Analiza czynników wpływających na wystąpienia cukrzycy

Jędrzej Sarna

28 stycznia 2024

1 Wprowadzenie

Tematem projektu jest badanie czynników wpływających na występowanie cukrzycy oraz stworzenie modelu predykcyjnego, pozwalającego zdiagnozować cukrzycę u pacjenta na podstawie odpowiednich badań. Populacja wybrana do tej analizy to populacja Indian Pima, mieszkających w pobliżu Phoenix, w Arizonie. Grupa ta jest pod ciągłą obserwacją od 1965 roku przez Narodowy Instytut Cukrzycy i Chorób Trawienia oraz Nerek z powodu wysokiego wskaźnika występowania cukrzycy. Każdy mieszkaniec społeczności w wieku powyżej 5 lat jest proszony o poddanie się standaryzowanemu badaniu co dwa lata, które obejmuje test tolerancji glukozy. Cukrzyca jest diagnozowana zgodnie z kryteriami Światowej Organizacji Zdrowia i zostaje stwierdzona gdy zajdzie dowolny z dwóch przypadków opisanych poniżej:

- 2 godziny po badaniu przesiewowym poziom glukozy we krwi wyniósł co najmniej 200 mg/dl
- szpital Indian Health Service obsługujący społeczność stwierdził stężenie glukozy co najmniej 200 mg/dl podczas rutynowej opieki medycznej.

2 Opis danych

Zbiór danych poddany analizie w tym projekcie opiera się na 5-letnim okresie badań instytutu i obejmuje jedynie kobiety powyżej 21 roku życia. Dla każdej z pacjentek zostało zebranych 8 zmiennych objaśniających, które zostały już wyszczególnione jako znaczące czynniki ryzyka cukrzycy wśród Pimasów lub innych populacji. Poniżej wybrane zmienne zostały opisane, a pogrubioną czcionką przedstawione odpowiadające im nazwy w środowisku R.

- Liczba ciąż - **Pregnancies**
- Poziom glukozy po 2 godzinach w teście tolerancji glukozy (po spożyciu 75 gramów roztworu węglowodanowego) - **Glucose**
- Ciśnienie krwi rozkurczowe (mm Hg) - **BloodPressure**
- Grubość fałdu skórnego tricepsu (mm) - wygodny pomiar do oceny ogólnej otyłości - **SkinThickness**
- Poziom insuliny po 2 godzinach (µU/ml) - **Insulin**
- Wskaźnik BMI (Masa w kg / (Wzrost w m)^2) - **BMI**
- 'Funkcja rodowodu cukrzycy' - **DiabetesPedigreeFunction**
- Wiek (lata) - **Age**

Kluczowy jest fakt, że w zbiorze wybrano tylko jedno, szczególne, badanie dla danego pacjenta. Wyniki stężenia glukozy w badaniu musiały być poniżej progu określającego występowanie cukrzycy (zatem poniżej 200 mg/dl). Oznacza to zatem, że w zbiorze, na moment badań, nie występują żadne osoby ze zdiagnozowaną cukrzycą. W zbiorze danych jest jednak dostępna zmienna określająca występowanie cukrzycy u pacjenta w ciągu 5 lat od badania (w środowisku R nazwana **Outcome**), będąca naszą zmienną objaśnianą. Precyzując, zmienna ta jest binarna i przyjmuje dwie wartości:

- **1**, gdy cukrzyca została zdiagnozowana w ciągu pięciu lat od badania
- **0**, gdy badanie glukozy wykonane pięć lub więcej lat później nie wykazało cukrzycy.

Podczas gdy większość zmiennych jest prosta w intepretacji, gdyż są to często spotykane pomiary w medycynie, zmienna 'Funkcja rodowodu cukrzycy' jest bardziej złożona. Jest ona między innymi związana z występowaniem cukrzycy w rodzinie pacjenta (więcej tutaj). Postanowiłem wykluczyć zmienną z analizy w ramach tego projektu, aby uniknąć pogłębiania się w skomplikowaną interpretację tej specyficznej zmiennej. Pozwoli to skoncentrować się i poświęcić szczególną uwagę pozostałym cechom, umożliwiając bardziej szczegółową analizę i lepsze zrozumienie ich wpływu na występowanie cukrzycy.
Zbiór danych obejmuje informacje pochodzące od 768 unikalnych pacjentów, z których każdy charakteryzuje się według 7 zmiennych objaśniających.

3 Wstępna eksploracja danych

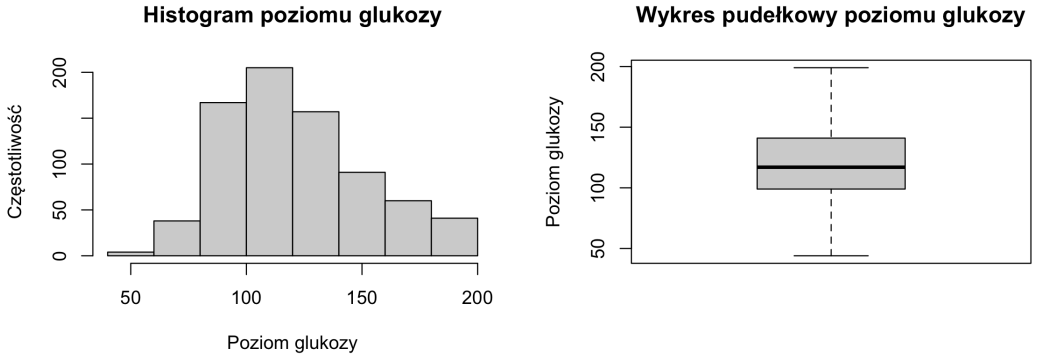
Na początku skupimy się na dokładnej analizie każdej ze zmiennych niezależnie. Przeanalizujemy rozkłady oraz sprawdzimy, czy występują anomalie czy brakujące dane. Dodatkowo rozważymy pogrupowanie niektórych cech, tworząc w ten sposób zmienne jakościowe. Taki wstępny przegląd pozwoli na lepsze zrozumienie charakterystyki każdej zmiennej, co jest niezbędne do dalszej analizy problemu.

Zmienna Pregnancies
Zmienna dotycząca liczby ciąż posiada wartości z zakresu od 0 do 17. Postanowiłem przemianować zmienną na jakościową uporządkowaną, opisującą różne liczebności dzieci. Zależało mi szczególnie na uwzględnieniu osobnej grupy, gdzie kobiety nie zaszły w ciąże, aby sprawdzić czy wyróżnia się ona w pewien sposób. Reszta podziałów powstawała z intencją zachowania podobnej liczności w grupach. Podział jest przedstawiony w tabeli poniżej.

Liczba ciąż	0	1	2	[3,4]	[5,7]	>7
Liczność grupy	111	135	103	143	152	124

Tabela 1: Podział badanej populacji ze względu na liczbę ciąż

Zmienna Glucose
Zmienna określająca poziom glukozy posiada 5 zerowych wpisów. Taki wynik nie jest możliwy, zatem biorąc pod uwagę niewielką liczbę takich obserwacji w stosunku do wielkości badanej populacji, postanowiłem usunąć je ze zbioru. Przedstawmy teraz graficznie rozkład zmiennej w naszym zbiorze.



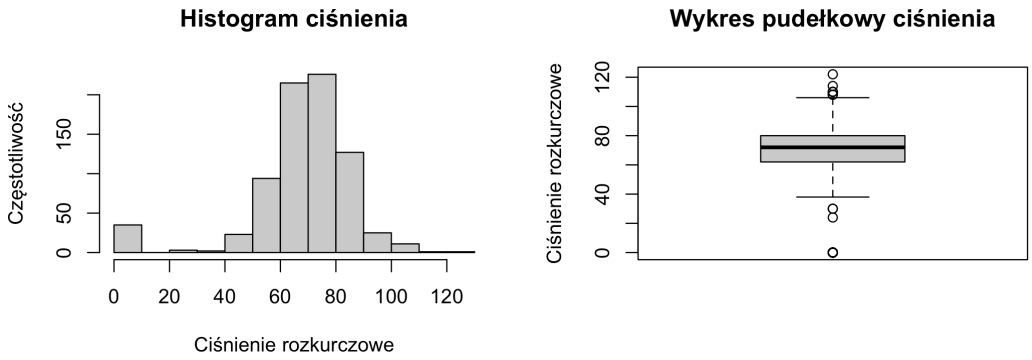
Rysunek 1: Rozkład zmiennej Glucose w zbiorze danych

Zgodnie z opisem, w zbiorze danych nie ma pacjentów o poziomie glukozy powyżej 200 mg/dl. Opierając się na powszechnych informacjach, prawidłowy poziom glukozy we krwi po zjedzeniu powinien wynosić poniżej 140 mg/dl. W zbiorze duża część pacjentów posiada prawidłowy poziom glukozy, jednak występuje również duża grupa o większym poziomie, wykazująca stan przedcukrzycowy. Z drugiej strony, występują również pojedyncze, niepokojąco niskie wyniki poziomu glukozy. Postanowiłem przemianować zmienną na jakościową uporządkowaną o dwóch poziomach, sugerując się powszechnie stosowanym progiem 140 mg/dl. Poniżej przedstawiony jest podział w postaci tabeli.

Poziom glukozy	prawidłowy (<140 mg/dl)	wysoki ($\geq$ 140 mg/dl)
Liczność grupy	566	197

Tabela 2: Podział badanej populacji ze względu na poziom glukozy

Zmienna BloodPressure
Analizę zmiennej opisującej ciśnienie rozkurczowe rozpocznijmy od przedstawienia jej rozkładu w zbiorze danych.



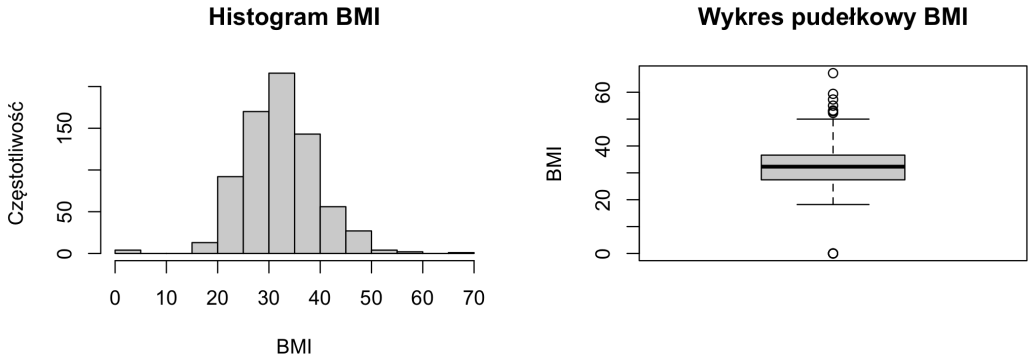
Rysunek 2: Rozkład zmiennej BloodPressure w zbiorze danych

Widać, że w zbiorze występują pomiary ciśnienia równe 0. Również taki pomiar nie jest możliwy, zatem, uwzględniając stosunkowo małą liczbę tego rodzaju wpisów, badanych z takim ciśnieniem usuwam ze zbioru danych. Biorąc pod uwagę, że prawidłowy zakres ciśnienia krwi mieści się między 80 a 90, obserwujemy, że badane plemię wykazuje tendencję do lekko zaniżonego ciśnienia, co widać zwłaszcza w świetle średniej wartości ciśnienia krwi wśród jego członków (około 72 mm Hg). Ogólnie rzecz biorąc, pomimo zauważalnie zaniżonych wartości, pomiary ciśnienia krwi wśród członków plemienia mieszczą się w granicach realistycznych wartości, co sugeruje ich wiarygodność i znaczenie dla dalszej analizy.

Zmienna SkinThickness
Zmienna reprezentująca grubość fałdu skórnego tricepsu zawiera znaczącą liczbę, aż 194, wartości zerowych. Choć niskie pomiary w tej kategorii są możliwe, wartość zerowa wydaje się nieprawdopodobna. Mając na uwadze, że tak liczne wartości zerowe mogłyby zakłócić analizę, zwłaszcza gdyby zastąpić je medianą czy średnią, postanowiłem wyeliminować tę zmienną z naszego zbioru danych. Taki krok nie tylko zwiększy wiarygodność naszych dalszych rozważań, ale również uprości proces analizy.

Zmienna Insulin
Podobnie jak w przypadku zmiennej reprezentującej grubość fałdu skórnego tricepsu, zmienna odzwierciedlająca poziom insuliny wśród badanych również wykazuję problematyczną charakterystykę. W zbiorze danych występuje aż 335 przypadków, gdzie pomiar insuliny wynosi zero, co jest niemożliwe. Wobec tej znacznej liczby nieprawdopodobnych wartości, zdecydowałem się usunąć tę zmienną ze zbioru danych, dążąc tym samym do zwiększenia spójności zbioru i wiarygodności analizy.

Zmienna BMI
Dla zmiennej określającej wskaźnik BMI rozkład w omawianym zbiorze prezentuje się następująco.



Rysunek 3: Rozkład zmiennej BMI w zbiorze danych

Histogram zmiennej BMI wykazuje kształt zbliżony do rozkładu normalnego, co może wydawać się intuicyjnie właściwe. Niemniej jednak, można zauważyć również obecność wartości zerowych dla tej zmiennej, które nie są realistyczne. Ponieważ takich przypadków jest zaledwie cztery, zdecydowałem się na usunięcie tych obserwacji ze zbioru.

Zmienna Age
Zmienna Age w naszym zbiorze danych obejmuje zakres wieku od 21 lat (co jest zgodne z opisem zebranych informacji) do 81 lat. Zdecydowałem się podzielić tę zmienną na kilka grup wiekowych, aby umożliwić dokładniejszą analizę w poszczególnych kategoriach wiekowych. Takie podejście pozwoli na głębsze wnioski dotyczące potencjalnych różnic w badanych zjawiskach w zależności od wieku badanych. Podział jest przedstawiony w tabeli poniżej.

Wiek kobiety	<24	[24,30]	[31,40]	>40
Liczność grupy	161	229	148	186

Tabela 3: Podział badanej populacji ze względu na wiek pacjentów

Ostatecznie w zbiorze danych mamy 724 różnych badanych oraz pięć zmiennych objaśniających, w skład których wchodzą trzy zmienne jakościowe oraz dwie ilościowe.

4 Analiza wpływu zmiennych objaśniających

Chociaż jestem świadom, że zmienne w naszym zbiorze danych zostały już wcześniej wyselekcjonowane na podstawie badań biologicznych jako istotne dla diagnozy cukrzycy, postanowiłem przeprowadzić własną, niezależną analizę statystyczną. Taki krok pozwoli na weryfikację i pogłębienie wiedzy o wpływie tych zmiennych na występowanie choroby, a także na identyfikację potencjalnych wzorców i zależności. Przejdźmy teraz do szczegółowej analizy każdej zmiennej z osobna ze zmienną objaśnianą. Rozpocznijmy od zmiennych jakościowych.

4.1 Zmienne jakościowe

Zmienna Age i występowanie cukrzycy
Na początek sprawdźmy, czy wiek wpływa na wystąpienie cukrzycy u badanego. W tym celu zbudujmy tablicę kontyngencji, jako że obie cechy są zmiennymi jakościowymi.

	cukrzyca	
wiek	nie	tak
<24	139	22
[24, 30]	170	59
[31, 40]	78	70
>40	88	98

Tabela 4: Tabela kontyngencji dla zmiennych Age i Outcome

Wyznaczmy estymowane prawdopodobieństwo zdiagnozowania cukrzycy w każdej z grup oraz wyliczmy przedziały ufności dla prawpodobieństwa.

	Estymowane p	Przedział typu Walda	Przedział typu Wilsona
<24	0.137	(0.084, 0.190)	(0.092, 0.198)
[24, 30]	0.258	(0.201, 0.314)	(0.205, 0.318)
[31, 40]	0.473	(0.393, 0.553)	(0.394, 0.553)
>40	0.527	(0.455, 0.599)	(0.455, 0.597)

Tabela 5: Estymowane prawdopodobieństwa występowania cukrzycy w poszczególnych grupach wiekowych

Po analizie tabeli, widać pewien trend wzrostu prawdopodobieństwa występowania cukrzycy wraz z wiekiem pacjentów. Przedziały ufności typu Walda i Wilsona są dość szerokie, ale zachowują spójny wzrost, co potwierdza wywnioskowaną tendencję. Dodatkowo, przeprowadzając test chi-kwadrat pozwalający ocenić, czy istnieje statystycznie istotna zależność między grupami wiekowymi a występowaniem cukrzycy, otrzymujemy p-wartość$< 2.2 \times 10^{-16}$.

Możemy zatem z dużą pewnością odrzucić hipotezę zerową mówiącą o braku związku między wiekiem a ryzykiem cukrzycy. Wyniki testu potwierdzają, że wiek ma statystycznie istotny wpływ na występowanie cukrzycy w badanej populacji.

Zmienna Glucose i występowanie cukrzycy
Następnie przeanalizujmy wpływ poziomu glukozy na występowanie cukrzycy. Wydaje się być oczywistym, że ta zmienna będzie miała duży wpływ, gdyż sama diagnoza cukrzycy opiera się właśnie na poziomie glukozy. Jednak może nadzwyczajnie większy wynik pojedynczego badania wcale nie świadczy o potencjalnej diagnozie cukrzycy w przyszłości? Rozpocznijmy analizę od przygotowania tablicy kontyngencji dla tych dwóch zmiennych jakościowych.

	cukrzyca	
poziom glukozy	nie	tak
prawidłowy	413	121
wysoki	62	128

Tabela 6: Tabela kontyngencji dla zmiennych Glucose i Outcome

Już po samych wartościach w tablicy widać, że poziom glukozy podczas badania może mieć znaczący wpływ na diagnozę cukrzycy w przyszłości. W grupie o wysokim poziomie glukozy zdiagnozowano więcej osób z cukrzycą niż w grupie z prawidłowym poziomem, mimo, że grupa jest znacznie mniej liczna. Wyliczmy estymowane prawdopodobieństwa wystąpienia cukrzycy w każdej z grup oraz odpowiadające im przedziały ufności.

	Estymowane p	Przedział typu Walda	Przedział typu Wilsona
prawidłowy	0.227	(0.191, 0.262)	(0.193, 0.264)
wysoki	0.674	(0.607, 0.740)	(0.604, 0.736)

Tabela 7: Estymowane prawdopodobieństwa występowania cukrzycy w grupach z różnym poziomem glukozy

Widzimy znaczące różnicę między grupami, zarówno przy estymowanych prawdopodobieństwach, jak i przedziałach ufności. Oznacza to zatem, że ta zmienna może być kluczowa przy predykcji występowania cukrzycy u pacjenta.

Zmienna Pregnancies i występowanie cukrzycy
Ostatnią zmienną jakościową, jaką poddamy analizie pod kątem wpływu na naszą zmienną objaśnianą jest liczba ciąży badanej. Wyliczmy estymowane prawdopodobieństwa dla każdego z poziomów tej cechy.

	Estymowane p	Przedział typu Walda	Przedział typu Wilsona
0	0.323	(0.231, 0.415)	(0.239, 0.420)
1	0.214	(0.144, 0.284)	(0.152, 0.292)
2	0.149	(0.085, 0.212)	(0.096, 0.223)
$[3, 4]$	0.313	(0.239, 0.388)	(0.245, 0.391)
$[5, 7]$	0.364	(0.290, 0.438)	(0.294, 0.441)
>7	0.387	(0.313, 0.461)	(0.317, 0.462)

Tabela 8: Estymowane prawdopodobieństwa występowania cukrzycy w grupach z różną liczbą ciąż

Widać, że zależność między prawdopodobieństwami w poszczególnych grupach wcale nie jest monotoniczna. Kobiety, które nie przeszły ciąży mają wciąż stosunkowe duże prawdopodobieństwo zdiagnozowania cukrzycy. Należy jednak zauważyć, że dla tej grupy przedziały ufnością są 'najszersze'. Dla grupy z jedną ciążą prawdopodobieństwo jest mniejsze, jednak zdecydowanie najmniejsze prawdopodobieństwo występuje dla grupy badanych, które przeszły 2 ciąże. Widać jednak, że dla kobiet, które wiele razy przeszyły ciąże występuje już większe prawdopodobieństwo zdiagnozowania cukrzycy.
W przypadku tej zmiennej, test chi-kwadrat również wskazuje bardzo niską p-wartość (9.893×10^{-9}), sugerując statystycznie istotny związek między liczbą ciąż, a wystąpieniem cukrzycy u badanego.

4.2 Zmienne jakościowe

Zmienna BMI i występowanie cukrzycy
Aby sprawdzić, czy zmienna ilościowa wpływa na zmienną objaśnianą posłużę się analizą modelu regresji logistycznej. Zbudujmy model, gdzie zmienną objaśniającą będzie cecha BMI i zbadajmy jego strukturę.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.87750	0.43753	-8.862	$< 2e{-}16$ ***
BMI	0.09783	0.01285	7.614	2.65e-14 ***

Tabela 9: Podsumowanie modelu ze zmienną objaśniającą BMI

Współczynnik dla BMI jest pozytywny i statystycznie istotny, co wskazuje, że wzrost wartości BMI jest powiązany ze wzrostem logarytmicznych szans na wystąpienie cukrzycy. Zmienna może mieć zatem znaczący wpływ na diagnozę cukrzycy.

Zmienna BloodPressure i występowanie cukrzycy
Podobną procedurę zastosujmy dla zmiennej określającej ciśnienie rozkurczowe. Zbudujmy model regresji logistycznej oparty jedynie na zmiennej BloodPressure i przeanalizujmy otrzymane wyniki.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.776126	0.493480	-5.626	1.85e-08 ***
BloodPressure	0.029159	0.006613	4.409	1.04e-05 ***

Tabela 10: Podsumowanie modelu ze zmienną objaśniającą BloodPressure

Widać, że zmienna BloodPressure również jest statystycznie istotna. Może ona jednak w mniejszy sposób wyjaśniać zmienną zależną w porównaniu do zmiennej BMI, gdyż zarówno dewiancja jak i wartość kryterium informacyjnego Akaike dla tego modelu jest mniejsza w stosunku do poprzedniego, zbudowanego ze zmienną BMI. Rozwinę ten wątek w dalszej części referatu, przy budowie kolejnych modeli.

5 Budowa modelu

W analizowanym zbiorze danych zmienna objaśniana przyjmuje postać binarną, co sugeruje, że model regresji logistycznej może być odpowiednim narzędziem do przewidywania jej wartości. Użycie tego modelu pozwoli na zrozumienie zależności i oszacowanie prawdopodobieństwa przynależności do jednej z dwóch grup.

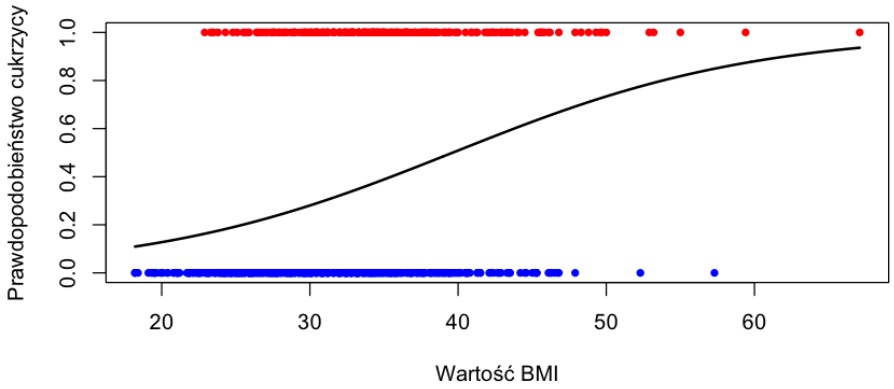
Rozpoczniemy proces budowy modelu od tak zwanego modelu pustego w regresji logistycznej, który nie zawiera żadnych predyktorów poza wyrazem wolnym. Omawiany model wszystkim badanym przypisuje to samo prawdopodobieństwo, równe średniej częstości występowania cukrzycy w analizowanej próbie. Ten model posłuży jako punkt wyjściowy, do którego będziemy systematycznie dodawać zmienne, dążąc do lepszego wyjaśnienia zmienności zmiennej zależnej.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.64586	0.07824	-8.255	$< 2e - 16$ ***

Tabela 11: Podsumowanie modelu pustego

Współczynnik dla wyrazu wolnego wyszedł oczywiście ujemny, co oznacza, że w badanej grupie mniejszość ma zdiagnozowaną cukrzycę. W tym przypadku, Null deviance i Residual deviance mają tę samą wartość (931.94), co jest spodziewane, ponieważ model zawiera tylko wyraz wolny. Wartości te są jednak dość wysokie, co może wskazywać, że model jest prosty i może nie w pełni odzwierciedlać złożoność danych.

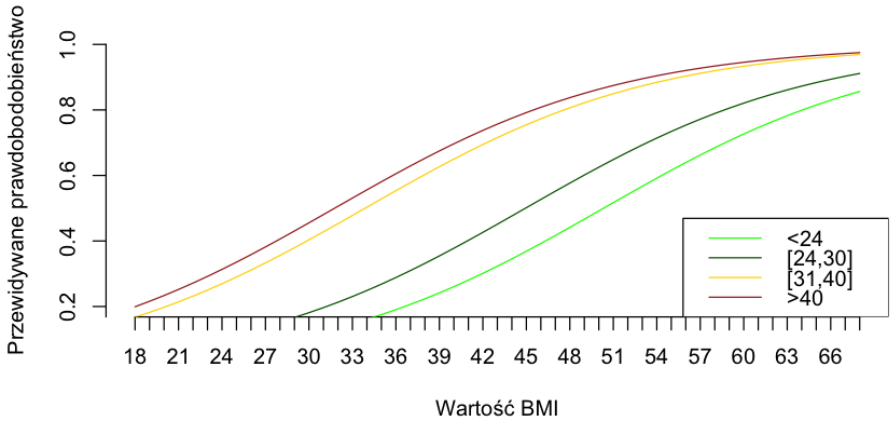
Następnie, w celu wzbogacenia analizy i potencjalnego zwiększenia jakości dopasowania modelu, uwzględnimy zmienną ilościową BMI. Wprowadzenie tej cechy, pozwoli na ocenę, w jaki sposób zmiany tej zmiennej wpływają na prawdopodobieństwo wystąpienia cukrzycy. Oczywiście omawiany model został już zbudowany przy analizie zmiennej BMI. W tym przypadku dewiancja zmniejszyła się (z 931.94 na 865.53), co oznacza lepszą jakość dopasowania modelu. Zwizualizujmy teraz działanie modelu na dostępnych danych.



Rysunek 4: Zależność między BMI a prawdopodobieństwem wystąpienia cukrzycy

Linia czarna reprezentuje przewidywane prawdopodobieństwo wystąpienia cukrzycy na podstawie zbudowanego modelu. Widzimy zatem, że wraz ze wzrostem wartości BMI wzrasta prawdopodobieństwo wystąpienia cukrzycy.

Następnie przejdziemy do dalszego rozszerzenia naszego modelu regresji logistycznej poprzez dodanie dwóch zmiennych jakościowych: wieku i poziomu glukozy. W pierwszym kroku stworzymy model uwzględniający kategorię wiekową, co pozwoli nam ocenić, jak grupy wiekowe różnią się pod względem prawdopodobieństwa zdiagnozowania cukrzycy.



Rysunek 5: Model regresji logistycznej wykorzystujący zmienne objaśniające BMI (ilościowa) i wiek (jakościowa)

Różnica między krzywymi wskazuje na istotny wpływ wieku na związek między BMI a cukrzycą. Oznacza to, że starsze osoby mają wyższe prawdopodobieństwo wystąpienia cukrzycy przy tym samym poziomie BMI w porównaniu z młodszymi osobami. Przeanalizujmy strukturę otrzymanego modelu.

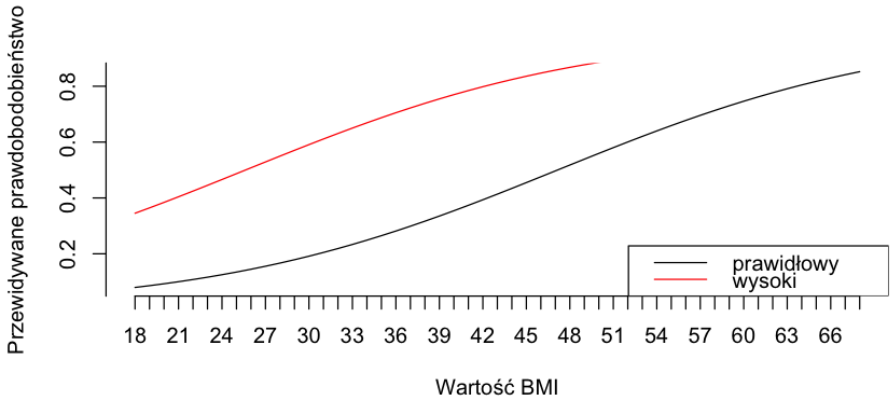
Zmienne	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.08420	0.52015	-9.775	$< 2e-16$ ***
BMI	0.10102	0.01375	7.349	1.99e-13 ***
Age category [24,30]	0.54457	0.28805	1.891	0.0587
Age category [31,40]	1.66247	0.29445	5.646	1.64e-08 ***
Age category >40	1.87474	0.28394	6.603	4.04e-11 ***

Tabela 12: Podsumowanie modelu ze zmienną BMI oraz zmienną jakościową Age

W tabeli zmienna "Age category [24,30]"ma p-wartość równą 0.0587, co jest wyższe niż standardowy poziom istotności statystycznej 0.05. Oznacza to, że związek między tą kategorią wiekową a prawdopodobieństwem wystąpienia cukrzycy nie jest statystycznie istotny na poziomie 0.05. Innymi słowy, nie ma wystarczających dowodów statystycznych, aby stwierdzić, że prawdopodobieństwo diagnozy cukrzycy dla grupy wiekowej [24,30] różni się od grupy referencyjnej (w tym przypadku grupy wiekowej <24), gdy uwzględnimy poziom ufności. Nie przekreśla to jednak znaczenia tej zmiennej całkowicie, gdyż być może w interakcji z dodatkowymi zmiennymi okaże się statystycznie

istotna. Model ten ma zdecydowanie niższą dewiancję od poprzedniego (790.25 zamiast poprzedniego 865.54) oraz niższą wartość kryterium Akaike, zatem ogólnie ma lepszą jakość dopasowania do danych.

Teraz do modelu ze zmienną objaśniającą BMI dodajmy zmienną jakościową określającą kategorię poziomu glukozy, co umożliwi analizę wpływu tego wskaźnika na zmienną objaśnianą. To rozszerzenie ma na celu również potencjalne zwiększenie jakości dopasowania modelu.



Rysunek 6: Model regresji logistycznej wykorzystujący zmienne BMI (ilościowa) i poziom glukozy (jakościowa)

Przedstawiony wykres podkreśla, jak istotny jest poziom glukozy przy diagnozie cukrzycy dla omawianego zbioru danych. Widać znaczącą różnicę w przewidywanych prawdopodobieństwach wystąpienia cukrzycy. Biorąc określony wskaźnik BMI, różnica w przewidywanych prawdopodobieństwach wystąpienia cukrzycy dla grup jest bardzo duża (krzywe są bardzo od siebie oddalone).

Zmienna	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.96081	0.46915	-8.443	$< 2e-16$ ***
BMI	0.08400	0.01371	6.128	8.89e-10 ***
Glucose category (wysoki)	1.80859	0.19177	9.431	$< 2e-16$ ***

Tabela 13: Podsumowanie modelu ze zmienną BMI oraz zmienną jakościową Glucose

Przy tym modelu możemy zauważyć, że współczynnik dla zmiennej jakościowej dotyczącej glukozy jest szczególnie wysoki, co sugeruje, że wysoki poziom glukozy we krwi jest silnym predyktorem wystąpienia cukrzycy u pacjenta. Również jak w poprzednich modelach, wyższe wartości BMI zwiększają logarytm szans na wystąpienie cukrzycy. Dla tego modelu wartości dewiancji oraz kryterium Akaike znacząco zmalały (odpowiednio 770.42 oraz 776.42). Oznacza to, że zmienna określająca poziom glukozy znacząco poprawia jakość modelu i należy ją rozważać przy modelowaniu zmiennej zależnej.

Teraz uwzględnijmy w modelu zarówno zmienną BMI jak i obie zmienne jakościowe Age oraz Glucose. Zbadajmy jak działa model przy takim zestawie zmiennych objaśniających.

Zmienna	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.00388	0.54179	-9.236	$< 2e-16$ ***
BMI	0.08735	0.01427	6.122	9.24e-10 ***
Age category [24,30]	0.51602	0.30320	1.702	0.0888 .
Age category [31,40]	1.53743	0.30969	4.964	6.89e-07 ***
Age category >40	1.60555	0.29963	5.358	8.40e-08 ***
Glucose category (wysoki)	1.61105	0.19900	8.096	5.70e-16 ***

Tabela 14: Podsumowanie modelu ze zmienną BMI oraz zmiennymi jakościowymi Age oraz Glucose

Wpływ wieku w dalszym ciągu jest zróżnicowany: grupy wiekowe [31,40] i >40 wykazują istotne statystycznie zwiększenie prawdopodobieństwa cukrzycy w porównaniu do grupy referencyjnej (<24 lat), podczas gdy grupa [24,30] nie wykazuje istotności statystycznej na poziomie 0.05. Zmienna jakościowa określająca poziom glukozy wciąż odgrywa ważną rolę w modelu. Ten model jest również lepszy od poprzednich, co potwierdza ponownie niższa dewiancja (721.45) oraz wartość kryterium Akaike (733.45).

Do tej pory przy budowie uwzględniliśmy trzy z pięciu dostępnych zmiennych objaśniających. Po analizie poszczególnych zmiennych w poprzednich rozdziałach mogliśmy dojść do wniosku, że pozostałe zmienne, zatem liczba ciąż oraz ciśnienie rozkurczowe, będą w najmniejszym stopniu wpływać na wyjaśnienie zmiennej objaśnianej. Przenalizujmy teraz modele, które dodatkowo zawierają omawiane cechy.

Variable	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-4.910	0.620	$< 2 * 10^{-16}$
BMI	0.088	0.015	$2.09 * 10^{-16}$
Age category [24,30]	0.479	0.313	0.126
Age category [31,40]	1.386	0.346	$6 * 10^{-16}$
Age category >40	1.412	0.357	$7 * 10^{-16}$
Glucose category (high)	1.605	0.201	$1.38 * 10^{-16}$
Pregnancies category 1	-0.286	0.357	0.424
Pregnancies category 2	-0.195	0.392	0.618
Pregnancies category [3,4]	0.097	0.344	0.779
Pregnancies category [5,7]	-0.105	0.355	0.768
Pregnancies category >7	0.326	0.380	0.392

Tabela 15: Model z dodaną zmienną Pregnancies

Variable	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-4.558	0.685	$2.82 * 10^{-16}$
BMI	0.092	0.015	$8.97 * 10^{-16}$
Age category [24,30]	0.511	0.303	0.091
Age category [31,40]	1.576	0.312	$4.30 * 10^{-16}$
Age category >40	1.676	0.307	$4.81 * 10^{-16}$
Glucose category (high)	1.630	0.200	$4.06 * 10^{-16}$
BloodPressure	-0.009	0.008	0.298

Tabela 16: Model z dodaną zmienną BloodPressure

Jak widać w tabelach, żadna z dodanych zmiennych nie jest statystycznie istotna w kontekście predykcji wystąpienia cukrzycy. Każda z cech ma p-wartość wyższą od poziomu istotności ustalonego na 0.05 (wartości czerwone w tabelach powyżej). Również w dalszym ciągu grupa wiekowa [24,30] nie jest jest statystycznie istotna.

Biorąc pod uwagę te wyniki, model zawierający trzy zmienne objaśniające, wskaźnik BMI, grupy wiekowe oraz poziomy glukozy jest najlepszym spośród rozważanych. Ma on dobrą jakość dopasowania do danych oraz zawiera głównie zmienne będące statystycznie istotne przy określaniu występowania cukrzycy. Dzięki temu, że składa się z tylko trzech zmiennych niezależnych, jest też modelem prostym w interpretacji.

6 Podsumowanie

Podczas tego referatu dokonałem dogłębnej analizy czynników wpływających na występowanie cukrzycy u badanych. Mimo, iż dostępne zmienne objaśniające były już sklasyfikowane jako znaczące w kontekście wpływu na cukrzycę, dodatkowa analiza statystyczna pozwoliła wyszczególnić zmienne będące najbardziej istotne statystycznie w kontekście objaśniania zmiennej zależnej. Zmienne określające wskaźnik BMI, grupę wiekową badanego czy poziom glukozy okazały się być najbardziej wpływowe. Dokładniej mówiąc, każda z opisanych zmiennych miała nawet monotoniczny wpływ na prawdpodobieństwo diagnozy cukrzycy - wraz ze wzrostem wartości czy też wartości określanych przez grupy uporządkowane, prawdopodobieństwo wystąpienia choroby wzrastało.

Zbudowany model regresji logistycznej oparty na trzech wymienionych zmiennych okazał się najlepszy spośród wszystkich omówionych. Oprócz poziomu [24,30] dla zmiennej jakościowej określającej wiek, każda ze zmiennych jest statystycznie istotna przy określaniu występowania cukrzycy. Dodatkowo, p-wartość dla tej zmiennej nie jest zdecydowanie wyższa od ustalonego poziomu istotności 0,05. Postanowiłem również przetestować jakość predykcyjną modelu, dzieląc zbiór danych na zbiór treningowy oraz testowy w stosunku 80/20. W ten sposób model estymuje wartości parametrów na zbiorze treningowym i zostaje oceniony poprzez predykcję na zbiorze testowym. Dokładność modelu wynosi w tym przypadku prawie 80%, co jest zadowalającym wynikiem. Warto dodać, że przy tym podejściu, żaden inny model nie uzyskał tak wysokiej dokładności.