Zdefiniowanie problemu predykcyjnego oraz scharakteryzowanie analizowanego zbioru danych

Mateusz Mazur

Jakub Sakowski

Aleksandra Wójcik

1. Definicja problemu predykcyjnego

Celem projektu jest zbudowanie modelu klasyfikacyjnego, który pozwoli przewidzieć ryzyko śmierci pacjenta z niewydolnością serca na podstawie dostępnych danych klinicznych. Zmienną objaśnianą (zmienną celu) jest DEATH_EVENT, która przyjmuje wartość:

- `1` – jeśli pacjent zmarł w trakcie okresu obserwacji,
- `0` – jeśli pacjent przeżył.

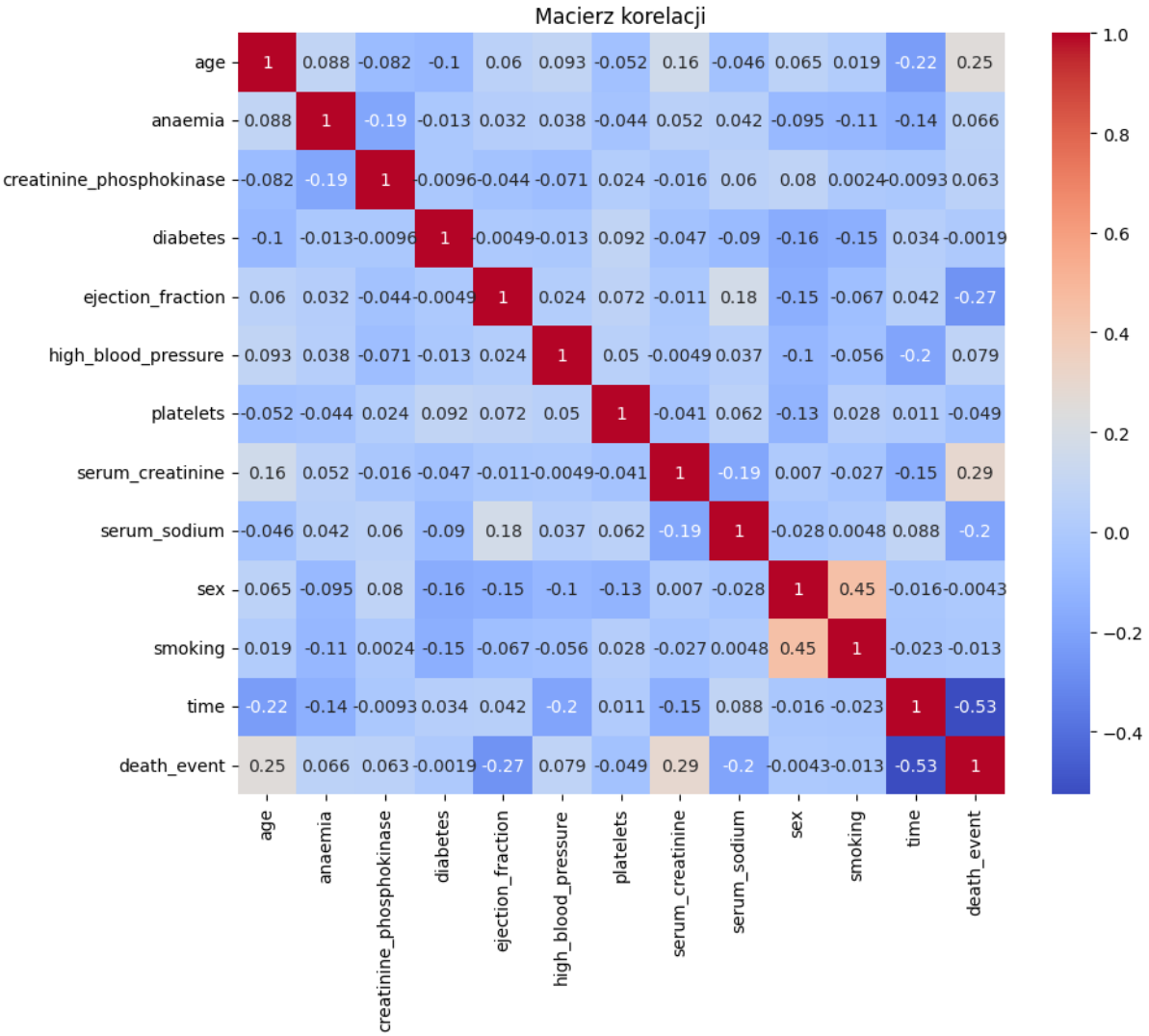
Zadanie to jest klasyfikacją binarną, w której model na podstawie danych wejściowych (cech medycznych i demograficznych pacjenta) ma określić prawdopodobieństwo wystąpienia zdarzenia śmiertelnego. Praktyczne zastosowanie takiego modelu może wspierać lekarzy w ocenie ryzyka i podejmowaniu decyzji klinicznych.

2. Charakterystyka zbioru danych

Analizowany zbiór danych pochodzi z repozytorium **UCI Machine Learning Repository** (ID: 519) i nosi nazwę **Heart Failure Clinical Records**. Zbiór został opracowany na podstawie danych z badań pacjentów z objawami przewlekłej niewydolności serca. Dane były pierwotnie używane w publikacji naukowej: *"Anaemia and Renal Dysfunction in Heart Failure"*.

2.1 ***Podstawowe informacje:***

- **Liczba rekordów (pacjentów):** 299
- **Liczba cech (atrybutów):** 13 (w tym 12 wejściowych + 1 docelowa)
- **Typ zadania:** Klasyfikacja binarna (DEATH_EVENT)
- **Brakujące dane:** brak



Rysunek 1: Macierz korelacji pomiędzy zmiennymi zbioru danych

2.2 ***Rodzaje cech:***

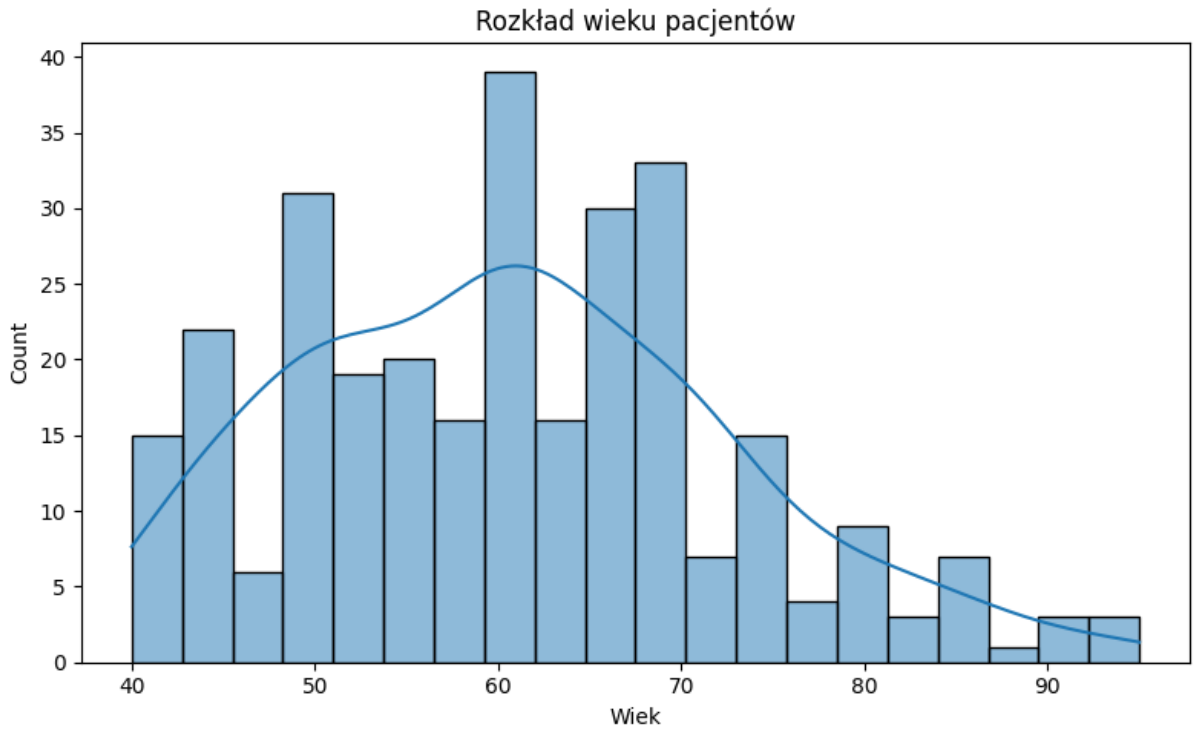
Rodzaj danych	Przykładowe cechy
Demograficzne	`age`, `sex`
Kliniczne	`anaemia`, `diabetes`, `high_blood_pressure`, `smoking`
Laboratoryjne	`serum_creatinine`, `serum_sodium`, `creatinine_phosphokinase`
Funkcjonalne	`ejection_fraction`, `platelets`
Obserwacyjne	`time` – liczba dni obserwacji pacjenta

2.3 ***Zmienna docelowa:***

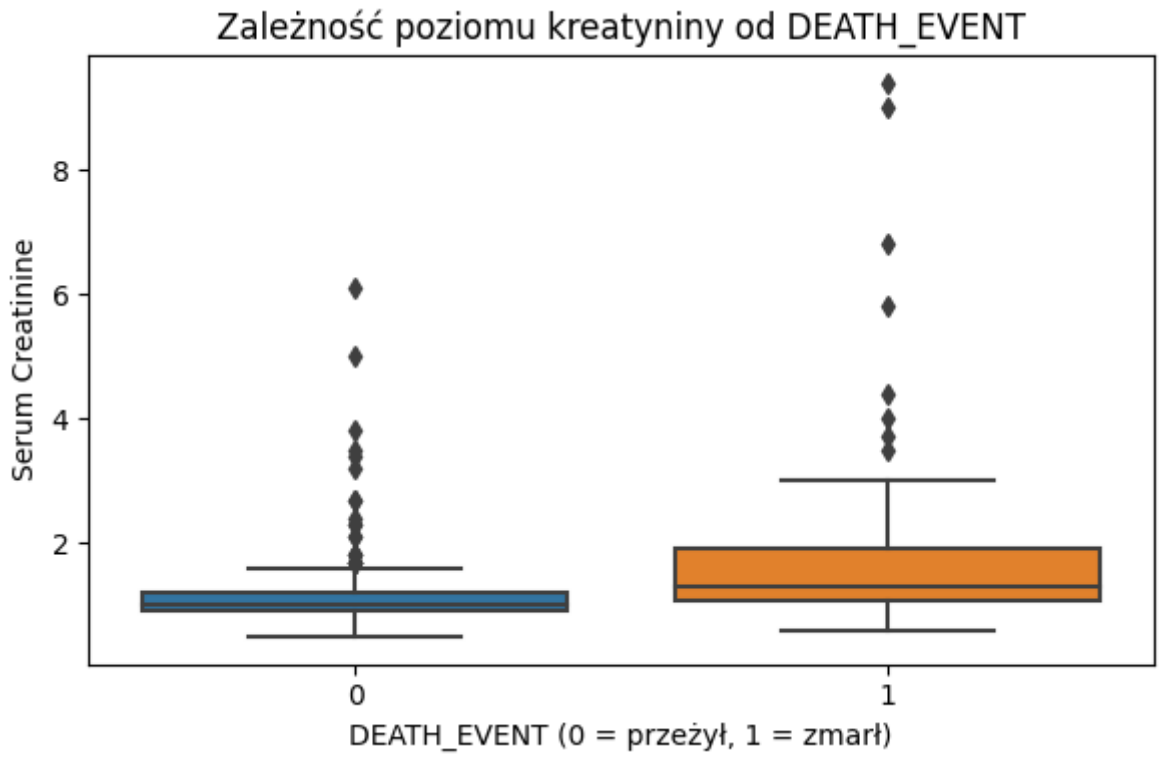
- **DEATH_EVENT** – informacja, czy pacjent zmarł podczas obserwacji (0 = przeżył, 1 = zmarł).

2.4 ***Uwagi:***

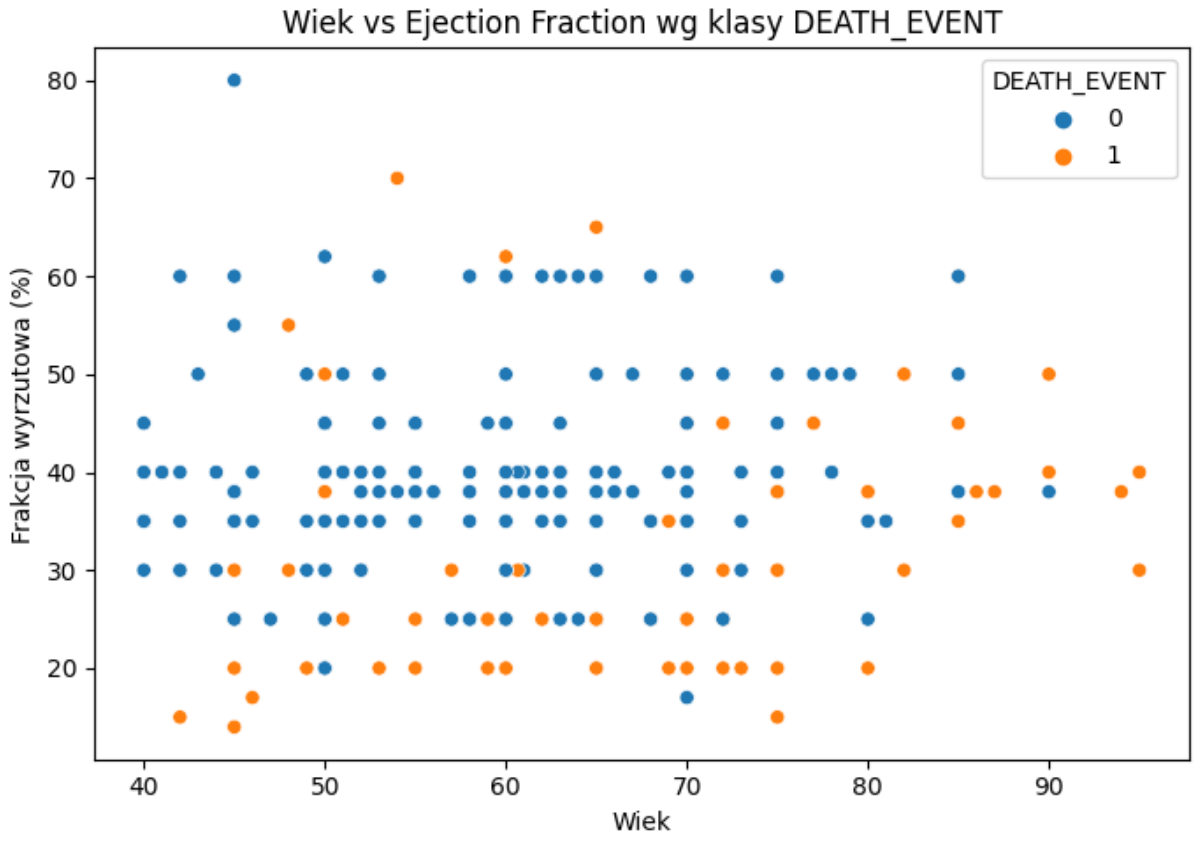
- Dane są stosunkowo zrównoważone, choć występuje pewna przewaga jednej z klas.
- Cechy są mieszane: binarne (np. `anaemia`), ciągłe (`serum_creatinine`), dyskretne (`ejection_fraction`) – co umożliwia zastosowanie szerokiego wachlarza technik eksploracyjnych i modelujących.



Rysunek 2: Rozkład wieku pacjentów. Oś pozioma przedstawia przedziały wiekowe (40-90 lat), a *oś pionowa - liczbę pacjentów w każdej grupie wiekowej*

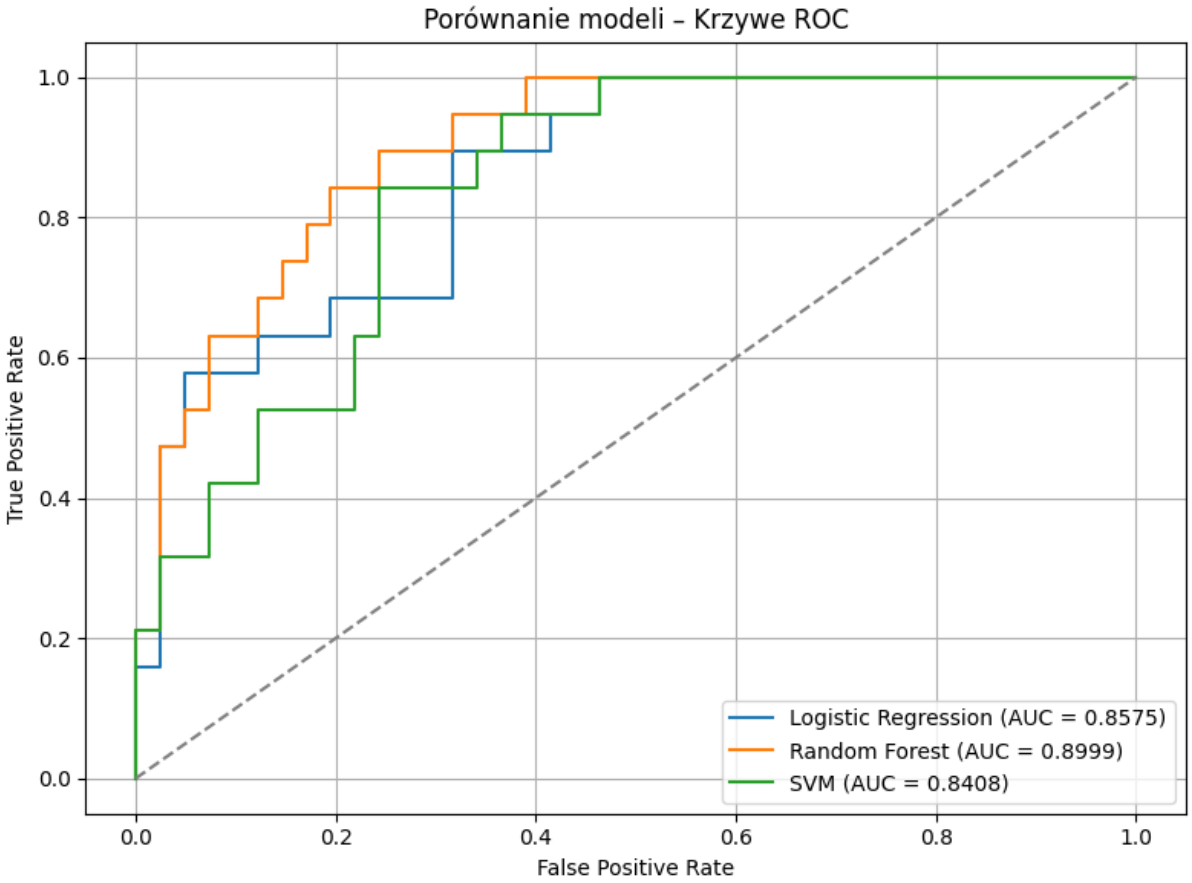


Rysunek 3: Rozkład zależności średniego poziomu kreatyniny w surowicy od zdarzenia śmierci



Rysunek 4: Zależność wieku od frakcji wyrzutowej (Ejection Fraction) w grupach pacjentów wg. Statusu zdarzenia śmierci.

3. Porównanie skuteczności modeli

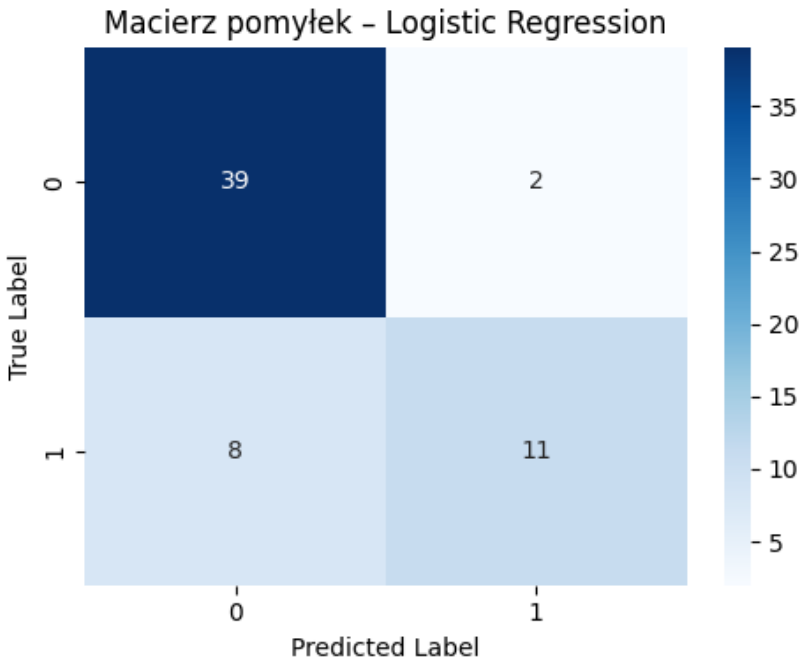


Rysunek 5: Krzywe ROC dla modeli regresji logistycznej, lasu losowego i SVM.

Logistic Regression

Metryka	***Class 0***	***Class 1***	***Accuracy***	***Macro Avg***	***Weighted Avg***
Precision	*0.83*	*0.85*	-	*0.84*	*0.83*
Recall	*0.95*	*0.58*	-	*0.77*	*0.83*
F1-score	*0.89*	*0.69*	*0.83*	*0.79*	*0.82*
Support	*41*	*19*	*60*	*60*	*60*

ROC AUC: 0.8575

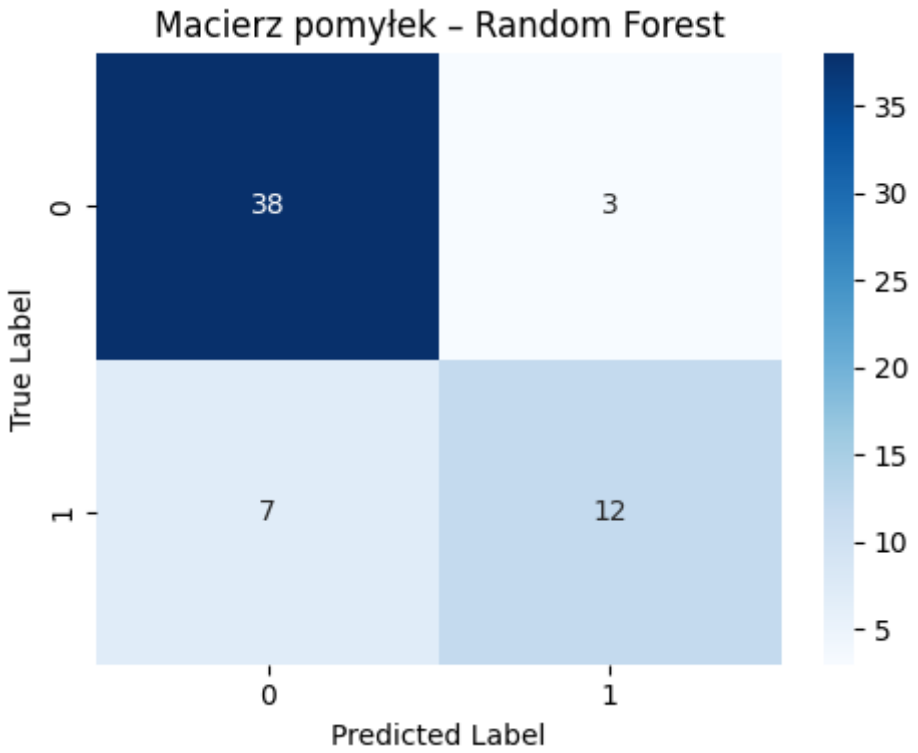


Rysunek 6: Macierz pomyłek (Confusion Matrix) dla modelu regresji logistycznej

Random Forest

Metryka	***Class 0***	***Class 1***	***Accuracy***	***Macro Avg***	***Weighted Avg***
Precision	*0.84*	*0.80*	-	*0.82*	*0.83*
Recall	*0.93*	*0.63*	-	*0.78*	*0.83*
F1-score	*0.88*	*0.71*	*0.83*	*0.79*	*0.83*
Support	*41*	*19*	*60*	*60*	*60*

ROC AUC: 0.8999

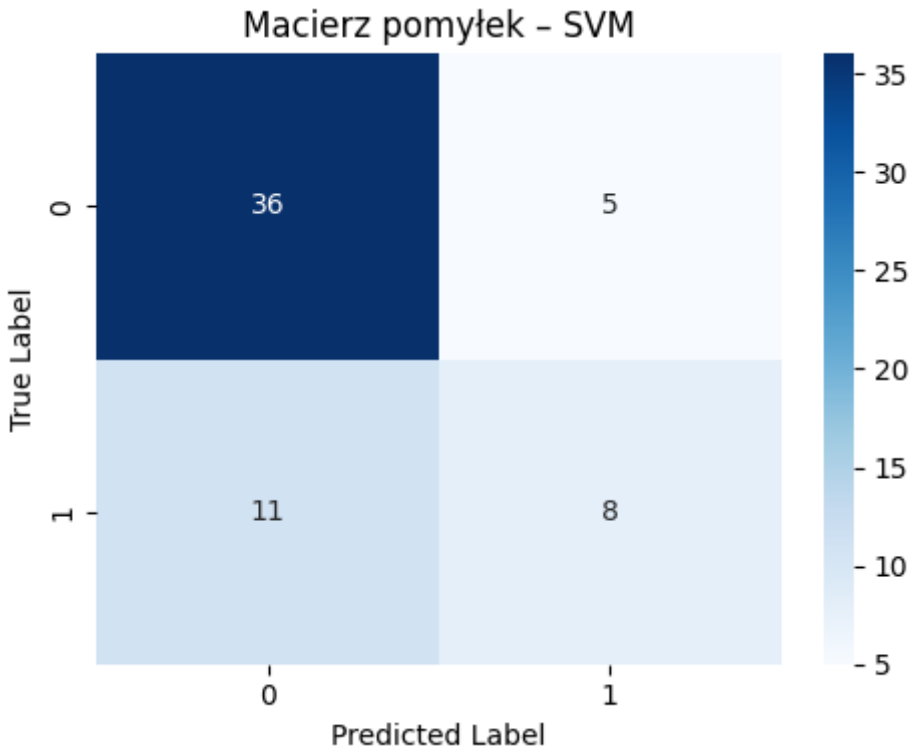


Rysunek 7: Macierz pomyłek (Confusion Matrix) dla modelu Random Forest.

SVM

Metryka	***Class 0***	***Class 1***	***Accuracy***	***Macro Avg***	***Weighted Avg***
Precision	*0.77*	*0.62*	*-*	*0.69*	*0.72*
Recall	*0.88*	*0.42*	*-*	*0.65*	*0.73*
F1-score	*0.82*	*0.50*	*0.73*	*0.66*	*0.72*
Support	*41*	*19*	*60*	*60*	*60*

ROC AUC: 0.8408



Rysunek 8: Macierz pomyłek (Confusion Matrix) dla modelu SVM.

4. **Podsumowanie**

Projekt klasyfikacji binarnej skupił się na przewidywaniu zdarzeń śmierci (DEATH_EVENT) w oparciu o dane medyczne, wykorzystując trzy różne modele uczenia maszynowego. Regresja logistyczna osiągnęła dokładność (accuracy) na poziomie 83%, z wartością ROC AUC wynoszącą 0.8575. Model ten wykazał wysoką precyzję (0.85) i czułość (0.95) dla klasy większościowej (przeżyli), ale znacznie niższą czułość (0.58) dla klasy mniejszościowej (zmarli), co przełożyło się na F1-score wynoszący odpowiednio 0.89 i 0.69.

Lasy losowe (Random Forest) wypadły nieco lepiej, osiągając dokładność 83% i najwyższą wartość ROC AUC (0.8999) spośród wszystkich testowanych modeli. Precyzja dla klasy zmarłych wyniosła 0.80, a czułość 0.63, co dało F1-score na poziomie 0.71. Dla klasy przeżyłych wyniki były zbliżone do regresji logistycznej, z precyzją 0.84 i czułością 0.93.

Model SVM okazał się najmniej skuteczny, z dokładnością 73% i ROC AUC równym 0.8408. Wykazał się szczególnie niską czułością (0.42) i F1-score (0.50) dla klasy zmarłych, co wskazuje na trudności w prawidłowej klasyfikacji tej grupy. Dla klasy przeżyłych precyzja wyniosła 0.77, a czułość 0.88, co przełożyło się na F1-score równy 0.82.

Wyniki pokazują, że lasy losowe są najlepszym wyborem pod względem ogólnej wydajności i zdolności do rozróżniania klas, choć wciąż wymagają poprawy w wykrywaniu przypadków zgonów. W dalszych etapach projektu warto skupić się na optymalizacji modeli, zwiększeniu

czułości dla klasy mniejszościowej oraz dokładniejszej analizie wpływu poszczególnych cech na prognozy, np. z wykorzystaniem metod interpretowalności AI, takich jak SHAP. Dodatkowo, zastosowanie bardziej zaawansowanych algorytmów, takich jak XGBoost, lub technik przetwarzania danych nierównowagowych może przynieść dalszą poprawę wyników.