

WOJSKOWA AKADEMIA TECHNICZNA
im. Jarosława Dąbrowskiego
WYDZIAŁ CYBERNETYKI



SPRAWOZDANIE
Metody Eksploracji Danych

Temat laboratorium: **MODELE LOGITOWE. REGRESJA LOGISTYCZNA**

INFORMATYKA

(kierunek studiów)

INŻYNIERIA SYSTEMÓW – ANALIZA DANYCH

(specjalność)

Zespół:

Michał ŚLĘZAK
Szymon OLEŚKIEWICZ

Prowadzący laboratorium:

Dr inż. Romuald Hoffmann, prof.
WAT

Warszawa 2025

Spis treści

Wstęp	4
Rozdział I. Podstawy teoretyczne.....	5
I.1. Model liniowy.....	5
I.2. Model logitowy	6
I.3. Ocena istotności zmiennych – test Walda.....	7
I.4. Metryki oceny jakości klasyfikacji modeli	7
I.5. Wykorzystane narzędzia	10
Rozdział II. Opis problemu	11
II.1. Treść zadania	11
Rozdział III. Implementacja modeli	12
III.1. Model liniowy	12
III.2. Model logitowy	14
Podsumowanie.....	18
Bibliografia	19
Spis tabel	20
Spis rysunków	20
Załączniki	21

Wstęp

Celem niniejszego sprawozdania jest porównanie modelu liniowego oraz logitowego w określaniu zależności stanu cywilnego w grupie studentów studiów jednolitych magisterskich z warszawskich uczelni

W pierwszej kolejności zbudowano model liniowy prawdopodobieństwa oparty o klasyczną Metodę Najmniejszych Kwadratów. Następnie, zastosowano model logitowy.

Rozdział I. Podstawy teoretyczne

I.1. Model liniowy

Model liniowy jednej zmiennej wyraża się poniższym wzorem.

$$\hat{y} = a_1 x + a_0 \quad (1)$$

Należy wyznaczyć takie wartości parametrów a_1 i a_0 , aby najlepiej dopasować przebieg prostej regresji do wartości danych. Aby tego dokonać, należy rozwiązać poniższe zadanie optymalizacyjne.

$$M(a_0, a_1) = \sum_{i=1}^n [y_i - (a_1 x_i + a_0)]^2 \quad (2)$$

Przyrównując pochodne funkcji sumy kwadratów reszt po zmiennych a_0 i a_1 do zera, otrzymamy układ równań, który w postaci macierzowej będzie mieć następującą postać.

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix} \quad (3)$$

Po lewostronnym przemnożeniu obu stron równania przez macierz odwrotną do stojącej przy macierzy parametrów, otrzymamy wzór na wartość macierzy tych parametrów.

$$\begin{bmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum_{i=1}^n x_i y_i \\ \sum_{i=1}^n y_i \end{bmatrix} = \begin{bmatrix} a_1 \\ a_0 \end{bmatrix} \quad (4)$$

Powyższy tok rozumowania można uogólnić dla regresji wielu zmiennych. W tym celu można wykorzystać zapis macierzowy, gdzie po przekształceniach otrzymamy ostatecznie wzór na wartości współczynników dla dowolnej liczby parametrów.

$$A = \begin{bmatrix} a_0 \\ \vdots \\ a_k \end{bmatrix} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y \quad (5)$$

W powyższym wzorze macierze X oraz Y zostały wyznaczone na podstawie zbioru danych, takich jak poniżej.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \dots \\ y_k \\ \dots \\ y_n \end{bmatrix} \quad (6)$$

To właśnie m.in. tym przypadkiem będziemy się zajmować w ramach tego laboratorium – modelem liniowym prawdopodobieństwa w celu klasyfikacji stanu cywilnego studentów w zależności od roku studiów.

W tym wypadku nasz model będzie wyglądał jak poniżej.

$$P(Y = 1|X) = a_0 + a_1 X \quad (7)$$

Gdzie parametr a_1 jest zmianą prawdopodobieństwa sukcesu przy wzroście X o jednostkę. Należy mieć jednak na uwadze istotne ograniczenia modelu liniowego prawdopodobieństwa. Głównym z nich jest to, że może on generować wartości prawdopodobieństwa mniejsze od 0 lub większe od 1, co w praktyce jest niemożliwe.

I.2. Model logitowy

Model logitowy przewiduje logarytm szans. W naszym przypadku, prawdopodobieństwo zawarcia związku małżeńskiego oznaczono jako p a rok studiów jako r.

$$\hat{L} = \ln \frac{\hat{p}}{1 - \hat{p}} = b_1 X + b_0 = b_1 r + b_0 \quad (8)$$

Gdzie p to prawdopodobieństwo sukcesu (bycia w związku małżeńskim), iloraz w logarytmie naturalnym to iloraz szans (stosunek prawdopodobieństwa sukcesu do prawdopodobieństwa porażki) a L to nasza liniowa postać modelu logitowego po oszacowaniu. Parametr b_1 mówi nam o zmianie logarytmu szans. Wskazuje on nam, ile razy zmieni się szansa na to, że student jest w związku małżeńskim przy wzroście roku studiów o 1. Warto zaznaczyć, że stosując przekształcenie odwrotne do transformacji powyższego modelu, otrzymamy oszacowanie prawdopodobieństwa p w postaci funkcji logistycznej.

$$\hat{p} = \frac{1}{1 + e^{-\hat{L}}} = \frac{1}{1 + e^{-(b_1X + b_0)}} \quad (9)$$

I.3. Ocena istotności zmiennych – test Walda

Do oceny istotności statystycznej poszczególnych zmiennych w modelu wykorzystano test Walda. Statystyka z sprawdza hipotezę zerową $H_0: b_i = 0$. Jeśli wartość (p-value, w naszym programie oznaczane jako $P > |z|$) jest mniejsza od przyjętego poziomu istotności (w naszym wypadku $\alpha = 0,05$), zmienną uznaje się za istotną statystycznie. W przypadku modelu liniowego prawdopodobieństwa zasada jest taka sama, ale zamiast testu Walda stosuje się test T-Studenta oraz statystykę t zamiast z . Wartości krytyczne statystyk obliczono z wykorzystaniem biblioteki Python – `scipy`. Statystyka Walda W_i jest równa kwadratowi z . Warto dodać, że statystyka Walda ma w przybliżeniu rozkład chi-kwadrat z liczbą stopni swobody równą 1 – to właśnie z tablic rozkładu chi-kwadrat będziemy brać wartość krytyczną.

Do weryfikacji skuteczności modeli w procesie klasyfikacji stanu cywilnego studentów wykorzystano wiele metryk, których implementacja znajduje się w bibliotekach Pythona. Metryki wykorzystane w obliczeniach znajdują się poniżej.

I.4. Metryki oceny jakości klasyfikacji modeli

Macierz pomyłek, która jest zestawieniem wartości rzeczywistych naszego zbioru danych z wartościami przewidywanymi przez nasz model. Ma postać tabeli kwadratowej, w której wyróżnia się cztery główne wartości.

Confusion Matrix

		Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)	
	False Negatives (FNs)	True Negatives (TNs)	

Rys. 1. Macierz pomyłek

1. True Positive – prawdziwie dodatnie: liczba przypadków, w których model poprawnie przewidział sukces (student jest w związku małżeńskim i model to potwierdził)
2. True Negative – prawdziwie ujemne: liczba przypadków, w których model poprawnie przewidział brak zdarzenia (student jest w stanie wolnym i model to potwierdził).
3. False Positive – fałszywie dodatnie: model przewidział sukces, mimo że w rzeczywistości zdarzenie nie miało miejsca (student w rzeczywistości jest wolny a model przewidział, że jest w związku małżeńskim).
4. False Negative – fałszywie ujemne: model przewidział brak zdarzenia, mimo że w rzeczywistości ono wystąpiło (student w rzeczywistości jest w związku małżeńskim a model przewidział, że jest wolny).

Na podstawie tych wartości są obliczane kolejne kluczowe metryki służące do oceny jakości modelu.

Dokładność (Accuracy) ukazuje nam ogólny odsetek poprawnych klasyfikacji (ile ogółem sklasyfikowano poprawnie).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

Czułość (Recall) ukazuje nam zdolność modelu do wykrywania klasy pozytywnej – ile trafiłem z rzeczywistych małżeństw (student w związku małżeńskim).

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

Precyzja (Precision) ukazuje nam wiarygodność klasyfikacji pozytywnych – ile trafiłem naprawdę małżeństw.

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

Specyficzność (Specificity) ukazuje nam zdolność do wykrywania klasy negatywnej- ile trafiłem z rzeczywistych wolnych studentów (student nie jest w związku małżeńskim).

$$Specificity = \frac{TN}{TN + FP} \quad (13)$$

F1-Score jest to średnia harmoniczna precyzji i czułości, która przydaje nam się w szczególności w niezbalansowanych klasach.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (14)$$

Dodatkowo, wykorzystano analizę krzywej ROC (Receiver Operating Characteristic). Jest ona graficzną reprezentacją skuteczności modelu klasyfikującego dla wszystkich możliwych progów klasyfikacji. Przedstawia ona zależność między 2 parametrami – czułością (True Positive Rate) oraz swoistością (False Positive Rate). Wykorzystano także metrykę AUC, która jest polem pod krzywą ROC i która mówi nam o jakości modelu przed wyborem konkretnego progu klasyfikacji. Wykorzystując właśnie krzywą ROC można wyznaczyć optymalny próg klasyfikacji poprzez maksymalizację odległości (różnicy) TPR i FPR.

Odchylenie standardowe modelu S_e uzyskamy pierwiastkując wariancję S_e^2 .

$$S_e^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k - 1} \quad (15)$$

Na podstawie wartości odchylenia standardowego S_e modelu jednej zmiennej można wyznaczyć odchylenia parametrów a_0 i a_1 , tak jak poniżej.

$$S_{a_0} = S_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x}_i)^2}} \quad (16)$$

$$S_{a_1} = \frac{S_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_i)^2}} \quad (17)$$

W celu zbadania jak dobrze wyznaczona prosta regresji oddaje przebieg zmiennej można wykorzystać współczynnik determinacji, który obliczany jest na podstawie poniższego wzoru. Współczynnik ten przyjmuje wartości od 0 do 1. Im bliżej 1, tym nasz model jest bardziej dopasowany.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (18)$$

Współczynnik Pseudo R^2 (McFaddena) pokazuje nam „dopasowanie modelu do danych”. W regresji logistycznej jest to miara opierająca się na funkcji wiarygodności (Likelihood), która informuje nas o tym, o ile lepszy jest nasz model od modelu zerowego, czyli takiego zawierającego tylko wyraz wolny.

$$\text{Pseudo } R^2 = 1 - \frac{\ln L_M}{\ln L_0} \quad (19)$$

Gdzie w liczniku mamy logarytm funkcji wiarygodności dla modelu ze zmiennymi objaśniającymi a w mianowniku logarytm funkcji wiarygodności dla modelu tylko z wyrazem wolnym. Wzór na funkcję wiarygodności jest taki jak poniżej.

$$L(\theta|y, X) = \prod_i P(y_i|X_i, \theta)^{y_i} \cdot P(y_i = 1|X_i, \theta)^{1-y_i} \quad (20)$$

Gdzie:

- θ – wektor wartości parametrów modelu
- y – wektor wartości zmiennej objaśnianej przyjmującej wartości 0 lub 1
- X – macierz zmiennych objaśniających i-numer obserwacji
- $P(y_i = 1|X_i, \theta)$ – prawdopodobieństwo, że dla i-tej obserwacji y przyjmuje wartość 1, przy założeniu wartości zmiennej niezależnej X_i oraz wartości parametrów regresji logistycznej.

W procesie budowania modeli, nie dzielono danych na zbiór testowy i treningowy. Potraktowano cały zbiór danych jako dane treningowe ze względu na bardzo małą licznosć zbioru danych.

I.5. Wykorzystane narzędzia

W celu wyznaczenia modeli liniowych wraz z ich parametrami na podstawie wykładów oraz wiedzy własnej przygotowano program w języku *Python* wykorzystujący biblioteki:

- **Pandas,**
- **NumPy,**
- **stastmodel,**
- **scikit-learn,**
- **scipy,**

które również zostały wykorzystane do obliczeń oraz wizualizacji wyników. Modele oraz ich parametry zostały wyznaczone za pomocą wyżej wspomnianych narzędzi, które oparte są o metodę MNK (Metoda Najmniejszych Kwadratów minimalizująca sumę kwadratów reszt RSS), gdzie wraz z dopasowaniem danych, wyznaczono wiele parametrów i metryk dla danego stworzonego modelu.

Rozdział II. Opis problemu

II.1. Treść zadania

W wybranej losowo grupie studentów jednolitych studiów magisterskich z warszawskich uczelni badano ich stan cywilny w zależności od roku studiów. Wyniki obserwacji zebrano w tabeli 1. Przyjęto przez M oznaczać osoby będące w związku małżeńskim, natomiast przez W – osoby stanu wolnego.

Tab. 1. Dane dotyczące stanu cywilnego badanych studentów w zależności od roku studiów

Rok studiów	1	2	5	1	4	3	2	1	5	2	3	4	1	2
Stan cywilny	W	W	M	W	M	W	M	W	M	W	M	M	W	W
Rok studiów	5	4	3	1	4	5	2	5	3	4	3	2	5	1
Stan cywilny	M	M	M	M	M	W	W	M	W	W	M	W	M	W

M oznacza osoby będące w związku małżeńskim, natomiast przez W – osoby stanu wolnego

W zadaniu proszę:

1. Wyznaczyć zależność stanu cywilnego badanych studentów od roku studiów, zakładając najpierw liniowy model prawdopodobieństwa, a następnie model logitowy.
2. Na podstawie opracowanych modeli i przeprowadzonych obliczeń sformułować własne wnioski.
3. Wyniki analizy proszę zatrzymać w postaci sprawozdania, do którego proszę dodać jako załączniki wszystkie pliki z obliczeniami (obliczenia można przeprowadzić w dowolnie wybranym narzędziu)

Rozdział III. Implementacja modeli

III.1. Model liniowy

Pierwszym modelem, który zbudowano jest model liniowy prawdopodobieństwa. Zmienną objaśniającą jest w tym przypadku rok studiów. Przyjęto **próg klasyfikacji na poziomie 0,5**. Wyznaczono **współczynnik przy zmiennej objaśniającej** oraz **wyraz wolny**, które są równe odpowiednio **0,1950** oraz **-0,0781**. Na podstawie testu T-Studenta oraz wartości poziomu istotności (p-value), zmienna objaśniająca – rok studiów jest istotna statystycznie, gdyż wartość p-value (0,002) jest mniejsza od zadanego poziomu istotności na poziomie 0,05.

Tab. 2. Tabela z parametrami modelu liniowego

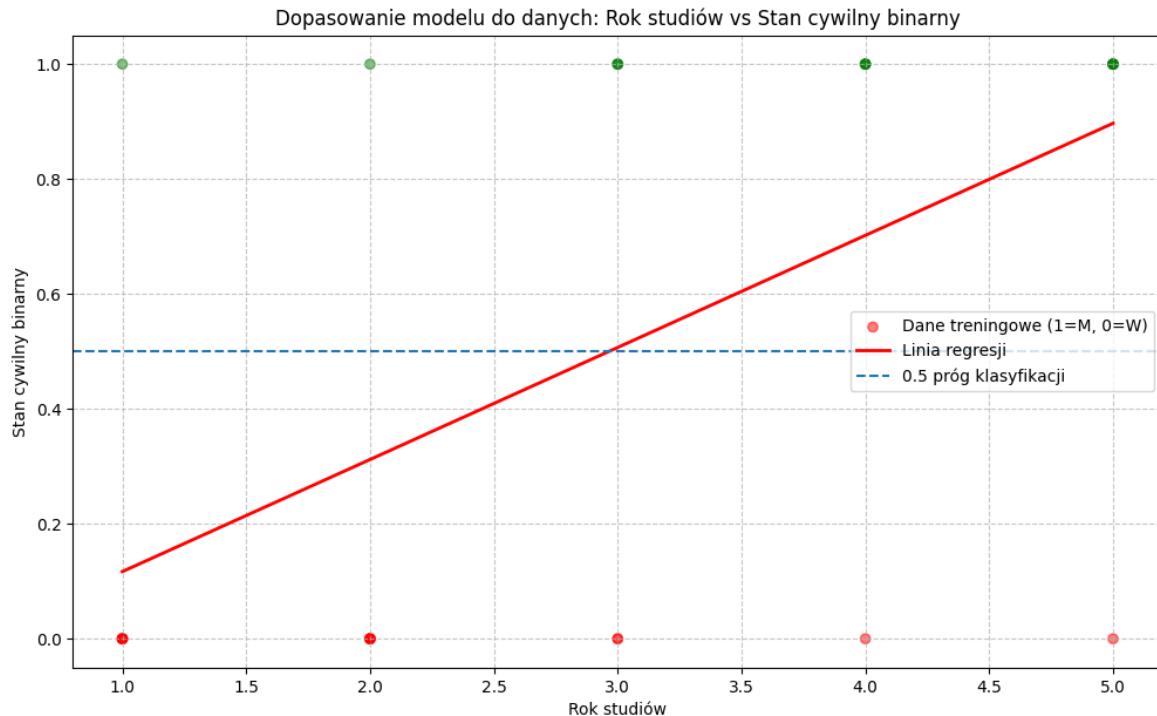
Parametr a_i	Odchylenie $S(a_i)$	$ t > t_{\text{kryt}} = 2,055$
$a_0 = -0,0781$	0,184	0,425
$a_1 = 0,1950$	0,056	3,501 – istotny

Otrzymany model:

$$P(Y = 1|X) = -0,0781 + 0,1950X$$

Metryki dopasowania:

$R^2 = 32\%$



zmiennej objaśniającej – Rok studiów, która okazała się być istotna statystycznie, gdyż wartość statystyki $|t|$ jest większa od wartości krytycznej.

Metryki oceny jakości klasyfikacji:

$$\text{Accuracy} \approx 78,6\%$$

$$\text{Precision} \approx 75\%$$

$$\text{Recall} \approx 85,7\%$$

$$\text{Specificity} \approx 71,4\%$$

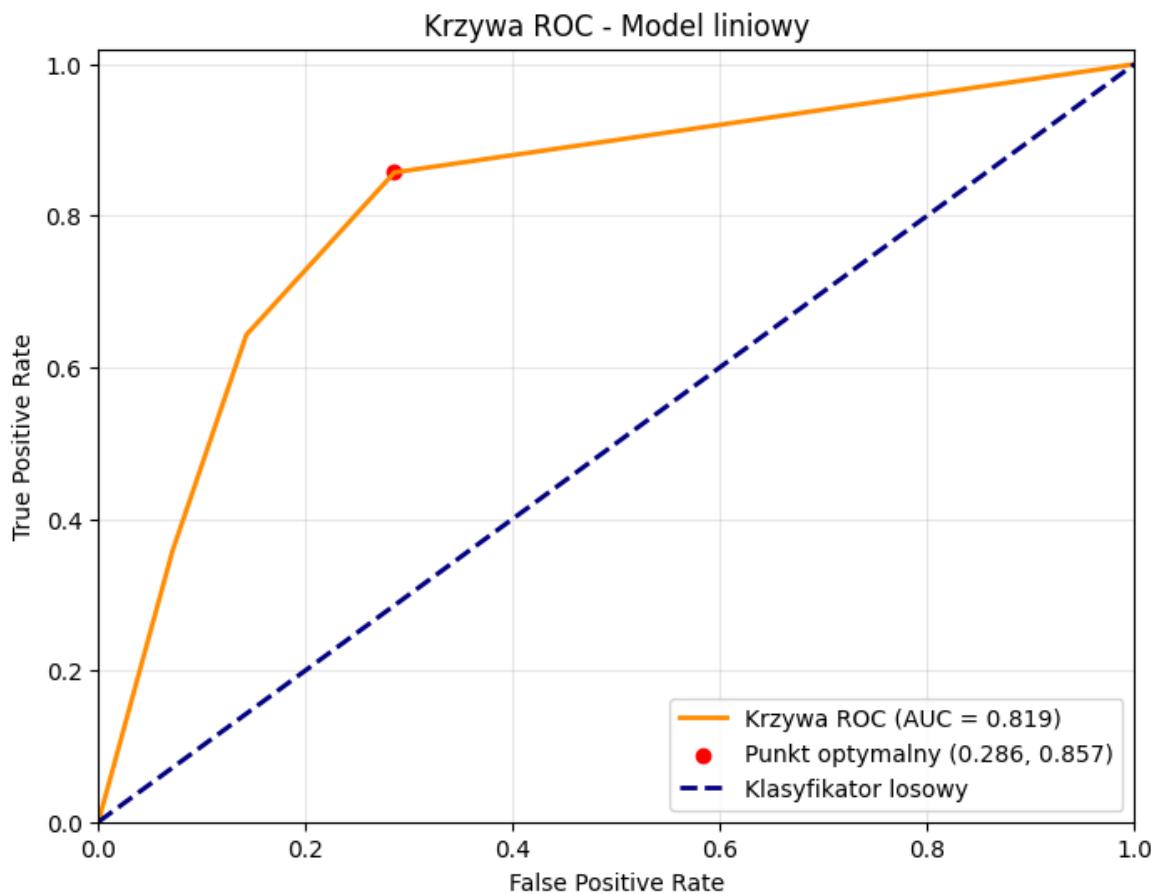
$$F1 \approx 80\%$$

$$AUC \approx 0,82$$

Tab. 3. Macierz pomyłek modelu liniowego prawdopodobieństwa

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 12	FP = 4
Przewidziana 0	FN = 2	TN = 10

Model liniowy prawdopodobieństwa osiąga dobre wyniki, aczkolwiek nie są one idealne. Jedną z lepszych metryk do oceny jakości klasyfikacji jest czułość ($\sim 85,7\%$), która mówi jak dobrze model klasyfikuje rzeczywiste małżeństwa. Model skutecznie klasyfikuje studentów w związku małżeńskim, chociaż ma tendencję do przeklasyfikowania studentów jako małżeństwa jeśli pod uwagę weźmiemy precyzję na poziomie $\sim 75\%$. Podsumowując, mając na uwadze powyższe metryki oraz wykres dopasowania, można zauważyc zależność: im wyższy rok studiów, tym większe prawdopodobieństwo, że student jest w związku małżeńskim.



Rys. 3. Wykres krzywej ROC dla modelu liniowego prawdopodobieństwa

Rysując krzywą ROC wyznaczono także optymalny próg klasyfikacji (0,507), który w przybliżeniu jest równy naszemu, odgórnie przyjętemu, równemu 0,5.

III.2. Model logitowy

Kolejnym modelem, który zbudowano jest model logitowy. Prawdopodobieństwo zawarcia związku małżeńskiego oznaczono jako p a rok studiów jako r .

$$\hat{L} = \ln \frac{\hat{p}}{1 - \hat{p}} = b_1 r + b_0 \quad (21)$$

Poniżej przedstawiono parametry i statystyki wraz z metrykami, które zostały wyznaczone dla tego konkretnego modelu. Skorzystano z biblioteki języka Python, która korzysta z metody maksymalnej estymacji prawdopodobieństwa. Przyjęto poziom istotności na poziomie $\alpha = 0,05$. Wartość wartości krytycznej statystyki Walda dla 1 stopnia swobody i zadanego poziomu istotności odczytano za pomocą biblioteki scipy - $\chi^2_{kryt} = 3,841$.

Tab. 4. Tabela z parametrami dla modelu logitowego

Parametr b_i	Odchylenie S(b_i)	p-value < 0,05	$W_i > \chi^2_{kryt}$
$b_0 = -2,8833$	1,171	0,014 – istotny	6,061 – istotny
$b_1 = 0,9770$	0,368	0,005 – istotny	7,049 – istotny

Otrzymany model:

$$\hat{L} = 0,9770r - 2,8833$$

$$Pseudo R^2 = 25,7\%$$

Metryki oceny jakości klasyfikacji:

$$Accuracy \approx 78,6\%$$

$$Precision \approx 75\%$$

$$Recall \approx 85,7\%$$

$$Specificity \approx 71,4\%$$

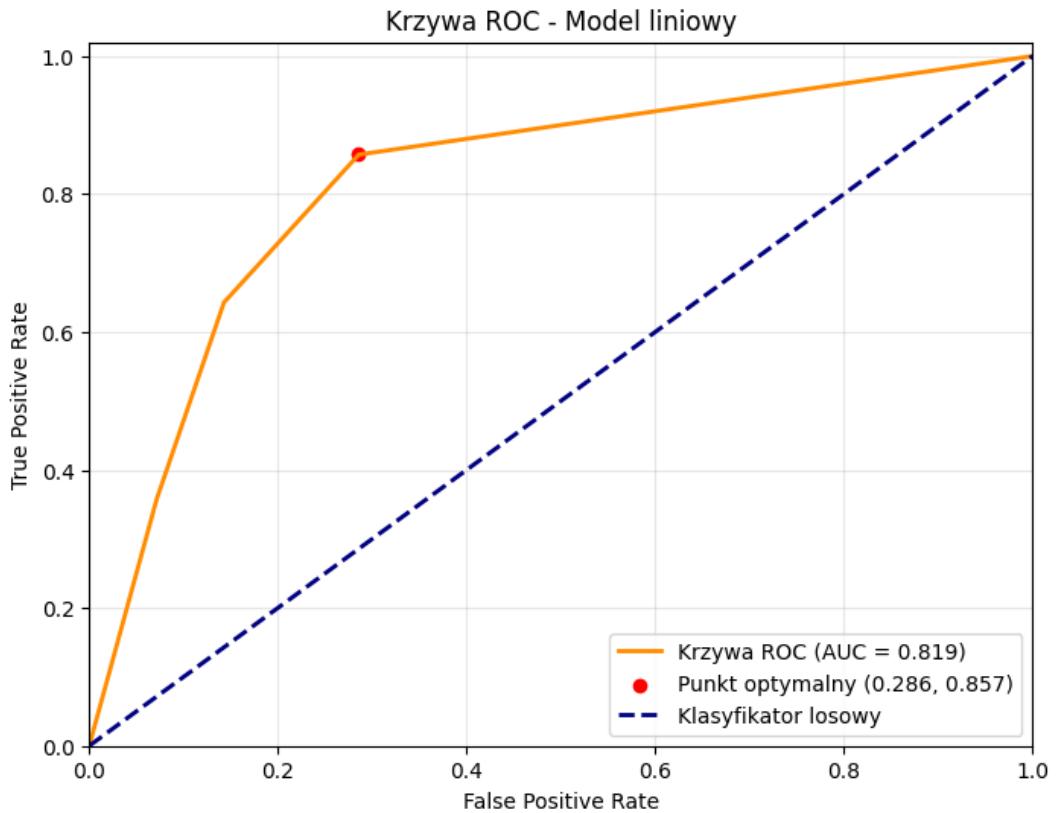
$$F1 \approx 80\%$$

$$AUC \approx 0,82$$

	Rzeczywista 1	Rzeczywista 0
Przewidziana 1	TP = 12	FP = 4
Przewidziana 0	FN = 2	TN = 10

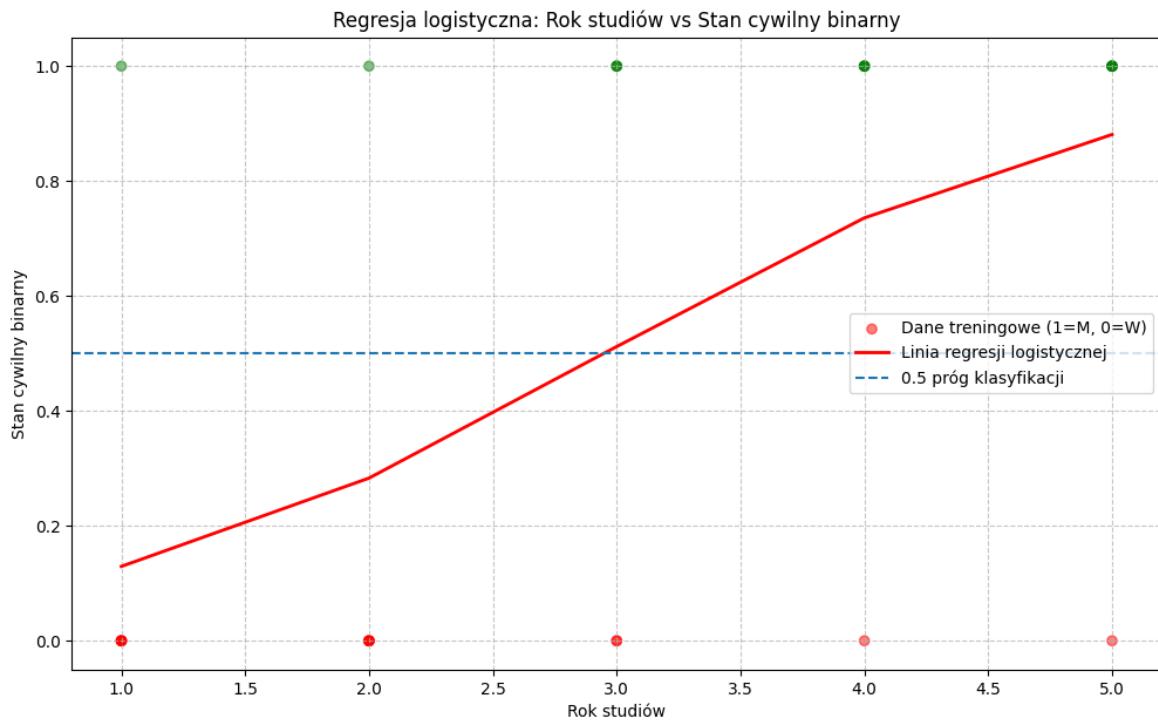
Tab. 5. Macierz pomylek modelu logitowego

Jak widać, dla modelu logitowego otrzymaliśmy te same wartości metryk jak dla modelu liniowego prawdopodobieństwa. Model logitowy ma jedynie gorszą metrykę Pseudo R^2 . Wniosek jest jeden, nie ma sensu budować dla tych danych modelu logitowego, bo jest bardziej skomplikowany a model liniowy daje te same wyniki de facto i nieco lepiej się dopasowuje do danych.



Rys. 4. Wykres krzywej ROC dla modelu logitowego

Rysując krzywą ROC wyznaczono także optymalny próg klasyfikacji (0,512), który w przybliżeniu jest równy naszemu, odgórnie przyjętemu, równemu 0,5.



Rys. 5. Wykres linii regresji logistycznej dla modelu logitowego

Podsumowanie

W wyniku analizy zbudowanych modeli – modelu liniowego prawdopodobieństwa i modelu logitowego, stwierdzono że model liniowego prawdopodobieństwa w zupełności wystarczy na potrzeby realizowanego zadania przy zadanych danych, gdyż model logitowy daje te same wartości metryk a jest nieco bardziej skomplikowany. Wykazuje się on dość dobrym dopasowaniem do danych oraz wysokimi metrykami oceny jakości klasyfikacji. Na podstawie przeprowadzonej analizy, zauważono zależność roku studiów od statusu cywilnego studentów – im wyższy rok studiów, tym większe prawdopodobieństwo, że student jest w związku małżeńskim.

Bibliografia

- [1] <https://en.wikipedia.org/wiki/Pseudo-R-squared>
- [2] <https://mirosławmamczur.pl/jak-dziala-regresja-logistyczna/>
- [3] <https://www.youtube.com/watch?v=idGS1gLIXTA>
- [4] <https://www.youtube.com/watch?v=2HpcV1sq0fE>
- [5] <https://en.wikipedia.org/wiki/Pseudo-R-squared>
- [6] https://pl.wikipedia.org/wiki/Funkcja_wiarygodno%C5%9Bci
- [7] https://www.naukowiec.org/wzory/statystyka/statystyka-walda--test_455.html

Spis tabel

Tab. 1.	Dane dotyczące stanu cywilnego badanych studentów w zależności od roku studiów	11
Tab. 2.	Tabela z parametrami modelu liniowego.....	12
Tab. 3.	Macierz pomyłek modelu liniowego prawdopodobieństwa	13
Tab. 4.	Tabela z parametrami dla modelu logitowego.....	15
Tab. 5.	Macierz pomyłek modelu logitowego.....	15

Spis rysunków

Rys. 1.	Macierz pomyłek.....	7
Rys. 2.	Wykres dopasowania modelu liniowego prawdopodobieństwa do danych	12
Rys. 3.	Wykres krzywej ROC dla modelu liniowego prawdopodobieństwa	14
Rys. 4.	Wykres krzywej ROC dla modelu logitowego	16
Rys. 5.	Wykres linii regresji logistycznej dla modelu logitowego.....	16

Załączniki

1. Plik notebook *Lab-2-Zadanie-1-Obliczenia.ipynb* – kompletny kod przeprowadzonych eksperymentów
- 2.