otrzymałam dla pozostałych testów.

5.1 Predykcja na zbiorze testowym

W celu zbadania mocy predykcyjnej powyższego modelu podzieliłam moja bazę danych na zbiór treningowy i testowy.

Na pierwszym z nich buduję model, który następnie sprawdzam na zbiorze testowym. Przedstawię teraz wyniki predykcji dla poszczególnych modeli:

- Model logit2

Model wykorzystuje zmienne takie jak: wiek, czas obserwacji pacjenta, poziom kreatyniny i sodu we krwi oraz frakcja wyrzutowa i jest to najlepszy model pod względem najniższego wyniku AIC równego 235.49 na całym zbiorze danych.

Tablica z jego wynikami wygląda następująco:

```
             Aktualne
Przewidziane  0  1
           0 73 10
           1 10 26
```

Dokładność tego modelu wynosi: 83.19%

- Model logit1 (pełny)

Wynik AIC dla tego modelu na pełnym zbiorze danych jest nieco wyższy, ponieważ wynosi 245.55. Natomiast wyniki jego predykcji wyglądają następująco:

```
             Aktualne
Przewidziane  0  1
           0 70 10
           1 13 26
```

Dokładność tego modelu wynosi: 80.67%

Podsumowując, model logit2 wykazał się najlepszymi zdolnościami predykcyjnymi w porównaniu do wszystkich pozostałych modeli.

6. Podsumowanie

Celem mojej analizy było zidentyfikowanie kluczowych czynników wpływających na śmierć pacjentów z powodu niewydolności serca oraz stworzenie skutecznego modelu predykcyjnego umożliwiającego identyfikację pacjentów zwiększonego ryzyka zgonu.

Analiza wykresów wskazała, że obecność anemii, nadciśnienie, wiek i czas obserwacji pacjenta są potencjalnie istotnymi czynnikami wpływającymi na ryzyko zgonu. Z kolei zmienne takie jak płeć, cukrzyca i palenie tytoniu nie wydają się istotnie wpływać na zmienną objaśnianą. Analiza tablic kontyngencji potwierdziła częściowo moje początkowe założenia, potwierdzając istotność zmiennych wieku i czasu obserwacji pacjenta.

Podczas analizy tablic kontyngencji zauważyłam też istotną statystyczną zależność między poziomem kreatyniny i sodu we krwi, frakcją wyrzutową, a ryzykiem zgonu pacjenta.

Ostatecznie wybrałam model bazując na kryterium informacyjnym Akaike (AIC). Model logit2 uwzględniający zmienne takie jak wiek, czas obserwacji pacjenta, poziom kreatyniny, poziom sodu we krwi oraz frakcja wyrzutowa jest najlepszym modelem z uwagi na uzyskanie najniższego AIC. Model logit2 osiąga najlepsze wyniki predykcji, co potwierdza jego skuteczność w identyfikacji pacjentów zwiększonego ryzyka zgonu z powodu niewydolności serca.