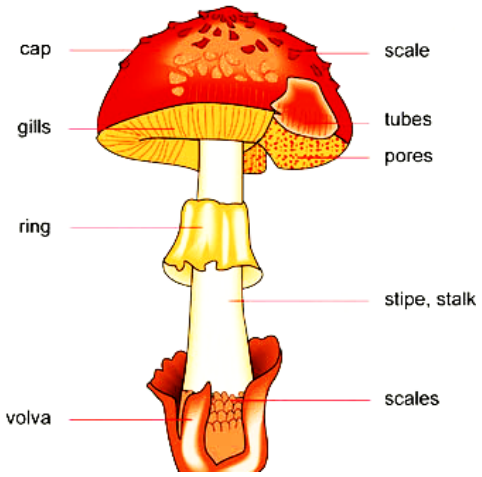
Naiwny Klasyfikator Bayesowski

Projekt autorstwa: Bartosz Kaganiec, Bartosz Ludwin

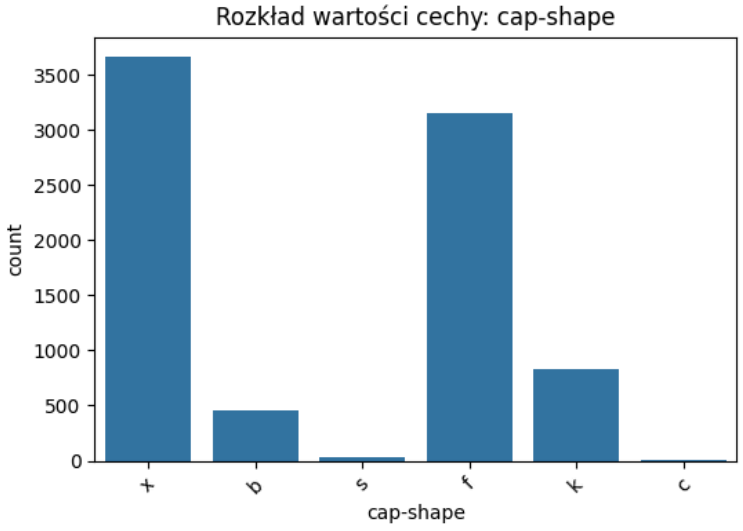
Wstępna analiza danych

Mushroom Classification

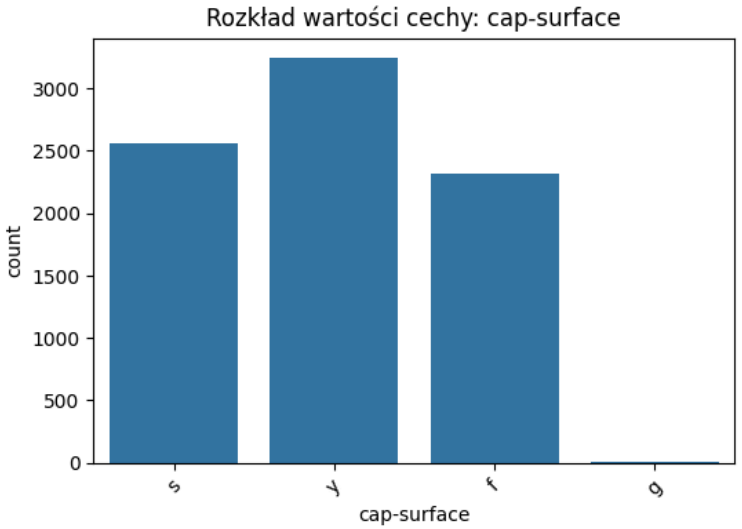
1. **Wizualizacja**:



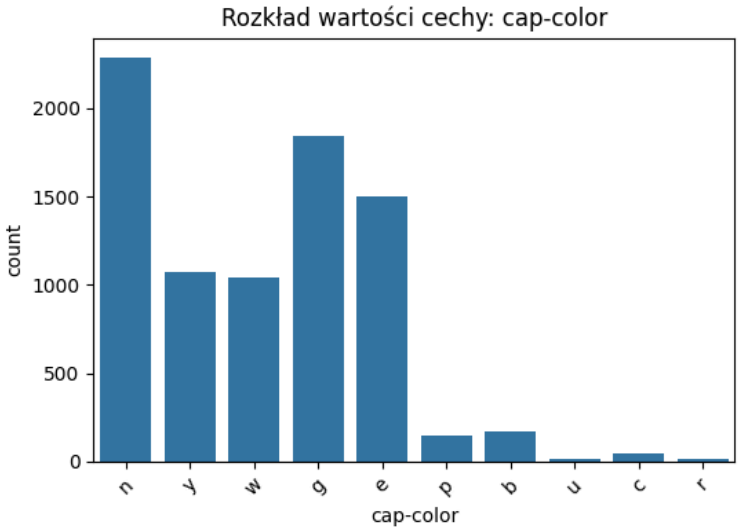
2. **Brakujące wartości:** W zbiorze występują brakujące wartości.
3. **Analiza cech:**
    - **class (klasa):**
        - Unikalne wartości: e (jadalny), p (trujący)
        - Najczęstsza wartość: e (**4208 wystąpień**)
        - Stosunek: około 13:12
    - **cap-shape (kształt kapelusza):**
        - Unikalne wartości:
            - b (bell),
            - c (conical),
            - x (convex),
            - f (flat),
            - k (knobbed),
            - s (sunken)
        - Najczęstsza wartość: x (convex) – **3656 wystąpień**



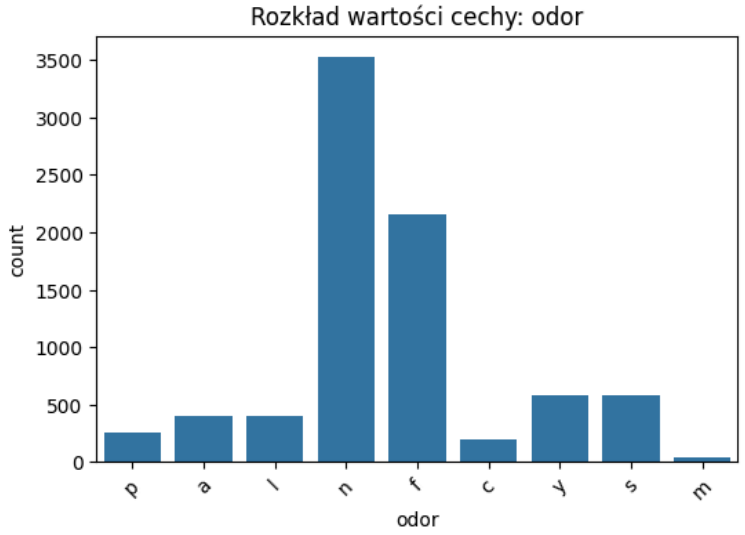
- **cap-surface (powierzchnia kapelusza):**
  - Unikalne wartości:
    - f (fibrous),
    - g (grooves),
    - y (scaly),
    - s (smooth)
  - Najczęstsza wartość: y (scaly) – **3244 wystąpienia**



- **cap-color (kolor kapelusza):**
  - Unikalne wartości: n (brown),
    - b (buff),
    - c (cinnamon),
    - g (gray),
    - r (green),
    - p (pink),
    - u (purple),
    - e (red),
    - w (white),
    - y (yellow)
  - Najczęstsza wartość: n (brown) – **2284 wystąpienia**

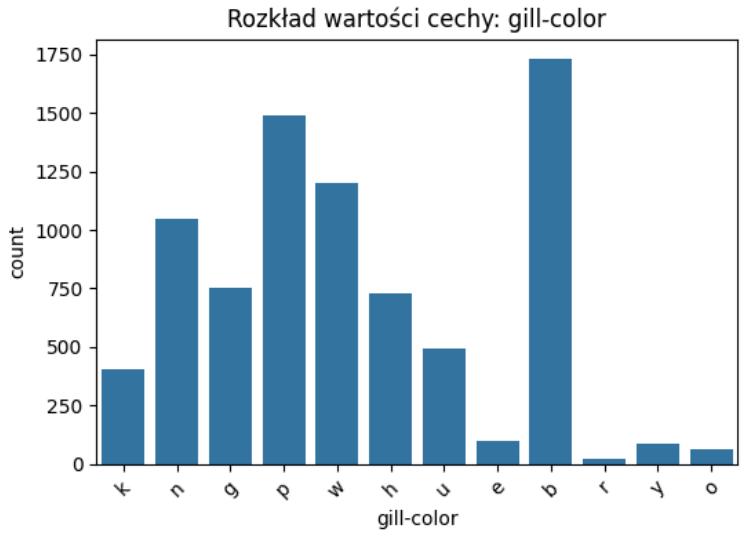


- **bruises (czy ma siniaki):**
  - Unikalne wartości: t (tak), f (nie)
  - Najczęstsza wartość: f (nie) – **4748 wystąpień**
  - Stosunek: około 4:6
- **odor (zapach):**
  - Unikalne wartości: a (almond), l (anise), c (creosote), y (fishy), f (foul), m (musty), n (none), p (pungent), s (spicy)
  - Najczęstsza wartość: n (none) – **3528 wystąpień**

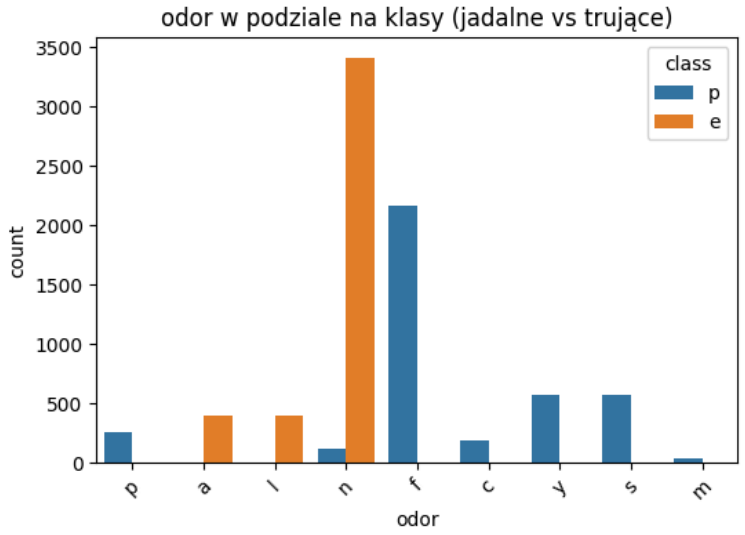


- **gill-attachment (przyczepność blaszek):**
  - Unikalne wartości: a (attached), f (free)
  - Najczęstsza wartość: f (free) – **7914 wystąpień**
  - Stosunek: około 1:19
- **gill-spacing (odstęp między blaszkami):**
  - Unikalne wartości: c (crowded), w (close)
  - Najczęstsza wartość: c (crowded) – **6812 wystąpień**
  - Stosunek: około 5:1
- **gill-size (rozmiar blaszek):**
  - Unikalne wartości: b (broad), n (narrow)
  - Najczęstsza wartość: b (broad) – **5612 wystąpień**
  - Stosunek: około 7:3

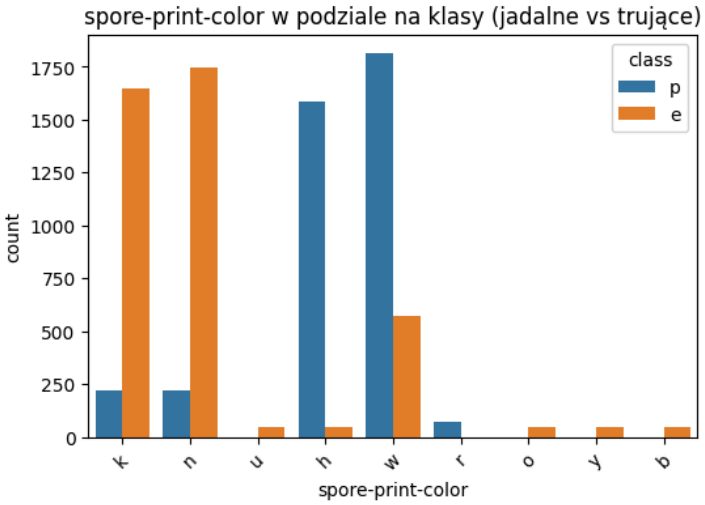
- **gill-color (kolor blaszek):**
  - Unikalne wartości: k, n, b, h, g, r, o, p, u, e, w, y
  - Najczęstsza wartość: b (buff) – **1728 wystąpień**



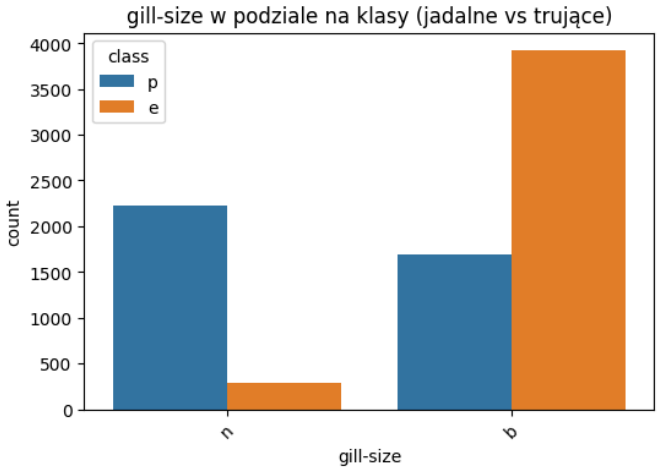
4. **Cechy wyróżniające grzyby jadalne i trujące:**
   - **odor (zapach):** Jest to najbardziej znacząca cecha w odróżnianiu grzybów jadalnych od trujących.
     - Rozkład zapachów:
       - **Grzyby trujące** (p): Najczęściej mają zapach f (foul - nieprzyjemny, **2160 wystąpień wśród trujących**) lub p (pungent - ostry, **2560 wystąpień wśród trujących**).
       - **Grzyby jadalne** (e): Najczęściej mają zapach n (none - brak zapachu, **3408 wystąpień wśród jadalnych**) lub a (almond - migdałowy, **400 wystąpień wśród jadalnych**).
     - **Wniosek:** Brak zapachu lub zapach migdałowy jest charakterystyczny dla grzybów jadalnych, podczas gdy ostry i nieprzyjemny zapach wskazuje na to, że grzyb jest trujący.



- **spore-print-color (kolor zarodników):** Ta cecha różnicuje klasy w znacznym stopniu.
  - Rozkład kolorów zarodników:
    - **Grzyby trujące** (p): Najczęstszy kolor zarodników to w (white - biały, **1200 wystąpień wśród trujących**) oraz r (green - zielony, **744 wystąpienia wśród trujących**).
    - **Grzyby jadalne** (e): Najczęstszy kolor zarodników to k (black - czarny, **1872 wystąpienia wśród jadalnych**) oraz n (brown - brązowy, **1312 wystąpień wśród jadalnych**).
  - **Wniosek:** Zielony i biały kolor zarodników częściej występują dla grzybów trujących, podczas gdy czarny i brązowy są bardziej typowe dla grzybów jadalnych.



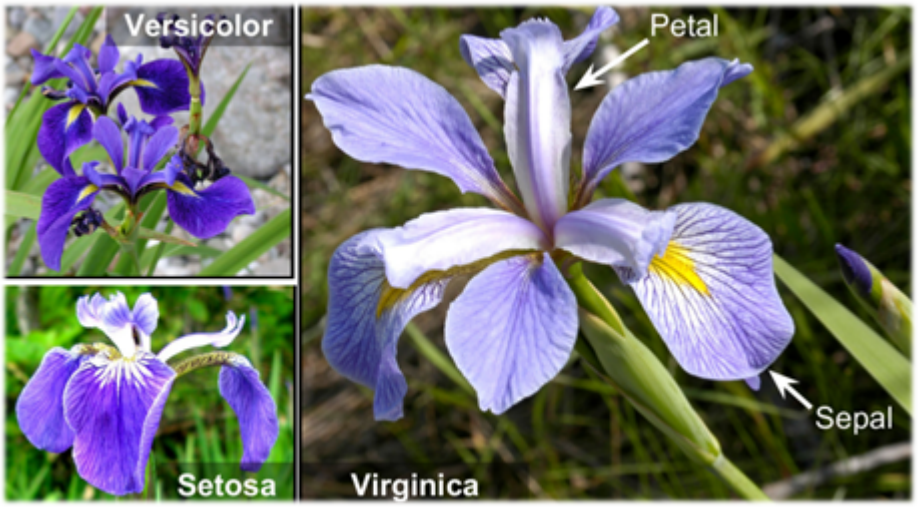
- **gill-size (rozmiar blaszek):** Rozmiar blaszek jest prostą i skuteczną cechą w klasyfikacji.
  - Rozkład:
    - **Grzyby jadalne** (e): W większości przypadków mają szerokie blaszki b (broad), **3360 wystąpień**.
    - **Grzyby trujące** (p): Najczęściej mają wąskie blaszki n (narrow), **3516 wystąpień**.



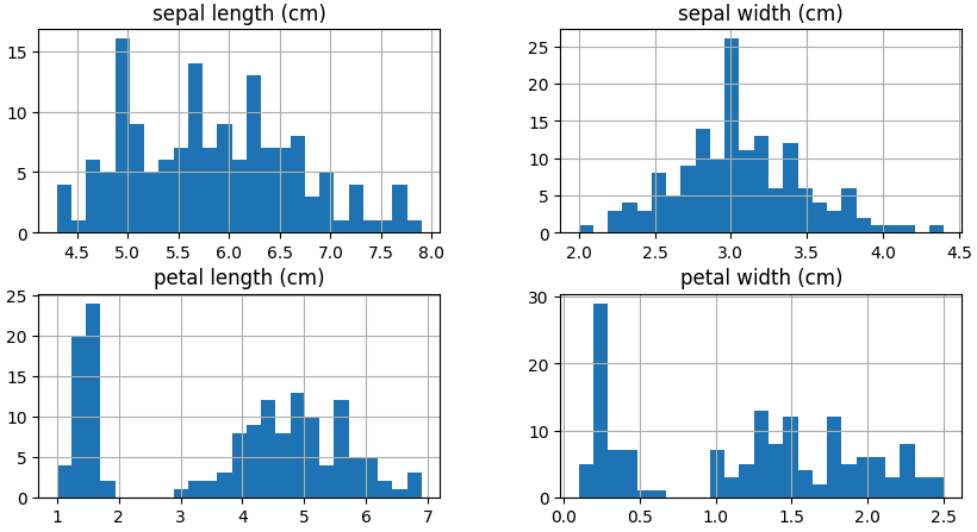
  - **Wniosek:** Szerokie blaszki są charakterystyczne dla grzybów jadalnych, a grzyb z wąskimi z większym prawdopodobieństwem jest trujący.

Iris

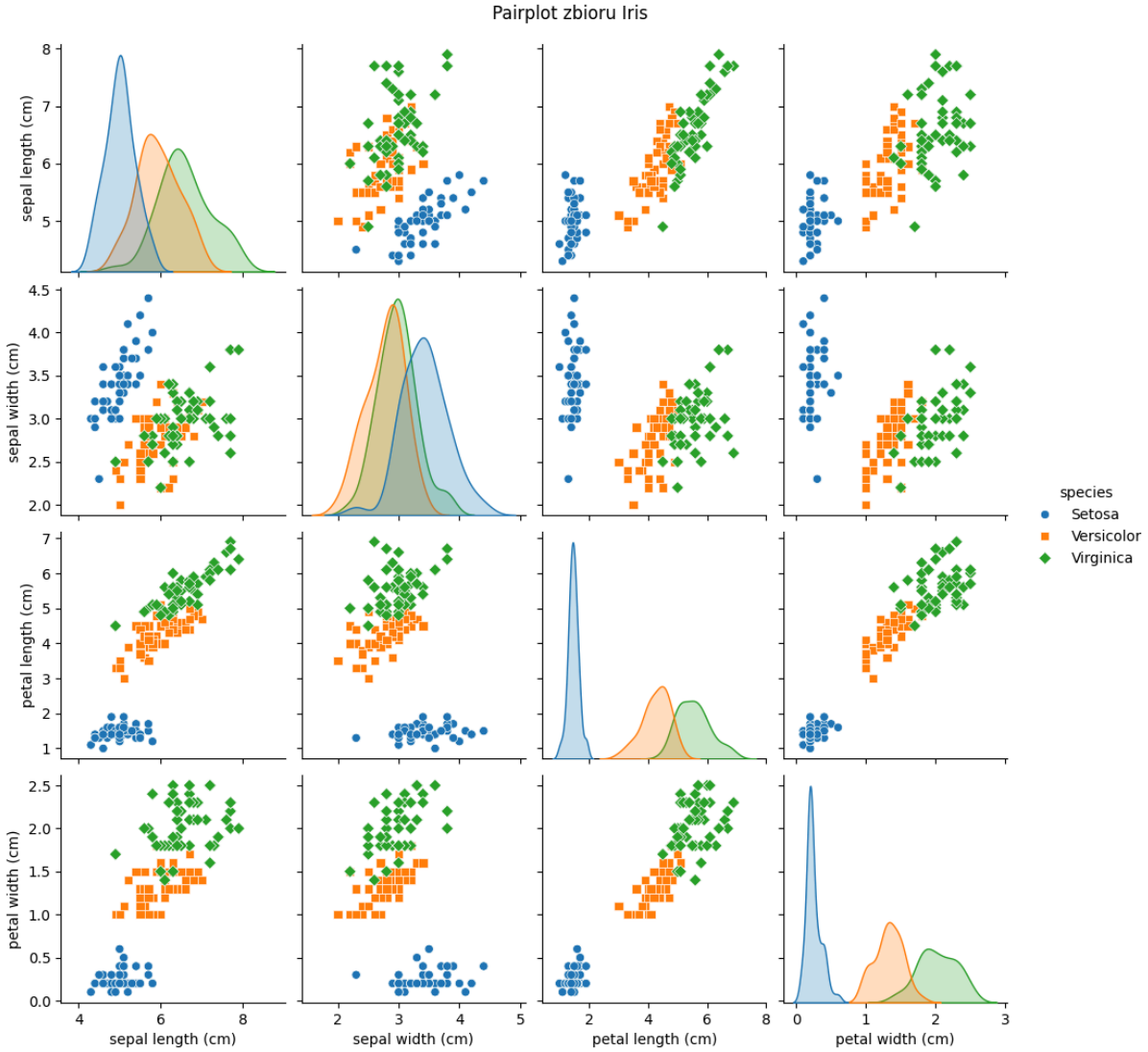
1. **Wizualizacja**:



2. **Brakujące wartości:** W zbiorze nie występują brakujące wartości (zbiór kompletny).
3. **Proporcja poszczególnych gatunków irysa:** Proporcja wynosi, **1:1:1**.
4. **Analiza cech:**
    - **sepal length (długość działki kielicha) [cm]:**
        - Średnia wartość: 5.84
        - Mediana: 5.8
        - Zakres: 4.3 - 7.9
    - **sepal width (szerokość działki kielicha) [cm]:**
        - Średnia wartość: 3.06
        - Mediana: 3.0
        - Zakres: 2.0 - 4.4
    - **petal length (długość płatka) [cm]:**
        - Średnia wartość: 3.758
        - Mediana: 4.35
        - Zakres: 1.0 - 6.9
    - **petal width (szerokość płatka) [cm]:**
        - Średnia wartość: 1.2
        - Mediana: 1.3
        - Zakres: 0.1 - 2.5
    - **species (gatunek kwiatu):**
        - Unikalne wartości: ***setosa*, *versicolor*, *virginica***



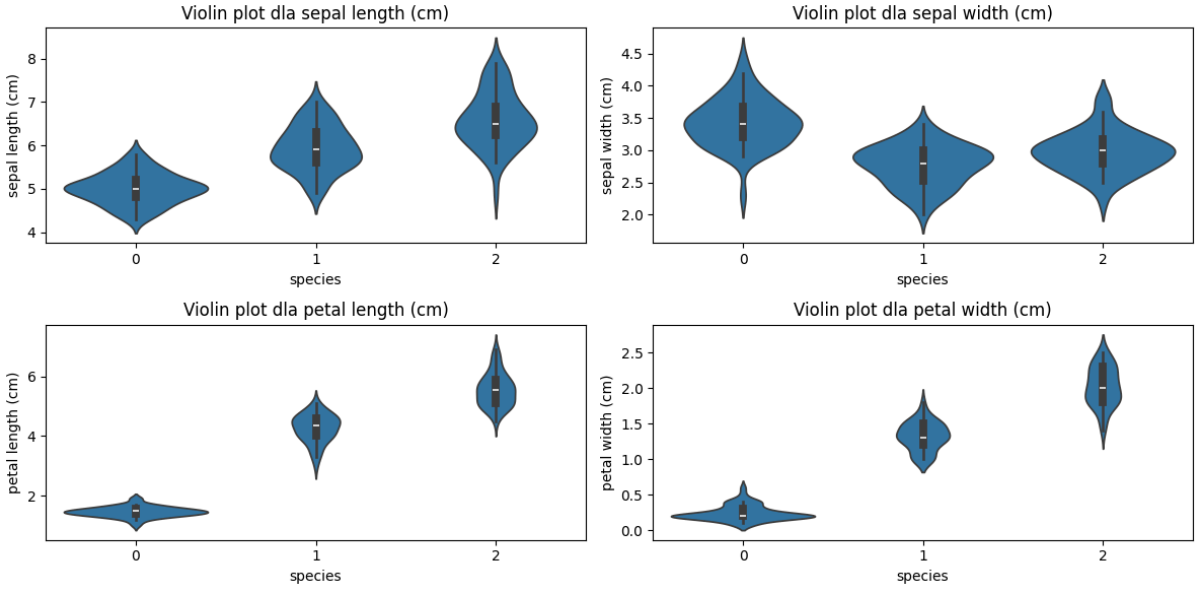
5. **Cechy wyróżniające poszczególne gatunki:**



- **petal length (długość płatka):**
  - **Iris setosa** charakteryzuje się najkrótszymi płatkami (zazwyczaj < 2 cm, średnia: **1.46 cm**)
  - **Iris versicolor** płatki są zazwyczaj dłuższe niż u setosy, ale krótsze niż u virginiki (zwykle 3–5 cm, średnia: **4.26 cm**)
  - **Iris virginica** średnio posiada najdłuższe płatki (mogą przekraczać 6 cm, średnia: **5.55 cm**)

- **petal width (szerokość płatka):**
  - **Iris setosa** wąskie płatki, średnia szerokość: **0.25 cm**
  - **Iris versicolor** średnia szerokość płatków większa niż u setosy, ale mniejsza niż u virginiki, średnia szerokość: **1.33 cm**
  - **Iris virginica** najszersze płatki, średnia szerokość: **2.03 cm**

- **sepal length (długość działki kielicha)**
  - **Iris setosa** zazwyczaj ma najkrótszą działkę kielicha w porównaniu do pozostałych gatunków (zakres: ok. 4.3–5.8 cm, średnia: **5.01 cm**)
  - **Iris versicolor** długość działki kielicha dłuższa niż u setosy, ale często krótsza niż u virginiki (zakres: ok. 4.9–7.0 cm, średnia: **5.94 cm**)
  - **Iris virginica** zwykle największa długość działki kielicha, sięgająca nawet 7.9 cm (zakres: ok. 4.9–7.9 cm, średnia: **6.59 cm**)
- **sepal width (szerokość działki kielicha)**
  - **Iris setosa** Zazwyczaj bywa wyższa niż u versicolor czy virginiki (średnia: **3.42 cm**)
  - **Iris versicolor** zbliżona do virginici (średnia: **2.77 cm**)
  - **Iris virginica** zbliżona do versicolor (średnia: **2.97 cm**)

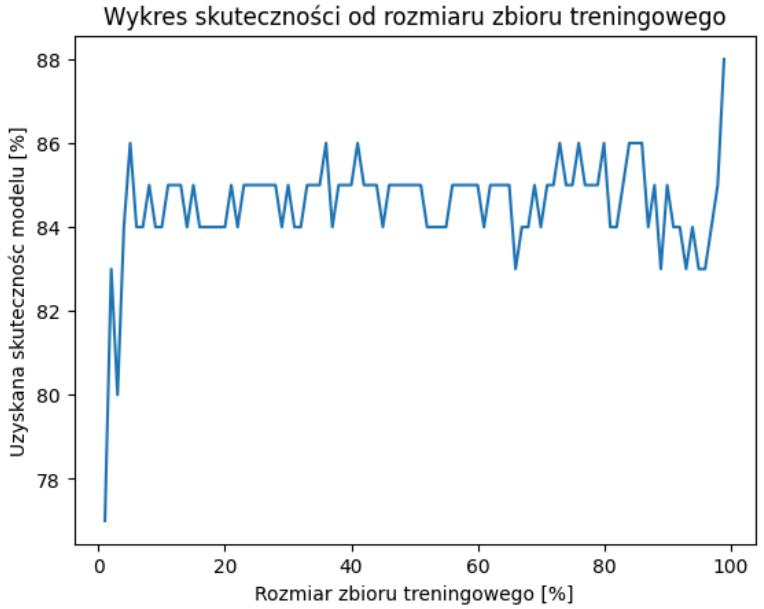


- **Wnioski:** Długość i szerokość płatka jest związana z gatunkiem irysa. Najmniejsze wymiary są w przypadku **setosy**, następnie **versicolor** a największe rozmiary ma **virginica**. Dodatkowo przypadku setosy szerokość działki kielicha często bywa wyższa niż u versicolor czy virginiki, co w połączeniu z krótszym sepal length daje dość charakterystyczne proporcje. Natomiast zakresy wartości długości działek wszystkich kielichów oraz ich szerokości w przypadku versicolor oraz virgnicici nachodzą na siebie, przez co te cechy nie nadają się do kategoryzacji.

Ocena jakości modelu

Mushrooms:

Model wykorzystujący rozkład wielomianowy w klasyfikatorze bayesowskim wykazał się dobrą skutecznością na poziomie około **85%**. Ciekawą obserwacją jest fakt że model wyznaczał się niezmienną skutecznością niezależnie od procentu rozmiaru zbioru treningowego (co można zauważyć na poniższym wykresie skuteczności od rozmiaru zbioru treningowego). Dzieje się tak prawdopodobnie przez wielkość zbioru Mushroom-Classification - jest on na tyle obszerny, że nawet jego kilka procent wystarcza na skuteczne działanie modelu.



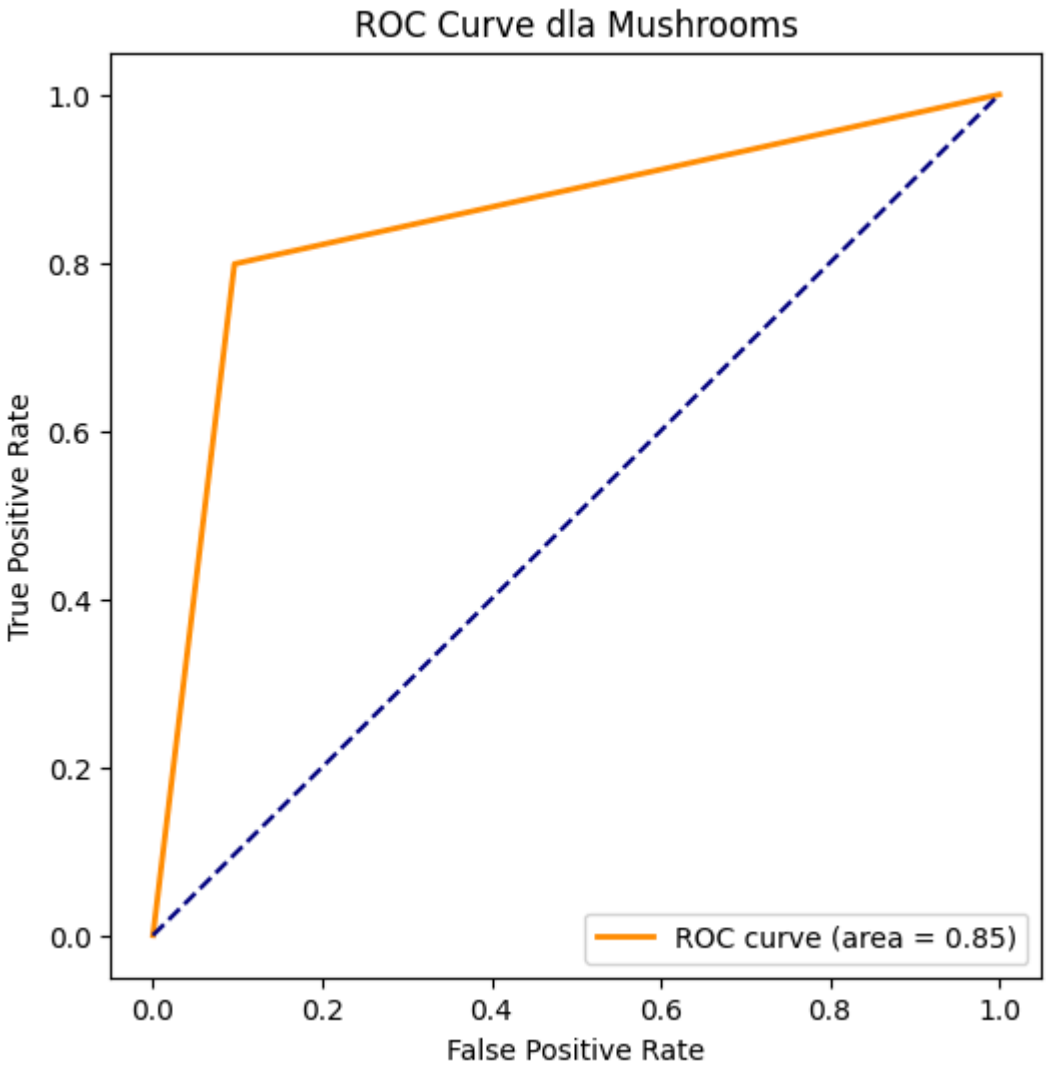
Klasyfikator charakteryzuje się również zadowalającą precyzją i czułością zarówno na cechę ‘e’ jak i na ‘p’. Jego F1-score wynosi w okolicach 0.85, co jest dobrym kompromisem między czułością a precyzją działania. Wynik jest podparty dużą próbą liczącą 2438 wartości.

```
Ocena klasyfikatora dla zbioru Mushrooms:
Accuracy: 0.8470
Confusion Matrix:
[[1130  124]
 [ 249  935]]
Classification Report:
              precision    recall  f1-score   support

           e       0.82      0.90      0.86      1254
           p       0.88      0.79      0.83      1184

    accuracy                           0.85      2438
   macro avg       0.85      0.85      0.85      2438
weighted avg       0.85      0.85      0.85      2438
```

Wyniki te potwierdza również wykres krzywej charakterystyki operacyjnej odbiornika (ROC Curve) jako że pole pod nią wynosi 0.85, oraz sama krzywa jest wyraźnie przesunięta ku lewego górnego rogu.



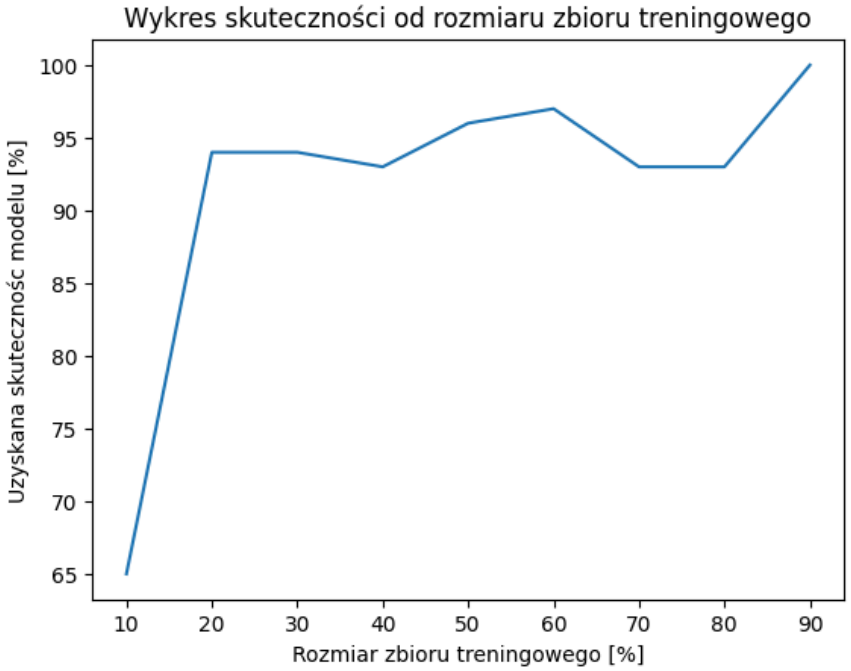
Lepsza skuteczność modelu nie została osiągnięta z kilku powodów. Jednym z nich jest obecność brakujących wartości w zbiorze Mushrooms-Classification, co mogło wpłynąć na jakość klasyfikacji, zwłaszcza jeśli dotyczyło to istotnych cech odróżniających grzyby jadalne od trujących.

Dodatkowym czynnikiem jest występowanie rzadkich wartości cech, które przy losowym podziale na zbiór treningowy i testowy mogą znaleźć się głównie w jednym z nich. W takim przypadku model nie jest w stanie odpowiednio nauczyć się klasyfikacji tych przypadków, co może skutkować większą liczbą błędów.

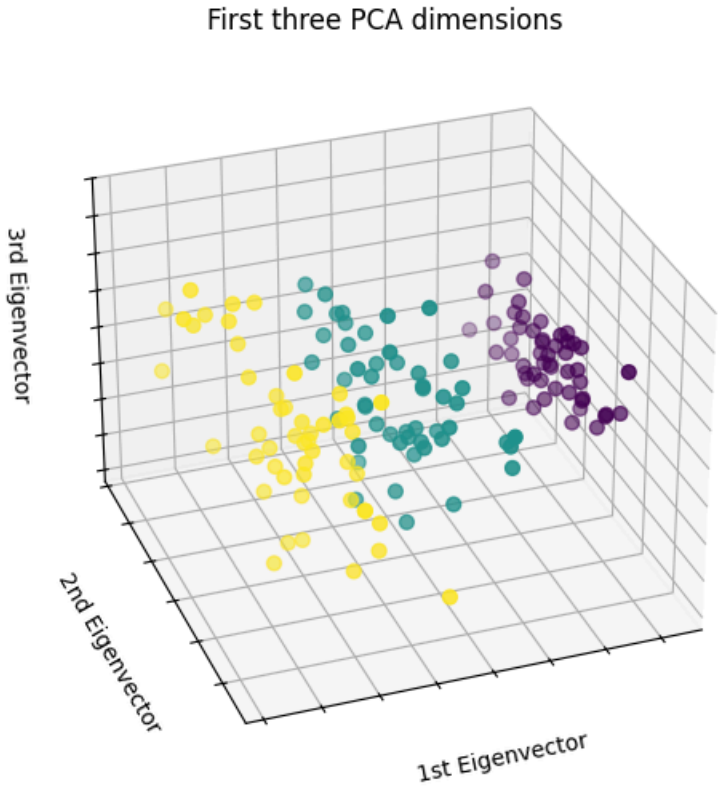
Kolejnym aspektem wpływającym na wyniki jest specyfika klasyfikatora Bayesowskiego, który zakłada niezależność cech. W rzeczywistości pewne atrybuty w zbiorze Mushrooms-Classification mogą być ze sobą skorelowane (np. kolor kapelusza i rodzaj blaszek), co może prowadzić do niedokładnych predykcji.

Iris:

Model oparty na gaussowskim klasyfikatorze bayesowskim znakomicie sprawdził się w przypadku zbioru Iris. Skuteczność modelu nie dość że szybko rośnie wraz ze zwiększaniem zbioru treningowego, to jeszcze jest wysoka nawet dla małej mocy zbioru treningowego (co widać na poniższym wykresie).



Dzieje się tak, ponieważ dane w zbiorze Iris pochodzą z rozkładu normalnego (który został użyty do tworzenia klasyfikatora), oraz są wyraźnie skategoryzowane w zależności od gatunku, co dobrze zostało zobrazowane na poniższej grafice prezentującej cechy wszystkich irysów (cztery cechy zostały zredukowane poprzez PCA aby stworzyć 3 wymiary)



Analizując model działający na proporcjach 70% danych treningowych oraz 30% testowych dało powtarzalny wynik dokładności na poziomie około **~95%**. Obecność prawie wszystkich wartości na przekątnej głównej w macierzy błędów również potwierdza skuteczność zaimplementowanego modelu w rozpoznawaniu gatunku Irysa.

```
Ocena klasyfikatora dla zbioru :
Accuracy: 0.9556
Confusion Matrix:
[[15  0  0]
 [ 0 13  2]
 [ 0  0 15]]
```

Źródła

- Grafiki do wizualizacji zostały zaczerpnięte ze strony: https://www.researchgate.net/
- Informacje o algorytmie:
  - https://scikit-learn.org/1.6/index.html
  - https://chatgpt.com/
  - https://www.kaggle.com/code/prashant111/naive-bayes-classifier-in-python
- Źródła użyte w celu pozyskania informacji o analizie danych:
  - https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=pl#:~:text=The%20ROC%20curve%20is%20a,holdover%20from%20WWII%20radar%20detection
  - https://matplotlib.org/stable/index.html
  - https://chatgpt.com
  - https://www.kaggle.com/datasets/uciml/mushroom-classification