Uniwersytet WSB Merito w Poznaniu

Ryzyko zawału serca - czy wiek ma tak duże znaczenie? Analiza danych i predykcja ataku serca.

Autor: Hanna Szreder
Nr albumu: 90924

Promotor: dr inż. Wojciech Sałabun

Studia Podyplomowe
Kierunek: Analiza danych w Python 3

Poznań, 30/06/2024r.

Streszczenie

W niniejszej pracy przedstawiony został zbiór pt. "Heart Attack Analysis & Prediction Dataset" - opis zawartych w nim danych, graficzne ich przedstawienie, porównanie oraz analiza. Metody analizy i przedstawienia danych są uzależnione od ich typu - to również jest w tej pracy zawarte i odpowiednio dostosowane. Na samym końcu zastosowane zostały algorytmy uczenia maszynowego, w celu zobaczenia jak zbiór może radzić sobie z klasyfikacją danych w późniejszym czasie.

Słowa kluczowe— zawał serca, analiza danych, predykcja, algorytmy

Abstract

In this work, the dataset "Heart Attack Analysis & Prediction Dataset" was shown and described using text but also graphics, plots. It was analyzed and compared. The methods used for analysis and comparing data depends on its type - it is also included in this work. At the very end, the algorithms of machine learning were used to observe how this set can do in the terms of predicting future results.

Keywords— heart attack, data analysis, prediction, algorithms

Spis treści

1 Wstęp

Celem tego projektu jest analiza czynników mogących wpływać na zawał serca, porównanie ich z ryzykiem jego występowania, oraz predykcja jego wystąpienia na ich podstawie. Głównym powodem, dla którego zdecydowałam się na taki temat, jest fakt, że zarówno moi, jak i rodzice moich rówieśników, wchodzą w wiek, w którym przyjmuje się, że ryzyko wystąpienia ataku serca jest najwyższe. Ale czy tylko wiek ma znaczenie? I czy tak naprawdę ma on aż tak duże znaczenie? Czy ta poważna dolegliwość może tak naprawdę dotknąć każdego, niezależnie od wieku?

Po przejrzeniu wielu zbiorów danych, zdecydowałam się na zbiór pt. "Heart Attack Analysis & Prediction Dataset", pobrany ze strony kaggle.com. Przy wyborze, kierowałam się wartością "Usability" - ten dataset ma wartość 10.0, co oznacza, że jest w pełni wiarygodny, przygotowany z niezwykłą starannością, oraz kompatybilny z zasadami tworzenia dobrej jakości zbioru danych. Autor wziął w nim pod uwagę nie tylko wiek czy płeć badanych, ale wiele innych czynników, które razem mogą zwiększać, lub zmniejszać, ryzyko zawału. Wartości poszczególnych kolumn zostały szczegółowo opisane, do czego chciałabym przejść w następnej części tej pracy.

2 Opis zbioru danych

2.1 Wartości w zbiorze

Zbiór danych pt. "Heart Attack Analysis & Prediction Dataset", jak już wcześniej zostało wspomniane, pobrany został ze strony kaggle.com, a dokładniej https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/discussion. Składa się on z 2 plików csv, skupiam się jednak na jednym, mianowicie "heart.csv", który składa się z 14 kolumn danych opisanych przez jego autora:
1. age - wiek badanego
2. sex - płeć badanego - 1: mężczyzna, 0: kobieta
3. cp - rodzaj bólu w klatce piersiowej; podzielony na 4 podkategorie - 1: ból typowo dusznicowy, 2: ból atypowo dusznicowy, 3: niedusznicowy, 4: asymptomatyczny. Dusznica jest chorobą serca, która powoduje bóle w klatce piersiowej i trudności z oddychaniem. Znana jest również pod nazwą "dławica piersiowa"
4. trtbps - ciśnienie spoczynkowe krwi podane w mm Hg
5. chol - poziom cholesterolu podany w mg/dl
6. fbs - poziom cukru we krwi na czczo - 1: $>$ 120 mg/dl, 2: $<$ 120 mg/dl
7. restecg - EKG serca spoczynkowe - 0: wartość normalna, 1: nieprawidłowość fali ST-T, 2: wykazujący prawdopodobny lub definitywny przerost lewej komory według kryteriów Estesa
8. thalachh - maksymalne tętno osiągnięte przez pacjenta
9. exng - dusznica wywołana ćwiczeniami - 1: tak, 0: nie

10. oldpeak - depresja ST wywołana wysiłkiem w stosunku do odpoczynku - poziom na elektrokardiogramie poniżej normy
11. slp - nachylenie segmentu ST (0-2)
12. caa - liczba głównych naczyń krwionośnych zabarwionych fluoroskopią (0-4)
13. thall - wynik testu na talasemię (0-3) - nie jest to opisane w datasecie; talasemia to grupa dziedzicznych chorób krwi:
14. output - ryzyko zawału mięśnia sercowego - 1: tak, 0: nie

2.2 Rozkład danych

W użytym zbiorze, każda kolumna zawiera 303 wpisy czy też wiersze. Każdy z nich zawiera liczby całkowite (integers) bądź też zmiennoprzecinkowe (floats). Rysunek 1 przedstawia ogólny rozkład danych i wielkość bazy.

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

Rysunek 1: Wielkość bazy danych

Dla pewności wykonałam jeszcze jedną operację sprawdzenia czy baza zawiera jakiekolwiek braki danych (Rysunek 2),

```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
```

Rysunek 2: Braki danych w zbiorze

oraz operację opisania samych typów danych występujących w kolumnach (Rysunek 3):

```
age           int64
sex           int64
cp            int64
trtbps        int64
chol          int64
fbs           int64
restecg       int64
thalachh      int64
exng          int64
oldpeak     float64
slp           int64
caa           int64
thall         int64
output        int64
```

Rysunek 3: Typy danych w zbiorze

Powyższe dane wskazują na to, że zbiór jest dobrze przygotowany - brak braków danych oraz ogólna spójność w typie danych zawartych w nim, zdecydowanie ułatwiają pracę na tym datasecie. Oczywiście przy analizie, pod uwagę muszą zostać wzięte również inne jego cechy.

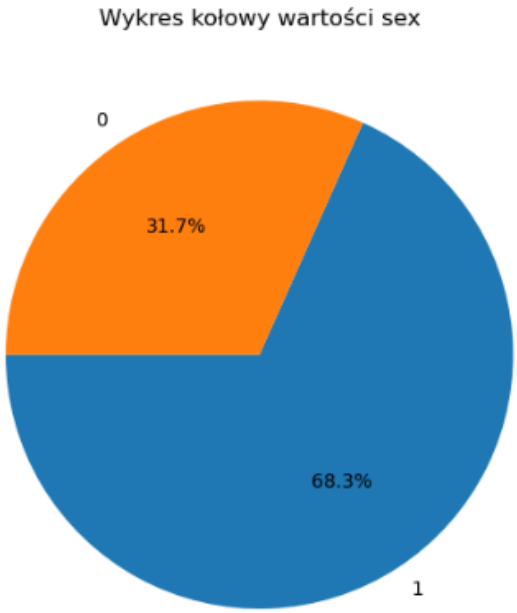
2.2.1 Rozkład poszczególnych cech - wartości kolumn z danymi kategorycznymi w datasecie

Jak widać wyżej, na tabelce z rozkładem danych oraz przy ich opisie, w tym datasecie wystepują różne dane, różne jednostki oraz wartości, niektóre mają kilka stałych opcji do wyboru, niektóre jednak są różne w każdym przypadku. Pierwsze to dane kategoryczne, drugie to dane numeryczne. Oznacza to, że, żeby przedstawić ich rozkład na zbiorze, trzeba użyć do tego różnych metod graficznych. Najpierw chciałabym przedstawić prosty rozkład z Jupyter Notebook oraz wykresy kołowe dla tych kategorycznych:

1. Płeć:

```
sex
1    207
0     96
```

Rysunek 4: Rozkład płci w zbiorze

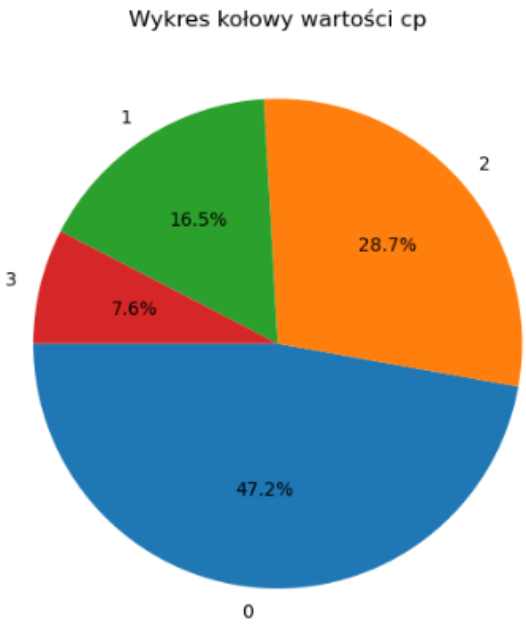


Rysunek 5: Wykres kołowy dla cechy "sex"

2. Ból w klatce piersiowej:

cp	
0	143
2	87
1	50
3	23

Rysunek 6: Rozkład typów bólu w klatce piersiowej



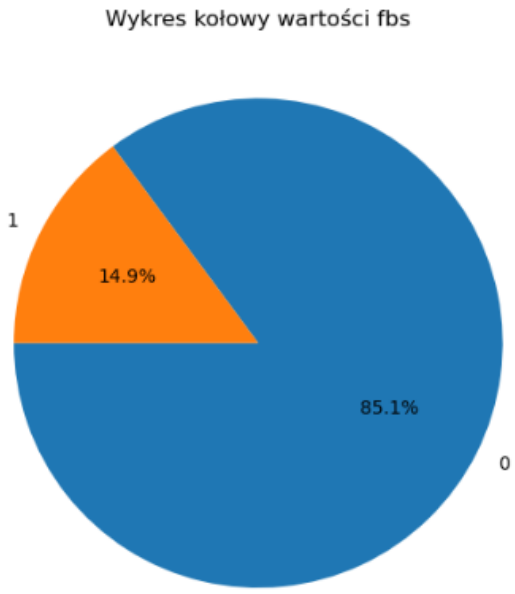
Rysunek 7: Wykres kołowy dla wartości "cp'

Jak widać na Rysunku 6, wartości nie zgadzają się z opisem podanym przez autora, dlatego przyjmuję (co jest oczywiście jednym z możliwych źródeł błędu), że wartość 0 na rysunku odpowiada wartości 1 z opisu, wartość 1 na rysunku to 2 w opisie, wartość 2 na rysunku to wartość 3 w opisie, a wartość 3 na rysunku to wartość 4 w opisie.

3. Poziom cukru we krwi na czczo:

```
fbs
0    258
1     45
```

Rysunek 8: Rozkład poziomu cukru we krwi na czczo



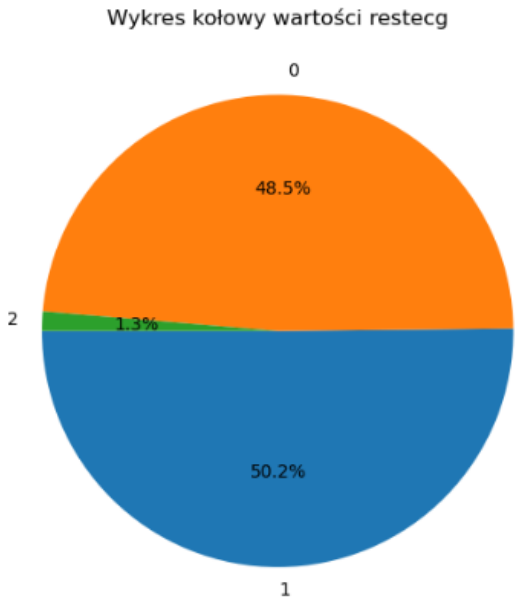
Rysunek 9: Wykres kołow dla cechy "fbs"

Tu też widać nieścisłości pomiędzy wartościami - przyjmuje, że 0 na rysunku to 1 w opisie, a 1 na rysunku to 2 w opisie.

4. EKG spoczynkowe:

```
restecg
1    152
0    147
2      4
```

Rysunek 10: Rozkład wyników EKG spoczynkowego

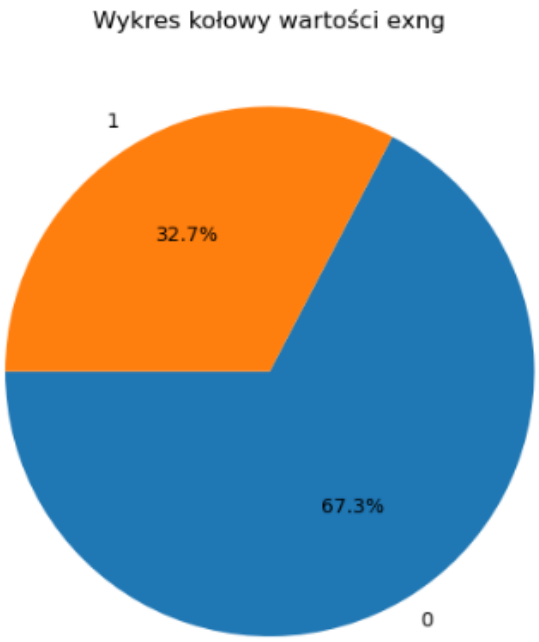


Rysunek 11: Wykres kołowy dla cechy "restecg"

5. Dusznica wywołana ćwiczeniami:

```
exng
0     204
1      99
```

Rysunek 12: Rozkład wyników dla dusznicy wywołanej ćwiczeniami

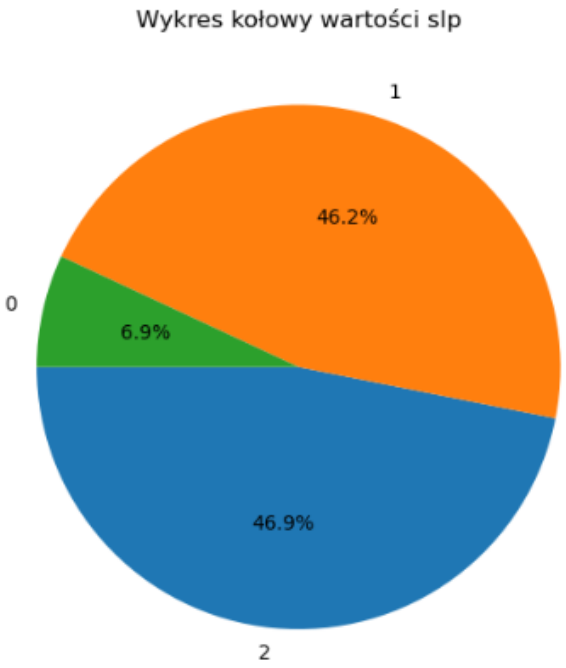


Rysunek 13: Wykres kołowy dla cechy "exng"

6. Nachylenie segmentu ST:

slp	
2	142
1	140
0	21

Rysunek 14: Rozkład nachylenia segmentu ST

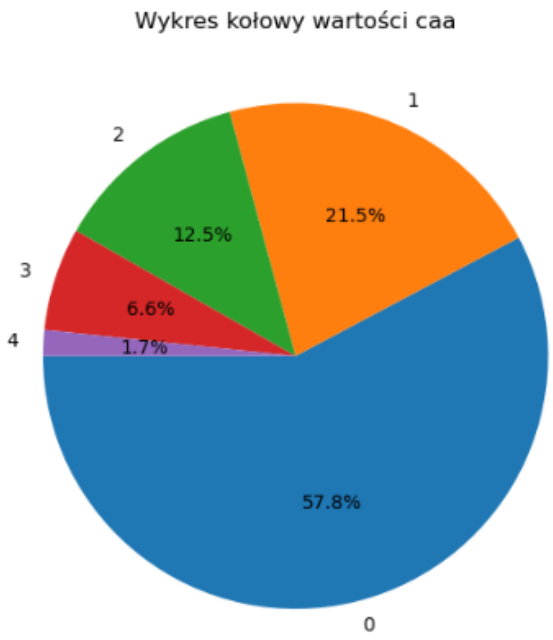


Rysunek 15: Wykres kołowy dla cechy "slp"

7. Liczba głównych naczyń krwionośnych zabarwionych fluoroskopią (0-4):

caa	
0	175
1	65
2	38
3	20
4	5

Rysunek 16: Rozkład liczby głównych naczyń krwionośnych zabarwionych fluoroskopią

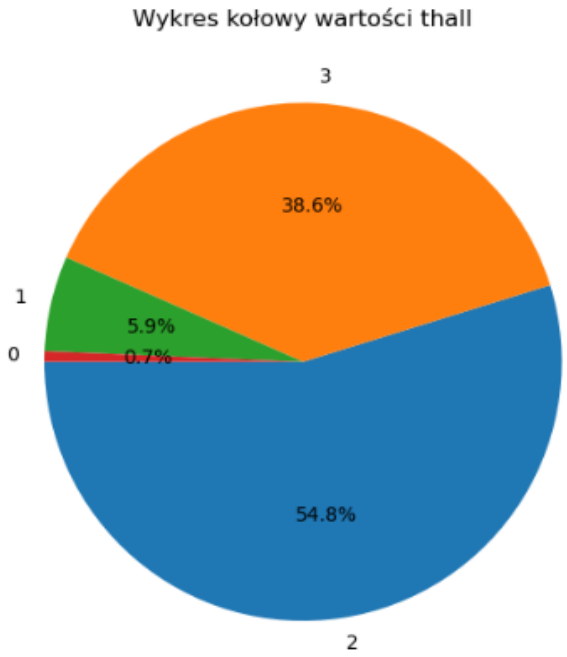


Rysunek 17: Wykres kołowy dla cechy "caa"

8. Wynik testu talasu:

```
thall
2    166
3    117
1     18
0      2
```

Rysunek 18: Rozkład wyników testu na talasemię

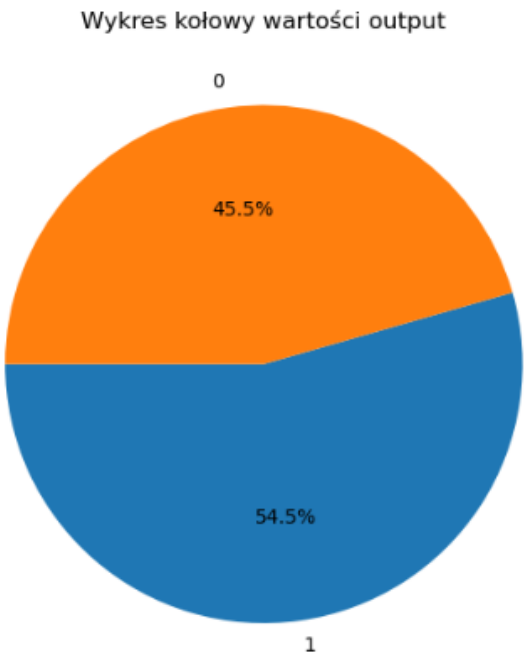


Rysunek 19: Wykres kołowy dla cechy "thall"

9. Output:

```
output
1    165
0    138
```

Rysunek 20: Rozkład wyników dla ryzyka zawału serca



Rysunek 21: Wykres kołowy dla cechy "output"

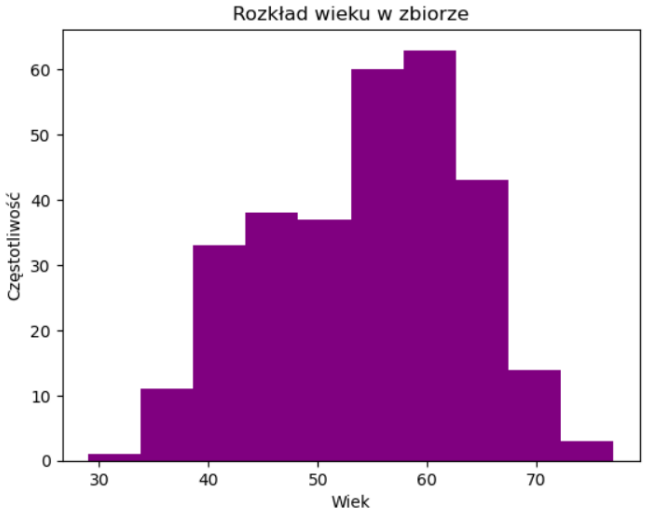
2.2.2 Opis rozkładu cech i ich interpretacja

Przejście po kolei przez rozkład każdej cechy, zdecydowanie pozwala na ujrzenie jaśniejszego obrazu zbioru oraz wzięcie pod uwagę potencjalnych problemów z nich wynikających.
Zaczynając od płci, zdecydowanie więcej mężczyzn (207, 68.3%) niż kobiet (96, 31.7%) zostało przebadanych - prawdopodobnie dlatego, że częściej słyszy się o zawałach u panów, jednak próbki powinny być podobnej wielkości by być rzetelne i miarodajne. W późniejszej analizie może okazać się, że przez to interpretacja wyników będzie trudniejsza
Jeśli chodzi o ból w klatce piersiowej to widać, że najczęściej wśród badanych występują bóle typowo dusznicowe (143, 47.2%) i niedusznicowe (87, 28.7%). Oczywiście na ten moment nie można powiedzieć, że dusznica sama w sobie w jakikolwiek sposób przyczynia się do ryzyka zawału. Jak wspomniane wyżej, wartości nie zgadzają się tu z opisem podanym przez autora, co może prowadzić do zaburzenia wyników; zdecydowałam się na swoją, opisaną wyżej, interpretację. Reszta wartości, tj. ból atypowo dusznicowy i asymptomatyczny to kolejno 50 badanych - 16.5% i 23 badanych, czyli 7.6%.
Następnie mamy przedstawiony poziom cukru we krwi - zdecydowana większość badanych ma poziom $>$ 120 mg/dl (258, 85.1%), co też na ten moment nie da-

je nam żadnych informacji o docelowych poszukiwaniach. Tu również opis nie zgadza się z właściwymi wartościami.
Jeśli chodzi o EKG spoczynkowe serca, u wiekszości badanych ukazała się nieprawidłowość fali ST-T (152, 50.2%), u dużej liczby pacjentów wartość była normalna (147, 48.5%), a tylko u kilku wykazano prawdopodobny lub definitywny przerost lewej komory według kryteriów Estesa (4, 1.3%).
Pomimo tego, że wielu pacjentów doświadczyło dusznicowego bólu w klatce piersiowej, zdecydowana większość badanych nie doświadczyła ataku dławicy piersiowej podczas aktywności fizycznej (204, 67.3%).
Pacjenci mają prawie równomiernie rozłożone nachylenia segmentu ST między typem 1 (140, 46.2%) i typem 2 (142, 46.9%). Pacjentów z typem 0 było 21 - 6.9%.
Większość badanych nie miała zabarwionych żadnych głównych naczyń krwionośnych (175, 57.8%). Jeśli chodzi o resztę - pacjentów z 1 zabarwionym naczyniem było 65, co daje 21.5%, z 2 - 38, czyli 12.5%. Z 3 zabarwionymi natomiast było natomiast 20, czyli 6.6%, a z 4 - 5, co daje 1.7%.
Najwięcej pacjentów ma typ talasemii klasyfikowany jako 2 (166, 54.8%), potem typ 3 (117, 38.6%) i typ 1 (18, 5.9%). Typ 3 jest natomiast najrzadszy (2, 0.7%)
Jeśli chodzi o wyniki końcowe, są one rozłożone mniej więcej równomiernie, mimo to jednak większość badanych jest w grupie ryzyka (165, 54.5%).

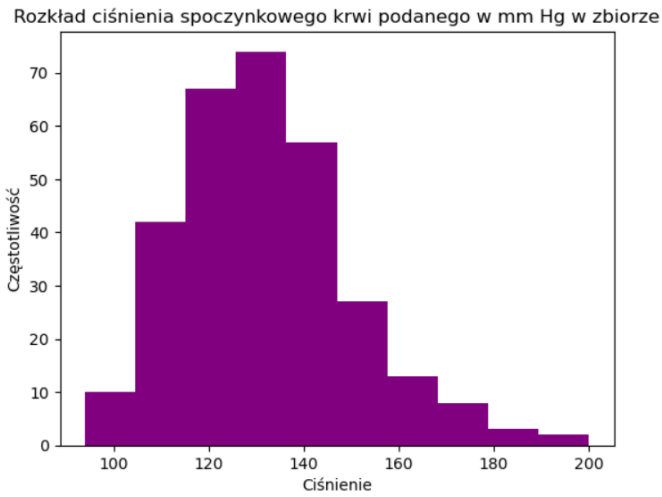
2.2.3 Rozkład poszczególnych cech - wartości kolumn numerycznych w datasecie

1. Wiek:



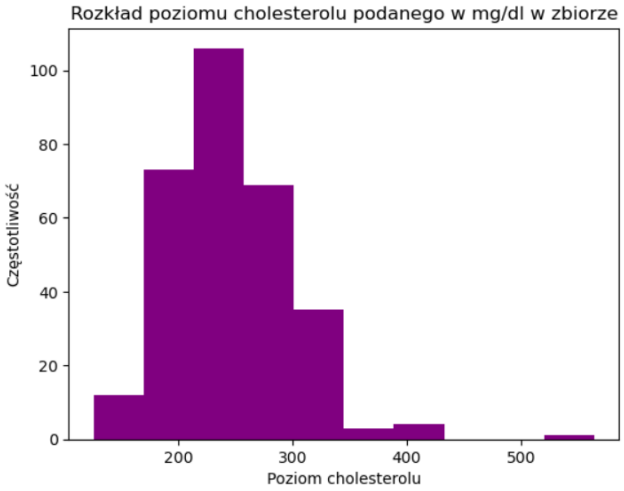
Rysunek 22: Rozkład wieku w zbiorze

2. Ciśnienie spoczynkowe:



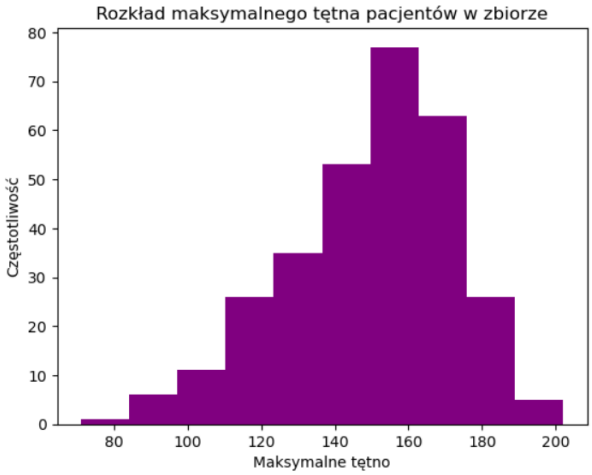
Rysunek 23: Rozkład ciśnienia spoczynkowego krwi podanego w mm Hg w zbiorze

3. Cholesterol:



Rysunek 24: Rozkład poziomu cholesterolu podanego w mg/dl w zbiorze

4. Maksymalne tętno pacjenta:



Rysunek 25: Rozkład maksymalnego tętna pacjentów w zbiorze

5. Depresja ST:

Rysunek 26: Rozkład depresji ST wywołanej wysiłkiem w stosunku do odpoczynku w zbiorze

2.2.4 Opis rozkładu cech i ich interpretacja

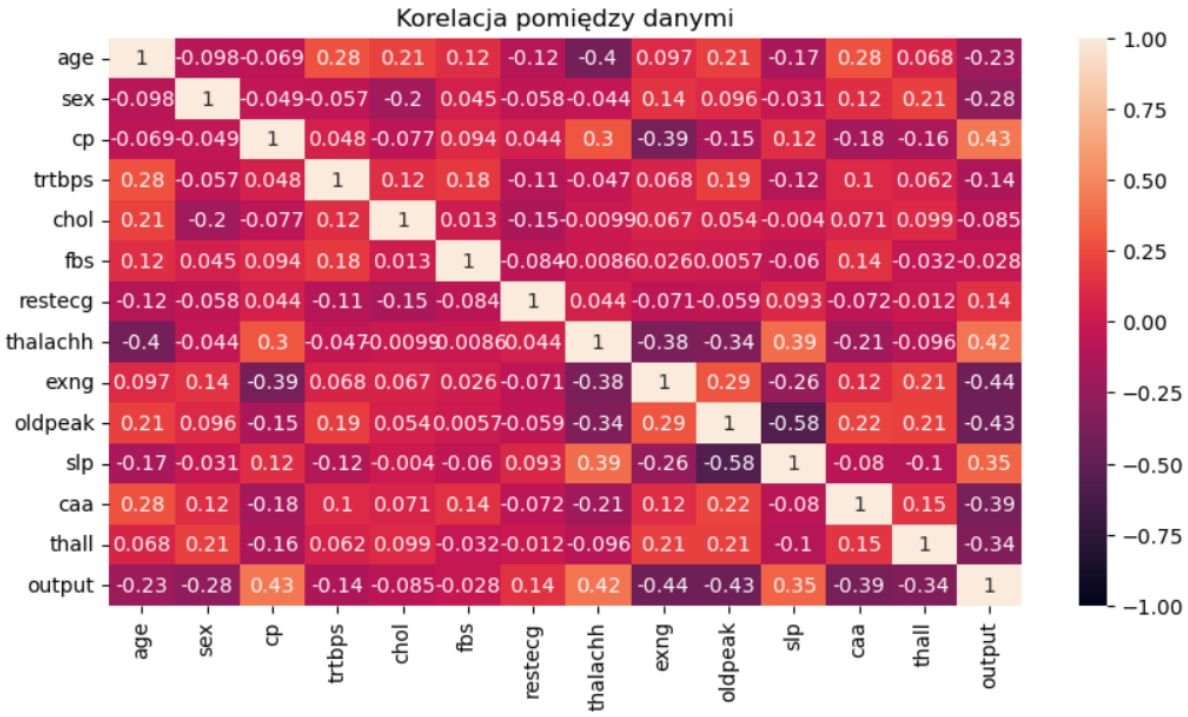
Jeśli chodzi o rozkład wieku w zbiorze, to widać wyraźnie, że zdecydowana większość badanych mieści się w przedziale 50-70 lat. Oznacza to, że, tak jak w przypadku płci, badanie tak naprawdę skupia się na ludziach w średnim wieku i starszych, co wynikać może z podwyższonego ryzyka, jednak to zostanie sprawdzone w późniejszych sekcjach.
Ciśnienie spoczynkowe badanych rozłożone jest głównie pomiędzy około 110 mmHG a 150 mmHg, z tendencją skosu w prawo.
Poziom cholesterolu natomiast rozkłada się głównie na pomiędzy około 180mg/dl a 350 mg/dl.
Na rozkładzie maksymalnego tętna pacjentów zaobserwować można wyraźną strukturę rosnącą, czy też skosu w lewo, na ten moment jednak nie mówi to za wiele o ryzyku zawału.
Jeśli chodzi o rozkład depresji ST wywołanej wysiłkiem w stosunku do odpoczynku w zbiorze, widać wyraźnie, że wykres jest skośny w prawo - zdecydowana większość pacjentów osiągała wartość od 0 do 1.

3 Analiza zbioru

W powyższych sekcjach opisałam zbiór "Heart Attack Analysis & Prediction Dataset" - jakie dane w nim wystepują, jak wygląda grupa, którą przebadano, a to ujawniło potencjalne związki, które chciałabym w tej sekcji poddać analizie.

3.1 Mapa cieplna korelacji pomiędzy wszystkimi cechami

Pierwszą rzeczą, którą warto pokazać, jest wykres, a raczej mapa cieplna, korelacji pomiędzy wszystkimi wartościami, każdej z każdą:



Rysunek 27: Mapa cieplna korelacji pomiędzy danymi

Ma ona na celu wstępne przedstawienie zależności danych pomiędzy sobą.

3.1.1 Obserwacje i interpretacja

1.Age - Wiek
Korelacja pomiędzy wiekiem a innymi cechami jest generalnie dość niska - pozytywne, zbliżające się do 0.3, można zauważyć z liczbą głównych naczyń krwionośnych zabarwionych fluoroskopią oraz z ciśnieniem spoczynkowym. Oznacza to, że z wiekiem, badani mieli więcej takich komórek oraz wyższe ciśnienie w

spoczynku.
Jesli chodzi o korelację negatywną, zauważyć można ją z maksymalnym tetnęm pacjenta (-0.4) - im starsza osoba, tym tętno maksymalne jest niższe.

2.Sex - Płeć
Korelacje są niskie, co może oznaczać, że płeć sama w sobie nie ma znaczącego związku z innymi danymi. Można wyodrębnić jedynie słabą negatywną korelację z ryzykiem zawału (-0.28), jednak ciężko na tej podstawie interpretować wyniki.

3.Cp - ból w klatce piersiowej
Jeśli chodzi o ból w klatce piersiowej, pozytywną korelację widać pomiędzy maksymalnym tętnem pacjenta (0.3) i wynikiem końcowym, czyli ryzykiem zawału (0.43). Oznacza to, że wraz z bólem w klatce, wzrasta tętno oraz ryzyko choroby serca.
Negatywną, znaczącą korelację, zaobserwować można pomiędzy bólami w klatce a dusznicą wywołaną ćwiczeniami (-0.39). Ta korelacja oznacza, że wraz z bólami w klatce piersiowej, zmniejsza się ryzyko dusznicy po wysiłku, co jest dość ciekawą obserwacją.

4. Trtbps - ciśnienie spoczynkowe
Tutaj wyodrębnić można korelację pozytywną z wiekiem (0.28) - im pacjent starszy, tym ciśnienie spoczynkowe jest wyższe.

5. Chol - poziom cholesterolu
Nie można wyodrębnić żadnych znaczących korelacji - poziom cholesterolu nie jest uzależniony od innych cech oraz inne od niego (czy raczej - nie można zauważyć żadnego potencjalnego związku pomiędzy nimi).

6. Fbs - poziom cukru we krwi
Nie można wyodrębnić żadnych znaczących korelacji - poziom cukru we krwi nie jest uzależniony od innych cech oraz inne od niego (czy raczej - nie można zauważyć żadnego związku pomiędzy nimi, jak w przypadku cholesterolu).

7. Restecg - EKG spoczynkowe
Nie można wyodrębnić żadnych znaczących korelacji - EKG spoczynkowe nie jest uzależnione od innych cech oraz inne od niego (czy raczej - nie można zauważyć żadnego związku pomiędzy nimi, tak jak w poprzednich dwóch przypadkach).

8. Thalachh - maksymalne tętno pacjenta
Pozytywną korelację można wyodrębnić tutaj z bólem w klatce piersiowej (0.3), nachyleniem segmentu ST (0.39) oraz z wynikiem końcowym (0.42). Oznacza to, że wraz z maksymalnym osiągalnym tętnem pacjenta, wzrasta prawdopodobieństwo bólu w klatce piersiowej, wiekszego nachylenia segmentu ST oraz ryzyko zawału serca.
Negatywną znaczącą korelację, można wyodrębnić z wiekiem (-0.4), dławicą piersiową po wysiłku (-0.38) oraz z depresją ST wywołana wysiłkiem w stosunku

do odpoczynku (-0.34). Maksymalne tętno maleje wraz z wiekiem, prawdopodobieństwem wystąpienia dusznicy po ćwiczeniach oraz depresji ST wywołanej wysiłkiem.

9. Exng - dusznica wywołana ćwiczeniami
Tutaj, co prawda słabą, ale pozytywną korelację widać z depresją ST wywołaną wysiłkiem w stosunku do odpoczynku (0.29) - dusznica wywołana ćwiczeniami może przyczyniać się do depresji ST.
Negatywna korelacja natomiast występuje z bólem w klatce piersiowej (-0.39), z wynikiem końcowym (-0.44), oraz z maksymalnym tętnem pacjenta (-0.38). Oznacza to, że dławica wywołana ćwiczeniami może wiązać się i/lub wpływać na mniejsze bóle w klatce, niższe ryzyko choroby serca oraz niższe maksymalne tętno pacjenta.

10. Oldpeak - depresja ST wywołana wysiłkiem w stosunku do odpoczynku
Tutaj słabą pozytywną korelację można zaobserwować z dusznicą wywołaną ćwiczeniami (0.29) - depresja ST może wpływać na dławicę, ale i odwrotnie, jak opisane wyżej.
Negatywne korelacje widać z maksymalnym tętnem pacjenta (-0.34), z nachyleniem segmentu ST (-0.58) oraz z wynikiem końcowym (-0.43). To oznaczać może, że im częstsza czy wyższa depresja ST, tym niższe maksymalne tętno, niższe nachylenie segmentu oraz mniejsze ryzyko zawału.

11. Slp - nachylenie segmentu ST
Pozytywną korelację widać tutaj z maksymalnym tętnem pacjenta (0.39), co oznacza, że istnieje możliwość, że im większe nachylenie segmentu ST, tym wyższe maksymalne tętno.
Negatywna korelacja występuje z depresją ST wywołaną wysiłkiem (-0.58). To oznaczać może, że im większe nachylenie segmentu, tym mniejsze prawdopodobieństwo depresji ST.

12. Caa - liczba głównych naczyń krwionośnych zabarwionych fluoroskopią
Tutaj pozytywną, słabą, korelację, widzimy jedynie z wiekiem (0.28) - czyli możliwy związek pomiędzy tymi dwoma cechami.
Negatywna korelacja natomiast występuje z wynikiem końcowym (-0.39) - im większa liczba głównych naczyń krwionośnych, tym mniejsze ryzyko choroby.

13. Thall - wynik testu na talasemię
Tutaj wyodrębnić można jedynie korelację negatywną z wynikiem końcowym (-0.34), która wskazuje na to, że im wyższy typ talasemii czy też wynik testu, to ryzyko zawału jest mniejsze.

14. Najważniejsze obserwacje, czyli korelacje z output - czy jest ryzyko choroby serca
Zaczynając od pozytywnych, które zostały już, co prawda, opisane wyżej, ale chciałabym zebrać je tutaj w jedno miejsce, mamy korelację z bólem w klatce

piersiowej (0.43), korelację z maksymalnym tętnem pacjenta (0.42) oraz z nachyleniem segmentu ST (0.35). Oznacza to, że im większe te wartości, tym ryzyko zawału jest większe oraz przy większym ryzyku choroby, te cechy również mają wyższe wartości.
Jeśli chodzi o korelacje negatywne - korelacja z dusznicą wywołaną ćwiczeniami (-0.44), korelacja z depresją ST wywołaną wysiłkiem w stosunku do odpoczynku (-0.43), z bólem w klatce piersiowej (-0.39) oraz z wynikiem testu na talasemię (-0.34). Można mówić też o słabej korelacji z płcią (-0.28). Oznacza to, że te cechy są raczej związane z niskim ryzykiem zawału.

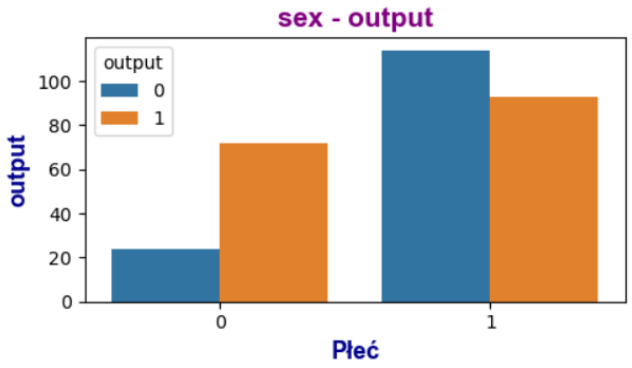
Patrząc na mapę i na korelacje należy pamiętać jednak, że nie są to związki przyczynowo-skutkowe, jedynie cechy, które mogą być ze sobą, i najprawdopodobniej są, powiązane. Co ciekawe również, patrząc na korelację wieku oraz ryzyka, mapa pokazuje, że z wiekiem ryzyko to powinno się zmniejszać.

3.2 Wykresy porównujące wartości do output

Ważnym elementem tej analizy jest przyrównanie każdej wartości z samym ryzykiem zawału, żeby móc lepiej przyjrzeć się związkami pomiędzy nimi.

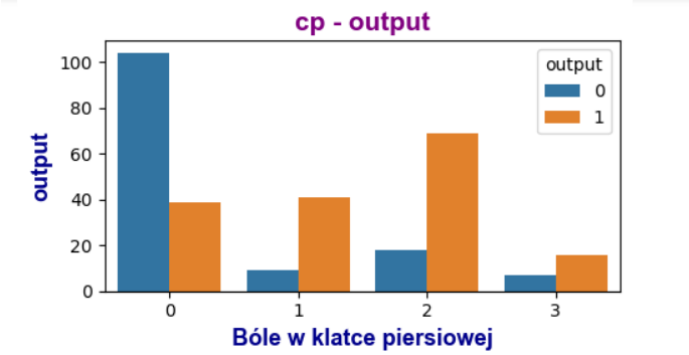
3.2.1 Wartości kategorialne

1. Płeć:



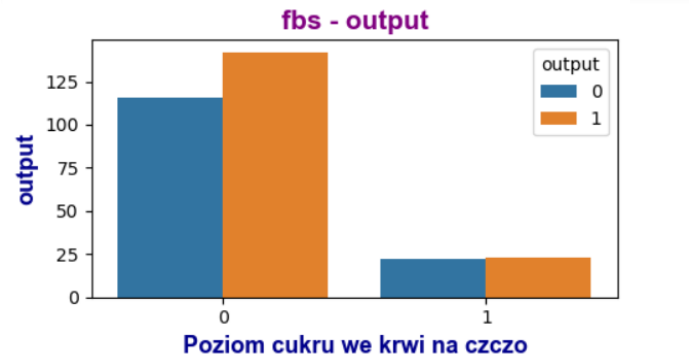
Rysunek 28: Porównanie płci i ryzyka zawału

2. Bóle w klatce piersiowej:



Rysunek 29: Porównanie bólów w klatce piersiowej i ryzyka zawału

3. Poziom cukru we krwi na czczo:

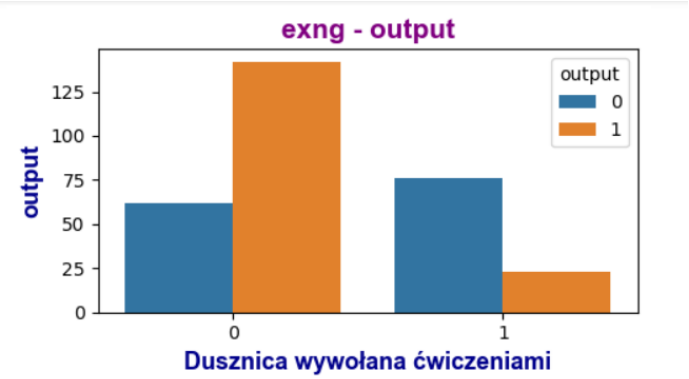


Rysunek 30: Porównanie poziomu cukru we krwi na czczo i ryzyka zawału

4. EKG spoczynkowe:

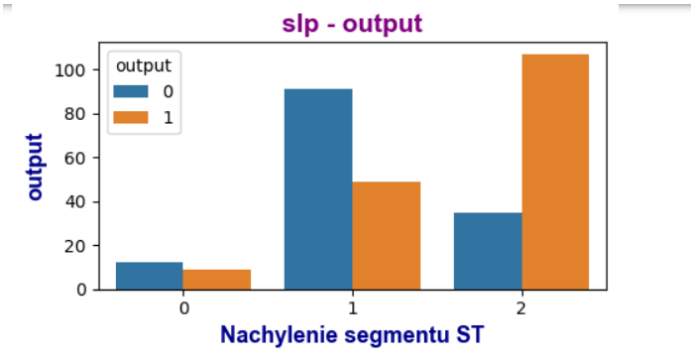
Rysunek 31: Porównanie wyników EKG spoczynkowego i ryzyka zawału

5. Dusznica wywołana ćwiczeniami:



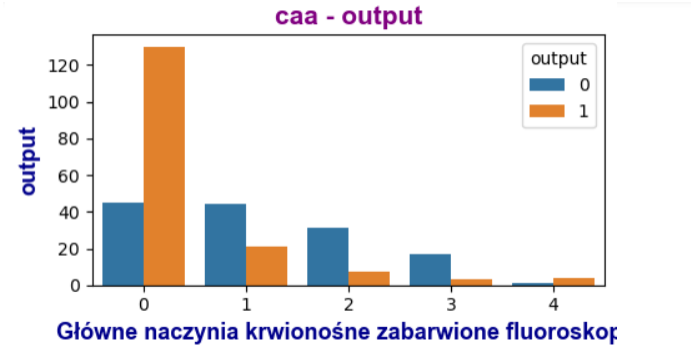
Rysunek 32: Porównanie dusznicy wywołanej ćwiczeniami i ryzyka zawału

6. Nachylenie segmentu ST:



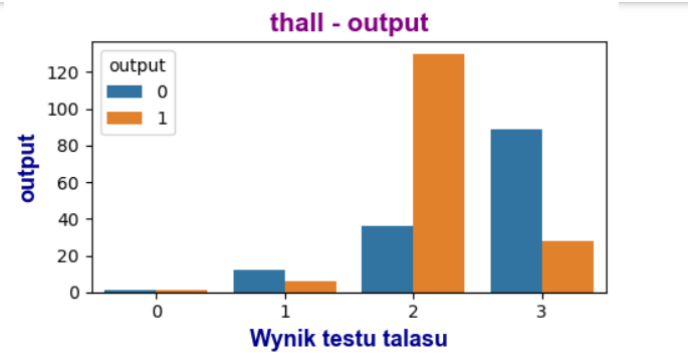
Rysunek 33: Porównanie nachylenia segmentu ST i ryzyka zawału

7. Liczba głównych naczyń krwionośnych zabarwionych fluoroskopią:



Rysunek 34: Porównanie liczby głównych naczyń krwionośnych zabarwionych fluoroskopią i ryzyka zawału

8. Wynik testu na talasemię:



Rysunek 35: Porównanie wyniku testu na talasemię i ryzyka zawału

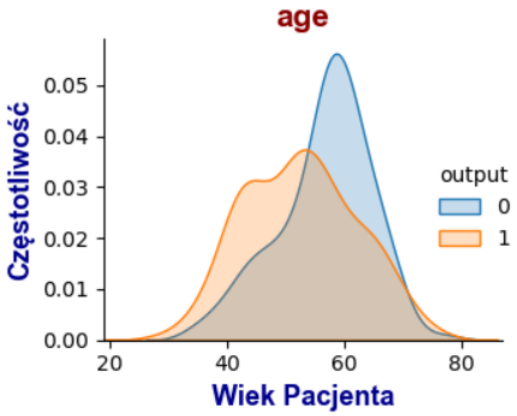
3.2.2 Opis i interpretacja związków pomiędzy wartościami kategorialnymi a ryzykiem zawału

Zaczynając od płci, zauważyć można ciekawe rzeczy - jeśli chodzi o ryzyko zawału, wśród kobiet więcej jest narażonych niż nie, jednak pomimo to, więcej przypadków tej choroby może zostać zauważonych u mężczyzn. Co również ciekawe, o wiele więcej mężczyzn niż kobiet nie jest w grupie ryzyka. Tutaj warto zatrzymać się na chwilę i wspomnieć o nierównej ilości mężczyzn i kobiet w badaniu, bo to właśnie z tego powodu powstać mogły takie a nie inne wyniki.
Jeśli chodzi o bóle w klatce piersiowej, największe ryzyko zawału występuje w grupie osób z bólami niedusznicowymi, co ciekawe, bo większość badanych zdiagnozowana była z dusznicowymi. Najmniejszą grupą ryzyka jest grupa z bólami asymptomatycznymi, ale powodem tego może być fakt, że ta grupa jest najmniejsza. Widać również, że u grupy z bólami typowo dusznicowymi, większość osób nie jest narażona na zawał.
Patrząc ma rozkład poziomu cukru we krwi na czczo, zdecydowanie można powiedzieć, że u osób z poziomem większym niż 120 mg/dl ryzyko jest większe. U osób z poziomem mniejszym niż 120 mg/dl, ryzyko i jego brak są na tym samym poziomie. Warto również przypomnieć, że pierwsza grupa była zdecydowanie większa, więc trudno tutaj o jednoznaczną interpretację.
Jeśli chodzi o EKG spoczynkowe, grupa z wynikiem 2 była bardzo mała, stąd najniższe wartości na wykresie. Ewidentnie jednak grupa 1, czyli grupa z nieprawidłowością fali ST-T, jest najbardziej narażona na zawał.
Przechodząc do dusznicy wywołanej aktywnością fizyczną - te wyniki są dla mnie wyjątkowo ciekawe, bo, pomimo tego, że po interpretacji bólów w klatce piersiowej, można było spodziewać się takich wyników, myślałam, że wysiłek i problemy z oddychaniem mogą wpłynąć bardziej na zawał mięśnia sercowego. Wykres mówi jednak inaczej - najbardziej na zawał narażeni są ci, których EKG jest w normie. Czyżby jednak wpływ na to również miała wielkość próbki?

Patrząc na nachylenie segmentu ST, zdecydowaną przewagę w ryzyku zachorowania ma grupa z typem 2, grupa z typem 1 ma większe szanse na niezachorowanie, natomiast grupa 0 mniej więcej równe.
Największe ryzyko zawału występuje o osób, u których nie było żadnych zabarwionych fluoroskopią naczyń krwionośnych. Pomimo tego, że ta grupa była największa, różnica pomiędzy wynikami może wskazywać na to, że może to mieć faktyczny wpływ na zachorowanie.
Jeśli chodzi o wyniki testu na talasemię - w grupie 0 były tylko dwie osoby, dlatego ten wynik niestety nic nam nie mówi. Spośród reszty, zdecydowanie najbardziej narażone są osoby w grupie 2. Z racji większej liczby osób w grupie 3 niż 1, można też podejrzewać, że ich wyniki mogą być dość do siebie podobne.

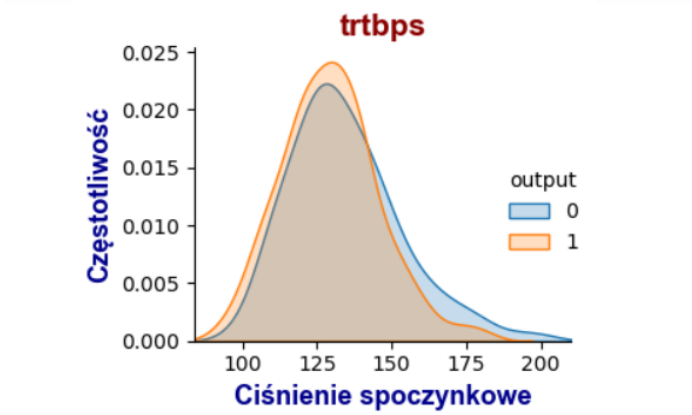
3.2.3 Wartości numeryczne

1. Wiek:



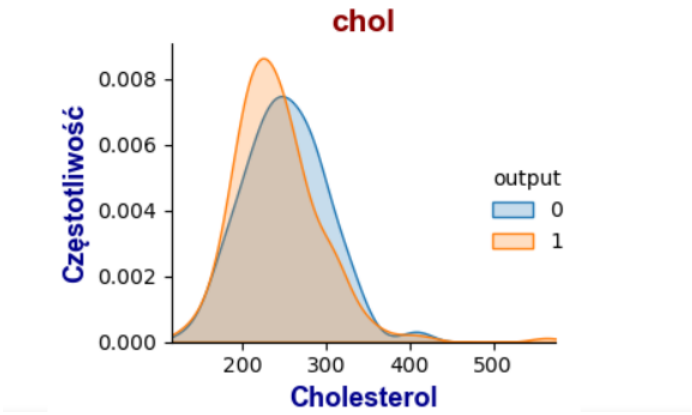
Rysunek 36: Porównanie wieku i ryzyka zawału

2. Ciśnienie spoczynkowe:



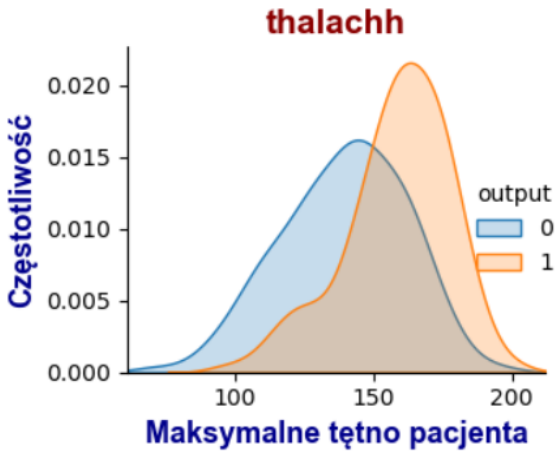
Rysunek 37: Porównanie ciśnienia spoczynkowego oraz ryzyka zawału

3. Cholesterol:



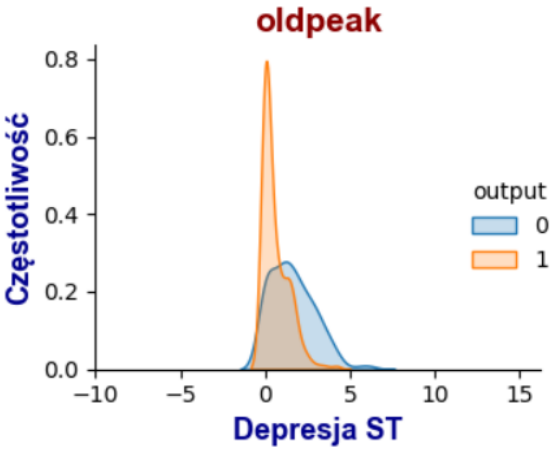
Rysunek 38: Porównanie poziomu cholesterolu oraz ryzyka zawału

4. Maksymalne tętno pacjenta:



Rysunek 39: Porównanie maksymalnego tętna pacjenta oraz ryzyka zawału

5. Depresja ST:



Rysunek 40: Porównanie depresji ST oraz ryzyka zawału

3.2.4 Opis i interpretacja związków pomiędzy wartościami numerycznymi a ryzykiem zawału

Przechodząc do wieku, czyli tak naprawdę wartości, która interesowała mnie najbardziej, muszę przyznać, że wyniki mnie zadziwiły. Oczywiście, jak widać na wykresie, generalnie ryzyko wzrasta po 40 roku życia, jednak wyraźny "peak" braku ryzyka około 60 roku życia jest zdumiewający. Oczywiście, jest to badanie na dość małej próbce i istnieje wiele innych czynników, które mogą podwyższać ryzyko, np. dieta.
Jeśli chodzi o ciśnienie spoczynkowe, tutaj można powiedzieć, że nie ma ono większego znaczenia - wykresy praktycznie nakładają się na siebie.
To samo można powiedzieć o poziomie cholesterolu, co też jest dosyć zaskakującym odkryciem. Widać również, że im wyższy poziom, tym niższe ryzyko - pamiętać należy jednak o różnych rodzajach cholesterolu.
Patrząc na wykres z maksymalnym ciśnieniem pacjenta, widać wyraźnie, że im jest ono wyższe, tym ryzyko również wzrasta.
Jeśli chodzi o depresję ST, widać wyraźnie, że niższy wskaźnik jest zdecydowanie bardziej powiązany z ryzykiem zawału serca.

4 Predykcja

Jeśli chodzi o predykcję i uczenie maszynowe, zdecydowałam się na algorytmy Logistic Regression, DecisionTreeClassifier, KNeighborsClassifier oraz SVC. Moim głównym celem było sprawdzenie jak dobrze poradzą sobie one z przewidywaniem ryzyka zawału serca na podstawie podanych danych, lecz niestety nie są one w swojej pracy idealne. Oczywiście biorę pod uwagę fakt, że z mojej strony, jako użytkownika, na pewno mogło to być lepiej dopracowane i podane algorytmom dokładniej. Przykładowo tzw. "outliers", czyli wyniki odstające od reszty - postanowiłam je zostawić, co mogło wpłynąć na rezultaty. Poniżej wyniki oraz krótki opis poszczególnych algorytmów:
1. Logistic Regression

Jest to algorytm uczenia nadzorowanego, który ma na celu dopasowanie danych do kategorii, tak jak na przykład w zbiorze użytym w tej pracy. Miał on przewidzieć czy dane zostaną skategoryzowane jako 0 lub 1, dlatego jest on jednym z lepszych w tym wypadku.

```
Test Accuracy: 0.8709677419354839
```

Rysunek 41: Dokładność zbioru testowego

Generalnie "Test Accuracy" na poziomie 0.87 wydaje mi się być wynikiem bardzo dobrym - oczywiście nie jest idealny, ale pokazuje, że ten algorytm radzi

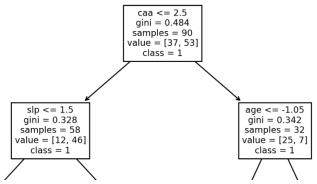
sobie przyzwoicie z przewidywaniem ryzyka zawału.

2. DecisionTreeClassifier

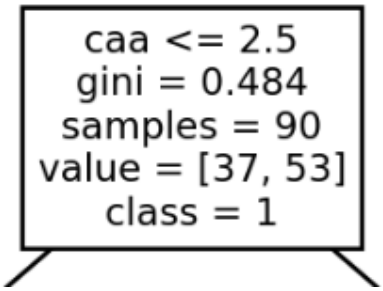
Ten algorytm uczenia nadzorowanego służy zwykle to klasyfikowania wielu klas czy cech, jednak chciałam sprawdzić jak poradzi sobie z podziałem binarnym.

```
Test: 0.812206572769953
Train: 1.0
```

Rysunek 42: Dokładność zbioru testowego i treningowego



Rysunek 43: Przykład z wygenerowanego drzewa



Rysunek 44: Przykład z wygenerowanego drzewa

Tutaj również widać, że dokładność jest w miarę wysoka, jednak nie na tyle, by móc poprawnie klasyfikować - wszystkie wartości na wygenerowanym przeze drzewie, zostały zakwalifikowane jako 1 - ryzyko zawału.

3. KNeighborsClassifier

Algorytm sprawdza najbliższych sąsiadów i w ten sposób uczy się klasyfikować dane.

```
Test: 0.8360655737704918
Train: 1.0
```

Rysunek 45: Dokładność zbioru testowego i treningowego

Tutaj, dokładnie tak jak wyżej, wyniki nie są wystarczająco dobre, ale do dopracowania.

4. SVC

Support Vector Classifier również jest algorytmem uczenia maszynowego służącym do klasyfikacji.

```
Test: 0.8360655737704918
Train: 1.0
```

Rysunek 46: Dokładność zbioru testowego i treningowego

Co ciekawe, wyniki SVC i algorytmu sąsiadów, są dokładnie takie same - to na pewno znak, że wiele może być tutaj ulepszone.

4.1 Podsumowanie

Celem tego projektu było głównie sprawdzenie czy wiek jest przyczyną zawałów oraz sprawdzenie jakie inne cechy mogą przyczyniać się do ryzyka tej przypadłości. Najpierw przedstawione zostały poszczególne dane - ich rozkład oraz korelacje pomiędzy nimi. Jak wiadomo, korelacje nie są związkami przyczynowo-skutkowymi, czyli mogą być absolutnie przypadkowe. Pomimo to, naświetlają jednak możliwości późniejszych interpretacji. Najważniejszą częścią tej pracy jest przyrównanie poszczególnych kolumn danych i ich rozkładu do rozkładu ryzyka zawału serca. Widać wiele zależności, które już można uznać za przyczynowo-skutkowe, a najważniejsze z nich to na pewno poziom cukru we krwi na czczo - jasno widać, że im wyższy poziom, tym wyższe ryzyko. Myślę, że nieprawidłowości na EKG spoczynkowym są również bardzo dobrym wskaźniekiem ryzyka zawału, jak i również dusznica spowodowana ćwiczeniami. Jest to choroba serca i układu oddechowego, ale najwyraźniej nie jest ona przyczyną tej przypadłości serca, czyli widać tu negatywny związek przyczynowo-skutkowy. Wysokie nachylenie segmentu ST również można uznać za możliwą przyczynę wyższego ryzyka, tak jak i brak zabarwionych fluoroskopią głównych naczyń

krwionośnych. Talasemia typu 2 też może być uznana za przyczynę bycia w grupie podwyższonego ryzyka. Jeśli chodzi o wiek pacjenta, tutaj widać wyraźnie, że po 40 roku życia możliwość zachorowania zdecydowanie się zwiększa, więc uważam, że również może zostać uznany za ważny czynnik. Kolejną przyczyną zawału jest najprawdopodobniej wysokie tętno - z racji tego, że serce pompuje krew, zbyt wysoki poziom może spowodować, że serce nie wytrzyma. Również poziom depresji ST pomiędzy 0 a 1 jest zdecydowanie jedną z przyczyn wyższego ryzyka zachorowania.
Jeśli chodzi o resztę cech tj. płeć, bóle w klatce piersiowej, ciśnienie spoczynkowe i poziom cholesterolu, one również mają wpływ na rozwinięcie się chorób serca, oraz atak tego mięśnia, ale nie nazwałabym ich bezpośrednimi jego przyczynami.
Krótko również o predykcji - tutaj wiele aspektów mogłoby być ulepszonych, jednak nie wyklucza to badań w tym kierunku w przyszłości.
Podsumowując, projekt ten przedstawia wiele czynników, które mogą wpłynąć na atak serca - oczywiście wiele części mogłoby być "lepszych", np. wielkość zbioru czy też zapewne moje, jako użytkownika, dokładniejsze badania, lecz zdecydowanie najważniejsze aspekty zostały tutaj ukazane.

5 Bibliografia

1. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/data

2. https://www.kaggle.com/code/oskarbeza/ataki-serca-analiza-i-predykcja#Podsumowanie

3. https://www.kaggle.com/code/akifahmednasifpurno/up-to-date-heart-attack-analysis-and-prediction

4. https://www.doz.pl/czytelnia/a15059-Dusznica_bolesna_dlawica_piersiowa__przyczyny_objawy_rozpoznanie_i_leczenie

5. https://www.dkms.pl/dawka-wiedzy/o-nowotworach-krwi/talasemia-czym-jest-jakie-sa-jej-objawy