



# Klasyfikacja typu ryżu na podstawie cech ziarna

## Cel projektu:

Analiza eksploracyjna danych oraz budowa i porównanie modeli kNN i Naiwnego Bayesa do klasyfikacji typu ryżu na podstawie numerycznych cech geometrycznych ziarna oraz ocena jakości modeli przy użyciu wybranych miar klasyfikacyjnych.

# Dane zbioru

**Dane – Rice Type Classification**  
**Źródło – Kaggle**

**Rozmiar zbioru:**

- 18 185 obserwacji
- 12 kolumn (11 cech + 1 etykieta)

**Typ danych:**

- wszystkie cechy numeryczne (float64 / int64)
- brak brakujących danych

**Cechy opisują wielkość i kształt ziarna:** powierzchnię (Area), wymiary (Major/MinorAxisLength), wydłużenie (Eccentricity), obszar wypukły (ConvexArea), średnicę równoważną (EquivDiameter), wypełnienie obszaru (Extent), obwód (Perimeter), okrągłość (Roundness) oraz proporcje wymiarów (AspectRatio).

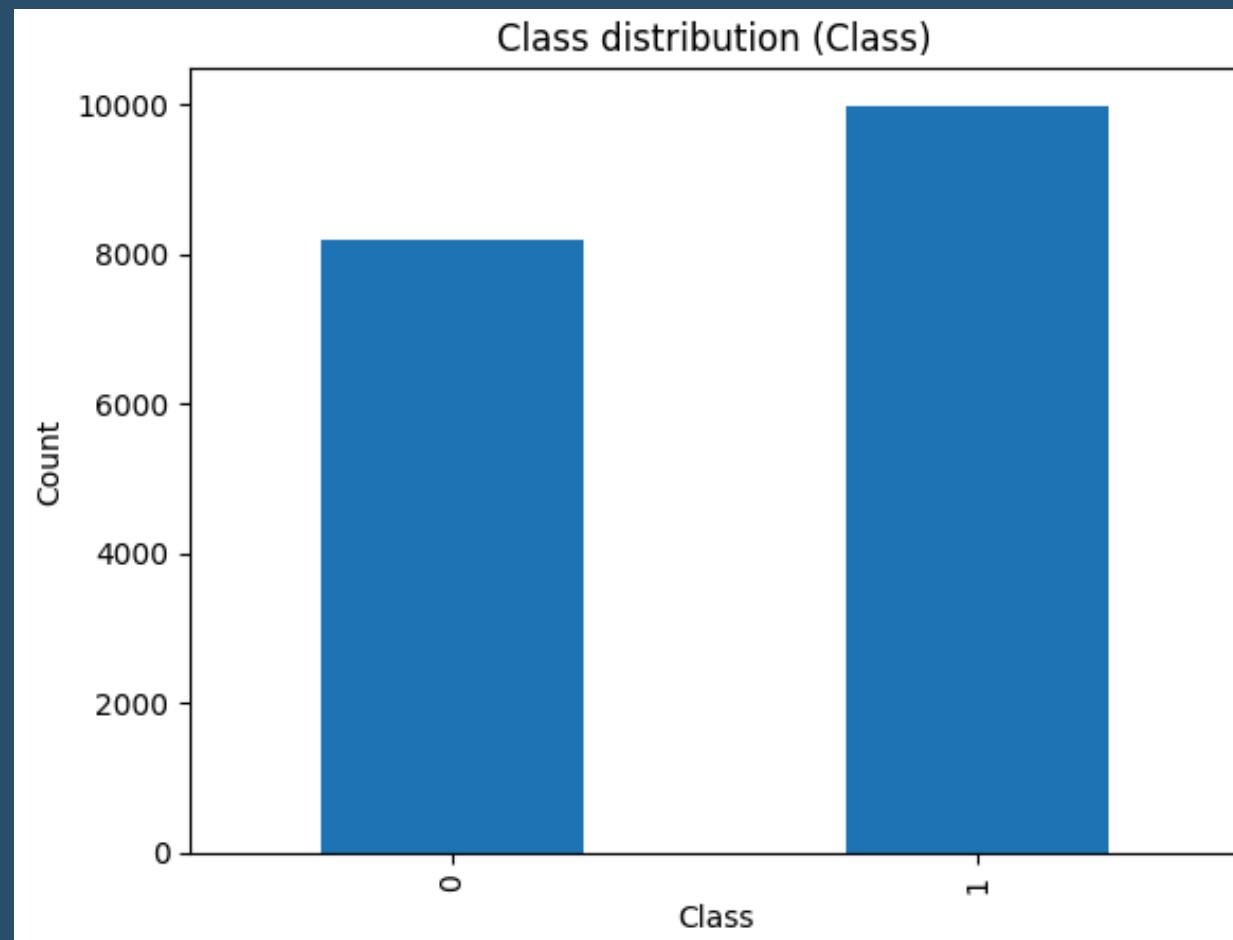
**Etykieta:**

**Class**

- 0 – Gonen
- 1 – Jasmine



# EDA (analiza eksploracyjna)

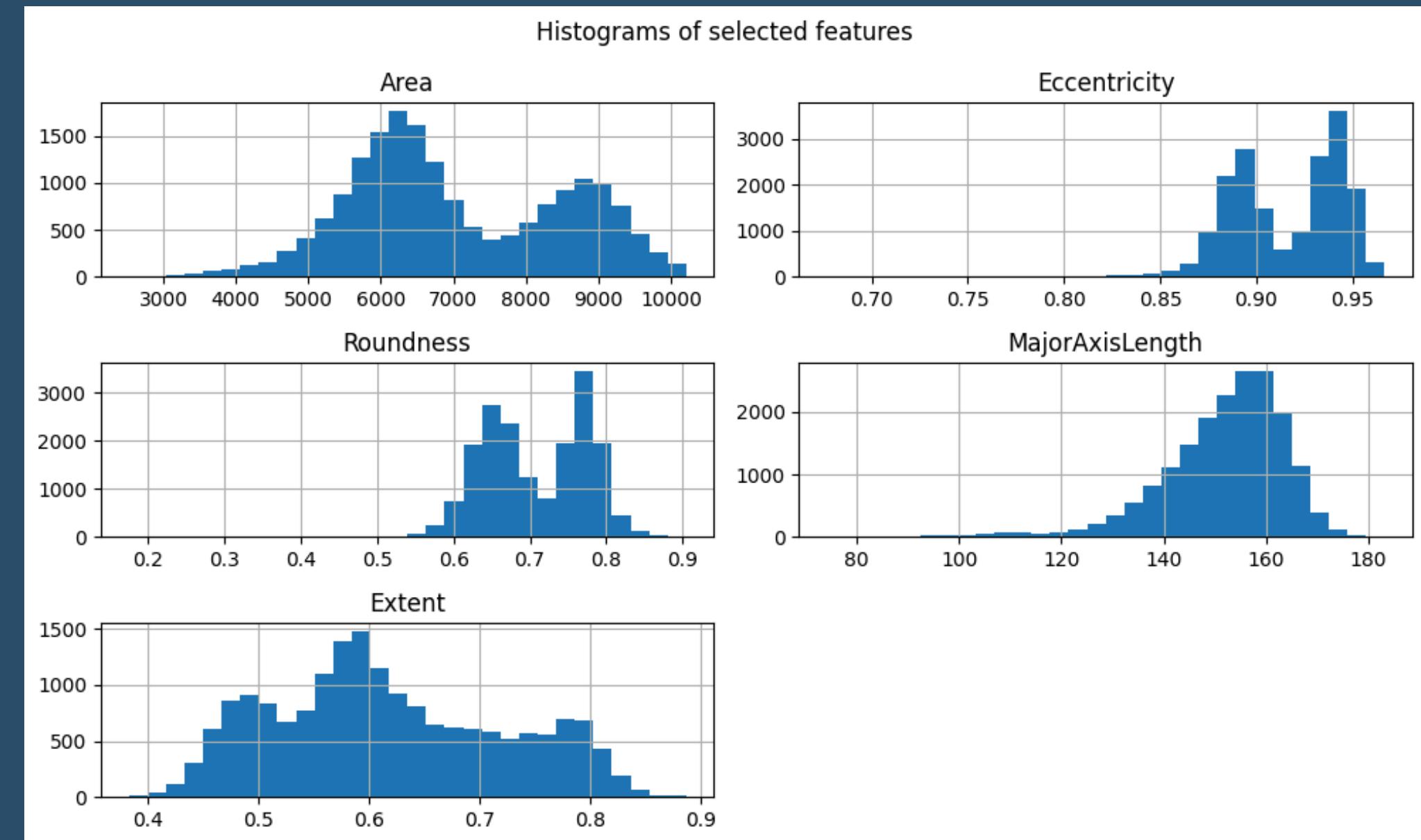


## WYKRES ROZKŁADU KLAS

### Zmienna docelowa (Class):

- Klasa 0 – Gonien: 8200 próbek ( $\approx 45\%$ )
- Klasa 1 – Jasmine: 9985 próbek ( $\approx 55\%$ )

Zbiór jest względnie zbalansowany. Różnica licznosci nie jest na tyle duża, aby wymagać dodatkowych technik balansowania klas.



## HISTOGRAMY CECH

Area, Eccentricity, Roundness,  
MajorAxisLength, Extent

Histogramy pokazują różne zakresy i rozkłady cech. Część cech wykazuje asymetrię oraz możliwe skupienia wartości, co może wpływać na skuteczność modeli i uzasadnia zastosowanie skalowania danych.

# EDA (analiza eksploracyjna)

# MACIERZ KORELACJI CECH

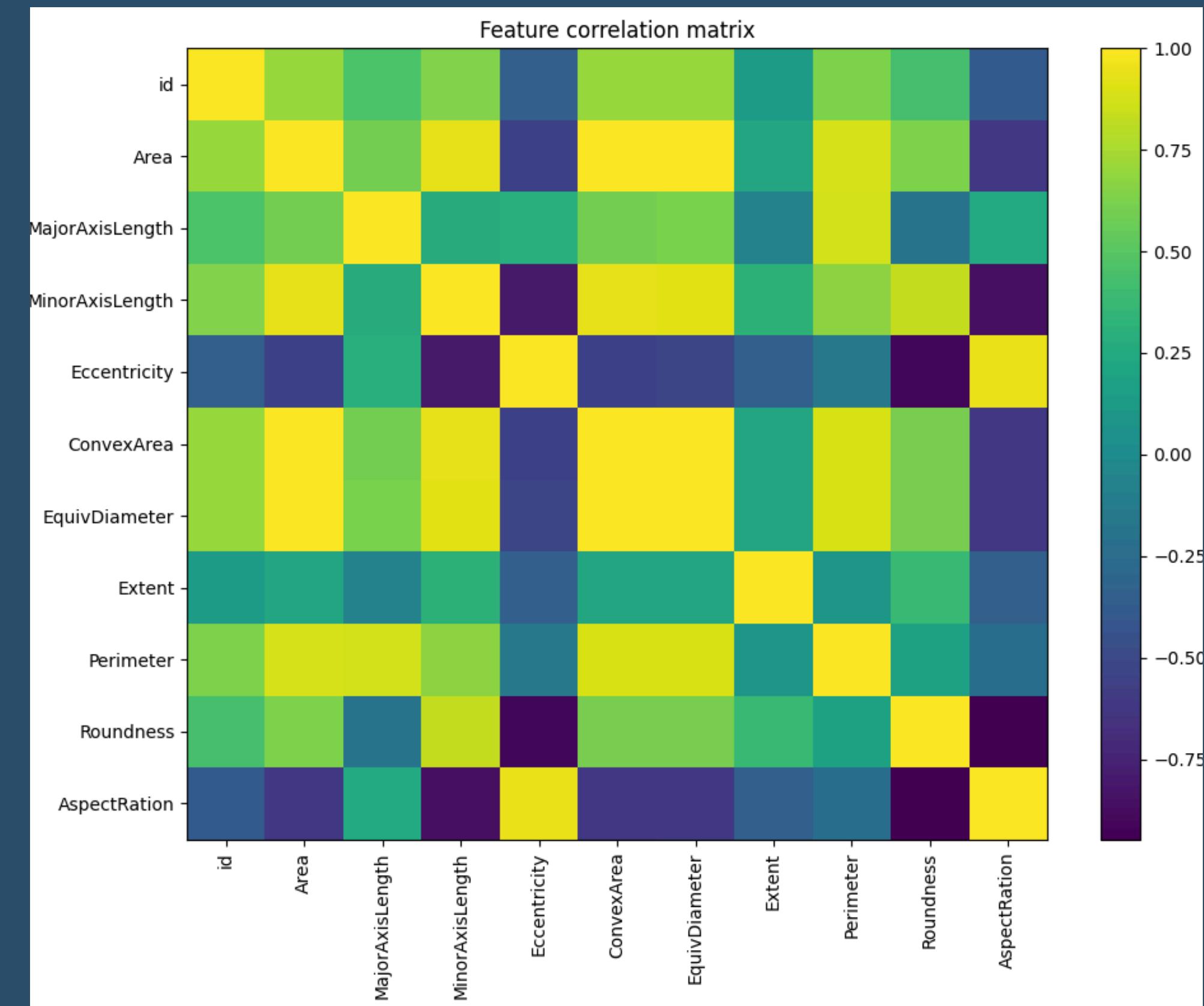
Macierz korelacji pokazuje silne zależności między niektórymi cechami, co wskazuje na redundancję informacji i uzasadnia ich redukcję przed modelowaniem.

## Najsilniejsze zależności:

- Area – ConvexArea ( $r \approx 0.999$ )
  - Area – EquivDiameter ( $r \approx 0.998$ )
  - Eccentricity – AspectRatio ( $r \approx 0.95$ )
  - Roundness – AspectRatio ( $r \approx 0.95$ )

## Znaczenie dla modelu:

Redukcja silnie skorelowanych cech zmniejsza ryzyko przeuczenia oraz poprawia stabilność i interpretowalność modelu.



# Przygotowanie danych

Celem przygotowania danych było ograniczenie nadmiarowości cech oraz poprawa jakości uczenia modeli.



## PRZYGOTOWANIE DANYCH

- brak brakujących danych – dane nie wymagały czyszczenia
- usunięcie kolumny id (brak wartości informacyjnej)
- redukcja silnie skorelowanych cech ( $|r| > 0.95$ )
- usunięto: ConvexArea, EquivDiameter, AspectRatio, Perimeter
- liczba cech po redukcji: 5
- standaryzacja cech (StandardScaler) – ujednolicenie skali zmiennych, istotne dla algorytmów opartych na odległości (kNN)

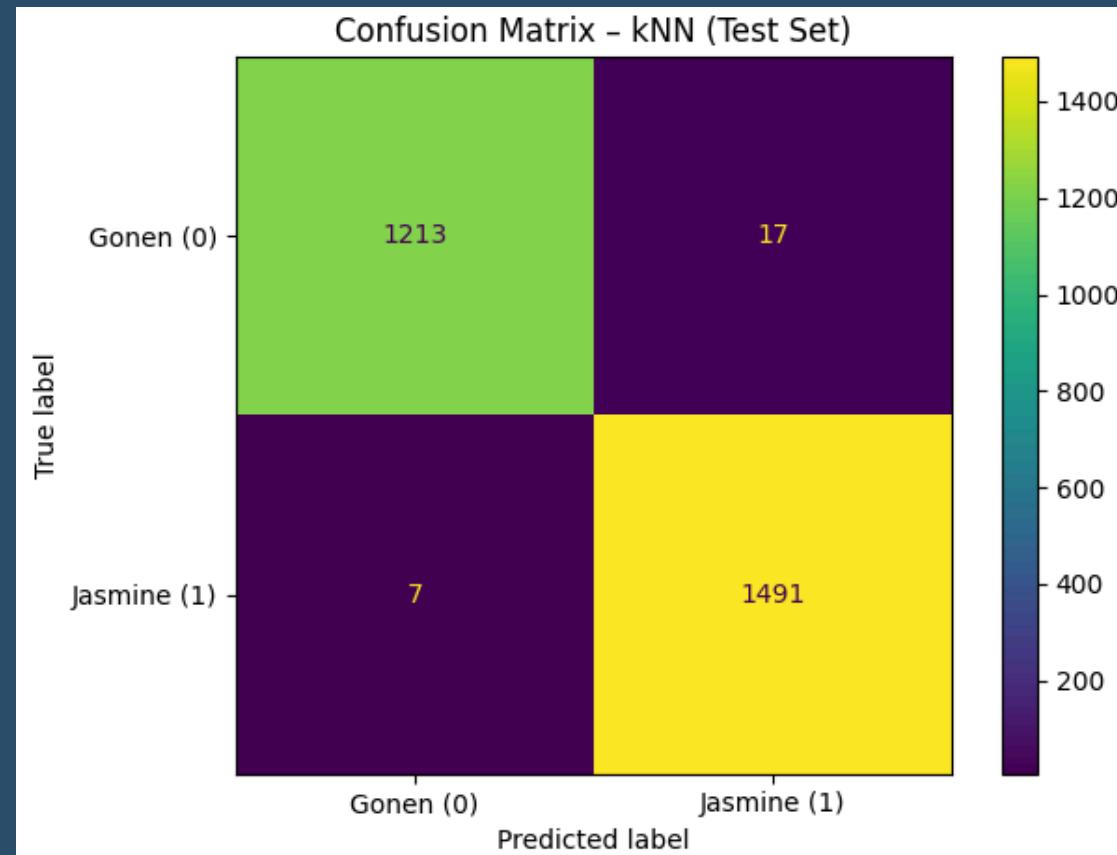
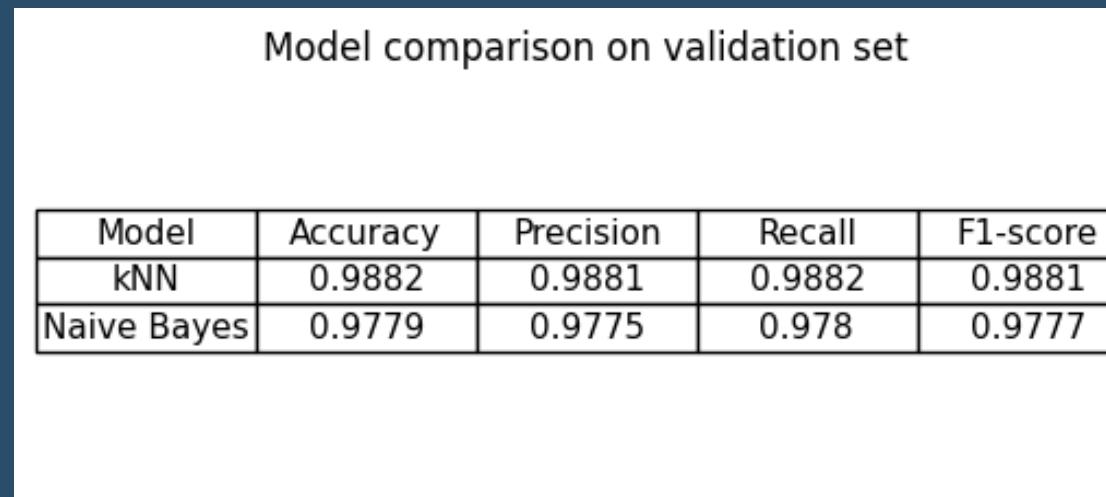
## PODZIAŁ DANYCH

- trening: 12 736 próbek (70%)
- walidacja: 2 721 próbek (15%)
- test: 2 728 próbek (15%)

## MODELE

- **kNN** – strojenie parametru k na zbiorze walidacyjnym
- → najlepsze k = 5
- **Naiwny Bayes** (GaussianNB) – model porównawczy

# Wyniki i wnioski



## WALIDACJA (SELEKCJA MODELU)

Selekcja modelu została wykonana na zbiorze walidacyjnym z wykorzystaniem metryk accuracy, precision, recall i F1-score. Oba modele osiągnęły wysoką skuteczność, jednak kNN uzyskał nieco lepsze wyniki.

Na podstawie walidacji dobrano hiperparametr  $k$  (najlepsze  $k = 5$ ).

## EWALUACJA (TEST – KNN)

Wybrany model kNN został oceniony na niezależnym zbiorze testowym, osiągając wysoką skuteczność: accuracy  $\approx 0.991$ , precision  $\approx 0.989$ , recall  $\approx 0.995$  oraz F1-score  $\approx 0.992$ .

Macierz pomyłek potwierdza niewielką liczbę błędnych klasyfikacji i dobrą generalizację modelu.

## WNIOSKI

Model kNN okazał się najlepszym rozwiązaniem, przewyższając Naiwny Bayes pod względem skuteczności i stabilności predykcji.

Redukcja cech, standaryzacja danych oraz strojenie parametru  $k$  pozytywnie wpłynęły na jakość klasyfikacji.