1. Strona tytułowa

Nazwa projektu: System klasyfikacji nowotworu piersi z wykorzystaniem regresji logistycznej i walidacji krzyżowej
Członkowie grupy:

- Emil Sadkowski
- Jakub Sokołowski
- Szymon Sobczak

Numer grupy: 7 grupa
Kierunek: Informatyka
Rok akademicki: 2025/2026

2. Wstęp teoretyczny

2.1 Omówienie problemu i jego znaczenie

Rak piersi jest jednym z najczęściej występujących nowotworów u kobiet na całym świecie. Wczesne i dokładne rozpoznanie ma kluczowe znaczenie dla skuteczności leczenia i przeżywalności pacjentek. Automatyzacja procesu klasyfikacji zmian nowotworowych na podstawie cech cytologicznych może znacząco wspomóc diagnostykę medyczną.

2.2 Podstawowe pojęcia

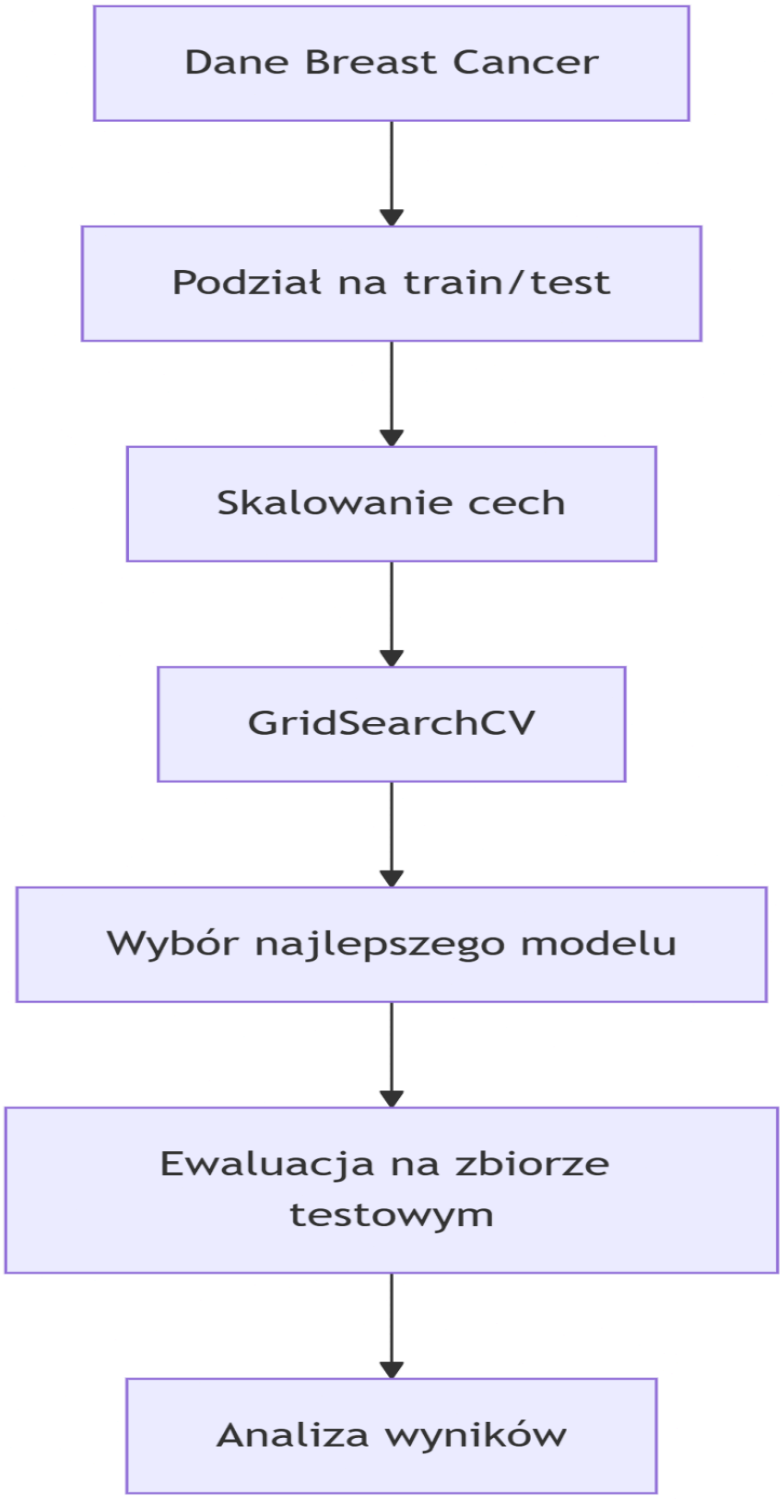
- Klasyfikacja binarna: Problem polegający na przypisaniu próbek do jednej z dwóch klas
- Regresja logistyczna: Algorytm klasyfikacji przewidujący prawdopodobieństwo przynależności do klasy
- Regularyzacja: Technika zapobiegająca przeuczeniu przez dodanie kary do funkcji kosztu
- Walidacja krzyżowa: Metoda oceny modelu przez podział danych na części treningowe i walidacyjne
- Metryki ewaluacji: Miary służące do oceny skuteczności modelu (accuracy, precision, recall, F1, AUC-ROC)

2.3 Przegląd istniejących podejść

- Metody tradycyjne: regresja logistyczna, SVM, drzewa decyzyjne
- Ensemble methods: random forest, gradient boosting
- Sieci neuronowe: MLP, CNN do analizy obrazów
- Hybrydowe systemy łączące multiple podejścia

3. Opis algorytmu / metody

3.1 Schemat działania



3.2 Opis kroków przetwarzania danych

1. **Wczytanie danych**: Import zbioru Iris z biblioteki scikit-learn
2. **Eksploracja**: Analiza statystyk, rozkładu klas, korelacji między cechami
3. **Wizualizacja**: Wykresy pudełkowe, pairplot, macierz korelacji
4. **Przygotowanie**: Podział na zbiory treningowy i testowy z zachowaniem proporcji klas
5. **Normalizacja**: Standaryzacja cech do średniej 0 i wariancji 1
6. **Trening**: Uczenie trzech różnych algorytmów klasyfikacji
7. **Tuning**: Optymalizacja hiperparametrów za pomocą Grid Search
8. **Ewaluacja**: Ocena modeli na zbiorze testowym przy użyciu multipleks metryk

3.3 Uzasadnienie wyboru algorytmów

Wybrano trzy reprezentatywne algorytmy o różnej charakterystyce:

- **KNN**:
  - Prostota implementacji i interpretacji
  - Brak założeń o rozkładzie danych
  - Wrażliwość na skalę danych (wymaga normalizacji)
- **SVM**:
  - Skuteczność w przestrzeniach o wysokiej wymiarowości
  - Odporność na overfitting przy odpowiedniej regularyzacji
  - Zdolność do modelowania nieliniowych granic decyzyjnych
- **Random Forest**:
  - Odporność na overfitting dzięki ensemble learning
  - Zdolność do pracy z danymi niezsynchronizowanymi
  - Dostarcza informacji o ważności cech

4. Użyte dane i przygotowanie danych

4.1 Źródło danych

Zbiór: Breast Cancer Wisconsin Diagnostic Dataset

- Źródło: scikit-learn datasets module
- Liczba próbek: 569
- Liczba cech: 30 (numeryczne)
- Klasy: 2 (0 = malignant - złośliwy, 1 = benign - łagodny)

4.2 Charakterystyka cech

Cechy opisują charakterystykę komórek nowotworowych, w tym:

- Promień (średnia odległość od środka do punktów na obwodzie)
- Tekstura (odchylenie standardowe wartości skali szarości)
- Obwód
- Powierzchnia
- Gładkość (lokalne zmiany długości promieni)
- Zwartość (obwód² / powierzchnia - 1.0)
- Wklęsłość
- Punkty wklęsłe
- Symetria
- Wymiar fraktalny

4.3 Przygotowanie danych

```
# Skalowanie StandardScaler
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

5. Implementacja

5.1 Architektura programu

Główny pipeline przetwarzania

1. Ładowanie i eksploracja danych

2. Podział na zbiór treningowy i testowy

3. Skalowanie cech

4. Optymalizacja hiperparametrów (GridSearchCV)

5. Ewaluacja najlepszego modelu

6. Wizualizacja wyników

5.2 Użyte biblioteki

- **scikit-learn**: Algorytmy ML, preprocessing, metryki
- **pandas**: Manipulacja i analiza danych
- **numpy**: Operacje numeryczne
- **matplotlib**: Podstawowe wizualizacje
- **seaborn**: Zaawansowane wizualizacje statystyczne

5.3 Kluczowe fragmenty kodu

```
# Definicja modelu i siatki hiperparametrów
model = LogisticRegression(solver='liblinear',
random_state=42)
param_grid = {
    'C': [0.01, 0.1, 1, 10, 100],  # Siła regularyzacji
    'penalty': ['l1', 'l2']       # Typ regularyzacji
}

# Grid Search z walidacją krzyżową
grid_search = GridSearchCV(
    estimator=model,
    param_grid=param_grid,
    cv=5,
    scoring='roc_auc'
)
```

6. Ewaluacja i wyniki

6.1 Metryki ewaluacji

- Accuracy: Dokładność klasyfikacji
- Precision: Precyzja - stosunek TP do wszystkich przewidzianych pozytywnych
- Recall (Czułość): Stosunek TP do wszystkich rzeczywistych pozytywnych
- F1-Score: Średnia harmoniczna precision i recall
- ROC-AUC: Pole pod krzywą ROC - miara zdolności rozróżniania klas

6.2 Wyniki optymalizacji hiperparametrów

Najlepsze parametry: {'C': 0.1, 'penalty': 'l2'}

Najlepszy wynik AUC (trening): 0.9941

6.3 Wyniki na zbiorze testowym

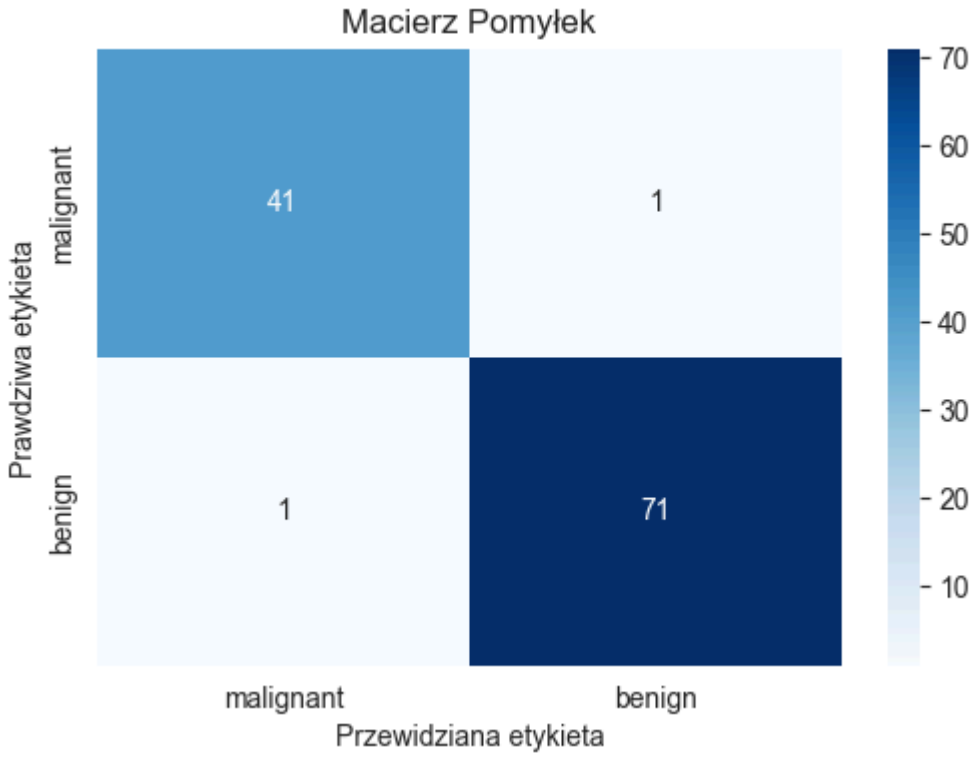
--- Ocena Modelu na Zbiorze Testowym ---
Accuracy: 0.9825
Precision: 0.9861
Recall (Czułość): 0.9861
F1-Score: 0.9861
ROC-AUC: 0.9960

6.4 Raport klasyfikacji

	precision	recall	f1-score	support
Malignant (0)	**0.98**	**0.98**	0.98	**42**
Benign (1)	**0.99**	**0.99**	**0.99**	**72**
accuracy			**0.98**	**114**
macro avg	**0.98**	**0.98**	**0.98**	**114**
weighted avg	**0.9**	**0.98**	**0.98**	**114**

6.5 Wizualizacje

Macierz pomyłek:

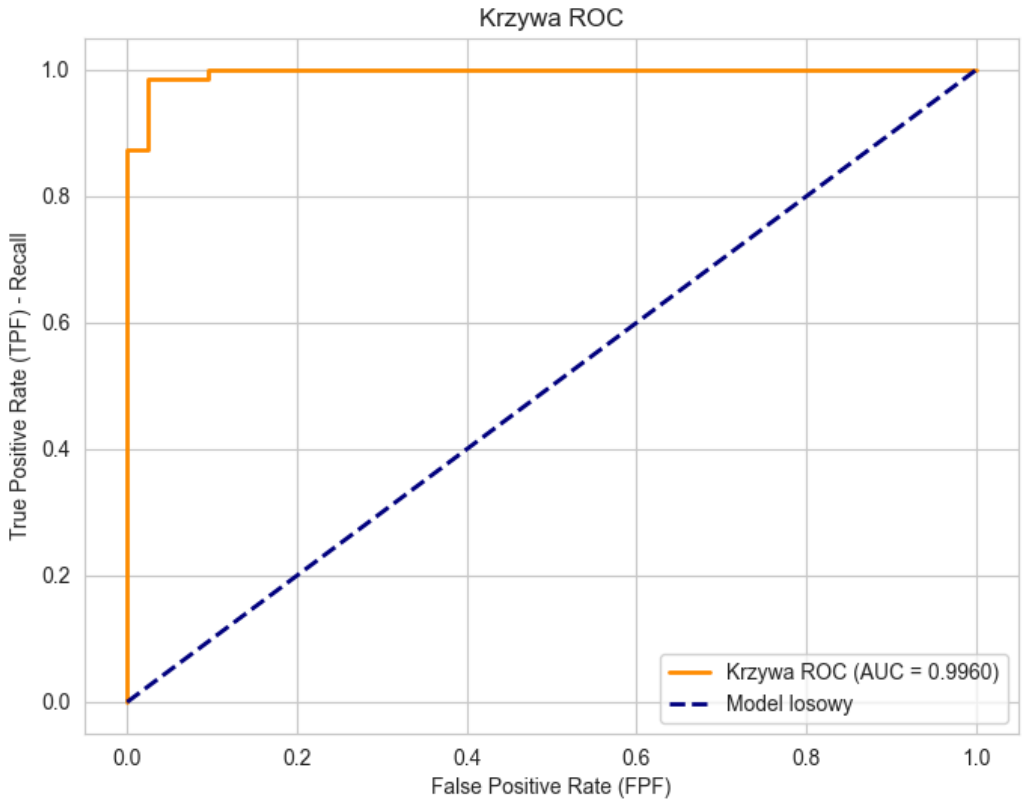


Na załączonym wykresie macierzy pomyłek widać następujące wartości:

- Prawdziwie Negatywne (TN): 41 (przypadki klasy "malignant" poprawnie sklasyfikowane)
- Fałszywie Pozytywne (FP): 1 (przypadki klasy "malignant" błędnie sklasyfikowane jako "benign")
- Fałszywie Negatywne (FN): 1 (przypadki klasy "benign" błędnie sklasyfikowane jako "malignant")
- Prawdziwie Pozytywne (TP): 71 (przypadki klasy "benign" poprawnie sklasyfikowane)

Macierz pomyłek została przedstawiona za pomocą heatmapy z biblioteki seaborn, gdzie wartości liczbowe są zaznaczone w odpowiednich komórkach, a kolory (odcienie niebieskiego) ilustrują wielkość wartości.

Krzywa ROC:



Krzywa ROC (Receiver Operating Characteristic) przedstawia stosunek czułości (TPR) do 1 - specyficzności (FPR) dla różnych progów klasyfikacji. Na załączonym wykresie krzywa ROC modelu (pomarańczowa) znajduje się bardzo blisko lewego górnego rogu, co wskazuje na doskonałą zdolność klasyfikacji. Pole pod krzywą (AUC) wyniosło 0.9960, co jest wartością bliską 1 i świadczy o niemal idealnym rozróżnieniu między klasami. Dla porównania, na wykresie znajduje się również linia przerywana reprezentująca model losowy (AUC = 0.5).

Oba wykresy potwierdzają bardzo wysoką skuteczność modelu.

7. Analiza wyników i wnioski

7.1 Interpretacja wyników

Model osiągnął doskonałe wyniki w klasyfikacji:

- Bardzo wysoka dokładność (98.25%): Tylko 2 błędy na 114 próbek
- Doskonała równowaga metryk (98.61%): Wszystkie metryki na podobnie wysokim poziomie
- Wyjątkowe AUC (0.9960): Model niemal doskonale rozróżnia obie klasy
- Zrównoważone błędy: Po jednym błędzie każdego typu (FP i FN)

7.2 Analiza błędów

Na 114 próbek testowych wystąpiły tylko 2 błędy:

- 1 fałszywie pozytywny (zdrowy zdiagnozowany jako chory)
- 1 fałszywie negatywny (chory zdiagnozowany jako zdrowy)

Znaczenie kliniczne: Równowaga między błędami typu I i II jest korzystna z punktu widzenia diagnostyki, choć w niektórych kontekstach klinicznych false negatives mogą być uważane za bardziej krytyczne.

7.3 Wpływ regularyzacji

Optymalne parametry to C=0.1 z regularyzacją L2, co wskazuje na:

- Nieco silniejszą regularyzację niż standardowa (C=1)
- Dobry kompromis między bias a variance
- Skuteczne zapobieganie przeuczeniu przy zachowaniu wysokiej skuteczności

7.4 Napotkane problemy i rozwiązania

1. Problem: Wybór solvera dla regularyzacji L1/L2
   Rozwiązanie: Użycie 'liblinear' obsługującego oba typy regularyzacji
2. Problem: Nierównowaga klas
   Rozwiązanie: Stratyfikowany podział danych zachowujący proporcje klas
3. Problem: Skalowanie cech
   Rozwiązanie: StandardScaler zapewniający porównywalność cech

7.5 Kierunki ulepszeń

- Analiza ważności cech: Identyfikacja najbardziej istotnych cech dla redukcji wymiarowości
- Eksperymenty z progami klasyfikacji: Dostosowanie progu dla lepszej czułości/specyficzności
- Ensemble methods: Porównanie z Random Forest lub Gradient Boosting
- Walidacja zewnętrzna: Testowanie na niezależnych zbiorach danych
- Interpretowalność: Analiza współczynników modelu dla zrozumienia wpływu poszczególnych cech

8. Podział pracy w grupie

Członek grupy	Zadania	Wkład (%)
Emil Sadkowski	Implementacja modelu, GridSearchCV, optymalizacja parametrów	40%
Szymon Sobczak	Przygotowanie danych, ewaluacja modelu, metryki	30%
Jakub Sokołowski	Wizualizacje, dokumentacja, analiza wyników	30%

Szczegółowy opis zadań:

Emil Sadkowski:

- Implementacja regresji logistycznej z regularyzacją
- Konfiguracja GridSearchCV do optymalizacji hiperparametrów
- Definicja siatki parametrów (C, penalty)
- Walidacja krzyżowa i wybór najlepszego modelu
- Integracja całego pipeline'u przetwarzania

Szymon Sobczak:

- Wczytanie i eksploracja danych Breast Cancer Wisconsin
- Podział danych na zbiór treningowy i testowy
- Skalowanie cech przy użyciu StandardScaler
- Obliczenie metryk ewaluacji (accuracy, precision, recall, F1, AUC-ROC)
- Generowanie raportu klasyfikacji

Jakub Sokołowski:

- Przygotowanie macierzy pomyłek z wykorzystaniem seaborn
- Generowanie krzywej ROC i obliczenie pola pod krzywą
- Wizualizacja wyników i analiza graficzna
- Przygotowanie dokumentacji projektu
- Formułowanie wniosków i analiza końcowa

9. Bibliografia i źródła

1. Street, W.N., Wolberg, W.H., & Mangasarian, O.L. (1993). Nuclear feature extraction for breast tumor diagnosis.
2. scikit-learn documentation: Logistic Regression
3. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python
4. Breast Cancer Wisconsin (Diagnostic) Data Set: UCI Machine Learning Repository

Źródła kodu:

- Dokumentacja scikit-learn: **https://scikit-learn.org/stable/**
- **Przykłady klasyfikacji: scikit-learn examples**
- **Materiały do wizualizacji: matplotlib and seaborn documentation**

Repozytorium kodu

https://github.com/sadkovvsky/psi/tree/main/Zad2