1. Cel

Celem laboratorium jest przeprowadzenie badań związanych z klasyfikacją binarną i wieloklasową, z zastosowaniem regresji logistycznej na wybranych zbiorach danych takich jak : Titanic (przewidywanie przeżycia pasażerów) oraz Iris (klasyfikacja gatunków irysów).

2. Materiały i metody

1. *Zbiory danych:*

    a) *Titanic: Zawiera dane na temat pasażerów statku Titanic.*
    Zmienna docelowa to "Survived", która wskazuje, czy pasażer przeżył (1) czy nie (0). Dane zawierają m.in. cechy takie jak wiek, klasa, płeć, liczba członków rodziny, cena biletu.

    b) *Iris: Klasyczny zbiór danych zawierający informacje o trzech gatunkach irysów (setosa, versicolor, virginica) opisanych za pomocą cech morfologicznych: długość i szerokość płatków oraz kielichów.*

2. *Techniki analizy:*

    a) *Regresja logistyczna: Stosowana zarówno do klasyfikacji binarnej (Titanic), jak i wieloklasowej (Iris).*

    b) *Przygotowanie danych: Wyczyszczenie danych (usunięcie braków i niepotrzebnych kolumn), kodowanie zmiennych kategorycznych, uzupełnianie brakujących wartości oraz podział na zbiór uczący i testowy.*

    c) *Ewaluacja: Wykorzystano krzywą ROC, krzywą precyzji/czułości, dokładność, macierz błędów oraz F1-score.*

3. Wyniki i dyskusja

Zadanie 1: Klasyfikacja binarna na zbiorze danych Titanic 1.

Analiza statystyk podstawowych i wizualizacja danych:

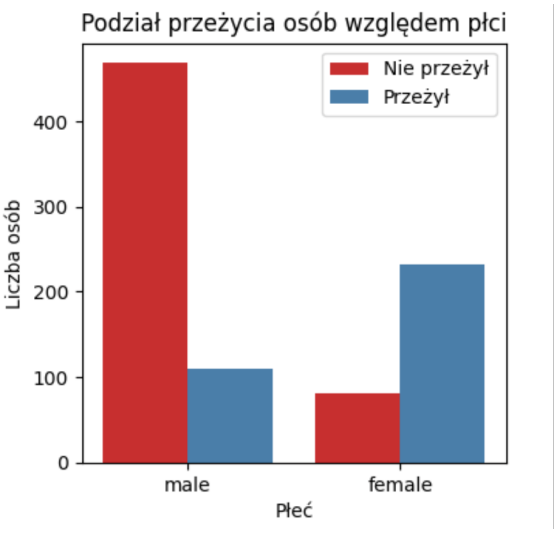
- *Rozkład cechy Survived:*

  Spośród wszystkich pasażerów 38% przeżyło katastrofę, co odzwierciedla stosunkowo niski odsetek pozytywnych przypadków (klasa 1). Nierównomierny rozkład klas może prowadzić do potencjalnych problemów z nierównowagą danych, gdzie model może być bardziej skłonny do przewidywania klasy negatywnej (klasa 0).

	PassengerId	Survived
count	891.000000	891.000000
mean	446.000000	0.383838
std	257.353842	0.486592
min	1.000000	0.000000
25%	223.500000	0.000000
50%	446.000000	0.000000
75%	668.500000	1.000000
max	891.000000	1.000000

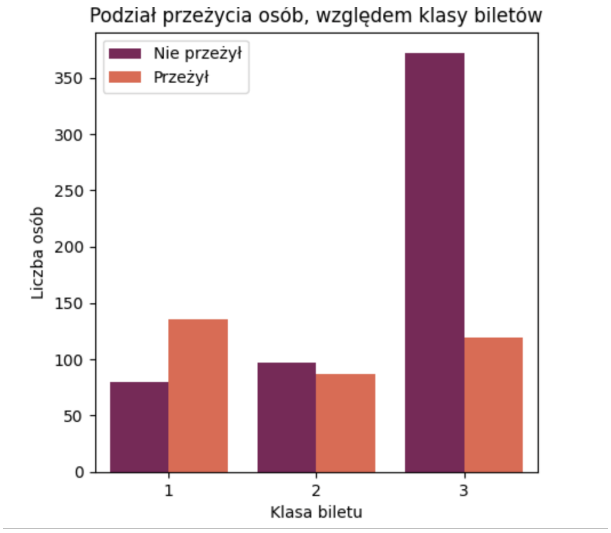
- *Analiza przeżycia względem płci:*

  Dane wykazały, że kobiety miały wyraźnie wyższy wskaźnik przeżycia niż mężczyźni. Stanowi to istotną wskazówkę, że płeć powinna być silnie uwzględniona jako cecha predykcyjna.



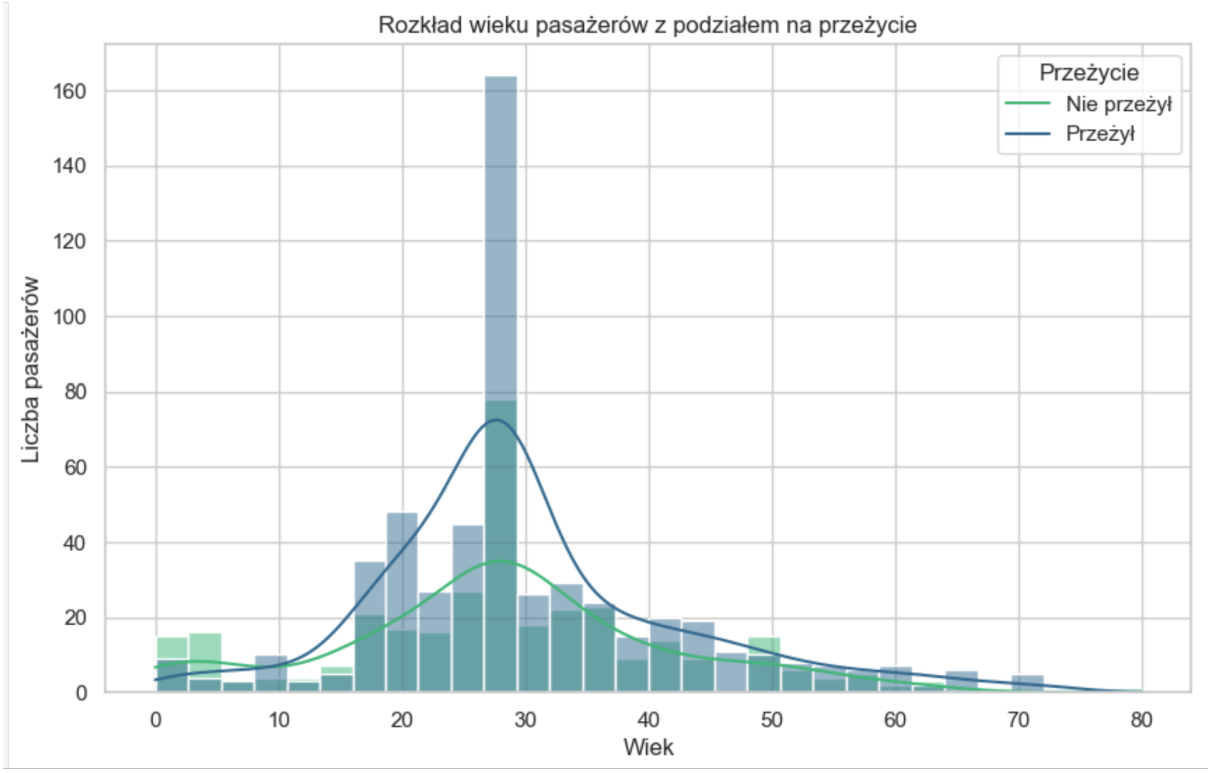
- *Analiza klasy biletu:*

  Pasażerowie podróżujący w pierwszej klasie mieli większe szanse na przeżycie niż ci z niższych klas, co może być związane z dostępnością łodzi ratunkowych i priorytetem ewakuacji.



- *Wiek:*

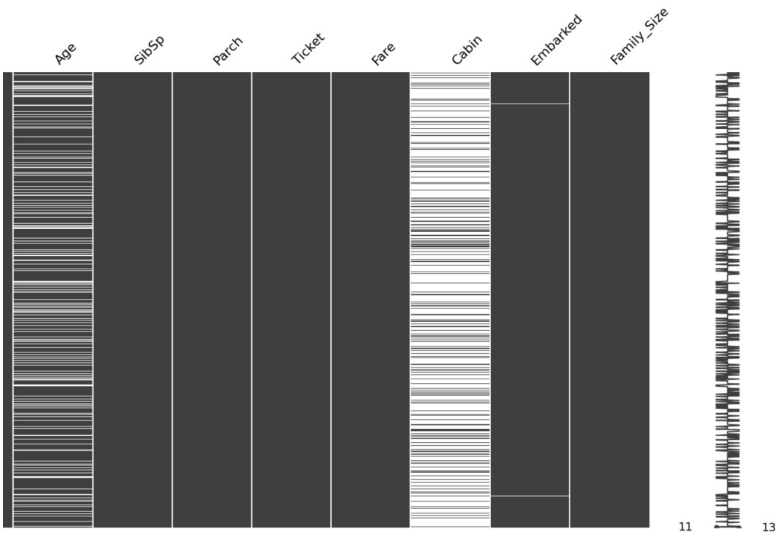
Rozkład wieku wskazuje, że dzieci miały niższy wskaźnik przeżycia w porównaniu do osób między 18 a 30 rokiem życia , co może sugerować brak procedur ratunkowych (np. "kobiety i dzieci pierwsze").
Rozkład ukazuje ,że ludzie „młodzi-dorośli" wykorzystując swoją sprawność zdołali w większości uratować swoje życie, natomiast dzieci i osoby coraz starsze miały ogromne problemy z ratowaniem swojego życia ze względu na swoją sprawność/ rozwój fizyczny.

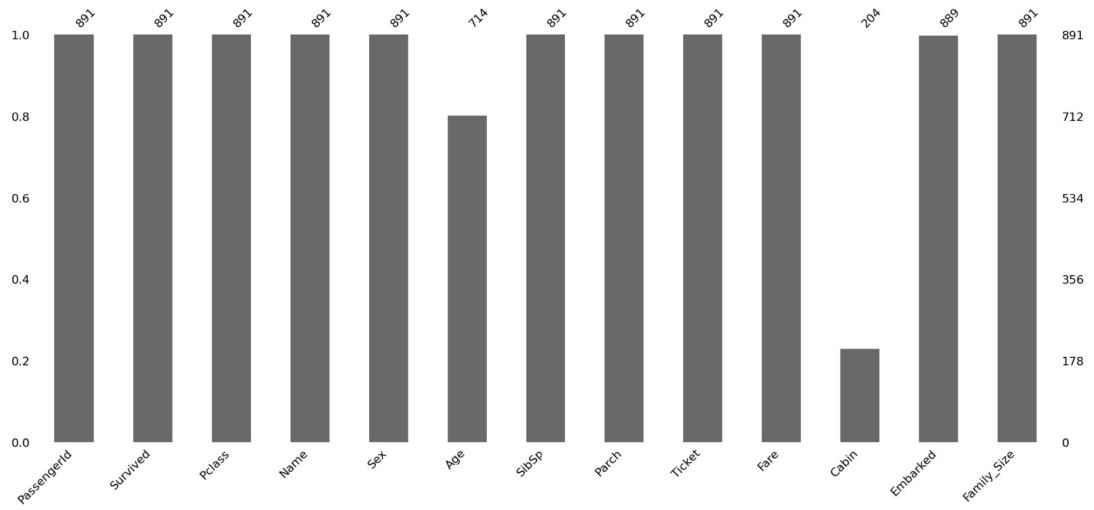


2. *Przygotowanie danych:*

- *Wprowadzono zmienne wskaźnikowe (dummy) dla cech kategorycznych, takich jak płeć i miejsce zaokrętowania (wejścia na pokład), oraz usunięto kolumny Name, Ticket i Cabin.*
Braki w wieku uzupełniono medianą, a niekompletne dane w Embarked usunięto (NaN).

```
Brakujące dane:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
Family_Size      0
dtype: int64
```





	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare	Family_Size	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
0	1	0	3	22	1	0	7.2500	1	False	True	False	False	True
1	2	1	1	38	1	0	71.2833	1	True	False	True	False	False
2	3	1	3	26	0	0	7.9250	0	True	False	False	False	True
3	4	1	1	35	1	0	53.1000	1	True	False	False	False	True
4	5	0	3	35	0	0	8.0500	0	False	True	False	False	True
...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	27	0	0	13.0000	0	False	True	False	False	True
887	888	1	1	19	0	0	30.0000	0	True	False	False	False	True
888	889	0	3	28	1	2	23.4500	3	True	False	False	False	True
889	890	1	1	26	0	0	30.0000	0	False	True	True	False	False
890	891	0	3	32	0	0	7.7500	0	False	True	False	True	False

3. *Trening modelu regresji logistycznej:*

- *Model trenowano na 80% zbiorze treningowym z ustawieniem random_state=101 dla replikowalności. Model miał za zadanie przewidzieć klasę Survived (przeżycie).*

4. *Ewaluacja modelu:*

- *Dokładność: Model uzyskał dokładność 81% na zbiorze testowym, co wskazuje na jego skuteczność w klasyfikacji.*

- *Macierz błędów: Wyniki wskazują na 98 prawidłowych klasyfikacji dla klasy 0 i 47 dla klasy 1. Zanotowano jednak 24 fałszywe negatywy, co sugeruje, że model nie rozpoznaje wszystkich przypadków pozytywnych (przeżycie).*

- *Precyzja i czułość: Precyzja dla klasy 1 wyniosła 84%, a czułość 66%. Niska czułość oznacza, że model stosunkowo rzadko identyfikuje pozytywne przypadki (przeżycie) i ma tendencję do fałszywego odrzucania.*

	precision	recall	f1-score	support
0	0.80	0.92	0.86	107
1	0.84	0.66	0.74	71
accuracy			0.81	178
macro avg	0.82	0.79	0.80	178
weighted avg	0.82	0.81	0.81	178

```
[[98  9]
 [24 47]]
```

- *Krzywa ROC i AUC:*

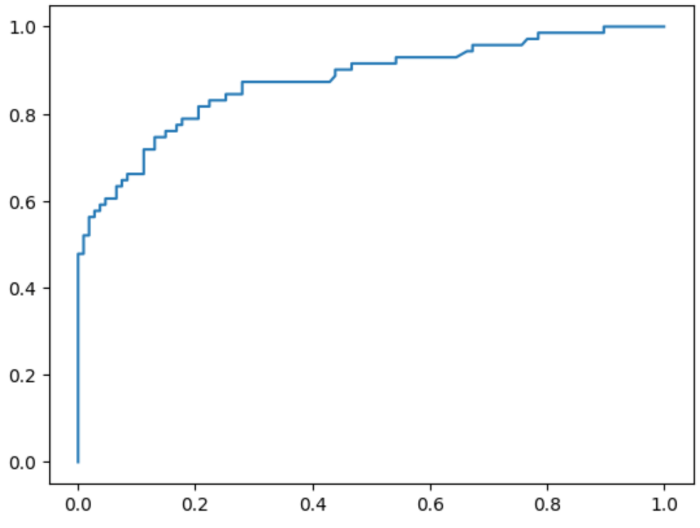
  FPR (False Positive Rate) — odsetek błędnych klasyfikacji przypadków negatywnych jako pozytywne (czyli osób, które nie przeżyły, a zostały zaklasyfikowane jako osoby, które przeżyły),

  TPR (True Positive Rate) — czułość, czyli odsetek poprawnych klasyfikacji przypadków pozytywnych (osób, które przeżyły, zaklasyfikowanych jako przeżywający).

  Idealna krzywa ROC powinna szybko wzrastać do wartości TPR bliskiej 1
  przy bardzo niskim FPR, co wskazuje na wysoki poziom trafnych klasyfikacji
  przy niskim odsetku fałszywych alarmów.
  Im bliżej górnego lewego rogu znajduje się krzywa, tym lepsza jest zdolność modelu do rozróżniania klas.

  Wartość AUC(pole powierzchni pod wykresem) wynosząca 0,5 oznacza losowy klasyfikator, podczas gdy AUC zbliżona do 1 oznacza model o wysokiej skuteczności. Wyższa wartość AUC sugeruje, że model skutecznie rozróżnia między pasażerami, którzy przeżyli, a tymi, którzy nie przeżyli.

  Analiza krzywej ROC wykazała umiarkowanie wysoki wynik AUC,
  co wskazuje na przyzwoitą jakość modelu.
  Model dobrze odróżnia klasy, ale istnieje pole do optymalizacji parametrów.

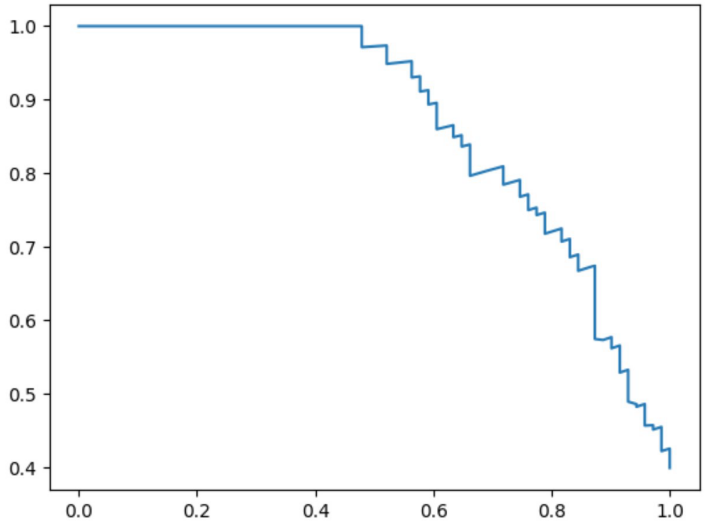


- *Wykres czułość/precyzja:*

Czułość reprezentuje zdolność modelu do poprawnego identyfikowania przypadków pozytywnych, czyli osób, które przeżyły katastrofę.

Precyzja mierzy dokładność przewidywań pozytywnych, tj. procent przewidywań przeżycia, które rzeczywiście odpowiadały osobom, które przeżyły.

W miarę zwiększania progu decyzyjnego precyzja rośnie, natomiast czułość maleje. Przy niskich progach model ma wyższą czułość kosztem precyzji, co oznacza, że przewiduje przeżycie dla większej liczby osób, ale większy procent tych przewidywań jest błędny.



Zadanie 2: Klasyfikacja wieloklasowa na zbiorze danych Iris

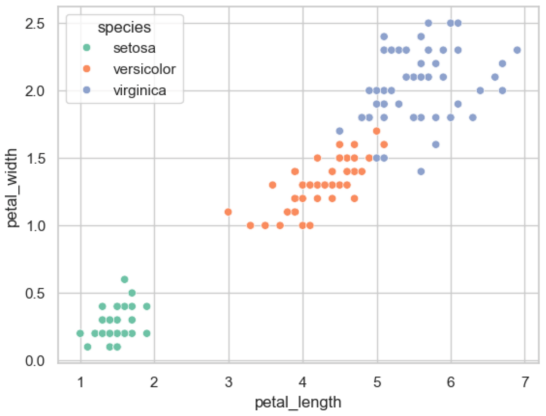
Zadanie 2a: Klasyfikacja gatunku Iris virginica na podstawie szerokości płatka

1. *Opis danych:*

Dane zawierają informacje o cechach kwiatów trzech gatunków irysów: setosa, versicolor, i virginica.
Kluczową cechą w tej analizie była szerokość płatka, która wyraźnie różnicuje Iris virginica od pozostałych gatunków.

2. *Wizualizacja i analiza cechy petal_width:*

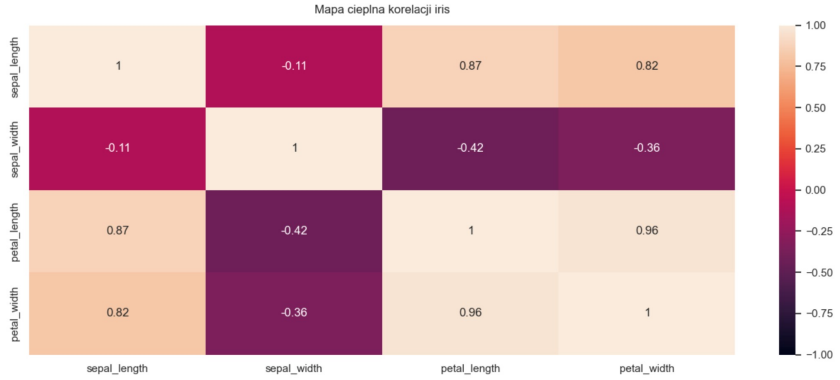
Rozkład szerokości płatków wskazał, że gatunek virginica posiada wyraźnie większe wartości tej cechy w porównaniu do innych gatunków.
Użycie tej cechy do klasyfikacji Iris virginica było więc uzasadnione.



	sepal_length	sepal_width	petal_length	petal_width
ount	150.000000	150.000000	150.000000	150.000000
nean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

3. *Mapa Korelacji:*

Analiza korelacji ujawniła, że szerokość płatka ma wysoką korelację z długością płatka (0.96), co sugeruje istotną rolę tej cechy w klasyfikacji Iris virginica.



4. *Przygotowanie danych i modelowanie:*

Dane zostały podzielone na zbiór treningowy i testowy, a model trenowano przy ustawieniu random_state=101 dla powtarzalności wyników. Wartości petal_width zostały przekształcone do formatu odpowiedniego do analizy jako X1 (szerokość płatka) i y1 (etykieta binarna wskazująca obecność gatunku virginica).

5. *Wyniki predykcji:*

Model skutecznie identyfikował gatunek virginica na podstawie szerokości płatka, osiągając wysoką precyzję i czułość. Zastosowanie szerokości płatka jako jedynej cechy było trafne, biorąc pod uwagę wyraźną różnicę tej cechy dla virginica.

6. Granica Decyzyjna dla Klasy Iris virginica:

Metoda Znalezienia Granicy Decyzyjnej

a) *Przygotowanie danych do prognozy:*

Utworzono zbiór 1000 hipotetycznych punktów, równomiernie rozłożonych w zakresie wartości szerokości płatka od 0 do 3.
Wartości te posłużyły do przetestowania modelu i uzyskania prawdopodobieństw dla klasy Iris virginica w funkcji szerokości płatka.

b) *Obliczenia prawdopodobieństw:*

Model regresji logistycznej wygenerował dla każdego punktu prawdopodobieństwo przynależności do klasy Iris virginica.
Wartość ta była rosnąca wraz ze wzrostem szerokości płatka.

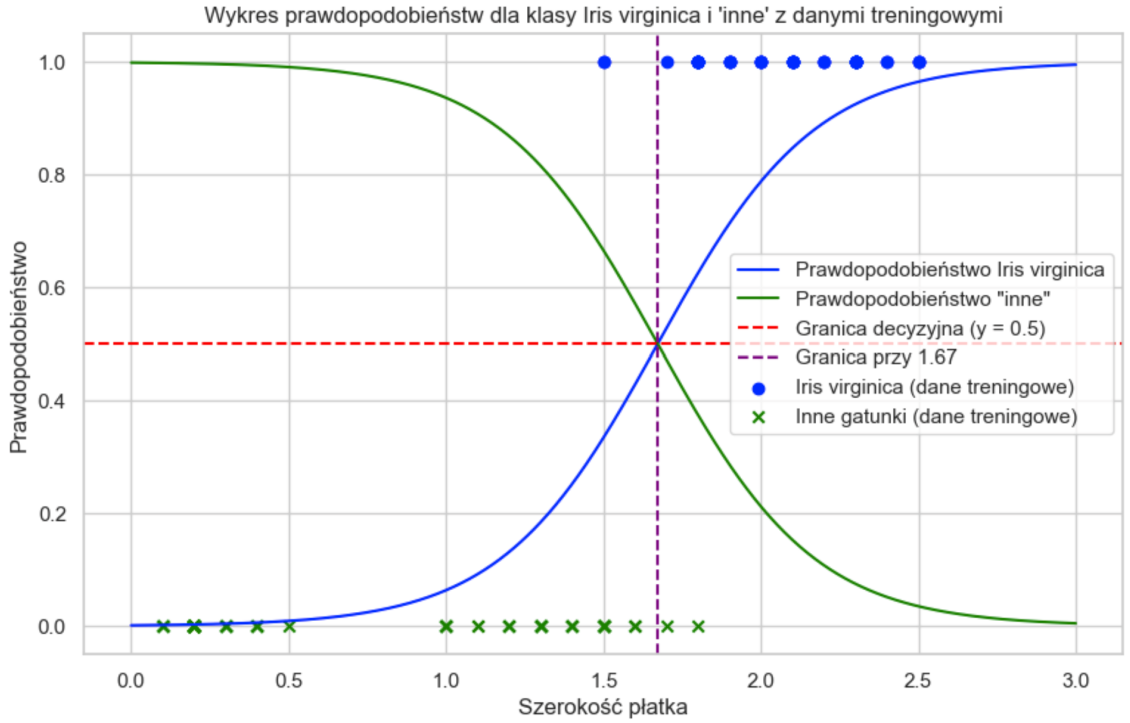
c) *Wyznaczenie granicy decyzyjnej:*

Granicę decyzyjną określono jako punkt szerokości płatka, dla którego model przewiduje prawdopodobieństwo $p \approx 0.5$.
Jest to miejsce, gdzie klasyfikacja przechodzi od klasy "inne" do klasy Iris virginica.

d) *Interpretacja Wyniku*

Granica decyzyjna na poziomie około 1.67 cm szerokości płatka oznacza, że irysy o szerokości płatka mniejszej niż 1.67 cm są klasyfikowane jako "inne" (czyli setosa lub versicolor), podczas gdy te o szerokości większej lub równej 1.67 cm są klasyfikowane jako Iris virginica.
Wynik ten zgadza się z biologiczną obserwacją, że Iris virginica ma większe płatki niż pozostałe dwa gatunki, co pozwala na skuteczne rozróżnienie tej klasy.

```
Granica decyzyjna szerokości płatka: 1.6726726726726726
```



Zadanie 2b: Klasyfikacja wieloklasowa dla wszystkich gatunków irysów

1. *Wieloklasowa klasyfikacja z cechami sepal_length, sepal_width, petal_length, petal_width:*

    a) *Zbiór danych Iris, zawierający wszystkie cechy, został wykorzystany do klasyfikacji trzech gatunków irysów w wieloklasowej analizie.*

    b) *Standaryzacja: Dane zostały poddane standaryzacji za pomocą StandardScaler, co zapewnia równoważny wpływ każdej cechy na model, co jest szczególnie ważne w przypadku regresji logistycznej.*

c) Regresja Logistyczna:

- *Wybrano regresję logistyczną jako metodę klasyfikacji, stosując solver liblinear, który obsługuje zarówno kary L1, jak i L2.*

- *Dostrajanie Modelu: Parametry C (kontrola regularizacji) i penalty (rodzaj kary) zostały dostrojone przy użyciu RandomizedSearchCV, co umożliwiło losowe wyszukiwanie optymalnych wartości parametrów, przyspieszając proces w porównaniu z pełnym przeszukiwaniem siatki.*

- *Optymalizacja Modelu: Wybór najlepszych parametrów pozwolił na osiągnięcie możliwie najlepszej wydajności klasyfikacyjnej.*

2. Raport Klasyfikacji:

	precision	recall	f1-score	support
setosa	1.00	0.90	0.95	10
versicolor	0.85	1.00	0.92	17
virginica	1.00	0.82	0.90	11
accuracy			0.92	38
macro avg	0.95	0.91	0.92	38
weighted avg	0.93	0.92	0.92	38

a) Dokładność ogólna: Model osiągnął dokładność na poziomie 92%.

b) Wyniki dla klas:

Setosa: Precyzja 1.00, czułość 0.90, wskaźnik F1 0.95 – Model skutecznie rozpoznawał kwiaty setosa, z nielicznymi błędami.

Versicolor: Precyzja 0.85, czułość 1.00, wskaźnik F1 0.92 – Versicolor był poprawnie klasyfikowany w 100% przypadków testowych.

Virginica: Precyzja 1.00, czułość 0.82, wskaźnik F1 0.90 – Gatunek virginica był dobrze klasyfikowany, choć niższa czułość wskazuje na większe prawdopodobieństwo błędnej klasyfikacji tej klasy w porównaniu z innymi. 3. Macierz Błędów:

```
[[ 9  1  0]
 [ 0 17  0]
 [ 0  2  9]]
```

Setosa: Model poprawnie sklasyfikował 9 na 10 przypadków, a jeden przypadek setosa został błędnie zaklasyfikowany do innej klasy.

Versicolor: Model poprawnie rozpoznał wszystkie 17 przypadków versicolor, co oznacza brak błędów dla tej klasy.

Virginica: Model błędnie sklasyfikował 2 przypadki virginica jako versicolor. To może wynikać z podobieństw między tymi dwoma klasami, co sugeruje trudność w ich rozróżnieniu.

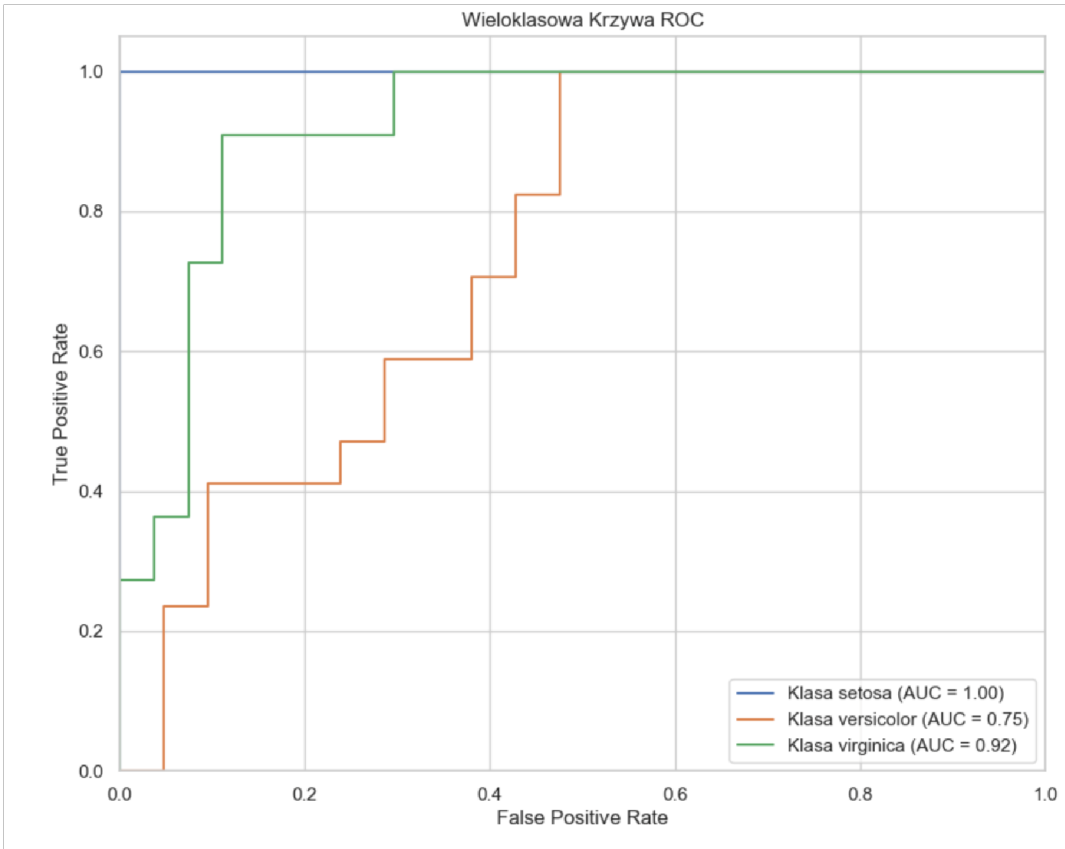
4. Wnioski

Skuteczność Klasyfikacji: Model jest skuteczny w rozpoznawaniu klasy setosa, jednak wykazuje pewne trudności przy rozróżnianiu versicolor i virginica, co jest widoczne w niższej czułości dla virginica.

Możliwość Poprawy: Dalsza optymalizacja modelu lub użycie dodatkowych cech mogłoby poprawić czułość dla virginica, minimalizując błędne klasyfikacje między virginica i versicolor.

Dobre Dopasowanie Modelu: Średnie wskaźniki (macro avg oraz weighted avg) potwierdzają ogólną wysoką jakość klasyfikacji, a dokładność na poziomie 92% wskazuje na dobrą sprawność modelu w klasyfikacji wieloklasowej.

5. Krzywa ROC:



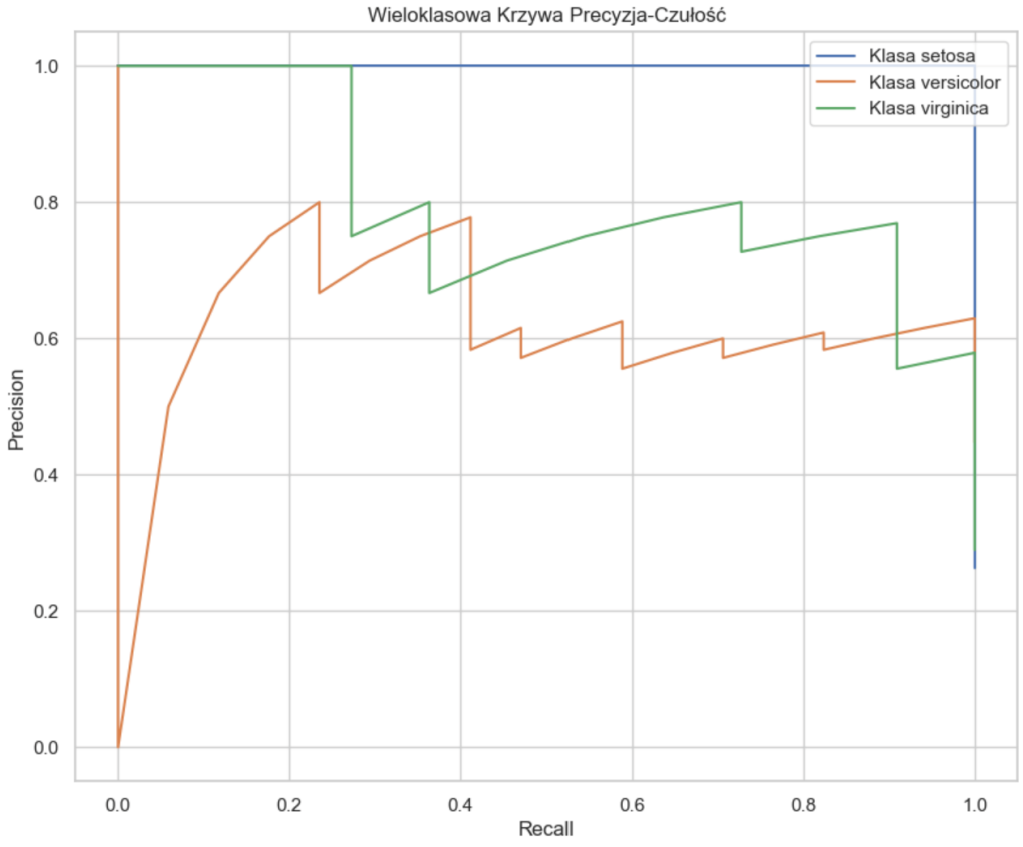
- *Wartości AUC bliskie 1 dla klas setosa, versicolor, i virginica sugerują, że model jest skuteczny w odróżnianiu każdej klasy od pozostałych.*

- *Najbardziej odrębna klasa (prawdopodobnie setosa, ze względu na jej odmienność morfologiczną) będzie miała najwyższą wartość AUC, co może wynikać z wyraźnych różnic morfologicznych w stosunku do pozostałych gatunków.*
- *Mniejsze różnice pomiędzy krzywymi versicolor i virginica mogą oznaczać większe podobieństwo tych dwóch klas i trudniejszą klasyfikację między nimi.*

6. Krzywa Czułość/Precyzja:

- *Krzywe dla klas versicolor i virginica mogą wykazywać większą zmienność, co sugeruje trudniejszą klasyfikację pomiędzy tymi gatunkami. Spadki precyzji przy wysokiej czułości mogą oznaczać większe ryzyko fałszywych alarmów przy próbie rozpoznania tych klas.*

- *Dla klasy setosa, krzywa wskazuje na wysoką precyzję i czułość, co jest zgodne z wyraźną rozróżnialnością tego gatunku w zbiorze danych.*



4. Podsumowanie

Zadanie 1: Klasyfikacja binarna na zbiorze Titanic

Analiza i przygotowanie danych: Przeprowadzono analizę przeżywalności względem cech, takich jak płeć, klasa i wiek, co ujawniło korelację między tymi atrybutami a przeżywalnością.

Modelowanie: Model regresji logistycznej osiągnął dokładność 81% na zbiorze testowym. Precyzja dla klasy 1 (przeżycie) wyniosła 84%, a czułość 66%.

Krzywa ROC: Wynik AUC sugerował umiarkowaną zdolność modelu do rozróżniania między osobami, które przeżyły, a tymi, które zginęły.

Krzywa czułość/precyzja: Wyższa precyzja przy wyższym progu decyzyjnym pokazuje kompromis między dokładnym rozpoznaniem przeżycia a minimalizacją fałszywych alarmów.

W przypadku zbioru Titanic, model regresji logistycznej osiągnął przyzwoite wyniki, ale wykazywał niższą czułość dla klasy przeżycia.
Optymalizacja parametrów mogłaby poprawić zdolność rozpoznawania przypadków przeżycia.

Zadanie 2: Klasyfikacja wieloklasowa na zbiorze Iris

Klasyfikacja Iris virginica: Użycie szerokości płatka jako cechy dało wysoką skuteczność, co było uzasadnione dużą różnicą tej cechy w porównaniu do innych gatunków.

Klasyfikacja wszystkich gatunków irysów: Model osiągnął dokładność 92%, a analiza raportu klasyfikacji wykazała, że model dobrze klasyfikował każdy z gatunków z różnymi poziomami precyzji i czułości.
Wyniki dla setosy i versicolor wskazują na ich dobrą rozróżnialność przez model.

W klasyfikacji wieloklasowej Iris, regresja logistyczna wykazała wysoką skuteczność dzięki dobrze oddzielającym cechom. Szerokość płatka była szczególnie skuteczna w identyfikacji gatunku virginica, a model osiągnął wysoką dokładność przy klasyfikacji wszystkich gatunków.

Bibliografia

Bilińska, A., Kosharnyi, B., & Smarzewska, F. (2023). „Sieci złożone w analizie bezpieczeństwa na przykładzie siatki terrorystycznej.” W: Kuczmaszewska, A. (red.), Koła naukowe etapem rozwoju kompetencji zawodowych i naukowych, Lublin: Politechnika Lubelska, s. 124-141, https://bc.pollub.pl/Content/13856/Ko%C5%82a_naukowe.pdf#page=124, dostęp: 04.11.2024.

A Novel Approach for Developing a Linear Regression Model within Logistic Cluster Using Scikit-Learn. Scientific Research Publishing, https://www.scirp.org/journal/paperinformation?paperid=134301, dostęp: 04.11.2024.

Murphy, K.P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press, dostęp: 04.11.2024, https://raw.githubusercontent.com/kerasking/book-1/master/ML%20Machine%20Learning-A%20Probabilistic%20Perspective.pdf