Badanie czynników wpływających na występowanie zawału serca

Mikołaj Pełszyk

Spis treści

Wstęp

Na przestrzeni ostatnich lat choroby związane z układem krążenia są bardzo częstym powodem śmierci, pomimo że odpowiednio wcześnie wprowadzone leczenie jest bardzo skuteczne. Według danych pozyskanych przez GUS w latach 1990-2013 Każdego roku w Polsce umiera 380 tys. osób a z powodu chorób kardiologicznych umiera aż 46% z nich. Co więcej, na przestrzeni lat 1990-2013 wzrosła liczba śmierci spowodowanych tego typu chorobami o 41%.[1] Na sam zawał serca natomiast umiera co piąta osoba w naszym kraju.[2] Te statystyki nie są znacznie rozbieżne z tendencją światową, czego wyrazem jest polityka WHO , w której zakłada się, że do roku 2025 co najmniej 50% chorych na choroby układu krążenia ma być odpowiednio diagnozowanych i z zapewnionym odpowiednim leczeniem. [3] Ryzyko zawału zwiększa się wraz z wiekiem i różni się pomiędzy płciami. Dla mężczyzn ryzyko zaczyna znacznie się zwiększać powyżej 32 roku życia a dla kobiet po 45 roku życia, przy czym ogólnie mężczyźni częściej zmagają się z tego typu problemami. Celem pracy jest zbadanie jakie czynniki wpływają na występowanie zawału mięśnia sercowego i wyjaśnienie, które z nich mają największy wpływ na prawdopodobieństwo

[1] A.Golande, P. Kumar, Heart Disease Prediction Using Effective Machine Learning Techniques Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab, [dostęp : 04.05.2022].
[2] Ec.europa.eu,Dane statystyczne dotyczące przyczyn zgonu - Statistics Explained, [dostęp: 04.05.2022].
[3] Thelancet.com, The changing patterns of cardiovascular diseases and their risk factors in the states of India: the Global Burden of Disease Study 1990–2056, [dostęp: 05.05.2022].

takiego toku wydarzeń. Na podstawie literatury można było spodziewać się istotności zmiennych dotyczących anomalii w zakresie odcinka ST w wynikach elektrokardiografii oraz anomalii w przepływie krwi w mięśniu serca. W literaturze wskazuje się również, że wpływ na powstawanie chorób układu krążenia mają otyłość, nieodpowiednie odżywianie, i brak aktywności fizycznej a brak ich leczenia prowadzi do zawału mięśnia sercowego.[4] Inne wspominane czynniki ryzyka to spożywanie alkoholu i palenie papierosów. Zawał serca bywa często ignorowany przez lekarzy, mimo zgłoszeń pacjentów o bólu w klatce piersiowej. Dysponując danymi o bólach wynikających z dusznicy oraz niezwiązanych z dusznicą autor postawił hipotezę badawczą, że bóle w klatce piersiowej są jednym z kluczowych czynników pozwalających określić, czy pacjent przechodzi zawał serca. W celu przeprowadzenia estymacji wykorzystano model Logit, często wykorzystywany w badaniach w dziedzinie medycyny, również w zakresie badań dotyczących tego typu schorzeń. Poprawne zdefiniowanie czynników ryzyka, szczególnie na podstawie szybkich do przeprowadzenia badań oraz wywiadu z pacjentem może przyczynić się do przedłużenia życia osób, które przejdą zawał serca w przyszłości. Praca składa się z pięciu części: Przeglądu literatury, w którym przedstawiono podobne badania wykonane przy użyciu modelu Logit oraz metod Machine Learning, opisu bazy danych i zmiennych, przedstawienia formy funkcyjnej modelu i procesu jej dopracowywania, wyników przeprowadzonego badania oraz podsumowania z wnioskami autora.

Przegląd literatury

Badanie “Investigation Risk Factors of Cardiovascular Disease in Khartoum State, Sudan: Case - Control Study 2015” także dotyczy identyfikacji czynników sprzyjających ryzyku zawału serca i ocenie ich znaczenia. W celu dokonania oszacowania wpływu parametrów na zmienną wyjaśnianą którą były choroby sercowo-naczyniowe autorzy użyli regresji logistycznej. Dane były pozyskane poprzez badanie ankietowe które zostało przeprowadzone w grupie 162 chorych pacjentów oraz w grupie kontrolnej, która składała się z 162 osób nie będących pacjentami i wolnych od tego rodzaju chorób. Badacze wyróżnili czynniki na które można wywrzeć wpływ poprzez wprowadzenie leczenia lub zmianę nawyków, czyli wysokie ciśnienie krwi, poziom glukozy, aktywność fizyczna i zmiana diety, która ma wpływ na nadwagę i cholesterol. Zbadany został także

[4] R. Hajar, Risk Factors for Coronary Artery Disease: Historical Perspectives, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5686931/, [dostęp: 04.05.2022].

wpływ czynników na które nie ma wpływu przy obecnym stanie wiedzy ani chory ani medycyna, czyli wiek, płeć historia tego typu chorób w rodzinie. Autorzy wnioskują, że wysokie ciśnienie krwi, występowanie choroby w rodzinie oraz brak aktywności fizycznej są głównymi przyczynami występowania chorób układu krążenia w grupie badanych pochodzących ze stolicy Sudanu.[5]

Kolejnym badaniem o zbliżonej tematyce oraz przeprowadzone za pomocą modelu Logitowego jest "Logistic Regression Analysis To Determine Cardiovascular Diseases Risk Factors A Hospital-Based Case-Control Study". Przeanalizowano w nim czynniki ryzyka wystąpienia chorób układu sercowo-naczyniowego. Badany zbiór zawiera informacje z 800 wywiadów z pacjentami szpitala specjalizującego się w tego typu chorobach zbieranych przez okres roku, od stycznia 2018 do stycznia 2019 roku. Próba badanych była randomizowana, wśród badanych przeważali mężczyźni, było ich 55% a kobiet 45%, ponadto kobiety były średnio młodsze. Autorzy chcieli przebadać nie tylko czynniki czysto medyczne, ale również ekonomiczno-społeczne i demograficzne. Najważniejsze wnioski badaczy to, że ryzyko wystąpienia chorób układu krążenia wzrasta u osób palących (odds = 9.84), z chorymi nerkami (odds = 6.19) i pijących alkohol (odds = 3.62).[6]

W badaniu "Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab" W celu zbadania prawdopodobieństwa zachorowania w przeciągu najbliższych 10 lat na choroby układu krążenia dokonano analizy zbioru danych pochodzących z Framingham Heart Study, w którym zebrano odpowiedzi od 4238 osób. W celu oszacowania wpływu parametrów wykorzystano model Logit oraz metody Machine Learning, które zyskują dużą popularność wśród badaczy w dziedzinie medycyny. Autorzy oszacowali ilorazy szans, wrażliwość i specyficzność. Wybrany przez badaczy model dobrze prognozuje badane zjawisko. Przyjęto punkt odcięcia na poziomie 0,4, Przy takim ustawieniu wrażliwość wyniosła 83,8%, natomiast specyficzność była bardzo wysoka, wyniosła 99.8%. Wyniki badania podobnie jak we wcześniej przytoczonym badaniu wskazują, że liczba papierosów, wiek oraz wysokość ciśnienia krwi są istotnymi determinantami badanych chorób, ponadto sugerują, że bardziej prawdopodobnym jest, że w przeciągu 10 lat od przeprowadzenia badania chorym będzie mężczyzna niż kobieta.[7]

[5] E. A Frah, Investigation Risk Factors of Cardiovascular Disease in Khartoum State, Sudan: Case - Control Study 2055, [dostęp: 03.05.2022].

[6] E. Mustafa, A. Alnory, Logistic Regression Analysis to Determine Cardiovascular Diseases Risk Factors A Hospital-Based Case-Control Study, 2059. | International Journal of Medical Science and Clinical Invention. [dostęp: 03.05.2022].

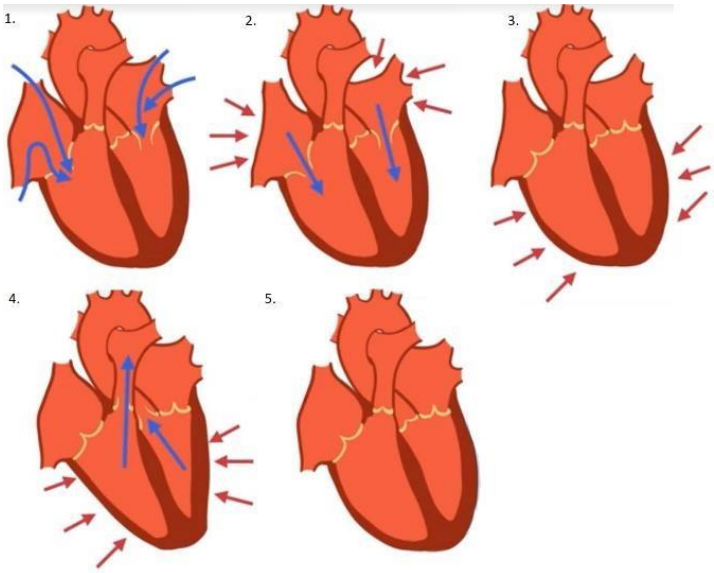
[7] A.Golande, P. Kumar, Heart Disease Prediction Using Effective Machine Learning Techniques
Predicting Heart Diseases in Logistic Regression of Machine Learning Algorithms By Python Jupyterlab [dostęp: 03.05.2022].

Wprowadzenie do problematyki badania

Prawidłowy cykl pracy serca jest następujący:

1. Rozkurcz komór, przedsionków, zastawki P-K otwarte, zastawki tętnicze zamknięte
2. Rozkurcz komór, **skurcz przedsionków**, zastawki P-K otwarte, zastawki tętnicze zamknięte
3. **Izowolumetryczny skurcz komór, rozkurcz przedsionków, zastawki P-K zamknięte**, zastawki tętnicze zamknięte
4. **Skurcz komór**, rozkurcz przedsionków, zastawki P-K zamknięte, **zastawki tętnicze otwarte**
5. **Rozkurcz izowolumetryczny komór**, rozkurcz przedsionków, zastawki PK zamknięte, **zastawki tętnicze zamknięte**

Rysunek 1. Prawidłowy cykl pracy serca



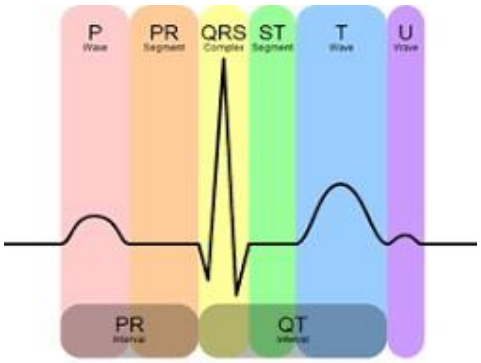
Źródło: https://www.youtube.com/watch?app=desktop&v=iF01He_Ecxs&fbclid=IwAR0tdCM-Px8FV3gmeyXx3cxSjzn4MQmjnmyrEENd0dzUHl1MGeH67TgC44 [dostęp: 04.05.2022].

Pęknięcie blaszki miażdżycowej często prowadzi do zawału serca. Spowodowane jest to tym, że w miejscu pęknięcia tworzy się skrzeplina, która doprowadza do zamknięcia naczynia wieńcowego przez co krew nie może przepływać przez mięsień. Kluczowy jest czas, im dłużej tętnica jest zamknięta, tym większa szansa na trwałą szkodę u pacjenta. Zmiany w obszarze z odciętym przepływem krwi są nieodwracalne już po okresie od 3 do 6 godzin[8].

[8] K Thygesen, Czwarta uniwersalna definicja zawału serca, https://journals.viamedica.pl/kardiologia_polska/article/download/KP.2058.0203/62413, Kardiologia Polska nr 76, str. 1405., ISSN 0022–9032.

W badaniu EKG badana jest elektryczna czynność serca. Każdy z przedstawionych na rysunku załamków odpowiada rozładowaniom elektrycznym lub ponownemu naładowaniu badanego obszaru serca. Czas trwania linii izoelektrycznej dzieli się na odcinki a łączny czas trwania odcinków i sąsiadującego załamka to odstępy.

Rysunek 2. Prawidłowy zapis EKG w spoczynku



Źródło: https://fizjotechnologia.com/przeglad-sprzetu/jak-rozumiec-i-interpretowac-krzywa-ekg.html/attachment/krzywaekg-odcinki [dostęp: 04.05.2022].

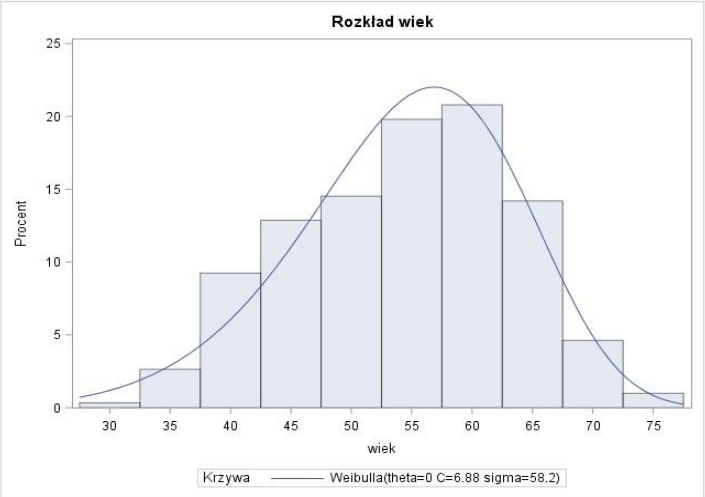
W badaniu użyta została zmienna związana z odcinkiem ST, który mierzony jest od końca zespołu QRS. Gdy serce działa prawidłowo znajduje się na poziomie linii izoelektrycznej, obniżenie jest nieprawidłowe a uniesienie występuje w ostrym zawale serca.[9]

Opis bazy danych

Baza danych, która została użyta w niniejszym badaniu zawiera dokumentację medyczną 303 pacjentów badanych ze szpitala w Cleveland u których istniało podejrzenie zawału serca. W badanej próbie było 207 mężczyzn oraz 96 kobiet.

[9] EKG W PIGUŁCE Elektrokardiografia – metoda obrazowania, https://www.ka.edu.pl/download/gfx/ksw/pl/defaultopisy/1054/5/1/mcr1_-ekg_w_pigulce.pdf [dostęp: 04.05.2022].

Wykres 1. Wiek badanych



Źródło: badanie własne.

Tabela 1. Statystyki opisowe zmiennej wiek

Zmienna analizowana: wiek				
N	Średnia	Odch. std.	Minimum	Maksimum
303	54.3663366	9.0821010	29.0000000	77.0000000

Źródło: badanie własne.

W próbie znalazły się osoby w przedziale wiekowym od 29 do 77 roku życia. Rozkład zmiennej pokrywa się z grupą ryzyka wystąpienia tej przypadłości, ponieważ ryzyko wzrasta u mężczyzn już w wieku 32 lat a u kobiet w wieku 45 lat. Najbardziej liczną grupą badanych są osoby w wieku między 5560 lat, średnia wieku wyniosła 54.36 lat z odchyleniem standardowym 9.08. Przeprowadzając wstępne oszacowania, wiek okazywał się zmienną nieistotną, czego przyczyną może być przynależność znacznej większości badanych do grupy ryzyka.

Tabela 2. Testy i kwantyle rozkładu zmiennej wiek

Testy dopasowania dla rozkładu Weibulla				
Test	Statystyka		Wartość p	
Cramer-von Mises	W-kwadr.	0.10440303	Pr. > W-kwadr.	0.092
Anderson-Darling	A-kwadr.	0.77596878	Pr. > A-kwadr.	0.044

Kwantyle (rozkład Weibulla)		
	Kwantyl	
Procent	Obserwowane	Szacunkowe
1.0	35.0000	29.8140
5.0	39.0000	37.7797
10.0	42.0000	41.9444
25.0	47.0000	48.5343
50.0	55.0000	55.1483
75.0	61.0000	60.9908
90.0	66.0000	65.6565
95.0	68.0000	68.2152
99.0	71.0000	72.6123

Źródło: badanie własne.

Kwantyle rozkładu wskazują, że ponad 95% badanych osób to osoby już o podwyższonym ryzyku zawału serca. Efekt wieku byłby istotny, gdyby w grupie badanych ujęto także osoby młode. Mediana w zbiorze wyniosła 55 lat. Uciąglony rozkład wieku jest zbliżony do rozkładu Weibulla, jednak testy podają sprzeczne wyniki. Według testu Cramera von Misesa jest to ten rozkład, jednak wynik testu Andersona-Darlinga odrzuca tę hipotezę na poziomie istotności 5%.

Dokładny opis zmiennych zawartych w zbiorze danych ujęty został w kolejnym podrozdziale. W zbiorze nie znajdowały się żadne braki danych.

Forma funkcyjna modelu oraz opis zmiennych

Na podstawie analizowanej literatury i wiedzy własnej utworzono model logit.

Do wstępnych estymacji użyto różnych kombinacji zmiennych dostępnych w zbiorze danych, były to następujące parametry:

x1 - wiek - wiek w latach**,**

x2 - mezczyzna - (1 - mężczyzna, 0 - kobieta),

x3 - bol_typ1 - ból typowy dla dusznicy bolesnej,

x4 - bol_typ2 - nietypowy ból dla dusznicy bolesnej,

x5 - bol_typ3 - bol niezwiazany z dusznicą bolesną,

poziom bazowy dla zmiennych ból to brak bólu

x6 - cisnienie - ciśnienie krwi mierzone w spoczynku w mm Hg,

x7 - cholesterol - poziom cholesterolu w surowicy (mg/dl),

x8 - cukier - czy poziom cukru we krwi na czczo przekracza 120mg/dl (1 - tak, 0 - nie), x9 - ecg1 (1 = zaburzenia - fale ST lub T odchylone o >0.05mV od normy 0 -brak zaburzeń), x10 - ecg2 Widoczna hipertrofia mięśnia sercowego w EKG (1 = hipertrofia, 0 - brak hipertrofii),

x11 - tetno - maksymalne tętno,

x12 - spadek_ST - wielkość spadku odcinka ST w EKG podczas wysiłku fizycznego w porównaniu do stanu spoczynku.

x13 - ST_nach1 - nachylenie w dol odcinka ST w EKG w spoczynku (1 - tak, 0 - nie),

x14 - ST_nach2 - nachylenie w gore odcinka ST w EKG w spoczynku (1 - tak, 0 - nie), poziom bazowy to płaskie (normalne) nachylenie odcinka ST w badaniu x15 - szer_zyly - liczba głównych żył pokolorowanych przez fluoroskopię**,**

x16 – thal1 Obserwowana Talasemia - odbiegający od normy przepływ krwi w niektórych częściach serca poziom bazowy to normalny przepływ krwi w sercu

y - target - diagnoza choroby serca (1 - zwezenie średnicy o > 50% 0 - zwezenie o <50%) **(output)**

Najbardziej nietypowy wynik to wyższa szansa przejścia zawału przez mężczyzn a nie kobiety, według jednego z estymowanych modeli (ze wszystkimi zmiennymi istotnymi statystycznie) bycie mężczyzną powoduje zmniejszenie szansy na zawał do 29.2% względem kobiet. Jest to wniosek specyficzny dla tego zbioru danych, gdyż spośród 96 badanych kobiet aż 72 (75%) przeszły zawał, wśród badanych mężczyzn 93 przeszło zawał spośród 207 (44,9%). Pomimo istotności statystycznej ze względu na niezgodność z literaturą przedmiotu autor postanowił odrzucić tę zmienną w celu uzyskania modelu bardziej ogólnego.

Po wyrzuceniu zmiennych nieistotnych z modelu ostateczna forma funkcyjna zawiera zmienne:

x1 - tetno - maksymalne tętno, x2 -

bol_typ3 - bol niezwiazany z dusznica,

x3 - bol_typ1 - ból typowy dla dusznicy,

x4 - spadek_ST - wielkość spadku odcinka ST w EKG podczas wysiłku fizycznego w porównaniu do stanu spoczynku,

x5 - ST_nach2 - nachylenie w gore odcinka ST w EKG w spoczynku (1 - tak, 0 - nie),

X6 - szer_zyly - liczba głównych żył pokolorowanych przez fluoroskopię,

x7 – thal1 - Obserwowana talasemia - odbiegający od normy przepływ krwi w niektórych częściach serca**,**

y - target - diagnoza choroby serca (1 - zwezenie średnicy o > 50% 0 - zwezenie o <50%)

Wyniki estymacji modelu zostały przedstawione w tabeli poniżej:

Tabela 3. Wyniki estymacji modelu Logit

Analiza ocen maksymalnej wiarygodności					
Parametr	DF	Ocena	Błąd standardowy	Chi-kwadrat Walda	Pr. > chi-kw.
Intercept	1	-3.2606	1.2859	6.4293	0.0112
tetno	1	0.0179	0.00844	4.4928	0.0340
bol_typ3	1	1.6661	0.4080	16.6724	<.0001
bol_typ1	1	1.4689	0.5850	6.3042	0.0120
spadek_ST	1	-0.5675	0.1984	8.1814	0.0042
ST_nach2	1	0.7632	0.3992	3.6559	0.0559
szer_zyly	1	-0.8564	0.1882	20.6986	<.0001
thal1	1	1.7989	0.3355	28.7512	<.0001

Źródło: badanie własne.

Wszystkie dobrane zmienne są istotne i zostaną poddane dalszej analizie.

Tabela 4. Testy hipotezy o Beta = 0

Testowanie globalnej hipotezy zerowej: BETA=0			
Test	Chi-kwadrat	DF	Pr. > chi-kw.
Iloraz wiarygod.	185.5463	7	<.0001
Wynik punktowy	147.6338	7	<.0001
Wald	82.5031	7	<.0001

Źródło: badanie własne.

Globalna hipoteza zerowa o becie równej 0 jest odrzucana na każdym poziomie istotności, czyli estymowany model jest łącznie istotny.

Tabela 5. Wyniki adjusted R2 dla modelu

R-kwadrat	0.4579	Maksymalnie przeskalowane R-kwadrat	0.6122

Źródło: badanie własne.

Adjusted R2 dla modelu wynosi 45.79%. Nie da się łatwo interpretować tego kryterium, jednak wiadomo, że model wyjaśnia zjawisko w ~40%.
Poniżej przedstawione zostały wyniki oszacowań ilorazu szans dla poszczególnych zmiennych w oszacowanym modelu:

Tabela 6. Wyniki oszacowań ilorazów szans dla modelu

Oceny ilorazu szans			
Efekt	Wynik punktowy	Przedział ufności Walda 95%	
tetno	1.018	1.001	1.035
bol_typ3	5.292	2.378	11.774
bol_typ1	4.345	1.380	13.675
spadek_ST	0.567	0.384	0.836
ST_nach2	2.145	0.981	4.691
szer_zyly	0.425	0.294	0.614
thal1	6.043	3.131	11.663

Źródło: badanie własne.

Aby ocenić prawdziwy wpływ zmiennych na wystąpienie zawału serca należy przeanalizować ilorazy szans dla każdej ze zmiennych. Wystąpienie bólu w klatce piersiowej niezwiązany z dusznicą sprawia, że szansa wystąpienia zawału stanowi 529.2% poziomu bazowego. Wystąpienie bólu typowego dla dusznicy zwiększa szansę wystąpienia zawału do 434.5%, względem braku tego typu objawów. Wzrost tętna o 1 punkt powoduje wzrost szans na zawał o 0.18% za każdy punkt. Spadek odcinka ST w EKG podczas wysiłku w porównaniu do stanu spoczynku zmniejsza szansę na zawał do 56.7% względem braku spadku. Nachylenie w górę odcinka ST w EKG w stanie spoczynku powoduje wzrost szansy na zawał względem poziomu bazowego o 214.5%. Wraz z liczbą pokolorowanych żył we fluoroskopii maleje szansa na zawał,

z każdą żyłą szansa bazowa mnożona jest przez 42.5%. Pacjenci z Talasemią i obserwowalnymi anomaliami w przepływie krwi w sercu mają 604.3% szans na zawał względem osób zdrowych.

Tabela 7. Tabela klasyfikacyjna (cut off point = 0.5)

Liczebność Procent Proc. wier. Proc. kol.	Tabela target od y		
	y		
target	0	1	Razem
0	108 35.64 78.26 86.40	30 9.90 21.74 16.85	138 45.54
1	17 5.61 10.30 13.60	148 48.84 89.70 83.15	165 54.46
Razem	125 41.25	178 58.75	303 100.00

Źródło: badanie własne.

Wrażliwość w modelu ma wartość 83.15% czyli w takiej części danych poprawnie określane dodatnie wyniki testu. Specyficzność wynosi 86.40% czyli z taką dokładnością określane są prawdziwie ujemne wyniki testu.

Tabela 8. Uproszczona tabela klasyfikacyjna

dobra_klasyfikacja	Liczebność	Procent	Liczebność skumulowana	Procent skumulowany
0	47	15.51	47	15.51
1	256	84.49	303	100.00

Źródło: badanie własne.

Łączna miara Accuracy (Count R2) wyniosła 84.49% dla domyślnego punktu odcięcia 0.5. Zmiany tego parametru nie powodowały wzrostu jakości oszacowań w odniesieniu do dwóch rodzajów błędów statystycznych.

Wnioski

W badaniu zastosowano model Logit do predykcji prawdopodobieństwa zawału serca u pacjentów, u których taki stan był podejrzewany przez lekarzy ze szpitala w Cleveland. Wyszczególniono 7 czynników. Na podstawie informacji z tabeli klasyfikacyjnej, Count R2 dla modelu wyniósł 84.49%, i z taką dokładnością pozwala określić kto przejdzie zawał serca a kto nie. Lepiej oszacowana została specyficzność, która wyniosła 86.4%, niewiele gorzej wrażliwość, 83.15%. Zbiór danych nie pozwolił na prawdziwe oszacowanie wpływu dwóch często szacowanych zmiennych w tego typu badaniach: wieku badanych, gdyż 95% badanych było w grupie ryzyka przejścia zawału oraz płci, gdyż ze względu na specyfikę doboru próby wyniki mimo istotności statystycznej były sprzeczne z wiedzą naukową. Przez odrzucenie zmiennej związanej z płcią model stał się bardziej ogólny, tracąc przy tym 1.3% miary adjusted R2. Z badania wynika, że bóle w klatce piersiowej, niezwiązane z dusznicą mają większy wpływ na szansę przejścia zawału serca niż bóle związane z dusznicą, jednak względem poziomu bazowego jest to kolejno ponad 5-krotnie i 4-krotnie większa szansa. Wniosek ten potwierdza hipotezę badawczą o bólu w klatce piersiowej jako jednym z kluczowych czynników pozwalających stwierdzić zawał mięśnia sercowego. Talasemia z widocznymi objawami niedoboru krwi w mięśniu sercowym także jest istotnym czynnikiem, zwiększa szansę na zawał ponad 6-krotnie względem osób zdrowych. Nachylenie odcinka ST w górę w badaniu EKG w stanie w spoczynku zwiększa szansę na zawał dwukrotnie. Istotne okazały się także wyniki Fluoroskopii, każda widoczna w badaniu żyła zmniejsza szansę na zawał do 42.5%. Spadek nachylenia odcinka ST w badaniu wysiłkowym względem badania w stanie spoczynku zmniejsza szansę o ponad połowę. Istotnym ograniczeniem w stwierdzeniu wpływu płci i wieku w badaniu była wielkość bazy danych i dobór próby, dlatego dalsze badania powinny opierać się na większych zbiorach danych, gdzie nie występuje problem selekcji próby. Badanie można rozwinąć także poprzez zwiększenie zestawu zmiennych niezależnych, które mogą mieć wpływ na choroby serca, np. Cechy fizyczne pacjentów jak wzrost, wskaźnik BMI oraz historia leczenia.

Literatura

A.Golande, P. Kumar, Heart Disease Prediction Using Effective Machine Learning Techniques Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab [dostęp: 03.01.2022]

Bredy C, Ministeri M, Kempny A, Alonso-Gonzalez R, Swan L, Uebing A, Diller G-P, Gatzoulis MA, Dimopoulos K. *New York Heart Association (NYHA) classification in adults with congenital heart disease: relation to objective measures of exercise and outcome*, Eur Heart J – Qual Care Clin Outcomes. 2017; 4(1):51–8 [dostęp: 03.01.2022]

E. A Frah, Investigation Risk Factors of Cardiovascular Disease in Khartoum State, Sudan: Case - Control Study 2015. [dostęp: 04.01.2022]

E. Mustafa, A. Alnory, Logistic Regression Analysis To Determine Cardiovascular Diseases Risk Factors A Hospital-Based Case-Control Study, 2019. | International Journal of Medical Science and Clinical Invention. [dostęp: 03.01.2022].

K Thygesen, Czwarta uniwersalna definicja zawału serca, https://journals.viamedica.pl/kardiologia_polska/article/download/KP.2018.0203/62413, Kardiologia Polska nr 76, str. 1405., ISSN 0022–9032.

R. Hajar, Risk Factors for Coronary Artery Disease: Historical Perspectives, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5686931/ [dostęp: 04.01.2022]

T. Saxena, Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning [dostęp: 05.06.2021]

Thelancet.com, The changing patterns of cardiovascular diseases and their risk factors in the states of India: the Global Burden of Disease Study 1990–2016, [dostęp: 05.01.2022].

Ec.europa.eu,Dane statystyczne dotyczące przyczyn zgonu - Statistics Explained, [dostęp: 05.01.2022].
Thelancet.com, The changing patterns of cardiovascular diseases and their risk factors in the states of India: the Global Burden of Disease Study 1990–2016.[dostęp: 04.01.2022]

Dane

Źródło bazy danych ze szpitala Cleveland: https://sci2s.ugr.es/keel/dataset.php?cod=57 [dostęp: 02.01.2022].

Spis Rysunków

Spis Wykresów

Spis Tabel

Aneks z kodem

```
x 'cd C:\Users\123\Desktop\EAD'; libname dane
'C:\Users\123\Desktop\EAD\dane';
proc import datafile='zawaly.csv' out=pacjenci dbms=csv replace;
run;
proc logistic data= pacjenci; model target(event='1') = mezczyzna
tetno bol_typ3 bol_typ1 spadek_ST
ST_nach2 szer_zyly thal1 /rsquare;
output out = pred_logit p=p; run;
data pred_logit;
set pred_logit;
      if p<=0.5
  then y=0;
if p>0.5
then y=1; run;

proc freq data=pred_logit;  tables
target*y / out=cross_results; run;
data cross_results;   set
cross_results;
  if target = 1 and y = 1 then result_type = 'TP';
if target = 1 and y = 0 then result_type = 'FN';
if target = 0 and y = 1 then result_type = 'FP';
```

```
if target = 0 and y = 0 then result_type = 'TN';
run;  data pred_logit; set pred_logit;
if target=y then dobra_klasyfikacja=1;
                    else dobra_klasyfikacja=0;
run;  proc freq data=pred_logit; tables
dobra_klasyfikacja; run;
proc univariate data=Pacjenci;
var wiek;
   histogram wiek/weibull; run;
```