Wpływ wybranych czynników na konsumpcję alkoholu wśród młodzieży w wieku 15 – 22 lat.

1. Wstęp

Według raportu Światowej Organizacji Zdrowia z 2007 roku spożywanie alkoholu znajduje się na trzecim miejscu wśród czynników ryzyka dla zdrowia społeczeństwa. Wyższe miejsca w haniebnym rankingu zajmują palenie tytoniu i nadciśnienie tętnicze. Wykazano ponadto, iż spożywanie alkoholu ma związek z ponad 60 rodzajami chorób i urazów. Na szybkość powstawania uzależnienia istotny wpływ wywiera stopień dojrzałości organizmu oraz najbliższe otoczenie. Dodatkowo z badań wynika, że znaczącą rolę w uzależnieniu odgrywa wiek, w którym rozpoczyna się intensywne picie alkoholu.

W związku z powyższym zdecydowaliśmy się zbadać wpływ wybranych czynników na konsumpcję alkoholu wśród młodzieży w wieku od 15 do 22 lat.

1.1 Cel pracy

Celem pracy jest przeprowadzenie analizy wielorakiej oraz skonstruowanie modelu regresji wielorakiej, który wykażę wpływ poszczególnych zmiennych objaśniających na konsumpcję alkoholu wśród młodzieży. Wyniki pozwolą ponadto na skonstruowanie interpretacji będącej podsumowaniem tematu.

1.2 Dane

Dane użyte w badaniu zostały pobrane ze strony www.kaggle.com będącej internetową społecznością naukowców zajmujących się danymi i praktyków uczenia maszynowego. Zostały one pozyskane przez Fabio Pagnotta i Hossaina Mohammada Amrana za pomocą ankiety przeprowadzonej wśród uczniów matematyki w szkole ponadpodstawowej. Arkusz na stronie internetowej zawiera dane o 33 cechach przypisanych do 395 studentów, lecz na potrzeby naszego badania zdecydowaliśmy się o subiektywne zredukowanie ilości cech do 5. Wybrane przez nas cechy związane są z warunkami w życiu osobistym oraz szkołą.

Wśród zbioru znajdują się 2 cechy ilościowe i 3 jakościowe. Na potrzeby badania zdecydowano o przekształceniu zmiennej jakościowej Walc (wielkość konsumpcji alkoholu w weekend) w taki sposób, aby była ona ponadto zmienną dychotomiczną. Celem przekształcenia jest użycie zmiennej Walc jako zmiennej objaśnianej w dwumianowym modelu logitowym, którego założenie wskazuje na użycie jako zmiennej objaśnianej zmiennej dychotomicznej. Zmienna objaśniana jest zmienną jakościową i dychotomiczną, w związku z czym przyjmuje

dwie wartości – 0 lub 1. Początkowo zmienna przyjmowała wartości od 1 (bardzo małe spożycie) do 5 (bardzo duże spożycie), lecz zostały one sprowadzone do opisanych wartości za pomocą mediany, która wynosiła 2. Rozgraniczenie polegało na przypisaniu wartości 0 wszystkim przypadkom poniżej mediany oraz wartości 1 przypadkom powyżej mediany. Poniżej znajdują się pełne opisy dobranych zmiennych wraz ze skalą oraz objaśnieniem.

- Tygodniowy czas poświęcony na nauce (studytime) – zmienna ilościowa wyrażona w godzinach
- Relacje z rodziną (famrel) – zmienna jakościowa wyrażona w skali od 1 (bardzo złe) do 5
- Ilość czasu poświęconego na wyjścia ze znajomymi (goout) – zmienna jakościowa wyrażona w skali od 1 (bardzo mało) do 5 (bardzo dużo)
- Liczba nieobecności w szkole (absences) – zmienna ilościowa wyrażona w dniach
- Wielkość konsumpcji alkoholu w weekend (Walc) – zmienna jakościowa, w której 0 oznacza brak spożycia alkoholu w weekend, natomiast 1 oznacza spożywanie alkoholu w weekend

1.3 Problem badawczy

Przygotowywany projekt ma na celu ukazanie, w jaki sposób poszczególne czynniki wpływają na weekendową konsumpcję alkoholu wśród młodzieży. Rozwiązanie problemu badawczego pomoże wskazać oraz podsumować siłę wpływu wybranych czynników na wzrost konsumpcji alkoholu, co przełoży się na sformułowanie odpowiednich wniosków w podsumowaniu.

1.4 Hipotezy badawcze

W celu dokonania badania, na podstawie poprzedniego punktu sformułowane zostały następujące hipotezy badawcze:

1. Wraz ze wzrostem tygodniowego czasu na nauce weekendowe spożycie alkoholu maleje z powodu mniejszej ilości wolnego czasu.
2. Lepsze relacje z rodziną pozwalają na uniknięcie używania nadmiaru używek w młodym wieku.
3. Większa ilość czasu poświęcona na wyjścia ze znajomymi doprowadza do większego spożycia alkoholu, co może być spowodowane presją otoczenia.

4. Wraz ze wzrostem liczby nieobecności w szkole zwiększa się weekendowa konsumpcja alkoholu poprzez mniejszy nacisk na edukację i obowiązki.

2. Dwumianowy model logitowy i probitowy

2.1 Korelacje między zmiennymi

Potencjalne zmienne objaśniające, które zostały wybrane do dwumianowego modelu logitowego, muszą spełniać założenie o braku występowania korelacji pomiędzy sobą, tzn. muszą być niezależne od siebie. W tabeli 1 przedstawiona została macierz korelacji między wstępnie wybranymi zmiennymi.

Tab. 1 Korelacje pomiędzy zmiennymi objaśniającymi.

Zmienna	**sex**	age	studytime	failures	famrel	freetime	goout	absences	G3
sex	1,000000	0,028606	0,306268	-0,044436	-0,058971	-0,238744	-0,075897	0,066962	-0,103456
age	0,028606	1,000000	-0,004140	0,243665	0,053940	0,016434	0,126964	0,175230	-0,161579
studytime	0,306268	-0,004140	1,000000	-0,173563	0,039731	-0,143198	-0,063904	-0,062700	0,097820
failures	-0,044436	0,243665	-0,173563	1,000000	-0,044337	0,091987	0,124561	0,063726	-0,360415
famrel	-0,058971	0,053940	0,039731	-0,044337	1,000000	0,150701	0,064568	-0,044354	0,051363
freetime	-0,238744	0,016434	-0,143198	0,091987	0,150701	1,000000	0,285019	-0,058078	0,011307
goout	-0,075897	0,126964	-0,063904	0,124561	0,064568	0,285019	1,000000	0,044302	-0,132791
absences	0,066962	0,175230	-0,062700	0,063726	-0,044354	-0,058078	0,044302	1,000000	0,034247
G3	-0,103456	-0,161579	0,097820	-0,360415	0,051363	0,011307	-0,132791	0,034247	1,000000

Źródło: Opracowanie własne z wykorzystaniem programu Statistica.

Na podstawie macierzy korelacji z tabeli 1 zdecydowano o usunięciu części zmiennych i nieuwzględnieniu ich w modelu logitowym, ponieważ wykazują one korelacje między sobą.

Ostatecznie wybrane zmienne to opisane w rozdziale pierwszym: studytime, famrel, goout i absences. W tabeli 2 przedstawiono korelacje między tymi zmiennymi, jak widać spełniają one założenie o braku występowania znaczącej korelacji między sobą. Można więc na ich podstawie utworzyć model logitowy dwumianowy.

Tab. 2 Korelacje pomiędzy wybranymi zmiennymi objaśniającymi.

Zmienna	**studytime**	famrel	goout	absences
studytime	1,000000	0,039731	-0,063904	-0,062700
famrel	0,039731	1,000000	0,064568	-0,044354
goout	-0,063904	0,064568	1,000000	0,044302
absences	-0,062700	-0,044354	0,044302	1,000000

Źródło: Opracowanie własne z wykorzystaniem programu Statistica.

2.2 Dwumianowy model logitowy

Wykorzystując oprogramowanie RStudio stworzony został dwumianowy model logitowy. Tabela 3 przedstawia estymację modelu dwumianowego logitowego dla zmiennej dychotomicznej, która wyjaśnia wielkość konsumpcji alkoholu w weekend.

Tab. 3 Estymacja modelu dwumianowego logitowego.

	parametr B	Błąd standardowy	statystyka testu Walda	p-wartość
wyraz wolny	-1,0354322	0,65704719	-1,575887	0,12
studytime	-0,5068557	0,14698631	-3,448319	0,00
famrel	-0,2751372	0,12894311	-2,133788	0,03
goout	0,79759294	0,11312768	7,050378	0,00
absences	0,03411633	0,01411448	2,417115	0,02

Źródło: Opracowanie własne z wykorzystaniem programu RStudio.

Na podstawie wartości z tabeli numer 3 można stwierdzić, że parametry B różnią się statystycznie od zera, co oznacza, że zmienne objaśniające wybrane do modelu statystycznie istotnie wpływają na zmienną objaśnianą (wielkość konsumpcji alkoholu). Ponadto wartości p- value wynoszą mniej, niż założony poziom istotności (0,05), w związku z czym można odrzucić hipotezę zerową, według której wybrane zmienne nie wpływają na zmienną objaśnianą.

Dodatkowo można wyznaczyć postać modelu logitowego, która wygląda następująco:

$$\textbf{logit(p)} = \ln\left(\frac{p}{1-p}\right) = \textbf{-1,0354 – 0,5069*studytime}$$

– 0,2751*famrel

+ 0,7976*goout

+ 0,0341*absences

Zmienne w modelu nie powinny być współliniowe, w związku z czym następnym krokiem jest sprawdzenie braku współliniowości zmiennych, co zostało przedstawione w tabeli 4.

Tab. 4 Wartości VIF dla zmiennych objaśniających

	studytime	famrel	goout	absences
wartość VIF	1,004138	1,023731	1,029369	1,004773

Źródło: Opracowanie własne z wykorzystaniem programu RStudio.

Wartości VIF świadczą o braku współliniowości zmiennych, ze względu na to, iż nie przekraczają one umownej granicy dla modelów logitowych (VIF > 2,5). Dodatkowo można

stwierdzić, że pomiędzy zmiennymi występuje bardzo słaby związek, ponieważ wartości są bliskie 1.

2.3 Dwumianowy model probitowy

Za pomocą programu RStudio utworzony został dwumianowy model probitowy. Tabela 5 przedstawia estymacje modelu probitowego dla zmiennej objaśnianej Walc.

Tab. 5 Estymacja modelu dwumianowego probitowego.

	parametr B	Błąd standardowy	statystyka testu Walda	p-wartość
wyraz wolny	-0,6221954	0,393788455	-1,580024	0,14
studytime	-0,3043285	0,085877967	-3,543732	0,00
famrel	-0,1645493	0,076944624	-2,138542	0,03
goout	0,47881941	0,065245368	7,338749	0,00
absences	0,02043182	0,008463986	2,413971	0,02

Źródło: Opracowanie własne z wykorzystaniem programu RStudio.

Podobnie jak w modelu logitowym można zauważyć, że wybrane zmienne objaśniające istotnie wpływają na zmienną objaśnianą. Wynika to z różniących się statystycznie od 0 parametrów B oraz p-wartości mniejszych od przyjętego poziomu istotności na poziomie 0,05.

Postać estymowanego modelu probitowego jest następująca:

probit(p) = -0,6222 – 0,3043*studytime
– 0,1645*famrel
+ 0,4788*goout
+ 0,0204*absences

2.4 Porównanie dopasowania modelu logitowego i probitowego

W poniższej tabeli zawarte są wartości kryteriów dopasowania modelu logitowego oraz probitowego. Na podstawie wartości z tabeli 6 można stwierdzić, że pod względem dopasowania nieco lepszym modelem jest model logitowy, ponieważ ma niższą wartość kryterium Akaike oraz wyższe wartości miar pseudo-R2. W związku z tym, rozpatrywanym modelem w dalszej części będzie model logitowy.

Tab. 6 Wartości kryteriów dopasowania dla modelu logitowego i probitowego.

	kryterium AIC	R^2 McFaddena	R^2 Cragga-Uhlera
model logitowy	457,3801	0,1598174	0,2618221
model probitowy	457,5347	0,1595271	0,2613959

Źródło: Opracowanie własne z wykorzystaniem programu RStudio.

2.5 Porównanie jakości predykcji modelu logitowego i probitowego

Tabele 7 i 8 przedstawiają tablice trafności modeli dla p*=0,5. Można zauważyć, iż dla obydwu modeli wartości nietrafionych jest tyle samo, w związku z tym na podstawie tablicy trafności niemożliwe jest wybranie lepszego modelu.

Tab. 7 Tablica trafności dla p=0.5 w modelu logitowym.*

	wartości przewidywane	
zaobserwowane	0	1
0	190	46
1	73	86

Źródło: Opracowanie własne z wykorzystaniem programu RStudio.

Tab. 8 Tablica trafności dla p=0.5 w modelu probitowym.*

	wartości przewidywane	
zaobserwowane	0	1
0	190	46
1	73	86

Źródło: Opracowanie własne z wykorzystaniem programu RStudio

W tabeli 9 przedstawiono miary oparte na tablicy trafności dla p*=0,5, według których skonstruowano następujące wnioski:

- **ACC** (zliczeniowy R^2)– udział liczby odpowiednio sklasyfikowanych jednostek w ogólnej liczbie jednostek wynosi 69,87%
- **ER** (wskaźnik błędu) – udział błędnie przypisanych jednostek w ogólnej liczbie jednostek wynosi 30,13%
- **SENS** (czułość) – udział liczby odpowiednio sklasyfikowanych 1 w ogólnej liczbie oszacowanych 1 wynosi 54,09%
- **SPEC** (swoistość) – udział liczby odpowiednio oszacowanych 0 w ogólnej liczbie oszacowanych 0 wynosi 80,51%
- **PPV** (dodatnia zdolność predykcyjna) – udział liczby odpowiednio oszacowanych 1 w liczbie wszystkich prognozowanych 1 wynosi 65,15%

- **NPV** (ujemna zdolność predykcyjna) – udział liczby odpowiednio oszacowanych 0 w liczbie wszystkich prognozowanych 0 wynosi 72,24%

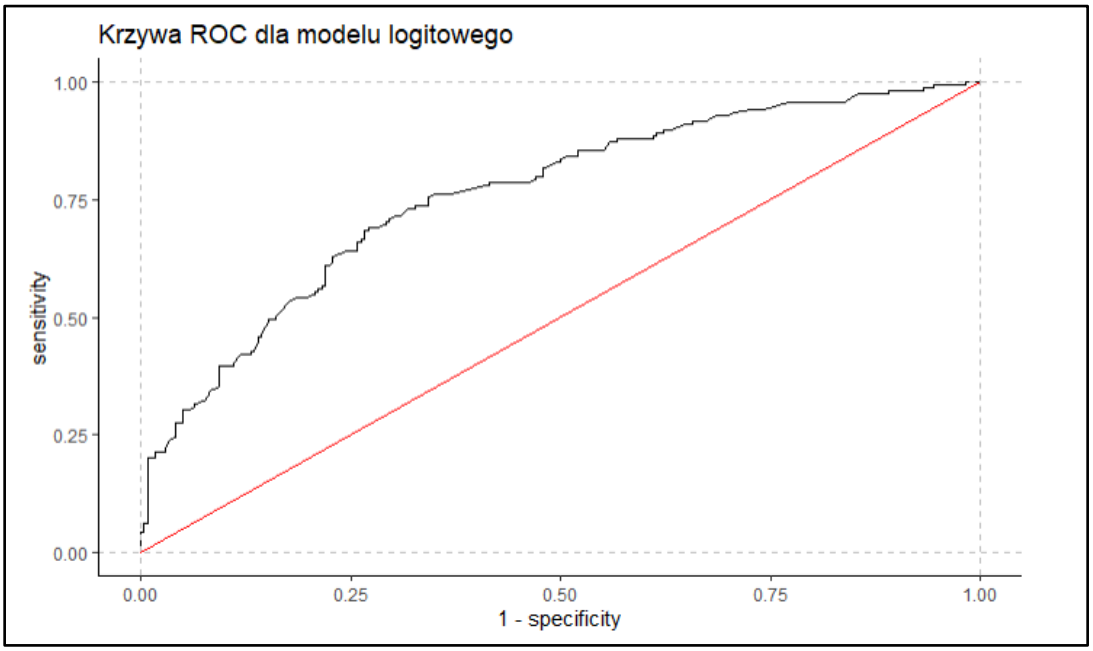
Tab. 9 Miary oparte na tablicy trafności dla p=0.5 w modelu logitowym.*

P*=0,5	ACC	ER	SENS	SPEC	PPV	NPV
model logitowy	0,698734	0,301266	0,540881	0,805085	0,651515	0,722434

Źródło: Opracowanie własne z wykorzystaniem programu RStudio.

Na podstawie powyższych wniosków można przyjąć, że punkt odcięcia p*=0,5 nie jest do końca odpowiedni, ponieważ jedynie w 54% określa odpowiednio sklasyfikowane 1 w ogólnej liczbie oszacowanych 1. Ponadto rysunek 1 wykazuje brak stanowczej wypukłości wykresu, co świadczy o złym doborze punktu odcięcia. Pole AUC, które jest polem powierzchni pod wykresem krzywej ROC przyjmujące wartości w przedziale od 0 do 1 wynosi 0,76, natomiast większość testów w diagnostyce reprezentuje moc diagnostyczną wyrażającą się wielkościami AUC pomiędzy 0,80 a 0,95.

Rysunek 1 Krzywa ROC dla modelu logitowego.



Źródło: Opracowanie własne z wykorzystaniem programu RStudio.

Pozytywny fakt, to określenie w ponad 80% udziału liczby odpowiednio sklasyfikowanych 0 w ogólnej liczbie oszacowanych 0 oraz pokrycie w 2/3 trafnie sklasyfikowanych jednostek w ich ogólnej liczbie. W celu poprawienia czułości, zdecydowano o skonstruowanie tablicy trafności dla p*=0,42, której wyniki przedstawiono w tabeli 10.

Tab. 10 Miary oparte na tablicy trafności dla p=0,42 w modelu logitowym.*

p*=0,42	ACC	ER	SENS	SPEC	PPV	NPV
model logitowy	0,7063291	0,2936709	0,6666667	0,7330508	0,6272189	0,7654867

Źródło: Opracowanie własne z wykorzystaniem programu RStudio.

W przypadku zastosowania punktu odcięcia p*=0,42 można zauważyć, iż udział liczby odpowiednio sklasyfikowanych jednostek w ogólnej liczbie jednostek wzrósł do 70,63%, natomiast ER określający wskaźnik błędu zmniejszył się do 29,37%. Ponadto czułość wzrosła do 66,67%, natomiast swoistość zmniejszyła się do 73,31%.