Porównanie modeli klasyfikacyjnych w analizie czynników wpływających na otyłość

Uniwersytet Ekonomiczny w Katowicach

Kosarski Ksawery

Inżynieria Procesu Odkrywania Wiedzy

Spis treści

Cele badawcze i metodyka

Celem badawczym projektu jest zidentyfikowanie, które z wybranych modeli klasyfikacyjnych – automatyczna sieć neuronowa czy drzewo decyzyjne – wykazuje większą skuteczność w predykcji poziomu otyłości w analizowanym zbiorze danych. Projekt dąży do oceny efektywności tych modeli w dwóch różnych podejściach: predykcji zmiennej porządkowej, gdzie otyłość przyjmuje kilka poziomów oraz zmiennej zredukowanej do postaci binarnej. Dodatkowym celem jest zbadanie, które zmienne mają największy wpływ na ostateczny wynik predykcji, co pozwoli na lepsze zrozumienie czynników związanych z otyłością.

Projekt wykonano zgodnie z metodyką SEMMA która jest systematycznym podejściem do analizy danych i budowy modeli. Pierwszy etap, "Sample", polega na wybraniu reprezentatywnej próbki danych. Drugi etap, "Explore", obejmuje eksploracyjną analizę danych, mającą na celu zrozumienie struktury zbioru, identyfikację wzorców, zależności oraz wykrycie potencjalnych problemów, takich jak wartości odstające czy asymetria rozkładów. Etap "Modify" to modyfikacja danych, która może obejmować czyszczenie zbioru, przekształcenia zmiennych, usunięcie obserwacji odstających oraz tworzenie nowych cech. Celem tego kroku jest optymalne dostosowanie danych do wymagań modeli predykcyjnych. Kolejny etap, "Model", polega na budowie różnych modeli statystycznych i uczenia maszynowego w celu przewidywania wartości zmiennej docelowej. W tym kroku wykorzystuje się dane treningowe, a modele są następnie dostrajane przy użyciu zbiorów walidacyjnych. Ostatni etap, "Assess", polega na ocenie jakości modeli, w tym ich zdolności do generalizacji, przy użyciu zbioru testowego.

Omówienie badanych danych

Dane pochodzą z UCI Machine Learning Repository. Zbiór danych został stworzony przez Fabio Mendozę Palechora i Alexisa De la Hoz Manotasa w 2019 roku w ramach badania nad czynnikami wpływającymi na otyłość. Szczegóły dotyczące tego zbioru danych, jak również badania, które posłużyło do jego utworzenia, zostały opublikowane w czasopiśmie Data in Brief. Zbiór danych zawiera informacje na temat 2111 osób, a celem badań było zidentyfikowanie czynników związanych z otyłością. Dane obejmują szereg zmiennych demograficznych, behawioralnych oraz dotyczących stylu życia, co pozwala na wieloaspektową analizę tego problemu zdrowotnego.

Zmiennie znajdujące się w bazie:

- Gender – Płeć osoby badanej;
- Age – Wiek osoby badanej;
- Height – Wzrost osoby badanej;
- Weight – Waga osoby badanej;
- family_history_with_overweight – Czy ktoś z rodziny osoby badanej ma bądź miał nadwagę;
- FAVC – Czy osoba badana spożywa często wysoko kaloryczne posiłki;
- FCVC – Częstotliwość spożywania warzyw w ciągu dnia przez osobę badaną;
- NCP – Ilość spożywanych głównych posiłków dziennie;
- CAEC – Częstość spożywania przekąsek między posiłkami;
- SMOKE – Czy osoba badana jest palaczem;
- CH2O – Dziennie spożycie wody;
- SCC – Czy osoba badana kontroluje spożycie kalorii;
- FAF – Aktywność fizyczna badanej osoby (godziny tygodniowo);
- TUE – Czas przed ekranem osoby badanej (godziny dziennie);
- CALC – Spożycie alkoholu;
- MTRANS – Preferowany środek transportu osoby badanej;
- NObesity – Poziom otyłości (Niedowaga, Waga normalna, Nadwaga poziom I, Nadwaga poziom II, Otyłość typu I, Otyłość typu II, Otyłość typu III).

Opis procesu przygotowania danych do analiz

Dane są dostępne w pliku CSV, który został załadowany do programu SAS Enterprise Miner za pomocą wewnętrznego węzła „Import pliku”. Następnie dane zostały przekształcone za pomocą kodu SAS-owego, który wykonuje szereg operacji mających na celu przygotowanie danych do dalszej analizy:

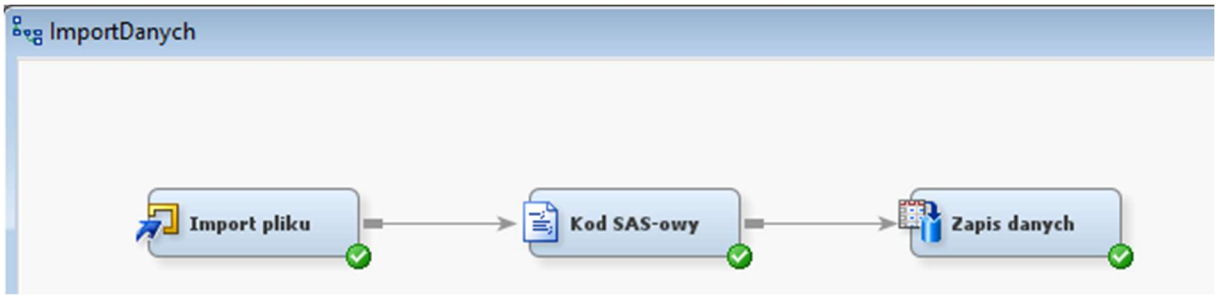
- Zmiennym nadano bardziej czytelne nazwy
- Zmienne liczbowe zostały zaokrąglone, ponieważ niektóre z nich były wygenerowane komputerowo i wymagały zaokrąglenia do realistycznych wartości.
- Zmienna płeć oraz preferowany środek transportu zostały zakodowane za pomocą kodowania one-hot, ponieważ są zmiennymi nominalnymi, które nie mają różnej wagi między wartościami.
- Inne zmienne kategoryczne, takie jak spożycie alkoholu czy częstość spożywania przekąsek, zostały przekształcone na zmienne liczbowe.
- Dodatkowo zmienne które przyjmowały wartości ‘no’ i ‘yes’ zostały przekształcone na zmienne binarne.

Na końcu oryginalne zmienne zostały usunięte.

```
data EMWS1.EMCODE_TRAIN;
set EMWS1.FIMPORT_train;

wiek = round(Age);
wzrost = round(Height, 0.01);
waga = round(Weight, 0.1);

Gender_male = (Gender = 'Male');
Gender_female = (Gender = 'Female');

MTRANS_automobile = (MTRANS = 'Automobile');
MTRANS_bike = (MTRANS = 'Bike');
MTRANS_motorbike = (MTRANS = 'Motorbike');
MTRANS_walking = (MTRANS = 'Walking');
MTRANS_public_transportation = (MTRANS = 'Public_Transportation');

if CAEC = 'no' then przekaski = 0;
else if CAEC = 'Sometimes' then przekaski = 1;
else if CAEC = 'Frequently' then przekaski = 2;
else if CAEC = 'Always' then przekaski = 3;

if CALC = 'no' then alkohol = 0;
else if CALC = 'Sometimes' then alkohol = 1;
else if CALC = 'Frequently' then alkohol = 2;
else if CALC = 'Always' then alkohol = 3;

if family_history_with_overweight = 'yes' then rodzinna0tylosc = 0;
else if family_history_with_overweight = 'no' then rodzinna0tylosc = 1;

if FAVC = 'yes' then kaloryczne Posilki = 1;
else if FAVC = 'no' then kaloryczne Posilki = 0;

warzywa = round(FCVC);
posilki = round(NCP);

if SMOKE = 'yes' then palacz = 1;
else if SMOKE = 'no' then palacz = 0;

woda = round(CH2O);

if SCC = 'yes' then monitorowanieKalori = 1;
else if SCC = 'no' then monitorowanieKalori = 0;

aktywnoscFizyczna = round(FAF);
czasPrzedEkranem = round(TUE);

if NObeyesdad = 'Insufficient_Weight' then otylosc = 0;
else if NObeyesdad = 'Normal_Weight' then otylosc = 1;
else if NObeyesdad = 'Overweight_Level_I' then otylosc = 2;
else if NObeyesdad = 'Overweight_Level_II' then otylosc = 3;
else if NObeyesdad = 'Obesity_Type_I' then otylosc = 4;
else if NObeyesdad = 'Obesity_Type_II' then otylosc = 5;
else if NObeyesdad = 'Obesity_Type_III' then otylosc = 6;

drop Age Height Weight family_history_with_overweight FAVC FCVC NCP Gender CALC CAEC MTRANS SMOKE CH2O SCC FAF TUE NObeyesdad;
run;
```

Rysunek 1. Kod SAS-owy przetwarzający dane

Kolejnym etapem jest zapisanie danych do tabeli SAS-owej za pomocą węzła zapisu danych i zaimportowanie jej do projektu.



Rysunek 2. Diagram importu danych

Tabela po imporcie prezentuje się następująco:

Zmienne - OBESITY_TRAIN

(brak) nie równe

Kolumny: Etykieta Eksploracja Podstawowe

Nazwa	Rola	Poziom	Raport	...	Porzucenie	Dolna granica	Górna granica
Gender_female	Wejście	Binarna	Nie		Nie	.	.
Gender_male	Wejście	Binarna	Nie		Nie	.	.
MTRANS_automobile	Wejście	Binarna	Nie		Nie	.	.
MTRANS_bike	Wejście	Binarna	Nie		Nie	.	.
MTRANS_motorbike	Wejście	Binarna	Nie		Nie	.	.
MTRANS_public_transportation	Wejście	Binarna	Nie		Nie	.	.
MTRANS_walking	Wejście	Binarna	Nie		Nie	.	.
aktywnoscFizyczna	Wejście	Porządkowa	Nie		Nie	.	.
alkohol	Wejście	Porządkowa	Nie		Nie	.	.
czasPrzedEkranem	Wejście	Porządkowa	Nie		Nie	.	.
kaloryczne Posilki	Wejście	Binarna	Nie		Nie	.	.
monitorowanieKalori	Wejście	Binarna	Nie		Nie	.	.
otylosc	Zmienna celu	Porządkowa	Nie		Nie	.	.
palacz	Wejście	Binarna	Nie		Nie	.	.
posilki	Wejście	Porządkowa	Nie		Nie	.	.
przekaski	Wejście	Porządkowa	Nie		Nie	.	.
rodzinnaOtylosc	Wejście	Binarna	Nie		Nie	.	.
waga	Wejście	Przedziałowa	Nie		Tak	.	.
warzywa	Wejście	Porządkowa	Nie		Nie	.	.
wiek	Wejście	Przedziałowa	Nie		Nie	.	.
woda	Wejście	Porządkowa	Nie		Nie	.	.
wzrost	Wejście	Przedziałowa	Nie		Nie	.	.

Rysunek 3. Tabela OBESITY

Wstępna analiza danych

W celu wstępnego zrozumienia zbioru danych, przeprowadzono analizę eksploracyjną wykorzystując kod w języku SAS oraz węzeł tworzący wykresy. Analiza ma na celu zrozumienie rozkładów zmiennych poprzez wyznaczenie podstawowych statystyk opisowych oraz wizualizację rozkładów, co pozwoli na wykrycie wartości odstających i asymetrii. Badane są również korelacje między zmiennymi numerycznymi oraz powiązania między zmiennymi kategorycznymi a poziomem otyłości.

```
proc univariate data=EMWS2.Ids_DATA;
   var aktywnoscFizyczna alkohol czasPrzedEkranem posilki przekaski woda warzywa waga wzrost wiek;
run;

proc freq data=EMWS2.Ids_DATA;
   tables Gender_female Gender_male MTRANS_automobile MTRANS_bike
          MTRANS_motorbike MTRANS_public_transportation MTRANS_walking
          kalorycznePosilki monitorowanieKalori palacz rodzinnaOtylosc otylosc;
run;
proc corr data=EMWS2.Ids_DATA spearman;
   var aktywnoscFizyczna alkohol czasPrzedEkranem posilki przekaski woda warzywa waga wzrost wiek;
run;
proc freq data=EMWS2.Ids_DATA;
   tables (Gender_female Gender_male MTRANS_automobile MTRANS_bike MTRANS_motorbike
           MTRANS_public_transportation MTRANS_walking kalorycznePosilki palacz rodzinnaOtylosc) * otylosc / chisq;
run;
```

Rysunek 4. Kod SAS-owy analizujący dane

Procedura univariate generuje dla każdej wskazanej zmiennej (aktywnoscFizyczna, alkohol, czasPrzedEkranem, posiłki, przekąski, woda, warzywa, waga, wzrost, wiek) liczbę obserwacji, sumę wag oraz oblicza średnią, odchylenie standardowe, wariancje, skośność, kurtozę, kwantyle i przeprowadza testy położenia w celu określenia czy średnia zmiennej różni się istotnie od zera a także wskazuje obserwacje odstające.

Przykładowo analiza zmiennej wieku pokazuje, że liczba obserwacji wynosi 2111 co jest zgodne z ilością wierszy w bazie więc nie występują braki danych. Średni wiek to 24,32 lat co wskazuje na to, że próba składa się głównie z osób młodszych, odchylenie standardowe wynoszące 6,36 oznacza, że zróżnicowanie wiekowe jest umiarkowanie. Skośność wynosząca 1,52 oznacza, że rozkład wieku jest prawostronnie asymetryczny – oznacza to, że w próbie znajduje się kilka osób starszych. Wartość kurtozy równej 2,8 wskazuje na wyższe zagęszczenie danych wokół średniej. Mediana wynosi 23 a moda wynosi 21 lat wskazując na to, że osoby badane są młodsze. Testy położenia wskazują na istotną różnicę od zera co jest oczywiste ze względu na to, że zmienna wiek nie może wynosić zero. Kwantyle pokazują bardziej szczegółowy rozkład oraz minimum oraz maximum zmiennej które wynosi odpowiednio 14 i 61 lat.

Procedura UNIVARIATE
Zmienna: wiek

Momenty

n	2111	Suma wag	2111
Średnia	24.315964	Suma obserwacji	51331
Odchylenie std.	6.35707808	Wariancja	40.4124417
Skośność	1.52132613	Kurtoza	2.79858182
Niesk. suma kw.	1333433	Skoryg. suma kw.	85270.252
Wsp. zmienności	26.14364	Błąd std. śr.	0.13836092

Bazowe miary statystyczne

Położenie		Zmienność	
Średnia	24.31596	Odchylenie std.	6.35708
Mediana	23.00000	Wariancja	40.41244
Moda	21.00000	Rozstęp	47.00000
		Rozstęp międzyćwiartkowy	6.00000

Testy położenia: mi0=0

Testowanie	-Statystyka-		-----Wartość p-----	
t Studenta	t	175.743	Pr. > |t|	<.0001
Znaków	M	1055.5	Pr. >= |M|	<.0001
Rangowanych znaków	S	1114608	Pr. >= |S|	<.0001

Kwantyle (definicja 5)

Poziom	Kwantyl
100% Maks.	61
99%	44
95%	38
90%	33
75% Q3	26
50% Mediana	23
25% Q1	20
10%	18
5%	18
1%	16
0% Min.	14

Obserwacje ekstremalne

----Najniższe---		----Najwyższe---	
Wartość	Obs.	Wartość	Obs.
14	416	55	1014
15	117	55	1089
16	1302	55	1159
16	960	56	253
16	954	61	134

Rysunek 5. Statystyki zmiennej wiek

Następna procedura freq stosowana dla zmiennych binarnych i kategorycznych oblicza udział poszczególnych kategorii w próbie, a także ich procentowy udział w całkowitej liczbie obserwacji. Przykładowo dla zmiennej Gender_female wartości równych zero jest 1068 co jest równe 50,59% próby a wartości równych jeden, czyli kobiet w podanej próbie jest 1043 co stanowi 49,41%. Wynik ten pokazuje, że próba jest zrównoważona pod względem płci i dane są dobrze reprezentowane.

Gender_female	Liczebność	Procent	Liczebność skumulowana	Procent skumulowany
0	1068	50.59	1068	50.59
1	1043	49.41	2111	100.00

Rysunek 6. Podsumowanie zmiennej Gender_female

Na podstawie wyników tej procedury widać, że dane w zmiennej wejściowej otylosc są równomiernie rozłożone

otylosc	Liczebność	Procent	Liczebność skumulowana	Procent skumulowany
0	272	12.88	272	12.88
1	287	13.60	559	26.48
2	290	13.74	849	40.22
3	290	13.74	1139	53.96
4	351	16.63	1490	70.58
5	297	14.07	1787	84.65
6	324	15.35	2111	100.00

Rysunek 7. Podsumowanie zmiennej otylosc

Następna procedura w kodzie – corr odpowiada za utworzenie macierzy korealcji dla zmiennych numerycznych, dzięki czemu można wstępnie stwierdzić ich niezależność.

Współczynniki korelacji Spearmana, N = 2111
Prawd. > |r| przy H0: rho=0

	aktywnosc Fizyczna	alkohol	czas Przed Ekranem	posilki	przekaski	woda	warzywa	waga	wzrost	wiek
aktywnoscFizyczna	1.00000	-0.07967	0.05770	0.14306	0.01537	0.11088	0.01500	-0.05116	0.31771	-0.19710
		0.0002	0.0080	<.0001	0.4802	<.0001	0.4909	0.0187	<.0001	<.0001
alkohol	-0.07967	1.00000	-0.03219	0.07560	-0.07648	0.09808	0.05941	0.21120	0.13557	0.09726
	0.0002		0.1392	0.0005	0.0004	<.0001	0.0063	<.0001	<.0001	<.0001
czasPrzedEkranem	0.05770	-0.03219	1.00000	0.02297	0.03779	-0.03539	-0.04900	-0.03449	0.07843	-0.29348
	0.0080	0.1392		0.2915	0.0826	0.1041	0.0244	0.1131	0.0003	<.0001
posilki	0.14306	0.07560	0.02297	1.00000	0.12037	0.05848	0.01780	0.04439	0.22492	-0.07267
	<.0001	0.0005	0.2915		<.0001	0.0072	0.4138	0.0414	<.0001	0.0008
przekaski	0.01537	-0.07648	0.03779	0.12037	1.00000	-0.15592	0.08008	-0.32379	-0.06035	-0.12205
	0.4802	0.0004	0.0826	<.0001		<.0001	0.0002	<.0001	0.0055	<.0001
woda	0.11088	0.09808	-0.03539	0.05848	-0.15592	1.00000	0.05178	0.19178	0.17586	0.02320
	<.0001	<.0001	0.1041	0.0072	<.0001		0.0173	<.0001	<.0001	0.2867
warzywa	0.01500	0.05941	-0.04900	0.01780	0.08008	0.05178	1.00000	0.15570	-0.07108	0.02963
	0.4909	0.0063	0.0244	0.4138	0.0002	0.0173		<.0001	0.0011	0.1736
waga	-0.05116	0.21120	-0.03449	0.04439	-0.32379	0.19178	0.15570	1.00000	0.46151	0.35666
	0.0187	<.0001	0.1131	0.0414	<.0001	<.0001	<.0001		<.0001	<.0001
wzrost	0.31771	0.13557	0.07843	0.22492	-0.06035	0.17586	-0.07108	0.46151	1.00000	-0.00255
	<.0001	<.0001	0.0003	<.0001	0.0055	<.0001	0.0011	<.0001		0.9069
wiek	-0.19710	0.09726	-0.29348	-0.07267	-0.12205	0.02320	0.02963	0.35666	-0.00255	1.00000
	<.0001	<.0001	<.0001	0.0008	<.0001	0.2867	0.1736	<.0001	0.9069	

Rysunek 8. Macierz korelacji

Analiza korelacji dostarcza wniosków dotyczących powiązań w danych. Spożycie alkoholu wskazuje umiarkowaną korelacje z wagą (r=0,21) oznacza to, że osoby spożywające więcej alkoholu mają tendencję do posiadania większej wagi. Może to wynikać z dodatkowych kalorii spożywanych z alkoholem lub zmniejszonej aktywności fizycznej u osób, które spożywają więcej alkoholu, co również potwierdza ujemna korelacja między spożyciem alkoholu a aktywnością fizyczną (r=-0,07967). Aktywność fizyczna jest również ujemnie skorelowana z wiekiem (r=-0,19710), co sugeruje, że starsze osoby są mniej aktywne fizycznie, co może

wpływać na inne aspekty zdrowia, takie jak waga i ogólna kondycja fizyczna. Czas spędzany przed ekranem wykazuje umiarkowaną ujemną korelację z wiekiem (r=-0,29348), co oznacza, że młodsze osoby spędzają więcej czasu przed ekranem. Liczba spożywanych posiłków jest dodatnio skorelowana z aktywnością fizyczną (r=0,14306) i wzrostem (r = 0,22492), co sugeruje, że osoby bardziej aktywne fizycznie i wyższe mają tendencję do spożywania większej liczby posiłków dziennie. Spożycie przekąsek wykazuje natomiast ujemną korelację z wagą (r=-0,32379), co może wskazywać, że osoby spożywające przekąski w umiarkowanych ilościach mają tendencję do lepszej kontroli wagi. Podobnie, spożycie przekąsek jest ujemnie skorelowane z wiekiem (r=-0,12205), co oznacza, że młodsze osoby spożywają więcej przekąsek, co może być związane z ich stylem życia. Spożycie wody jest dodatnio skorelowane z wagą (r=0,19178) oraz wzrostem (r=0,17586), co sugeruje, że osoby o większej masie ciała i wyższe spożywają więcej wody. Wiek jest dodatnio skorelowany z wagą (r=0,35666), co sugeruje, że starsze osoby mają tendencję do wyższej wagi.

Ostatnia procedura z kodu SAS wykonuje analizę związku między zmienną otyłość z różnymi cechami binarnymi.

Występuje lekkie zróżnicowanie otyłości w zależności od płci. W przypadku mężczyzn, liczba osób z wyższym poziomem otyłości jest generalnie większa w porównaniu do kobiet. Chi-kwadrat wynosi 657,7462, co sugeruje istotną zależność między płcią a poziomem otyłości. Statystyka V Cramera wynosząca 0,5582 wskazuje na silną zależność między tymi zmiennymi, co potwierdza, że płeć wpływa na poziom otyłości w badanej próbie.

Tabela Gender_female od otylosc

Gender_female otylosc

Liczebność / Procent / Proc. wier. / Proc. kol.	0	1	2	3	4	5	6	Suma
0	99	146	145	187	195	295	1	1068
	4.69	6.92	6.87	8.86	9.24	13.97	0.05	50.59
	9.27	13.67	13.58	17.51	18.26	27.62	0.09	
	36.40	50.87	50.00	64.48	55.56	99.33	0.31	
1	173	141	145	103	156	2	323	1043
	8.20	6.68	6.87	4.88	7.39	0.09	15.30	49.41
	16.59	13.52	13.90	9.88	14.96	0.19	30.97	
	63.60	49.13	50.00	35.52	44.44	0.67	99.69	
Suma	272	287	290	290	351	297	324	2111
	12.88	13.60	13.74	13.74	16.63	14.07	15.35	100.00

Statystyki dla tabeli przedstawiającej Gender_female od otylosc

Statystyka	DF	Wartość	Prawd.
Chi-kwadrat	6	657.7462	<.0001
Chi-kw. ilorazu wiarygodn.	6	872.5465	<.0001
Chi-kwadrat Mantela-Haenszela	1	2.0888	0.1484
Współczynnik FI		0.5582	
Współczynnik kontyngencji		0.4874	
V Cramera		0.5582	

Rysunek 9. Statystyki Gender_female od otylosc

Zmienna dotycząca rodzinnej historii otyłości wykazuje silną zależność z poziomem otyłości badanych osób. Osoby, które zgłaszają, że w ich rodzinie występowała otyłość, rzadziej mają wyższy poziom otyłości. Chi-kwadrat wynosi 621,9794, co potwierdza istotność tej zależności. Współczynnik V Cramera wynosi 0.5428, co oznacza silną zależność.

Tabela rodzinnaOtylosc od otylosc

rodzinnaOtylosc otylosc

Liczebność / Procent / Proc. wier. / Proc. kol.	0	1	2	3	4	5	6	Suma
0	126	155	209	272	344	296	324	1726
	5.97	7.34	9.90	12.88	16.30	14.02	15.35	81.76
	7.30	8.98	12.11	15.76	19.93	17.15	18.77	
	46.32	54.01	72.07	93.79	98.01	99.66	100.00	
1	146	132	81	18	7	1	0	385
	6.92	6.25	3.84	0.85	0.33	0.05	0.00	18.24
	37.92	34.29	21.04	4.68	1.82	0.26	0.00	
	53.68	45.99	27.93	6.21	1.99	0.34	0.00	
Suma	272	287	290	290	351	297	324	2111
	12.88	13.60	13.74	13.74	16.63	14.07	15.35	100.00

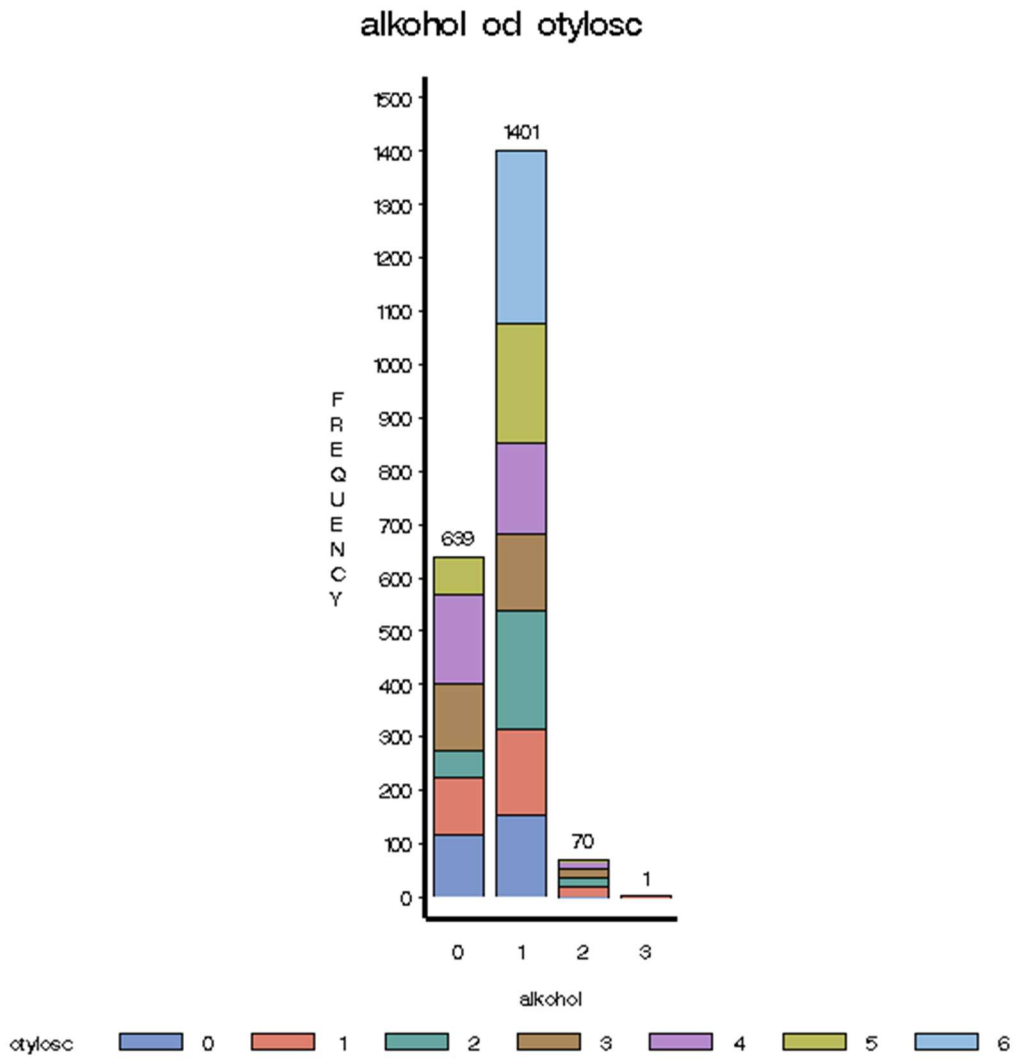
Statystyki dla tabeli przedstawiającej rodzinnaOtylosc od otylosc

Statystyka	DF	Wartość	Prawd.
Chi-kwadrat	6	621.9794	<.0001
Chi-kw. ilorazu wiarygodn.	6	673.2379	<.0001
Chi-kwadrat Mantela-Haenszela	1	538.4191	<.0001
Współczynnik FI		0.5428	
Współczynnik kontyngencji		0.4771	
V Cramera		0.5428	

Rysunek 10. Statystyki rodzinnaOtylosc od otylosc

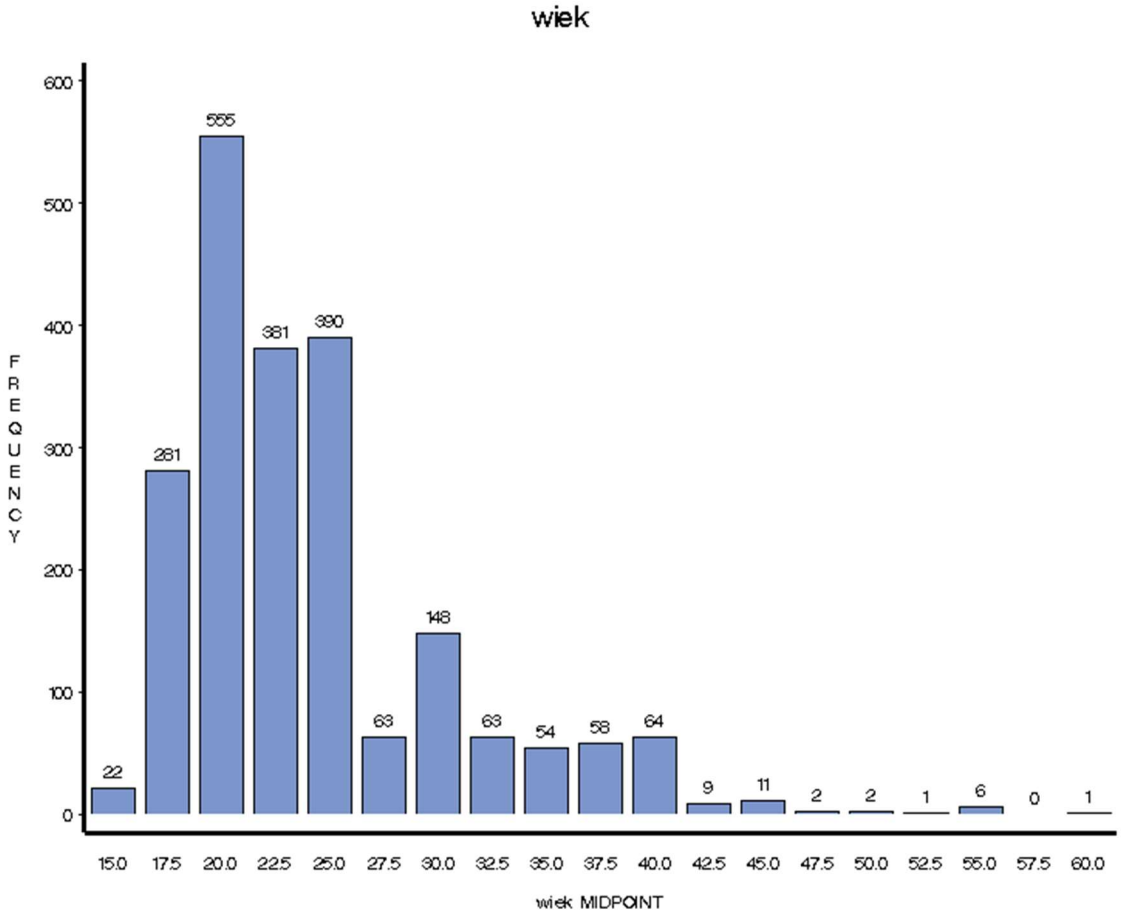
W celach analizy rozkładów zmiennych został również użyty węzeł ‘wykresy różne’ generujący histogramy dla wielu zmiennych, omówione zostaną wnioski dotyczące trzech zmiennych: alkoholu, wzrostu oraz wieku.

Na wykresie dotyczącym zmiennej "alkohol" można zauważyć wyraźną asymetrię w rozkładzie. Większość obserwacji dotyczy osób, które nie spożywają alkoholu lub spożywają go niewiele (wartość 1) podczas gdy pozostałe poziomy konsumpcji alkoholu (wartość 2 i 3) mają znacznie mniej obserwacji. Tak duża asymetria może wpłynąć na jakość modelu predykcyjnego, dlatego w kolejnym kroku ta zmienna porządkowa zostanie przekształcona na zmienną binarną. Dzięki temu uproszczeniu, zmienna "alkohol" będzie miała dwie wartości: brak spożycia alkoholu oraz spożycie alkoholu.



Rysunek 11. Wykres alkohol od otylosc

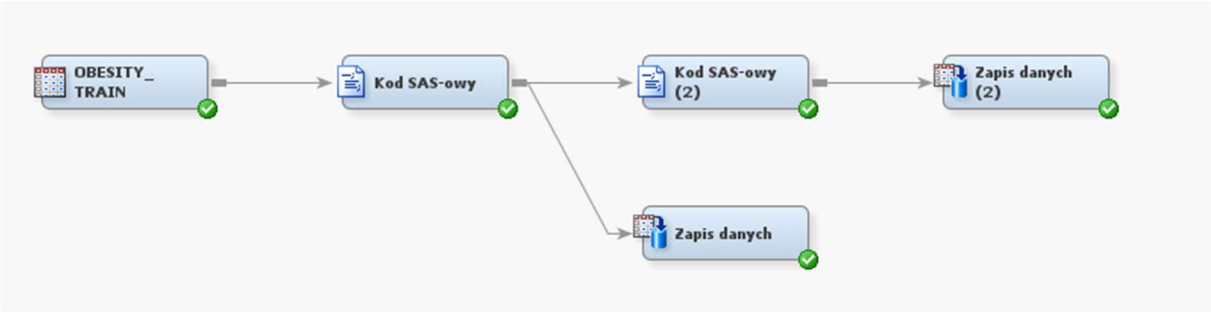
Analiza zmiennej „wiek” wykazała asymetrię oraz obecność wartości odstających, szczególnie dotyczących bardzo młodych oraz znacznie starszych respondentów. W celu poprawy reprezentatywności i stabilności zmiennej, planowane jest usunięcie tych nietypowych wartości, co umożliwi lepsze odzwierciedlenie rzeczywistego rozkładu wieku w badanej próbie i przyczyni się do bardziej precyzyjnych predykcji.



Rysunek 12. Histogram rozkładu wieku

Modyfikacji danych

Schemat przedstawia proces przekształcania i przygotowywania danych do analizy w programie SAS, który obejmuje kilka kroków.



Rysunek 13. Schemat modyfikacji danych

W pierwszym etapie, kod SAS-owy odpowiedzialny jest za modyfikację wybranych zmiennych, tj. „wiek", „wzrost" oraz „alkohol". Dane są filtrowane w taki sposób, aby w zbiorze pozostawiono jedynie osoby w przedziale wiekowym od 18 do 55 lat oraz wzrostem mieszczącym się między 1,5 a 1,92 m, eliminując tym samym obserwacje odstające w tych zmiennych. Dodatkowo zmienna „alkohol" zostaje przekształcona na zmienną binarną, gdzie wartość 1 przypisano osobom spożywającym alkohol, a 0 osobom, które tego nie robią.

```
DATA EMWS6.EMCODE2_TRAIN;
set EMWS6.Ids_DATA;
if wiek >= 18 and wiek <=55;
if wzrost >= 1.5 and wzrost <= 1.92;
if alkohol <> 0 then alkohol = 1;
run;
```

Rysunek 14. Kod SAS-owy modyfikujący dane

Następnie w kolejnym etapie kod SAS-owy 2 dokonuje kluczowej modyfikacji na zmiennej celu „otyłość". Zmienna ta, która pierwotnie miała charakter porządkowy, zostaje przekształcona na zmienną binarną. Zmiennym o wartościach odpowiadających "Niedowadze" i "Wadze normalnej" przypisywana jest wartość 0, natomiast pozostałe kategorie otyłości (Nadwaga, Otyłość typu I, II, III) zostają zgrupowane i przekształcone w wartość 1, co umożliwia modelowanie na zmiennej binarnej.

```
DATA EMWS6.EMCODE2_TRAIN;
set EMWS6.Ids_DATA;
if otylosc IN(0, 1) then otylosc = 0;
else otylosc = 1;
run;
```

Rysunek 15. Kod SAS-owy przekształcający zmienną otylosc na zmienną binarną

Zwieńczeniem procesu jest etap zapisu danych. Pierwszy zapis tworzy nową tabelę, w której zmienna „otyłość” pozostaje w swojej pierwotnej formie porządkowej, a obserwacje odstające nie są uwzględnione. Drugi zapis danych tworzy tabelę, w której dane również zostały pozbawione obserwacji odstających, a zmienna „otyłość” jest już zredukowana do formy binarnej.

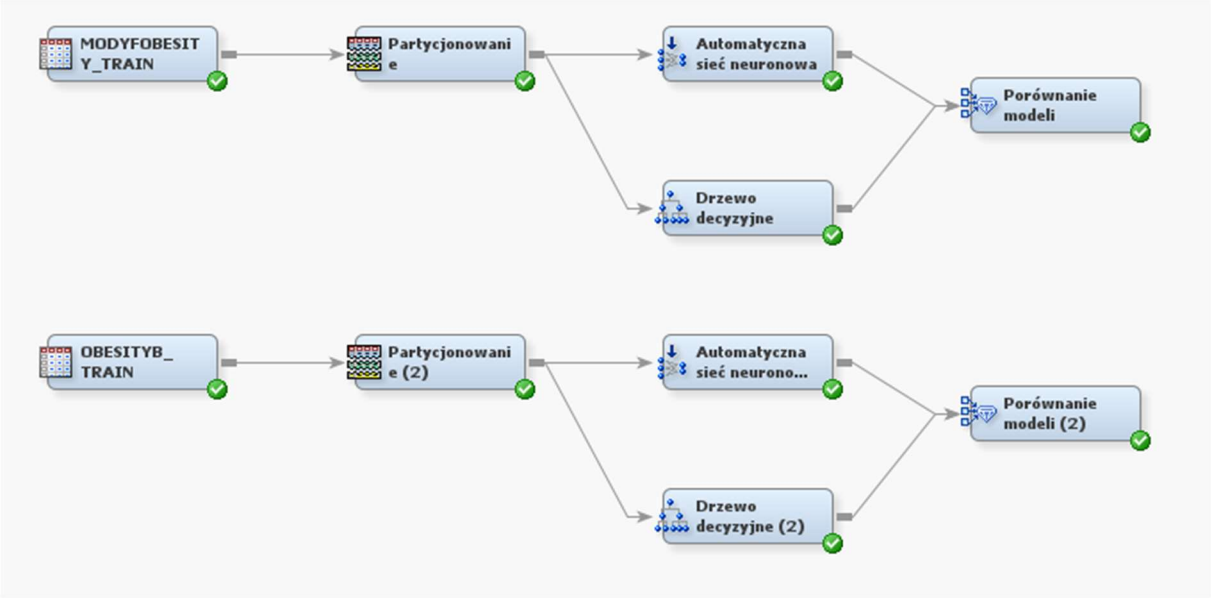
Modelowanie

Zastosowano dwa typy modeli: automatyczną sieć neuronową oraz drzewo decyzyjne. Oba modele zostały wykorzystane w kontekście predykcji zarówno zmiennej porządkowej, jak i binarnej, w celu oceny ich skuteczności i możliwości zastosowania w analizie danych dotyczących otyłości.

Automatyczna sieć neuronowa jest zaawansowanym narzędziem uczenia maszynowego, które umożliwia modelowanie skomplikowanych i nieliniowych zależności w danych. W modelu tym wykorzystuje się wielowarstwową strukturę neuronów, które przetwarzają dane wejściowe w celu wykrycia istotnych wzorców i zależności. Sieci neuronowe są szczególnie efektywne przy pracy z danymi, gdzie relacje pomiędzy zmiennymi są trudne do uchwycenia za pomocą tradycyjnych metod statystycznych. W przypadku predykcji zmiennej porządkowej sieć neuronowa pozwala na uwzględnienie złożonych interakcji między zmiennymi, natomiast w przypadku predykcji zmiennej binarnej model skupia się na dokładnym rozróżnieniu między tymi dwoma klasami.

Drzewo decyzyjne jest metodą opartą na iteracyjnym podziale danych na podstawie wartości atrybutów. Tworzy ono strukturę drzewa, gdzie każdy węzeł odpowiada testowi na wartość określonej zmiennej, a gałęzie reprezentują możliwe wyniki testu. Drzewa decyzyjne są interpretowalne, umożliwiając zrozumienie, które zmienne i ich wartości mają największy wpływ na określony poziom zmiennej celu. W kontekście zmiennej porządkowej drzewo pozwala zidentyfikować hierarchię cech wpływających na różne poziomy otyłości. Dla zmiennej binarnej model ten umożliwia skuteczną klasyfikację jednostek na te z obecnością lub brakiem otyłości, opierając się na określonych regułach decyzyjnych.

Oba modele zostały zastosowane w dwóch podejściach, zgodnie z przedstawionym diagramem: do predykcji zmiennej porządkowej (MODYFOBESITY_TRAIN) oraz do predykcji zmiennej binarnej (OBESITYB_TRAIN). Takie podejście umożliwia kompleksowe porównanie ich efektywności oraz zdolności generalizacji w różnorodnych kontekstach klasyfikacyjnych.



Rysunek 16. Schemat budowy modeli

Węzeł partycjonowania został zastosowany w celu podzielenia danych na trzy zbiory: uczący (70%), walidacyjny (15%) i testowy (15%). Podział ten umożliwia odpowiednią ocenę i dostrojenie modeli predykcyjnych. Zbiór uczący służy do wytrenowania modelu, umożliwiając mu identyfikację wzorców w danych. Zbiór walidacyjny wykorzystywany jest podczas procesu uczenia w celu dostrojenia hiperparametrów modelu oraz uniknięcia problemu nadmiernego dopasowania. Ostateczna ocena jakości modelu jest przeprowadzana na zbiorze testowym, który zawiera dane, z którymi model wcześniej nie miał kontaktu, co pozwala na ocenę jego zdolności generalizacyjnych.

W pierwszym modelu sieci neuronowej do klasyfikacji zmiennej porządkowej wybrano architekturę wielowarstwowej sieci lejkowej. Architektura lejkowa umożliwia stopniowe wyodrębnianie kluczowych cech z danych wejściowych. Sieć rozpoczyna od większej liczby neuronów w pierwszej warstwie i stopniowo zmniejsza ich liczbę w kolejnych warstwach ukrytych, co pozwala na efektywne przechwytywanie złożonych wzorców w danych i ich agregację w bardziej zrozumiałe reprezentacje.

Zastosowano funkcje aktywacji "logistyczną", "softmax" oraz "tanh", które zostały dobrane ze względu na ich właściwości nieliniowe, kluczowe w kontekście uczenia się złożonych zależności między zmiennymi wejściowymi a zmienną docelową. W modelu przewidziano maksymalnie 50 iteracji procesu uczenia, a także ustawiono 5 dodatkowych iteracji końcowych, które pozwalają na precyzyjne dostrojenie wag sieci.

Opcje modeli	
Architektura	Wielowarstwowa lejkowata
Zakończenie	Przeuczenie
Czynność uczenia	Wyszukiwanie
Funkcja błędu warstwy docelowej	Domyślna
Maksymalnie iteracji	50
Liczba jednostek ukrytych	8
Tolerancja	Średnia
Czas całkowity	Godzina
Opcje przyrostu i wyszukiwania	
Koryguj iteracje	Tak
Zablokuj połączenia	Nie
Całkowita liczba jednostek ukrytych	30
Uczenie końcowe	Tak
Końcowe iteracje	5
Funkcje aktywacji	
Bezpośrednia	Nie
Wykładnicza	Nie
Tożsamościowa	Nie
Logistyczna	Tak
Normalna	Nie
Odwrotna	Nie
Sinus	Nie
Softmax	Nie
Kwadratowa	Nie
Tanh	Tak

Rysunek 17. Parametry automatycznej sieci neuronowej

W drzewie decyzyjnym skonfigurowano szereg parametrów w celu optymalizacji klasyfikacji zmiennej porządkowej. Zastosowano jako kryterium podziału współczynnik Giniego, który służy do oceny niejednorodności klas w każdym podziale. Model ogranicza maksymalne rozgałęzienie każdego węzła do dwóch, co sprzyja przejrzystości struktury drzewa i ułatwia interpretację wyników. Minimalna wielkość liścia została określona na 30, co zapewnia, że w każdym końcowym węźle drzewa znajduje się wystarczająca liczba obserwacji, by uznać go za istotny. Wykorzystano walidację krzyżową z liczbą podzbiorów ustawioną na 8 i czterema powtórzeniami. Pozwala to na dokładną ocenę wydajności modelu oraz minimalizuje ryzyko przeuczenia.

Reguła podziału	
Kryterium przedziałowej zmiennej celu	ProbF
Kryterium nominalnej zmiennej celu	ProbChisq
Kryterium porządkowej zmiennej celu	Wsp. Giniego
Poziom istotności	0.05
Braki danych	Użyj w wyszukiwaniu
Używaj jednorazowo	Nie
Maksymalne rozgałęzienie	2
Maksymalna głębia	5
Minimalna wielkość zmiennej kategoryzującej	5
Węzeł	
Wielkość liścia	30
Liczba reguł	5
Liczba reguł zastępczych	0
Wielkość podziału	20
Poszukiwanie podziału	
Użyj decyzji	Nie
Użyj prawdopodobieństw a priori	Nie
Wyczerpujące	5000
Próba węzła	20000
Poddrzewo	
Metoda	Ocena
Liczba liści	1
Miara oceny	Decyzja
Ułamek ocen	0.25
Walidacja krzyżowa	
Wykonuj walidację krzyżową	Tak
Liczba podzbiorów	8
Liczba powtórzeń	4
Ziarno	12345

Rysunek 18. Parametry drzewa decyzyjnego

W przypadku modelu sieci neuronowej klasyfikującej zmienną binarną wprowadzono jedną zmianę a mianowicie wyłączenie funkcji aktywacji „softmax” ponieważ jest ona stosowana w przypadku problemów klasyfikacji wieloklasowej. W przypadku modelu drzewa decyzyjnego nie zmodyfikowano żadnych parametrów.

Analiza wyników

Model sieci neuronowej charakteryzuje się bardzo wysokimi wartościami AUR, które wynoszą odpowiednio 0,996 dla zbioru uczącego, 0,989 dla zbioru walidacyjnego oraz 0,992 dla zbioru testowego. Współczynnik Giniego dla tego modelu osiąga również znaczące wartości: 0,992, 0,978 i 0,985. Tak wysokie wartości tych wskaźników wskazują na bardzo dobrą zdolność modelu sieci neuronowej do rozróżniania klas zmiennej porządkowej "otyłość."

Obs.	TARGET	TARGETLABEL	_AUR_	_GINI_	KS	_KS_PROB_ CUTOFF	_KS_BIN_	BINNED_KS_ PROB_ CUTOFF
1	otylosc		0.996	0.992	0.993	0.369	0.949	0.533

Obs.	TARGET	TARGETLABEL	_VAUR_	_VGINI_	VKS	_VKS_ PROB_ CUTOFF_	_VKS_ BIN_	_VBINNED_ KS_PROB_ CUTOFF_
1	otylosc		0.989	0.978	0.985	0.275	0.937	0.451

Obs.	TARGET	TARGETLABEL	_TAUR_	_TGINI_	TKS	_TKS_ PROB_ CUTOFF_	_TKS_ BIN_	_TBINNED_ KS_PROB_ CUTOFF_
1	otylosc		0.992	0.985	0.979	0.335	0.934	0.532

Rysunek 19. Wyniki modelu automatycznej sieci neuronowej dla zmiennej porządkowej

W przypadku modelu drzewa decyzyjnego wartości AUR są nieco niższe i wynoszą 0,976, 0,973 oraz 0,979. Współczynnik Giniego uzyskany dla drzewa decyzyjnego wynosi odpowiednio 0,952, 0,945 i 0,959. Chociaż wartości te są niższe niż w przypadku sieci neuronowej, nadal wskazują na dobrą zdolność drzewa decyzyjnego do separacji klas.

Obs.	TARGET	TARGETLABEL	_AUR_	_GINI_	KS	_KS_PROB_ CUTOFF	_KS_BIN_	BINNED_KS_ PROB_ CUTOFF
1	otylosc		0.976	0.952	0.948	0.024	0.944	0.804

Obs.	TARGET	TARGETLABEL	_VAUR_	_VGINI_	VKS	_VKS_ PROB_ CUTOFF_	_VKS_ BIN_	_VBINNED_ KS_PROB_ CUTOFF_
1	otylosc		0.973	0.945	0.946	0.024	0.937	0.414

Obs.	TARGET	TARGETLABEL	_TAUR_	_TGINI_	TKS	_TKS_ PROB_ CUTOFF_	_TKS_ BIN_	_TBINNED_ KS_PROB_ CUTOFF_
1	otylosc		0.979	0.959	0.959	0.024	0.953	0.414

Rysunek 20. Wyniki drzewa decyzyjnego dla zmiennej porządkowej otylosc

Analiza statystyki Kołmogorowa-Smirnowa również potwierdza wyższość modelu sieci neuronowej w klasyfikacji zmiennej porządkowej. Wartości KS dla sieci neuronowej wynoszą odpowiednio 0,993, 0,985 oraz 0,979, co oznacza doskonałą zdolność do separacji klas. Dla modelu drzewa decyzyjnego wartości te są niższe, osiągając 0,948, 0,946 i 0,959. Pomimo nieco niższych wartości w porównaniu z siecią neuronową, drzewo decyzyjne nadal wykazuje dobrą zdolność do klasyfikacji zmiennej porządkowej.

Sieć neuronowa charakteryzuje się niższym odsetkiem błędnych klasyfikacji w porównaniu z drzewem decyzyjnym. W przypadku sieci neuronowej odsetek błędnych klasyfikacji wynosi 29,45% dla zbioru uczącego oraz 35,59% dla zbioru walidacyjnego. Dodatkowo, przeciętny błąd kwadratowy dla tego modelu to 0,07 zarówno dla zbioru uczącego, jak i walidacyjnego, co świadczy o wysokiej precyzji przewidywań. Dla modelu drzewa decyzyjnego odsetek błędnych klasyfikacji jest wyższy i wynosi 43,29% oraz 46,87% w zbiorze walidacyjnym. Ponadto, przeciętny błąd kwadratowy jest również wyższy, osiągając 0,08 dla zbioru uczącego i 0,09 dla walidacyjnego, co wskazuje na mniejszą dokładność tego modelu w porównaniu do sieci neuronowej.

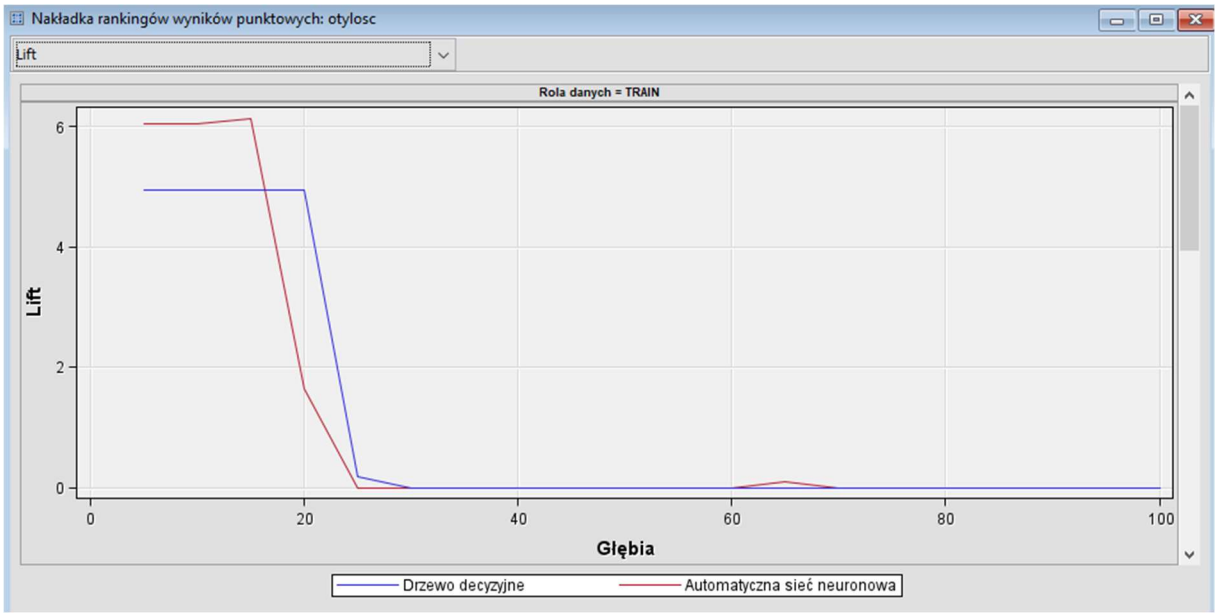
Statystyki dopasowania
Wybór modelu w zależności od Walidacja: Odsetek błędnych klasyfikacji (_VMISC_)

Wybrany model	Węzeł modelu	Opis modelu	Walidacja: Odsetek błędnych klasyfikacji	Uczenie: Przeciętny błąd kwadratowy	Uczenie: Odsetek błędnych klasyfikacji	Walidacja: Przeciętny błąd kwadratowy
Y	AutoNeural	Automatyczna sieć neuronowa	0.35589	0.066197	0.29446	0.073727
	Tree	Drzewo decyzyjne	0.46867	0.080717	0.43289	0.086313

Rysunek 21. Statystyki dopasowania modeli dla zmiennej porządkowej otylosc

Analiza wskaźnika liftu i zysku potwierdza wyższość modelu sieci neuronowej nad drzewem decyzyjnym. Dla sieci neuronowej skumulowany lift wynosi 6,04 w zbiorze uczącym i 6,14 w zbiorze walidacyjnym, co oznacza, że model jest znacznie skuteczniejszy niż losowe klasyfikowanie przypadków. Zysk osiągnięty przez sieć neuronową wynosi 504,19 i 513,85.

Dla drzewa decyzyjnego wskaźnik liftu jest niższy i wynosi 4,94 oraz 5,10 w zbiorze walidacyjnym. Zysk uzyskany z zastosowania drzewa decyzyjnego wynosi 394,11 i 410,21, co również jest niższe niż w przypadku modelu sieci neuronowej.



Rysunek 22. Wykres liftu

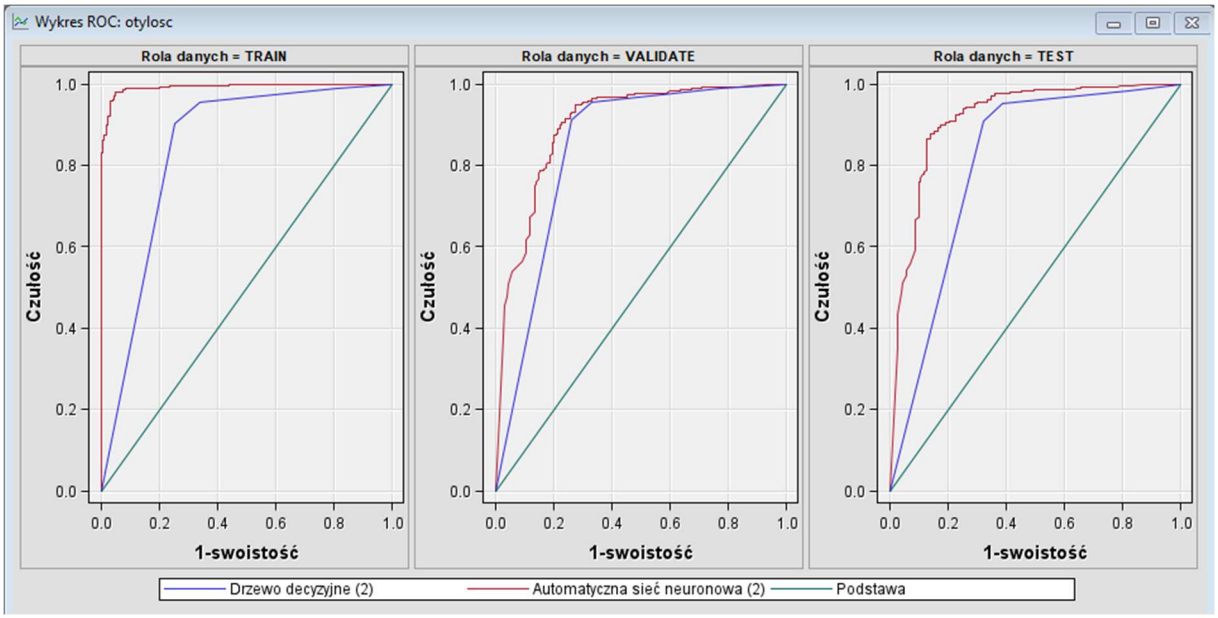
Model sieci neuronowej do klasyfikacji zmiennej binarnej charakteryzuje się bardzo wysokimi wartościami AUR, które wynoszą odpowiednio 0,99 dla zbioru uczącego, 0,90 dla zbioru walidacyjnego oraz 0,92 dla zbioru testowego. Współczynnik Giniego dla tego modelu osiąga również znaczące wartości: 0,985, 0,80 i 0,83. Tak wysokie wartości tych wskaźników wskazują na bardzo dobrą zdolność modelu sieci neuronowej do rozróżniania klas zmiennej binarnej "otyłość."

Obs.	TARGET	TARGETLABEL	_AUR_	_GINI_	KS	_KS_PROB_ CUTOFF	_KS_BIN_	BINNED_KS_ PROB_ CUTOFF
1	otylosc		0.993	0.985	0.928	0.563	0.922	0.717

Obs.	TARGET	TARGETLABEL	_VAUR_	_VGINI_	VKS	_VKS_ PROB_ CUTOFF_	_VKS_ BIN_	_VBINNED_ KS_PROB_ CUTOFF_
1	otylosc		0.9	0.801	0.675	0.846	0.666	0.831

Obs.	TARGET	TARGETLABEL	_TAUR_	_TGINI_	TKS	_TKS_ PROB_ CUTOFF_	_TKS_ BIN_	_TBINNED_ KS_PROB_ CUTOFF_
1	otylosc		0.915	0.83	0.737	0.917	0.726	0.914

Rysunek 23. Wyniki modelu automatycznej sieci neuronowej dla zmiennej binarnej otylosc



Rysunek 24. Wykres ROC dla obu modeli

W przypadku modelu drzewa decyzyjnego wartości AUR są niższe i wynoszą 0,84 dla zbioru uczącego, 0,838 dla zbioru walidacyjnego oraz 0,8 dla zbioru testowego. Współczynnik Giniego uzyskany dla drzewa decyzyjnego wynosi odpowiednio 0,68, 0,677 i 0,61. Chociaż wartości te są niższe niż w przypadku sieci neuronowej, nadal wskazują na umiarkowaną zdolność drzewa decyzyjnego do separacji klas.

Obs.	TARGET	TARGETLABEL	_AUR_	_GINI_	KS	_KS_PROB_ CUTOFF	_KS_BIN_	BINNED_KS_ PROB_ CUTOFF
1	otylosc		0.84	0.68	0.65	0.646	0.645	0.781

Obs.	TARGET	TARGETLABEL	_VAUR_	_VGINI_	VKS	_VKS_ PROB_ CUTOFF_	_VKS_ BIN_	_VBINNED_ KS_PROB_ CUTOFF_
1	otylosc		0.838	0.677	0.648	0.646	0.648	0.916

Obs.	TARGET	TARGETLABEL	_TAUR_	_TGINI_	TKS	_TKS_ PROB_ CUTOFF_	_TKS_ BIN_	_TBINNED_ KS_PROB_ CUTOFF_
1	otylosc		0.804	0.607	0.587	0.646	0.577	0.916

Rysunek 25. Wyniki drzewa decyzyjnego dla zmiennej binarnej otylosc

Analiza statystyki Kołmogorowa-Smirnowa również potwierdza wyższość modelu sieci neuronowej w klasyfikacji zmiennej binarnej. Wartości KS dla sieci neuronowej wynoszą odpowiednio 0,93, 0,68 oraz 0,74, co oznacza bardzo dobrą zdolność do separacji klas. Dla modelu drzewa decyzyjnego wartości te są niższe, osiągając 0,65 zarówno dla zbioru uczącego, jak i walidacyjnego, oraz 0,58 dla zbioru testowego. Pomimo niższych wartości w porównaniu z siecią neuronową, drzewo decyzyjne nadal wykazuje pewną zdolność do klasyfikacji zmiennej binarnej.

Sieć neuronowa charakteryzuje się niższym odsetkiem błędnych klasyfikacji w porównaniu z drzewem decyzyjnym. W przypadku sieci neuronowej odsetek błędnych klasyfikacji wynosi 3% dla zbioru uczącego oraz 11% dla zbioru walidacyjnego. Dodatkowo, przeciętny błąd kwadratowy dla tego modelu to 0,03 dla zbioru uczącego i 0,10 dla walidacyjnego, co świadczy o wysokiej precyzji przewidywań. Dla modelu drzewa decyzyjnego odsetek błędnych klasyfikacji jest wyższy i wynosi 12% zarówno w zbiorze uczącym, jak i walidacyjnym. Ponadto, przeciętny błąd kwadratowy jest również wyższy, osiągając 0,10 dla zbioru uczącego i 0,10 dla walidacyjnego, co wskazuje na mniejszą dokładność tego modelu w porównaniu do sieci neuronowej.

Statystyki dopasowania
Wybór modelu w zależności od Walidacja: Odsetek błędnych klasyfikacji (_VMISC_)

Wybrany model	Węzeł modelu	Opis modelu	Walidacja: Odsetek błędnych klasyfikacji	Uczenie: Przeciętny błąd kwadratowy	Uczenie: Odsetek błędnych klasyfikacji	Walidacja: Przeciętny błąd kwadratowy
Y	AutoNeural2	Automatyczna sieć neuronowa (2)	0.11037	0.027135	0.03011	0.096073
	Tree2	Drzewo decyzyjne (2)	0.11538	0.099257	0.11794	0.098501

Rysunek 26. Statystyki dopasowania

Wnioski

Model sieci neuronowej wykazuje wysoką skuteczność w rozpoznawaniu otyłości, o czym świadczą wartości wskaźników jakości, takich jak AUR i współczynnik Giniego, uzyskane dla zbiorów uczącego, walidacyjnego i testowego. Analiza statystyki Kołmogorowa-Smirnowa (KS) dodatkowo potwierdza zdolność tego modelu do precyzyjnej separacji klas. Sieć neuronowa cechuje się niższym odsetkiem błędnych klasyfikacji oraz mniejszym przeciętnym błędem kwadratowym, co podkreśla jej precyzję w przewidywaniu zmiennej celu.

Model drzewa decyzyjnego również wykazuje zadowalające rezultaty, choć wskaźniki jakości, takie jak AUR i współczynnik Giniego, sugerują umiarkowaną skuteczność w klasyfikacji przypadków otyłości. Statystyka Kołmogorowa-Smirnowa dla drzewa decyzyjnego wskazuje na nieco niższą zdolność rozróżniania klas, co przekłada się na wyższy odsetek błędnych klasyfikacji. Jednak w kontekście wskaźnika liftu i zysku, sieć neuronowa wyraźnie wykazuje przewagę, osiągając wyższe wartości.

Podsumowując, analiza wyników jednoznacznie wskazuje na subtelną wyższość modelu sieci neuronowej w rozpoznawaniu przypadków otyłości. Sieć neuronowa charakteryzuje się lepszą zdolnością do rozróżniania klas, wyższą precyzją przewidywań oraz większą efektywnością w klasyfikacji w porównaniu z modelem drzewa decyzyjnego. W związku z tym, model sieci neuronowej można uznać za bardziej odpowiedni do zastosowań związanych z klasyfikacją otyłości na poziomie zmiennej binarnej jak i porządkowej.