Studium Licencjackie

Kierunek: Metody Ilościowe w Ekonomii i Systemy Informacyjne

Imię i nazwisko autora: Paulina Skalik
Nr albumu: 108617

Wykorzystanie metod uczenia maszynowego w ocenie ryzyka zachorowania na cukrzycę typu II

Praca licencjacka
pod kierunkiem naukowym
dr hab., prof. SGH Barbara Kowalczyk
Instytut Ekonometrii

Warszawa 2023

Spis treści

Wstęp

Według International Diabetes Fundation w 2021 na cukrzycę chorowało ponad pół miliarda osób, czyli 8.5% dorosłej populacji świata. Choroba ta nie tylko wpływa negatywnie na jakość życia pacjenta, lecz niesie ze sobą zwiększone ryzyko wystąpienia poważnych chorób serca i nerek, utraty wzroku oraz także przedwczesnej śmierci. Wczesne wykrycie choroby wpływa pozytywnie na rokowania pacjenta, jednak prawie co drugi chory na świecie i co czwarty w Stanach Zjednoczonych nie jest świadomy występowania u niego choroby.[4] Z tego powodu ważne jest wyróżnienie czynników ryzyka sprzyjających powstawaniu choroby i objęcie zagrożonej grupy odpowiednią profilaktyką.

Z perspektywy ekonomicznej bezpośrednie koszty leczenia cukrzycy na świecie szacowane są na 966 miliardów dolarów. Przewlekłe choroby takie jak cukrzyca niosą za sobą także koszty pośrednie w postaci kosztów utraconej produktywności chorych do których należy między innymi ograniczona wydajność, czas przeznaczony na wizyty lekarskie, całkowita niezdolność do pracy czy przedwczesna śmiertelność. Z uwzględnieniem kosztów pośrednich, całkowite koszty cukrzycy szacowane były w 2015 roku na 1.3 tryliona dolarów rocznie i według prognoz do 2035 roku mogą wzrosnąć nawet do 2.5 tryliona dolarów co stanowi 2% światowego PKB.[13] Szacunki te nie uwzględniają jednak kosztów jakie stanowi utracona produktywność osób opiekujących się chorymi, których wciąż przybywa. Prognozowane jest, że do 2045 roku głównie z powodu starzejącego się społeczeństwa liczba chorych na cukrzycę wzrośnie o 16%.[4]

Aby walczyć ze skutkami jakie niesie cukrzyca należy skupić się na wykrywaniu czynników, które stawiają osobę w grupie ryzyka oraz tych, które minimalizują ryzyko zachorowania. W ten sposób można nie tylko objąć odpowiednie osoby profilaktyką, ale także wdrożyć do tej profilaktyki odpowiednie działania prewencyjne oraz rekomendować je całej populacji.

Celem mojej pracy jest identyfikacja czynników pozwalających zakwalifikować pacjenta do grupy ryzyka jak i czynników zapobiegających rozwojowi choroby. Analizy zostały przeprowadzone na danych pochodzących z corocznej ankiety przeprowadzanej przez amerykańskie Center of Disease Control dotyczącej głównie stanu zdrowia i praktyk zdrowotnych obywateli Stanów Zjednoczonych.[32] Wszystkie analizy zostały przeprowadzane z wykorzystaniem języka programowania R.

Pierwszy rozdział przedstawia obraz kliniczny cukrzycy, jej rozpowszechnienie w Polsce i Stanach Zjednoczonych oraz wyniki podobnych analiz czynników ryzyka oraz czynników prewencyjnych. W podrozdziale 1.1 omawiana jest diagnostyka, przebieg oraz

konsekwencje cukrzycy. Kolejne trzy podrozdziały dotyczą rozprzestrzenienia choroby kolejno na świecie, w Polsce oraz Stanach Zjednoczonych. Podrozdział 1.5 opisuje bezpośrednie i pośrednie koszty ekonomiczne cukrzycy z uwzględnieniem potencjalnych braków w przeprowadzanych szacunkach. W kolejnym podrozdziale nakreślone są potencjalne czynniki zwiększające ryzyko zachorowania. Podrozdział 1.6 skupia się na rekomendowanych przez różne organizacje i badaczy działaniach prewencyjnych, a podrozdział 1.7 jest przeglądem badań prowadzonych nad modelowaniem ryzyka zachorowania na cukrzycę z wykorzystaniem metod uczenia maszynowego.

W rozdziale drugim znajduje się opis charakterystyki zbudowanych modeli oraz opis danych wykorzystanych do ich budowy. Podrozdział 2.1 skupia się na metodzie pozyskania danych, a kolejne dwa opisują kolejno sposób transformacji zmiennej objaśnianej oraz wybranych zmiennych objaśniających. Podrozdział 2.4 pokrywa zasady działania i budowy wykorzystywanych modeli, a kolejny podrozdział pokrywa metody oceny ich jakości.

Rozdział trzeci obejmuje sposób przygotowania czterech zbiorów wykorzystywanych do budowy modeli oraz wyniki i ocenę jakości zbudowanych już modeli. Rozdział 3.1 zawiera opisy sposobu czyszczenia danych, radzenia sobie z brakami danych oraz analizę rozkładu zmiennej objaśnianej i zmiennych objaśniających. Następne cztery rozdziały przedstawiają model logitowy, model logitowy z wykorzystaniem algorytmu *stepwise*, drzewo decyzyjne oraz las losowy budowane na każdym z czterech zbiorów danych oraz ocenę ich jakości. Zbudowane model pozwalają zaklasyfikować pacjenta do grupy wysokiego ryzyka występowania cukrzycy za pomocą odpowiedzi na 24 pytania.

Zakończenie skupia się na porównaniu między sobą modeli pod kątem wybranych kryteriów i wyborze najlepszego klasyfikatora pod ich względem.

Rozdział I

Epidemiologia cukrzycy

1.1 Cukrzyca i jej konsekwencje zdrowotne

Światowa Organizacja Zdrowia definiuje cukrzycę jako przewlekłą chorobę metaboliczną charakteryzującą się podwyższonym poziomem glukozy we krwi, nazywanym inaczej hiperglikemią.[1] Przyczyną hiperglikemii u pacjentów chorujących na cukrzycę jest nieprawidłowa praca trzustki, która nie produkuje wystarczającej ilości insuliny regulującej poziom glukozy we krwi lub nieefektywne wykorzystanie wyprodukowanej insuliny przez organizm chorego. Według Polskiego Towarzystwa Diabetologicznego najbardziej charakterystycznymi objawami hiperglikemii są: wielomocz, wzmożone pragnienie, utrata masy ciała bez zmiany nawyków żywieniowych i stylu życia oraz inne mniej typowe objawy jak np. osłabienie, wzmożona senność, zmiany ropne na skórze oraz stan zapalny narządów moczowo-płciowych. Utrzymująca się hiperglikemia prowadzi do uszkodzenia, dysfunkcji i niewydolności serca, naczyń krwionośnych, oczu, nerek oraz układu nerwowego. Pacjenci z cukrzycą mają 2-3 krotnie większe ryzyko wystąpienia zawału serca jak i stanowią większość pacjentów z zaburzeniami pracy nerek. Dodatkowym powikłaniem jest także występowanie tzw. stopy cukrzycowej. Zjawisko to może prowadzić nawet do konieczności amputacji kończyny dolnej.[2] Choroba ta także zwiększa prawdopodobieństwo ciężkiego przebiegu chorób zakaźnych, w tym także COVID-19. Badania pokazały także, że cukrzyca wpływa nie tylko na zdrowie fizyczne chorych, ale także na zdrowie psychiczne. W Polsce pacjenci z cukrzycą gorzej od osób zdrowych oceniają jakość swojego życia oraz częściej zapadają na depresję.[3] Najczęściej występujące rodzaje cukrzycy to cukrzyca typu I, cukrzyca typu II oraz cukrzyca ciążowa. Powyżej 95% pacjentów diagnozowanych z cukrzycą choruje na cukrzycę typu II.

Według wytycznych WHO cukrzyca jest diagnozowana w momencie, kiedy spełniane jest co najmniej jedno z trzech kryteriów:

- występują symptomy hiperglikemii oraz poziom cukru we krwi mierzony w dowolnym momencie w ciągu dnia jest na poziomie wyższym niż 200 mg/dl,

---

[1] WHO, *Diabetes,* https://www.who.int/news-room/fact-sheets/detail/diabetes (dostęp: 5.12.2022)

[2] Araszkiewicz, A., Bandurska-Stankiewicz, E., Borys, S., Budzyński, A., Cyganek, K., Cypryk, K., Czech, A., Czupryniak, L., Drzewoski, J., Dzida, G., Dziedzic, T., Franek, E., Gajewska, D., Gawrecki, A., Górska, M., Grzeszczak, W., Gumprecht, J., Idzior-Waluś, B., Jarosz-Chobot, P., . . . Moczulski, D. (2021). 2021 Guidelines on the management of patients with diabetes. A position of Diabetes Poland*, Clinical Diabetology*, 10(1), s.1–113.

[3] Bąk, E., Nowak - Kapusta, Z., Dobrzyn-Matusiak, D., Marcisz-Dyla, E., Marcisz, C., & Krzemińska, S. (2019). An assessment of diabetes-dependent quality of life (ADDQoL) in women and men in Poland with type 1 and type 2 diabetes, *Annals of Agricultural and Environmental Medicine*, 26(3), s.429–438.

- poziom cukru mierzonego na czczo (po 8 godzinnym niespożywaniu kalorii) jest wyższy niż 126 mg/dl podczas dwóch pomiarów,
- doustny test obciążenia glukozą w 120 minucie badania wskazuje poziom glukozy na poziomie wyższym niż 200 mg/dl,
- stężenie HbA1c (hemoglobiny glikowanej) jest wyższe niż 6.5%.

Dodatkowymi czynnikami pozwalającymi zdecydować o dokładnej diagnozie są m.in. wiek badanego, jego masa ciała oraz historia występowania choroby w rodzinie. Dzięki dodatkowym informacjom można rozróżnić cukrzycę typu I od cukrzycy typu II. Pierwsza z nich objawia się głównie u dzieci, nastolatków i młodych dorosłych, a czynniki sprzyjające pojawieniu się jej nie są dokładnie znane. Naukowcy jednak zgadzają się co do tego, że są to nieokreślone czynniki genetyczne i środowiskowe. Cukrzyca typu II pojawia się najczęściej w późniejszym wieku i jej czynniki ryzyka, które będą przedstawione w dalszym podrozdziale, są dużo lepiej znane.

1.2 Epidemiologia cukrzycy na świecie

Według International Diabetes Federation[4] na całym świecie żyje około 537 milionów dorosłych w wieku od 20-79 chorujących na cukrzycę. Stanowi to 9.8% całej populacji. Ponad 75% zdiagnozowanych pacjentów pochodzi z krajów średnio oraz mało zamożnych, co sugeruje wpływ czynników społeczno-ekonomicznych na występowanie choroby. W państwach zaklasyfikowanych przez Bank Światowy jako kraje o wysokich dochodach cukrzyca dotyka 8.4% populacji, gdy w państwach zaklasyfikowanych jako państwa o niskich lub średnich dochodach jest to 10.1% populacji.[5] Dodatkowo statystyki dotyczące krajów mniej zamożnych mogą być znacząco zaniżone, jako że wykrywalność cukrzycy w tych regionach jest prawdopodobnie na poziomie poniżej 50% przez niską jakość opieki zdrowotnej oraz jej niską dostępność. IDF prognozuje także, że liczba chorych wzrośnie do 643 milionów do roku 2030 i do 783 milionów do roku 2045. Największe nasilenie choroby obserwowane jest na Bliskim Wschodzie i Północnej Afryce, gdzie dotyka ona 1 na 6 dorosłych. W tych miejscach oraz w Afryce środkowej i południowej IDF przewiduje najwyższe wzrosty zachorowalności w ciągu najbliższych 25 lat. Jest to 87% więcej przypadków dla Bliskiego Wschodu i Afryki

[4] International Diabetes Federation, *IDF Diabetes Atlas 10th Edition*, https://diabetesatlas.org/data/en/ (dostęp: 15.12.2022)
[5] World Bank, *Diabetes prevalence (% of population ages 20 to 79),* https://data.worldbank.org/indicator/SH.STA.DIAB.ZS (dostęp: 15.12.2022)

Północnej oraz 134% dla pozostałej części Afryki. W 2021 roku cukrzyca była odpowiedzialna za 6.7 miliona zgonów, co jest równoznaczne z jedną śmiercią z powodu cukrzycy co 5 sekund.

1.3 Epidemiologia cukrzycy w Polsce

W Polsce cukrzyca zajmuje dziewiąte miejsce w klasyfikacji najczęściej występujących chorób przewlekłych i dotyka 8.1% populacji w wieku powyżej 15 lat.[6] W 2018 roku liczba zarejestrowanych chorych wyniosła 2,65 mln pacjentów.[7] Według raportu Głównego Urzędu Statystycznego „Zdrowie i Ochrona Zdrowia" wydanego w 2020 roku zachorowalność obliczana według osoby na 10 000 ludności wynosiła w 2016 roku 112.2 a chorobowość okresowa 815.7. Zachorowalność dotyczy nowo stwierdzonych przypadków w danym roku, a chorobowość ogólnego rozprzestrzenienia choroby wśród społeczeństwa. Na cukrzycę w Polsce częściej zapadają kobiety niż mężczyźni, chorobowość okresowa na 10000 mieszkańców wyniosła 879,1 dla kobiet oraz 744,9 dla mężczyzn. W latach 2013-2018 zachorowalność na cukrzycę wzrosła o 13.7%. Jest to częściowo spowodowane zmianami demograficznymi i starzeniem się polskiego społeczeństwa, lecz nie jest to jedyny powód, ponieważ dynamika zmian demograficznych nie jest tak wysoka, jak dynamika wzrostu zachorowań na cukrzycę.[8] Dodatkowo, w ostatnich 10 latach wzrosła liczba zgonów z powodu cukrzycy z 16.9 zgonów na 100 tys. mieszkańców do 31.7 zgonów na 100 tys. mieszkańców.[9] Największy wzrost jest obserwowany między rokiem 2019 a 2020, jest to wzrost o 30% rok do roku w sytuacji, kiedy w poprzednich latach dynamika wzrostu przyjmowała wartości jednocyfrowe. Prawdopodobnie tak gwałtowny wzrost zgonów spowodowany był pandemią Covid-19, która nie tylko przyczyniła się do zatamowania służby zdrowia, jak i sama choroba była dużo bardziej dotkliwa dla osób z chorobami współistniejącymi.[10] Jednak w okresie od 2013 do 2018 roku różnica między nowo zdiagnozowanymi przypadkami cukrzycy a liczbą zgonów z jej powodu zmalała.[11] Dodatkowo, mimo że liczba osób diagnozowanych z cukrzycą wciąż wzrasta, to dynamika tego wzrostu spowalnia.

---

[6] GUS, *Zdrowie i Ochrona Zdrowia w 2020 roku,* https://stat.gov.pl/files/gfx/portalinformacyjny/pl/defaultaktualnosci/5513/1/11/1/zdrowie_i_ochrona_zdrowia_2020_korekta.pdf , (dostęp: 15.12.2022)

[7] Ministesrtwo Zdrowia, *Cukrzyca – Mapy potrzeb zdrowotnych,* https://basiw.mz.gov.pl/analizy/problemy-zdrowotne/cukrzyca-wersja-polska/, (dostęp:15.12.2022)

[8] Ministerstwo Zdrowia, *Cukrzyca w liczbach,* https://pacjent.gov.pl/artykul/cukrzyca-w-liczbach

[9] GUS, *Realizacja Celów Zrównoważonego Rozwoju w Polsce. Raport 2018 - Wskaźnik 3.1.e - Liczba zgonów w wyniku cukrzycy na 100 tys. ludności*, https://sdg.gov.pl/statistics_nat/3-1-e/ (dostęp: 15.12.2022)

[10] Magliano, D., & Boyko, E. J. (2021). *IDF Diabetes Atlas*. International Diabetes Federation.

[11] Instytut Ochrony Zdrowia, *Rekomendacje w zakresie kompleksowej opieki nad pacjentami z retinopatią cukrzycową, https://www.ioz.org.pl/_files/ugd/e91ac2_5d595422c5cf46c69cce6e3ee3aa37db.pdf*

1.4 Epidemiologia cukrzycy w Stanach Zjednoczonych

W Stanach Zjednoczonych według CDC żyje około 28.5 miliona dorosłych ze zdiagnozowaną cukrzycą, a liczba niezdiagnozowanych przypadków szacowana jest na około 8.5 miliona. Rocznie diagnozuje się tam około 1.4 miliona nowych przypadków choroby. Osoby zdiagnozowane chore na cukrzycę stanowią 11.3% całej populacji co pokazuje, że nasilenie tego zjawiska w tym kraju jest nieznacznie większe niż w Polsce. Dodatkową różnicę obserwujemy w rozkładzie płci chorych. W przeciwieństwie do Polski, w Stanach Zjednoczonych na cukrzycę zapadają częściej mężczyźni niż kobiety. Wśród kobiet 10.2% zostało zdiagnozowanych z cukrzycą, gdy wśród mężczyzn jest to 12.6%. Pod kątem etnicznym najbardziej dotknięci chorobą są natywni mieszkańcy terenów Stanów Zjednoczonych, a najniższe współczynniki zachorowalności jak i niezdiagnozowanych przypadków obserwowane są wśród białej populacji. Zjawisko cukrzycy wśród dorosłych w ciągu ostatnich 20 lat przybrało na sile. Szacunkowa liczba pacjentów chorujących na cukrzycę, z uwzględnieniem niezdiagnozowanych przypadków, wzrosła z 10.3% populacji w latach 2001-2004 do 13.2% procent w latach 2017-2020.[12] W grupie wiekowej powyżej 65 roku życia szacuje się, że na cukrzycę choruje prawie co trzeci Amerykanin.

1.5 Ekonomiczne koszty cukrzycy

W 2021 roku według szacunków IDF globalne wydatki na opiekę zdrowotną dla osób z cukrzycą wyniosły 966 miliardów dolarów, co jest wzrostem o 316% procent w ciągu ostatnich 15 lat. Część tego wzrostu mogła być spowodowana poprawą jakości gromadzonych danych, jednak prognozowane są dalsze wzrosty bezpośrednich wydatków na ten cel, do poziomu 1.03 tryliona dolarów w 2030 roku i 1.05 tryliona dolarów w 2045 roku.[4] Koszty cukrzycy to nie tylko bezpośrednie koszty leczenia, ale także koszty utraconej produkcyjności chorych. Według artykułu, który ukazał się w magazynie naukowym *Diabetes Care* w 2015 całkowity koszt ekonomiczny cukrzycy wyniósł 1.3 tryliona dolarów. W tej publikacji przedstawiono także trzy predykcje globalnego ekonomicznego kosztu cukrzycy w 2030. Pierwszy wariant zakłada wzrost zachorowalności i śmiertelności spowodowany tylko starzeniem się społeczeństwa oraz urbanizacją z pozostałymi czynnikami niezmienionymi i przewiduje koszt na poziomie 2.2 tryliona dolarów. W drugim wariancie wzięto pod uwagę wcześniej obserwowane trendy w danych dotyczących zachorowalności oraz śmiertelności i otrzymano prognozowane wydatki na poziomie 2.5 tryliona dolarów. Ostatni

[12] Center of Disease Control, *Prevalence of Both Diagnosed and Undiagnosed Diabetes*, https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html (dostęp: 5.12.2022)

scenariusz zakładał osiągnięcie celu zapadalności na cukrzycę oraz inne choroby niezakaźne wyznaczonego przez WHO na lata 2013-2020 i w tym wypadku koszt cukrzycy prognozowany jest na 2.1 tryliona dolarów. Przewiduje się, że koszty ekonomiczne cukrzycy będą stanowić 2% przewidywanego światowego PKB na rok 2030, z czego najwyższe w stosunku do PKB będą w krajach o średnich dochodach.[13]

Wydatki na leczenie cukrzycy najwyższe są w Stanach Zjednoczonych. Ten kraj zajmuje także drugie miejsce w rankingu wydatków na leczenie cukrzycy przypadających na jednego obywatela, gdzie wyprzedza go jedynie Szwajcaria. Według badania[14] przeprowadzonego przez American Diabetes Association w 2017 roku cukrzyca kosztowała Amerykańską gospodarkę 327 miliardów dolarów, na co złożyło się 237 miliardów dolarów wydanych na opiekę zdrowotną dla chorych oraz 90 miliardów dolarów utraconej produktywności pacjentów. Bezpośrednie koszty leczenia cukrzycy estymowane są jako nadwyżka kosztów opieki nad pacjentem z cukrzycą w porównaniu z przeciętnymi wydatkami na opiekę zdrowotną nad zdrowym pacjentem. Zawiera to między innymi koszty leczenia chorób, na które większe prawdopodobieństwo zapadnięcia mają osoby z cukrzycą. Dodatkowo osoby chore ponoszą zwiększone koszty leczenia chorób niepowiązanych bezpośrednio z cukrzycą z powodu ich cięższego przebiegu. Na koszty utraconej produktywności w omawianej estymacji składają się:

- opuszczone dni w pracy,
- ograniczona produktywność osób zatrudnionych z powodu objawów chorobowych,
- znaczna niezdolność do pracy i w konsekwencji bezrobocie,
- ograniczona produktywność osób poza siłą roboczą w pracach takich jak np. opieka nad dzieckiem czy prace domowe,
- przedwczesna śmiertelność.

Dodatkowe koszty utraconej produktywności, które nie zostały ujęte w omawianym modelu to m.in. koszty utraconej produktywności członków rodzin osób chorych na cukrzycę z powodu opieki nad chorym oraz koszty podróży w celach uzyskania opieki medycznej. Według omawianych predykcji 25% budżetu przeznaczonego na opiekę medycznych zostaje wydane na leczenie osób z cukrzycą, z czego połowa tego budżetu przeznaczana jest na leczenie samej

[13] Bommer, C., Sagalova, V., Heesemann, E., Manne-Goehler, J., Atun, R., Bärnighausen, T., Davies, J., & Vollmer, S. (2018). Global Economic Burden of Diabetes in Adults: Projections From 2015 to 2030, *Diabetes Care*, 41(5), s. 963–970.

[14] American Diabetes Association (2018). Economic Costs of Diabetes in the U.S. in 2017, *Diabetes Care, 41(5),* s. 917–928.

cukrzycy. Także średnie wydatki na opiekę zdrowotną osób z cukrzycą są 2.3 raza większe niż wydatki na ten sam cel osób bez cukrzycy. Bezpośrednio najwyższe koszty ekonomiczne cukrzycy ponoszą ubezpieczyciele oraz pracodawcy, jednak problem ten dotyka także całe społeczeństwo i objawia się w postaci wyższych składek ubezpieczeniowych, wyższych podatków, niższych zarobków oraz obniżonego standardu życia.

1.6 Czynniki ryzyka zachorowania na cukrzycę typu II

Dokładne przyczyny powstawania cukrzycy typu I nie są tak dobrze znane jak w przypadku typu II. Nieznane są sposoby zapobiegania rozwinięcia się choroby. Na ten moment znane czynniki zwiększające ryzyko zachorowania to pokrewieństwo z osobą z cukrzycą typu I oraz wiek, ponieważ choroba rozwija się głównie wśród dzieci i nastolatków. Obserwuje się także zwiększoną zapadalność wśród rasy białej w porównaniu z innymi rasami.

W przypadku cukrzycy typu II zostało zidentyfikowane wiele czynników zwiększających ryzyko zachorowania jak i zostały opracowane programy profilaktyczne, które mają na celu obniżenie tego ryzyka, nawet w przypadku silnych obciążeń genetycznych. Głównymi czynnikami ryzyka wymienianymi przez amerykańskie Center of Disease Control są:

- stan przedcukrzycowy,
- nadwaga (BMI powyżej 25.0),
- wiek powyżej 45 lat,
- cukrzyca typu II występująca u rodzica lub rodzeństwa,
- aktywność fizyczna podejmowana rzadziej niż 3 razy w tygodniu,
- wystąpienie cukrzycy ciążowej w trakcie ciąży lub urodzenia dziecka o wadze wyższej niż 4kg,
- bycie Afroamerykaninem, Latynosem, Amerykańskim Indianinem lub rdzennym mieszkańcem Alaski.

International Diabetes Association zwraca także uwagę na obwód pasa powyżej 80 cm u kobiet oraz powyżej 94 cm u mężczyzn, jako że duża ilość tłuszczu brzusznego zwiększa ryzyku zachorowania. Brytyjski NHS dodaje dodatkowo do listy czynniki ryzyka takie jak podwyższone ciśnienie krwi, palenie papierosów, syndrom policystycznych jajników, spożywanie więcej niż 14 porcji alkoholu tygodniowo, za małą ilość snu, siedzący tryb życia oraz choroby psychiczne takie jak schizofrenia, depresja i zaburzenia afektywne dwubiegunowe.[15] Oprócz chorób psychicznych, także długotrwały stres prowadzący do

[15] Diabetes UK, *Diabetes risk factors*, https://www.diabetes.org.uk/preventing-type-2-diabetes/diabetes-risk-factors (dostęp: 16.12.2022)

podwyższonego poziomu kortyzolu we krwi może wpłynąć na zwiększenie prawdopodobieństwa zachorowania, jako że istnieją badania popierające tezę o negatywnym wpływie wysokiego poziomu hormonów stresu na komórki produkujące insulinę.[16]

Tematem badań jest także wpływ czynników środowiskowych na zwiększone ryzyko zapadnięcia na cukrzycę typu II. Badania[17] przeprowadzone w dobrze rozwiniętych państwach wskazują wpływ zanieczyszczenia powietrza związkami chemicznymi takimi jak NO_2 i $PM_{2.5}$ oraz hałasu na zwiększenie się ryzyka zachorowania. Istnieją także badania sugerujące, że mieszkanie w okolicy, w której znajduje się dużo zieleni oraz tzw. *walkability* jest na wysokim poziomie, może obniżyć ryzyko wystąpienia choroby. *Walkability* definiowane jest jako łatwość poruszania się po okolicy bez wykorzystania środków transportu takich jak samochód czy komunikacja miejska.

Badania wskazują także, że na zwiększone ryzyko choroby nie wpływa tylko ilość spożywanych kalorii, a co za tym idzie wysokość BMI pacjenta, ale także konkretna kompozycja składników odżywczych w diecie. Wykazano, że zwiększone spożycie tłuszczy zwierzęcych może zwiększać ryzyko zachorowania.[18] Ponadto, duży udział błonnika pokarmowego w diecie może wpłynąć na obniżenie omawianego ryzyka.[19] Między innymi z powodu wcześniej wspominanych proporcji składników odżywczych badany jest wpływ diety śródziemnomorskiej na zmniejszenie szansy zapadnięcia na cukrzycę typu II oraz spowolnienie rozwoju choroby po diagnozie. Meta-analiza z 2014 roku potwierdza, że istnieją silne dowody optujące za rekomendowaniem tej diety, jako ograniczającej ryzyko zachorowania oraz dalsze postępowanie choroby.[20]

Analizowana jest również relacja między statusem socjoekonomicznym a ryzykiem zachorowania na cukrzycę typu II. Meta-analiza 23 badań opublikowana w International

---

[16] Merabet, N., Lucassen, P. J., Crielaard, L., Stronks, K., Quax, R., Sloot, P. M., la Fleur, S. E., & Nicolaou, M. (2022). How exposure to chronic stress contributes to the development of type 2 diabetes: A complexity science approach, *Frontiers in Neuroendocrinology*, 65.

[17] Dendup T, Feng X, Clingan S, Astell-Burt T. (2018). Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review. *International Journal of Environmental Research and Public Health*,15(1), s.78.

[18] Anastasia C. Thanopoulou, Basil G. Karamanos, Francesco V. Angelico, Samir H. Assaad-Khalil, Alfredo F. Barbato, Maria P. Del Ben, Predrag B. Djordjevic, Vesna S. Dimitrijevic-Sreckovic, Cristina A. Gallotti, Nikolaos L. Katsilambros, Ilias N. Migdalis, Mansouria M. Mrabet, Malina K. Petkova, Demetra P. Roussi, Maria-Teresa P. Tenconi, (2003). Dietary Fat Intake as Risk Factor for the Development of Diabetes : Multinational, multicenter study of the Mediterranean Group for the Study of Diabetes (MGSD). *Diabetes Care,*26 (2), s.302–307.

[19] Weickert, M., Pfeiffer, A. (2018). Impact of Dietary Fiber Consumption on Insulin Resistance and the Prevention of Type 2 Diabetes, *The Journal of Nutrition*, 148(1), s.7–12.

[20] Koloverou, E., Esposito, K., Giugliano, D., & Panagiotakos, D. (2014). The effect of Mediterranean diet on the development of type 2 diabetes mellitus: A meta-analysis of 10 prospective studies and 136,846 participants, *Metabolism*, 63(7), s.903–911.

Journal of Epidemiology wskazuje niskie wykształcenie oraz niskie dochody jako czynniki zwiększające prawdopodobieństwo zapadnięcia na omawianą chorobę. Relacja ta została zaobserwowana w krajach o niskich, średnich oraz wysokich dochodach, jednak wymaga dalszych badań w krajach o niskich i średnich dochodach. [21]

1.7 Działania prewencyjne

Czynniki ryzyka zachorowania na cukrzycę można podzielić na możliwe do zmiany przez człowieka oraz te, na które ludzie nie mają wpływu. Czynniki niezależne od człowieka to między innymi: historia choroby w rodzinie, pochodzenie etniczne, starzenie się organizmu oraz wystąpienie cukrzycy ciążowej. Zalecenia profilaktyczne mają na celu zminimalizowanie ryzyka zachorowania przez eliminację czynników ryzyka, na które pacjent ma wpływ. Zaliczamy do nich przede wszystkim: wagę, aktywność fizyczną, ciśnienie krwi, poziom cholesterolu, dietę, palenie wyrobów tytoniowych, spożycie alkoholu, stres oraz ilość snu.[22]

International Diabetes Fedaration jako główne działania prewencyjne przeciw cukrzycy podaje podejmowanie co najmniej 30 minut aktywności fizycznej każdego dnia oraz utrzymywanie poziomu BMI poniżej 25.[23] Działania te zmniejszają także szansę rozwinięcia się nadciśnienia i podwyższonego poziomu cholesterolu, które także zostały zidentyfikowane jako czynniki sprzyjające rozwojowi choroby. Dodatkową rekomendacją dietetyczną jest także ograniczenie spożycia tłuszczy nasyconych do poziomu 7% konsumowanych dziennie kalorii. [24] British Diabetic Association dodaje do tych rekomendacji zaprzestanie konsumpcji nikotyny oraz ograniczenie spożywania alkoholu.[25] Również podwyższony poziom stresu jest czynnikiem zwiększającym ryzyko zachorowania. W przypadku wysokiej ekspozycji na niego zalecane jest wykonywanie ćwiczeń relaksacyjnych i korzystanie z innych technik ograniczających stres.

Pod koniec 2022 roku została opublikowana meta-analiza i systematyczny przegląd badań dotyczących wpływu różnego rodzaju nagród, w postaci głównie gotówki, kart podarunkowych oraz nagród rzeczowych, na minimalizowanie czynników ryzyka zachorowania na cukrzycę

---

[21] Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T., & Sidorchuk, A. (2011). Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. *International Journal of Epidemiology*, 40(3), s.804–818.
[22] American Heart Association, *Diabetes Risk Factors*, https://www.heart.org/en/health-topics/diabetes/understand-your-risk-for-diabetes (dostęp: 21.12.2022).
[23]IDF, *Prevention*. https://www.idf.org/aboutdiabetes/prevention.html, (dostęp: 21.12.2022).
[24] Steyn, N., Mann, J., Bennett, P., Temple, N., Zimmet, P., Tuomilehto, J., Louheranta, A. (2004). Diet, nutrition and the prevention of type 2 diabetes, *Public Health Nutrition*, 7(1a), s.147-165.
[25]IDF, *Diabetes risk factors,* https://www.diabetes.org.uk/preventing-type-2-diabetes/diabetes-risk-factors (dostęp: 21.12.2022).

typu II takich jak wysokie BMI oraz wysokie ciśnienie tętnicze.[26] Badanie pokazało, że grupy, które dostawały nagrody osiągały lepsze wyniki od grup kontrolnych. Osoby z grup nagradzanych traciły średnio o 1.85 kg więcej wagi i redukowały swoje BMI o 0.47 więcej oraz obniżały swoje ciśnienie skurczowe i rozkurczowe o około 2.6 mm HG więcej. Rodzaj nagrody jak i jej wartość nie wpływają znacząco na ilość utraconej wagi i poziom o jaki obniżone zostało ciśnienie.

1.8 Modelowanie ryzyka zachorowania na cukrzycę

Wczesne zidentyfikowanie cukrzycy skutkuje lepszymi rokowaniami co do jakości życia pacjenta oraz niższymi kosztami leczenia, dlatego ważna jest identyfikacja czynników ryzyka zachorowania, aby objąć odpowiednie osoby profilaktyką.

Podjęto wiele prób modelowania ryzyka zachorowania na cukrzycę, jednak aktualnie dostępne dane mają ograniczenia w postaci trudnych do zmierzenia czynników ryzyka oraz braków danych. Do ciężkich pomiarowo na dużej grupie czynników ryzyka możemy zaklasyfikować między innymi dokładną dietę oraz dzienne spożycie kalorii, jako że czynniki te nie są analizowane codziennie przez ogół społeczeństwa. Dodatkowo wyniki modelowania może zaburzyć niewiedza badanego o chorobie będącej czynnikiem ryzyka jak nadciśnienie, podwyższony poziom cholesterolu lub zespól policystycznych jajników. W większości zbiorów danych cukrzyca typu I i II nie jest rozróżniana, ale jako że cukrzyca typu II stanowi od 90% do 95% przypadków cukrzycy i rozwija się w późniejszym wieku w większości badań za przypadki cukrzycy typu II uznaje się te zdiagnozowane przed 30 rokiem życia. Omawiane w dalszej części badania zawierają głównie analizy czynników ryzyka, które nie wymagają specjalistycznych badań laboratoryjnych. Modelowanie ryzyka zachorowani na podstawie wyników badań laboratoryjnych zwykle wykazuje większą precyzję, jednak tego rodzaju badania są niemożliwe do przeprowadzenia na większej próbie pacjentów, jak i są o wiele bardziej kosztowne od badania ankietowego. Center of Disease Control opublikowało w ostatnim czasie dwa badania dotyczące modelowania ryzyka zachorowania w Preventing Chronic Disease Journal. W badaniu z 2017 roku opartym na National Health and Nutrition Examination Survey (NHANES) przeprowadzanym na próbie 5471 obywatelach skorzystano z dwóch rodzajów nieparametrycznej metody regresji statystycznej MARS oraz modelu logitowego. W przypadku modelów MARS otrzymano AUC na poziomie około 0.85,

[26] Hulbert, L. R., Michael, S. L., Charter-Harris, J., Atkins, C., Skeete, R. A., & Cannon, M. J. (2022). Effectiveness of Incentives for Improving Diabetes-Related Health Indicators in Chronic Disease Lifestyle Modification Programs: a Systematic Review and Meta-Analysis, *Preventing Chronic Disease*, 19.

a w przypadku modelu logitowego na poziomie 0.84. Definicja pojęcia AUC oraz pojęć takich jak swoistość i trafność wykorzystywanych w dalszej części rozdziału do porównywania modeli znajduje się w podrozdziale 2.5. Wyniki sugerują, że największymi czynnikami ryzka jest wiek pacjenta, a następnie historia choroby w rodzinie.[27] Z danych pochodzących z NHANES korzystano także w badaniu opublikowanym w BMC Medical Informatics and Decision Making Journal w 2019 roku, gdzie przy użyciu drzew decyzyjnego z *boostingiem* gradientowym (XGBoost) otrzymano AUC na poziomie 0.862, czyli minimalnie wyższym od poziomu ze wcześniej omawianego badania. Najważniejsze predyktory cukrzycy według tego badania to kolejno: obwód pasa, wiek, waga, długość nóg oraz ilość spożywanej soli.[28] Drugie badanie CDC z 2019 roku korzysta z metod uczenia maszynowego do przewidywania ryzyka zachorowania na cukrzycę, a następnie porównuje ich jakość pod względem statystyk wyliczonych na podstawie macierzy błędów oraz wartości AUC. Badanie jest przeprowadzone na zbiorze danych z corocznego BRFSS z 2014 roku. Rekordy z brakującymi odpowiedziami, te w których respondent jest w ciąży oraz te w których występowała cukrzyca o osoby poniżej 30 roku życia zostały usunięte z analizowanego zbioru danych. Ostatecznie badany zbiór danych składał się z 138 146 rekordów, z czego w przypadku 20 467 stwierdzono cukrzycę. Wszystkie modele miały skuteczność na podobnym, wysokim poziomie zawierającym się w przedziale 74.3% – 82.4% oraz wysokie wartość AUC na poziomie około 0.72 – 0.79. Sieć neuronowa wykazywała najwyższą trafność – 82.4%, swoistość – 90.2% i AUC – 0.79, ale jej czułość – 37.8% była najniższa spośród badanych modeli. Drzewo decyzyjne miało najwyższą czułość – 51.6%, lecz najniższą trafność – 74.3% , swoistość – 78.2% i AUC – 0.71. Badania potwierdziło wpływa dobrze znanych czynników jak BMI i wiek na zwiększone ryzyko zachorowania oraz także zidentyfikowało nowy możliwy czynnik ryzyka zachorowania jakim jest za długi sen na poziomie powyżej 9 godzin na dobę.[29] W magazynie naukowym *Fronitiers in Genetics* także zostało opublikowane porównanie różnych metod uczenia maszynowego w przewidywaniu ryzyka zachorowania na cukrzycę typu II, jednak oparte są one na zbiorach danych zawierających dane laboratoryjne takie jak poziom glukozy i insuliny. Badanie przeprowadzono na dwóch zbiorach danych – na zbiorze danych pochodzących ze szpitala z Luzzhou w Chinach oraz na zbiorze danych dotyczących kobiet z indiańskiego plemienia

[27] Turi, K. N., Buchner, D. M., & Grigsby-Toussaint, D. S. (2017). Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors*, Preventing Chronic Disease,* 14.
[28] Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning, *BMC Medical Informatics and Decision Making*, 19(1).
[29] Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques, *Preventing Chronic Disease*, 16.

Pima. Analiza wykazała podobną skuteczność drzewa decyzyjnego, sieci neuronowej oraz lasu losowego, z minimalną przewagą lasu losowego nad pozostałymi modelami.[30] Podobne wyniki uzyskali w swoim badaniu Dutta, Paul i Ghosh, gdzie na podstawie tych samych danych z plemienia Indian Pima porównywali ze sobą regresję logitową, metodę wektorów nośnych (SVM) oraz las losowy. W tym badaniu także najwyższą skuteczność miał las losowy.[31] Nie znaleziono idealnego modelu do modelowania ryzyka zachorowania na cukrzycę, dlatego ważne jest zastosowanie więcej niż jednego modelu i następnie porównanie wyników. Mimo to, we wszystkich wspomnianych badaniach AUC był na poziomie powyżej 0.7 co klasyfikuje modele jako co najmniej dobre pod względem tego kryterium. Co więcej, nawet w przypadku modeli bez danych laboratoryjnych modele osiągały takie wyniki.

[30] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques, *Frontiers in Genetics*, 9.
[31] Dutta, D., Paul, D., & Ghosh, P. (2018). Analysing Feature Importances for Diabetes Prediction using Machine Learning. 2018 IEEE 9th Annual Information Technology, *Electronics and Mobile Communication Conference (IEMCON).*

Rozdział II

Opis danych źródłowych i metod estymacji modelu

2.1 Źródło danych

Analizowane dane pochodzą z The Behavioral Risk Factor Surveillance System (BRFSS) z 2021 roku. Jest to ankieta telefoniczna wspierana komputerowo (CATI) przeprowadzana corocznie od 1984 roku na próbie około 400 000 dorosłych respondentów przez amerykańską agencję rządową - Centers for Disease Control and Prevention we współpracy z władzami wszystkich stanów członkowskich oraz amerykańskich terytoriów zależnych.[32] Respondenci są losowo wybierani na podstawie numerów telefonicznych zarejestrowanych w analizowanym regonie. Celem ankiety jest zebranie danych na temat podejmowanych przez obywateli Stanów Zjednoczonych praktyk zdrowotnych oraz ryzykownych zachowań prowadzących do chorób przewlekłych i innych możliwych do uniknięcia schorzeń. Ankieta składa się z pytań głównych zadawanych wszystkim respondentom oraz pytań opcjonalnych dotyczących w większości konkretnego schorzenia badanego i prowadzonej związanej z nim profilaktyki. Na każde pytanie ankietowany może odmówić odpowiedzi lub udzielić odpowiedzi „nie wiem". Odpowiedzi „nie wiem" odpowiada liczba 7, a odmowie odpowiedzi odpowiada liczba 9, jeśli zmienna nie przyjmuje wartości powyżej 6. Jeśli zmienna przyjmuje wartości większe od 9, lecz mniejsze od 77, to odpowiedzi „nie wiem" odpowiada liczba 77, a odmowie odpowiedzi liczba 99. Analogiczna notacja zachodzi dla wszystkich większych wartości. Zbiór danych stworzony na podstawie ankiety dostępny jest na stronie CDC w dwóch formatach: XPT oraz ASCII. Na potrzebę analiz zdecydowałam się na wykorzystanie danych w formacie XPT i odczytanie ich w programie RStudio z wykorzystaniem biblioteki *haven*. W ten sposób otrzymałam ramkę danych zawierającą 438 693 wierszy i 304 kolumny. Część kolumn zawiera odpowiedzi na zadawane w trakcie wywiadu pytania, a część to zmienne pochodne stworzone na podstawie odpowiedzi na pytania.

2.2 Zmienna objaśniana

Zmienna objaśniana *diabetes* w oryginalnym zbiorze danych przyjmuje sześć wartości i jest jedną z możliwych odpowiedzi na pytanie dotyczące posiadania cukrzycy. W przypadku tej zmiennej w całym zbiorze występują 3 braki danych. Cukrzyca została stwierdzona u 13% ankietowanych, co oznacza, że analizowany zbiór nie jest zbiorem zbilansowanym i mniejszość

[32] CDC, *2021 BRFSS Survey Data and Documentation,* https://www.cdc.gov/brfss/annual_data/annual_2021.html (dostęp: 05.02.2023).

obserwacji należy do klasy pozytywnej. W Tabeli II.1 zostały przedstawione wartości przyjmowane przez zmienną objaśnianą.

Tabela II.1 Wartości przyjmowane przez zmienną objaśnianą i ich liczebność

Wartość	**Interpretacja**	**Liczebność**
1	u badanego została stwierdzona cukrzyca poza ciążą	57 616
2	u badanego została stwierdzona cukrzyca jedynie w trakcie trwania ciąży	3 808
3	u badanego nie została stwierdzona cukrzyca	366 342
4	u badanego został stwierdzony stan przedcukrzycowy	9 942
7	badany udzielił odpowiedzi „nie wiem”	613
9	badany odmówił odpowiedzi	369

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Wartości 2 i 3 zostały zastąpione zerami, co oznacza brak cukrzycy i przynależność do klasy negatywnej. Rekordy, dla których wartość zmiennej *diabetes* wynosiła 4 zostały usunięte ze zbioru danych, jako że zmienna informuje nas o bardzo wysokim ryzyku rozwinięcia się cukrzycy u badanego w najbliższym czasie, jednak tego nie gwarantuje. Uwzględnienie tej zmiennej w klasie pozytywnej lub negatywnej w zbiorze uczącym może negatywnie wpłynąć na jakość modelu, jako że osoby te mogą posiadać dużo charakterystyk podobnych do osób chorych jak i zdrowych. W zbiorze danych cukrzyca typu I i II nie jest rozróżniana, więc skoro cukrzyca typu I stanowi poniżej 5% przypadków, wszystkie przypadki cukrzycy uznawane są za przypadki cukrzycy typu I.

2.3 Zmienne objaśniające

Spośród 304 potencjalnych zmiennych objaśniających zostały wybrane 24, których potencjalny związek z ryzykiem zachorowania na cukrzycę został w większości wskazany w analizowanej literaturze. W zbiorze znajdują się zmienne odpowiadające charakterystykom demograficznym badanego, jego stanowi zdrowia, stylowi życia oraz sytuacji socjoekonomicznej. Zmienne dotyczące stanu zdrowia o przede wszystkim przewlekłe schorzenia pacjenta, które obciążają organizm czyniąc go bardziej podatnym na inne choroby. Zmienne dotyczące stylu życia pacjenta uwzględniają jego praktyki żywieniowe oraz podejście do aktywności fizycznej. Zmienne socjoekonomiczne dotyczą dochodów badanego, jego edukacji oraz dostępu do opieki zdrowotnej, jako że wczesne wykrycie choroby minimalizuje jej negatywne skutki. Zmienne demograficzne związane są z pochodzeniem etnicznym, płcią oraz miejscem zamieszkania, jako że wspomniane badania wskazują na relację między środowiskiem a ryzkiem zachorowania. Przy doborze zmiennych wzięta została także pod uwagę liczba ich kategorii. Wynika to z faktu, że niektóre klasyfikatory, a w szczególności las losowy, preferują zmienne przyjmujące większą liczbę wartości.[33] Wszystkie zmienne objaśniane zawierające się w ostatecznym zbiorze testowym są zmiennym kategorycznymi, gdzie liczba kategorii jest w przedziale od 2 do maksymalnie 13. Dodatkowym czynnikiem, który wpłynął na dobór zmiennych jest udział braków danych w odpowiedziach na poszczególne pytania. Wybrane zostały odpowiedzi na pytania, na które odpowiedź udzieliła ponad połowa badanych respondentów. Na dwóch zmiennych zostały przeprowadzone znaczące transformacje. Zmienna *fries* w oryginalnym zbiorze danych opisywała dzienna ilość spożywanych porcji frytek i była zmienną ciągłą. Została przetransformowana w zmienną binarną, dla której punktem odcięcia była wartość 25, równoważna spożywaniu 0.25 porcji frytek dziennie, co przekłada się na około 2 porcje frytek tygodniowo, i która to ilość potencjalnie zwiększa ryzyko zachorowania na cukrzycę.[34] Drugą zmienna poddaną transformacji jest zmienna *physhealth.* W początkowej formie jest to zmienna opisująca ilość dni w miesiącu podczas których badany ocenia stan swojego zdrowia fizycznego jako zły. Po transformacji zmienna informuje czy podczas jakiegokolwiek dnia w ciągu ostatniego miesiąca ankietowany ocenił stan swojego zdrowia jako zły. Zmienne z oryginalnego zbioru przyjmujące wartości 1 lub 2, nie uwzględniając odpowiedzi „nie wiem” lub odmowy odpowiedzi na

[33] Strobl C., Boulesteix A. L., Zeileis A., & Hothorn T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics*, 8(1).
[34] Shmerling, R. H., MD. (2019). *In defense of French fries*, Harvard Health, https://www.health.harvard.edu/blog/in-defense-of-french-fries-2019020615893 (dostęp: 03.02.2022).

pytanie, zostały przetransformowane na zmienne binarne. Nazwy oraz wartości przyjmowane przez zmienne objaśniające znajdują się w Tabeli II.2.

Tabela II.2 Nazwy zmiennych objaśniających i przyjmowane przez nie wartości

Nazwa zmiennej	**Przyjmowane wartości**
highBP	0 – brak nadciśnienia 1 – obecność nadciśnienia
highchol	0 – poziom cholesterolu w normie 1 – podwyższony poziom cholesterolu
bmi	1 – niedowaga 2 – waga w normie 3 – nadwaga 4 – otyłość
smoker	1 – badany pali aktualnie papierosy codziennie 2 – badany pali, ale nie codziennie 3 – badany palił w przeszłości 4 – badany nigdy nie palił
stroke	1 – badany pali aktualnie papierosy codziennie 2 – badany pali, ale nie codziennie 3 – badany palił w przeszłości 4 – badany nigdy nie palił
heartdis	0 – brak choroby wieńcowej serca oraz zawału 1 – występowanie choroby wieńcowej serca lub przejście zawału
physact	0 – brak aktywności fizycznej niezwiązanej z pracą w ciągu ostatnich 30 dni 1 – aktywność fizyczna niezwiązana z pracą w ciągu ostatnich 30 dni
fruit	0 – brak spożycia owoców co najmniej raz dziennie 1 – spożywanie owoców co najmniej raz dziennie
veggie	0 – brak spożycia warzyw co najmniej raz dziennie 1 – spożywanie warzyw co najmniej raz dziennie
alco	0 – spożycie poniżej 14 drinków tygodniowo w przypadku mężczyzn i poniżej 7 drinków w przypadku kobiet 1 – spożycie powyżej 14 drinków tygodniowo w przypadku mężczyzn

	i powyżej 7 drinków w przypadku kobiet Drink definiowany jest jako równowartość 30 ml mocnego alkoholu.
healthcare	0 – brak ubezpieczenia zdrowotnego 1 – posiadanie ubezpieczenia zdrowotnego
medcost	0 – badany w ciągu ostatnich 30 dni nie zrezygnował z wizyty u lekarza z powodu jej kosztu 1 - badany w ciągu ostatnich 30 dni zrezygnował z wizyty u lekarza z powodu jej kosztu
genhealth	1 – bardzo dobry ogólny stan zdrowia 2 – dobry ogólny stan zdrowia 3 – średni ogólny stan zdrowia 4 – zły ogólny stan zdrowia 5 – bardzo zły ogólny stan zdrowia Ocena ogólnego stanu zdrowia jest subiektywną oceną badanego.
physhealth	0 – badany w ciągu ostatniego miesiąca nie doświadczył złego stanu zdrowia fizycznego 1 – badany w ciągu ostatniego miesiąca ocenił swój stan zdrowia fizycznego jako zły przez co najmniej jeden dzień
diffwalk	0 – brak trudności z chodzeniem lub wchodzeniem po schodach 1 – trudności z chodzeniem lub wchodzeniem po schodach
sex	0 – kobieta 1 – mężczyzna
age	1 – 18-24 lata 2 – 25–29 lat 3 – 30–34 lata 4 – 35–39 lat 5 – 40–44 lata 6 – 45-49 lat 7 – 50-54 lata 8 – 55-59 lat 9 – 60-64 lata 10 – 65-69 lat 11 – 70-74 lata

	12 – 75-79 lat 13 – 80 i więcej lat
education	1 – brak edukacji 2 – ukończone 8 lat edukacji 3 – ukończone 9-11 lat edukacji 4 – ukończone 12 lat edukacji 5 – ukończone 13-15 lat edukacji 6 – ukończone studia wyższe
income	1 – roczny dochód gospodarstwa domowego poniżej 10 000$ 2 – roczny dochód gospodarstwa domowego w przedziale 10 000 – 15 000$ 3 – roczny dochód gospodarstwa domowego w przedziale 15 000 – 20 000$ 4 – roczny dochód gospodarstwa domowego w przedziale 20 000 – 25 000$ 5 – roczny dochód gospodarstwa domowego w przedziale 25 000 – 35 000$ 6 – roczny dochód gospodarstwa domowego w przedziale 35 000 – 50 000$ 7 – roczny dochód gospodarstwa domowego w przedziale 50 000 – 75 000$ 8 – roczny dochód gospodarstwa domowego w przedziale 75 000 – 100 000$ 9 – roczny dochód gospodarstwa domowego w przedziale 100 000 – 150 000$ 10 – roczny dochód gospodarstwa domowego w przedziale 150 000 – 200 000$ 11 – roczny dochód gospodarstwa domowego powyżej 200 000$
fries	0 – spożycie frytek poniżej 2 porcji tygodniowo 1 – spożycie frytek powyżej 2 porcji tygodniowo
depression	0 – niestwierdzona depresja lub pochodna choroba psychiczna 1 – stwierdzona depresja lub pochodna choroba psychiczna

asthma	0 – brak astmy 1 – stwierdzona astma
urban	0 – zamieszkanie poza miastem 1 – zamieszkanie w mieście
race	*white* – rasa biała *black* – rasa czarna *asian* – osoba pochodzenia azjatyckiego *native* – osoba bedąca potomkiem natywnych mieszkańców Stanów Zjednoczonych *hispanic* – osoba o pochodzeniu latynoskim *other* – osoba nie identyfikuje się z żadną z powyższych grup etnicznych

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

2.4 Metody estymacji modeli

Zmienna objaśniana w modelu jest zmienną jakościową przyjmującą dwie wartości z przedziału od 0 do 1, gdzie 0 oznacza brak choroby, a 1 jej występowanie. Zmienne objaśniające w modelu są zmiennymi kategorycznymi. W celu zidentyfikowania czynników ryzyka i prawdopodobieństwa zapadnięcia na cukrzycę zbudowane zostaną 4 modele: model logitowy, model logitowy z wykorzystaniem metody *stepwise*, drzewo klasyfikacyjne oraz las losowy. W przypadku modelu logitowego do estymacji posłuży funkcja *glm()* z biblioteki *stats*. Następnie model ten zostanie dopasowany z użyciem metody *stepwise* przy pomocy funkcji *stepAIC()* z biblioteki *MASS.* Drzewo decyzyjne będzie zbudowane z wykorzystaniem biblioteki *caret* z i użyciem funkcji *train()* z wyborem metody *rpart.* Las losowy zbudowany zostanie z użyciem pakietu *randomForest.*i funkcji *randomForest.*

2.4.1 Regresja logistyczna

Regresja logistyczna jest szczególnym przypadkiem modelu regresji liniowej, w którym zmienna objaśniana jest zmienną dychotomiczną, czyli przyjmuje tylko dwie wartości. Model zwraca wartość prawdopodobieństwa przyjęcia przez zmienna objaśnianą Y wartości 1, w zależności od wartości zmiennych objaśniających X_i. Zmienne objaśniające mogą być

zmiennymi jakościowymi lub ilościowymi. Równanie modelu[35] możemy przedstawić następująco:

$$p = \frac{\exp\left(\beta_0 + \sum_{i=1}^{k} \beta_i X_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{k} \beta_i X_i\right)}$$

Gdzie:

$$p - prawdodpodobieństwo\ zdarzenia\ Y = 1$$
$$\beta_i - parametr\ strukturalny\ w\ modelu$$
$$X_i - wartość\ zmiennej\ objaśnianej\ w\ modelu$$

Parametry strukturalne informują nas o tym jak zmieni się iloraz szans po zwiększeniu X_i o jednostkę. Jako iloraz szans definiujemy stosunek prawdopodobieństwa przyjęcia przez zmienną objaśnianą wartości 1 do prawdopodobieństwa przyjęcia wartości 0. Iloraz szans dla zmiennej jest równy e^{β_i} . Najczęściej parametry są estymowane z użyciem Metody Największej Wiarygodności. Podstawowe założenia jakie muszą spełniać dane kategoryczne używane do budowania modelu logitowego to odpowiednio duży rozmiar próby, brak silnych korelacji między zmiennymi objaśniającymi, brak wartości odstających oraz niezależność kolejnych obserwacji.

2.4.2 Algorytm *stepwise selection*

Algorytm *stepwise selection*[36] jest algorytmem automatycznego doboru zmiennych do modelu. Do selekcji zmiennych przy dopasowaniu modelu logitowego został wybrany dwukierunkowy algorytm *stepwise* korzystający naprzemiennie ze *stepwise forward selection* oraz *stepwise backward selection*. Algorytm *stepwise forward selection* rozpoczyna budowę od pustego modelu i następnie dobiera zmienne po kolei według kryterium najwyższej istotności statystycznej do momentu gdy nie jest obserwowana wyraźna poprawa jakości modelu. Do oceny jakości modelu stosowana jest najczęściej statystyka R – kwadrat, wskaźnik AIC, BIC lub ocena ogólnej istotności statystycznej całego modelu. *Stepwise backward selection* polega na usuwaniu kolejno zmiennych objaśnianych i sprawdzaniu jakości modelu po każdej takiej operacji. Jeśli jakość modelu polepszyła się po usunięciu danej zmiennej objaśniającej, to badana zmienna wykluczana jest z modelu. Proceder kończy się w sytuacji, w której usunięcie kolejnych zmiennych objaśniających nie przynosi znaczącej poprawy kryteriów oceny modelu

[35] Jackowska, B., (2011). Efekty interakcji między zmiennymi objaśniającymi w modelu logitowym w analizie zróżnicowania ryzyka zgonu, *Przegląd Statystyczny,* 58(1-2), s.24-41.
[36] Choueiry, G., *Understand Forward and Backward Stepwise Regression*, https://quantifyinghealth.com/stepwise-selection/ (dostęp: 10.01.2023)

analogicznych jak w przypadku *stepwise forward selection.* Dwukierunkowy algorytm *stepwise* korzysta z obu tych metod w celu uzyskania najlepszego dopasowania modelu.

2.4.3 Drzewo decyzyjne

W teorii grafów drzewo definiowane jest jako spójny i acykliczny graf. Drzewo klasyfikacyjne jest szczególnym przypadkiem drzewa, a w którym węzły, liście oraz gałęzie są interpretowalne. Węzły drzewa odpowiadają przeprowadzanym testom, gałęzie odpowiadają możliwym wynikom testów. Liście informują o przypisaniu obserwacji do konkretnej klasy według kryterium przypisanego do łączącej go z węzłem gałęzi.[37] Algorytm budowy drzewa dzieli wyjściowy zbiór na podzbiory na podstawie wyników testów tworząc kolejne węzły, aż do momentu spełnienia kryterium stopu. W momencie, kiedy kryterium stopu jest spełnione tworzony jest liść. Celem podziału jest uzyskanie jak najbardziej jednorodnych, co do klasy zmiennej objaśnianej podzbiorów. Do kryteriów stopu budowy drzewa należą między innymi: pusty zbiór testów lub obserwacji oraz jednorodność obserwacji w podzbiorze przekraczająca ustalone kryterium. Testy na węzłach dobierane są w sposób, który maksymalizuje zróżnicowania powstałych po ich wykonaniu podzbiorów, czyli ich entropii. Do zalet drzew decyzyjnych można zaliczyć interpretowalność wyników oraz wysoką dokładność. Ich wadą jest częste zbyt mocne dopasowanie drzewa do zbioru uczącego.[38] Z tego powodu przy budowie drzewa została wykorzystana 5-krotna walidacja krzyżowa. Walidacja ta polega na podziale zbioru danych wykorzystywanego do budowy drzew, w tym przypadku na 5 warstw. Podczas trenowania modelu jedna warstwa jest wykorzystywana jako zbiór testowy, a pozostałe jako zbiór treningowy. Operacja ta jest wykonywana w tym wypadku 5 razy, do momentu, kiedy każda warstwa będzie wykorzystana jako zbiór testowy.[39]

2.4.4 Las losowy

Las losowy jest jedną z metod zespołowego uczenia maszynowego. Algorytm polega na budowaniu wielu drzew decyzyjnych na próbach bootrsapowych, dzięki czemu minimalizowany jest problem nadmiernego dopasowania modelu do zbioru uczącego. Próby bootstrapowe są próbami o mniejszym rozmiarze losowanymi z głównego zbioru ze zwracaniem.[40] Największą wadą lasu losowego jest brak łatwo dostępnej informacji na temat

---

[37] Gromada, M. (2006). *Drzewa klasyfikacyjne, ich budowa, problemy złożoności i skalowalności*

[38] Jadhev, S. D., & Channe, H. (2016). P.Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques*, International Journal of Science and Research*, 5(1), s.1842-1845.

[39] Blockeel, H., & Struyf, J. (2003). Efficient algorithms for decision tree cross-validation, *Journal of Machine Learning Research*, 3, s.621–650.

[40] Ranganathan, S., Nakai, K., & Schonbach, C. (2018*). Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics.* Wydawinictwo Elsevier.

kierunku relacji między zmienną objaśnianą a zmienną objaśniającą. Znana jest jedynie siła tej relacji. Nie jesteśmy też w stanie wymienić reguł decyzyjnych decydujących o klasyfikacji zmiennej, jako że las losowy agreguje wyniki wielu pojedynczych drzew klasyfikacyjnych.

2.5 Badanie jakości modeli

W celu możliwości walidacji jakości modelu zbiór wyjściowy został podzielony na zbiór uczący i testowy w proporcjach odpowiednio 30% i 70%. Aby porównać ze sobą omawiane modele wykorzystane zostaną wskaźniki pomiaru jakości aplikowalne w przypadku wszystkich czterech modeli. Początkowo do wyliczenia podstawowych statystyk posłuży macierz błędów porównująca wartości prognozowane z wartościami rzeczywistymi ze zbioru testowego. Do statystyk podstawowych zaliczamy:

- TP (True Positive) - liczba obserwacji poprawnie zaklasyfikowanych do klasy pozytywnej,
- FP (False Positive) - liczba obserwacji błędnie zaklasyfikowanych do klasy pozytywnej,
- TN (True Negative) - liczba obserwacji poprawnie zaklasyfikowanych do klasy negatywnej,
- FN (False Negative) - liczba obserwacji błędnie zaklasyfikowanych do klasy negatywnej.

Na ich podstawie wyznaczone zostaną statystki pochodne takie jak: dokładność, specyficzność, precyzja i czułość. Do wyznaczenia ich posłuży funkcja *confusionMatrix()* z biblioteki *caret*. Do najważniejszych statystyk pochodnych możemy zaliczyć:

$$czułość = TPR = \frac{TP}{TP + FN}$$

$$specyficzność = TNR = \frac{TN}{TN + FP}$$

$$dokładność = ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precyzja = PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{FN + TN}$$

$$FPR = 1 - TPR$$

Zbudowana zostanie także krzywa ROC, obrazująca zależność między TPR oraz FPR. Na jej podstawie zostanie wyliczona wartość AUC, informująca o polu pod powierzchnią krzywej

ROC.[41] Do narysowanie krzywej ROC oraz wyliczenia AUC posłuży biblioteka *ROCR*. Za zmienne nieistotne statystycznie w modelu uznane będą zmienne dla których wartość p przyjmuje wartości powyżej 0,05.

[41] Hossin, M., & Sulaiman, M. R. (2015). A Review on Evaluation Metrics for Data Classification Evaluations, *International Journal of Data Mining & Knowledge Management Process*, 5(2), s.1–11.

Rozdział III

Budowa modelu

3.1 Przygotowanie i eksploracja danych

3.1.1 Zbiór pierwszy

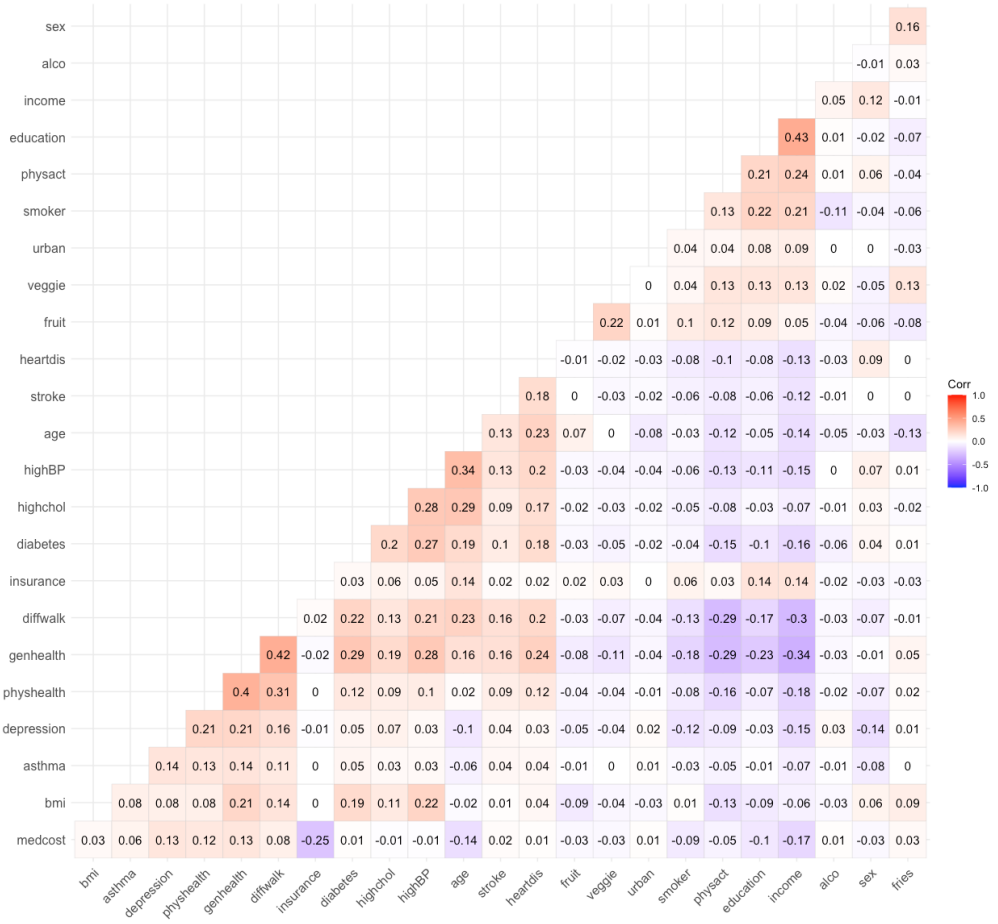
Zbiór pierwszy powstał wskutek usunięcia z wyjściowego zbioru wszystkich wierszy zawierających braki danych, jak i wierszy, w których badany odmówił odpowiedzi na pytanie lub odpowiedział na pytanie „nie wiem". W ten sposób w zbiorze pozostało 227 099 rekordów. Do klasy pozytywnej należy 32 876 obserwacji co stanowi 14% całego zbioru.

Rysunek III.1 Rozkład zmiennej objaśnianej w zbiorze pierwszym

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Między zmiennymi objaśniającymi nie występują silne korelacje. Z macierzy korelacji przedstawionej na Rysunku III. 2 możemy wywnioskować, że najwyższa pozytywna korelacja występuje między zmienna objaśnianą a generalnym stanem zdrowia, nadciśnieniem, trudnością w poruszaniu się i wysokim cholesterolem. Największa ujemna korelacja występuje między występowaniem cukrzycy a dochodami i aktywnością fizyczną. Z perspektywy socjo-ekonomicznej warto zauważyć, że dochody są ujemnie skorelowane z większością problemów zdrowotnych i dodatnio skorelowane z poziomem edukacji.

Rysunek III. 2 Macierz korelacji zmiennych



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Większość zmiennych objaśniających nie ma zbalansowanego rozkładu. Szczególnie widać to przy zmiennych *medcost, alco* oraz *insurance*, gdzie udział klasy pozytywnej jest na poziomie niższym niż 10%, co może wpłynąć negatywnie na określenie kierunku zależności między tymi zmiennymi a zmienną objaśnianą. Wykresy z rozkładami zmiennych objaśniających we wszystkich zbiorach znajdują się Załączniku 3.

3.1.2 Zbiór drugi

Zmienna objaśniana nie jest zbalansowana, więc zbiór pierwszy został zbalansowany z użyciem biblioteki *ROSE*. Na danych został przeprowadzony losowy *oversampling,* czyli duplikacja obserwacji z klasy pozytywnej oraz *undersampling,* czyli usunięcie obserwacji należących do klasy negatywnej, tak aby sumaryczna liczba obserwacji pozostała niezmieniona w porównaniu do wyjściowego zbioru. W rezultacie otrzymaliśmy rozkład zmiennej objaśnianej, gdzie klasy są prawie równoliczne. Rozkłady zmiennych objaśniających nie uległy istotnym zmianom. Różnica pojawiła się jedynie w proporcji klas zmiennej objaśnianej

w poszczególnych obserwacjach. W zbiorze także nadal nie występują silne korelacje mogące negatywnie wpłynąć na model logitowy.

3.1.3 Zbiór trzeci

Trzeci zbiór danych powstał na podstawie danych źródłowych. W pierwszym kroku wszystkie odmowy odpowiedzi i odpowiedzi „nie wiem" zostały zastąpione brakami danych. Następnie na tym zbiorze zostały przeprowadzone dwie iteracje z wykorzystaniem biblioteki *mice* i metodą *defaultMethod.* Metoda ta dopasowuje metodę prognozowania brakującej wartości do typu zmiennej. Jako, że wszystkie dane były kategoryczne braki danych zostały uzupełnione jedną z dwóch metod – regresją logistyczną w przypadku zmiennych binarnych oraz wielomianowa analizą logitową w przypadku zmiennych z liczbą kategorii wyższą niż dwie. Ostateczny zbiór zawiera odpowiedzi pochodzące od 438 693 osób. Proporcje zmiennej objaśnianej nie uległy znaczącym zmianom – 15% obserwacji należy do klasy pozytywnej, a 85% do klasy negatywnej. Znaczącej zmianie nie uległy także rozkłady zmiennych objaśnianych oraz korelacje między nimi.

3.1.4 Zbiór czwarty

Aby zbalansować klasy zmiennej objaśnianej zbiór po wypełnieniu braków danych z użyciem pakietu *mice* został analogicznie jak w przypadku zbioru drugiego zbalansowany z wykorzystaniem biblioteki *ROSE.* Liczba obserwacji nie uległa zmianie i pozostała na poziomie wyjściowym 438 693 obserwacji. Tak jak w przypadku zbioru drugiego rozkład zmiennych objaśniających nie uległ znaczącym zmianom. Zmianie uległ jedynie rozkład klas zmiennej objaśnianej w poszczególnych klasach zmiennych objaśnianych.

3.2 Interpretacja oraz ocena wyników modelowania na pierwszym zbiorze

3.2.1 Model logitowy

Model logitowy, tak jak wszystkie kolejne modele został zbudowany na zbiorze uczącym stanowiącym 25% całego zbioru. Pod Tabelą III.1 znajduje się sposób oznaczenia istotności zmiennych. Jak widać w Tabeli III.1 przy przyjęciu p=0.05, jako punktu odcięcia wszystkie zmienne oprócz: *racenative, medcost, asthma, urban, racehispanic, fruit, veggie, raceblack, physhealth* oraz *education* są statystycznie istotne. Największy wpływ na zwiększenie prawdopodobieństwa zaklasyfikowania obserwacji do klasy pozytywnej mają zmienne *highBP, highchol, bmi* oraz *genhealth.* Zwiększenie wartości tych zmiennych o jednostek powoduje wzrost ilorazu szans o ponad 60%. Wśród zmiennych istotnych statystycznie największy spadek prawdopodobieństwa zaklasyfikowania zmiennej do klasy pozytywnej po zwiększeniu wartości zmiennej o jednostkę występuje w przypadku zmiennych

racewhite, raceother, physact oraz *alco.* Oznacza to, że największymi czynnikami ryzyka zachorowania na cukrzycę według modelu jest nadciśnienie, podwyższonych poziom cholesterolu, słaba ocena generalnego stanu zdrowia oraz podwyższony wskaźnik BMI. Główną charakterystyką obniżającą ryzyko zachorowania jest identyfikowanie się z rasą białą lub inną, niż wymienione w kwestionariuszu. Bardzo wysoki wynik w tej kategorii ma zmienna *alco* informująca o nadużywaniu alkoholu. Badania wskazują relacje o przeciwnym z kierunku, z tego też powodu wpływ na taki wynik mógł mieć niski udział klasy pozytywnej (poniżej 10%) w zmiennej wykorzystanej w modelu. Do istotnych czynników obniżających ryzyko zachorowania na cukrzycę możemy zaliczyć także aktywność fizyczną i wysokie dochody.

Tabela III.1 Parametry modelu logitowego zbudowanego na zbiorze pierwszym

Zmienna	Parametr strukturalny	Błąd standardowy	p-value	Istotność
highBP	0.707074	0.030157	< 2e-16	***
highchol	0.603303	0.027912	< 2e-16	***
bmi	0.542846	0.018652	< 2e-16	***
genhealth	0.513145	0.016576	< 2e-16	***
heartdis	0.336345	0.038397	< 2e-17	***
insurance	0.252891	0.085953	0.003259	**
sex	0.190364	0.028350	1.88e-11	***
diffwalk	0.142790	0.035413	5.53e-05	***
age	0.132618	0.005652	< 2e-18	***
stroke	0.12550	0.054239	0.020677	*
racenative	0.105661	0.132654	0.425731	
depression	0.089166	0.033851	0.008437	**
fries	0.076820	0.029820	0.009992	**
medcost	0.051035	0.056703	0.368098	
asthma	0.038569	0.037133	0.298965	
smoker	0.038240	0.015278	0.012317	*
urban	0.025450	0.037222	0.494148	
alco	-0.615005	0.071163	< 2e-16	***
racehispanic	-0.183805	0.111367	0.098851	
physact	-0.151200	0.030962	1.04e-06	***
fruit	-0.016963	0.028180	0.547220	
veggie	-0.011510	0.035531	0.745980	
racewhite	-0.627039	0.099816	3.34e-10	***
raceother	-0.463888	0.124940	0.000205	***
raceblack	-0.190435	0.107696	0.077017	
physhealth	-0.05904	0.03122	0.058658	
income	-0.044247	0.00681	8.37e-11	***
education	-0.006392	0.015440	0.678886	

‘***’ p= 0.001 ‘**’ p = 0.01 ‘*’ p = 0.05

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.2.2 Model logitowy z wykorzystaniem algorytmu *stepwise*

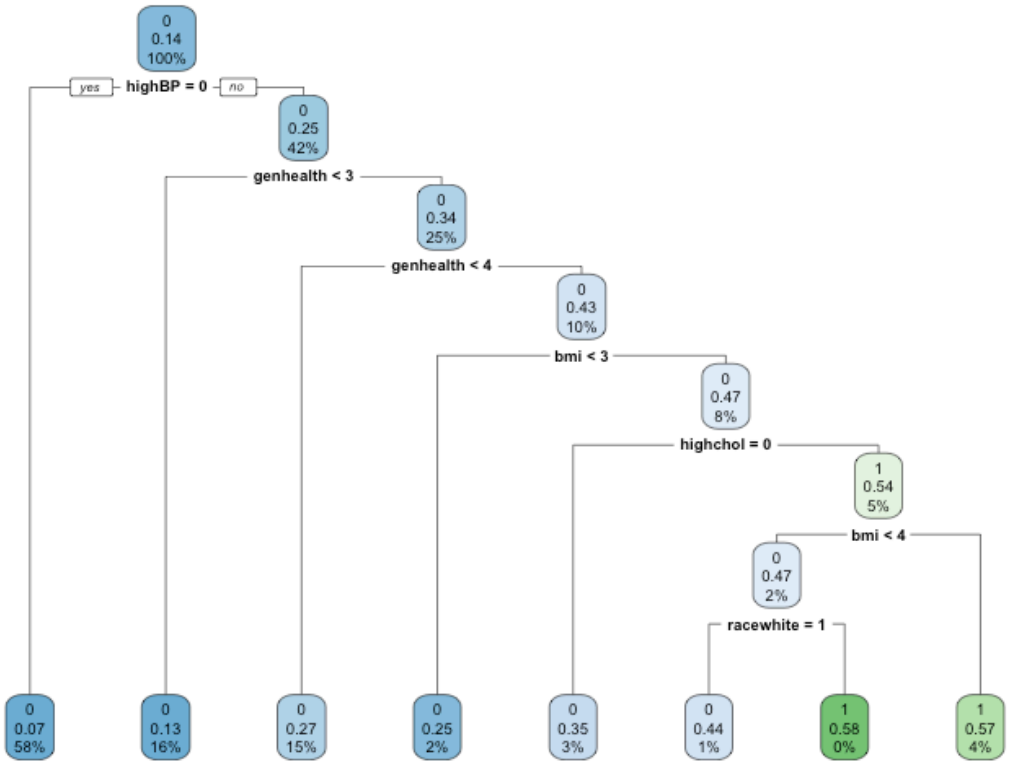
Po wykonaniu dwukierunkowego algorytm *stepwise* na modelu logitowym, ze zbioru zmiennych objaśnianych zostało usunięte sześć zmiennych. Z modelu usunięte zostały zmienne *medcost, asthma, urban, fruit, veggie* oraz *education.* Oznacza to, że wykluczenie większej ilości zmiennych nie zwiększyłoby wartości AIC, która jest wskaźnikiem dopasowania modelu. W wartościach parametrów modelu nie zaszły znaczące zmiany w stosunku do wartości z

Tabeli III.1, zmieniające wartość parametru o więcej niż 0.05. Istotność zmiennych także nie uległa zmianom.

3.2.3 Drzewo decyzyjne

Powstałe drzewo decyzyjne przedstawione na Rysunku III.3 generuje 7 reguł decyzyjnych i jego głębokość wynosi 6. W pierwszej regule sprawdzany jest brak nadciśnienia u pacjenta. Jeśli pracjent nie ma nadciśnienia zaliczany jest do klasy negatywnej, w przeciwnym wypadku przeprowadzany jest kolejny test. Druga reguła decyzyjna sprawdza generalną ocene stanu zdrowia. Jeśli stan zdrowia został oceniony powyżej średniej obserwacja klasyfikowana jest do klasy negatywnej, w przeciwnym przypadku przeprowadzamy kolejny test. Trzeci test także sprawdza ogólny stan zdrowia i klasyfikuje obserwacje do klasy negatywnej jeśli stan zdrowia został oceniony jako „bardzo dobry", „dobry" lub „przeciętny". W przeciwnym wypadku przeprowadzany jest kolejny test sprawdzający BMI. Jeśli pacjent ma niedowagę lub wagę w normie obserwacja klasyfikowana jest do klasy negatywnej. W przypadku BMI ponad normę przeprowadzany jest kolejny test sprawdzający klasę zmiennej informującej o wysokim poziomie cholesterolu. Jeśli nie występuje podwyższony poziom cholesterolu obserwacja klasyfikowana jest do klasy negatywnej, w przeciwnym wypadku przeprowadzany jest kolejny test, w którym ponownie testowana jest zmienna BMI. Jeśli wartość BMI jest równa 5, co oznacza otyłość, obserwacja klasyfikowana jest do klasy pozytywnej. W przeciwnym wypadku przeprowadzany jest ostatni test. Jeśli ankietowany deklaruje przynależność do rasy białej, to obserwacja klasyfikowana jest do klasy negatywnej. W przeciwnym wypadku obserwacja klasyfikowana jest do klasy negatywnej. Na podstawie tego drzewa decyzyjnego możemy wywnioskować, że na klasyfikacji obserwacji do klasy pozytywnej najbardziej sprzyja nadciśnienie, nadwaga, podwyższony cholesterol, zły stan zdrowia oraz nadwaga.

Rysunek III.3 Drzewo decyzyjne zbudowane na zbiorze pierwszym

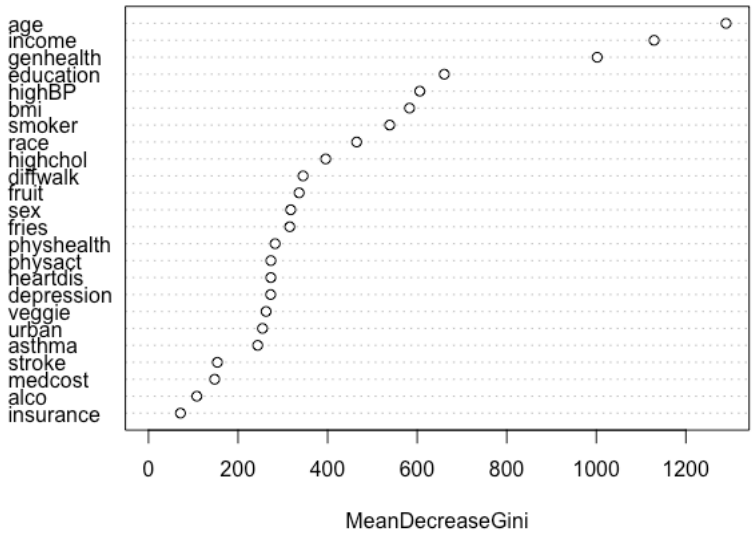


Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.2.4 Las losowy

Las losowy według Rysunku III.4, jako najbardziej istotne dla klasyfikacji zmienne wskazał wiek, dochód, generalny stan zdrowia oraz edukację. Lasy losowe faworyzują zmienne o większej ilości klas, dlatego warto zwrócić uwagę na najwyżej sklasyfikowane zmienne binarne którymi jest obecność nadciśnienia, wysoki cholesterol, trudność z poruszaniem się oraz spożycie owoców. Wysoko sklasyfikowane zostały dwie zmienne informujące o spożyciu owoców oraz warzyw, które były wskazane jako nieistotne w modelu logitowym, co wraz z wysoką klasyfikacją zmiennej *fries* może potwierdzać istotność diety w zapobieganiu cukrzycy.

Rysunek III.4 Istotność zmiennych w lesie losowym zbudowanym na zbiorze pierwszym



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.2.5 Ewaluacja jakości modeli

Wszystkie modele mają bardzo zbliżone wyniki pod względem analizy statystyk z macierzy błędów, co zostało przedstawione w Tabeli III.2. Dokładność wszystkich klasyfikatorów wynosi około 0.86, co oznacza, że 86% wszystkich obserwacji ze zbioru testowego zostało poprawnie sklasyfikowanych. Klasyfikatory identyfikują klasę pozytywną ze skutecznością 97-98%. W przypadku klasy negatywnej wartość jest znacząco niższa. Najlepiej klasę negatywną identyfikuje model logitowy ze skutecznością 18%, a najgorzej las losowy ze skutecznością 14%. Wszystkie klasyfikatory bardzo dobrze wykrywają występowanie cukrzycy, a zarazem bardzo słabo wykrywają klasę negatywną.

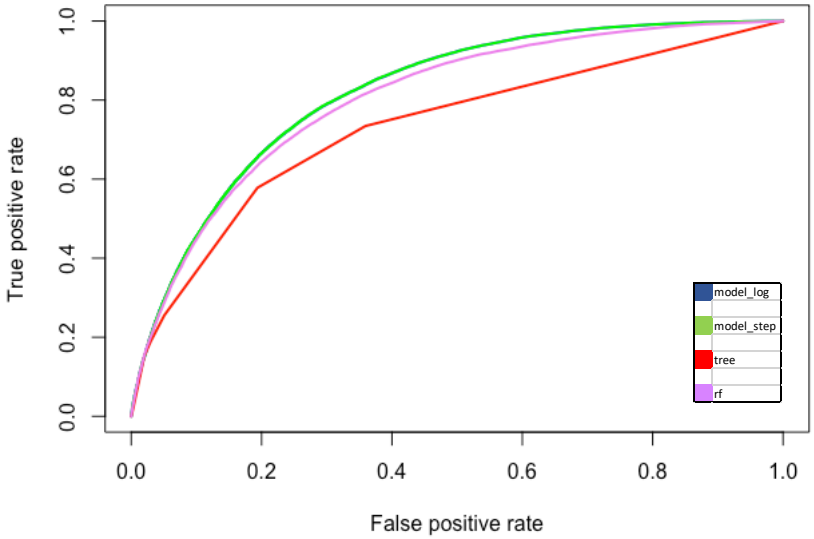
Tabela III. 2 Porównanie statystyk pochodnych modeli zbudowanych na zbiorze pierwszym

	Model logitowy	**Model logitowy ze stepwise**	**Drzewo decyzyjne**	**Las losowy**
dokładność	0.8595	0.8593	0.8601	0.8603
czułość	0.9744	0.9743	0.9798	0.9822
specyficzność	0.1802	0.1797	0.1529	0.14
PPV	0.8753	0.8753	0.8723	0.8709
NPV	0.5440	0.5424	0.5620	0.5708

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Biorąc pod uwagę krzywą ROC przedstawioną na Rysunku III.5 najlepszym klasyfikatorem jest model logitowy oraz model logitowy z optymalizacją stepwise. Wyniki są tak zbliżone, że różnica między modelami jest nieobserwowalna na wykresie. AUC w przypadku obu tych klasyfikatorów wynosi 0.819. Fioletowa krzywa odpowiada lasowi losowem i pole pod jej wykresem wynosi 0.803. Czerwona i najniżej położona krzywa odpowiada drzewu decyzyjnemu i AUC dla niej wynosi 0.731.

Rysunek III. 5 Krzywe ROC modeli zbudowanych na zbiorze pierwszym



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.3 Interpretacja oraz ocena wyników modelowania na drugim zbiorze

3.3.1 Model logitowy

W modelu na zbiorze zbalansowanym, którego parametry zostały przedstawione w Tabeli III.3 można zaobserwować, że podobnie jak w przypadku modelu na zbiorze niezbalansowanym największy wpływa na zwiększenie ryzyka zachorowania na cukrzycę ma nadciśnienie, wysoki cholesterol, wysokie BMI oraz zły generalny stan zdrowia. W przypadku tych czterech zmiennych zwiększenie wartości zmiennej o 1 powoduje wzrost iloraz szans o ponad 60%. Największy spadek szansy zaklasyfikowania obserwacji do klasy pozytywnej powoduje zwiększenie zmiennych *alco* oraz *racewhite* o jednostkę. Po raz kolejny zmienna dotycząca nadużywania alkoholu jest uznana za istotną i znacząco obniża ryzyko zachorowania na cukrzycę. Jak było wcześniej wspomniane może być to spowodowane dużą dysproporcją

w rozkładzie klas tej zmiennej. Do zmiennych nie istotnych należą zmienne *insurance, medcost, urban, smokre* oraz *raceblack*. W przypadku zmiennej *insurance* zaszła największa zmiana w porównaniu do modelowania na zbiorze pierwszym, ponieważ wówczas zmienna ta była istotna i miała szóstą w kolejności dodatnią siłę wpływu na zmienną objaśnianą.

Tabela III.3 Parametry modelu logitowego zbudowanego na zbiorze drugim

Zmienna	**Parametr strukturalny**	**Błąd standardowy**	**p-value**	**Istotność**
highBP	0.729479	0.021786	< 2e-16	***
highchol	0.581477	0.020875	< 2e-16	***
bmi	0.554212	0.013708	< 2e-16	***
genhealth	0.535324	0.012527	< 2e-16	***
heartdis	0.336147	0.033223	< 2e-16	***
sex	0.252381	0.021520	< 2e-16	***
stroke	0.242507	0.047599	3.49e-07	***
age	0.163699	0.004234	< 2e-16	***
diffwalk	0.114733	0.028700	6.40e-05	***
asthma	0.106255	0.029075	0.000258	***
fries	0.098549	0.022767	1.50e-05	**
insurance	0.092290	0.062443	0.139412	
depression	0.067263	0.026405	0.010853	*
medcost	0.045401	0.044302	0.305457	
urban	0.040145	0.028785	0.163128	
smoker	0.019112	0.011536	0.097575	.
alco	-0.813862	0.050999	< 2e-16	***
racewhite	-0.587631	0.072223	4.07e-16	***
raceother	-0.347287	0.091603	0.000150	***
racehispanic	-0.155647	0.081215	0.055303	.
raceblack	-0.127875	0.079166	0.106253	
fruit	-0.051827	0.021519	0.016023	*
physact	-0.161768	0.024400	3.36e-11	***
physhealth	-0.066577	0.023682	0.004934	**
education	-0.034264	0.012014	0.004345	**
income	-0.033477	0.005202	1.23e-10	***
veggie	-0.009492	0.027517	0.730138	
racenative	-0.006249	0.101282	0.950804	

‘***’ p= 0.001 ‘**’ p = 0.01 ‘*’ p = 0.05

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

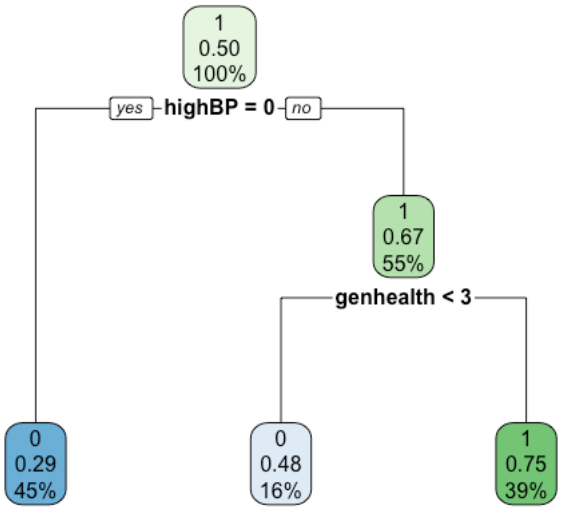
3.3.2 Model logitowy z wykorzystaniem algorytmu *stepwise*

W porównaniu do zastosowania algorytmu na zbiorze pierwszym, w przypadku zbioru zbalansowanego zostało wykluczone mniej zmiennych. Ze zbioru zmiennych objaśniających wykluczone zostały zmienne *insurance, veggie* oraz *medcost.* Parametry strukturalne w modelu nie uległy znaczącym zmianom w porównaniu do poprzedniego modelu.

3.3.3 Drzewo decyzyjne

Drzewo decyzyjne przedstawione na Rysunku III.6 i zbudowane na zbiorze zbalansowanym generuje znacząco mniej reguł decyzyjnych. Z tego powodu dostarczane jest mniej informacji na temat istotności poszczególnych zmiennych, jednak na tym zbiorze jest to najlepiej dopasowany model pod względem dokładności klasyfikatora. W pierwszym kroku sprawdzana jest obecność nadciśnienia. Jeśli nie występuje nadciśnienie obserwacja jest klasyfikowana jest do klasy negatywnej. W przeciwnym przypadku przeprowadzany jest kolejny test sprawdzający ocenę ogólnego stanu zdrowia. Jeśli ogólny stan zdrowia jest oceniany jako bardzo dobry lub dobry obserwacja jest klasyfikowana do klasy negatywnej, a w przeciwnym wypadku do klasy pozytywnej.

Rysunek III.6 Drzewo decyzyjne zbudowane na zbiorze drugim



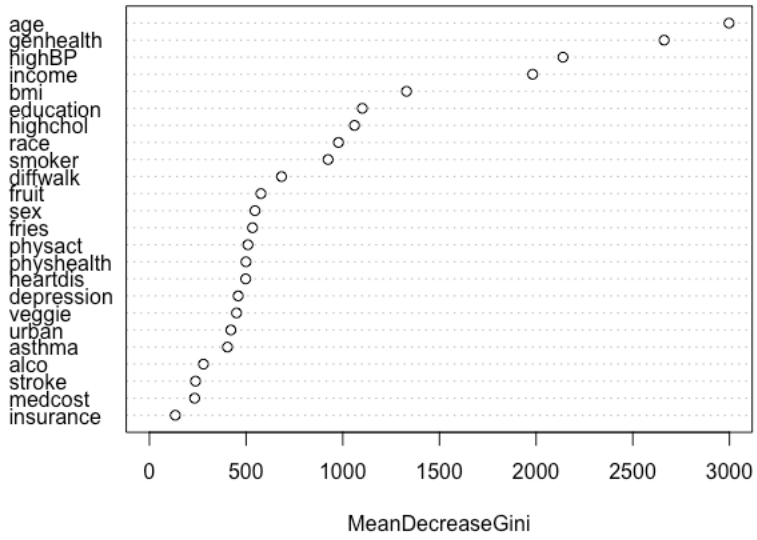
Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.3.4 Las losowy

Las losowy według Rysunku III.7 jako najważniejsze zmienne wyróżnia wiek, ogólny stan zdrowia, nadciśnienie, dochód, BMI, edukację oraz podwyższony poziom cholesterolu.

Warto też zwrócić uwagę na zmienne posiadające dwie klasy, które zostały sklasyfikowane niżej i jest to dodatkowo trudność z poruszaniem się, spożycie owoców i frytek oraz płeć. Zmienna *alco,* które była uznawana za bardzo znaczącą w modelu logitowym jest tutaj znacząco niżej sklasyfikowana.

Rysunek III.7 Istotność zmiennych w lesie losowym zbudowanym na zbiorze drugim



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.3.5 Ewaluacja jakości modeli

Po zbalansowaniu zbioru danych możemy zauważyć znaczące różnice między dokładnością klasyfikatorów. Według analiz, których wyniki są przedstawione w Tabeli III. 4 najniższą dokładność ma drzewo decyzyjne i klasyfikator ten poprawnie klasyfikuje 69% obserwacji. Najwyższą dokładność ma las losowy i klasyfikuje on prawidłowo 81% obserwacji. Najlepiej klasę pozytywną identyfikuje drzewo decyzyjne ze skutecznością 80%, a najgorzej las losowy ze skutecznością 77%. Klasa negatywna jest najlepiej identyfikowana przez las lasowy na poziomie 85%, a najgorzej przez drzewo decyzyjne na poziomie 58%. Patrząc holistycznie na statystyki las losowy jest najlepszym klasyfikatorem i najlepiej rozpoznaje klasę negatywną, jednak najwyższą skuteczność w rozpoznawaniu klasy pozytywnej ma drzewo decyzyjne.

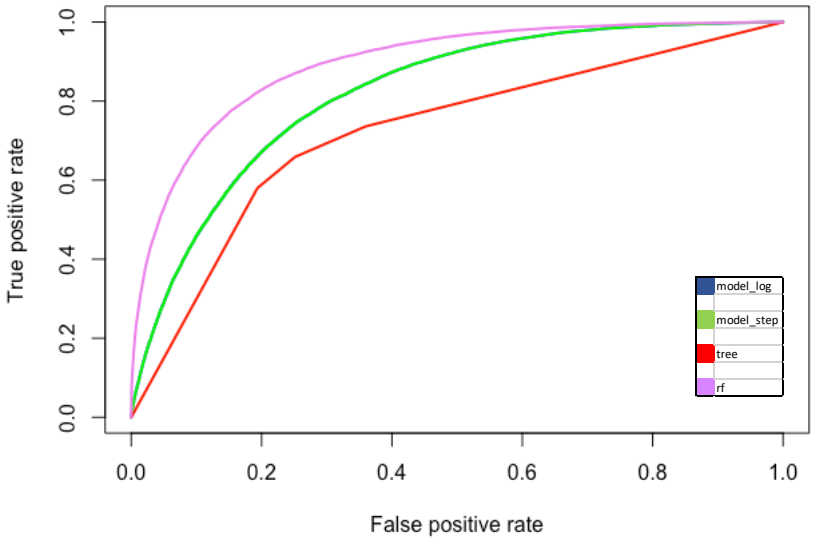
Tabela III.4 Porównanie statystyk pochodnych modeli zbudowanych na zbiorze drugim

	Model logitowy	**Model logitowy z stepwise**	**Drzewo decyzyjne**	**Las losowy**
dokładność	0.7462	0.7464	0.6912	0.8103
czułość	0.7272	0.7274	0.8053	0.7746
specyficzność	0.7651	0.7654	0.5775	0.8458
PPV	0.7550	0.7552	0.6548	0.8333
NPV	0.7382	0.7380	0.7247	0.7938

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Pod względem krzywych ROC znajdujących się na Rysunku III.8 najlepszym klasyfikatorem jest las losowy, a najgorszym drzewo decyzyjne. Wartość AUC dla lasu losowego wynosi 0.89. Pole pod krzywą ROC dla obu modeli logitowych wynosi 0.82, a dla lasu losowego 0.72.

Rysunek III. 8 Krzywe ROC modeli zbudowanych na zbiorze drugim



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.4 Interpretacja oraz ocena wyników modelowania na trzecim zbiorze

3.4.1 Model logitowy

Parametry modelu logitowego zostały przedstawione w Tabeli III.5. Najistotniejszy wpływ na klasyfikacje zmiennej do klasy pozytywnej ma nadciśnienie. Jego występowanie

zwiększa iloraz szans o 113%. Parametr strukturalny powyżej 0.5 występuje także przy zmiennych *highchol* oraz *bmi*. Ponownie można zaobserwować wysoką wartość parametru strukturalnego zmiennej oznaczającej nadużywanie alkoholu. Do zmiennych znacząco obniżających prawdopodobieństwo zachorowania na cukrzycę możemy zaliczyć przynależenie do rasy białej oraz wykonywanie aktywności fizycznej.

Tabela III.5 Parametry modelu logitowego zbudowanego na zbiorze trzecim

Zmienna	**Parametr strukturalny**	**Błąd standardowy**	**p-value**	**Istotność**
highBP	0.764850	0.022518	< 2e-16	***
bmi	0.564024	0.013562	< 2e-16	***
highchol	0.563210	0.020659	< 2e-16	***
genhealth	0.457401	0.011976	< 2e-16	***
heartdis	0.275838	0.028494	< 2e-16	***
insurance	0.265976	0.056336	2.34e-06	***
sex	0.206425	0.020871	< 2e-16	***
age	0.149280	0.004047	< 2e-16	***
stroke	0.146474	0.039129	0.000182	***
diffwalk	0.118345	0.025610	3.82e-06	***
depression	0.112297	0.025406	9.86e-06	*
fries	0.083146	0.022329	0.000196	**
asthma	0.079017	0.027331	0.003839	**
medcost	0.053961	0.039510	0.172014	
smoker	0.031384	0.011243	0.005250	**
urban	0.022481	0.027464	0.819	
veggie	0.002132	0.025612	0.933670	
alco	-0.786720	0.058265	< 2e-16	***
racewhite	-0.645951	0.070227	< 2e-16	***
raceother	-0.533390	0.087892	1.29e-09	***
raceblack	-0.222030	0.075833	0.003413	**
racehispanic	-0.154268	0.076886	0.044807	*
racenative	-0.097384	0.096973	0.315261	
education	-0.042122	0.010748	8.89e-05	***
physhealth	-0.038770	0.023100	0.093283	
fruit	-0.03726	0.020946	0.075236	
income	-0.030370	0.004912	6.30e-10	***
physact	-0.167076	0.022237	5.77e-14	***

‘***’ p= 0.001 ‘**’ p = 0.01 ‘*’ p = 0.05

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

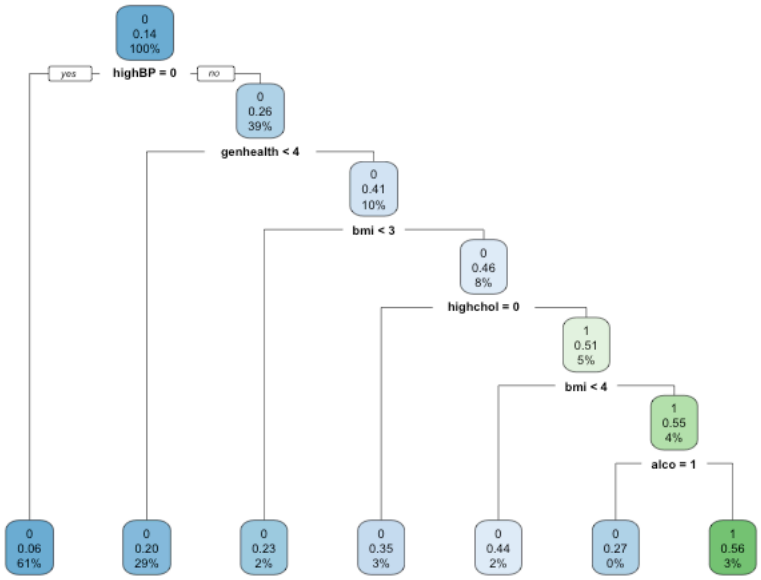
3.4.2 Model logitowy z wykorzystaniem algorytmu *stepwise*

Parametry strukturalne w przypadku wszystkich poprzednich optymalizacji nie uległy znaczącym zmianom. Algorytm wykluczył z modelu trzy zmienne: *medcost, urban* oraz *veggie.* Oznacza to, że ich wykluczenie z modelu nie prowadzi do spadku AIC. Zmienne te charakteryzowały się wartością p powyżej 0.15.

3.4.3 Drzewo decyzyjne

Omawiane drzewo decyzyjne przedstawione jest na Rysunku III.9. W pierwszym kroku sprawdzana jest wartość zmiennej odpowiadającej występowaniu nadciśnienia. Jeśli nadciśnienie nie występuje obserwacja klasyfikowana jest do klasy negatywnej, w przeciwnym wypadku obserwacja trafia do kolejnego węzła, gdzie sprawdzana jest wartość zmiennej odpowiadającej ocenie ogólnego stanu zdrowia. Jeśli stan zdrowia został oceniony jako bardzo dobry, dobry lub przeciętny obserwacja jest klasyfikowana do klasy negatywnej, w przeciwnym wypadku obserwacja trafia do kolejnego węzła, gdzie badane jest BMI. Jeśli występuje niedowaga lub waga jest w normie obserwacja jest klasyfikowana do klasy negatywnej, w przeciwnym wypadku w kolejnym węźle sprawdzana jest wartość zmiennej *highchol.* Jeśli nie występuje podwyższony poziom cholesterolu obserwacja klasyfikowana jest do klasy negatywnej, w przeciwnym wypadku po raz kolejny sprawdzany jest poziom BMI. Gdy występuje nadwaga obserwacja jest klasyfikowana do klasy negatywnej, jeśli występuje otyłość sprawdzana jest ocena generalnego stanu zdrowia. Jeśli stan zdrowia został oceniony jako bardzo zła obserwacja jest klasyfikowana do klasy pozytywnej. W przeciwnym wypadku sprawdzana jest wartość zmiennej *alco*. W sytuacji, gdy badany nadużywa alkoholu zmienna klasyfikowana jest do klasy negatywnej, w przeciwnym wypadku obserwacja trafia do klasy pozytywnej.

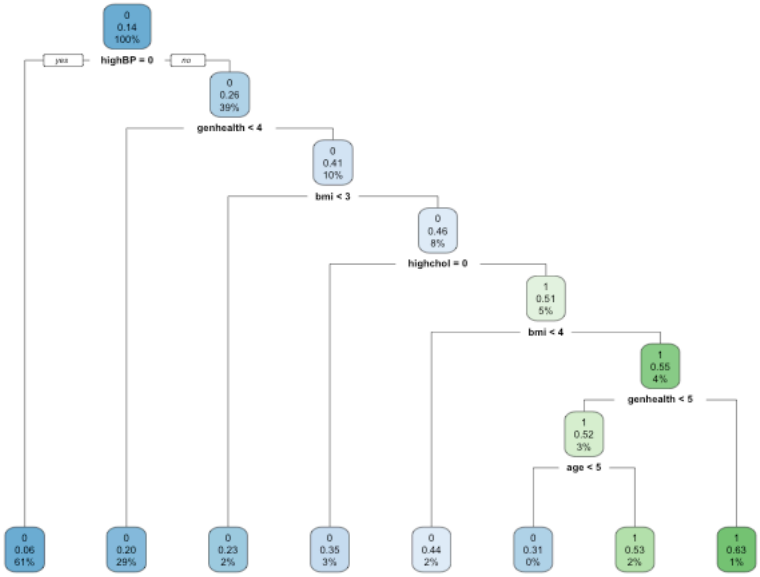
Rysunek III.9 Drzewo decyzyjne zbudowane na zbiorze trzecim



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Alkohol negatywnie wpływa na zdrowie, dlatego powstała reguła decyzyjna jest prawdopodobnie wynikiem niskiego udziału klasy pozytywnej we wszystkich obserwacjach. Z tego powodu zostało także zbudowane drzewo z wykluczeniem z tej zmiennej ze zbioru zmiennych objaśniających. Drzewo to jest przedstawione na Rysunku III.10. W nowo powstałym drzewie reguła decyzyjna dotycząca zmiennej *alco* została zastąpiona regułą decyzyjną dotyczącą zmiennej *age.* W przypadku wieku poniżej 40 lat zmienna klasyfikowana jest do klasy negatywnej, a w przypadku wyższego wieku do klasy pozytywnej.

Rysunek III.10 Drzewo decyzyjne zbudowane na zbiorze trzecim z wykluczeniem zmiennej *alco*

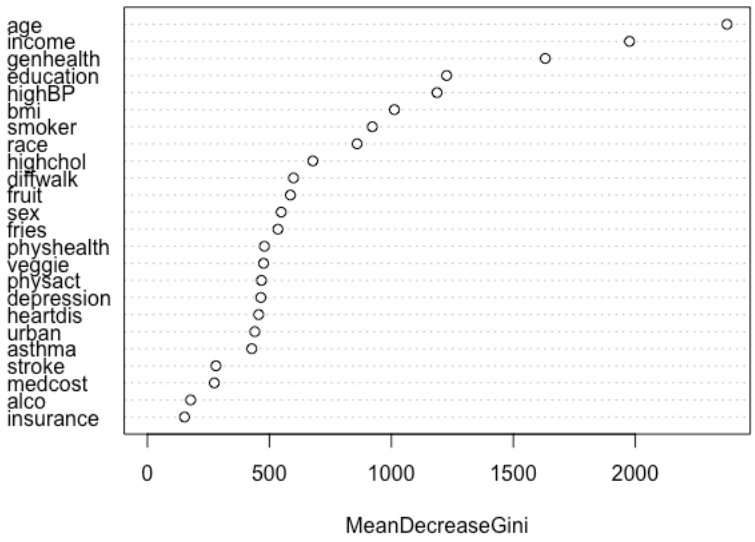


Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.4.4 Las losowy

Według Rysunku III.11 w grupie zmiennych posiadających więcej niż dwie klasy największe znaczenie ma wiek, dochód, edukacja, BMI, palenie papierosów oraz rasa. W kategorii zmiennych binarnych największe znaczenie ma nadciśnienie, wysoki cholesterol, trudność z poruszaniem, spożycie owoców, płeć oraz spożycie frytek. Warto także zauważyć, że w przypadku lasu losowego wpływ zmiennej *alco* odpowiadającej wysokiemu spożyciu alkoholu jest niski w porównaniu z pozostałymi zmiennymi objaśniającymi.

Rysunek III.11 Istotność zmiennych w lesie losowym zbudowanym na zbiorze trzecim



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.4.5 Ewaluacja jakości modeli

Jak można zauważyć w Tabeli III.6 wszystkie klasyfikatory charakteryzują się podobną dokładnością na poziomie w przybliżeniu 87%. Do klasy pozytywnej poprawnie klasyfikowane jest 98% obserwacji. Pod względem specyficzności najlepsze są oba modele logitowe, które poprawnie klasyfikują 17% obserwacji należących do klasy negatywnej, co wciąż nie jest satysfakcjonującym wynikiem. Dla wszystkich klasyfikatorów prawidłowe klasyfikacje stanowią około 88% klasyfikacji w klasie pozytywnej. Najwyższy udział prawidłowych klasyfikacji w klasie negatywnej występuje w lesie losowym, a najmniejszy w modelach logitowych. Drzewo decyzyjne po wykluczeniu zmiennej objaśnianej *alco* uzyskuje wyniki bardzo zbliżonej jakości do wyników wyjściowego drzewa.

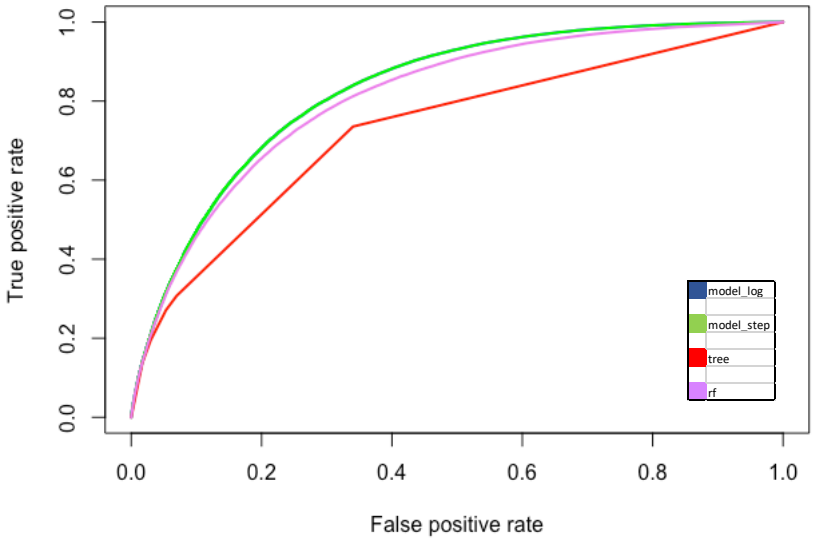
Tabela III.6 Porównanie statystyk pochodnych modeli zbudowanych na zbiorze trzecim

	Model logitowy	**Model logitowy z stepwise**	**Drzewo decyzyjne**	**Las losowy**
dokładność	0.8669	0.8668	0.8672	0.8675
czułość	0.9777	0.9777	0.9830	0.9844
specyficzność	0.1679	0.1678	0.1372	0.1304
PPV	0.8811	0.8811	0.8778	0.8771
NPV	0.5443	0.5441	0.5610	0.5698

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Krzywe ROC dla wszystkich modeli przedstawione są na Rysunku III.12. Pod względem krzywej ROC najlepszymi klasyfikatorami są modele logitowe z AUC na poziomie 0.83, którym odpowiada zielona krzywa ROC. Bardzo zbliżone wartości AUC, na poziomie 0.81 ma las losowy. Najgorszym klasyfikatorem pod względem tego kryterium jest drzewo decyzyjne z AUC na poziomie 0.73.

Rysunek III.12 Krzywe ROC modeli zbudowanych na zbiorze trzecim



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.5 Interpretacja oraz ocena wyników modelowania na czwartym zbiorze

3.5.1 Model logitowy

Tabela III.7 prezentuje parametry omawianego modelu logitowego. Zmienne, z największym wpływem na ryzyko zachorowania nie uległy zmianom w stosunku do zbioru trzeciego. Największy wpływ na zaklasyfikowanie zmiennej do klasy pozytywnej mają zmienne *highBP, bmi* i *highchol*, a największy wpływ na zaklasyfikowanie do klasy negatywnej mają zmienne *alco, racewhite* oraz *physact.*

Tabela III.7 Parametry modelu logitowego zbudowanego na zbiorze czwartym

Zmienna	Parametr strukturalny	Błąd standardowy	p-value	Istotność
highBP	0.765660	0.015825	< 2e-16	***
bmi	0.569445	0.009771	< 2e-16	***
highchol	0.559490	0.015209	< 2e-16	***
genhealth	0.482010	0.008851	< 2e-16	***
heartdis	0.355796	0.024158	< 2e-16	***
insurance	0.346955	0.039892	< 2e-17	***
sex	0.220622	0.015572	< 2e-16	***
age	0.173027	0.002906	< 2e-17	***
stroke	0.158910	0.033081	1.56e-06	***
diffwalk	0.147712	0.020392	4.37e-13	***
depression	0.125490	0.019673	1.79e-10	***
medcost	0.073481	0.030137	0.0148	*
smoker	0.050690	0.008345	1.24e-09	***
asthma	0.044702	0.021204	0.0350	*
fries	0.018698	0.01666	0.2618	**
veggie	0.017736	0.019586	0.3652	
urban	0.015147	0.02073	0.4650	
racewhite	-0.728367	0.049527	< 2e-15	***
alco	-0.701683	0.038078	< 2e-16	***
raceother	-0.519155	0.062603	< 2e-16	***
raceblack	-0.255702	0.054645	2.88e-06	***
racehispanic	-0.249178	0.055191	6.34e-06	***
physact	-0.140279	0.017215	1.56e-06	***
fruit	-0.077772	0.015696	7.23e-07	***
education	-0.041741	0.008260	4.34e-0	***
income	-0.036664	0.003669	< 2e-16	***
physhealth	-0.019670	0.017179	0.2522	
racenative	-0.011340	0.071430	-0.159	

‘***’ p= 0.001 ‘**’ p = 0.01 ‘*’ p = 0.05

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

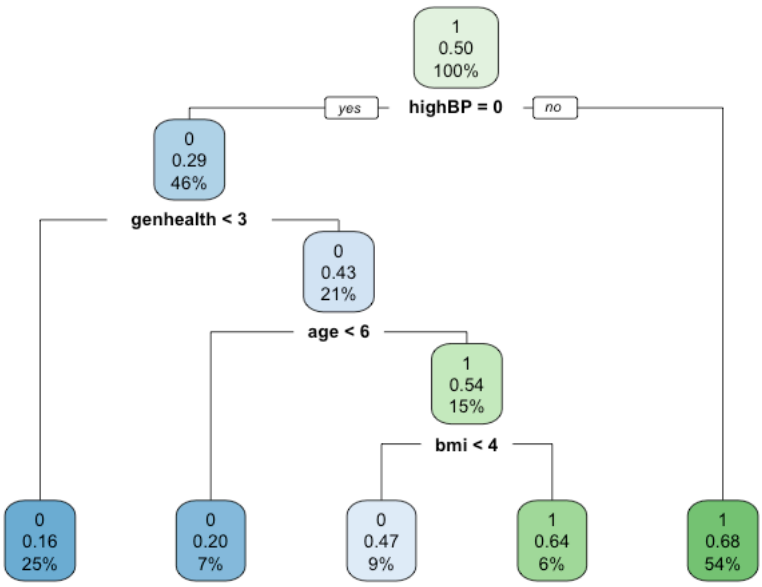
3.5.2 Model logitowy z wykorzystaniem algorytmu stepwise

Po zastosowaniu algorytmu stepwise z modelu logitowego zostały usunięte cztery zmienne objaśniające. Trzy z nich były nieistotne i były to zmienne *veggie, urban* oraz *physhealth*. Usunięta została także zmienna objaśniająca *fries,* która była zmienna statystycznie istotną.

3.5.3 Drzewo decyzyjne

W drzewie decyzyjnym z Rysunku III.13 zbudowanym na zbalansowanym zbiorze występuje mniej reguł decyzyjnych niż w drzewie zbudowanym na niezbalansowanym zbiorze. Jak w przypadku wszystkich pozostałych drzew najpierw sprawdzana jest wartość zmiennej *highBP*. Jeśli występuje nadciśnienie zmienna klasyfikowana jest do klasy pozytywnej. W przeciwnym wypadku sprawdzana jest ocena generalnego stanu zdrowia. Jeśli stan zdrowia został oceniony jako bardzo dobry lub dobry obserwacja trafia do klasy negatywnej, w przeciwnej sytuacji sprawdzana jest wartość zmiennej *age*. W przypadku wieku poniżej 45 lat obserwacji przypisywana jest klasa negatywna, w sytuacji, kiedy wiek jest równy lub wyższy 45 lat testowana jest wartość BMI. Jeśli występuje otyłość obserwacja klasyfikowana jest do klasy pozytywnej, a w przypadku jej braku zmienna trafia do klasy negatywnej.

Rysunek III.13 Drzewo decyzyjne zbudowane na zbiorze czwartym



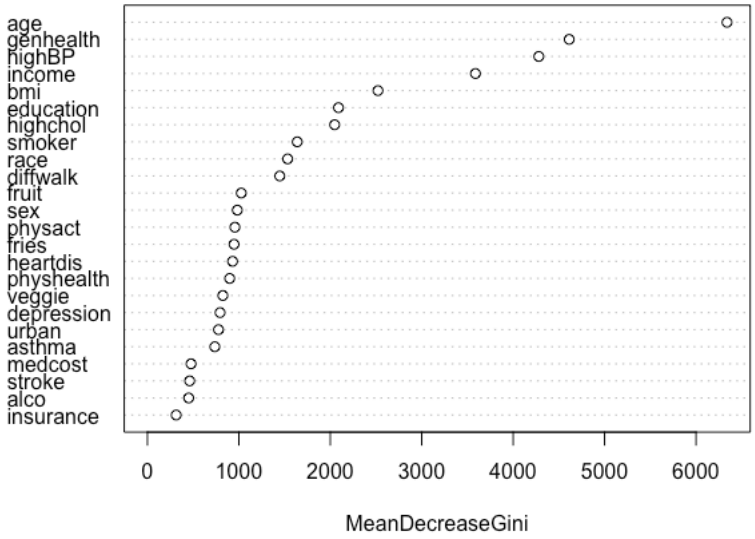
Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.5.4 Las losowy

Analizując Rysunek III.14 można zauważyć, że spośród zmiennych binarnych największe znaczenie mają zmienne odpowiadające nadciśnieniu, podwyższonemu cholesterolowi i trudności z poruszaniem się. Wśród zmiennych z większą liczbą klas

największe znaczenie ma wiek, ogólny stan zdrowia, dochód, edukacja, palenie papierosów oraz rasa.

Rysunek III.14 Istotność zmiennych w lesie losowym zbudowanym na zbiorze czwartym



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

3.5.5 Ewaluacja jakości modeli

Statystyki pochodne wyliczone na podstawie macierzy błędów zostały przedstawione w Tabeli III.8. Pod względem poprawnych klasyfikacji najlepszym klasyfikatorem jest las losowy ze dokładnością na poziomie 0.81. Klasyfikator ten najlepiej rozpoznaje klasę pozytywną jak i negatywną. Najniższą dokładność ma drzewo decyzyjne oraz najgorzej rozpoznaje klasę pozytywną.

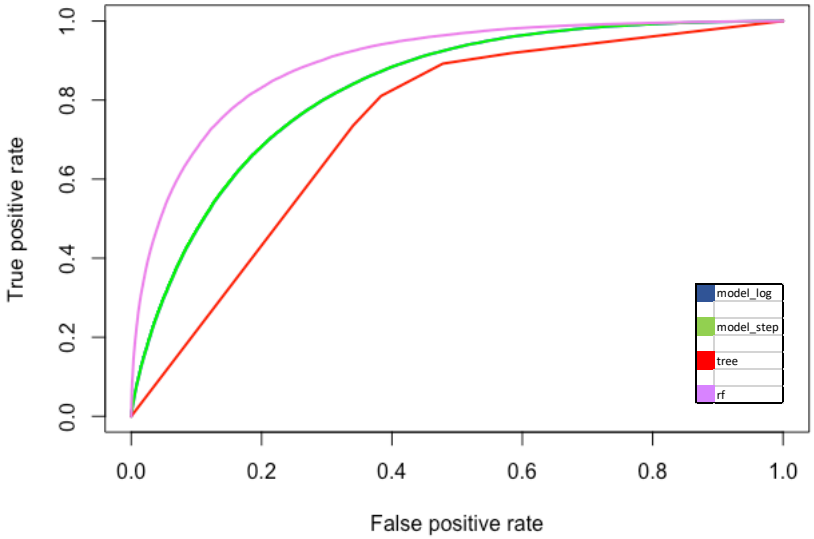
Tabela III.8 Porównanie statystyk pochodnych modeli zbudowanych na zbiorze czwartym

	Model logitowy	**Model logitowy z stepwise**	**Drzewo decyzyjne**	**Las losowy**
dokładność	0.7522	0.7524	0.7138	0.8139
czułość	0.7289	0.7292	0.6167	0.7639
specyficzność	0.7754	0.7755	0.8104	0.8638
PPV	0.7636	0.7638	0.7641	0.8481
NPV	0.7418	0.7421	0.6799	0.7861

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Krzywe ROC przedstawione są na Rysunku III.15. Najwyżej znajduje się krzywa ROC odpowiadająca lasowi losowemu z polem pod krzywą na poziomie 0.9, co kwalifikuje klasyfikator jako bardzo dobry. Kolejne krzywe odpowiadają modelom logitowym z AUC na poziomie 0.83. Najgorszym modelem pod tym względem jest drzewo decyzyjne z AUC równym 0.73.

Rysunek III.15 Krzywe ROC modeli zbudowanych na zbiorze czwartym



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Zakończenie

W pracy zostały przedstawione cztery różne modele budowane na czterech różnych zbiorach. W ten sposób możliwe było zbadanie wpływu różnych metod imputacji danych oraz poziomu zbalansowania zbioru na jakość omawianych modeli uczenia maszynowego. Pod względem dokładności najlepszym klasyfikatorem jest las losowy zbudowany na niezbalansowanym zbiorze z wypełnieniem braków danych z wykorzystaniem biblioteki *mice.* Identyfikuje on także najlepiej klasę pozytywną ze skutecznością powyżej 98%, jednak tak jak wszystkie modele zbudowane na zbiorach niezbalansowanych bardzo słabo identyfikuje klasę negatywną ze skutecznością jedynie 13%. Zastosowanie takiego klasyfikatora do wyboru osób, które powinny być objęte profilaktyką, z powodu przynależności do grupy wysokiego ryzyka sprawiłoby, że większość pacjentów zostałaby objętą profilaktyką. Klasyfikacja mimo brak wystarczających przesłanek przyczyniłby się do generowania nadmiernych kosztów. Przy tej interpretacji zakładamy, że osoba jest wysoce zagrożona, jeśli na podstawie odpowiedzi na pytania, którym odpowiadają zmienne objaśniające w modelu, model zaklasyfikował ją do klasy pozytywnej. Może to wskazywać na niezdiagnozowaną cukrzycę, która według szacunków w Stanach Zjednoczonych stanowi jedną czwartą przypadków. Pod względem wartość AUC najlepszym klasyfikatorem jest las losowy zbudowany na zbalansowanym zbiorze z brakami danych wypełnionymi z wykorzystaniem biblioteki *mice.* Wartość AUC w tym modelu jest na poziomie 0.9, co sprawia, że pod względem tego parametru klasyfikator może zostać uznany jako bardzo dobry. Wartość ta przekracza wartości AUC obserwowane w przywoływanych badaniach, jednak zbiory w tamtych przypadkach nie były zbalansowane. Omawiany klasyfikator wykrywa klasę pozytywna ze skutecznością 76% i klasę negatywną ze skutecznością 86%. W przypadku zastosowania tego klasyfikatora do identyfikacji pacjentów z grupy wysokiego ryzyka znaczna część przypadków nie zostałaby wykryta. Wartość AUC tego samego modelu zbudowanego na zbiorze niezbalansowanym z wypełnieniem braków z wykorzystaniem biblioteki *mice* wynosi 0.81. Można także zauważyć, że zastosowanie biblioteki *mice* poprawiło jakość wszystkich klasyfikatorów.

Pod względem czynników ryzyka sprzyjających występowaniu cukrzycy jako najważniejszy oba modele logitowy oraz drzewo decyzyjne wskazały występowanie nadciśnienia. Jako również ważne stany chorobowe zostały wskazane wysoki poziom cholesterolu, nadwaga oraz otyłość. Może to być przesłanka do rekomendacji częstszych badań poziomu glukozy osobom cierpiącym na te przypadłości. Modele wskazały także istotny, lecz nie równie silny, wpływ diety na ryzyko zapadnięcia na cukrzycę, a konkretnie

wpływ spożywania owoców na zmniejszenie tego ryzyka i przeciwny wpływ spożywania powyżej dwóch porcji frytek dziennie, które są produktem z wysoką zawartością węglowodanów i tłuszczy. Pod względem czynników socjoekonomicznych dochód oraz edukacja zostały wskazane przez modele logitowe jako czynniki ograniczające zachorowanie na cukrzycę. Do czynników, które mają prewencyjne działanie w przypadku tej choroby można też zaliczyć aktywność fizyczną. Pod względem czynników demograficznych według reguł decyzyjnych występujących w drzewach decyzyjnych zbudowanych na zbiorach niezbalansowanych wiek powyżej 40 lat jest także wskazaniem do częstszego badania się pod kątem cukrzycy. Dodatkowo dużo bardziej zagrożone są osoby należące do grup etnicznych innych niż rasa biała.

Zbudowane modele mogą posłużyć do oceny ryzyka zachorowania lub występowania niestwierdzonej cukrzycy na podstawie odpowiedzi na 24 pytania ankietowe. Dodatkowo parametry strukturalne w modelach logitowych oraz reguły decyzyjne w drzewie decyzyjnym pozwalają zidentyfikować wagę danych czynników ryzyka lub zachowań prewencyjnych. Wczesne wykrycie cukrzycy minimalizuje jej negatywny wpływ na zdrowie chorego i co za tym idzie z perspektywy ekonomicznej obniża koszty jakie ponosi gospodarka oraz jednostka z powodu choroby.

Bibliografia:

Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T., & Sidorchuk, A. (2011). Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. *International Journal of Epidemiology*, 40(3), s.804–818.

American Diabetes Association (2018). Economic Costs of Diabetes in the U.S. in 2017, *Diabetes Care, 41(5),* s. 917–928.

American Heart Association, *Diabetes Risk Factors*, https://www.heart.org/en/health-topics/diabetes/understand-your-risk-for-diabetes (dostęp: 21.12.2022).

Anastasia C. Thanopoulou, Basil G. Karamanos, Francesco V. Angelico, Samir H. Assaad-Khalil, Alfredo F. Barbato, Maria P. Del Ben, Predrag B. Djordjevic, Vesna S. Dimitrijevic-Sreckovic, Cristina A. Gallotti, Nikolaos L. Katsilambros, Ilias N. Migdalis, Mansouria M. Mrabet, Malina K. Petkova, Demetra P. Roussi, Maria-Teresa P. Tenconi, (2003). Dietary Fat Intake as Risk Factor for the Development of Diabetes : Multinational, multicenter study of the Mediterranean Group for the Study of Diabetes (MGSD). *Diabetes Care,*26 (2), s.302–307.

Araszkiewicz, A., Bandurska-Stankiewicz, E., Borys, S., Budzyński, A., Cyganek, K., Cypryk, K., Czech, A., Czupryniak, L., Drzewoski, J., Dzida, G., Dziedzic, T., Franek, E., Gajewska, D., Gawrecki, A., Górska, M., Grzeszczak, W., Gumprecht, J., Idzior-Waluś, B., Jarosz-Chobot, P., . . . Moczulski, D. (2021), 2021 Guidelines on the management of patients with diabetes. A position of Diabetes Poland*, Clinical Diabetology*, 10(1), s.1–113.

Bąk, E., Nowak - Kapusta, Z., Dobrzyn-Matusiak, D., Marcisz-Dyla, E., Marcisz, C., & Krzemińska, S. (2019). An assessment of diabetes-dependent quality of life (ADDQoL) in women and men in Poland with type 1 and type 2 diabetes, *Annals of Agricultural and Environmental Medicine*, 26(3), s.429–438.

Blockeel, H., & Struyf, J. (2003). Efficient algorithms for decision tree cross-validation, *Journal of Machine Learning Research*, 3, s.621–650.

Bommer, C., Sagalova, V., Heesemann, E., Manne-Goehler, J., Atun, R., Bärnighausen, T., Davies, J., & Vollmer, S. (2018). Global Economic Burden of Diabetes in Adults: Projections From 2015 to 2030, *Diabetes Care*, 41(5), s. 963–970.

CDC, *2021 BRFSS Survey Data and Documentation,* https://www.cdc.gov/brfss/annual_data/annual_2021.html (dostęp: 05.02.2023).

Center of Disease Control, *Prevalence of Both Diagnosed and Undiagnosed Diabetes*, https://www.cdc.gov/diabetes/data/statistics-report/diagnosed-undiagnosed-diabetes.html (dostęp: 5.12.2022)

Choueiry, G., *Understand Forward and Backward Stepwise Regression*, https://quantifyinghealth.com/stepwise-selection/ (dostęp: 10.01.2023).

Dendup T, Feng X, Clingan S, Astell-Burt T. (2018). Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review. *International Journal of Environmental Research and Public Health*,15(1), s.78.

Diabetes UK, *Diabetes risk factors*, https://www.diabetes.org.uk/preventing-type-2-diabetes/diabetes-risk-factors (dostęp: 16.12.2022)

Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning, *BMC Medical Informatics and Decision Making*, 19(1).

Dutta, D., Paul, D., & Ghosh, P. (2018). Analysing Feature Importances for Diabetes Prediction using Machine Learning. 2018 IEEE 9th Annual Information Technology, *Electronics and Mobile Communication Conference (IEMCON).*

Gromada, M. (2006). *Drzewa klasyfikacyjne, ich budowa, problemy złożoności i skalowalności*

GUS, *Realizacja Celów Zrównoważonego Rozwoju w Polsce. Raport 2018 - Wskaźnik 3.1.e - Liczba zgonów w wyniku cukrzycy na 100 tys. ludności*, https://sdg.gov.pl/statistics_nat/3-1-e/ (dostęp: 15.12.2022).

GUS, *Zdrowie i Ochrona Zdrowia w 2020 roku,* https://stat.gov.pl/files/gfx/portalinformacyjny/pl/defaultaktualnosci/5513/1/11/1/zdrowie_i_ochrona_zdrowia_2020_korekta.pdf , (dostęp: 15.12.2022)

Hossin, M., & Sulaiman, M. R. (2015). A Review on Evaluation Metrics for Data Classification Evaluations, *International Journal of Data Mining & Knowledge Management Process*, 5(2), s.1–11.

Hulbert, L. R., Michael, S. L., Charter-Harris, J., Atkins, C., Skeete, R. A., & Cannon, M. J. (2022). Effectiveness of Incentives for Improving Diabetes-Related Health Indicators in Chronic Disease Lifestyle Modification Programs: a Systematic Review and Meta-Analysis, *Preventing Chronic Disease*, 19.

IDF, *Prevention*. https://www.idf.org/aboutdiabetes/prevention.html, (dostęp: 21.12.2022).

Instytut Ochrony Zdrowia, *Rekomendacje w zakresie kompleksowej opieki nad pacjentami z retinopatią cukrzycową,* *https://www.ioz.org.pl/_files/ugd/e91ac2_5d595422c5cf46c69cce6e3ee3aa37db.pdf*

International Diabetes Federation, *IDF Diabetes Atlas 10th Edition*, https://diabetesatlas.org/data/en/ (dostęp: 15.12.2022)

Jackowska, B., (2011). Efekty interakcji między zmiennymi objaśniającymi w modelu logitowym w analizie zróżnicowania ryzyka zgonu, *Przegląd Statystyczny,* 58(1-2), s.24-41.

Jadhev, S. D., & Channe, H. (2016). P.Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques, *International Journal of Science and Research*, 5(1), s.1842-1845.

Koloverou, E., Esposito, K., Giugliano, D., & Panagiotakos, D. (2014). The effect of Mediterranean diet on the development of type 2 diabetes mellitus: A meta-analysis of 10 prospective studies and 136,846 participants, *Metabolism*, 63(7), s.903–911.

Magliano, D., & Boyko, E. J. (2021). *IDF Diabetes Atlas*. International Diabetes Federation.

Merabet, N., Lucassen, P. J., Crielaard, L., Stronks, K., Quax, R., Sloot, P. M., la Fleur, S. E., & Nicolaou, M. (2022). How exposure to chronic stress contributes to the development of type 2 diabetes: A complexity science approach, *Frontiers in Neuroendocrinology*, 65.

Ministerstwo Zdrowia, *Cukrzyca w liczbach,* https://pacjent.gov.pl/artykul/cukrzyca-w-liczbach

Ministesrtwo Zdrowia, *Cukrzyca – Mapy potrzeb zdrowotnych,* https://basiw.mz.gov.pl/analizy/problemy-zdrowotne/cukrzyca-wersja-polska/, (dostęp:15.12.2022)

Ranganathan, S., Nakai, K., & Schonbach, C. (2018*). Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics.* Wydawinictwo Elsevier.

Shmerling, R. H., MD. (2019). *In defense of French fries*, Harvard Health, https://www.health.harvard.edu/blog/in-defense-of-french-fries-2019020615893 (dostęp: 03.02.2022).

Steyn, N., Mann, J., Bennett, P., Temple, N., Zimmet, P., Tuomilehto, J., Louheranta, A. (2004). Diet, nutrition and the prevention of type 2 diabetes, *Public Health Nutrition*, 7(1a), s.147-165.

IDF, *Diabetes risk factors,* https://www.diabetes.org.uk/preventing-type-2-diabetes/diabetes-risk-factors (dostęp: 21.12.2022).

Strobl C., Boulesteix A. L., Zeileis A., & Hothorn T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC Bioinformatics*, 8(1).

Turi, K. N., Buchner, D. M., & Grigsby-Toussaint, D. S. (2017). Predicting Risk of Type 2 Diabetes by Using Data on Easy-to-Measure Risk Factors*, Preventing Chronic Disease,* 14.

Weickert, M., Pfeiffer, A. (2018). Impact of Dietary Fiber Consumption on Insulin Resistance and the Prevention of Type 2 Diabetes, *The Journal of Nutrition*, 148(1), s.7–12.

WHO, *Diabetes,* https://www.who.int/news-room/fact-sheets/detail/diabetes (dostęp: 5.12.2022)

World Bank*, Diabetes prevalence (% of population ages 20 to 79),* https://data.worldbank.org/indicator/SH.STA.DIAB.ZS (dostęp: 15.12.2022)

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques, *Preventing Chronic Disease*, 16.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques, *Frontiers in Genetics*, 9.

Spis tabel :

Spis rysunków:

Załącznik 1: Skrypt w R dla zbioru pierwszego i drugiego

```
library(readr)
library(tidyr)
library(stats)
library(dplyr)
library(haven)
library(tidyverse)
library(caret)
library(MASS)
library(randomForest)
library(ROCR)
library(ROSE)
library(ggplot2)
library(rpart)
library(rpart.plot)
library(caret)
library(ggcorrplot)
library(ROCR)

df <- read_xpt("dane_sas/2021.xpt", col_select = c(DIABETE4, SEXVAR, GENHL
TH, PHYSHLTH,
                                                   MEDCOST1, EXERANY2,"_RF
HYPE6" ,"_RFCHOL3" , CVDSTRK3, "_MICHD", ASTHMA3,  ADDEPEV3,
                                                  EDUCA, INCOME3, DIFFWALK
, "_SMOKER3","_URBSTAT", "_IMPRACE","_HLTHPLN","_AGEG5YR",
                                                    "_BMI5CAT","_RFDRHV7",
"_FRTLT1A", "_VEGLT1A", "FRNCHDA_"))

df <- as.data.frame(df)
df <- na.omit(df)

colnames(df)[colnames(df)=="DIABETE4"] = "diabetes"
colnames(df)[colnames(df)=="SEXVAR"] = "sex"
colnames(df)[colnames(df)=="GENHLTH"] = "genhealth"
colnames(df)[colnames(df)=="PHYSHLTH"] = "physhealth"
colnames(df)[colnames(df)=="MEDCOST1"] = "medcost"
colnames(df)[colnames(df)=="EXERANY2"] = "physact"
colnames(df)[colnames(df)=="_RFHYPE6"] = "highBP"
colnames(df)[colnames(df)=="_RFCHOL3"] = "highchol"
colnames(df)[colnames(df)=="CVDSTRK3"] = "stroke"
colnames(df)[colnames(df)=="_MICHD"] = "heartdis"
colnames(df)[colnames(df)=="ASTHMA3"] = "asthma"
colnames(df)[colnames(df)=="ADDEPEV3"] = "depression"
colnames(df)[colnames(df)=="EDUCA"] = "education"
colnames(df)[colnames(df)=="INCOME3"] = "income"
colnames(df)[colnames(df)=="DIFFWALK"] = "diffwalk"
colnames(df)[colnames(df)=="_SMOKER3"] = "smoker"
colnames(df)[colnames(df)=="_URBSTAT"] = "urban"
colnames(df)[colnames(df)=="_IMPRACE"] = "race"
colnames(df)[colnames(df)=="_HLTHPLN"] = "insurance"
colnames(df)[colnames(df)=="_AGEG5YR"] = "age"
colnames(df)[colnames(df)=="_BMI5CAT"] = "bmi"
colnames(df)[colnames(df)=="_RFDRHV7"] = "alco"
```

```
colnames(df)[colnames(df)=="_FRTLT1A"] = "fruit"
colnames(df)[colnames(df)=="_VEGLT1A"] = "veggie"
colnames(df)[colnames(df)=="FRNCHDA_"] = "fries"

#cukrzyca
df <- df[df$diabetes !=7,] #dont know
df <- df[df$diabetes !=9,] #refused
df <- df[df$diabetes !=4,] #prediabetes
df$diabetes[df$diabetes == 3] <- 0 #in pregnancy
df$diabetes[df$diabetes == 2] <- 0 #no diabetes
summary(df$diabetes)
#płeć
df$sex[df$sex == 2] <- 0 #kobieta
df <- df[df$sex !=7,] #dont know
df <- df[df$sex !=9,] #refused
summary(df$sex)
df$sex <-as.numeric(df$sex)

#generalny stan zdrowia 1 - good 5 - bad
df <- df[df$genhealth !=7,] #dont know
df <- df[df$genhealth !=9,] #refused
summary(df$genhealth)


#zdrowie fizyczne - ilość dni ze złym stanem
df$physhealth[df$physhealth == 88] <- 0 #no days
df <- df[df$physhealth !=77,] #dont know
df <- df[df$physhealth !=99,] #refused
df$physhealth[df$physhealth != 0] <- 1
summary(df$physhealth)

#rezygnacja z opieki zdrowotnej
df$medcost[df$medcost == 2] <- 0 #no access
df <- df[df$medcost !=7,] #dont know
df <- df[df$medcost !=9,] #refused
summary(df$medcost)

#aktywność fizyczna
df$physact[df$physact== 2] <- 0 #not
df <- df[df$physact !=7,] #dont know
df <- df[df$physact!=9,] #refused
summary(df$physact)

#nadciśnienie
df$highBP[df$highBP == 1] <- 0
df$highBP[df$highBP == 2] <- 1
df <- df[df$highBP !=9,] #refused
summary(df$highBP)

#cholesterol - tylko osoby które miały badany cholesterol
df$highchol[df$highchol == 1] <- 0
df$highchol[df$highchol == 2] <- 1 #not high cholesterol
df <- df[df$highchol !=9,] #refused
summary(df$highchol)
```

```
#wylew
df$stroke[df$stroke == 2] <- 0 #no stroke
df <- df[df$stroke !=7,] #dont know
df <- df[df$stroke !=9,] #refused
summary(df$stroke)

#choroba serca lub zawal
df$heartdis[df$heartdis== 2] <- 0 #not
df <- df[df$heartdis !=7,] #dont know
df <- df[df$heartdis!=9,] #refused
summary(df$heartdis)

#astma
df$asthma[df$asthma== 2] <- 0 #not
df <- df[df$asthma !=7,] #dont know
df <- df[df$asthma !=9,] #refused
summary(df$asthma)

#depresja
df$depression[df$depression == 2] <- 0 #no
df <- df[df$depression !=7,] #dont know
df <- df[df$depression !=9,] #refused
summary(df$depression)

#edukacja
df <- df[df$education !=9,] #refused
summary(df$education)

#zarobki
df <- df[df$income !=77,] #dont know
df <- df[df$income !=99,] #refused
summary(df$income)

#trudność w chodzeniu po schodach
df$diffwalk[df$diffwalk == 2] <- 0 #no
df <- df[df$diffwalk !=7,] #dont know
df <- df[df$diffwalk !=9,] #refused
summary(df$diffwalk)

#palenie papierosów
df <- df[df$smoker !=9,] #refused
summary(df$smoker)

#miasto czy wieś
df$urban[df$urban == 2] <- 0 #no
summary(df$urban)

#rasa

df$race[df$race == 1] <- "white"
df$race[df$race == 2] <- "black"
df$race[df$race == 3] <- "asian"
df$race[df$race == 4] <- "native"
```

```
df$race[df$race == 5] <- "hispanic"
df$race[df$race == 6] <- "other"
df$race <- as.factor(df$race)

#ubezpieczenie zdrowotne
df$insurance[df$insurance == 2] <- 0 #no
df <- df[df$insurance !=9,] #refused
summary(df$insurance)

#wiek
df <- df[df$age != 14,] #missing

 #bmi
summary(df$bmi)

#alkohol
df$alco[df$alco == 1] <- 0 #no
df$alco[df$alco == 2] <- 1 #no
df <- df[df$alco !=9,] #refused
summary(df$alco)

#owocki
df$fruit[df$fruit == 2] <- 0 #no
df <- df[df$fruit !=9,] #refused
summary(df$fruit)

#warzywa
df$veggie[df$veggie == 2] <- 0 #no
df <- df[df$veggie !=9,] #refused
summary(df$veggie)

#frytki
df$fries <- df$fries/100
df$fries[df$fries < 0.25] <- 0
df$fries[df$fries >= 0.25] <- 1
summary(df$fries)


#podział na zbiór treningowy i testowy
set.seed(3)
df$diabetes <- as.factor(df$diabetes)
inTraining <- createDataPartition(df$diabetes, p = .25, list = FALSE)
training <- df[ inTraining,]
testing  <- df[-inTraining,]
summary(training)

#walidacka krzyzowa
fitControl <- trainControl(
  method = "cv",
  number = 5)

barplot(prop.table(table(df$diabetes)),
        col = rainbow(4),
        ylim = c(0, 1),
```

```
        main = "Rozkład zmiennej objaśnianej")

#model logitowy
model_log <- glm(diabetes~., data = training, family = 'binomial')
summary(model_log)

#model logitowy + stepwise
model_step <- stepAIC(model_log, direction="both", trace = FALSE)
summary(model_step)


#drzewo
tree <- train(diabetes ~ ., data = training,
              method = "rpart",
              trControl = fitControl)
rpart.plot(tree$finalModel)

#las losowy
rf <- randomForest(diabetes ~.,
                   data = training)
varImpPlot(rf)

#wykres zmienna objaśniająca
ggplot(df,
       aes(x = as.factor(df$veggie), y = ..count.. /sum(..count..),
           fill = diabetes)) +
  geom_bar(position = "stack") +
  labs(x ="veggie", y ="Procent", title ="Rozkład zmiennej veggie") + scal
e_y_continuous(labels = scales::percent)

#macirrz korelacji
df2 <- dplyr::select_if(df, is.numeric)
r <- cor(df2, use="complete.obs")
ggcorrplot(r, hc.order = TRUE, type = "lower", lab = TRUE)

x<-table(predict(rf, new = testing), testing$diabetes)

### Regresje

cm1 <- table(ifelse(predict(model_log, newdata = testing, type = "response
") > 0.5, 1, 0), testing$diabetes)
confusionMatrix(cm1)
cm2 <- table(ifelse(predict(model_step, newdata = testing, type = "respons
e") > 0.5, 1, 0),testing$diabetes)
confusionMatrix(cm2)
### Drzewa
cm3 <- table(predict(tree, new = testing, type = "raw"), testing$diabetes)
confusionMatrix(cm3)
### Las
cm4 <- table(predict(rf, new = testing, type = "class"), testing$diabetes)
confusionMatrix(cm4)

#ROC
roc_log <- predict(model_log, newdata = testing, type = "response")
```

```
roc_log <- as.vector(roc_log)

roc_step <- predict(model_step, newdata = testing, type = "response")
roc_step <- as.vector(roc_step)

roc_tree <- predict(tree, newdata = testing, type = "prob")
roc_tree <- as.vector(roc_tree[,2])

roc_rf <- predict(rf, newdata = testing, type = "prob")
roc_rf <- as.vector(roc_rf[,2])

#krzywa ROC
pred <- prediction(roc_step, testing$diabetes)
perf <- performance(pred ,"tpr","fpr")
plot(perf, lwd=2, col ="blue")

pred2 <- prediction(roc_log, testing$diabetes)
perf2 <- performance(pred2 ,"tpr","fpr")
plot(perf2, lwd = 2, col ="green", add = TRUE)

pred3 <- prediction(roc_tree, testing$diabetes)
perf3 <- performance(pred3 ,"tpr","fpr")
plot(perf3, lwd = 2, col ="red", add = TRUE)

pred4 <- prediction(roc_rf, testing$diabetes)
perf4 <- performance(pred4 ,"tpr","fpr")
plot(perf4, lwd = 2, col ="violet", add = TRUE)

auc_log <- (performance(prediction(roc_log, testing$diabetes), "auc")@y.va
lues[[1]])
auc_step <- (performance(prediction(roc_step, testing$diabetes), "auc")@y.
values[[1]])
auc_tree <- (performance(prediction(roc_tree, testing$diabetes), "auc")@y.
values[[1]])
auc_rf <- (performance(prediction(roc_rf, testing$diabetes), "auc")@y.valu
es[[1]])

##drugi zbiór zbalansowany
df_balanced <- ovun.sample(diabetes~., data=df, method = "both",
                           p = 0.5,
                           seed = 222)$data

#podzial na zbior uczacy i testowy
set.seed(3)
df_balanced$diabetes <- as.factor(df_balanced$diabetes)
inTraining <- createDataPartition(df_balanced$diabetes, p = .25, list = FA
LSE)
training2 <- df_balanced[ inTraining,]
fitControl <- trainControl(
  method = "cv",
  number = 5)
testing2  <- df_balanced[-inTraining,]
summary(training2)
```

```
#wykres zmiennej objaśnianej
barplot(prop.table(table(training2$diabetes)),
        col = rainbow(4),
        ylim = c(0, 1),
        main = "Rozkład zmiennej objaśnianej")

#wykres objaśniajace
ggplot(df_balanced,
       aes(x = as.factor(veggie), y = ..count.. /sum(..count..),
           fill = diabetes)) +
  geom_bar(position = "stack") +
  labs(x ="veggie", y ="Procent", title ="Rozkład zmiennej veggie") + scal
e_y_continuous(labels = scales::percent)

#model logitowy
model_log <- glm(diabetes~., data = training2, family = 'binomial')
summary(model_log)

#model logitowy + stepwise
model_step <- stepAIC(model_log, direction="both", trace = FALSE)
summary(model_step)

#drzewo
tree <- train(diabetes ~ ., data = training2,
              method = "rpart",
              trControl = fitControl)
rpart.plot(tree$finalModel)

#las losowy
rf <- randomForest(diabetes ~.,
                   data = training2)
varImpPlot(rf)

#macierze bledow

### model logitowy
cm1 <- table(ifelse(predict(model_log, newdata = testing2, type = "respons
e") > 0.5, 1, 0), testing2$diabetes)
confusionMatrix(cm1)
cm2 <- table(ifelse(predict(model_step, newdata = testing2, type = "respon
se") > 0.5, 1, 0),testing2$diabetes)
confusionMatrix(cm2)
### Drzewa
cm3 <- table(predict(tree, new = testing2, type = "raw"), testing2$diabete
s)
confusionMatrix(cm3)
### Las
cm4 <- table(predict(rf, new = testing2, type = "class"), testing2$diabete
s)
confusionMatrix(cm4)

#ROC
roc_log <- predict(model_log, newdata = testing2, type = "response")
roc_log <- as.vector(roc_log)
```

```
roc_step <- predict(model_step, newdata = testing2, type = "response")
roc_step <- as.vector(roc_step)

roc_tree <- predict(tree, newdata = testing2, type = "prob")
roc_tree <- as.vector(roc_tree[,2])

roc_rf <- predict(rf, newdata = testing2, type = "prob")
roc_rf <- as.vector(roc_rf[,2])

pred <- prediction(roc_step, testing2$diabetes)
perf <- performance(pred ,"tpr","fpr")
plot(perf, lwd=2, col ="blue")

pred2 <- prediction(roc_log, testing2$diabetes)
perf2 <- performance(pred2 ,"tpr","fpr")
plot(perf2, lwd = 2, col ="green", add = TRUE)

pred3 <- prediction(roc_tree, testing2$diabetes)
perf3 <- performance(pred3 ,"tpr","fpr")
plot(perf3, lwd = 2, col ="red", add = TRUE)

pred4 <- prediction(roc_rf, testing2$diabetes)
perf4 <- performance(pred4 ,"tpr","fpr")
plot(perf4, lwd = 2, col ="violet", add = TRUE)

auc_log <- (performance(prediction(roc_log, testing2$diabetes), "auc")@y.v
alues[[1]])
auc_step <- (performance(prediction(roc_step, testing2$diabetes), "auc")@y
.values[[1]])
auc_tree <- (performance(prediction(roc_tree, testing2$diabetes), "auc")@y
.values[[1]])
auc_rf <- (performance(prediction(roc_rf, testing2$diabetes), "auc")@y.val
ues[[1]])
```

Załącznik 2: Skrypt w R dla zbioru trzeciego i czwartego

```
 library(mice)
 library(readr)
 library(tidyr)
 library(stats)
 library(dplyr)
 library(haven)
 library(tidyverse)
 library(caret)
 library(MASS)
 library(randomForest)
 library(ROCR)
 library(ROSE)
 library(ggplot2)
 library(rpart)
 library(rpart.plot)
 library(caret)
 library(ggcorrplot)
 library(ROCR)

 df2 <- read_xpt("dane_sas/2021.xpt", col_select = c(DIABETE4, SEXVAR, GEN
HLTH, PHYSHLTH,
                                                     MEDCOST1, EXERANY2,"_R
FHYPE6" ,"_RFCHOL3" , CVDSTRK3, "_MICHD", ASTHMA3,  ADDEPEV3,
                                                     EDUCA, INCOME3, DIFFWA
LK, "_SMOKER3","_URBSTAT", "_IMPRACE","_HLTHPLN","_AGEG5YR",
                                                     "_BMI5CAT","_RFDRHV7",
"_FRTLT1A", "_VEGLT1A", "FRNCHDA_"))
 df2 <- as.data.frame(df2)

 colnames(df2)[colnames(df2)=="DIABETE4"] = "diabetes"
 colnames(df2)[colnames(df2)=="SEXVAR"] = "sex"
 colnames(df2)[colnames(df2)=="GENHLTH"] = "genhealth"
 colnames(df2)[colnames(df2)=="PHYSHLTH"] = "physhealth"
 colnames(df2)[colnames(df2)=="MEDCOST1"] = "medcost"
 colnames(df2)[colnames(df2)=="EXERANY2"] = "physact"
 colnames(df2)[colnames(df2)=="_RFHYPE6"] = "highBP"
 colnames(df2)[colnames(df2)=="_RFCHOL3"] = "highchol"
 colnames(df2)[colnames(df2)=="CVDSTRK3"] = "stroke"
 colnames(df2)[colnames(df2)=="_MICHD"] = "heartdis"
 colnames(df2)[colnames(df2)=="ASTHMA3"] = "asthma"
 colnames(df2)[colnames(df2)=="ADDEPEV3"] = "depression"
 colnames(df2)[colnames(df2)=="EDUCA"] = "education"
 colnames(df2)[colnames(df2)=="INCOME3"] = "income"
 colnames(df2)[colnames(df2)=="DIFFWALK"] = "diffwalk"
 colnames(df2)[colnames(df2)=="_SMOKER3"] = "smoker"
 colnames(df2)[colnames(df2)=="_URBSTAT"] = "urban"
 colnames(df2)[colnames(df2)=="_IMPRACE"] = "race"
 colnames(df2)[colnames(df2)=="_HLTHPLN"] = "insurance"
 colnames(df2)[colnames(df2)=="_AGEG5YR"] = "age"
 colnames(df2)[colnames(df2)=="_BMI5CAT"] = "bmi"
 colnames(df2)[colnames(df2)=="_RFDRHV7"] = "alco"
 colnames(df2)[colnames(df2)=="_FRTLT1A"] = "fruit"
```

```
 colnames(df2)[colnames(df2)=="_VEGLT1A"] = "veggie"
 colnames(df2)[colnames(df2)=="FRNCHDA_"] = "fries"

 #cukrzyca
 df2$diabetes[df2$diabetes == 7] <- NA
 df2$diabetes[df2$diabetes == 9] <- NA
 df2$diabetes[df2$diabetes == 4] <- NA
 df2$diabetes[df2$diabetes == 3] <- 0 #in pregnancy
 df2$diabetes[df2$diabetes == 2] <- 0 #no diabetes
 df2$diabetes <- as.factor(df2$diabetes)
 summary(df2$diabetes)

 #płec
 df2$sex[df2$sex == 7] <- NA
 df2$sex[df2$sex == 9] <- NA
 df2$sex[df2$sex == 2] <- 0 #kobieta
 df2$sex <- as.factor(df2$sex)
 summary(df2$sex)

 finished_imputed_data$sex <- as.numeric(finished_imputed_data$sex)
 finished_imputed_data$sex[finished_imputed_data$sex == 1] <- 0
 finished_imputed_data$sex[finished_imputed_data$sex == 2] <- 1
 summary(finished_imputed_data$sex)

 #generalny stan zdrowia
 df2$genhealth[df2$genhealth == 7] <- NA
 df2$genhealth[df2$genhealth == 9] <- NA
 df2$genhealth <- as.factor(df2$genhealth)
 finished_imputed_data$genhealth <- as.numeric(finished_imputed_data$genhe
alth)
 summary(finished_imputed_data$genhealth)

 #zdrowie fizyczne
 df2$physhealth[df2$physhealth == 77] <- NA
 df2$physhealth[df2$physhealth == 99] <- NA
 df2$physhealth[df2$physhealth == 88] <- 0
 df2$physhealth[df2$physhealth != 0 & !is.na(df2$physhealth)] <- 1
 df2$physhealth <- as.factor(df2$physhealth)
 summary(df2$physhealth)

 finished_imputed_data$physhealth <- as.numeric(finished_imputed_data$phys
health)
 finished_imputed_data$physhealth[finished_imputed_data$physhealth == 1] <
- 0
 finished_imputed_data$physhealth[finished_imputed_data$physhealth == 2] <
- 1
 summary(finished_imputed_data$physhealth)

 #rezygnacja z lekarza z powodu kosztu
 df2$medcost[df2$medcost == 7] <- NA
 df2$medcost[df2$medcost == 9] <- NA
 df2$medcost[df2$medcost == 2] <- 0
 df2$medcost <- as.factor(df2$medcost)
 summary(df2$medcost)
```

```
 finished_imputed_data$medcost <- as.numeric(finished_imputed_data$medcost
)
 finished_imputed_data$medcost[finished_imputed_data$medcost == 1] <- 0
 finished_imputed_data$medcost[finished_imputed_data$medcost == 2] <- 1
 summary(finished_imputed_data$medcost)

 #aktywność fizyczna
 df2$physact[df2$physact == 7] <- NA
 df2$physact[df2$physact == 9] <- NA
 df2$physact[df2$physact == 2] <- 0
 df2$physact <- as.factor(df2$physact)

 finished_imputed_data$physact <- as.numeric(finished_imputed_data$physact
)
 finished_imputed_data$physact[finished_imputed_data$physact == 1] <- 0
 finished_imputed_data$physact[finished_imputed_data$physact == 2] <- 1
 summary(finished_imputed_data$physact)

 #wysokie ciśnienie
 df2$highBP[df2$highBP == 7] <- NA
 df2$highBP[df2$highBP == 9] <- NA
 df2$highBP[df2$highBP == 1] <- 0
 df2$highBP[df2$highBP == 2] <- 1
 df2$highBP <- as.factor(df2$highBP)

 finished_imputed_data$highBP <- as.numeric(finished_imputed_data$highBP)
 finished_imputed_data$highBP[finished_imputed_data$highBP == 1] <- 0
 finished_imputed_data$highBP[finished_imputed_data$highBP == 2] <- 1
 summary(finished_imputed_data$highBP)

 #wyoski cholesterol
 df2$highchol[df2$highchol == 7] <- NA
 df2$highchol[df2$highchol == 9] <- NA
 df2$highchol[df2$highchol == 1] <- 0
 df2$highchol[df2$highchol == 2] <- 1
 df2$highchol <- as.factor(df2$highchol)

 finished_imputed_data$highchol <- as.numeric(finished_imputed_data$highch
ol)
 finished_imputed_data$highchol[finished_imputed_data$highchol == 1] <- 0
 finished_imputed_data$highchol[finished_imputed_data$highchol == 2] <- 1
 summary(finished_imputed_data$highchol)

 #wylew
 df2$stroke[df2$stroke == 7] <- NA
 df2$stroke[df2$stroke == 9] <- NA
 df2$stroke[df2$stroke == 2] <- 0
 df2$stroke <- as.factor(df2$stroke)

 finished_imputed_data$stroke <- as.numeric(finished_imputed_data$stroke)
 finished_imputed_data$stroke[finished_imputed_data$stroke == 1] <- 0
 finished_imputed_data$stroke[finished_imputed_data$stroke == 2] <- 1
 summary(finished_imputed_data$stroke)
```

```
 #choroba serca lub zawał
 df2$heartdis[df2$heartdis == 7] <- NA
 df2$heartdis[df2$heartdis == 9] <- NA
 df2$heartdis[df2$heartdis == 2] <- 0
 df2$heartdis <- as.factor(df2$heartdis)

 finished_imputed_data$heartdis <- as.numeric(finished_imputed_data$heartd
is)
 finished_imputed_data$heartdis[finished_imputed_data$heartdis == 1] <- 0
 finished_imputed_data$heartdis[finished_imputed_data$heartdis == 2] <- 1
 summary(finished_imputed_data$heartdis)

 #astma
 df2$asthma[df2$asthma == 7] <- NA
 df2$asthma[df2$asthma == 9] <- NA
 df2$asthma[df2$asthma == 2] <- 0
 df2$asthma <- as.factor(df2$asthma)

 finished_imputed_data$asthma <- as.numeric(finished_imputed_data$asthma)
 finished_imputed_data$asthma[finished_imputed_data$asthma == 1] <- 0
 finished_imputed_data$asthma[finished_imputed_data$asthma == 2] <- 1
 summary(finished_imputed_data$asthma)

 #depresja
 df2$depression[df2$depression == 7] <- NA
 df2$depression[df2$depression == 9] <- NA
 df2$depression[df2$depression == 2] <- 0
 df2$depression <- as.factor(df2$depression)

 finished_imputed_data$depression <- as.numeric(finished_imputed_data$depr
ession)
 finished_imputed_data$depression[finished_imputed_data$depression == 1] <
- 0
 finished_imputed_data$depression[finished_imputed_data$depression == 2] <
- 1
 summary(finished_imputed_data$depression)

 #edukacja
 df2$education[df2$education == 9] <- NA
 df2$education <- as.factor(df2$education)

 finished_imputed_data$education <- as.numeric(finished_imputed_data$educa
tion)
 summary(finished_imputed_data$education)

 #zarobki
 df2$income[df2$income == 77] <- NA
 df2$income[df2$income == 99] <- NA
 df2$income <- as.factor(df2$income)

 finished_imputed_data$income <- as.numeric(finished_imputed_data$income)
 summary(finished_imputed_data$income)
```

```
 #trudność w poruszaniu się
 df2$diffwalk[df2$diffwalk == 7] <- NA
 df2$diffwalk[df2$diffwalk == 9] <- NA
 df2$diffwalk[df2$diffwalk == 2] <- 0
 df2$diffwalk <- as.factor(df2$diffwalk)

 finished_imputed_data$diffwalk <- as.numeric(finished_imputed_data$diffwa
lk)
 finished_imputed_data$diffwalk[finished_imputed_data$diffwalk == 1] <- 0
 finished_imputed_data$diffwalk[finished_imputed_data$diffwalk == 2] <- 1
 summary(finished_imputed_data$diffwalk)

 #palenie papierosów
 df2$smoker[df2$smoker == 9] <- NA
 df2$smoker <- as.factor(df2$smoker)

 finished_imputed_data$smoker <- as.numeric(finished_imputed_data$smoker)
 summary(finished_imputed_data$smoker)

 #miasto czy wieś
 df2$urban[df2$urban == 2] <- 0
 df2$urban <- as.factor(df2$urban)

 finished_imputed_data$urban <- as.numeric(finished_imputed_data$urban)
 finished_imputed_data$urban[finished_imputed_data$urban == 1] <- 0
 finished_imputed_data$urban[finished_imputed_data$urban == 2] <- 1
 summary(finished_imputed_data$urban)

 #ubezpieczenie
 df2$insurance[df2$insurance == 9] <- NA
 df2$insurance[df2$insurance == 2] <- 0
 df2$insurance <- as.factor(df2$insurance)

 finished_imputed_data$insurance <- as.numeric(finished_imputed_data$insur
ance)
 finished_imputed_data$insurance[finished_imputed_data$insurance == 1] <-
0
 finished_imputed_data$insurance[finished_imputed_data$insurance == 2] <-
1
 summary(finished_imputed_data$insurance)

 #rasa
 df2$race <- as.factor(df2$race)

 #wiek
 df2$age[df2$age == 14] <- NA
 df2$age <- as.factor(df2$age)

 finished_imputed_data$age <- as.numeric(finished_imputed_data$age)
 summary(finished_imputed_data$age)

 #alkohol
 df2$alco[df2$alco == 1] <- 0
 df2$alco[df2$alco == 2] <- 1
```

```
df2$alco[df2$alco == 9] <- NA
df2$alco <- as.factor(df2$alco)

finished_imputed_data$alco <- as.numeric(finished_imputed_data$alco)
finished_imputed_data$alco[finished_imputed_data$alco == 1] <- 0
finished_imputed_data$alco[finished_imputed_data$alco == 2] <- 1
summary(finished_imputed_data$alco)

#frytki
df2$fries[df2$fries < 25] <- 0
df2$fries[df2$fries >= 25] <- 1
df2$fries <- as.factor(df2$fries)

finished_imputed_data$fries <- as.numeric(finished_imputed_data$fries)
finished_imputed_data$fries[finished_imputed_data$fries == 1] <- 0
finished_imputed_data$fries[finished_imputed_data$fries == 2] <- 1
summary(finished_imputed_data$fries)

#owoce
df2$fruit[df2$fruit == 2] <- 0
df2$fruit[df2$fruit == 9] <- NA
df2$fruit <- as.factor(df2$fruit)

finished_imputed_data$fruit <- as.numeric(finished_imputed_data$fruit)
finished_imputed_data$fruit[finished_imputed_data$fruit == 1] <- 0
finished_imputed_data$fruit[finished_imputed_data$fruit == 2] <- 1
summary(finished_imputed_data$fruit)

#warzywa
df2$veggie[df2$veggie == 2] <- 0
df2$veggie[df2$veggie == 9] <- NA
df2$veggie <- as.factor(df2$veggie)

finished_imputed_data$veggie <- as.numeric(finished_imputed_data$veggie)
finished_imputed_data$veggie[finished_imputed_data$veggie == 1] <- 0
finished_imputed_data$veggie[finished_imputed_data$veggie == 2] <- 1
summary(finished_imputed_data$veggie)

#bmi
df2$bmi <- as.factor(df2$bmi)
summary(df2)

finished_imputed_data$bmi <- as.numeric(finished_imputed_data$bmi)
summary(finished_imputed_data$bmi)

#race
finished_imputed_data$race <- as.numeric(finished_imputed_data$race)
summary(finished_imputed_data$race)
finished_imputed_data$race[finished_imputed_data$race == 1] <- "white"
finished_imputed_data$race[finished_imputed_data$race == 2] <- "black"
finished_imputed_data$race[finished_imputed_data$race == 3] <- "asian"
finished_imputed_data$race[finished_imputed_data$race == 4] <- "native"
finished_imputed_data$race[finished_imputed_data$race == 5] <- "hispanic"
finished_imputed_data$race[finished_imputed_data$race == 6] <- "other"
```

```
 imputed_data <- mice(df2, m=2, defaultMethod = c("pmm", "logreg", "polyre
g", "polr"))
 summary(imputed_data)
 finished_imputed_data <- complete(imputed_data, 2:25)
 summary(finished_imputed_data)

 write_csv(finished_imputed_data,"/Users/paulinaskalik/dane_licencjat/mice
.csv")

 #podzial na zbiory
 set.seed(3)
 inTraining <- createDataPartition(finished_imputed_data$diabetes, p = .25
, list = FALSE)
 training3 <- finished_imputed_data[ inTraining,]
 testing3  <- finished_imputed_data[-inTraining,]
 fitControl <- trainControl(
         method = "cv",
         number = 5)


 #rozkład zmiennej objaśnianej
 barplot(prop.table(table(finished_imputed_data$diabetes)),
         col = rainbow(4),
         ylim = c(0, 1),
         main = "Rozkład zmiennej objaśnianej")

 finished_imputed_data_2 <- dplyr::select_if(finished_imputed_data, is.num
eric)
 r <- cor(finished_imputed_data_2, use="complete.obs")
 ggcorrplot(r, hc.order = TRUE, type = "lower", lab = TRUE)


 #wykres zmiennych objaśniających
 ggplot(fiished_imputed_data,
        aes(x = as.factor(veggie), y = ..count.. /sum(..count..),
            fill = diabetes)) +
         geom_bar(position = "stack") +
         labs(x ="veggie", y ="Procent", title ="Rozkład zmiennej veggie")
+ scale_y_continuous(labels = scales::percent)

 #model logitowy
 model_log <- glm(diabetes~., data = training3, family = 'binomial')
 summary(model_log)

 #stepwise
 model_step <- stepAIC(model_log, direction="both", trace = FALSE)
 summary(model_step)

 #drzewo z alco
 tree <- train(diabetes ~ ., data = training3,
               method = "rpart",
               trControl = fitControl)
```

```
 rpart.plot(tree$finalModel)

 #drzewo bez alco
 training_ex <- training3
 training_ex[c("alco")] <- list(NULL)

 tree2 <- train(diabetes ~ ., data = training_ex,
                method = "rpart",
                trControl = fitControl)
 rpart.plot(tree2$finalModel)


 #las losowy
 rf <- randomForest(diabetes ~.,
                    data = training3)
 varImpPlot(rf)

 #macierze błędów

 #modele logitowe
 cm1 <- table(ifelse(predict(model_log, newdata = testing3, type = "respon
se") > 0.5, 1, 0), testing3$diabetes)
 confusionMatrix(cm1)
 cm2 <- table(ifelse(predict(model_step, newdata = testing3, type = "respo
nse") > 0.5, 1, 0),testing3$diabetes)
 confusionMatrix(cm2)
 ### Drzewa
 cm3 <- table(predict(tree, new = testing3, type = "raw"), testing3$diabet
es)
 confusionMatrix(cm3)

 cm5 <- table(predict(tree2, new = testing3, type = "raw"), testing3$diabe
tes)
 confusionMatrix(cm5)
 ### Las
 cm4 <- table(predict(rf, new = testing3, type = "class"), testing3$diabet
es)
 confusionMatrix(cm4)

 #ROC
 roc_log <- predict(model_log, newdata = testing3, type = "response")
 roc_log <- as.vector(roc_log)

 roc_step <- predict(model_step, newdata = testing3, type = "response")
 roc_step <- as.vector(roc_step)

 roc_tree <- predict(tree, newdata = testing3, type = "prob")
 roc_tree <- as.vector(roc_tree[,2])

 roc_rf <- predict(rf, newdata = testing3, type = "prob")
 roc_rf <- as.vector(roc_rf[,2])

 pred <- prediction(roc_step, testing3$diabetes)
 perf <- performance(pred ,"tpr","fpr")
```

```
 plot(perf, lwd=2, col ="blue")

 pred2 <- prediction(roc_log, testing3$diabetes)
 perf2 <- performance(pred2 ,"tpr","fpr")
 plot(perf2, lwd = 2, col ="green", add = TRUE)

 pred3 <- prediction(roc_tree, testing3$diabetes)
 perf3 <- performance(pred3 ,"tpr","fpr")
 plot(perf3, lwd = 2, col ="red", add = TRUE)

 pred4 <- prediction(roc_rf, testing3$diabetes)
 perf4 <- performance(pred4 ,"tpr","fpr")
 plot(perf4, lwd = 2, col ="violet", add = TRUE)


 auc_log <- (performance(prediction(roc_log, testing3$diabetes), "auc")@y.
values[[1]])
 auc_step <- (performance(prediction(roc_step, testing3$diabetes), "auc")@
y.values[[1]])
 auc_tree <- (performance(prediction(roc_tree, testing3$diabetes), "auc")@
y.values[[1]])
 auc_rf <- (performance(prediction(roc_rf, testing3$diabetes), "auc")@y.va
lues[[1]])


##zbiór zbalansowany
 mice_balanced <- ovun.sample(diabetes~., data=finished_imputed_data, meth
od = "both",
                              p = 0.5,
                              seed = 222)$data

 #podzial na zbior testowy
 set.seed(3)
 mice_balanced$diabetes <- as.factor(mice_balanced$diabetes)
 inTraining <- createDataPartition(mice_balanced$diabetes, p = .25, list =
FALSE)
 training4 <- mice_balanced[ inTraining,]
 fitControl <- trainControl(
         method = "cv",
         number = 5)
 testing4  <- mice_balanced[-inTraining,]
 summary(training4)

 #wykres zmienna objaśniająca
 ggplot(mice_balanced,
        aes(x = as.factor(bmi), y = ..count.. /sum(..count..),
            fill = diabetes)) +
         geom_bar(position = "stack") +
         labs(x ="bmi", y ="Procent", title ="Rozkład zmiennej bmi") + sca
le_y_continuous(labels = scales::percent)

 #model logitowy
 model_log <- glm(diabetes~., data = training4, family = 'binomial')
 summary(model_log)
```

```
#stepwise
model_step <- stepAIC(model_log, direction="both", trace = FALSE)
summary(model_step)

#drzewo
tree <- train(diabetes ~ ., data = training4,
              method = "rpart",
              trControl = fitControl)
rpart.plot(tree$finalModel)

#las losowy
rf <- randomForest(diabetes ~.,
                   data = training4)
varImpPlot(rf)

#modele logitowe
cm1 <- table(ifelse(predict(model_log, newdata = testing4, type = "respon
se") > 0.5, 1, 0), testing4$diabetes)
confusionMatrix(cm1)
cm2 <- table(ifelse(predict(model_step, newdata = testing4, type = "respo
nse") > 0.5, 1, 0),testing4$diabetes)
confusionMatrix(cm2)
### Drzewa
cm3 <- table(predict(tree, new = testing4, type = "raw"), testing4$diabet
es)
confusionMatrix(cm3)
### Las
cm4 <- table(predict(rf, new = testing4, type = "class"), testing4$diabet
es)
confusionMatrix(cm4)

#ROC
roc_log <- predict(model_log, newdata = testing4, type = "response")
roc_log <- as.vector(roc_log)

roc_step <- predict(model_step, newdata = testing4, type = "response")
roc_step <- as.vector(roc_step)

roc_tree <- predict(tree, newdata = testing4, type = "prob")
roc_tree <- as.vector(roc_tree[,2])

roc_rf <- predict(rf, newdata = testing4, type = "prob")
roc_rf <- as.vector(roc_rf[,2])

pred <- prediction(roc_step, testing4$diabetes)
perf <- performance(pred ,"tpr","fpr")
plot(perf, lwd=2, col ="blue")

pred2 <- prediction(roc_log, testing4$diabetes)
perf2 <- performance(pred2 ,"tpr","fpr")
plot(perf2, lwd = 2, col ="green", add = TRUE)

pred3 <- prediction(roc_tree, testing4$diabetes)
```

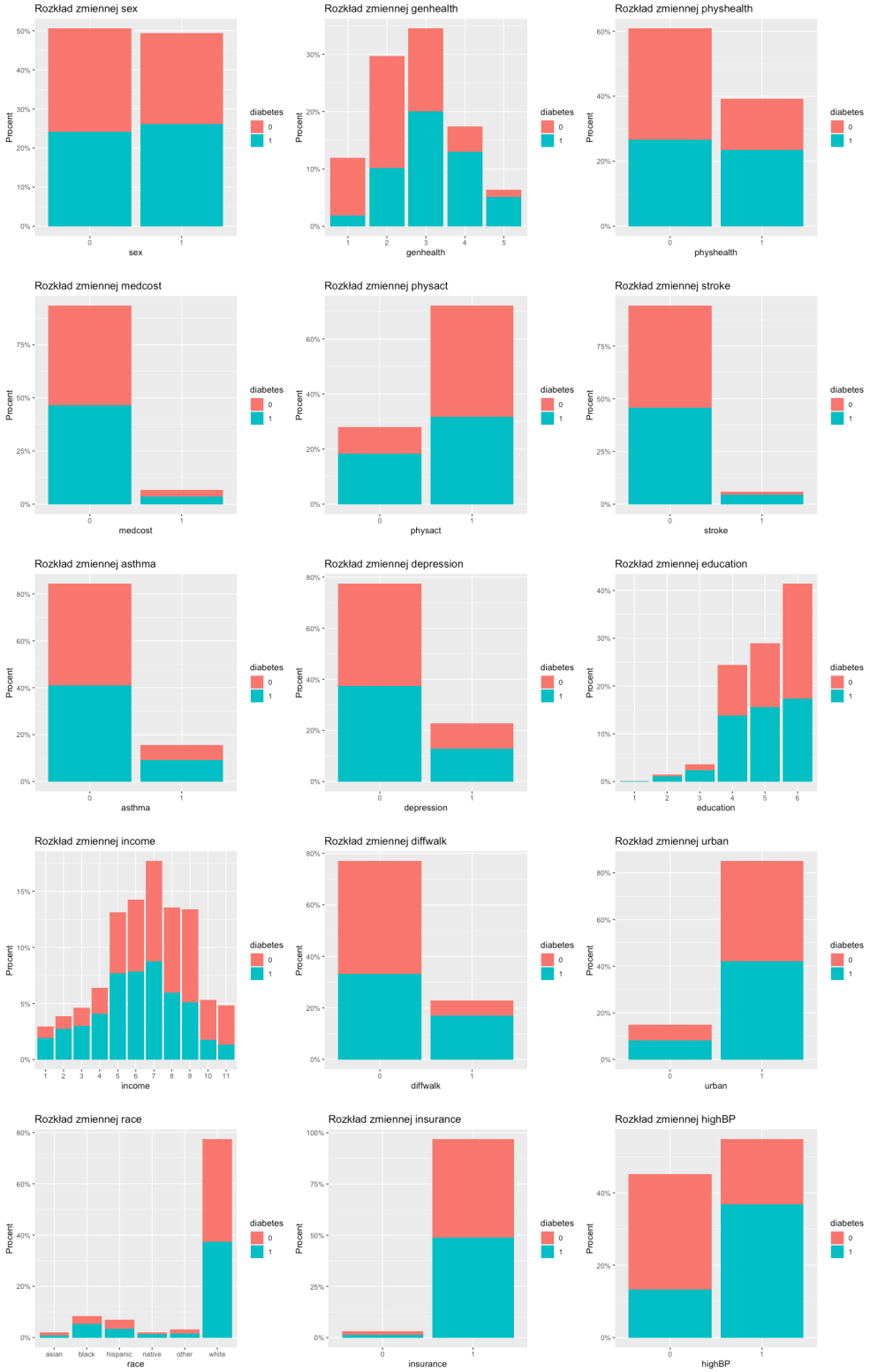
```
 perf3 <- performance(pred3 ,"tpr","fpr")
 plot(perf3, lwd = 2, col ="red", add = TRUE)


 pred4 <- prediction(roc_rf, testing4$diabetes)
 perf4 <- performance(pred4 ,"tpr","fpr")
 plot(perf4, lwd = 2, col ="violet", add = TRUE)


 auc_log <- (performance(prediction(roc_log, testing4$diabetes), "auc")@y.
values[[1]])
 auc_step <- (performance(prediction(roc_step, testing4$diabetes), "auc")@
y.values[[1]])
 auc_tree <- (performance(prediction(roc_tree, testing4$diabetes), "auc")@
y.values[[1]])
 auc_rf <- (performance(prediction(roc_rf, testing4$diabetes), "auc")@y.va
lues[[1]])
```

Załącznik 3: Rozkłady zmiennych objaśniających

1. Rozkład zmiennych objaśnianych w pierwszym zbiorze

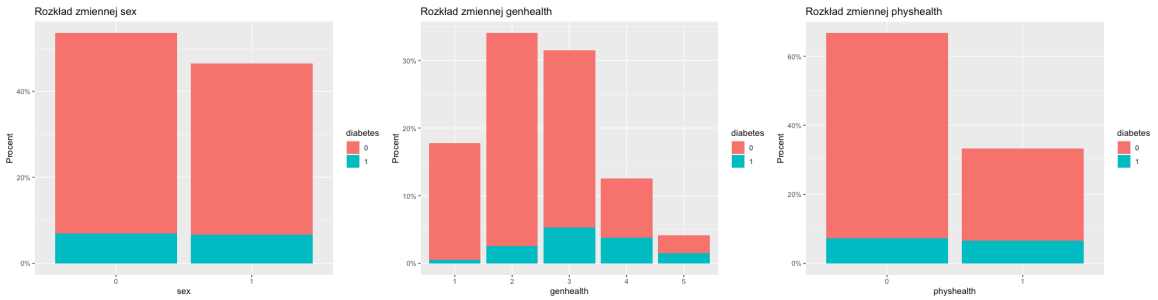
Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

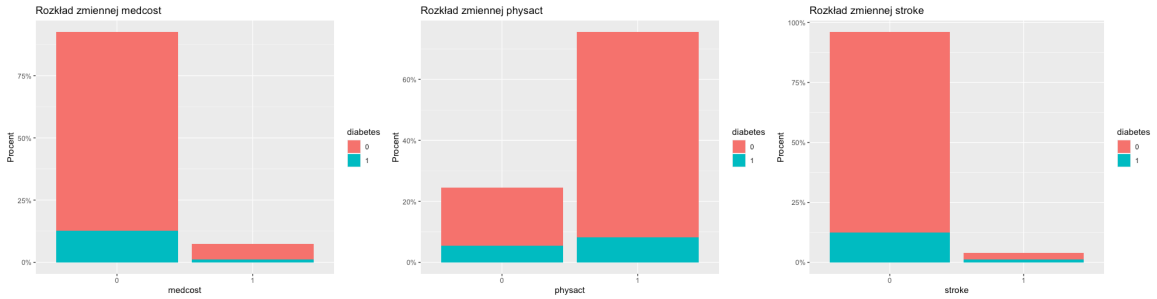
2. Rozkład zmiennych objaśnianych w drugim zbiorze



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

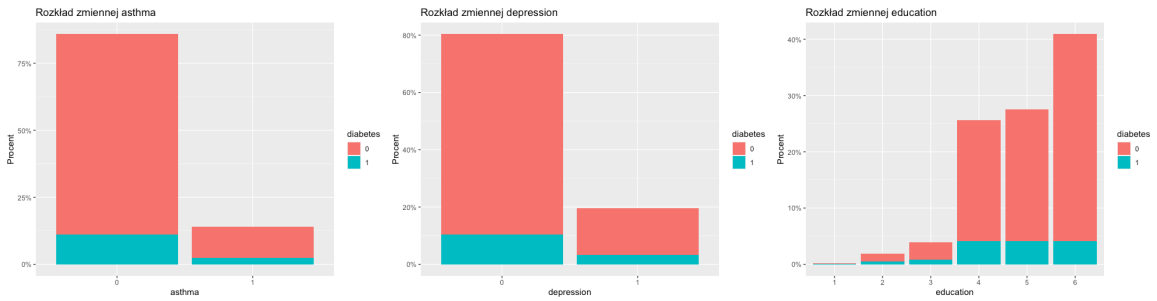
3. Rozkład zmiennych objaśnianych w trzecim zbiorze




Rozkład zmiennej medcost
Procent
medcost
diabetes
0
1

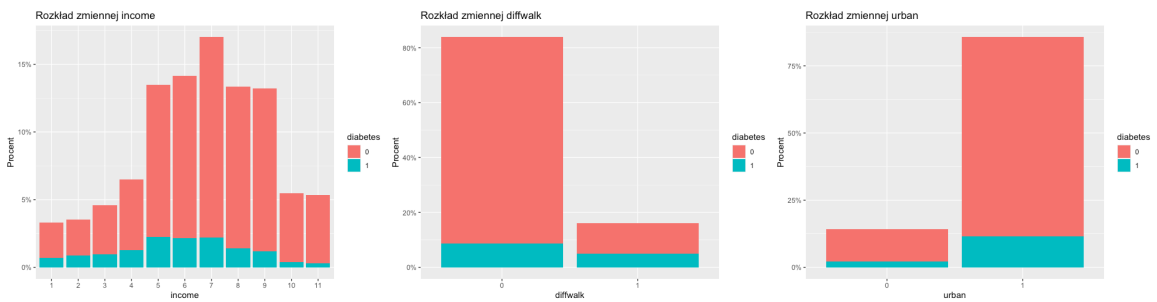
Rozkład zmiennej physact
Procent
physact
diabetes
0
1

Rozkład zmiennej stroke
Procent
stroke
diabetes
0
1


Rozkład zmiennej asthma
Procent
asthma
diabetes
0
1

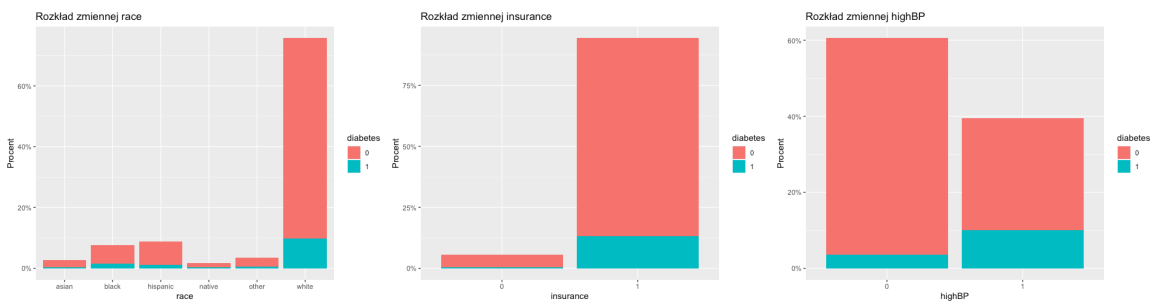
Rozkład zmiennej depression
Procent
depression
diabetes
0
1

Rozkład zmiennej education
Procent
education
diabetes
0
1


Rozkład zmiennej income
Procent
income
diabetes
0
1

Rozkład zmiennej diffwalk
Procent
diffwalk
diabetes
0
1

Rozkład zmiennej urban
Procent
urban
diabetes
0
1

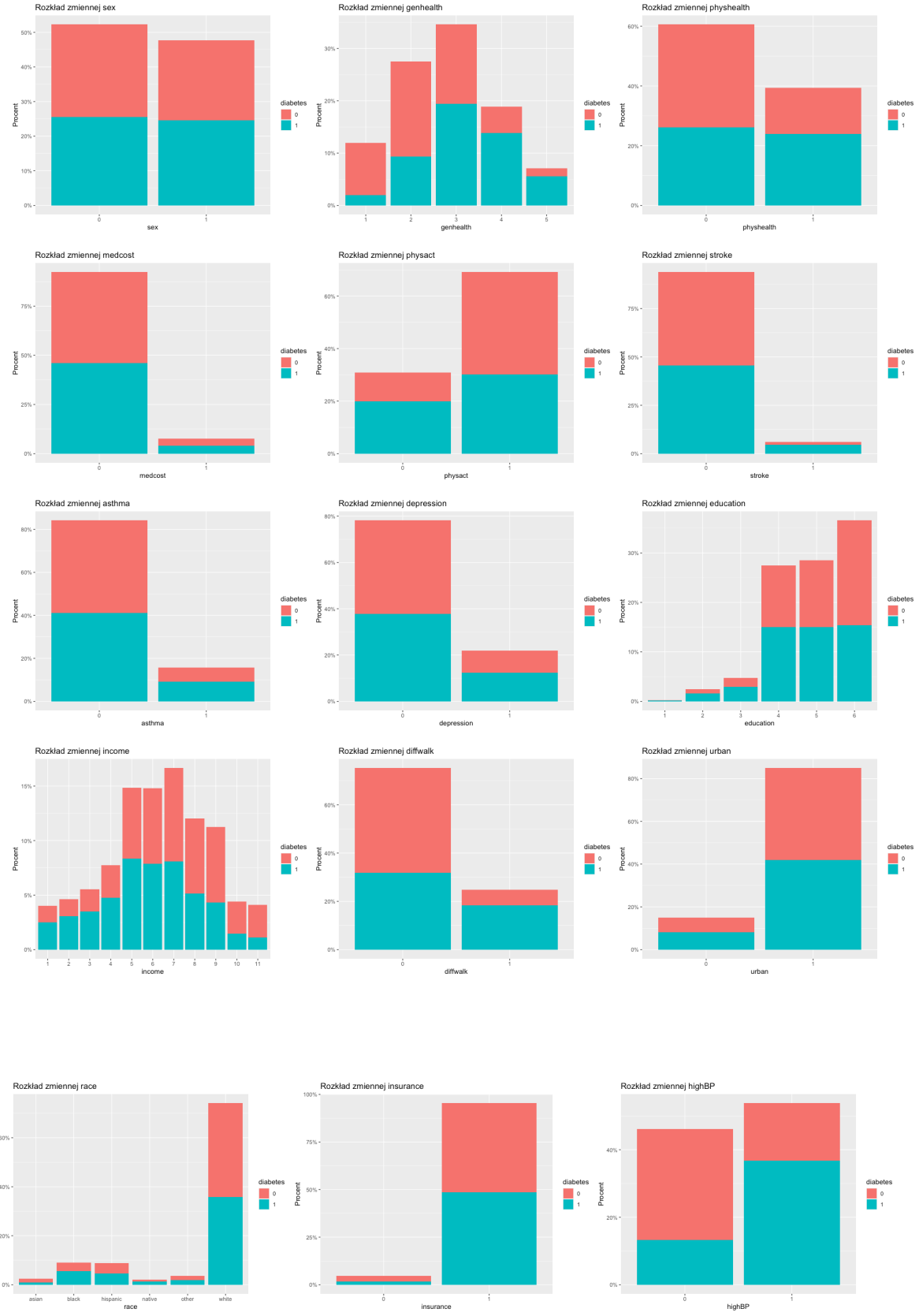

Rozkład zmiennej race
Procent
asian
black
hispanic
native
other
white
race
diabetes
0
1

Rozkład zmiennej insurance
Procent
insurance
diabetes
0
1

Rozkład zmiennej highBP
Procent
highBP
diabetes
0
1

Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

4. Rozkład zmiennych objaśnianych w czwartym zbiorze



Źródło: Opracowanie własne na podstawie wyników ankiety BRFSS

Streszczenie:

Przedmiotem pracy jest prognozowanie ryzyka zachorowania na cukrzycę typu II z wykorzystaniem różnych modeli uczenia maszynowego i języka R. Analizy przeprowadzane są na zbiorze danych pochodzących z ankiety dotyczącej stanu zdrowia amerykańskich obywateli przeprowadzonej przez amerykańską agencję rządową Center of Disease Control w 2021 roku. Pierwszy rozdział przedstawia obraz kliniczny cukrzycy, jej epidemiologię, koszty ekonomiczne oraz dotychczasowe badania prowadzone w omawianym zakresie. Drugi rozdział dotyczy metodologii budowy modeli oraz metod ewaluacji ich jakości. Trzeci rozdział skupia się na interpretacji i ewaluacji powstałych modeli. Do budowy modeli zostały wykorzystane cztery różne zbiory danych, różniących się sposobami eliminacji braków danych oraz poziomem zbalansowania zbioru pod względem klas zmiennej objaśnianej. Wykorzystanie imputacji danych za pomocą biblioteki *mice* wpłynęło pozytywnie na jakość modeli. Pod względem dokładności najlepszym modelem okazał się las losowy zbudowany na zbiorze trzecim ze skutecznością w rozpoznawaniu cukrzycy na poziomie 98%. Pod względem kryterium AUC najlepszy model został uzyskany na zbiorze czwartym i był to las losowy.

Słowa kluczowe: uczenie maszynowe, imputacja danych, model logitowy, drzewo decyzyjne, las losowy, metody ilościowe, cukrzyca