Sprawozdanie z analizy zbiorów danych medycznych
Medyczne zastosowania uczenia maszynowego

Temat: Analiza zbiorów danych dotyczących raka prostaty, zawierających próbki moczu, surowicy i tkanek

Katarzyna Pieczonka, nr indeksu 132785
Informatyka II stopień, semestr zimowy
rok akademicki 2024/25

1. Biblioteki i zbiory danych

Biblioteki używane podczas analizy:

- pandas - biblioteka do analizy danych w języku Python, która oferuje struktury danych i narzędzia do manipulacji tabelami (DataFrames). Pandas ułatwia wczytywanie, przetwarzanie i analizowanie danych
- numpy - biblioteka do obliczeń naukowych w Pythonie, oferująca wsparcie dla wielowymiarowych tablic oraz funkcje matematyczne i statystyczne. Jest podstawą dla wielu innych bibliotek analizy danych
- scikit-learn (sklearn) - biblioteka uczenia maszynowego w Pythonie, która oferuje różnorodne algorytmy do klasyfikacji, regresji, klasteryzacji i redukcji wymiarów, a także narzędzia do przetwarzania wstępnego danych i ewaluacji modeli.

Zbiory danych:

Dane pochodzą z trzech zbiorów dotyczących raka prostaty. W plikach znajdują się wyniki badań próbek moczu, tkanek oraz surowicy.
Zbiory zawierają dane pobrane od około 278 przypadków pacjentów. Klasa docelowa to decyzja, czy pacjent jest chory (oznaczone za pomocą 1), czy zdrowy (oznaczone za pomocą 0). Oprócz klasy docelowej dane zawierają 2076 atrybutów, których nazwy to długości fal np. 802.24219.

2. Przygotowanie zbiorów danych

Podczas przygotowania zbiorów danych, sprawdzane są one pod względem brakujących wartości, a następnie usuwane są duplikaty.

```
#Sprawdzanie braków danych w dataset1
print("Liczba brakujących wartości dla poszczególnych kolumn w dataset1:")
missing_values = dataset1.isnull().sum()
for col, value in missing_values.items():
    if value != 0:
        print(f"{col}, brak: {value} wartości")

#Sprawdzanie duplikatów w dataset1
duplicates = dataset1.duplicated().sum()
print(f"Duplikaty: {duplicates}")
dataset1 = dataset1.drop_duplicates()
```

```
Liczba brakujących wartości dla poszczególnych kolumn w dataset1:
Duplikaty: 2
Liczba brakujących wartości dla poszczególnych kolumn w dataset2:
Duplikaty: 0
Liczba brakujących wartości dla poszczególnych kolumn w dataset3:
Duplikaty: 2
```

Jak pokazuje powyższy zrzut ekranu, żaden ze zbiorów nie posiada brakujących wartości, ale w zbiorze 1 (próbki moczu), oraz 3 (próbki tkanki) znajdują się duplikaty, które następnie zostają usunięte.

Każdy ze zbiorów danych przed analizą zostaje skalowany za pomocą funkcji `StandardScaler()` z biblioteki scikit-learn.

3. Model Random Forest

Każdy ze zbiorów danych trenowany jest za pomocą modelu Random Forest z ustawioną ilością drzew 100.

```
model = RandomForestClassifier(n_estimators=100, random_state=42)
model.fit(features_train, np.ravel(labels_train)) #Uczenie klasyfikatora na części treningowej
labels_predicted = model.predict(features_test) #Generowania decyzji dla części testowej
```

4. Miary jakości

Do oceny modelu stosowane są następujące miary:

- dokładność (accuracy) - procent poprawnych predykcji spośród wszystkich predykcji

```
accuracy_score(y_test_classes, y_pred_classes))
```

- F1-score - miara, która uwzględnia zarówno precyzję, jak i czułość - jest to średnia harmoniczna tych dwóch wartości

```
f1_score(y_test_classes, y_pred_classes, average='weighted')
```

- precyzja (precision)

```
precision_score(y_test_classes, y_pred_classes,average='weighted')
```

- czułość (recall)

```
recall_score(y_test_classes, y_pred_classes, average='weighted')
```

Parameter ‘weighted’ używany podczas określania tych miar oznacza, że miara jest obliczana jako średnia ważona dla wszystkich klas. Waga każdej klasy jest proporcjonalna do liczby wystąpień tej klasy w zbiorze testowym.
Średnia ważona jest obliczana z następującego wzoru:

$$M_{\text{weighted}} = \frac{\sum_{i=1}^{n}(w_i \times M_i)}{\sum_{i=1}^{n} w_i}$$

Gdzie:
n - liczba klas
w_i - ilość próbek w klasie i
M_i - miara jakości dla klasy i

Średnia ważona pozwala na uwzględnienie wpływu niezrównoważonych klas na ostateczny wynik. Dzięki temu miara jest bardziej reprezentatywna dla ogólnej wydajności modelu na całym zbiorze danych, szczególnie gdy liczby próbek w różnych klasach znacząco się różnią.

Macierz pomyłek:

```
***Macierz pomyłek***
[[69  0]
 [ 2 95]]
```

Macierz pomyłek obrazuje, czy klasy zostały poprawnie rozpoznane. Ponieważ w zbiorze danych istnieją tylko dwie klasy, więc macierz ta ma tylko dwie kolumny i dwa wiersze. W tym wypadku klasa pierwsza została rozpoznana prawidłowo we wszystkich przypadkach, natomiast klasa druga została nieprawidłowo rozpoznana w dwóch przypadkach.

5. Wyniki analizy

<u>Wyniki analizy próbek moczu:</u>

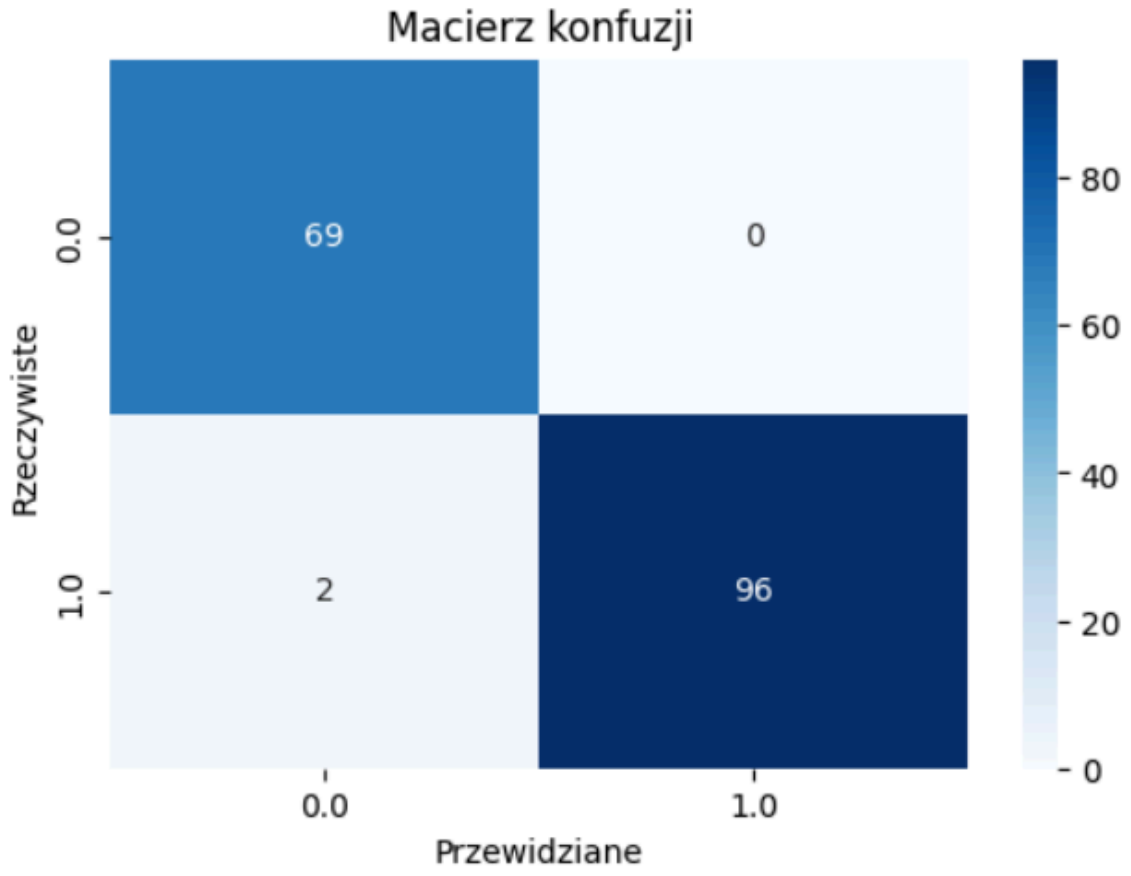
Próba 1:

```
***Kroswalidacja***
Dokładności w poszczególnych foldach: [1.         1.         1.         1.         1.         1.
 1.         1.         1.         0.81818182]
Średnia dokładność: 98.18%
Odchylenie standardowe dokładności: 5.45%
***Wyniki jakości***
Dokładność (Accuracy): 98.80%
Precyzja (Precision): 0.99
Czułość (Recall): 0.99
F1-Score: 0.99
***Macierz pomyłek***
[[69  0]
 [ 2 95]]
***Raport klasyfikacji***
              precision    recall  f1-score   support

         0.0       0.97      1.00      0.99        69
         1.0       1.00      0.98      0.99        97

    accuracy                           0.99       166
   macro avg       0.99      0.99      0.99       166
weighted avg       0.99      0.99      0.99       166
```

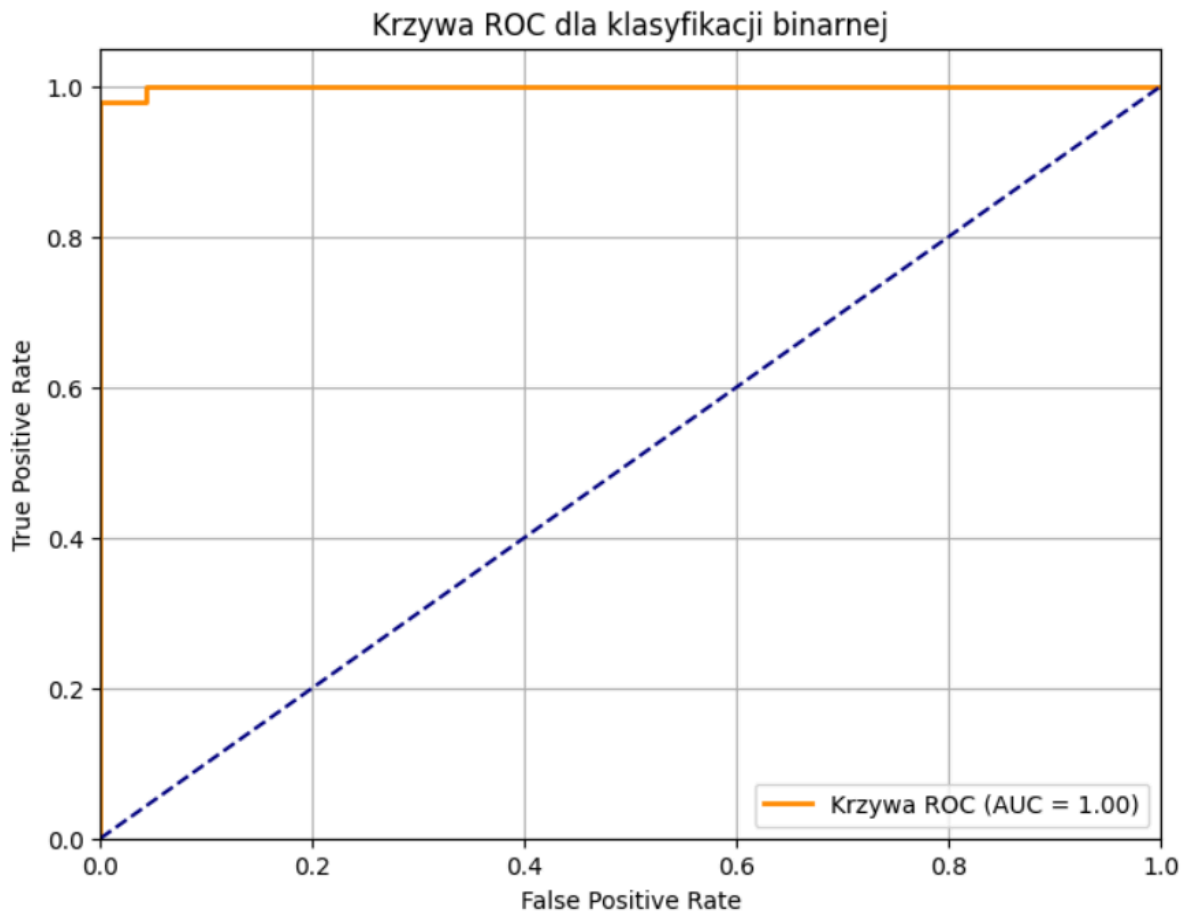
Próba 2:

```
***Kroswalidacja***
Dokładności w poszczególnych foldach: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Średnia dokładność: 100.00%
Odchylenie standardowe dokładności: 0.00%
***Wyniki jakości***
Dokładność (Accuracy): 98.80%
Precyzja (Precision): 0.99
Czułość (Recall): 0.99
F1-Score: 0.99
***Macierz pomyłek***
[[69  0]
 [ 2 96]]
***Raport klasyfikacji***
              precision    recall  f1-score   support

         0.0       0.97      1.00      0.99        69
         1.0       1.00      0.98      0.99        98

    accuracy                           0.99       167
   macro avg       0.99      0.99      0.99       167
weighted avg       0.99      0.99      0.99       167
```

Graficzne przedstawienie macierzy pomyłek:



Oraz krzywa ROC:

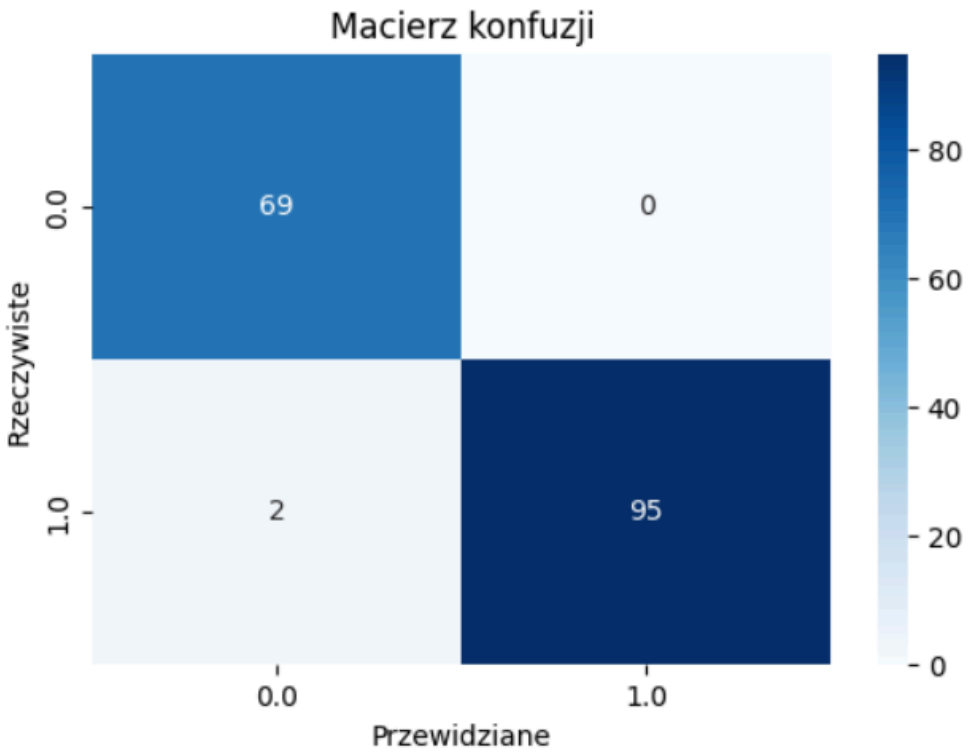


Wysokie wyniki dokładności mogą wskazywać na nadmierne dopasowanie modelu.

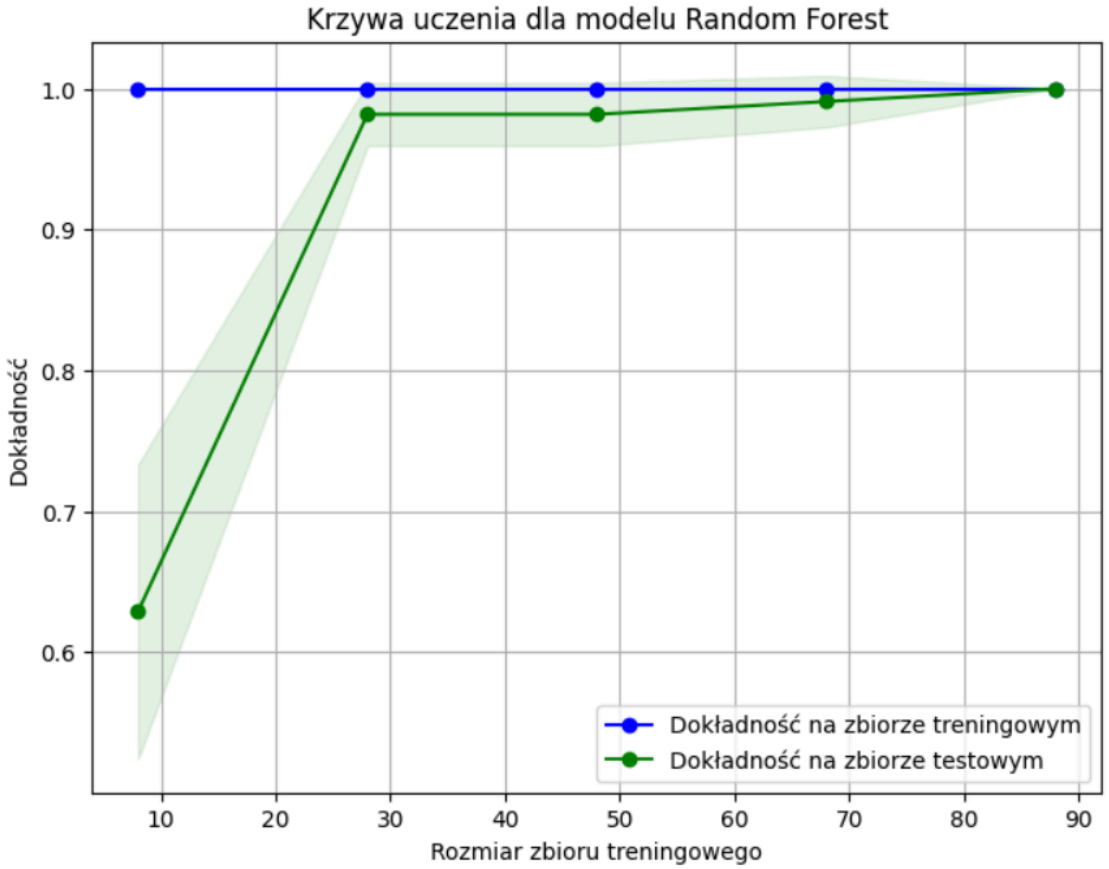
Próba 3:

```
***Kroswalidacja***
Dokładności w poszczególnych foldach: [1.         1.         1.         1.         1.         1.
 1.         1.         1.         0.81818182]
Średnia dokładność: 98.18%
Odchylenie standardowe dokładności: 5.45%
***Wyniki jakości***
Dokładność (Accuracy): 98.80%
Precyzja (Precision): 0.99
Czułość (Recall): 0.99
F1-Score: 0.99
***Macierz pomyłek***
[[69  0]
 [ 2 95]]
***Raport klasyfikacji***
              precision    recall  f1-score   support

         0.0       0.97      1.00      0.99        69
         1.0       1.00      0.98      0.99        97

    accuracy                           0.99       166
   macro avg       0.99      0.99      0.99       166
weighted avg       0.99      0.99      0.99       166
```

Wizualizacja graficzna macierzy pomyłek:



Wykres uczenia dla tego modelu i tego zbioru danych:

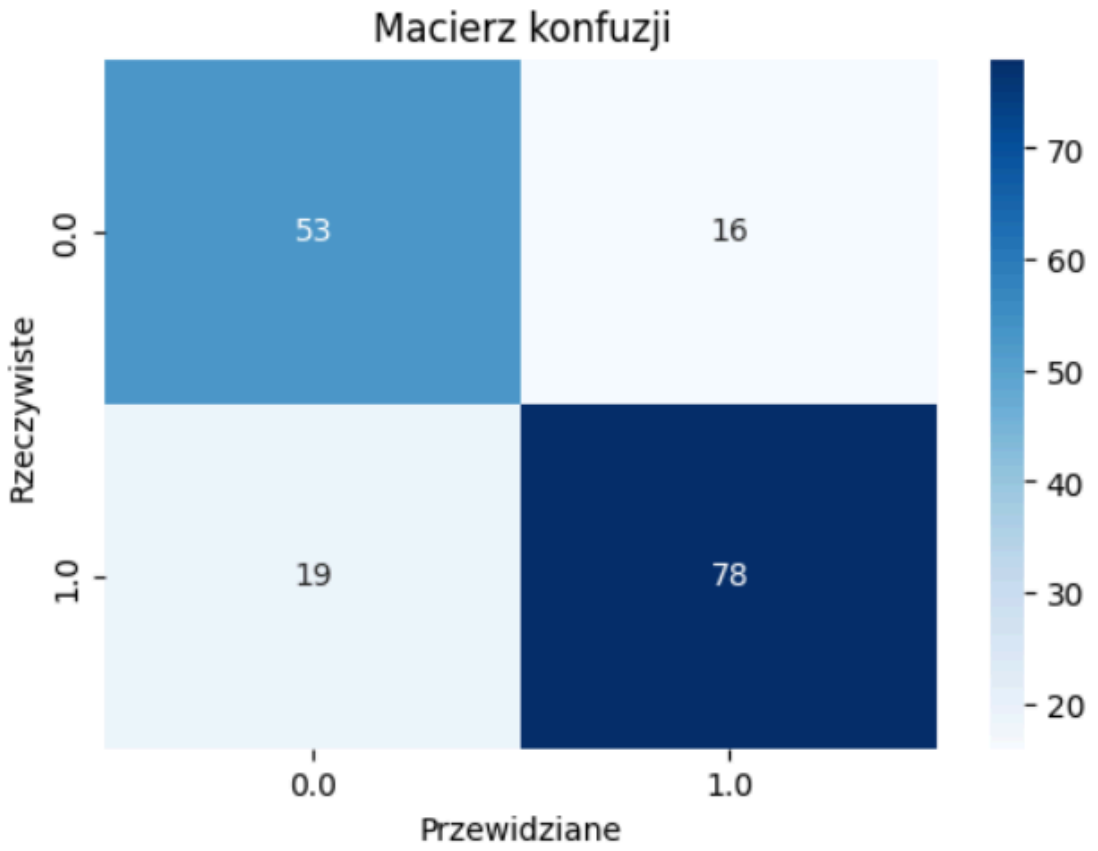


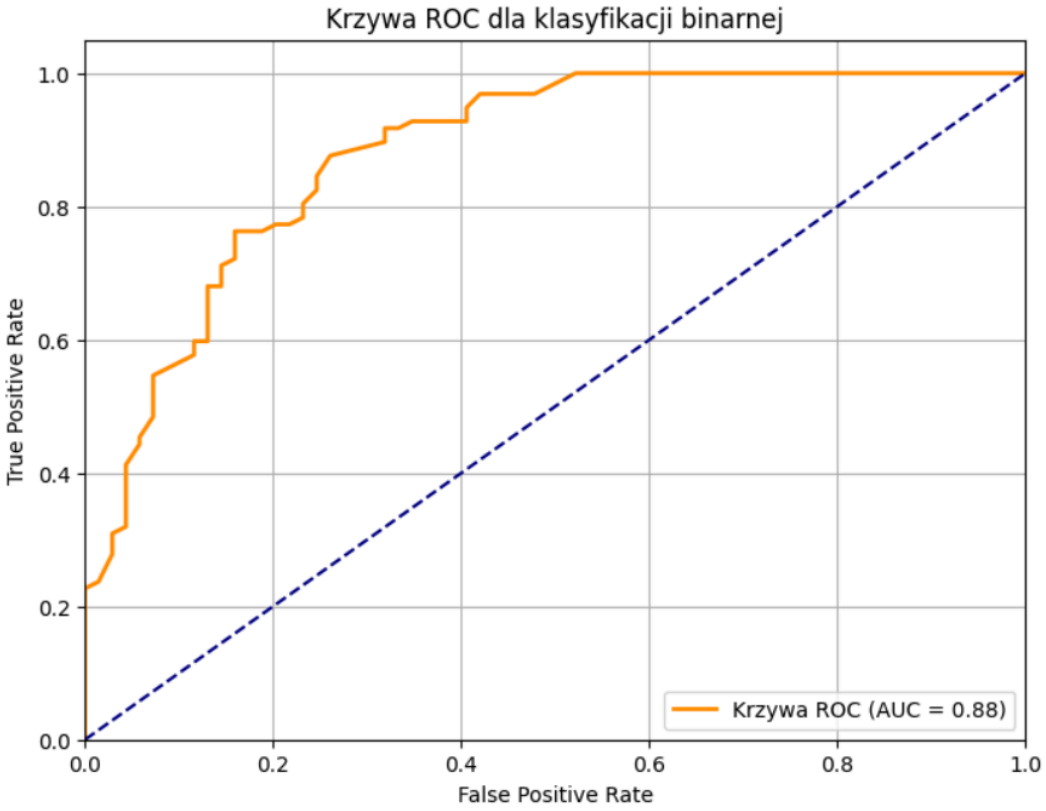
Wykres ten wskazuje, że dla większych zbiorów treningowych, model osiąga wysoką dokładność, co pokazuje jego zdolność do generalizacji. Przy mniejszych zbiorach, dokładność nieco spada, jednak jest to normalne w tym przypadku.
Analiza tego wykresu wskazuje, że nie występuje tutaj problem z przeuczeniem.

<u>Wyniki analizy próbek surowicy:</u>

Próba 1:

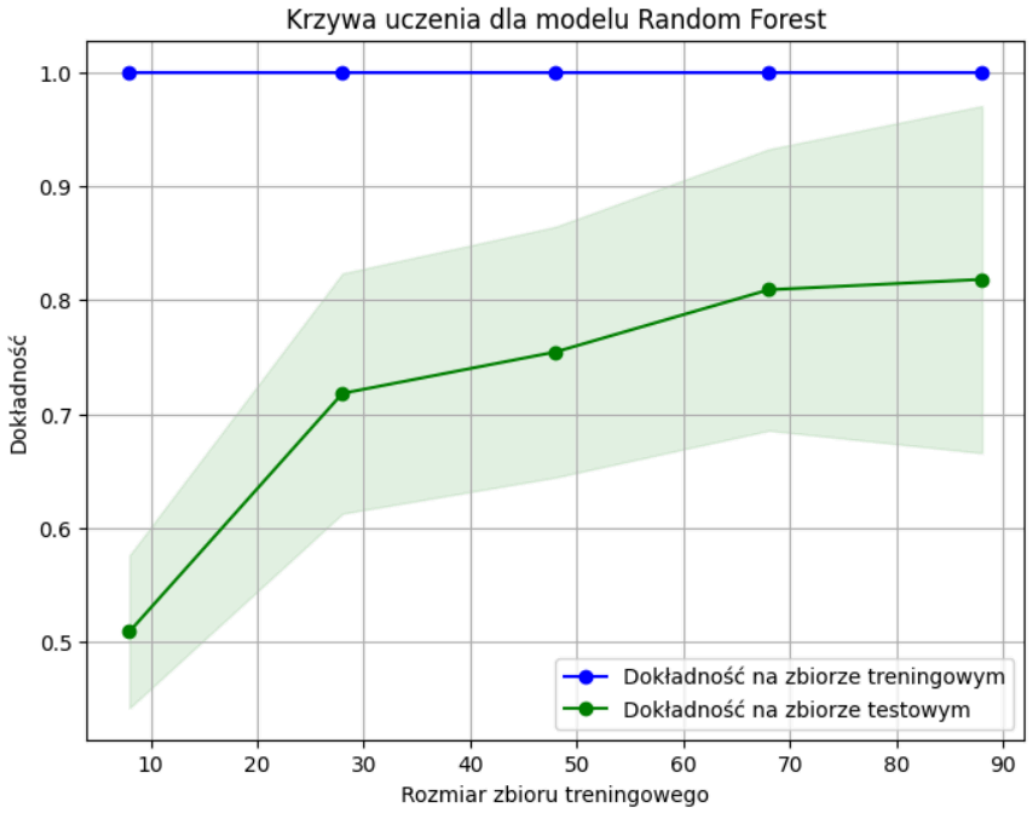
```
***Kroswalidacja***
Dokładności w poszczególnych foldach: [0.72727273 0.81818182 0.90909091 0.81818182 0.63636364 0.72727273
 0.90909091 1.         0.81818182 0.72727273]
Średnia dokładność: 80.91%
Odchylenie standardowe dokładności: 10.33%
***Wyniki jakości***
Dokładność (Accuracy): 78.92%
Precyzja (Precision): 0.79
Czułość (Recall): 0.79
F1-Score: 0.79
***Macierz pomyłek***
[[53 16]
 [19 78]]
***Raport klasyfikacji***
              precision    recall  f1-score   support

         0.0       0.74      0.77      0.75        69
         1.0       0.83      0.80      0.82        97

    accuracy                           0.79       166
   macro avg       0.78      0.79      0.78       166
weighted avg       0.79      0.79      0.79       166
```





Nieco niższe wyniki mogą wskazywać, że model Random Forest nie jest odpowiednio dobrany do klasyfikacji tych danych.

Krzywa uczenia:



Wykres ten wskazuje na niską dokładność nawet dla większych próbek zbioru treningowego. Model nie wydaje się być poprawnie dobrany do danych.

Próba 2, z wykorzystaniem modelu KNN (K-najbliższych sąsiadów):

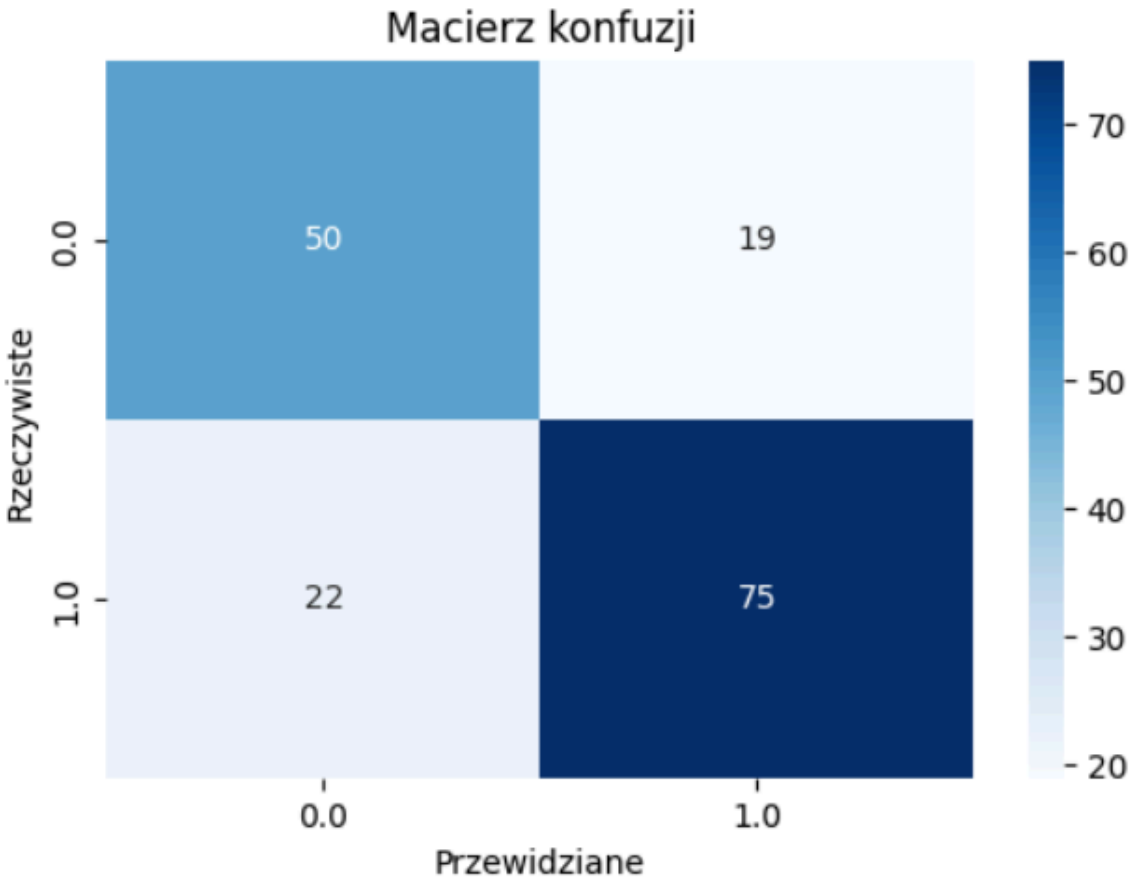
```
# Tworzenie modelu K-najbliższych sąsiadów
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(features_train, np.ravel(labels_train)) #Uczenie klasyfikatora na części treningowej
labels_predicted = knn.predict(features_test) #Generowania decyzji dla części testowej
```

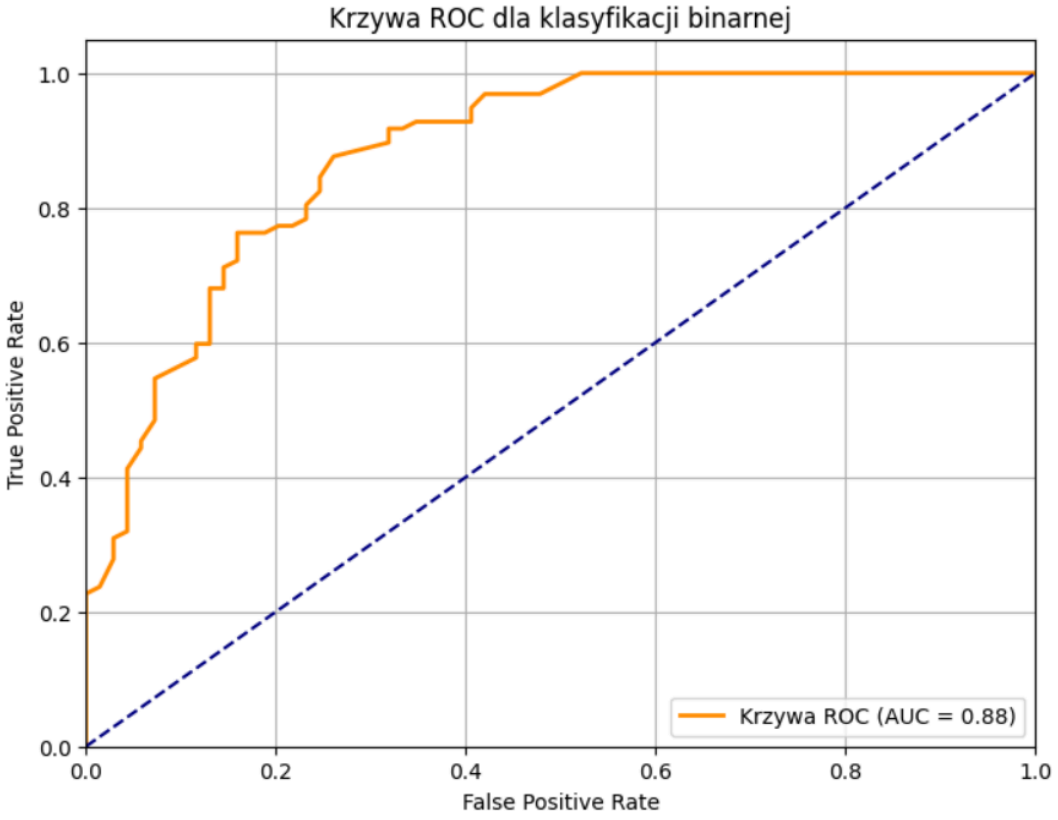
```
***Kroswalidacja***
Dokładności w poszczególnych foldach: [1.         1.         1.         0.84615385 1.         1.
 1.         1.         1.         1.        ]
Średnia dokładność: 98.46%
Odchylenie standardowe dokładności: 4.62%
***Wyniki jakości***
Dokładność (Accuracy): 100.00%
Precyzja (Precision): 1.00
Czułość (Recall): 1.00
F1-Score: 1.00
***Macierz pomyłek***
[[93  0]
 [ 0 97]]
***Raport klasyfikacji***
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        93
         1.0       1.00      1.00      1.00        97

    accuracy                           1.00       190
   macro avg       1.00      1.00      1.00       190
weighted avg       1.00      1.00      1.00       190
```

Model KNN uzyskał w tej próbie lepsze wyniki dla tego zbioru danych, niż model Random Forest.

Próba 2:

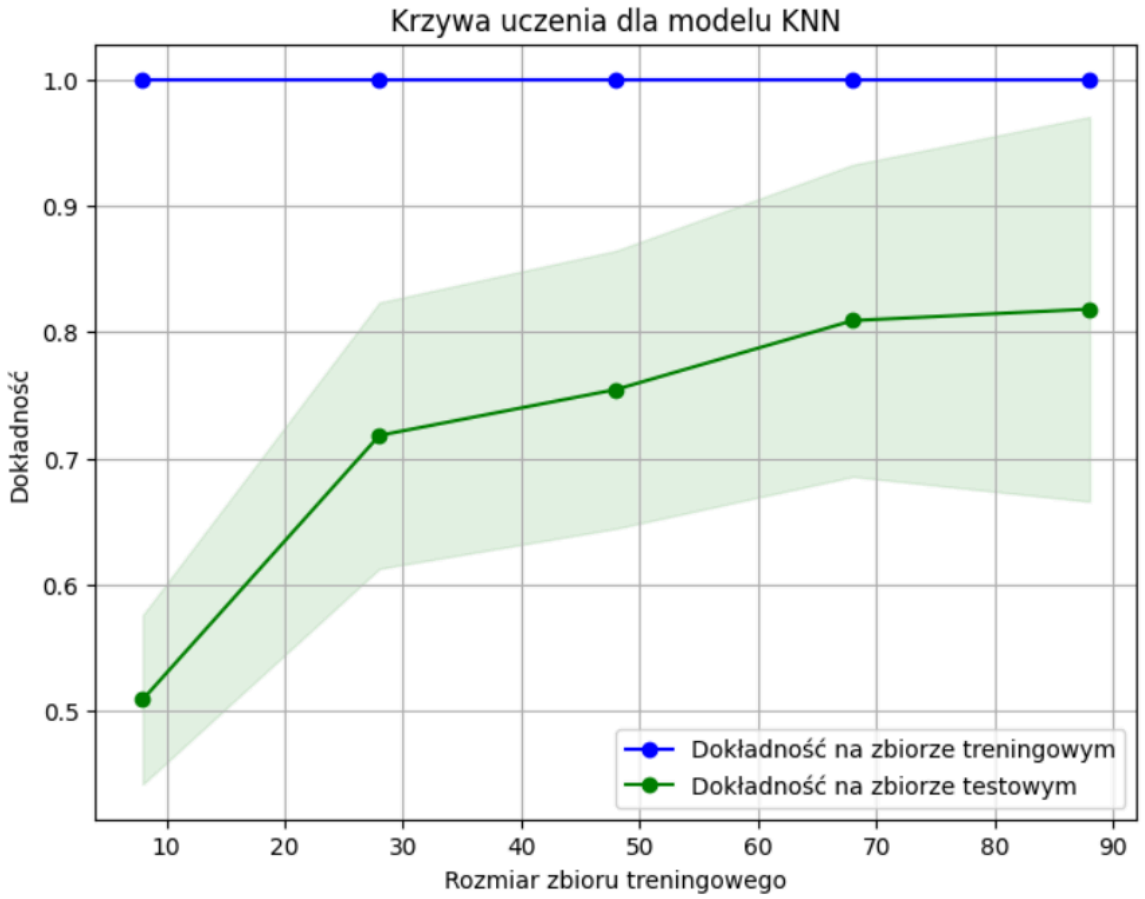
```
***Kroswalidacja***
Dokładności w poszczególnych foldach: [0.81818182 0.72727273 0.81818182 0.63636364 0.81818182 0.81818182
 0.81818182 0.81818182 0.45454545 0.72727273]
Średnia dokładność: 74.55%
Odchylenie standardowe dokładności: 11.35%
***Wyniki jakości***
Dokładność (Accuracy): 75.30%
Precyzja (Precision): 0.75
Czułość (Recall): 0.75
F1-Score: 0.75
***Macierz pomyłek***
[[50 19]
 [22 75]]
***Raport klasyfikacji***
              precision    recall  f1-score   support

         0.0       0.69      0.72      0.71        69
         1.0       0.80      0.77      0.79        97

    accuracy                           0.75       166
   macro avg       0.75      0.75      0.75       166
weighted avg       0.75      0.75      0.75       166
```





W tej próbie, wyniki były niższe niż przy wykorzystaniu modelu Random Forest.

Krzywa uczenia:



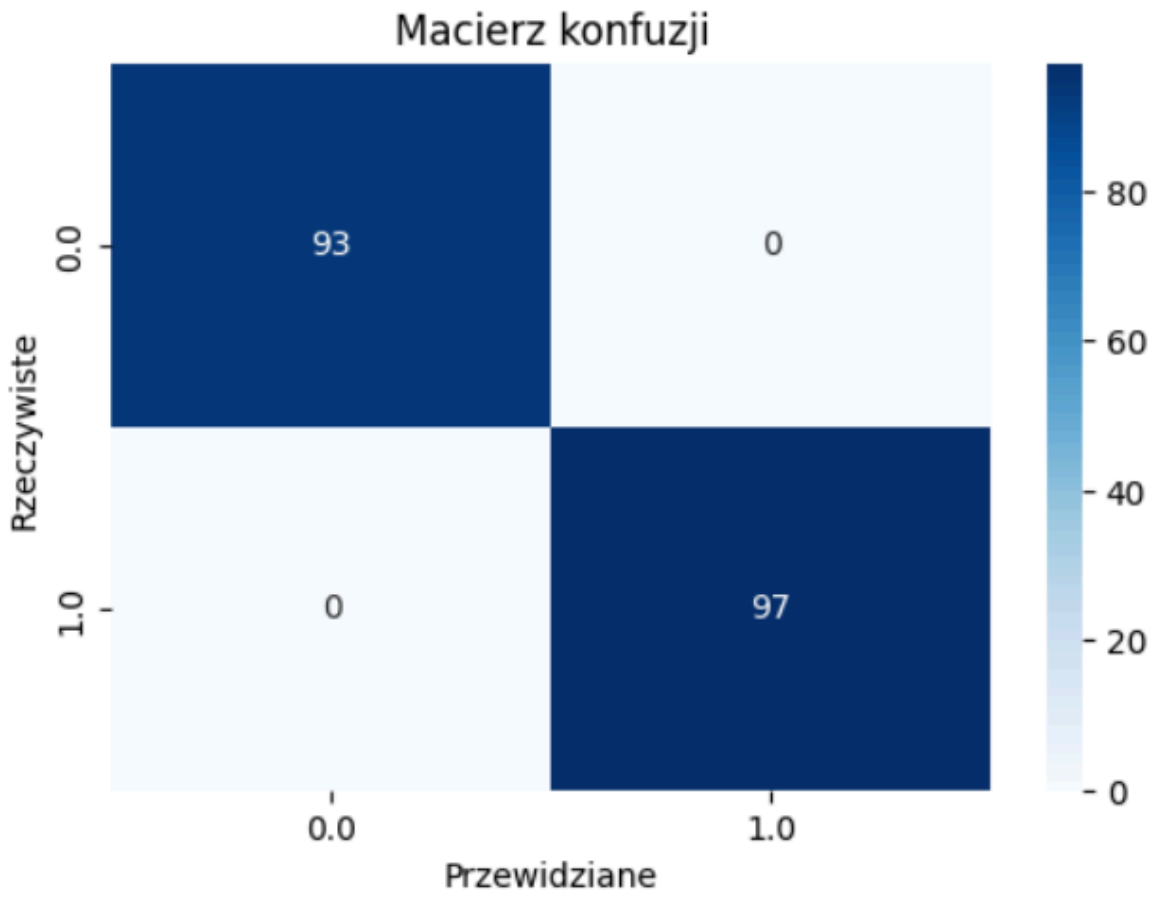
Wyniki analizy próbek tkanek:

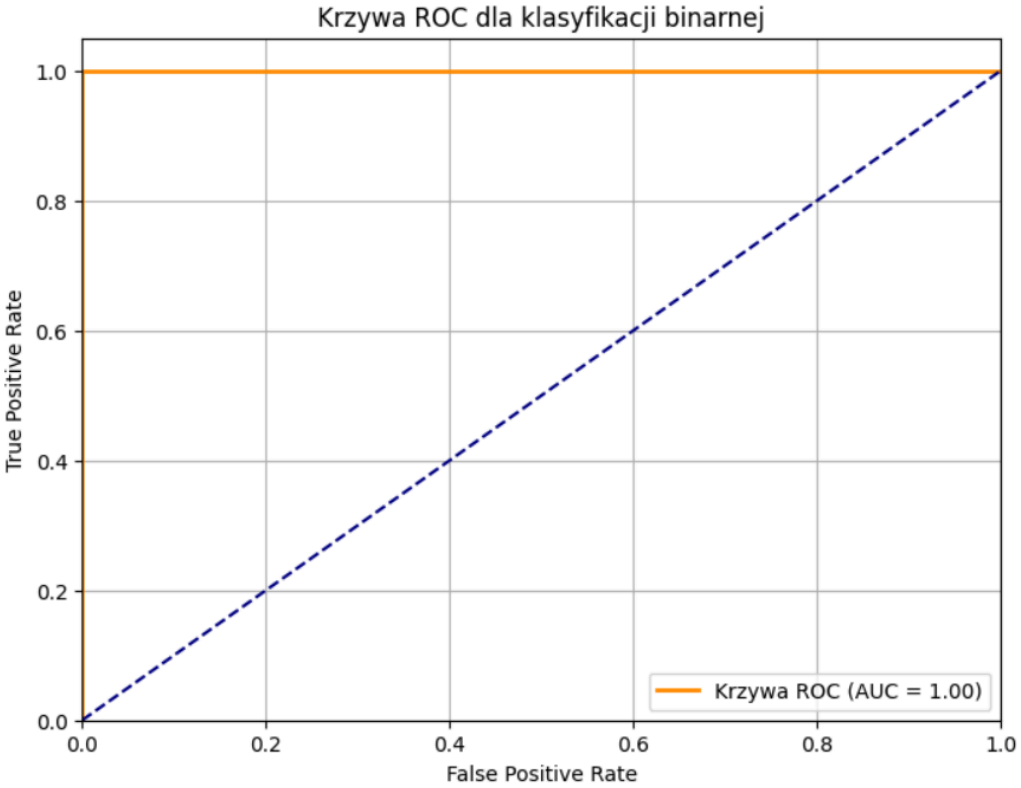
Próba 1:

```
***Kroswalidacja***
Dokładności w poszczególnych foldach: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
Średnia dokładność: 100.00%
Odchylenie standardowe dokładności: 0.00%
***Wyniki jakości***
Dokładność (Accuracy): 100.00%
Precyzja (Precision): 1.00
Czułość (Recall): 1.00
F1-Score: 1.00
***Macierz pomyłek***
[[93  0]
 [ 0 96]]
***Raport klasyfikacji***
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        93
         1.0       1.00      1.00      1.00        96

    accuracy                           1.00       189
   macro avg       1.00      1.00      1.00       189
weighted avg       1.00      1.00      1.00       189
```

Próba 2:

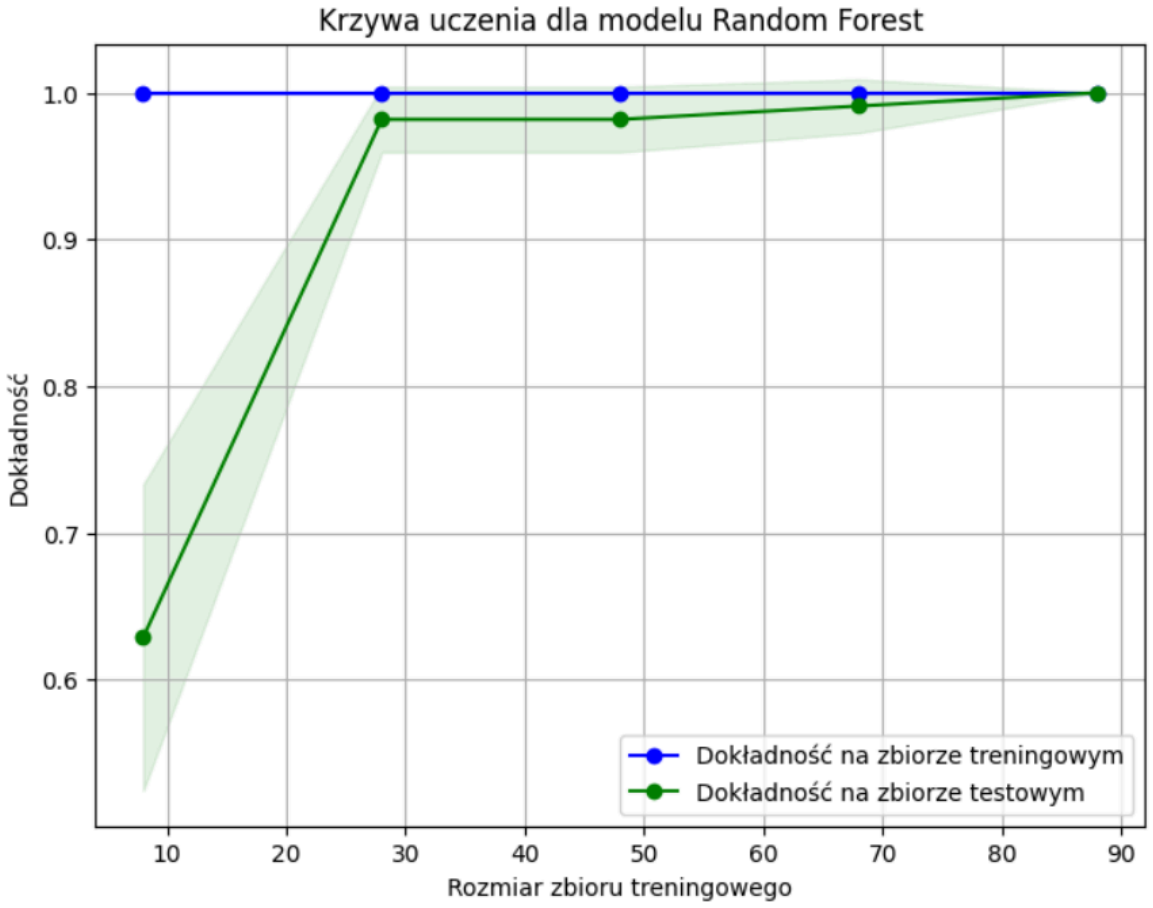
```
***Kroswalidacja***
Dokładności w poszczególnych foldach: [1.         1.         1.         0.92307692 1.         1.
 1.         1.         1.         1.        ]
Średnia dokładność: 99.23%
Odchylenie standardowe dokładności: 2.31%
***Wyniki jakości***
Dokładność (Accuracy): 100.00%
Precyzja (Precision): 1.00
Czułość (Recall): 1.00
F1-Score: 1.00
***Macierz pomyłek***
[[93  0]
 [ 0 97]]
***Raport klasyfikacji***
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        93
         1.0       1.00      1.00      1.00        97

    accuracy                           1.00       190
   macro avg       1.00      1.00      1.00       190
weighted avg       1.00      1.00      1.00       190
```





100% dokładność dopasowania modelu może wskazywać na overfitting. W takim przypadku model może nie działać poprawnie na nowych danych.

Krzywa uczenia:



Wykres ten wskazuje, że dla większych zbiorów treningowych, model osiąga wysoką dokładność, co pokazuje jego zdolność do generalizacji. Analiza tego wykresu wskazuje, że nie występuje tutaj problem z przeuczeniem.

6. Wnioski

Wysoka dokładność może wskazywać na overfitting, czyli nadmierne dopasowanie danych, co może skutkować niezdolnością modelu do prawidłowej klasyfikacji w przypadku nowych danych. Macierze pomyłek, oraz wykresy krzywej uczenia wskazują jednak, że model prawidłowo rozpoznaje klasy na danych testowych, oraz posiada zdolność do generalizacji.
Raport klasyfikacji wskazuje także, że obie klasy są dobrze rozpoznawane, więc nie występuje w tym przypadku raczej problem niezbalansowanych danych.