Klasyfikacja jadalności grzybów

Jacek Podanowski
Informatyka stosowana
Politechnika Wrocławska
MSiD Lab 7:30 TN
272658@student.pwr.edu.pl

1 Wstęp

W ostatnim czasie wśród młodzieży można zaobserwować wzrost zainteresowania tematem grzybobrania. Niestety, niedoświadczeni amatorzy mogą pomylić się podczas klasyfikacji znalezionych grzybów, co może nieść ze sobą katastrofalne skutki. Dlatego zdecydowałem się przeprowadzić badania w zakresie klasyfikacji grzybów.

Współczesne technologie i metody analizy danych mogą znacząco zwiększyć dokładność i efektywność rozpoznawania gatunków trujących i jadalnych. W tym raporcie skupię się na próbie klasyfikacji jadalności grzybów na podstawie ich cech przy użyciu algorytmów uczenia maszynowego.

Głównym celem tej pracy jest zbadanie, w jakim stopniu mierzalne cechy grzybów mogą być wykorzystane do klasyfikacji ich jadalności. Analiza obejmie również ocenę różnych algorytmów uczenia maszynowego, takich jak SVM, sieci neuronowe oraz lasy losowe, które zostaną sprawdzone pod kątem ich dokładności i efektywności.

2 Zbiór danych i jego przetwarzanie

2.1 Zbiór danych

Zbiór dnaych został pozyskany ze strony kaggle.com. Zawiera on około 50 tysięcy rekordów, które opisującą 9 cech badanego grzyba :

1. średnica kapleusza (podana w cm)
2. kształt kapleusza (jako liczba całkowita od 0 do 6)
3. typ połączenia blaszek z trzonem (jako liczba całkowita od 0 do 6)
4. kolor blaszek (jako liczba całkowita od 0 do 11)
5. wysokość trzonu (podana w cm)
6. szerokość trzonu (podana w mm)
7. kolor trzonu (jako liczba całkowita od 0 do 12)
8. sezon zberania (jako liczba rzeczywista od 0.02 do 1.804)
9. klasa grzyba (0 lub 1 zależnie od jadalności)

Tabela 1: Przykładowy wycinek danych

śred kap	**ksz kap**	**poł bla**	**kol bla**	**wys trz**	**szer trz**	**kol trz**	**sezon**	**klasa**
1372	2	2	10	3.807467	1545	11	1.804273	1
1461	2	2	10	3.807467	1557	11	0.943194	1
1371	6	0	10	3.612496	1566	11	1.804273	0

2.2 Przetwarzanie wstępne

2.2.1 Tabela sezon

Wartości w tej tabeli przyjmują 4 wartości :

Wartości w tabeli sezon
1.80427271
0.94319455
0.88845029
0.02737213

Metadane dotyczące tego zbioru informują, że ta tabela określa porę roku w którym został zebrany dany grzyb. Aby poprawić czytelność zamieniłem je na liczby całkowite od 1 do 4.

2.2.2 Tabela klasa

Aby poprawić czytelność danych nazwa została zmieniona na "jadalność", jako że lepiej przedstawia opisywaną wartość.

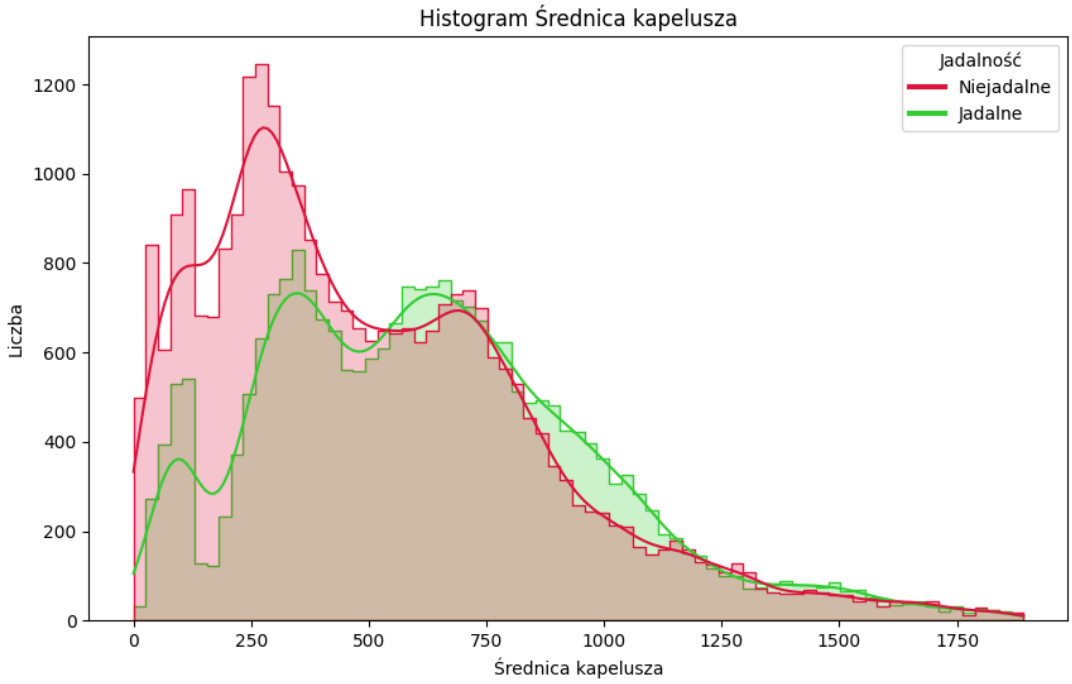
2.2.3 Przetworzone dane

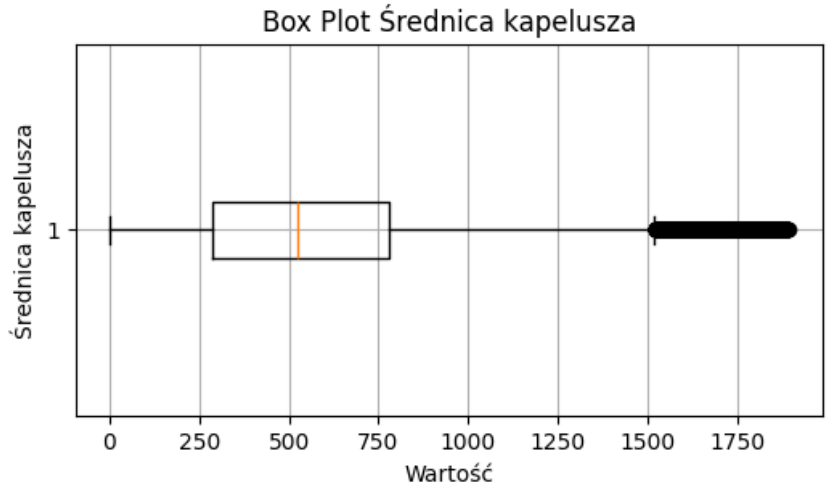
Tabela 2: Przykładowy wycinek przetworzonych danych

śred kap	**ksz kap**	**poł bla**	**kol bla**	**wys trz**	**szer trz**	**kol trz**	**sezon**	**jadalność**
1372	2	2	10	3.807467	1545	11	1	1
1461	2	2	10	3.807467	1557	11	2	1
1371	6	0	10	3.612496	1566	11	1	0

2.3 Analiza eksploracyjna

2.3.1 Średnica kapelusza

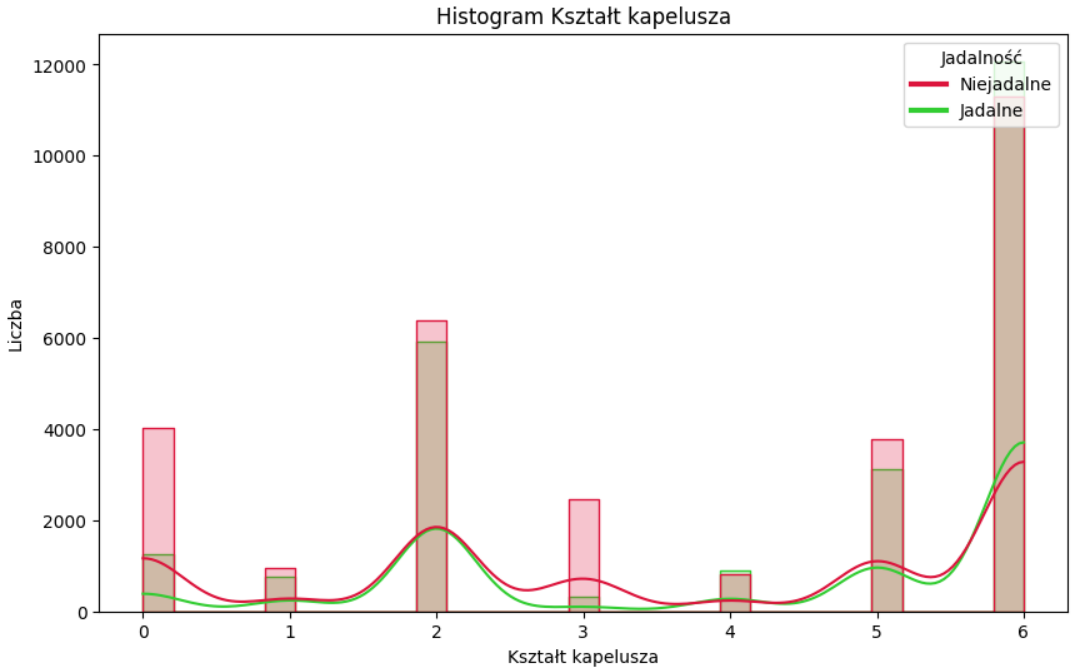


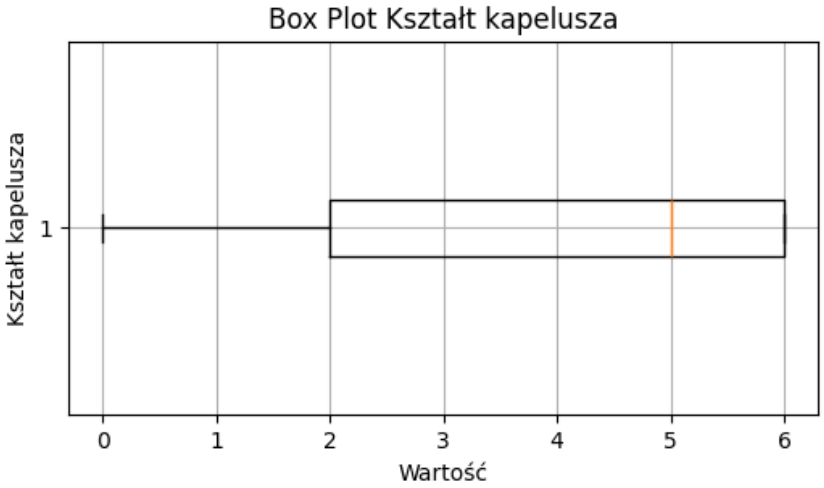


Cecha	Min	Max.	Średnia	Std	Wartości unikalne
Średnica kapelusza	0.0	1891.0	567.25	359.88	1847

Tabela 3: Podstawowe statystyki dla średnicy kapelusza

2.3.2 Kształt kapelusza

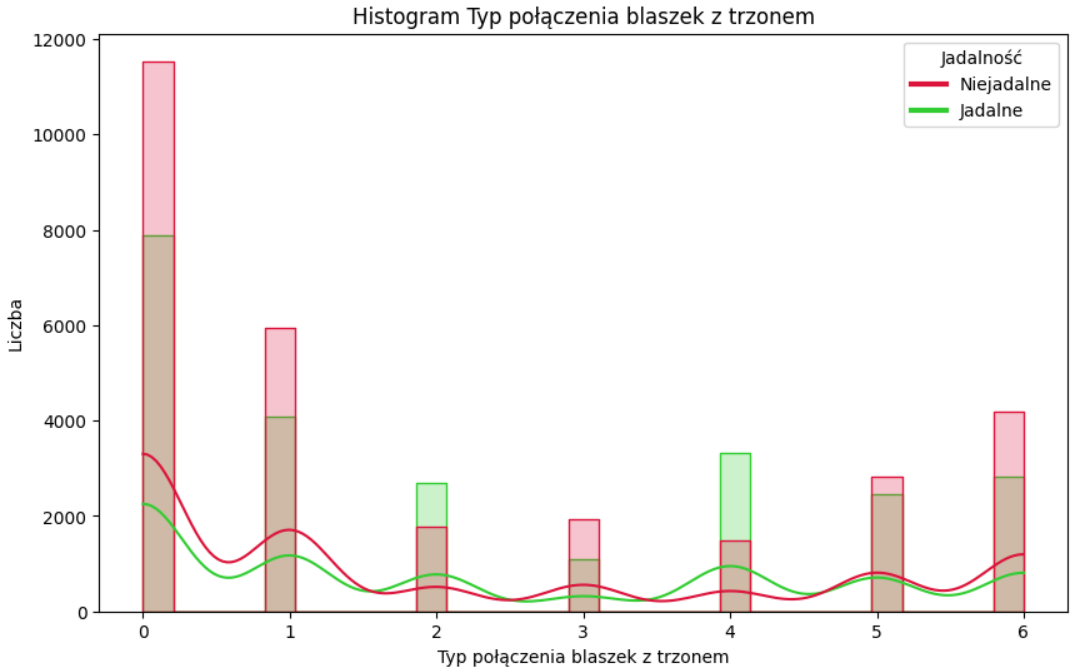


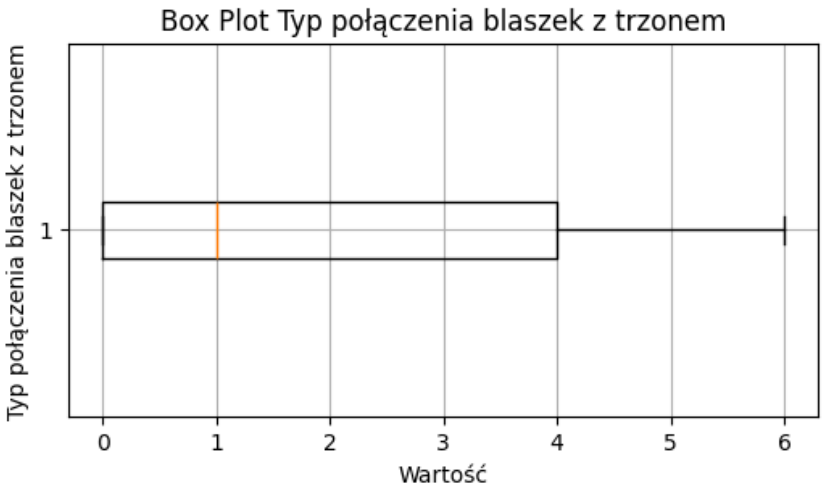


Cecha	Min	Max.	Wartości unikalne
Kształt kapelusza	0	6	7

Tabela 4: Podstawowe statystyki dla kształtu kapelusza

2.3.3 Typ połączenia blaszek z trzonem

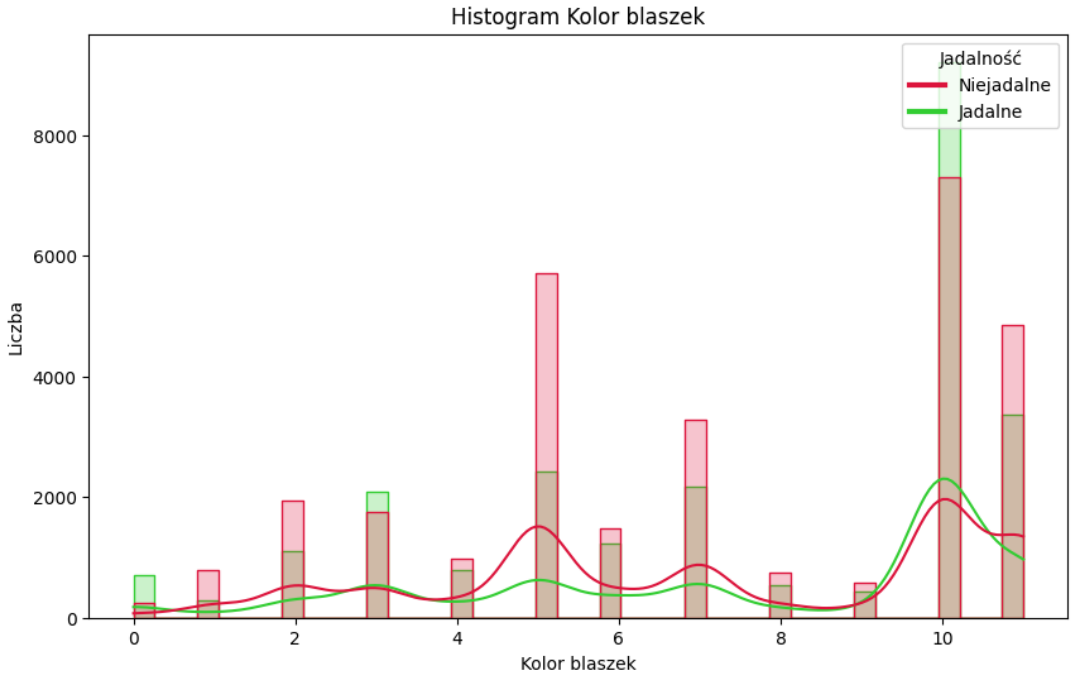


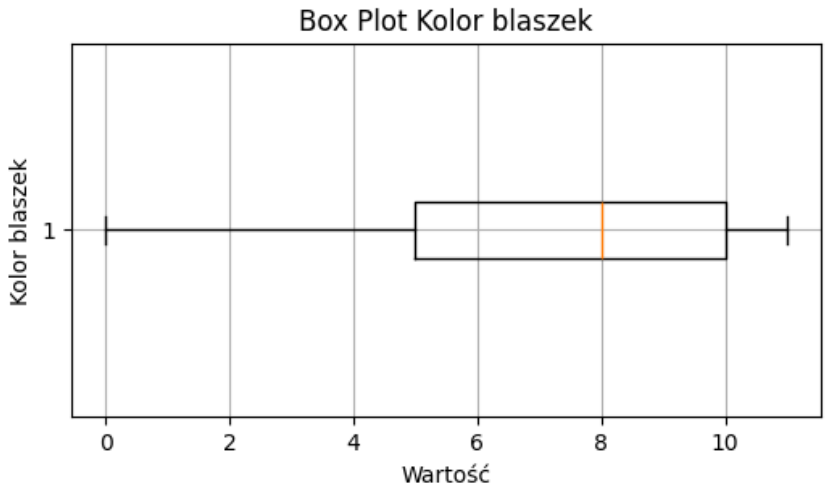


Cecha	Min	Max.	Wartości unikalne
Typ połączenia blaszek	0	6	7

Tabela 5: Podstawowe statystyki dla typu połączenia blaszek

2.3.4 Kolor blaszek

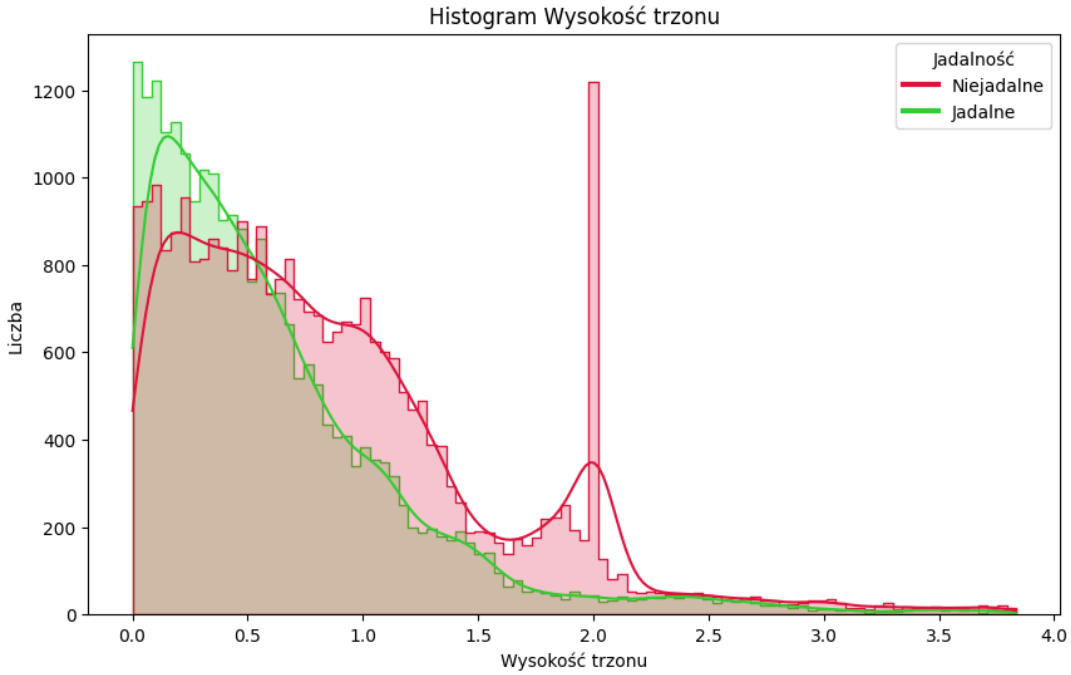


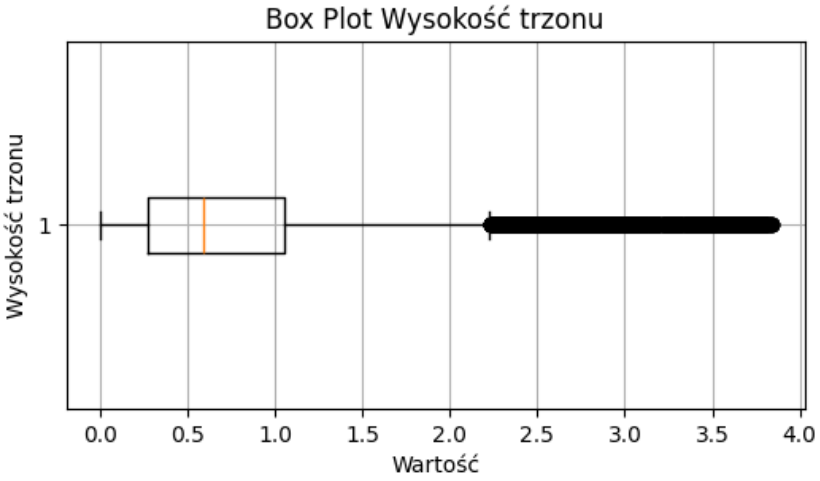


Cecha	Min	Max.	Wartości unikalne
Kolor blaszek	0	11	12

Tabela 6: Podstawowe statystyki dla koloru blaszek

2.3.5 Wysokość trzonu

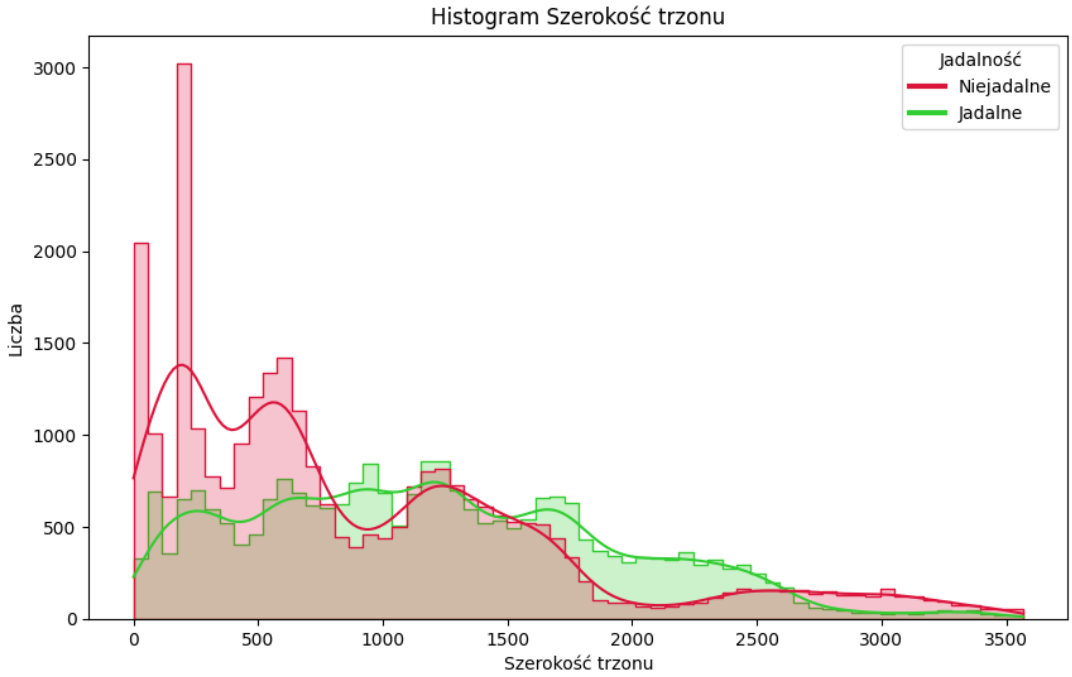


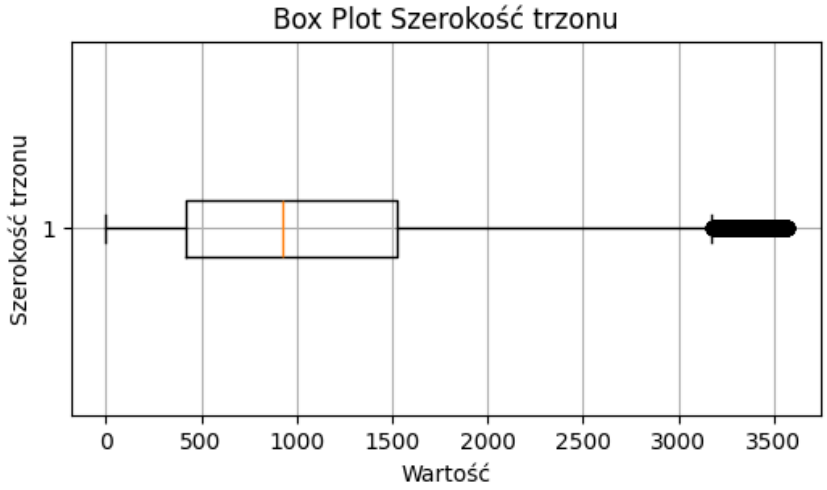


Cecha	Min	Max.	Średnia	Std	Wartości unikalne
Wysokość trzonu	0.0	3.83	0.76	0.65	1454

Tabela 7: Podstawowe statystyki dla wysokości trzonu

2.3.6 Szerokość trzonu

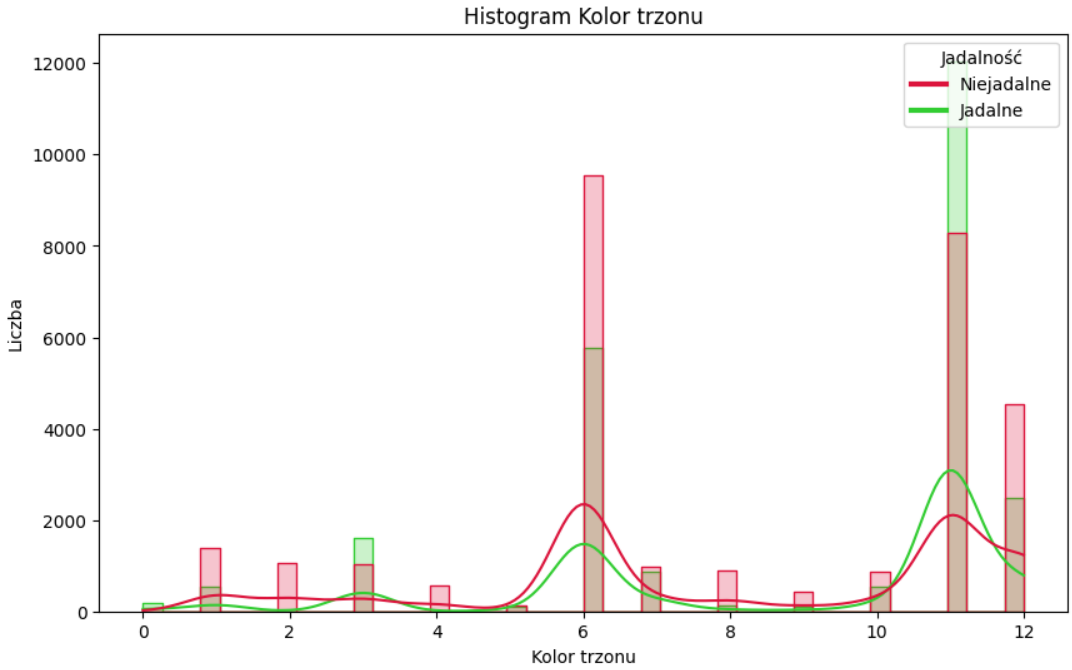


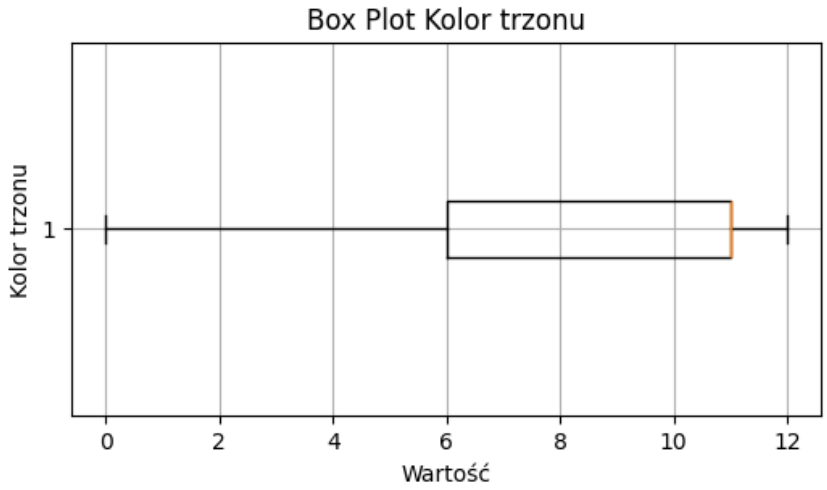


Cecha	Min	Max.	Średnia	Std	Wartości unikalne
Szerokość trzonu	1.0	3569.0	1072.16	775.42	3509

Tabela 8: Podstawowe statystyki dla szerokości trzonu

2.3.7 Kolor trzonu

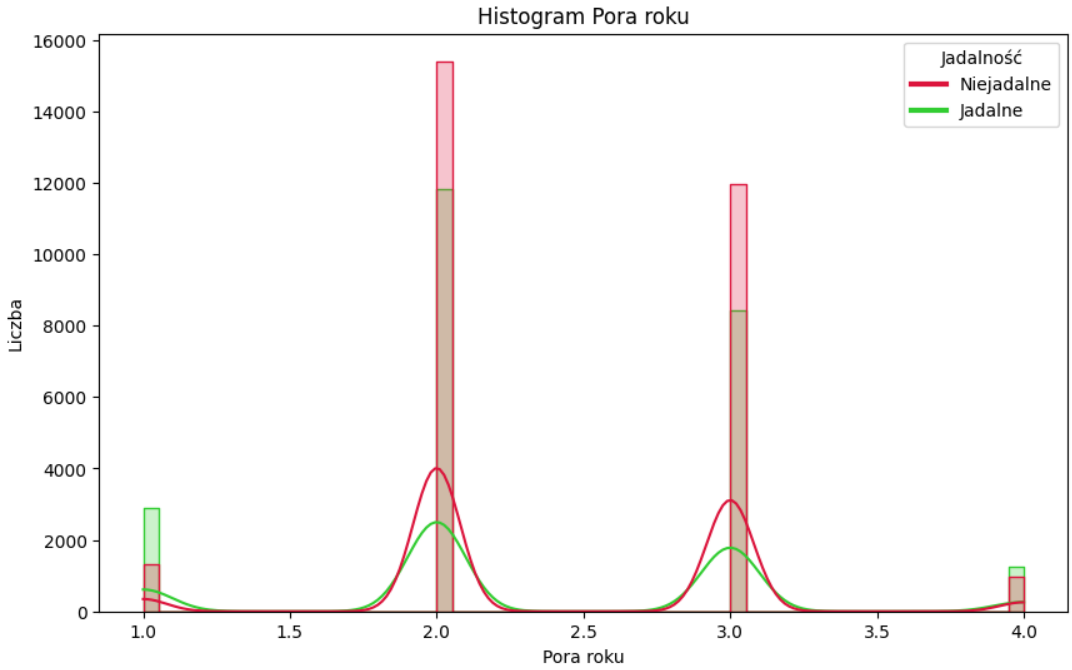


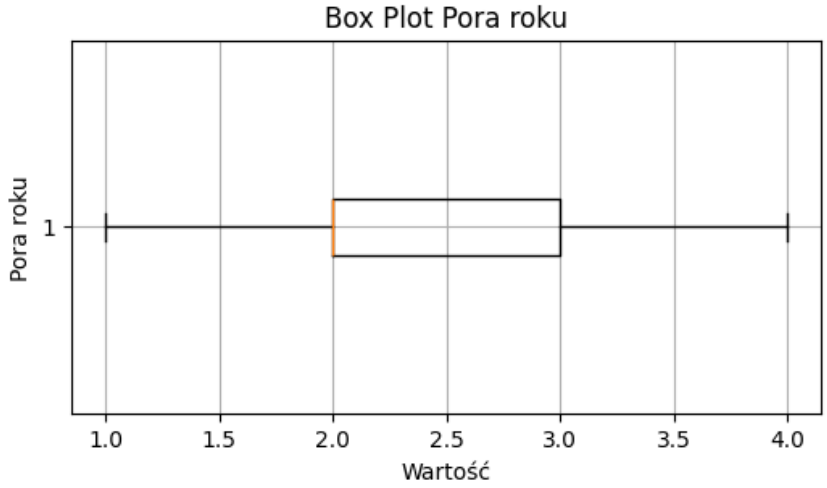


Cecha	Min	Max.	Wartości unikalne
Kolor trzonu	0	12	13

Tabela 9: Podstawowe statystyki dla koloru trzonu

2.3.8 Pora roku

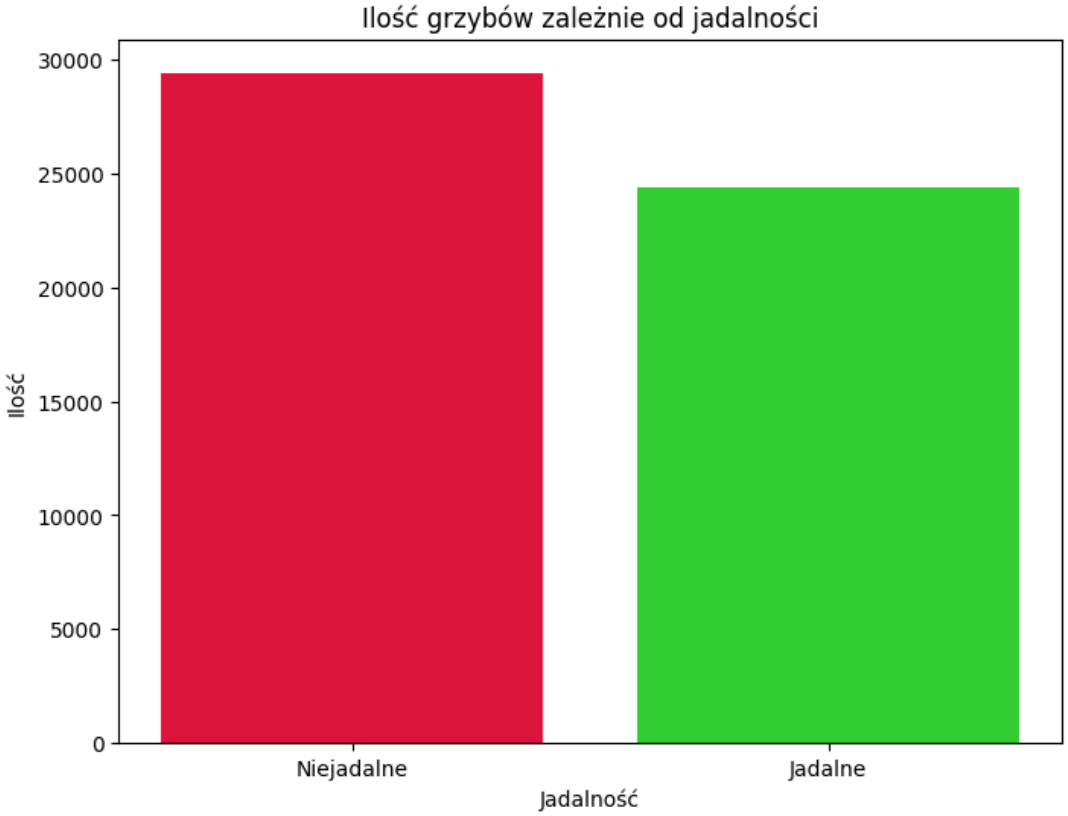




Cecha	Min	Max.	Wartości unikalne
Szerokość trzonu	1	4	4

Tabela 10: Podstawowe statystyki dla pory roku

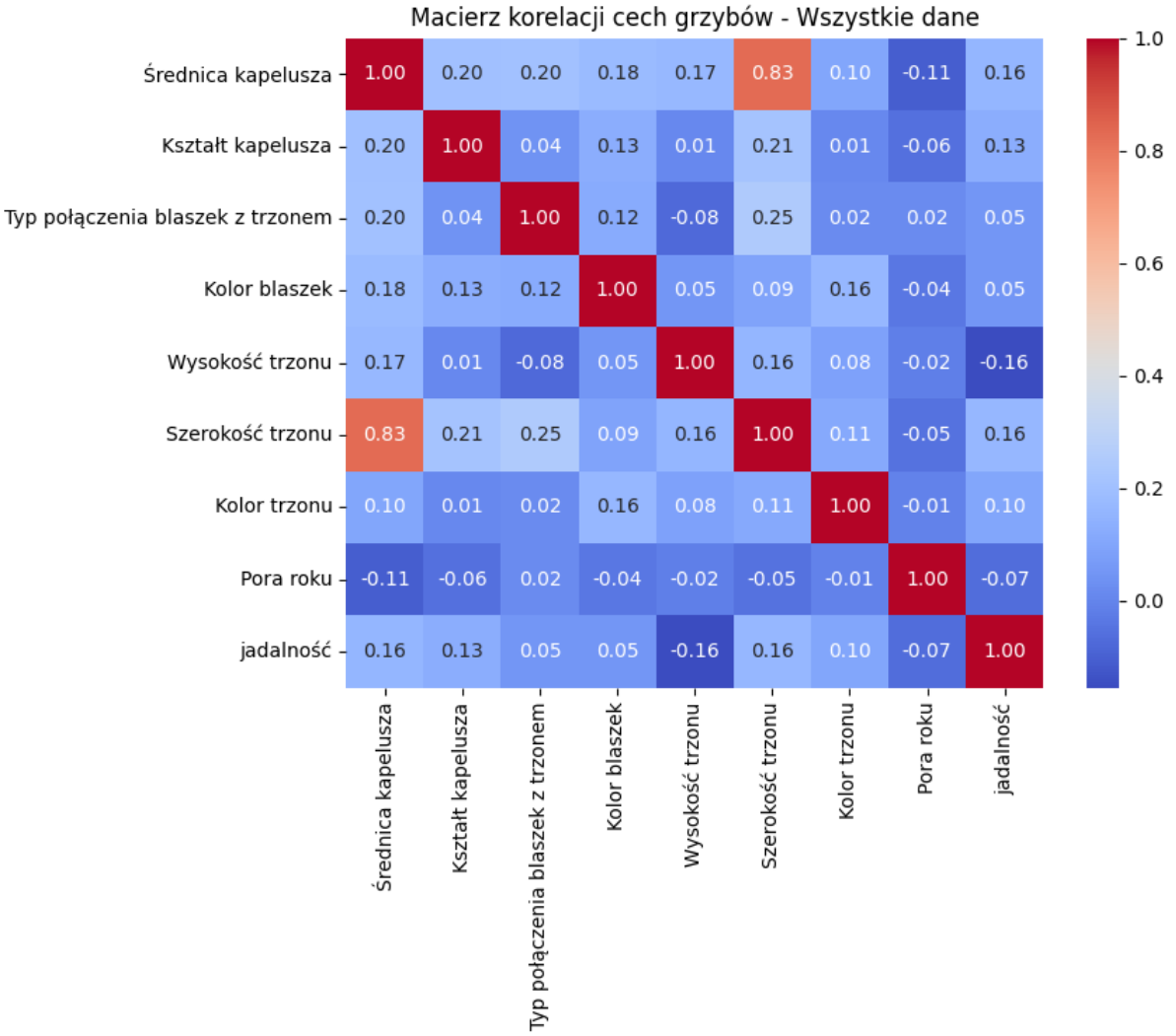
2.3.9 Jadalność

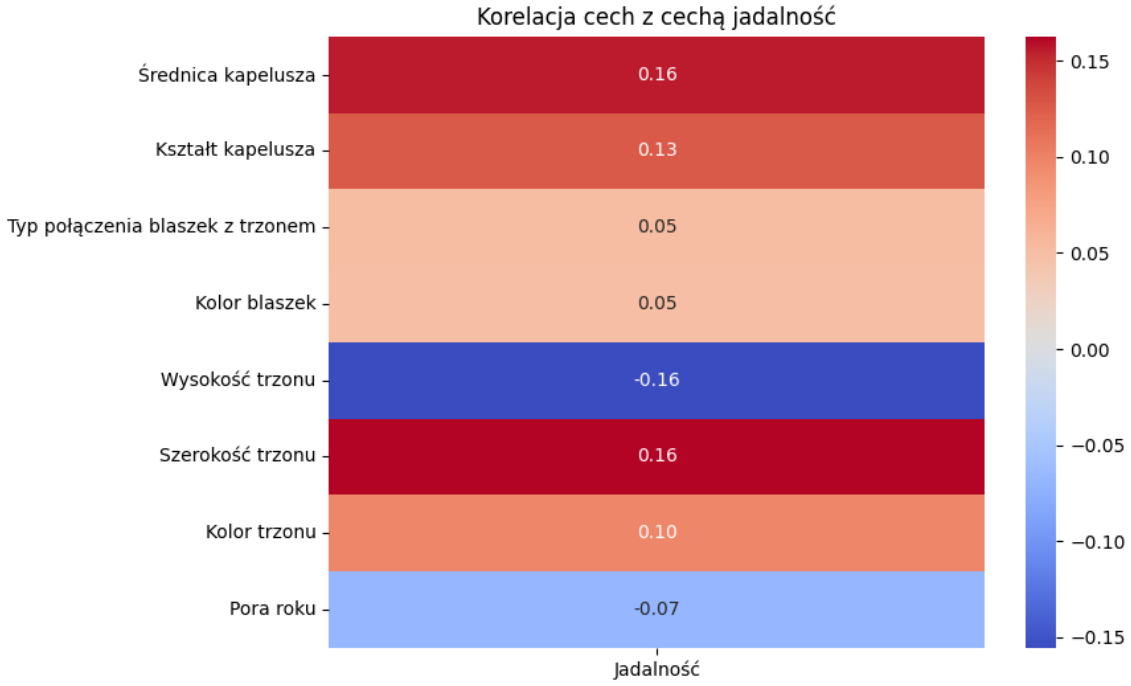


Cecha	Niejadalne	Jadalne	% Jadalnych
Jadalność	28612	24360	45,98 %

Tabela 11: Podstawowe statystyki dla jadalności

2.3.10 Zależnosci między cechami





2.3.11 Podsumowanie analizy danych

Analiza danych wykazała, że określenie jadalności grzyba jest złożonym problemem. Na histogramach można zauważyć niewielkie różnice między cechami grzybów jadalnych i trujących, ale oba zbiory są do siebie bardzo podobne.

Najwyższa zaobserwowana korelacja między jadalnością a inną cechą wynosiła **0.16**, co oznacza, że żadna pojedyncza cecha nie jest wystarczająco silnie związana z jadalnością, aby samodzielnie przewidywać, czy grzyb jest jadalny czy trujący.

Subtelne różnice sugerują jednak, że klasyfikacja jest możliwa przy zastosowaniu odpowiednich metod analizy wielowymiarowej.

Box ploty zostały wykonane w celu zilustrowania rozkładu danych. Wartości uznane za odstające **nie zostały usunięte**, ponieważ pochodzą z pomiarów realnych grzybów i ich zachowanie pozwala na bardziej wiarygodną analizę, oddającą pełne spektrum zróżnicowania w populacji badanych okazów. Usunięcie tych wartości mogłoby zniekształcić wyniki i zafałszować obraz rzeczywistości. Wartości skrajne mogą dostarczać cennych informacji na temat ekstremalnych przypadków, które mogą być kluczowe w kontekście bezpieczeństwa żywnościowego i toksykologii.

3 Klasyfikacja

W celu klasyfikacji jadalności grzybów początkowo zdecydowałem się na wybór trzech różnych metod uczenia maszynowego: **SVM, Sieci neuronowych oraz Drzew losowych**. Z powodu trudności na jakie natknąłem się przy SVM zdecydowałem się na dodanie dwóch kolejnych modeli : **Regresji logistycznej i Naive Bayes**.

3.1 SVM

Przy tej próbie klasyfikacji została zastosowana implementacja SVM z biblioteki scikit-learn. Niestety proces trenowania modelu dla tego zbioru danych przekroczył możliwości obliczeniowe mojego komputera, po 20 godzinach oczekiwania zakończyłem proces i zdecydowałem się na poszukanie alternatyw.

3.2 Sieć Neuronowa

Zastosowałem implementację sieci neuronowej przy użyciu biblioteki Keras. Sieć ma strukturę 8-16-8-1, używa funkcji aktywacji ReLU i sigmoidy. Model był trenowany przez 500 epok.
W różnych eksperymentach jej skuteczność wynosiła **do 75 % do 85 %**.

3.3 Drzewa Losowe

Model drzew losowych (Random Forest Classifier) z biblioteki scikit-learn osiągnął imponującą skuteczność na poziomie **99%**, co czyni go najlepszym z testowanych klasyfikatorów. Może to wynikać z dużej ilości danych kategorycznych w badanym zbiorze.

3.4 Regresja Logistyczna

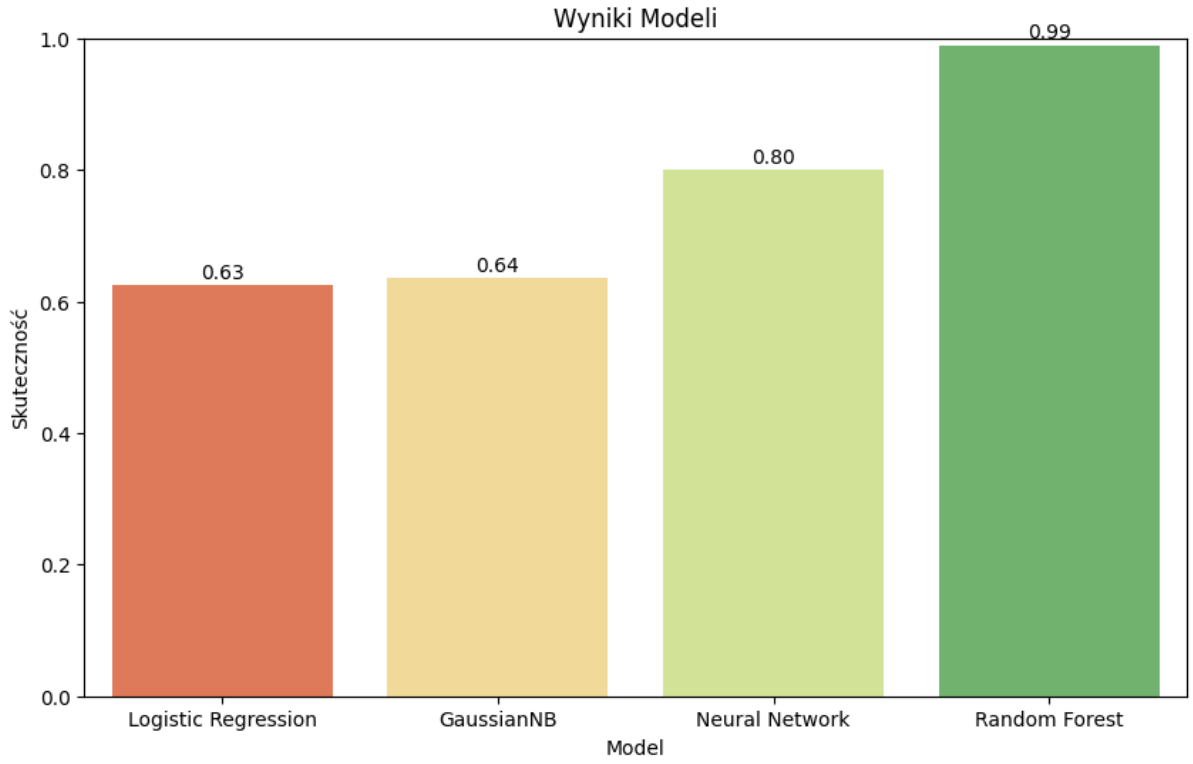
Model regresji logistycznej z biblioteki scikit-learn osiągnął skuteczność na poziomie **63%**, co jest zauważalnie niższym wynikiem w porównaniu do modelu drzew losowych i sieci neuronowej. Skuteczność regresji logistycznej może być ograniczona przez jej liniowy charakter, co utrudnia modelowanie bardziej złożonych nieliniowych zależności w danych.

3.5 Naive Bayes (Naiwny klasyfikator bayesowski)

Użyłem implementacji GaussianNB z biblioteki scikit-learn, którego skuteczność wyniosła **64%**. Słaby wynik mogą wynikać z tego że, model ten jest skuteczny w klasyfikacji gdy dane mają rozkład normalny a opracowywany zbiór grzybów nie ma takich właściwości.

4 Podsumowanie

Na koniec przeprowadziłem eksperyment polegający na 5-krotnej kroswalidacji modeli. Średnie wyniki każdego z nich zostały przedstawione na poniższym wykresie.



Celem tej pracy było zbadanie możliwości klasyfikacji jadalności grzybów na podstawie ich cech za pomocą różnych algorytmów uczenia maszynowego. Najlepszym i jedynym realnie skutecznym modelem okazał się Random Forest z niemal 100% skutecznością.

Z przeprowadzonej analizy wynika więc, że przy użyciu Random Forest jesteśmy w stanie niemal bezbłędnie określić jadalność grzyba.