

POLITECHNIKA ŚWIĘTOKRZYSKA

Wydział Elektrotechniki, Automatyki i Informatyki

Maksymilian Sowula, Paweł Marek

Numer grupy dziekańskiej: 1ID21A

**Selekcja cech i klasyfikacja danych numerycznych z
zastosowaniem wybranego klasyfikatora**

Analiza i Wizualizacja Danych - Projekt

1. Opis danych

Poniższy rozdział zawiera opis danych zawartych w zbiorach użytych do realizacji projektu – „Wine Quality” oraz „Heart Disease” czyli zbiorach dotyczących jakości wina oraz występowania chorób serca.

1.1. Zbiór „Wine Quality”

Zbiór danych „Wine Quality” jest to zbiór zawierający próbki danych dotyczących właściwości chemicznych czerwonego i białego wina wytwarzanego w północnej części Portugalii. Zawiera on 6497 próbek, które opisuje 11 atrybutów wejściowych oraz jedna cecha wyjściowa. Cechy wejściowe reprezentują wybrane parametry fizykochemiczne wina takie jak zawartość siarczanów czy kwasowości, a atrybut wyjściowy określa ocenę jakości wina w skali od 0 do 10. Poniższa tabela przedstawia zestawienie wszystkich atrybutów z opisem.

Tabela 1.1. Opis atrybutów zbioru „Wine Quality” [1].

Oryginalna nazwa atrybutu	Przetłumaczona nazwa atrybutu	Rola	Typ	Opis	Czy brakuje danych?
fixed_acidity	Kwasowość stała	Cecha wejściowa	Numeryczny	-	Nie
volatile_acidity	Kwasowość lotna	Cecha wejściowa	Numeryczny	-	Nie
citric_acid	Kwas cytrynowy	Cecha wejściowa	Numeryczny	-	Nie
residual_sugar	Cukier resztkowy	Cecha wejściowa	Numeryczny	-	Nie
chlorides	Chlorki	Cecha wejściowa	Numeryczny	-	Nie
free_sulfur_dioxide	Wolny SO ₂	Cecha wejściowa	Numeryczny	-	Nie
total_sulfur_dioxide	Całkowity SO ₂	Cecha wejściowa	Numeryczny	-	Nie
density	Gęstość	Cecha wejściowa	Numeryczny	-	Nie

pH	pH	Cecha wejściowa	Numeryczny	-	Nie
sulphates	Siarczany	Cecha wejściowa	Numeryczny	-	Nie
alcohol	Alkohol	Cecha wejściowa	Numeryczny	-	Nie
quality	Jakość	Cecha wyjściowa	Numeryczny	Skala od 0 do 10	Nie
color	Kolor wina	Inne	Tekstowy	-	Nie

W poniższej tabeli zaprezentowano pięć wybranych rekordów ze zbioru „Wine Quality” pokazujące wartości poszczególnych atrybutów wejściowych i wyjściowego.

Tabela 1.2. Pięć przykładowych rekordów ze zbioru „Wine Quality”.

-	Pierwszy rekord	Drugi rekord	Trzeci rekord	Czwarty rekord	Piąty rekord
Kwasowość stała	6.300	5.900	6.000	7.000	6.300
Kwasowość lotna	0.510	0.645	0.310	0.270	0.300
Kwas cytrynowy	0.130	0.120	0.470	0.360	0.340
Cukier resztkowy	2.300	2.000	3.600	20.700	1.600
Chlorki	0.076	0.075	0.067	0.045	0.049
Wolny SO2	29.000	32.000	18.000	45.000	14.000
Calkowity SO2	40.000	44.000	42.000	170.000	132.000
Gęstość	0.99574	0.99547	0.99549	1.001	0.994
pH	3.420	3.570	3.390	3.000	3.300
Siarczany	0.750	0.710	0.660	0.450	0.490
Alkohol	11.000	10.200	11.000	8.800	9.500
Jakość	6	5	6	6	6
Kolor wina	Red	Red	Red	White	White

1.2. Zbiór „Heart Disease”

Zbiór danych „Heart Disease” jest to zbiór zawierający próbki danych klinicznych oraz pomiarów fizjologicznych pacjentów, które zostały zebrane w Klinice Cleveland oraz w kilku innych ośrodkach medycznych. Zbiór ten jest szeroko stosowany w badaniach naukowych i edukacyjnych do przewidywania schorzeń kardiologicznych.

Zawiera on 297 próbek (pacjentów), które opisuje 13 atrybutów wejściowych oraz jedna cecha wyjściowa. Cechy wejściowe reprezentują wybrane parametry kliniczne pacjenta, takie jak: wiek, ciśnienie krwi, poziom cholesterolu, wyniki badań EKG oraz obecność różnych typów dławicy piersiowej. Atrybut wyjściowy (w tym pliku kolumna condition) określa diagnozę choroby serca i przyjmuje dwie wartości: 0 (brak choroby) lub 1 (obecność choroby).

Tabela 1.3. Opis atrybutów zbioru „Heart Disease” [2].

Oryginalna nazwa atrybutu	Przetłumaczona nazwa atrybutu	Rola	Typ	Opis	Czy brakuje danych?
age	Wiek	Cecha (Feature)	Numeryczny	Wiek pacjenta w latach.	Nie
sex	Płeć	Cecha (Feature)	Binarny	1 = Mężczyzna, 0 = Kobieta.	Nie
cp	Typ bólu w klatce piersiowej	Cecha (Feature)	Kategoryczny	Rodzaj bólu klatki piersiowej: 0: typowa dławica piersiowa, 1: nietypowa dławica piersiowa, 2: ból nieanginalny, 3: bezobjawowy.	Nie
trestbps	Spoczynkowe ciśnienie krwi	Cecha (Feature)	Liczba ciągła	Ciśnienie krwi w spoczynku (mm Hg przy przyjęciu do szpitala).	Nie
chol	Cholesterol w surowicy	Cecha (Feature)	Liczba ciągła	Poziom cholesterolu w surowicy (mg/dl).	Nie

fbs	Poziom cukru na czczo	Cecha (Feature)	Binarny	Cukier we krwi na czczo $> 120 \text{ mg/dl}$. 1 = prawda, 0 = fałsz.	Nie
restecg	Wyniki EKG spoczynkowego	Cecha (Feature)	Kategoryczny	Wyniki badania elektrokardiograficznego w spoczynku: 0: prawidłowy, 1: nieprawidłowości ST-T, 2: prawdopodobny lub pewny przerost lewej komory.	Nie
thalach	Maksymalne tętno	Cecha (Feature)	Liczba ciągła	Maksymalne tętno osiągnięte podczas testu wysiłkowego.	Nie
exang	Dławica piersiowa indukowana wysiłkiem	Cecha (Feature)	Binarny	Czy wystąpiła dławica piersiowa wywołana wysiłkiem. 1 = tak, 0 = nie.	Nie
oldpeak	Depresja ST	Cecha (Feature)	Liczba ciągła	Obniżenie odcinka ST wywołane wysiłkiem w stosunku do spoczynku.	Nie
slope	Nachylenie odcinka ST	Cecha (Feature)	Kategoryczny	Nachylenie szczytowego odcinka ST podczas wysiłku: 0: nachylenie w górę, 1: płaskie, 2: nachylenie w dół.	Nie

ca	Liczba głównych naczyń	Cecha (Feature)	Kategoryczny	Liczba dużych naczyń (0–3) pokolorowanych fluoroskopią.	Nie
thal	Talasemia	Cecha (Feature)	Kategoryczny	Wynik testu talasemii: 0: (nieznane, brak), 1: wada stała, 2: prawidłowy, 3: wada odwracalna. (Uwaga: w tym pliku wartości to 0, 1, 2)	Nie
condition	Stan/Diagnoza	Zmienna docelowa (Target)	Binarny	Obecność choroby serca (wynik binarnej klasyfikacji): 0 = brak choroby, 1 = obecność choroby.	Nie

W poniższej tabeli zaprezentowano pięć wybranych rekordów ze zbioru „Heart Disease” pokazujące wartości poszczególnych atrybutów wejściowych i wyjściowego.

Tabela 1.4. Pięć przykładowych rekordów ze zbioru „Heart Disease”.

-	Pierwszy rekord	Drugi rekord	Trzeci rekord	Czwarty rekord	Piąty rekord
age	69	69	66	65	64
sex	1	0	0	1	1
cp	0	0	0	0	0
trestbps	160	140	150	138	110
chol	234	239	226	282	211
fbs	1	0	0	1	0
restecg	2	0	0	2	2
thalach	131	151	114	174	144

exang	0	0	0	0	1
oldpeak	0.1	1.8	2.6	1.4	1.8
slope	1	0	2	1	1
ca	1	2	0	1	0
thal	0	0	0	0	0
condition	0	0	0	1	0

2. Opis realizowanego projektu

Poniższy rozdział zawiera opis algorytmów sztucznej inteligencji użytych w projekcie, metryk wybranych do oceny modeli oraz opis użytych bibliotek.

2.1. Opis algorytmów sztucznej inteligencji

2.1.1. Random Forest

Algorytm Random Forest jest zespołową metodą uczenia maszynowego, której celem jest poprawa jakości predykcji poprzez agregacje wyników generowanych przez wiele modeli bazowych, którymi są drzewa decyzyjne. Algorytm można opisać następującymi krokami:

- losowanie próbek – z oryginalnego zbioru uczącego D o liczności n generowanych jest B zbiorów danych D_1, D_2, \dots, D_B poprzez losowanie ze zwracaniem. Każdy z podzbiorów służy do niezależnej budowy jednego drzewa decyzyjnego,
- losowanie podprzestrzeni cech – podczas konstruowania drzewa na każdym węźle analizowana jest losowo wybrana podprzestrzeń cech o rozmiarze m gdzie $m < p$, a p to całkowita liczba atrybutów. Ze zbioru tych cech wybierana jest ta, która maksymalizuje kryterium jakości podziału,
- budowa zbioru drzew decyzyjnych – każde drzewo budowane jest niezależnie i bez przycinania,
- agregacja predykcji – w zadaniach klasyfikacji wynik modelu zespołowego określany jest na podstawie głosowania większościowego:

$$\hat{y} = mode(T_1(x), T_2(x), \dots, T_B(x))$$

gdzie:

- \hat{y} - przewidywany wynik końcowy modelu,
- $T_1(x), T_2(x), \dots, T_B(x)$ - wyniki poszczególnych drzew decyzyjnych dla próbki x ,
- B – liczba drzew w lesie,
- $\text{mode}(z)$ – najczęściej występująca wartość w lesie.

Zaś w zadaniach regresyjnych predykcja stanowi średnią arytmetyczną wartości przewidzianych przez wszystkie drzewa, opisuje to wzór:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x)$$

gdzie:

- \hat{y} - przewidywany wynik końcowy modelu,
- B – liczba drzew w lesie,
- $T_b(x)$ – predykcja drzewa o indeksie b dla próbki x .

2.1.2. XGBoost

Algorytm XGBoost (ang. eXtreme Gradient Boosting) jest zaawansowaną metodą zespołową opartą na technice wzmacnienia gradientowego (gradient boosting). W przeciwieństwie do algorytmu Random Forest, gdzie drzewa budowane są niezależnie, w XGBoost model konstruowany jest w sposób sekwencyjny i addytywny, dając do minimalizacji zdefiniowanej funkcji celu. Algorytm można opisać następującymi krokami:

- sekwencyjne uczenie – proces budowy modelu odbywa się iteracyjnie. W każdym kroku t dodawane jest nowe drzewo decyzyjne, którego zadaniem jest skorygowanie błędów (rezyduów) popełnionych przez sumę drzew utworzonych w krokach poprzednich,
- optymalizacja drugiego rzędu – podczas wyboru najlepszych podziałów w węzłach drzewa, algorytm wykorzystuje rozwinięcie Taylora drugiego rzędu funkcji straty (używając pierwszej i drugiej pochodnej – gradientu i hesjanu), co pozwala na szybszą i dokładniejszą zbieżność modelu,

- regularyzacja – funkcja celu, którą optymalizuje algorytm, zawiera wbudowane składniki regularyzacyjne (kary za liczbę liści oraz wielkość wag w liściach), co skutecznie zapobiega przeuczeniu modelu (ang. overfitting),
- agregacja predykcji – wynik końcowy jest sumą wartości predykcji wszystkich utworzonych drzew (tzw. model addytywny).

Ostateczna predykcja dla danej próbki x stanowi sumę wyników zwróconych przez wszystkie K drzew w zespole, co opisuje wzór:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i)$$

gdzie:

- * \hat{y}_i – przewidywany wynik końcowy modelu dla i -tej próbki,
- * K – całkowita liczba drzew w modelu,
- * $f_k(x_i)$ – wynik (waga liścia) zwrócony przez k -te drzewo dla próbki x_i .

Proces aktualizacji wyniku w t -tej iteracji, z uwzględnieniem współczynnika uczenia (ang. learning rate), przedstawia się następująco:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i)$$

gdzie:

- * $\hat{y}_i^{(t)}$ – predykcja modelu po t -tej iteracji,
- * $\hat{y}_i^{(t-1)}$ – predykcja modelu z poprzedniej iteracji,
- * $f_t(x_i)$ – predykcja nowego drzewa dodanego w kroku t ,
- * η – współczynnik uczenia (zakres $0 < \eta \leq 1$), skalujący wpływ nowego drzewa na cały model.

W zadaniach klasyfikacji ostateczna suma \hat{y}_i poddawana jest zazwyczaj transformacji (np. funkcją sigmoidalną w klasyfikacji binarnej), aby uzyskać prawdopodobieństwo przynależności do klasy.

2.2. Opis metryk wybranych do oceny analizowanych modeli

Do oceny wydajności analizowanych modeli wykorzystano następujące metryki:

- **dokładność** – podstawowa miara jakości klasyfikacji, która określa odsetek wszystkich poprawnych przewidywań modelu względem liczby wszystkich obserwacji,
- **precyzja** – miara określająca jaki procent obserwacji zaklasyfikowanych przez model jako pozytywne rzeczywiście należy do klasy pozytywnej,
- **czułość** – miara oceniająca zdolność modelu do wykrywania wszystkich rzeczywistych przypadków pozytywnych,
- **błąd średniokwadratowy (en. MSE)** – miara jakości obliczana jako średnia kwadratów różnic między wartościami przewidywanymi i rzeczywistymi,
- **średni błąd bezwzględny (en. MAE)** – miara obliczana jako średnia wartość bezwzględna różnic między wartościami przewidywanymi, a rzeczywistymi,
- **pierwiastek błędu średniokwadratowego (en. RMSE)** – jest to pierwiastek kwadratowy z błędu średniokwadratowego dzięki czemu metryka wyrażona jest w tych samych jednostkach co przewidywana zmienna,
- **miara F1** - to średnia harmoniczna dwóch innych miar: precyzji i czułości.
- **krzywa ROC-AUC** – krzywa ilustrująca zależność między odsetkiem wyników prawdziwie pozytywnych oraz prawdziwie negatywnych.

2.3. Opis użytych bibliotek

Biblioteki zewnętrzne

1. **pandas** - Biblioteka używana do manipulacji i analizy danych, w szczególności do pracy z ramkami danych (DataFrames).
2. **numpy** - Podstawowy pakiet do obliczeń naukowych w Pythonie, zapewniający wsparcie dla dużych, wielowymiarowych tablic i macierzy.

3. **scipy** - Biblioteka wykorzystywana do zaawansowanych obliczeń naukowych i technicznych, bazująca na NumPy.
4. **scikit-learn** - Kluczowa biblioteka do uczenia maszynowego, która dostarcza proste i wydajne narzędzia do eksploracji i analizy danych.
5. **matplotlib** - Kompleksowa biblioteka do tworzenia statycznych, animowanych i interaktywnych wizualizacji w Pythonie.
6. **seaborn** - Biblioteka do wizualizacji danych oparta na matplotlib, która zapewnia wysokopoziomowy interfejs do rysowania atrakcyjnych grafik statystycznych.
7. **jupyter** - Interaktywne środowisko obliczeniowe (w formie notatników), które umożliwia tworzenie i udostępnianie dokumentów zawierających kod na żywo, wizualizacje i tekst.

Wbudowane moduły Pythona

1. **os** - Moduł zapewniający sposób na korzystanie z funkcjonalności zależnych od systemu operacyjnego, np. do zarządzania ścieżkami plików.
2. **sys** - Moduł zapewniający dostęp do specyficznych dla systemu parametrów i funkcji interpretera Pythona.
3. **warnings** - Moduł służący do kontrolowania sposobu obsługi komunikatów ostrzegawczych generowanych przez program.
4. **time** - Moduł dostarczający różne funkcje związane z czasem, np. do mierzenia długości wykonania operacji.
5. **abc** - Moduł dostarczający infrastrukturę do definiowania abstrakcyjnych klas bazowych (Abstract Base Classes).

3. Opis najważniejszych fragmentów kodu

Poniżej przedstawiono kluczowe fragmenty kodu zrealizowane w ramach projektu, podzielone na analizę statystyczną oraz modelowanie uczenia maszynowego.

A. Analiza Statystyczna i Regresja Liniowa

Analiza została przeprowadzona w pliku src/analysis/heart_disease_statistics.ipynb. Obejmuje ona czyszczenie danych, statystyki opisowe, analizę korelacji oraz regresję liniową.

1. Czyszczenie i obsługa brakujących danych

Funkcja `clean_and_handle_missing_data` odpowiada za identyfikację braków danych, duplikatów oraz wartości odstających.

```
def clean_and_handle_missing_data(df): # ... (analiza braków i duplikatów)

# Usuwanie duplikatów
if duplicates > 0:
    df_clean = df_clean.drop_duplicates()

# Wypełnianie brakujących wartości średnią
if missing_values.sum() > 0:
    df_clean = fill_missing_values(df_clean, method='mean')

return df_clean
```

Listing 3.1. Czyszczenie i obsługa brakujących danych.

2. Statystyki opisowe

Funkcja `calculate_descriptive_statistics` generuje szczegółowe statystyki dla zmiennych numerycznych, w tym miary tendencji centralnej, rozrzutu oraz kształtu rozkładu (skośność, kurtoza).

```
def calculate_descriptive_statistics(df): numeric_columns =
df.select_dtypes(include=[np.number]).columns
stats_dict = {}

for column in numeric_columns:
    data = df[column]
    stats_dict[column] = {
        'Min': data.min(),
        'Max': data.max(),
        'Średnia': data.mean(),
        'Odch. std': data.std(),
        'Mediana': data.median(),
        'Skośność': data.skew(),
        'Kurtoza': data.kurtosis()
    }

# ... (tworzenie DataFrame i zapis do CSV)
return stats_df
```

Listing 3.2. Statystyki opisowe.

3. Analiza korelacji

Funkcja `analyze_correlations` oblicza macierz korelacji Pearsona i identyfikuje silne zależności między zmiennymi a zmienną celową (`condition`).

```
def analyze_correlations(df): numeric_columns =  
df.select_dtypes(include=[np.number]).columns correlation_matrix =  
df[numeric_columns].corr(method='pearson')  
  
# Identifikacja korelacji ze zmienną celową  
if 'condition' in numeric_columns:  
    quality_correlations = []  
    for column in numeric_columns:  
        if column != 'condition':  
            corr_coef, p_value = pearsonr(df[column], df['condition'])  
            quality_correlations.append({  
                'Zmienna': column,  
                'Korelacja': corr_coef,  
                'p-wartość': p_value  
            })  
return correlation_matrix, strong_corr_df, quality_corr_df
```

Listing 3.3. Analiza korelacji.

4. Prosta i Wielokrotna Regresja Liniowa

Funkcja `perform_linear_regression_analysis` realizuje regresję liniową dla pojedynczych zmiennych oraz regresję wielokrotną dla najlepszych predyktorów, wykorzystując bibliotekę scikit-learn.

```
def perform_linear_regression_analysis(df): # Prosta regresja liniowa dla  
każdej zmiennej for predictor in top_predictors: X = df[[predictor]] y =  
df['condition']  
  
model = LinearRegression()  
model.fit(X, y)  
y_pred = model.predict(X)  
  
r_squared = r2_score(y, y_pred)
```

```

mse = mean_squared_error(y, y_pred)

# Regresja wielokrotna dla top 3 predyktorów
top_3_vars = [x[0] for x in top_predictors[:3]]
X_multi = df[top_3_vars]
model_multi = LinearRegression()
model_multi.fit(X_multi, y)

return correlations_with_condition, model_multi, top_predictors[0][0]

```

Listing 3.4. Prosta i Wielokrotna Regresja Liniowa.

B. Modelowanie Machine Learning - XGBoost

Poniżej przedstawiono kluczowe fragmenty kodu z pliku src/analysis/heart_xgb_ml.ipynb, odpowiedzialne za realizację analizy danych dotyczących chorób serca przy użyciu modelu XGBoost.

1. Wczytanie i przygotowanie danych

Pierwszym krokiem jest wczytanie zbioru danych oraz podział na cechy (X) i zmienną celową (y).

```

df = pd.read_csv('datasets/heart_cleveland_upload.csv') X =
df.drop('condition', axis=1) y = df['condition']

```

Listing 3.5. Wczytanie i przygotowanie danych.

Następnie dane są dzielone na zbiór treningowy i testowy, a cechy numeryczne są normalizowane. Wykorzystywane są do tego klasy pomocnicze zdefiniowane w module src.preprocessing.

```

splitter = DataSplitter(test_size=0.1, random_state=42) X_train, X_test,
y_train, y_test = splitter.split(X, y)

normalizer = StandardNormalizer() label_encoder = LabelEncoder() preprocessor
= DataPreprocessor(normalizer, label_encoder)

X_train_norm, X_test_norm = preprocessor.preprocess_features(X_train, X_test)
y_train_enc, y_test_enc = preprocessor.preprocess_labels(y_train, y_test)

```

Listing 3.6. Dzielenie na zbiór treningowy i testowy.

2. Inicjalizacja modelu i walidacja krzyżowa

W analizie wykorzystano model XGBoost (XGBoostModel). W celu wstępnej oceny jakości modelu na domyślnych parametrach zastosowano 5-krotną walidację krzyżową (CrossValidator).

```
xgb_model = XGBoostModel(n_estimators=100, random_state=42) cv =  
CrossValidator(n_splits=5, random_state=42) cv_scores = cv.validate(xgb_model,  
X_train_norm, y_train_enc)
```

Listing 3.7. Inicjalizacja modelu i walidacja krzyżowa.

3. Optymalizacja hiperparametrów

Kluczowym etapem jest optymalizacja hiperparametrów modelu w celu uzyskania jak najlepszych wyników. Wykorzystano klasę HyperparameterTuner do przeszukiwania siatki (grid search) w przestrzeni parametrów zdefiniowanej przez model.

```
tuner = HyperparameterTuner(search_type='grid', cv=5) param_space =  
xgb_model.get_param_grid() xgb_model, cv_results = tuner.tune(xgb_model,  
X_train_norm, y_train_enc, param_space)
```

Listing 3.8. Optymalizacja hiperparametrów.

4. Analiza statystyczna

4.1. Zbiór „Wine Quality”

Podczas dokonywania analizy statystycznej w zbiorze „Wine Quality” wykryto 1177 zduplikowanych rekordów. Poniższa tabela przedstawia zakres oraz udział procentowy wartości odstających w zbiorze.

Tabela 4.1. Zakres oraz udział procentowy wartości odstających w zbiorze.

Atrybut	Minimum	Maksimum	Ilość wartości odstających
Kwasowość stała	3.800	15.9	357 (5.49%)
Kwasowość lotna	0.080	1.580	377 (5.80%)
Kwas cytrynowy	0.000	1.660	509 (7.83%)
Cukier resztkowy	0.600	65.800	118 (1.82%)

Chlorki	0.009	0.611	286 (4.40%)
Wolny SO ₂	1.000	289.000	62 (0.95%)
Całkowity SO ₂	6.000	440.000	10 (0.15%)
Gęstość	0.987	1.039	3 (0.05%)
pH	2.720	4.010	73 (1.12%)
Siarczany	0.220	2.000	191 (2.94%)
Alkohol	8.000	14.900	3 (0.05%)
Jakość	3.000	9.000	228 (3.51%)

Wyniki zapisane w powyższej tabeli wykazują, iż najbardziej nieregularny rozkład wartości występuje w atrybucie kwas cytrynowy, gdzie wartości odchodzące od normy stanowią odsetek 7.83 procenta wszystkich wartości. Kolejno, w procesie analizy obliczono wartości odchylenia standardowego, średniej, mediany oraz kwartyli. Poniższa tabela przedstawia uzyskane wyniki.

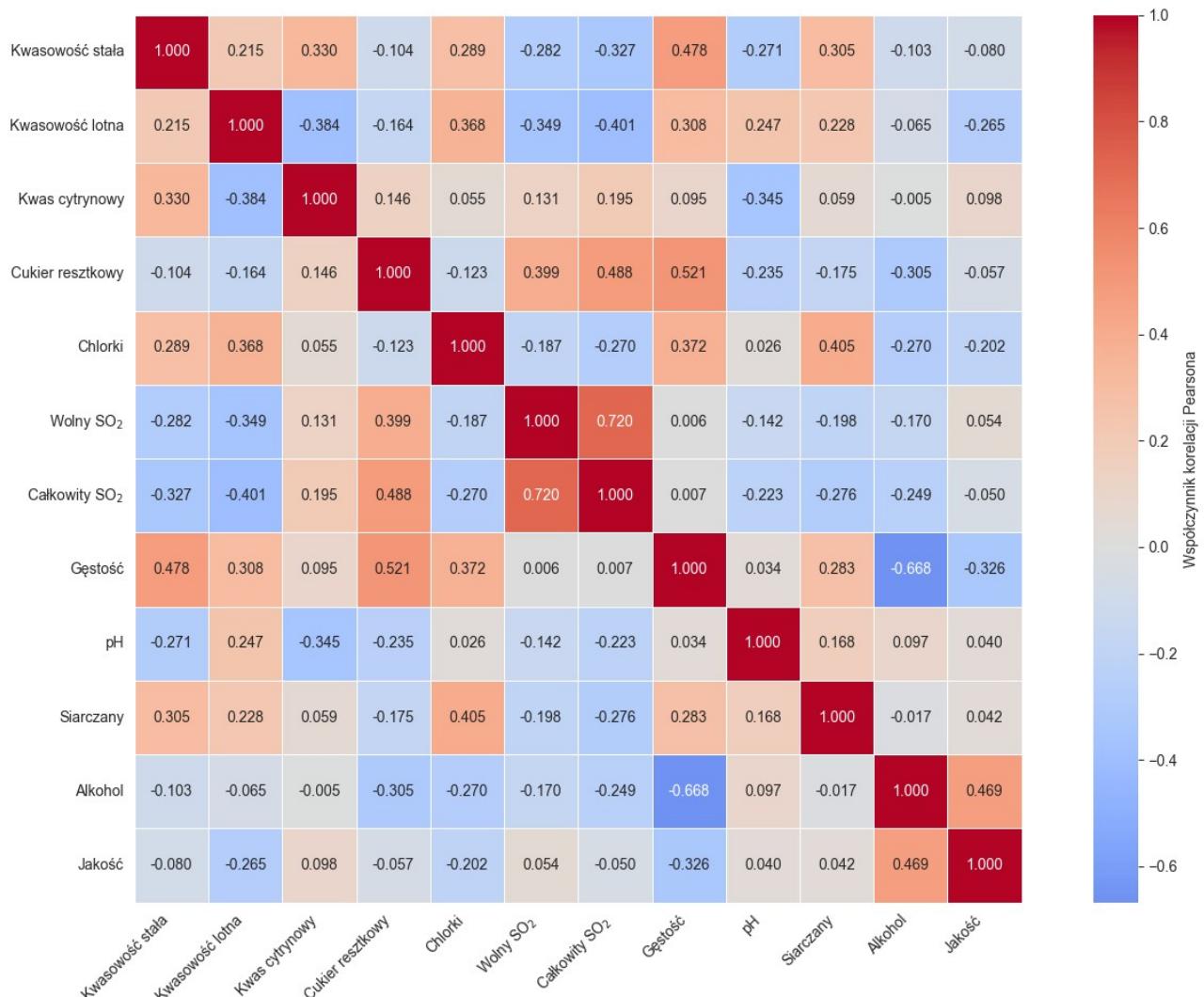
Tabela 4.2. Statystyki opisowe.

Atrybut	Średnia	Odchylenie standardowe	Medianą	Q1 (25%)	IQR (Rozstęp ćwiartkowy)	Q3 (75%)	Skośność	Kurtoza
Kwasowość stała	7.215	1.320	7.000	6.400	1.300	7.700	1.650	4.589
Kwasowość lotna	0.344	0.168	0.300	0.230	0.180	0.410	1.505	2.863
Kwas cytrynowy	0.318	0.147	0.310	0.240	0.160	0.400	0.484	2.582
Cukier resztkowy	5.048	4.500	2.700	1.800	5.700	7.500	1.707	7.026
Chlorki	0.057	0.037	0.047	0.038	0.028	0.066	5.338	48.261
Wolny SO ₂	30.037	17.805	28.000	16.000	25.000	41.000	1.363	9.521
Całkowity SO ₂	114.109	56.774	116.000	74.000	79.250	153.250	0.064	-0.300
Gęstość	0.995	0.003	0.995	0.992	0.005	0.997	0.666	8.711
pH	3.225	0.160	3.210	3.110	0.220	3.330	0.390	0.432

Siarczany	0.533	0.150	0.510	0.430	0.170	0.600	1.809	8.613
Alkohol	10.549	1.186	10.400	9.500	1.900	11.400	0.546	-0.538
Jakość	5.796	0.880	6.000	5.000	1.000	6.000	0.147	0.298

Analiza struktury danych ujawnia dominację rozkładów prawostronne asymetrycznych i leptokurtycznych, co w przypadku chlorków (ekstremalna kurtoza 48,26) wskazuje na silną koncentrację wyników wokół średniej przy jednoczesnym występowaniu odległych wartości odstających. Na tym tle wyróżniają się całkowity SO₂ oraz pH, które jako jedyne wykazują cechy rozkładu normalnego (symetria i mezokurtyczność), stanowiąc najbardziej stabilne parametry w badanym zbiorze. Kolejnym etapem analizy statystycznej była analiza korelacji liniowej Pearsona. Uzyskane wyniki przedstawiono na rysunku 4.1.

Mapa ciepła korelacji między zmiennymi



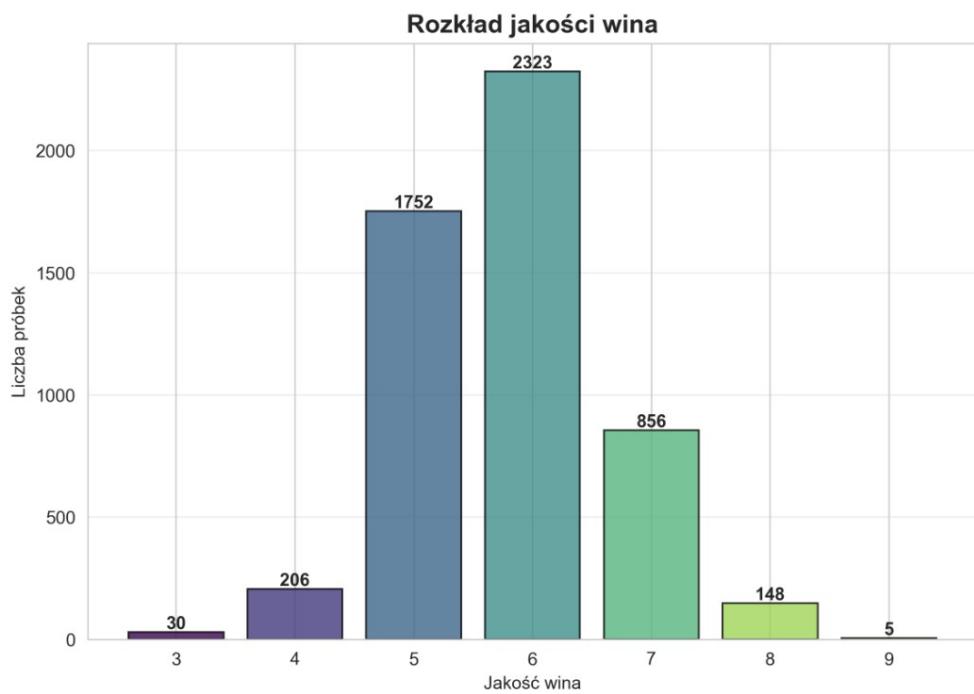
Rysunek 4.1. Wyniki uzyskane podczas analizy korelacji liniowej Pearsona.

Analiza macierzy Pearsona ujawnia logiczne współzależności fizykochemiczne, w tym silną ujemną korelację między gęstością a zawartością alkoholu ($r=-0.668$) oraz techniczną zależność frakcji dwutlenku siarki ($r=0.720$). W kontekście oceny organoleptycznej, kluczowym pozytywnym determinantem jakości okazało się stężenie alkoholu ($r=0.469$), podczas gdy wzrost gęstości i kwasowości lotnej wykazuje umiarkowany wpływ negatywny na ocenę końcową wina. Kolejnym etapem analizy było dokonanie regresji liniowej między zmiennymi aby znaleźć najsilniejszego predyktora jakości wina. Poniższa tabela przedstawia porównanie modeli regresji od najsilniejszego do najsłabszego.

Tabela 4.3. Uzyskane wyniki dla dokonanej regresji liniowej.

Model/Predyktor	Równanie regresji ($y = ax + b$)	R^2 (dopasowanie)	RMSE (Błąd)	Trend
Alkohol	Jakość = $0.35 * \text{Alkohol} + 2.12$	22.0%	0.777	Dodatni
Gęstość	Jakość = $-96.84 * \text{Gęstość} + 102.11$	10.7%	0.832	Ujemny
Kwasowość lotna	Jakość = $-1.39 * \text{Kwasowość lotna} + 6.27$	7.0%	0.848	Ujemny
Model wielokrotny	Jakość = $0.39 * \text{Alkohol} + 30.31 * \text{Gęstość} - 1.37 * \text{Kwasowość lotna}$	28.1%	0.746	Mieszany

Analiza regresji wykazuje, że zawartość alkoholu jest najsilniejszym pojedynczym predyktorem jakości ($R^2=22\%$), przy czym zależność ta ma charakter liniowy – zastosowanie wielomianu drugiego stopnia nie przyniosło istotnej poprawy modelu (wzrost R^2 o zaledwie 0,01 pp). Włączenie do modelu gęstości oraz kwasowości lotnej w ramach regresji wielokrotnej pozwoliło na wyjaśnienie 28,1% całkowitej wariancji, co sugeruje, że choć parametry te są istotne, ocena jakości wina zależy również od czynników nieobjętych tymi trzema zmiennymi. Warto odnotować silną ujemną korelację między gęstością a alkoholem ($r=-0.668$), co potwierdza fizykochemiczną spójność danych (alkohol obniża gęstość roztworu). Następnym etapem była wizualizacja danych w postaci wykresów rozkładu, pudełkowych, skrzypcowych, rozrzutu oraz histogramów. Poniższe rysunki przedstawiają uzyskane wykresy wraz z interpretacją.



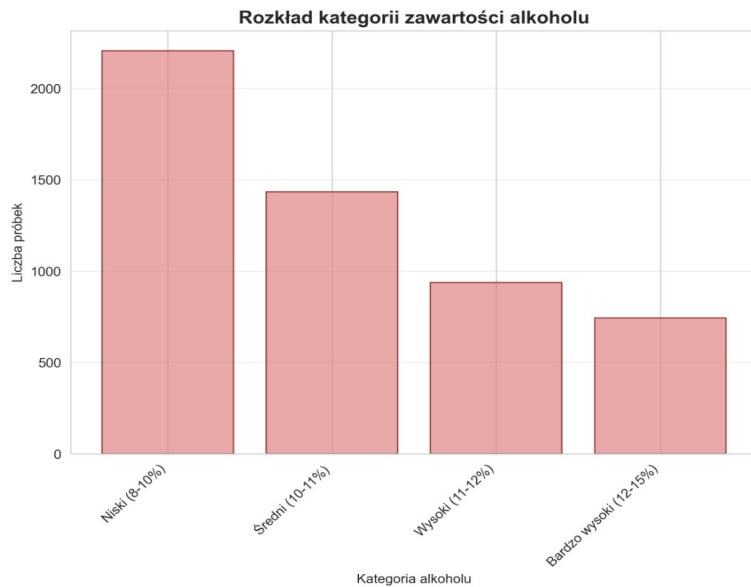
Rysunek 4.2. Rozkład jakości wina.

Z powyższego rozkładu można wywnioskować, iż zmienna celu (jakość) wykazuje rozkład zbliżony do normalnego z silną koncentracją wokół wartości średnich – aż 76,6% wszystkich próbek stanowią wina ocenione na 5 lub 6 punktów. Klasy skrajne są niedoreprezentowane (szczególnie oceny 3 i 9), co oznacza, że model predykcyjny może mieć trudności z poprawną klasyfikacją win bardzo dobrych oraz bardzo słabych. Poniższy rysunek przedstawia top 10 wartości siarczanów.



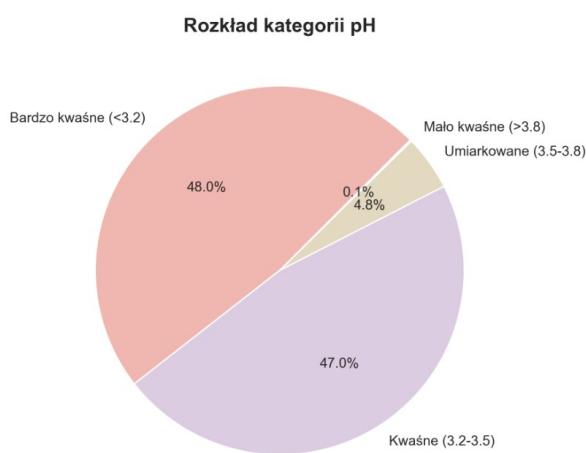
Rysunek 4.3. Wykres kołowy reprezentujący 10 najlepszych wartości siarczanów.

Powyższy wykres przedstawia najczęściej występujące wartości (mody) dla zmiennej siarczanów. Obserwuje się koncentrację wyników w przedziale 0,28–0,32, przy czym trzy najliczniejsze wartości (0,32; 0,31; 0,30) stanowią łącznie dominującą część (ok. 70%) zaprezentowanego podzbioru danych. Poniższy rysunek przedstawia rozkład kategorii zawartości alkoholu.



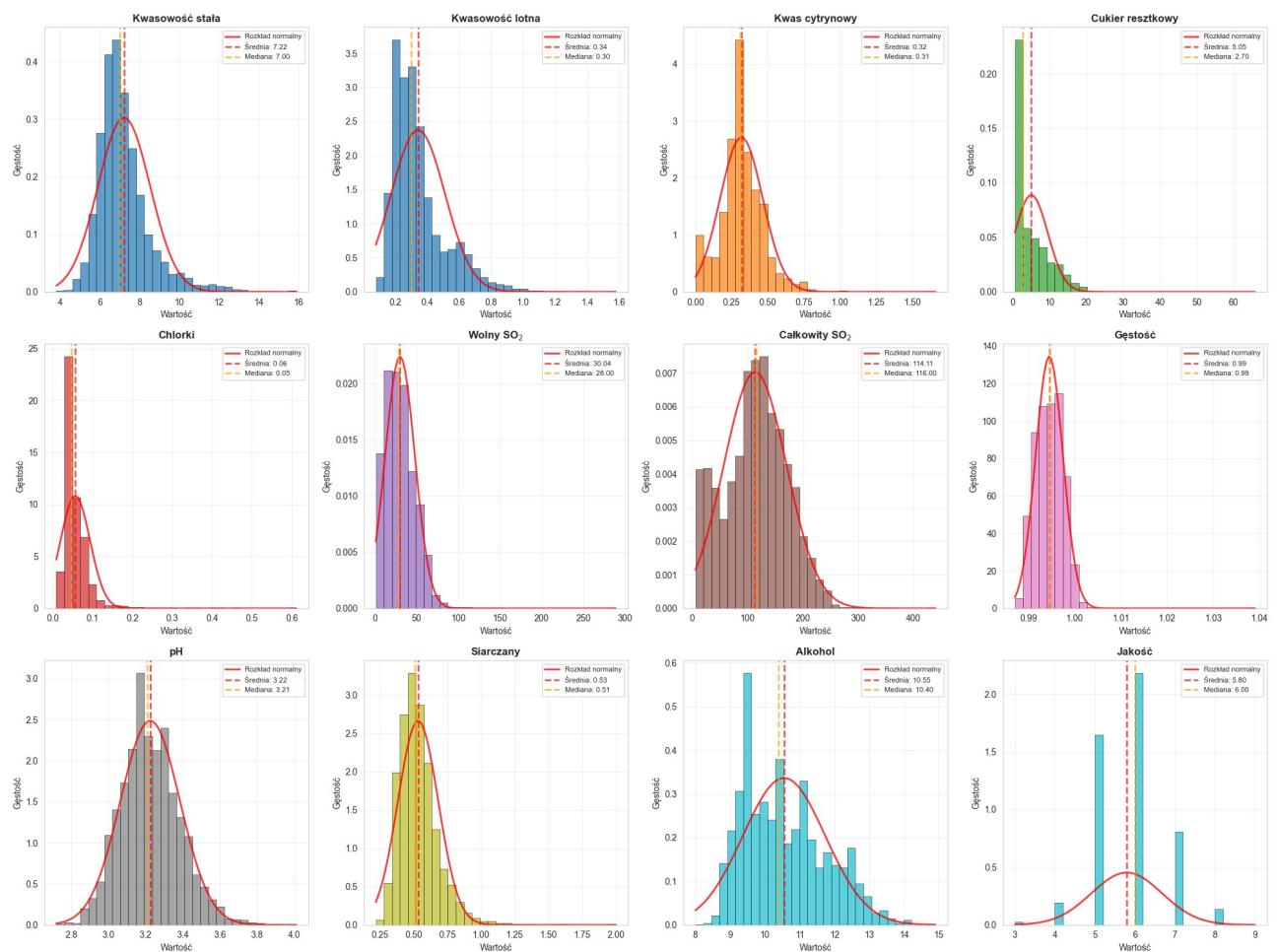
Rysunek 4.4. Rozkład kategorii zawartości alkoholu.

Na wykresie zawartym powyżej można zauważyć, iż dane wskazują na odwrotną zależność między liczebnością próbek a zawartością alkoholu. Najliczniejszą grupę stanowią wina o najniższej zawartości alkoholu (8-10%), natomiast wraz ze wzrostem procentowym liczba obserwacji maleje, osiągając minimum w przedziale 12-15%. Rozkład ten wykazuje cechy asymetrii prawostronnej. Poniższy rysunek przedstawia rozkład kategorii pH.



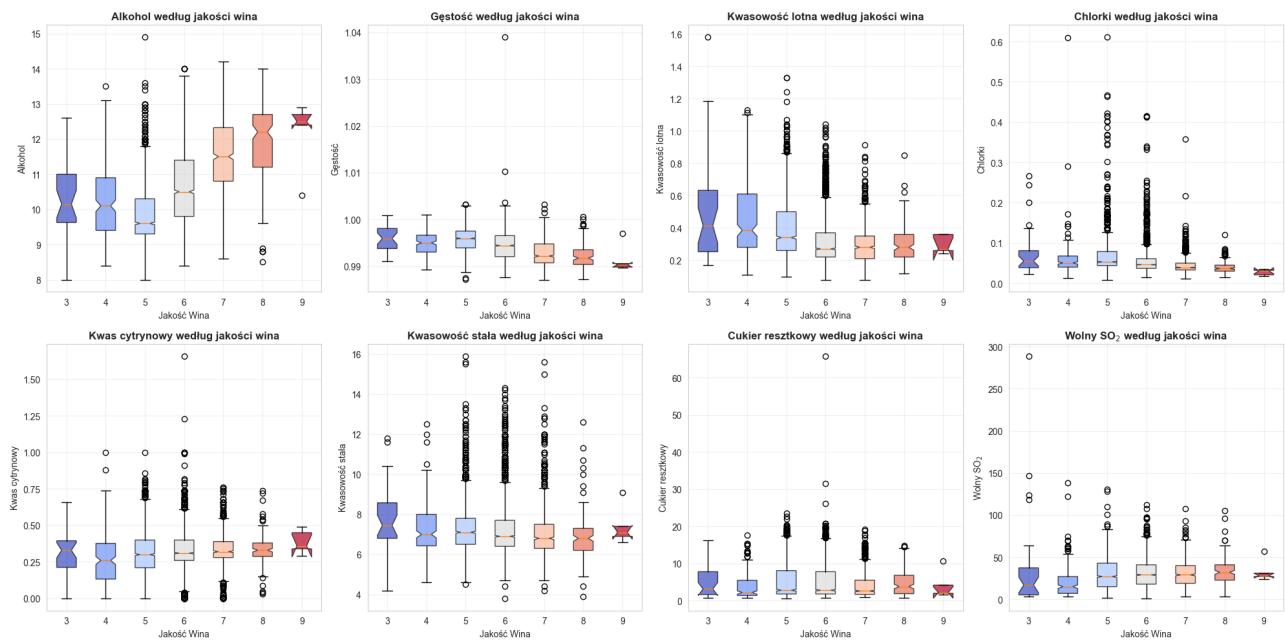
Rysunek 4.5. Rozkład kategorii pH.

Z powyższego rozkładu można odczytać, że zdecydowana większość badanych próbek (95%) charakteryzuje się wartościami pH poniżej 3,5, co odpowiada kategoriom "Bardzo kwaśne" oraz "Kwaśne". Odsetek próbek o wyższym pH (powyżej 3,8) jest marginalny, co wskazuje na niskie zróżnicowanie badanej populacji w zakresie wyższych wartości tego parametru. Poniższy rysunek przedstawia histogramy zmiennych zawartych w zbiorze.



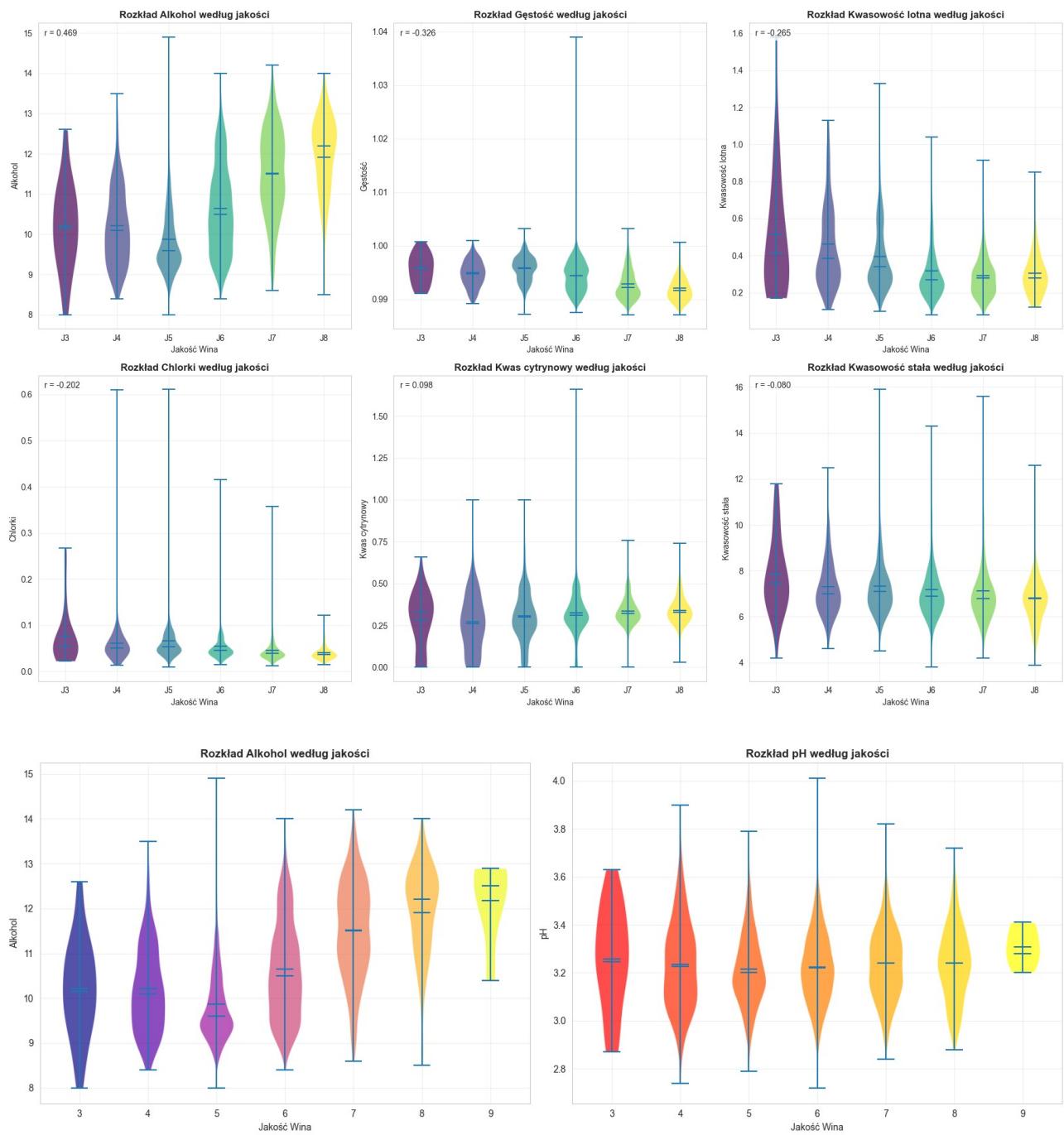
Rysunek 4.6. Histogramy pojedynczych atrybutów w zbiorze.

Analiza wizualna wykazuje, że większość zmiennych (w szczególności cukier resztkowy i chlorki) charakteryzuje się wyraźną asymetrią prawostronną z długimi ogonami wartości odstających. Jedynie parametry takie jak pH, gęstość oraz całkowity SO₂ posiadają rozkłady symetryczne, wykazujące duże podobieństwo do teoretycznej krzywej normalnej. W przypadku kwasu cytrynowego obserwuje się nietypowy, wielomodalny kształt rozkładu, co wskazuje na występowanie kilku lokalnych skupień danych. Zmienna celu (Jakość) ma charakter dyskretny i jest silnie skoncentrowana wokół wartości środkowych (5 i 6), przy wyraźnym niedoborze próbek dla ocen skrajnych. Poniższy rysunek przedstawia wykresy pudełkowe zależności poszczególnych atrybutów od jakości wina.



Rysunek 4.7. Wykresy pułapkowe zależności poszczególnych atrybutów od jakości wina.

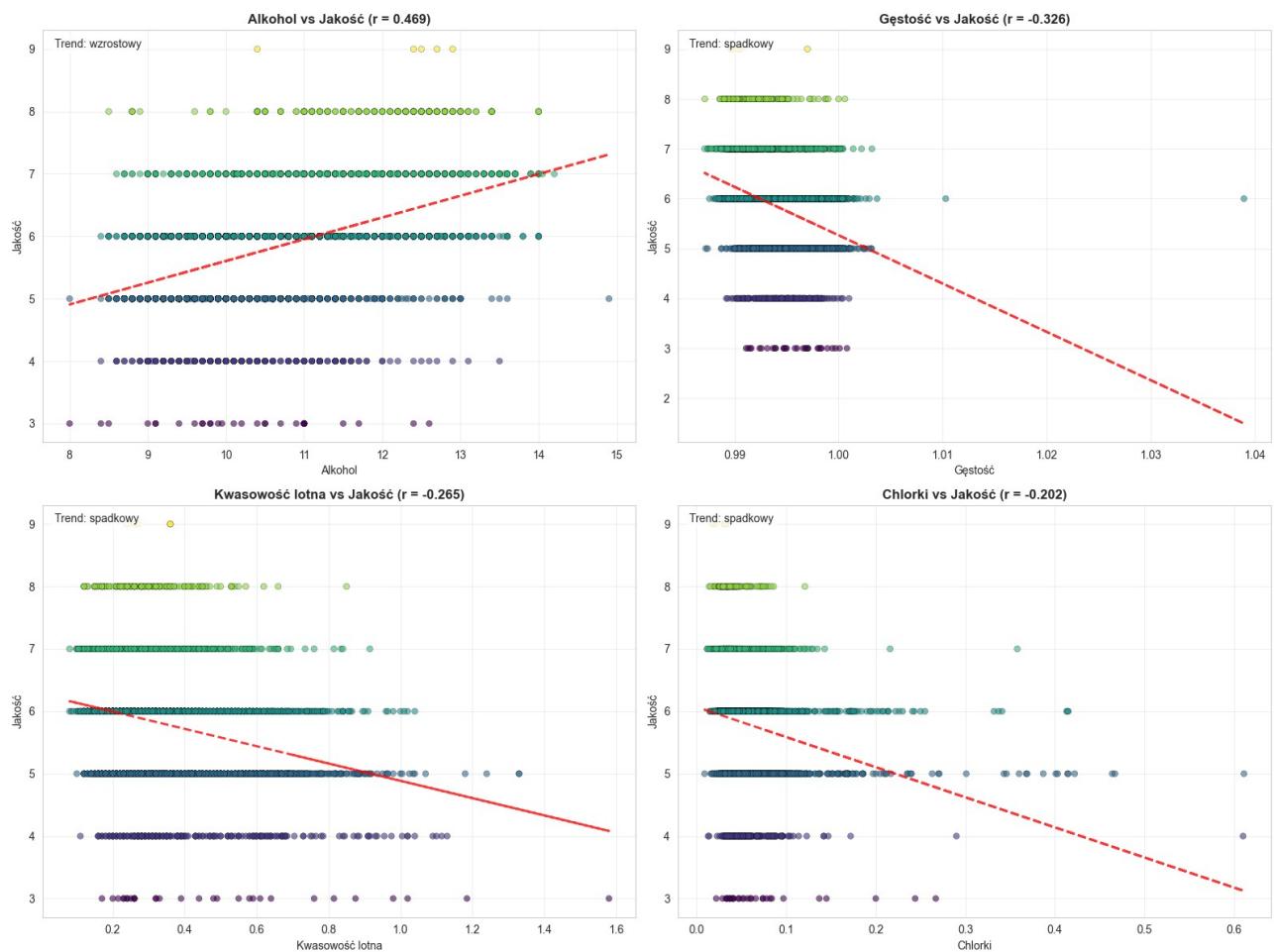
Prezentowane wykresy pułapkowe wizualizują rozkład parametrów fizykochemicznych wina w podziale na oceny jakości (od 3 do 9), pozwalając na ocenę tendencji centralnej oraz rozrzutu danych. Każde "pułapko" obejmuje środkowe 50% obserwacji (rozstęp międzykwartylowy), a pozioma linia wewnętrz rysunku przedstawia medianę, co pozwala zidentyfikować kluczowe trendy – wyraźny wzrost stężenia alkoholu oraz spadek kwasowości lotnej wraz z poprawą jakości trunku. Punkty znajdujące się poza "wąsami" reprezentują wartości odstające (anomie), które są najliczniejsze w klasach średnich (5 i 6), co świadczy o dużej niejednorodności win w tym segmencie cenowym. Wysokość pułapek informuje o stabilności procesu – spłaszczone pułapki dla gęstości i chlorków w najwyższych ocenach (8 i 9) wskazują na wysoką powtarzalność i precyzję parametrów w winach klasy premium. Poniższy rysunek prezentuje wykresy skrzypcowe zależności poszczególnych atrybutów od jakości wina.



Rysunek 4.8. Wykresy skrzypcowe zależności poszczególnych atrybutów od jakości wina.

Prezentowane wykresy skrzypcowe, łączące statystykę pozycji z gęstością rozkładu, potwierdzają, że alkohol jest najsilniejszym dyskryminatorem jakości ($r=0.469$), co widać po wyraźnym przesunięciu "brzuchów" skrzypiec (największego skupiska danych) w górę osi Y dla wyższych ocen. W przypadku zmiennych negatywnie skorelowanych, takich jak kwasowość lotna i chlorki, niższe klasy jakościowe charakteryzują się silnie wydłużonymi górnymi "szyjkami", co sygnalizuje częste występowanie ekstremalnie wysokich stężeń tych substancji w słabszych winach. Dla kontrastu, wina najwyższej jakości (oceny 7 i 8) cechują się znacznie bardziej zwartymi i płaskimi kształtami wykresów przy dolnych wartościach osi, co świadczy o wysokiej

powtarzalności procesu i eliminacji wad chemicznych. Parametry takie jak pH oraz kwasowość stała wykazują niemal identyczny, symetryczny profil rozkładu we wszystkich grupach, co wizualnie potwierdza ich znakomą przydatność w prognozowaniu jakości wina. Poniższy rysunek przedstawia wykresy rozrzutu silnie skorelowanych atrybutów ze zmienną docelową.



Rysunek 4.9. Wykresy rozrzutu silnie skorelowanych atrybutów ze zmienną docelową.

Z powyższego rysunku można wywnioskować, że alkohol posiada dominującą rolę jako jedyny parametr o wyraźnym trendzie wzrostowym ($r=0.469$), gdzie czerwona linia regresji jednoznacznie wskazuje, że wyższa zawartość procentowa sprzyja lepszym ocenom. Pozostałe zmienne – gęstość, kwasowość lotna i chlorki – charakteryzują się nachyleniem ujemnym (trend spadkowy), co wizualnie dowodzi, że niższe stężenia tych substancji korelują z wyższą jakością wina. Specyficzny, pasmowy układ punktów wynika z dyskretnej skali ocen (liczby całkowite 3–9), jednak linia dopasowania skutecznie przecina te skupiska, wyznaczając ogólną tendencję statystyczną. Warto zauważyć pustą przestrzeń w prawych górnym rogach wykresów dla kwasowości lotnej i chlorków, co sugeruje, że wina o wysokim stężeniu tych wadliwych składników praktycznie nie mają szans na osiągnięcie najwyższych not (8 i 9).

4.2. Zbiór „Heart Disease”

Podczas dokonywania analizy statystycznej w zbiorze „Heart Disease” (Cleveland) nie wykryto zduplikowanych rekordów (całych wierszy). Poniższa tabela przedstawia zakres oraz udział procentowy wartości odstających w zbiorze dla atrybutów ciągłych.

Tabela 4.4. Zakres oraz udział procentowy wartości odstających w zbiorze „Heart Disease”.

Atrybut	Minimum	Maksimum	Ilość wartości odstających
Wiek (age)	29.000	77.000	0 (0.00%)
Spoczynkowe ciśnienie krwi (trestbps)	94.000	200.000	9 (3.03%)
Cholesterol (chol)	126.000	564.000	5 (1.68%)
Maksymalne tętno (thalach)	71.000	202.000	1 (0.34%)
Depresja ST (oldpeak)	0.000	6.200	5 (1.68%)

Wyniki zapisane w powyższej tabeli wykazują, iż najbardziej nieregularny rozkład wartości (tj. największy udział wartości odstających) występuje w atrybucie Spoczynkowe ciśnienie krwi, gdzie wartości odchodzące od normy (powyżej 170 mm Hg lub poniżej 90 mm Hg) stanowią odsetek 3.03 procenta wszystkich wartości.

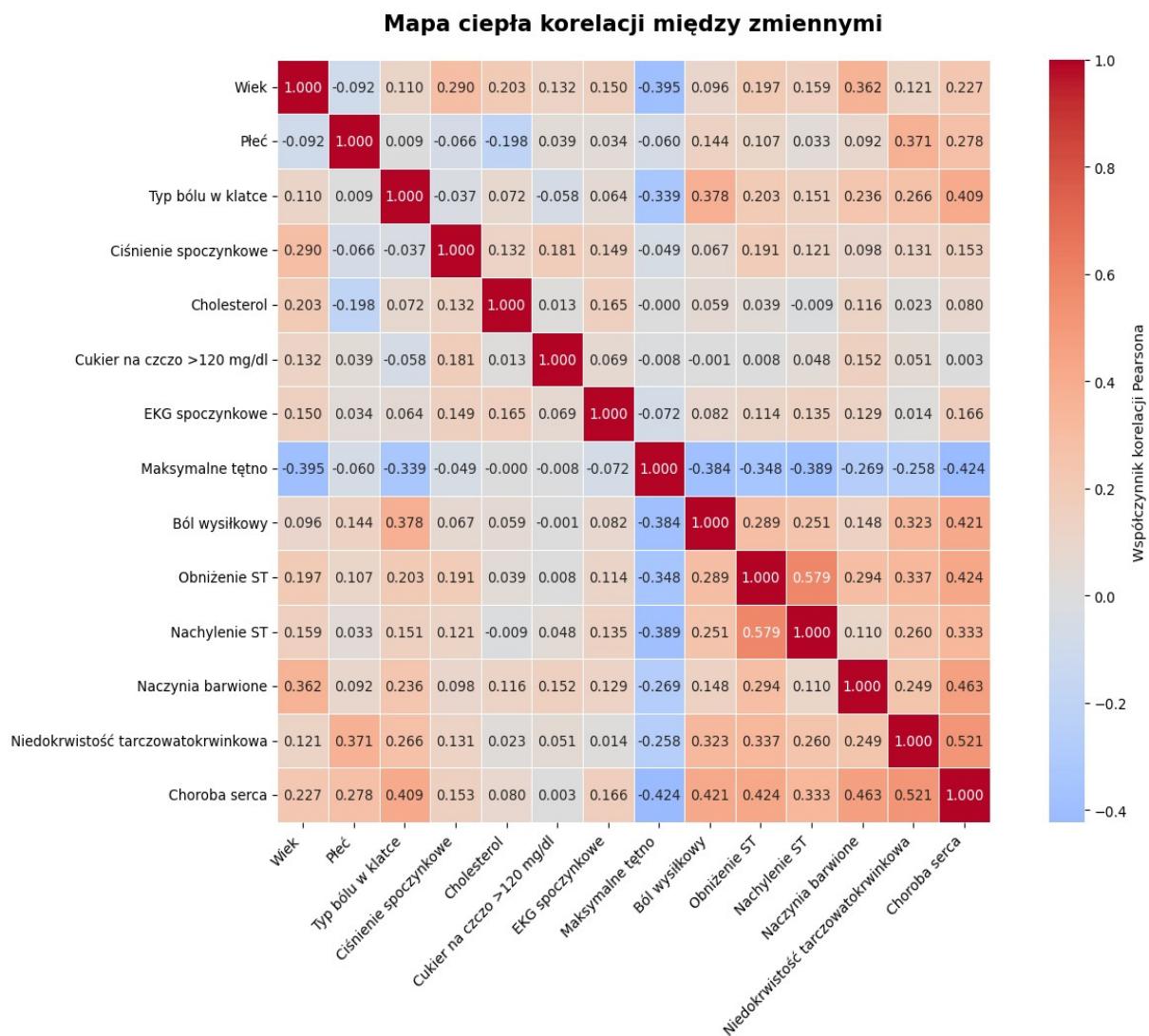
Tabela 4.5. Statystyki opisowe.

Atrybut	Średnia	Odchylenie standardowe	Medianą	Q1 (25%)	Q3 (75%)	IQR (Rozstęp ćwiartkowy)	Skośność	Kurtoza
Wiek	54.54	9.05	56.0	48.0	61.0	13.0	-0.22	-0.52
Płeć	0.68	0.47	1.0	0.0	1.0	1.0	-0.76	-1.43
Typ bólu w klatce (cp)	2.16	0.96	2.0	2.0	3.0	1.0	-0.84	-0.41
Ciśnienie spoczynkowe (trestbps)	131.69	17.76	130.0	120.0	140.0	20.0	0.70	0.81
Cholesterol (chol)	247.35	52.00	243.0	211.0	276.0	65.0	1.12	4.44
Cukier na czczo >120 mg/dl (fbs)	0.14	0.35	0.0	0.0	0.0	0.0	2.03	2.13
EKG spoczynkowe (restecg)	1.00	0.99	1.0	0.0	2.0	2.0	0.01	-2.00
Maksymalne tętno (thalach)	149.60	22.94	153.0	133.0	166.0	33.0	-0.54	-0.05
Ból wysiłkowy (exang)	0.33	0.47	0.0	0.0	1.0	1.0	0.74	-1.46
Obniżenie ST (oldpeak)	1.06	1.17	0.8	0.0	1.6	1.6	1.25	1.51
Nachylenie ST (slope)	0.60	0.62	1.0	0.0	1.0	1.0	0.51	-0.63
Naczynia barwione (ca)	0.68	0.94	0.0	0.0	1.0	1.0	1.18	0.24
Niedokrwistość tarczowatokrwinkowa (thal)	0.84	0.96	0.0	0.0	2.0	2.0	0.34	-1.83
Choroba serca (condition)	0.46	0.50	0.0	0.0	1.0	1.0	0.16	-1.99

Analiza struktury danych ujawnia dominację rozkładów o dodatniej skośności (prawostronnie asymetrycznych) i leptokurtycznych (z ostrym szczytem i grubymi ogonami), co w przypadku Cholesterolu (chol) (ekstremalna kurtoza 4.44) oraz Cukru na czczo (fbs) (ekstremalna skośność 2.03) wskazuje na występowanie znaczającej liczby pacjentów z wysokimi wynikami, odległymi od średniej (wartości odstające).

Na tym tle wyróżnia się Wiek (age), który jako jedyny wykazuje cechy rozkładu bliskiego normalnemu (skośność -0.22 i kurtoza -0.52), stanowiąc najbardziej stabilny parametr demograficzny w badanym zbiorze.

Kolejnym etapem analizy statystycznej była analiza korelacji liniowej Pearsona. Uzyskane wyniki przedstawiono na rysunku 4.10.



Rysunek 4.10. Wyniki uzyskane podczas analizy korelacji liniowej Pearsona.

Analiza macierzy Pearsona ujawnia kluczowe współzależności kliniczne w kontekście diagnozy. Najsilniejsze korelacje ze Stanem/Diagnozą (condition) wykazują Typ bólu w klatce piersiowej (cp) (silna korelacja ujemna) oraz Dławica indukowana wysiłkiem (exang), Liczba naczyń barwionych (ca) i Obniżenie ST (oldpeak) (silne korelacje dodatnie). W kontekście predykcyjnym, kluczowym pozytywnym determinantem choroby serca okazała się Dławica indukowana wysiłkiem oraz liczba zablokowanych naczyń, podczas gdy występowanie atypowego bólu w klatce piersiowej wykazuje silny wpływ negatywny na ostateczną diagnozę (osoby z nietypowym bólem rzadziej mają zdiagnozowaną chorobę serca).

Kolejnym etapem analizy było dokonanie regresji liniowej między zmiennymi, aby znaleźć najsilniejszego predyktora choroby serca. Poniższa tabela przedstawia porównanie modeli regresji od najsilniejszego do najsłabszego.

Tabela 4.6. Uzyskane wyniki dla dokonanej regresji liniowej.

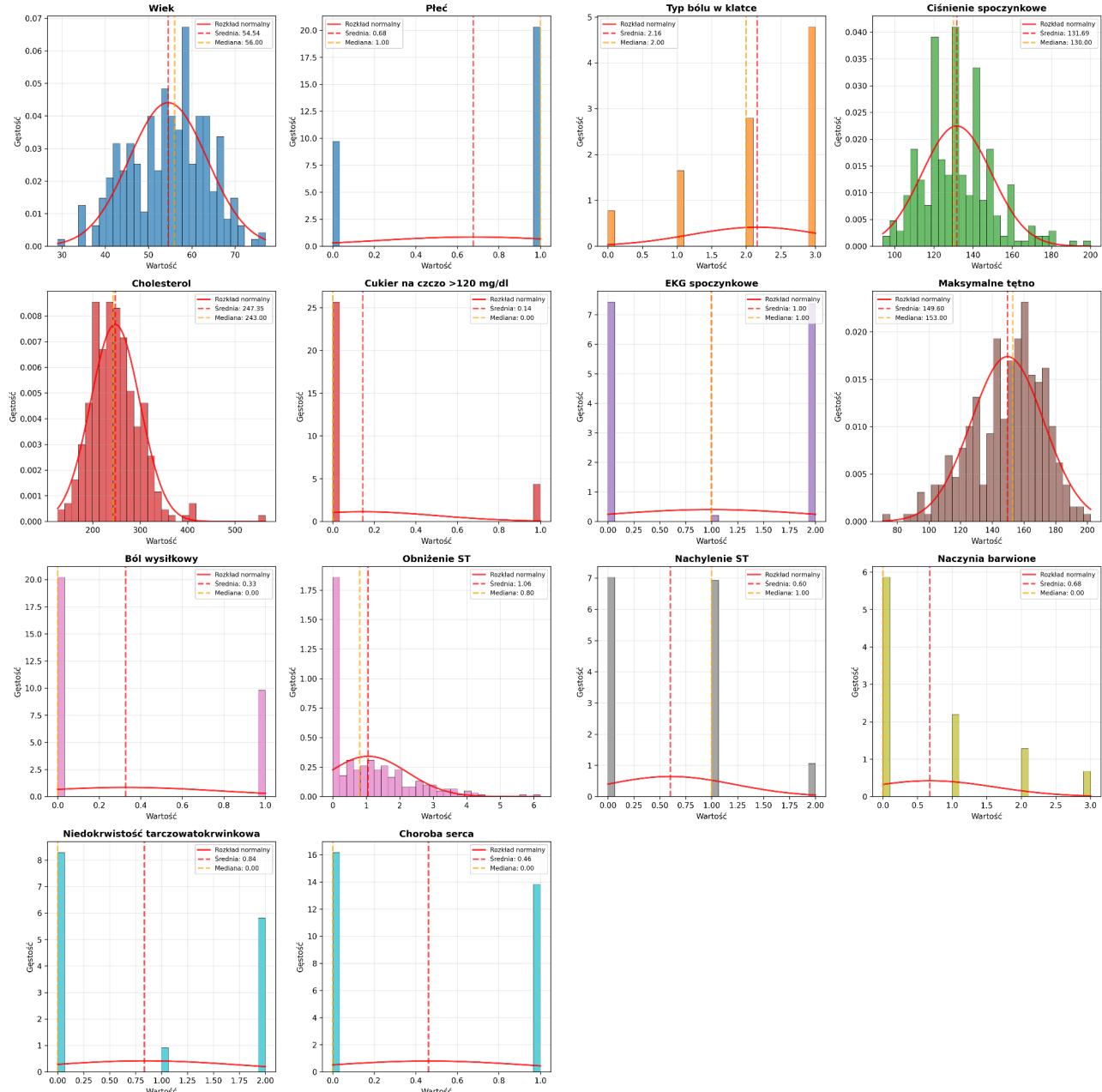
Model/Predyktor	Równanie regresji ($y=ax+b$)	R2 (dopasowanie)	RMSE (Błąd)	Trend
Niedokrwistość tarczowatokrwinkowa (thal)	$y = 0.2717x + 0.2344\$$	0.2709	0.4256	Dodatni
Naczynia barwione (ca)	$y = 0.2463x + 0.2946\$$	0.2145	0.4418	Dodatni
Obniżenie ST (oldpeak)	$y = 0.1816x + 0.2696\$$	0.1798	0.4515	Dodatni

Analiza regresji wykazuje, że Niedokrwistość tarczowatokrwinkowa (**thal**) jest najsilniejszym pojedynczym predyktorem Choroby serca ($R^2 = 27.1\%$), przy czym zależność ta ma charakter liniowy.

Włączenie do modelu Liczby naczyń barwionych (**ca**) oraz Obniżenia ST (**oldpeak**) w ramach regresji wielokrotnej pozwoliło na wyjaśnienie 42.56% całkowitej wariancji ($R^2 = 0.4256\$$), co sugeruje, że choć te parametry kliniczne są bardzo istotne, diagnoza choroby serca zależy również od czynników nieuwzględnionych w tym trójzmiennym modelu.

Warto odnotować spójność predykcyjną cech: wszystkie trzy atrybuty (Thal, Ca, Oldpeak) wykazują dodatni trend regresji, co jest klinicznie zgodne – ich wzrost wartości (np. większe obniżenie ST lub większa liczba zablokowanych naczyń) przekłada się na wyższe prawdopodobieństwo diagnozy choroby serca.

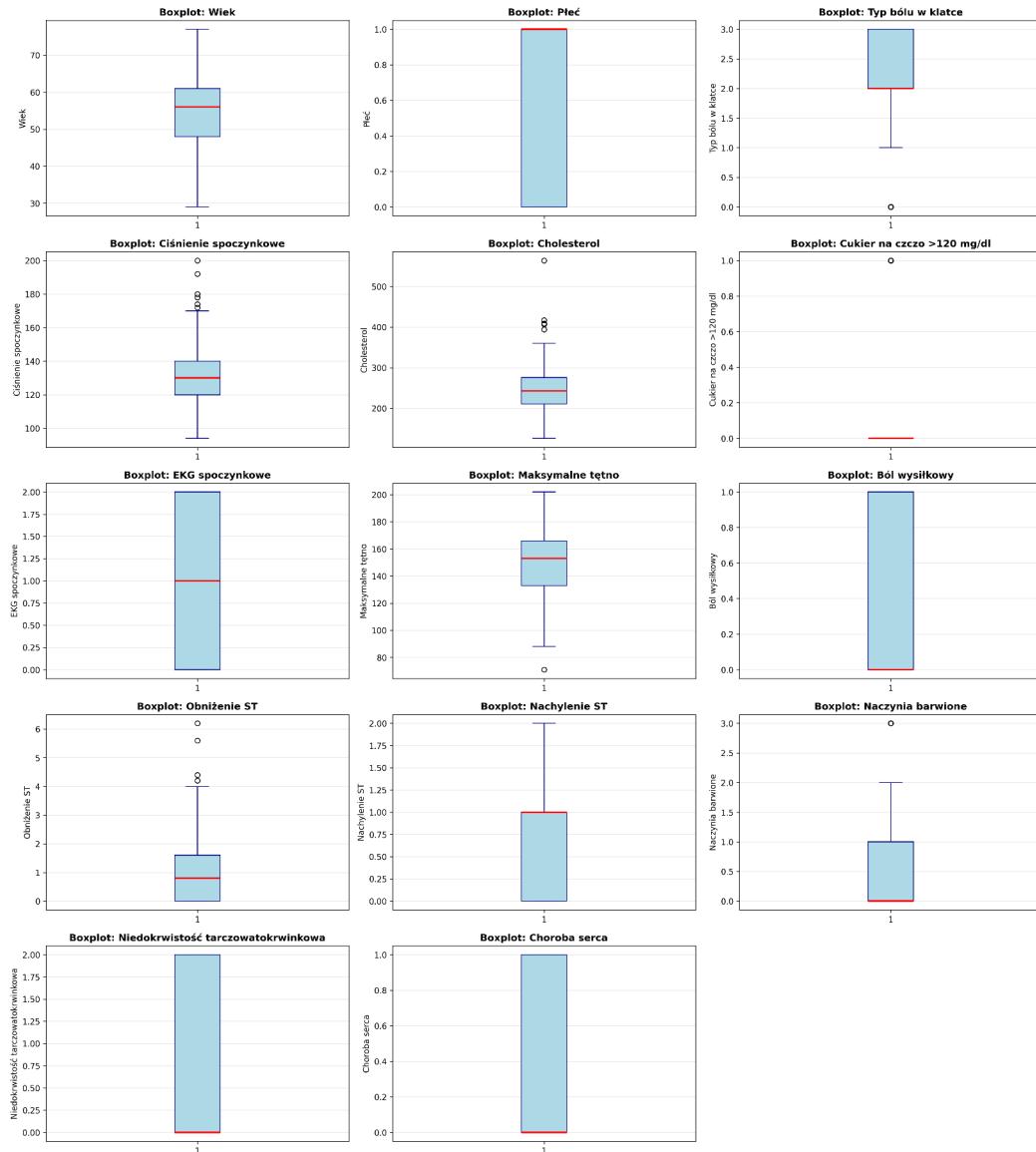
Następnym etapem była wizualizacja danych w postaci wykresów rozkładu, pudełkowych, skrzypcowych, rozrzutu oraz histogramów. Poniższe rysunki przedstawiają uzyskane wykresy wraz z interpretacją.



Rysunek 4.11. Rozkłady Częstości (Histogramy) Atrybutów Ciągłych

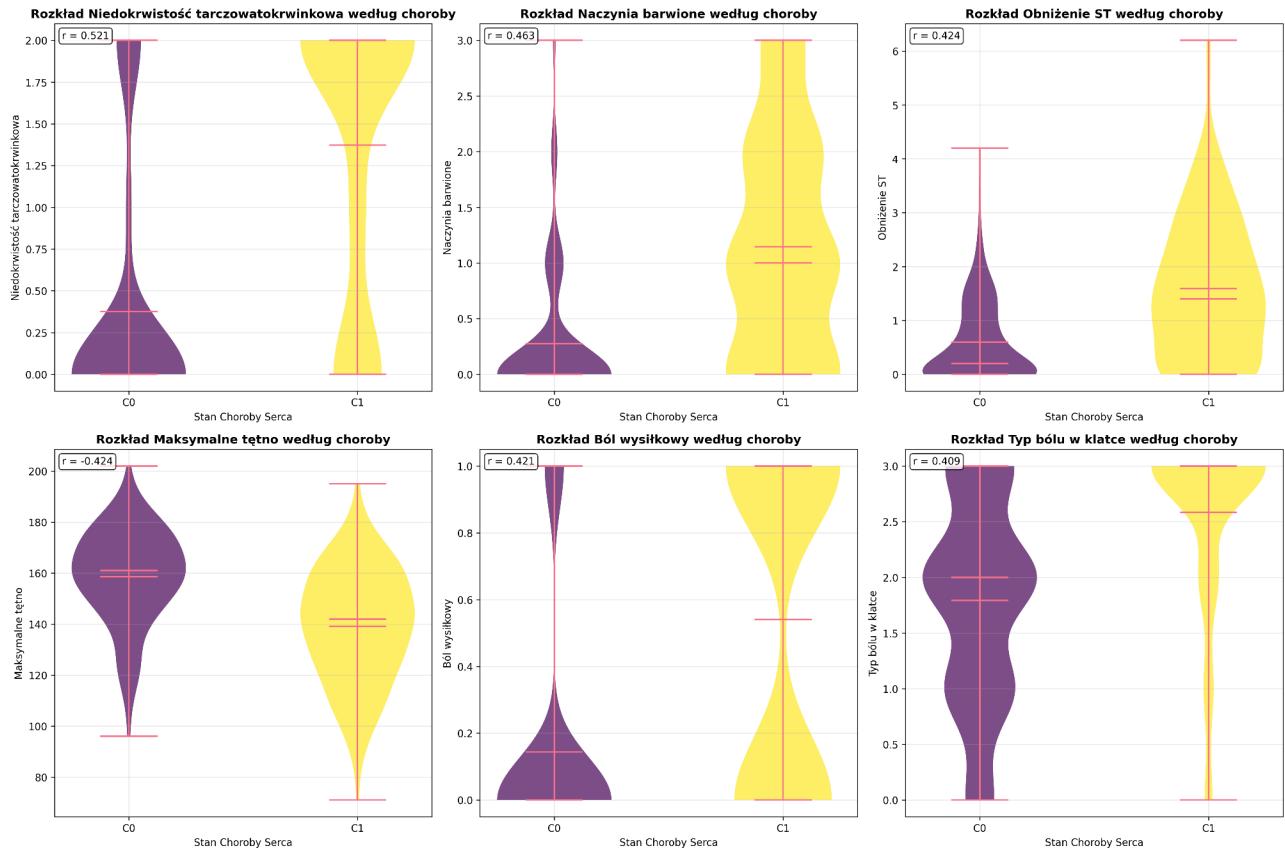
W rozkładach pięciu kluczowych zmiennych ciągłych zauważalna jest silna prawostronna asymetria Cholesterolu (chol) i Obniżenia ST (oldpeak), gdzie większość pacjentów ma niskie lub średnie wartości, ale występuje grupa z ekstremalnie wysokimi wynikami, co powoduje długi ogon po prawej stronie. Natomiast rozkład Wieku (age) oraz Maksymalnego tētna (thalach) jest

najbardziej zbliżony do symetrycznego, przy czym age wykazuje lekką lewostronną asymetrię, a thalach ma główną koncentrację wyników w zakresie 130-170. Z kolei Spoczynkowe ciśnienie krwi (trestbps) ma umiarkowaną prawostronną asymetrię, skupiając się w zakresie normatywnym, ale z widocznymi przypadkami nadciśnienia.



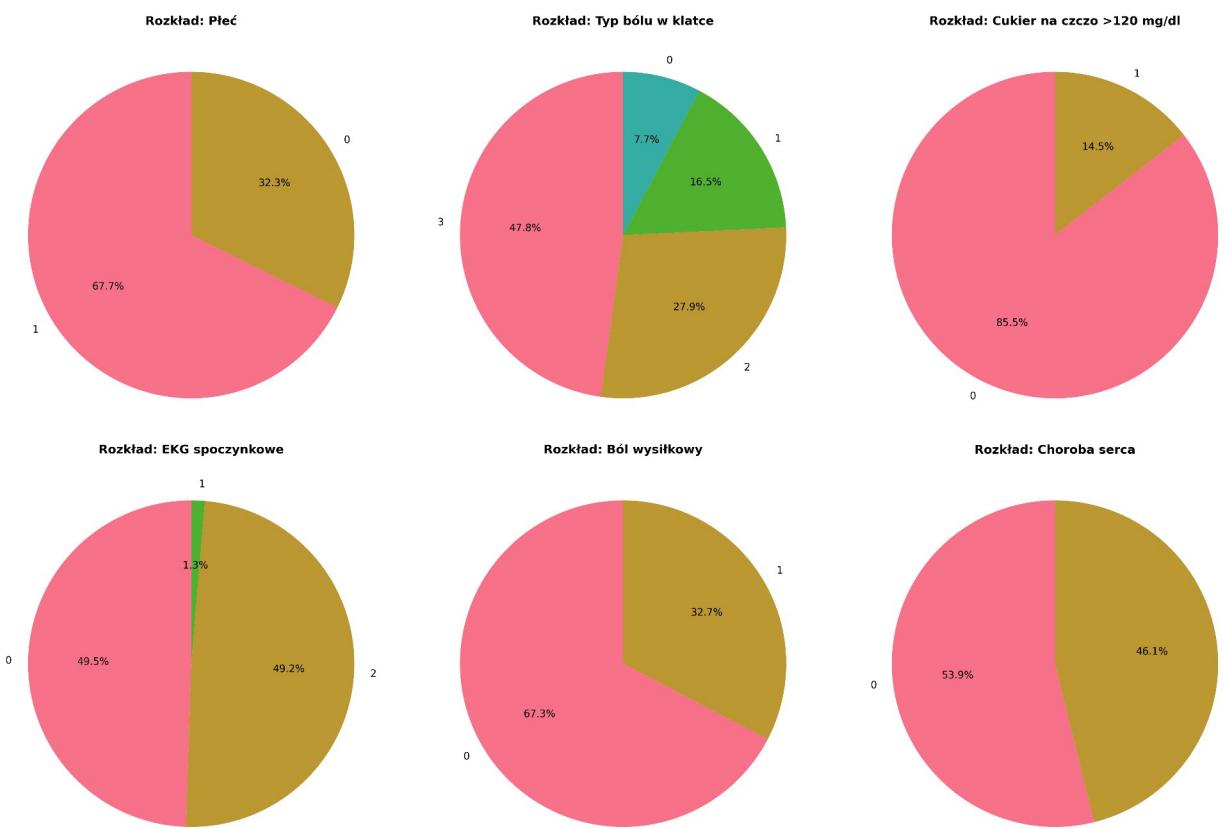
Rysunek 4.12. Wykresy Pudełkowe

Wizualna ocena danych za pomocą wykresów pudełkowych potwierdza obecność wartości odstających w czterech z pięciu atrybutów. Wykres dla Wieku (age) jest stabilny i nie posiada żadnych punktów poza wąsami. Pozostałe zmienne, w tym Spoczynkowe ciśnienie krwi (trestbps), Cholesterol (chol), Maksymalne tętno (thalach) i Obniżenie ST (oldpeak), wykazują występowanie wartości ekstremalnych, z których zwłaszcza pacjenci z ekstremalnie wysokim cholesterollem stanowią wyzwanie analityczne. Położenie mediany względem kwartyli w wykresach dla chol i oldpeak dodatkowo podkreśla prawostronną asymetrię tych rozkładów.



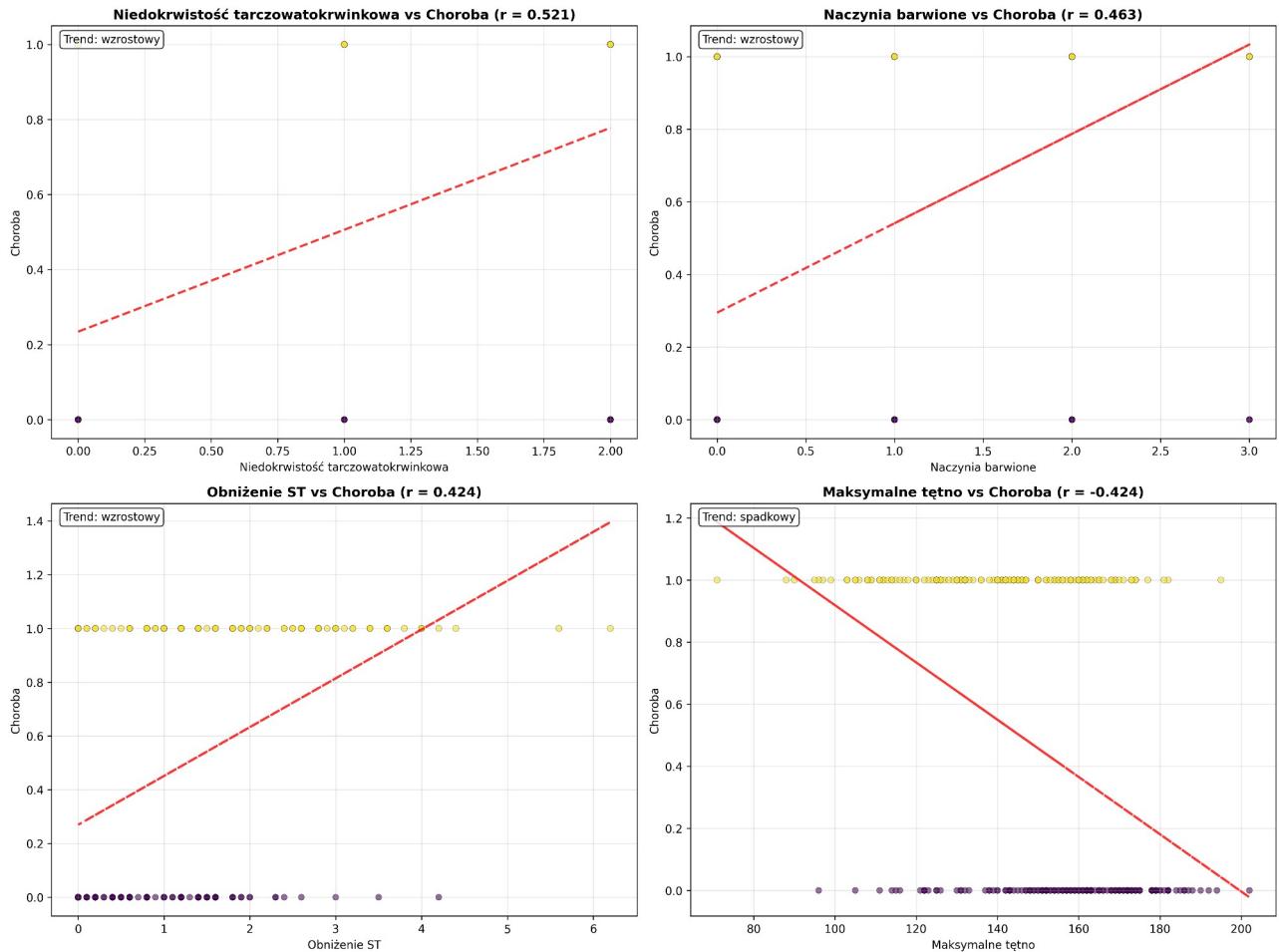
Rysunek 4.13. Wykresy Skrzypcowe

Wykresy skrzypcowe dostarczają szczegółowej informacji o gęstości rozkładu atrybutów. Dla **Cholesterolu** (chol) kształt jest najszerzy w dolnej części, co wskazuje na silną koncentrację danych w zakresie niskich i średnich wyników; górna część jest bardzo wąska i rozciągnięta, co wizualizuje rzadkie, ekstremalnie wysokie wartości. Rozkład **Maksymalnego tętna** (thalach) jest najbardziej regularny, z gęstością skoncentrowaną symetrycznie wokół mediany. Z kolei **Obniżenie ST** (oldpeak) ma swój kształt mocno ściśnięty przy wartości 0, a następnie gwałtownie zwężający się, co jest typowe dla rozkładów o silnej skośności.



Rysunek 4.14. Wykresy Kołowe

Analiza atrybutów kategorycznych za pomocą wykresów kołowych ujawnia znaczące niezbalansowanie w niektórych cechach demograficznych i klinicznych. Pacjenci to głównie mężczyźni (około 68%) i zdecydowana większość ma cukier na czczo poniżej 120 mg/dl oraz nie doświadcza dławicy piersiowej indukowanej wysiłkiem. Kluczowa zmienna docelowa, Stan/Diagnoza (condition), jest natomiast dobrze zbalansowana, z podziałem około 54% pacjentów zdrowych i 46% pacjentów z chorobą serca, co jest korzystne dla budowy modeli klasyfikacyjnych.



Rysunek 4.15. Wykresy Rozrzutu

Wykresy rozrzutu wizualizują zależności między kluczowymi atrybutami ciągłymi. Obserwacja relacji między Cholesterolem (chol) a Maksymalnym tętnem (thalach) ujawnia, że punkty są rozrzucone w sposób losowy i równomierny, co potwierdza brak istotnej korelacji liniowej między tymi dwiema cechami. Natomiast relacja między Wiekiem (age) a Maksymalnym tętnem (thalach) ukazuje wyraźny negatywny trend: im starszy jest pacjent, tym niższe maksymalne tętno jest w stanie osiągnąć. Wykres ten jest zgodny ze znaną wiedzą fizjologiczną i sugeruje, że age jest umiarkowanym predyktorem thalach.

5. Analiza eksploracyjna

5.1. Zbiór „Wine Quality”

5.1.1. Algorytm Random Forest

Poniższa tabela przedstawia wyniki dokładności, precyzji, czułości oraz miary F1 uzyskane dla pięciokrotnej walidacji krzyżowej.

Tabela 5.1. Uzyskane wyniki dla pięciokrotnej walidacji krzyżowej.

Miara	Wynik
Średnia dokładność	0.6797
Odchylenie standardowe dokładności	0.0152
Średnia precyzja	0.5730
Odchylenie standardowe precyzji	0.0635
Średnia czułość	0.3913
Odchylenie standardowe czułości	0.0318
Średnia miara F1	0.4277
Odchylenie standardowe miary F1	0.0362

Mimo stosunkowo wysokiej ogólnej dokładności (Accuracy $\approx 68\%$), niska wartość F1-score ($\approx 43\%$) ujawnia, że model nie radzi sobie z poprawną klasyfikacją rzadziej występujących ocen (klas mniejszościowych). Wyraźna dysproporcja między precyzją (57%) a czułością (39%) wskazuje, że model jest "konserwatywny" – rzadziej podejmuje ryzyko wskazania nietypowej oceny, przez co pomija znaczną część win bardzo dobrych lub bardzo słabych (bardzo niska czułość). Wyniki te jednoznacznie potwierdzają negatywny wpływ niebalansowania zbioru danych, gdzie dominacja klas środkowych (5 i 6) statystycznie "zagłusza" sygnały płynące z nielicznych próbek skrajnych. Kolejno wykonano optymalizację hiperparametrów aby znaleźć parametry trenowanego modelu, dla których metryki wydajnościowe będą miały najlepsze wyniki. Czas trenowania wynosił 0.663 sekundy. Najlepszymi parametrami okazały się:

- maksymalna głębokość drzewa – bez ograniczeń (None),
- maksymalna liczba cech brana pod uwagę przy podziale – pierwiastek z liczby wszystkich cech (sqrt),
- minimalna liczba próbek w liściu – 1,
- minimalna liczba próbek wymagana do dokonania podziału – 2,

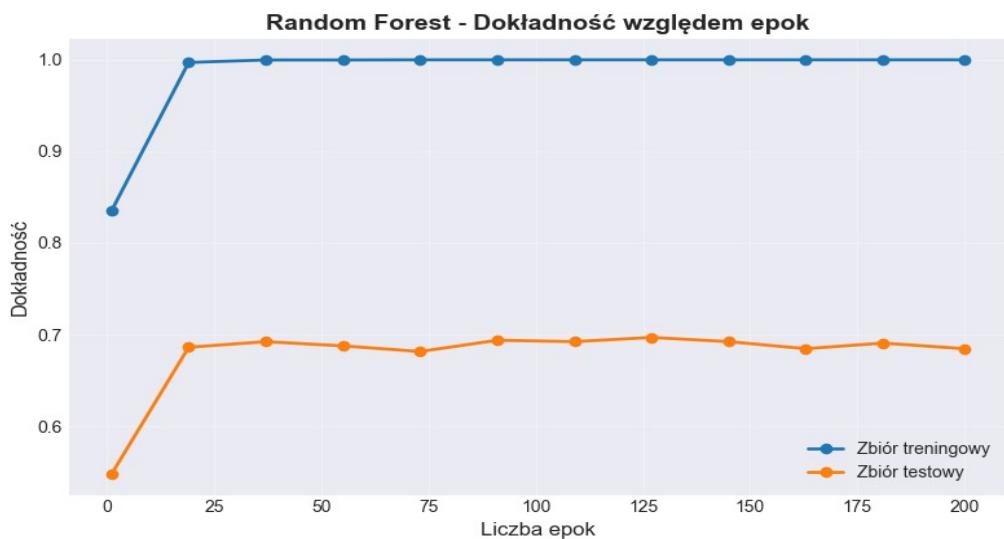
- liczba drzew w lesie – 200.

Poniższa tabela przedstawia uzyskane wartości metryk dla wskazanych parametrów.

Tabela 5.2. Wartości metryk uzyskane dla najlepszych parametrów modelu.

Metryka	Wartość
Średnia dokładność	0.6891
Średnia precyzyja	0.5441
Średnia czułość	0.3608
Średni błąd kwadratowy	0.4293
Średni błąd bezwzględny	0.3482

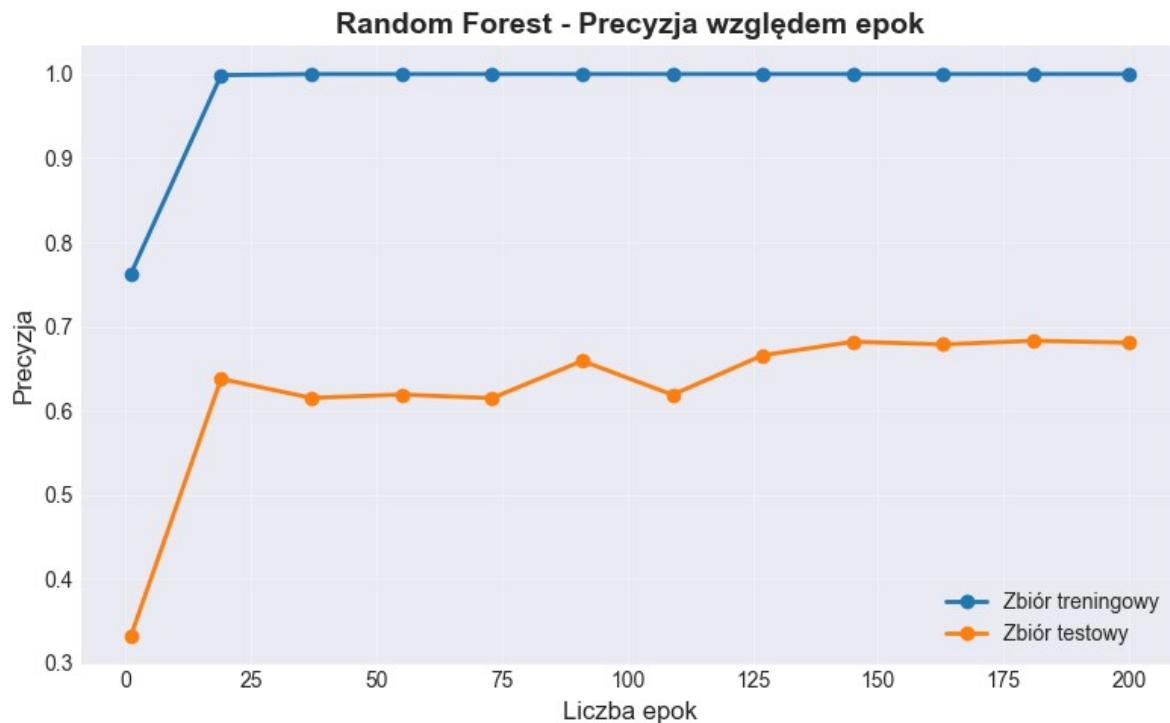
Dobór parametrów (brak ograniczeń głębokości, minimalna próbka w liściu = 1) wskazuje na budowę silnie dopasowanego, złożonego modelu, który przy 200 estymatorach osiąga stabilną dokładność na poziomie blisko 69%. Choć niska czułość (36%) potwierdza, że model wciąż ma trudności z "trafieniem" w rzadkie oceny skrajne, to bardzo niski błąd bezwzględny (MAE = 0,35) jest kluczowym, pozytywnym sygnałem. Oznacza on, że nawet gdy model się myli, jego pomyłki są minimalne i zazwyczaj oscylują zaledwie o +/- 1 stopień na skali jakości (np. myli ocenę 5 z 6), co w praktycznym zastosowaniu czyni go użytecznym narzędziem wspomagającym. Poniżej przedstawiono wykresy zależności liczby epok od uzyskanych parametrów.



Rysunek 5.1. Wykres zależności dokładności względem liczby epok.

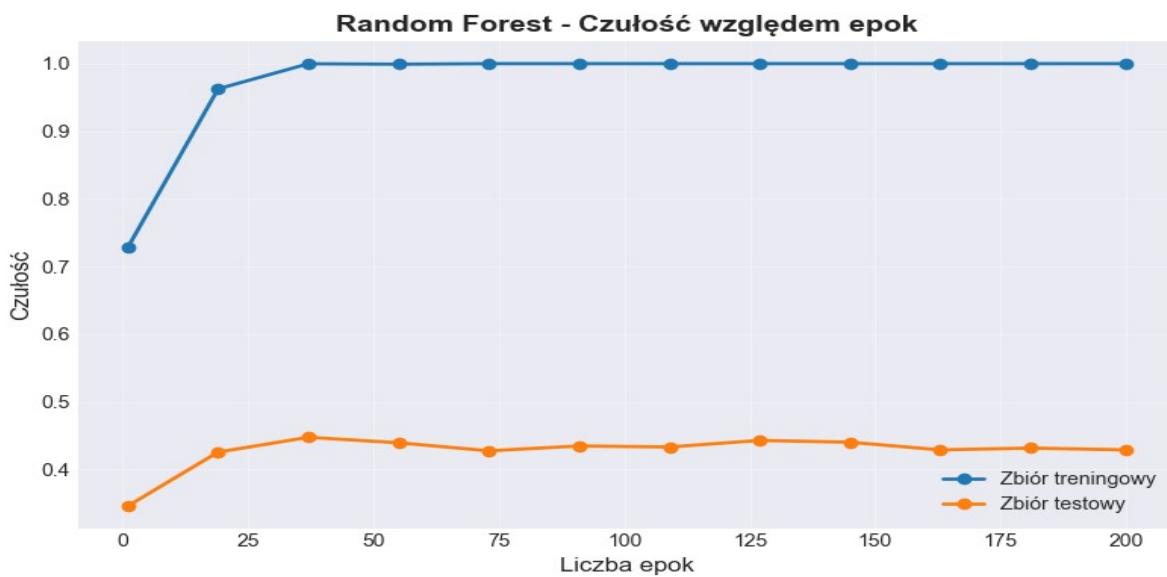
Z powyższego wykresu można wywnioskować, że ogólna skuteczność modelu wykazuje szybką zbieżność. Maksymalna dokładność na zbiorze testowym, wynosząca około 69%, jest osiągana już w przedziale 20–25 epoki. W dalszym przebiegu eksperymentu (do 200 epok) wartość

ta pozostaje na stałym poziomie. Świadczy to o tym, że optymalne parametry predykcyjne uzyskiwane są we wczesnej fazie treningu, a kontynuacja procesu obliczeniowego nie przekłada się na redukcję błędu generalizacji. Poniższy wykres przedstawia zależność wartości precyzji od liczby epok.



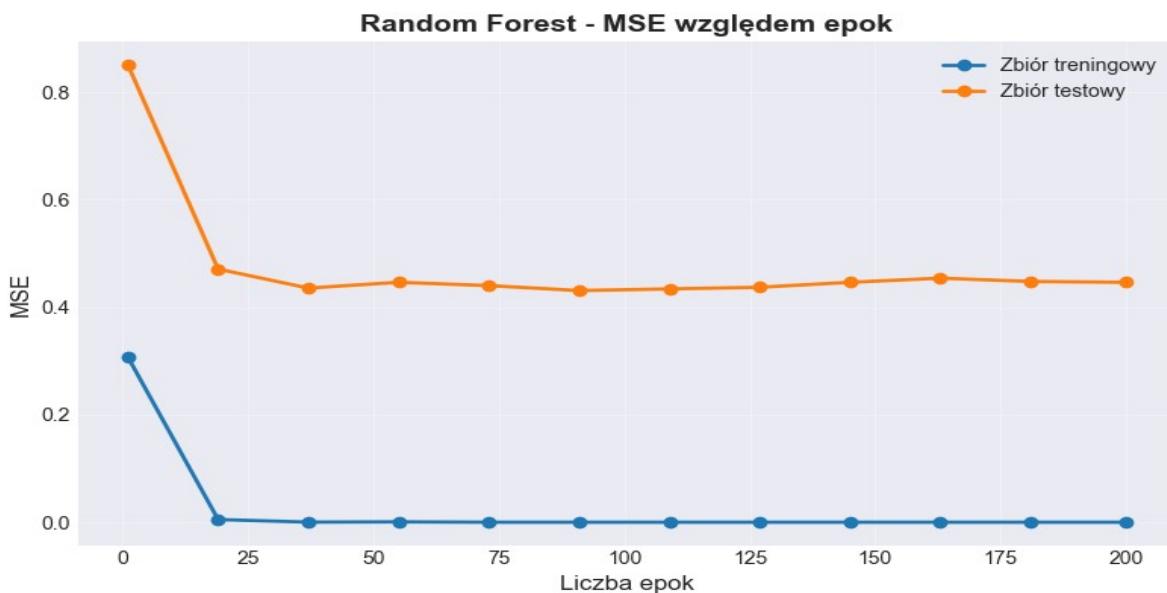
Rysunek 5.2. Wykres zależności precyzji względem liczby epok.

Powyższy wykres obrazuje dynamikę zmian precyzji w trakcie procesu uczenia. W początkowej fazie (do ok. 25. epoki) obserwuje się gwałtowny wzrost wartości metryki dla zbioru testowego, która następnie stabilizuje się na poziomie ok. 68%. Dalsze zwiększanie liczby epok (w przedziale 25–200) nie skutkuje istotną poprawą wyników, co wskazuje na osiągnięcie przez model maksymalnej zdolności dyskryminacyjnej przy relatywnie niewielkiej liczbie estymatorów. Poniższy wykres przedstawia zależność czułości względem numeru epoki.



Rysunek 5.3. Wykres zależności czułości względem liczby epok.

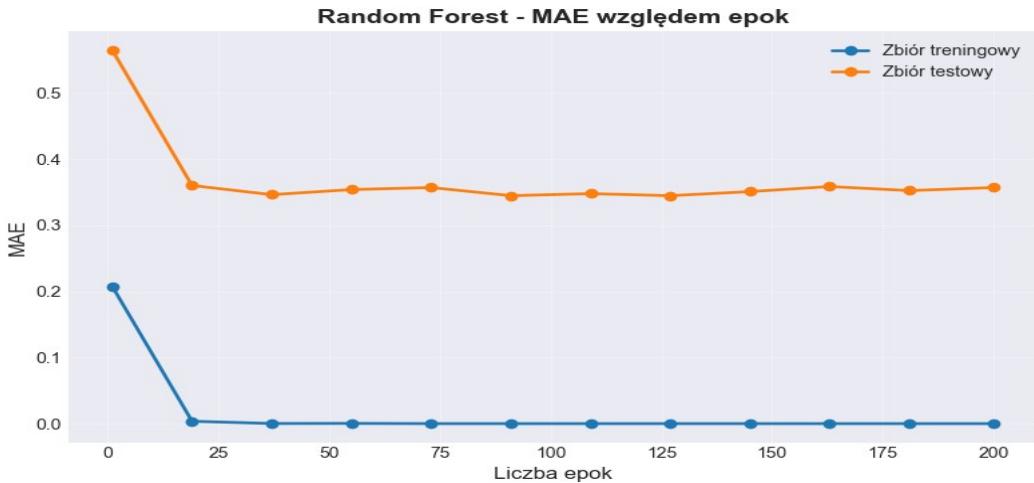
Przebieg krzywej czułości zawartej na powyższym rysunku charakteryzuje się szybkim nasyceniem. Po fazie wzrostowej w pierwszych 20 epokach, metryka osiąga poziom asymptotyczny w granicach 43% na zbiorze testowym. Dalszy proces treningowy nie przynosi przyrostu czułości, co sugeruje, że model wyczerpał swoje możliwości generalizacji w zakresie identyfikacji wszystkich instancji klas, a problem niskiej czułości wynika prawdopodobnie z charakterystyki danych, a nie ze zbyt krótkiego czasu uczenia. Na poniższym rysunku przedstawiono zależność błędu średniokwadratowego w zależności od numeru epoki.



Rysunek 5.4. Wykres zależności średniego błędu kwadratowego względem liczby epok.

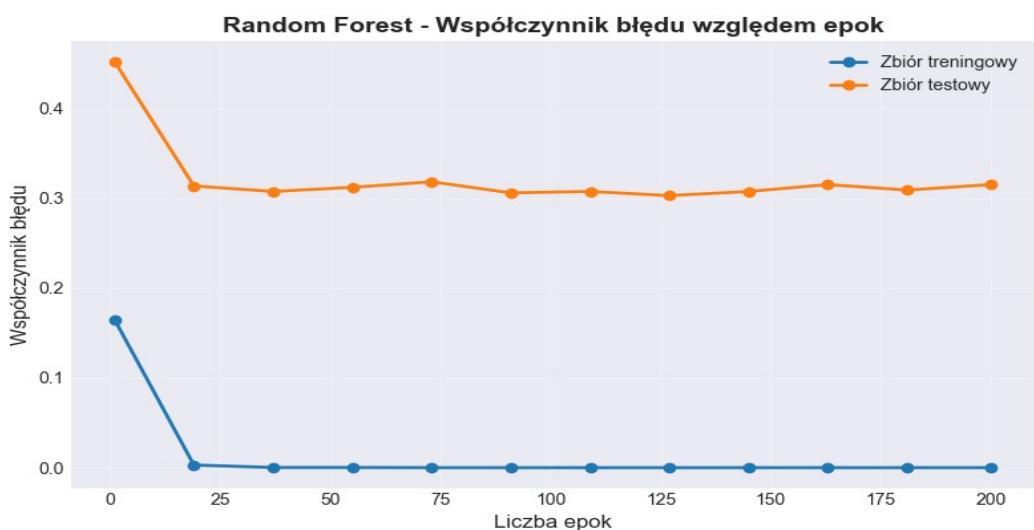
Na powyższym wykresie widoczne jest, iż po około 40 epokach wartość błędu średniokwadratowego na zbiorze testowym zatrzymuje się w przedziale 0.43-0.45 i pozostaje niewrażliwa na dalszy przebieg uczenia. Widoczna, duża i stała odległość między linią niebieską

(trening) a pomarańczową (test) przez wszystkie epoki sygnalizuje mocne dopasowanie do danych treningowych (overfitting). Poniższy rysunek przedstawia zależność średniego błędu bezwzględnego względem liczby epok.



Rysunek 5.5. Wykres zależności średniego błędu bezwzględnego względem liczby epok.

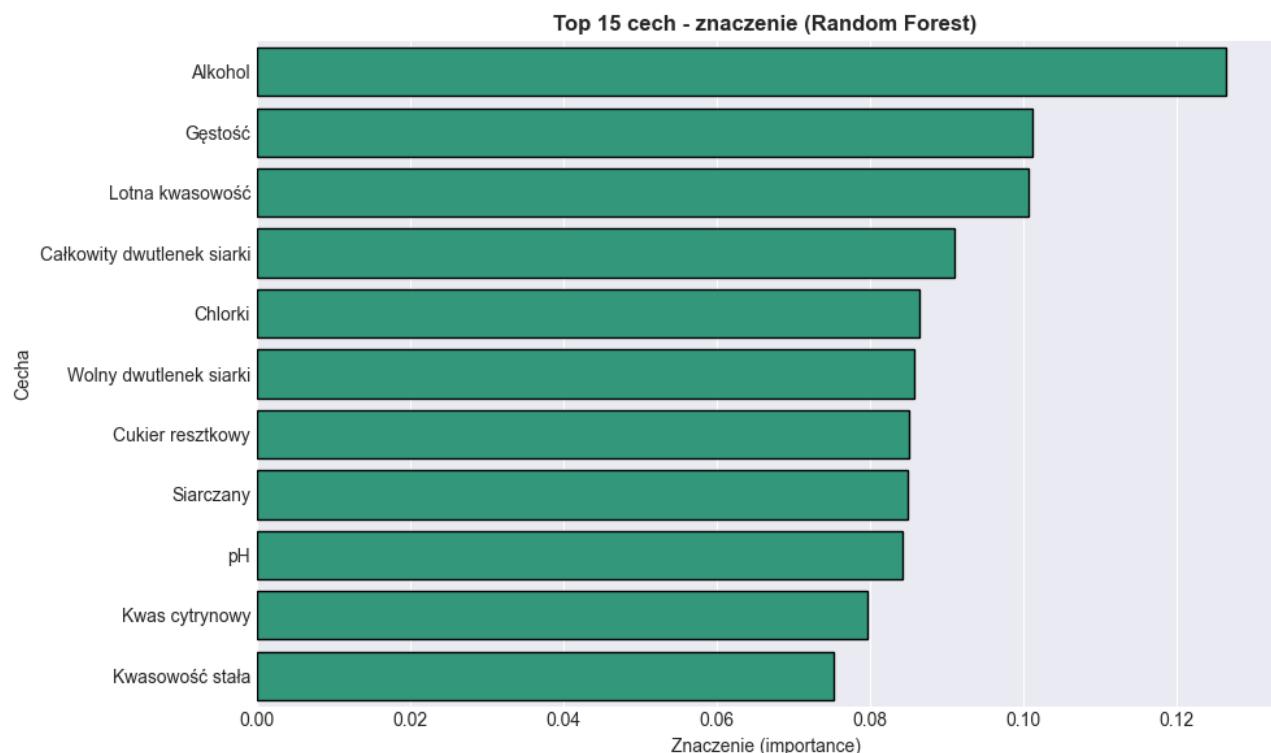
Z powyższego rysunku można odczytać, że Wstępna faza uczenia (pierwsze 25 epok) przynosi drastyczną redukcję średniego błędu na zbiorze testowym do poziomu ok. 0.35. Po tym punkcie krzywa staje się płaska. Oznacza to, że wydłużanie treningu o kolejne epoki nie sprawia, że model myli się mniej – błąd predykcji oceny pozostaje stały i wynosi średnio ok. 1/3 punktu. Poniższy rysunek przedstawia wykres zależności współczynnika błędu uczenia/testowania względem numeru epoki.



Rysunek 5.6. Wykres zależności błędu uczenia/testowania względem liczby epok.

Powyższy wykres pokazuje dynamikę uczenia się modelu. Błąd na zbiorze testowym (linia pomarańczowa) gwałtownie spada w pierwszych 20-25 epokach, po czym następuje stabilizacja na poziomie ok. 0.31. Dalsze zwiększanie liczby epok (nawet do 200) nie przynosi już istotnej

poprawy, co sugeruje, że model bardzo szybko osiąga swoje maksimum możliwości predykcyjnych dla tych danych. Poniższy rysunek przedstawia 15 najważniejszych cech w zbiorze.



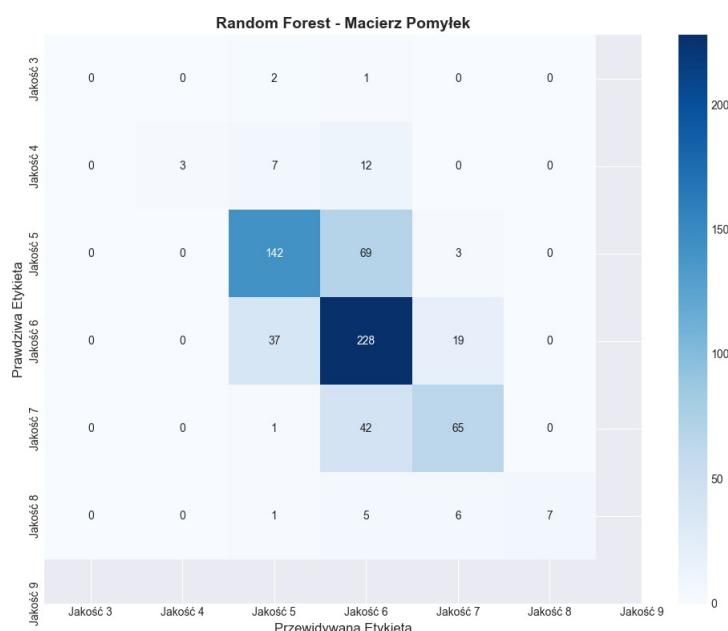
Rysunek 5.7. Piętnaście najważniejszych cech w zbiorze „Wine Quality” uzyskanych w procesie uczenia za pomocą algorytmu Random Forest.

Analiza ważności cech jednoznacznie wskazuje na alkohol jako dominujący predyktor, którego wpływ na decyzje modelu ($waga > 0,12$) znacznie przewyższa pozostałe parametry. Na kolejnych miejscach uplasowała się gęstość oraz kwasowość lotna, co potwierdza ich kluczową rolę w kształtowaniu jakości wina, zidentyfikowaną już wcześniej w analizie korelacji liniowej. Istotnym odkryciem jest wysoka pozycja całkowitego dwutlenku siarki, co dowodzi, że model Lasu Losowego skutecznie wykorzystuje nieliniowe zależności, które w prostych metodach statystycznych wydawały się nieistotne. Stosunkowo wyrównany poziom ważności dla pozostałych zmiennych (od chlorków po pH) sugeruje, że ostateczna klasyfikacja jest wynikiem złożonej interakcji wielu parametrów chemicznych, a nie oparciem się wyłącznie na kilku wiodących cechach. Kolejno wykonano testy modelu na zbiorze testowym. Poniższe rysunki oraz tabele przedstawiają uzyskane wyniki testów.

Tabela 5.3. Uzyskane wartości metryk dla zbioru testowego.

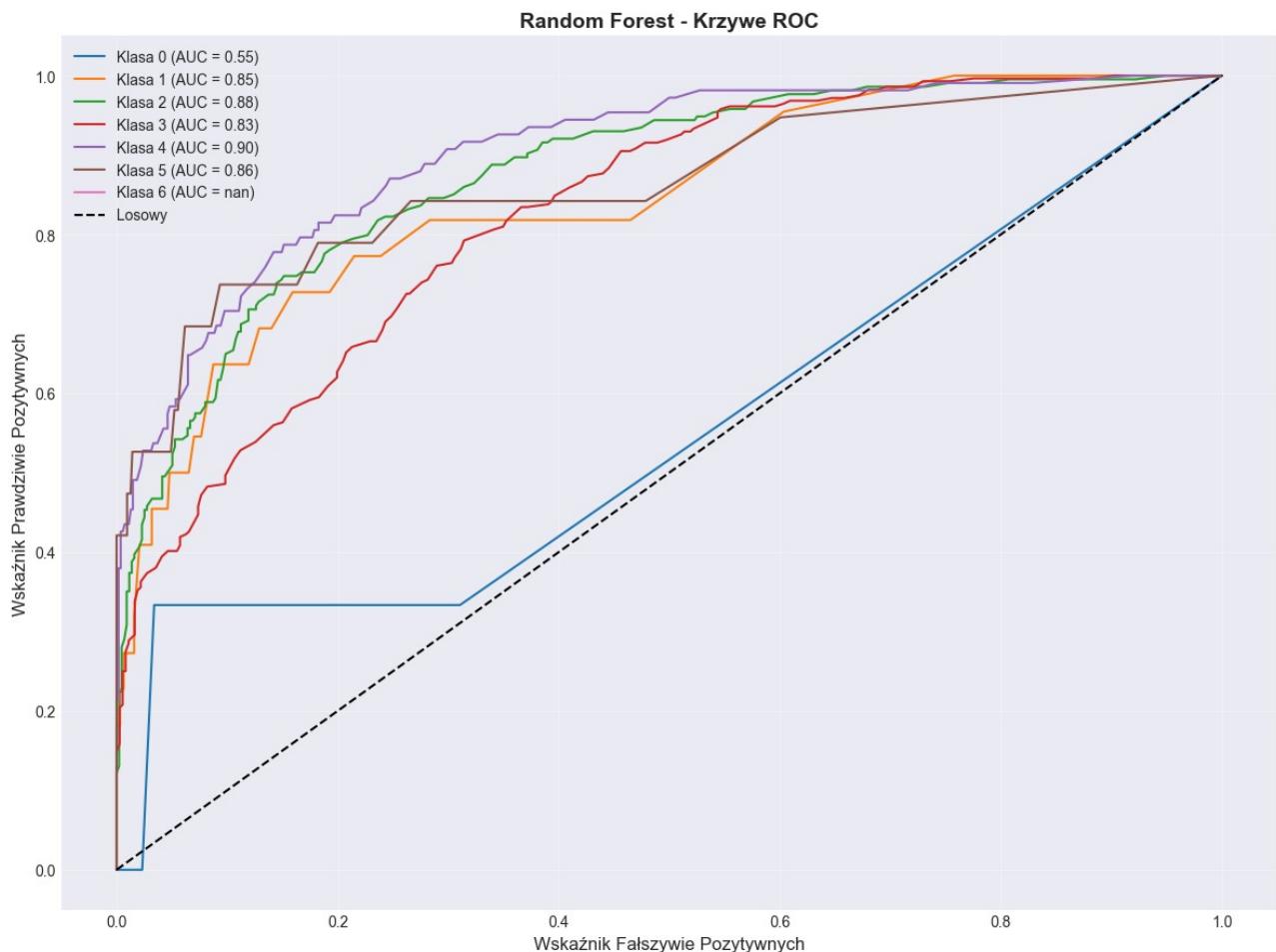
Metryka	Wartość
Dokładność	0.6846
Precyza	0.6808
Czułość	0.4288
Średni błąd kwadratowy	0.4462
Średni błąd bezwzględny	0.3569
Pierwiastek ze średniego błędu kwadratowego	0.6679

Ostateczna weryfikacja na zbiorze testowym potwierdza zdolność generalizacji modelu, który osiągnął dokładność na poziomie 68,46%, przy zbliżonej wartości precyzyji (68,08%). Istotna dysproporcja między wysoką precyją a niską czułością (42,88%) wskazuje jednak na konserwatywną charakterystykę klasyfikatora, który ma trudności z kompletną detekcją przypadków w klasach niedoreprezentowanych, minimalizując ryzyko fałszywych alarmów kosztem pominięcia niektórych trafień. Należy jednak podkreślić bardzo korzystny wynik średniego błędu bezwzględnego (MAE = 0,3569), który dowodzi, że pomyłki modelu są niewielkie i zazwyczaj nie przekraczają jednego stopnia na skali jakości. Wartość pierwiastka z błędem średniokwadratowym (RMSE = 0,6679) utrzymująca się poniżej jedności dodatkowo potwierdza, że model unika rażących błędów (odchyleń o 2 lub więcej klas), co czyni go wiarygodnym narzędziem wspomagającym ocenę enologiczną. Poniższy rysunek przedstawia uzyskaną macierz pomyłek.



Rysunek 5.8. Macierz pomyłek dla zbioru „Wine Quality” oraz algorytmu Random Forest.

Macierz pomyłek jednoznacznie obrazuje wpływ niezbalansowania zbioru danych na proces decyzyjny modelu. Obserwuje się silną koncentrację poprawnych klasyfikacji w obrębie klas większościowych (oceny 5 i 6), które stanowią dominantę w zbiorze treningowym. Kluczową charakterystyką błędów jest ich lokalność – pomyłki niemal wyłącznie oscylują wokół głównej przekątnej (różnica ± 1 klasy), co wskazuje, że model poprawnie identyfikuje ogólny trend jakości, a błędy wynikają z subtelnych różnic fizykochemicznych między sąsiadującymi ocenami. Całkowity brak predykcji dla klas skrajnych (3 i 9) potwierdza, że model nie wykształcił zdolności dyskryminacyjnych dla próbek niedoreprezentowanych. Poniższy rysunek prezentuje krzywe ROC-AUC.



Rysunek 5.9. Krzywe ROC-AUC dla zbioru „Wine Quality” oraz algorytmu Random Forest.

Przebieg krzywych ROC dla większości klas (szczególnie 4, 5, 7 i 8) wykazuje wysoki potencjał separacyjny modelu, z wartościami AUC w przedziale 0,83–0,90. Oznacza to, że model skutecznie rankinguje prawdopodobieństwa przynależności do tych klas, mimo niższej czułości wynikającej z doboru progu odcienia. Wyraźnym odstępstwem jest klasa 3 (AUC = 0,55), dla której zdolność dyskryminacyjna modelu jest bliska losowej, co stanowi bezpośredni skutek

niewystarczającej liczby próbek treningowych dla tej kategorii. Poniższa tabela prezentuje wartość metryk dla wskazanych etykiet atrybutu wyjściowego dla zbioru testowego.

Tabela 5.4. Wartość metryk dla wskazanych etykiet atrybutu wyjściowego dla zbioru testowego.

Etykietą	Precyza	Czułość	Wartość miary F1
Jakość 3	0.00	0.00	0.00
Jakość 4	1.00	0.14	0.24
Jakość 5	0.75	0.66	0.70
Jakość 6	0.64	0.80	0.71
Jakość 7	0.70	0.60	0.65
Jakość 8	1.00	0.37	0.54
Jakość 9	0.00	0.00	0.00

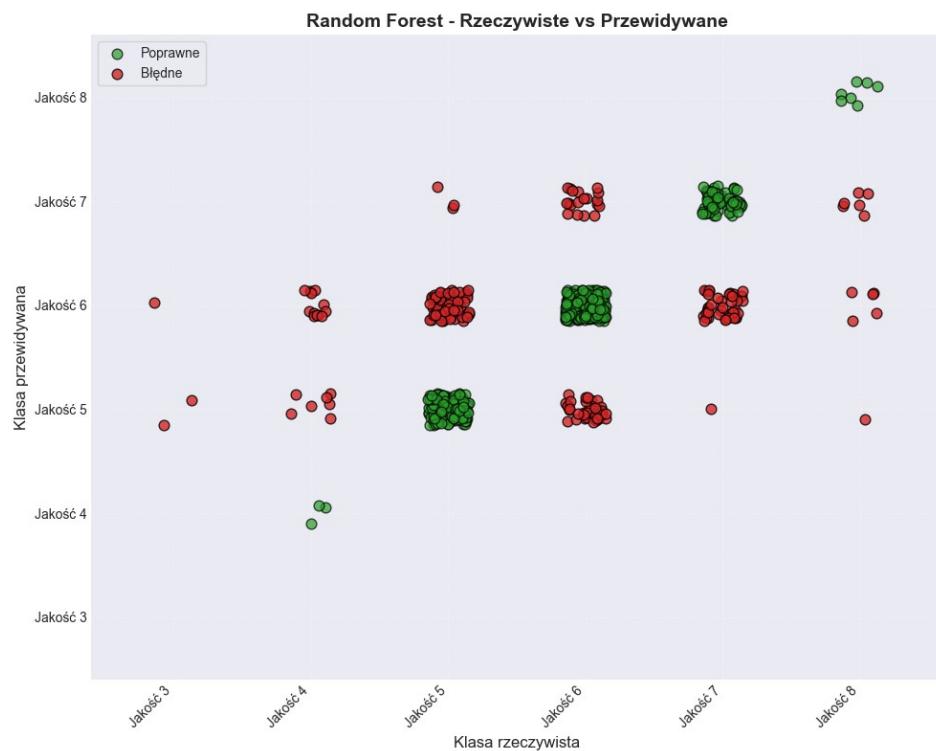
Zestawienie metryk dla poszczególnych etykiet ujawnia dychotomię w skuteczności modelu. Klasy środkowe (5, 6, 7) charakteryzują się wysoką stabilnością, osiągając zrównoważone wartości miary F1 (powyżej 0,65). W przypadku klas rzadkich (4 i 8) model przyjmuje strategię konserwatywną, co objawia się maksymalną precyza (1.00) przy jednoczesnej niskiej czułości – model dokonuje predykcji tych klas rzadko, lecz z bardzo wysoką pewnością. Zerowe wartości metryk dla klas 3 i 9 wskazują na konieczność zastosowania technik oversamplingu w przyszłych pracach w celu poprawy detekcji anomalii. Poniższa tabela prezentuje przykładowe predykcje wytrenowanego modelu.

Tabela 5.5. Przykładowe predykcje uzyskane dla zbioru „Wine Quality” oraz algorytmu Random Forest.

Klasa rzeczywista	Klasa przewidywana	Poprawna	Prawdopodobieństwo
Jakość 6	Jakość 7	Nie	0.475
Jakość 5	Jakość 5	Tak	0.450
Jakość 4	Jakość 4	Tak	0.690
Jakość 6	Jakość 6	Tak	0.610
Jakość 7	Jakość 7	Tak	0.495
Jakość 6	Jakość 6	Tak	0.465

Jakość 6	Jakość 6	Tak	0.565
Jakość 6	Jakość 6	Tak	0.345
Jakość 7	Jakość 7	Tak	0.860
Jakość 6	Jakość 5	Nie	0.560
Jakość 6	Jakość 6	Tak	0.405
Jakość 6	Jakość 6	Tak	0.630
Jakość 6	Jakość 6	Tak	0.855
Jakość 7	Jakość 6	Nie	0.580
Jakość 7	Jakość 6	Nie	0.630

Analiza poziomu pewności (prawdopodobieństwa) predykcji wskazuje, że model w przypadkach poprawnych klasyfikacji często osiąga wysokie wskaźniki pewności ($> 0,80$). W sytuacjach błędnych, prawdopodobieństwa są zazwyczaj niższe i bardziej rozmyte (w granicach 0,45–0,60), co sugeruje, że błędne decyzje zapadają w obszarach o dużej niejednoznaczności granic decyzyjnych (decision boundaries), gdzie charakterystyki chemiczne sąsiadujących klas jakościowych (np. 6 i 7) znaczaco się pokrywają. Poniższy rysunek przedstawia wizualizację przewidywań klas w modelu.



Rysunek 5.10. Wizualizacja predykcji modelu utworzonego ze zbioru „Wine Quality” oraz algorytmu Random Forest.

Wizualizacja predykcji potwierdza niski poziom błędu średniego bezwzględnego (MAE \approx 0,36). Rozkład punktów błędnych (kolor czerwony) wykazuje małą wariancję względem wartości referencyjnych, co oznacza, że model unika rażących pomyłek (np. klasyfikacji wina słabego jako bardzo dobrego). "Puste" strefy dla wartości skrajnych na osi Y (przewidywane) korelują z wnioskami z macierzy pomyłek, ilustrując tendencję modelu do "bezpiecznego" uśredniania wyników w kierunku centrum rozkładu.

5.1.2. Algorytm XGBoost

Poniższa tabela przedstawia wyniki dokładności, precyzji, czułości oraz miary F1 uzyskane dla pięciokrotnej walidacji krzyżowej.

Tabela 5.6. Uzyskane wyniki dla pięciokrotnej walidacji krzyżowej.

Miara	Wynik
Średnia dokładność	0.6634
Odchylenie standardowe dokładności	0.0138
Średnia precyzja	0.5276
Odchylenie standardowe precyzji	0.0601
Średnia czułość	0.3970
Odchylenie standardowe czułości	0.0347
Średnia miara F1	0.4279
Odchylenie standardowe miary F1	0.0390

Wyniki walidacji krzyżowej ujawniają stabilność modelu, ale także jego ograniczenia wynikające z nierównomiernego rozkładu klas. Średnia dokładność na poziomie 66,34% przy niskim odchyleniu standardowym (1,38 p.p.) świadczy o powtarzalności wyników pomiędzy różnymi podziałami zbioru danych. Jednakże, istotna różnica między dokładnością a średnią miarą F1 (42,79%) wskazuje, że wysoki wynik ogólny jest "napompowany" przez dobre rozpoznawanie klas większościowych. Stosunkowo wysokie odchylenie standardowe dla precyzji (6,01 p.p.) sugeruje, że model w niektórych iteracjach walidacji radził sobie zauważalnie gorzej z minimalizacją fałszywych alarmów (False Positives). Czas trenowania wynosił 1.321 sekundy. Najlepszymi parametrami okazały się:

- procent cech losowo wybieranych do zbudowania każdego drzewa – 0,8,
- współczynnik uczenia – 0,1,
- maksymalna głębokość pojedynczego drzewa – 9,

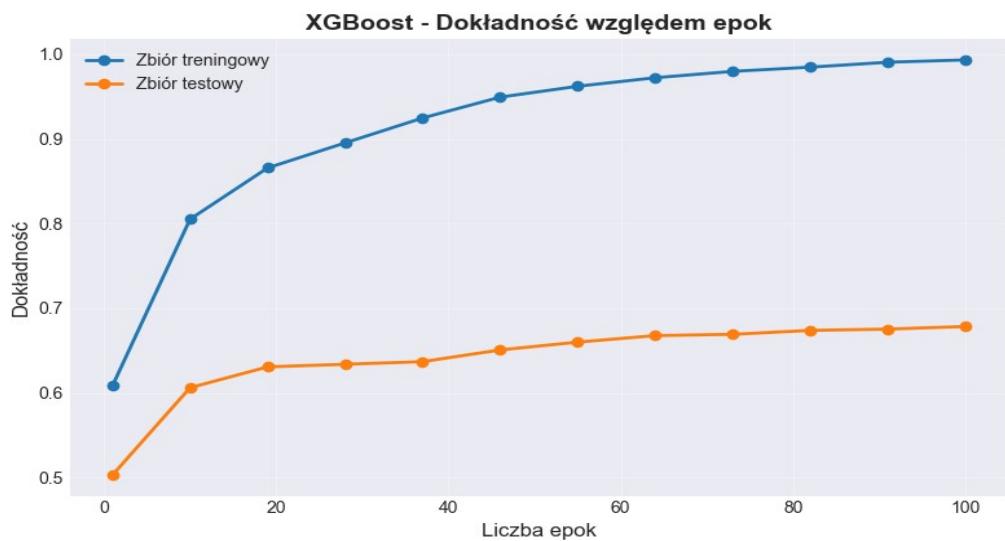
- liczba drzew w lesie – 100,
- ułamek próbek treningowych losowany do zbudowania każdego drzewa – 0.8.

Poniższa tabela przedstawia uzyskane wartości metryk dla wskazanych parametrów.

Tabela 5.7. Wartości metryk uzyskane dla najlepszych parametrów modelu.

Metryka	Wartość
Średnia dokładność	0.6836
Średnia precyzyja	0.4984
Średnia czułość	0.3581
Średni błąd kwadratowy	0.4438
Średni błąd bezwzględny	0.3566

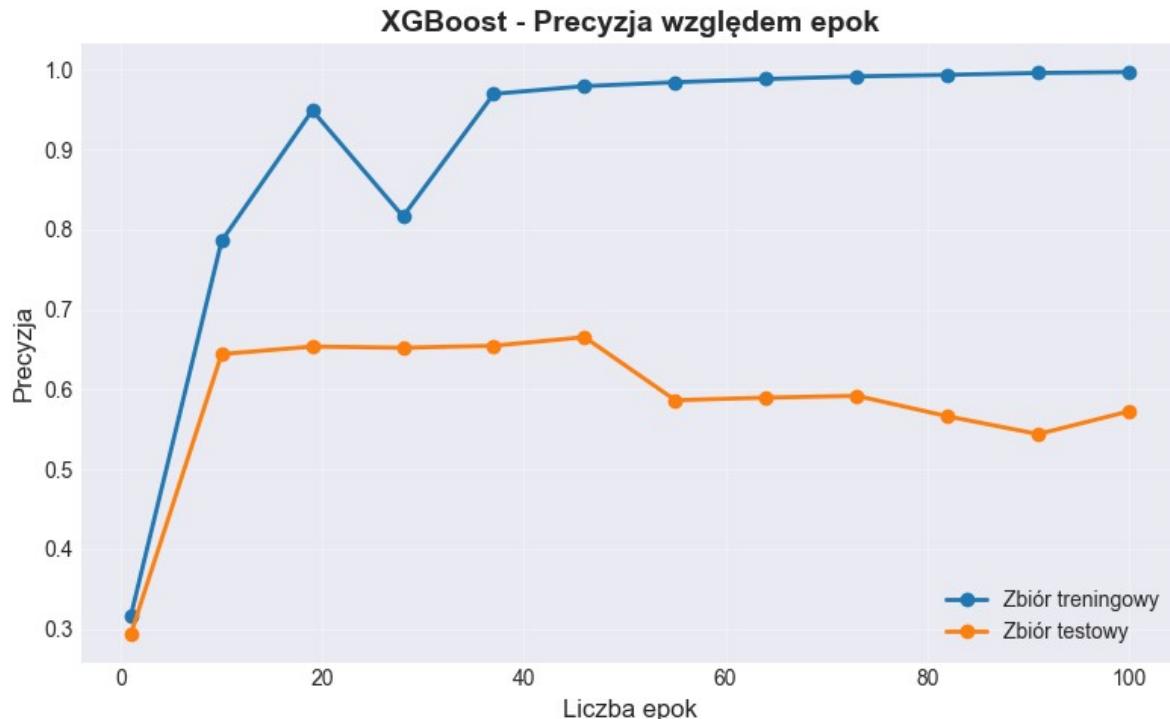
Proces optymalizacji hiperparametrów pozwolił na podniesienie skuteczności modelu. Po dostrojeniu (m.in. głębokość drzewa = 9, learning rate = 0.1), średnia dokładność wzrosła do 68,36%. Najważniejszym wskaźnikiem jest tutaj niski średni błąd bezwzględny (MAE = 0,3566). Oznacza on, że model po optymalizacji "kalibruje" swoje predykcje tak, aby średnie odchylenie od rzeczywistej oceny wynosiło zaledwie około 1/3 punktu, co jest wynikiem bardzo zadowalającym w kontekście praktycznym. Poniżej przedstawiono wykresy zależności liczby epok od uzyskanych parametrów.



Rysunek 5.11. Wykres zależności dokładności względem liczby epok.

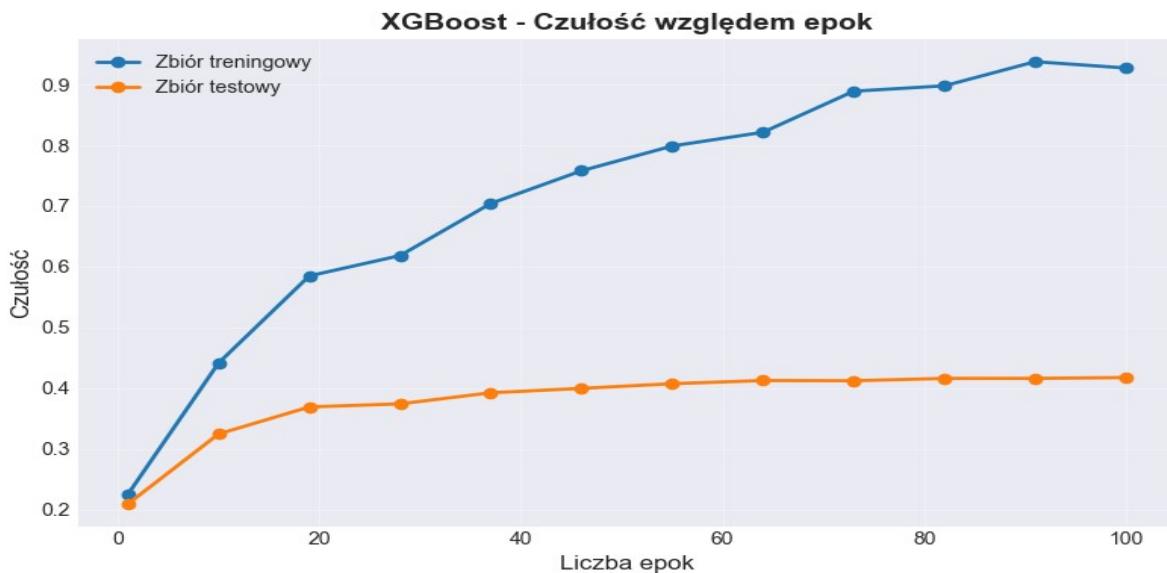
Analiza przebiegu uczenia wskazuje na bardziej stopniowy charakter przyrostu wiedzy w porównaniu do Lasów Losowych. Dokładność na zbiorze testowym rośnie sukcesywnie, osiągając stabilizację dopiero w okolicy 60–80 epoki na poziomie ok. 68%. Brak gwałtownego skoku w

początkowej fazie i powolny wzrost w późniejszych iteracjach sugeruje, że algorytm boostingowy (XGBoost) efektywnie koryguje błędy poprzedników w dłuższym horyzoncie czasowym, a proces trenowania przez 100 epok jest uzasadniony. Poniższy wykres przedstawia zależność wartości precyzji od liczby epok.



Rysunek 5.12. Wykres zależności precyzji względem liczby epok.

Wykres precyzji ujawnia interesującą dynamikę – po początkowym wzroście do poziomu ok. 65% (ok. 20-40 epoka), następuje lekki spadek i stabilizacja w granicach 57-60% w końcowych fazach treningu. Jest to zachowanie odmienne od Lasu Losowego i może wskazywać, że w miarę jak model stara się poprawić czułość (wykryć więcej trudnych przypadków), nieznacznie traci na pewności swoich predykcji, generując nieco więcej fałszywych alarmów (False Positives). Poniższy wykres przedstawia zależność czułości względem numeru epoki.



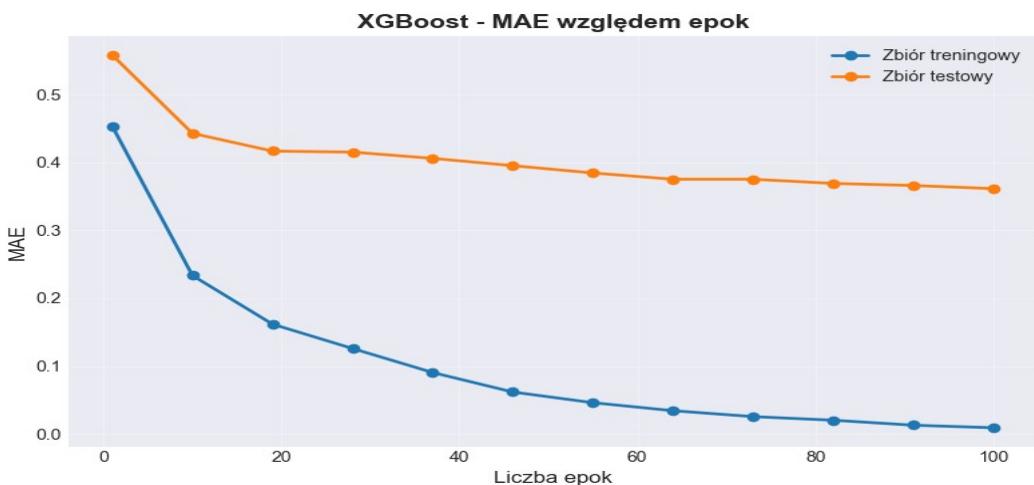
Rysunek 5.13. Wykres zależności czułości względem liczby epok.

W przeciwieństwie do modelu Random Forest, krzywa czułości dla XGBoost wykazuje trend rosnący przez niemal cały okres treningu, nie osiągając tak szybkiego nasycenia. Czułość na zbiorze testowym wzrasta z poziomu 20% do ponad 41% przy setnej epoce. Oznacza to, że boosting (iteracyjne uczenie na błędach) pozwala modelowi z każdą epoką coraz skuteczniej "wyławiać" próbki, które wcześniej były pomijane, choć odbywa się to kosztem wspomnianego wcześniej spadku precyzji. Na poniższym rysunku przedstawiono zależność błędu średniokwadratowego w zależności od numeru epoki.



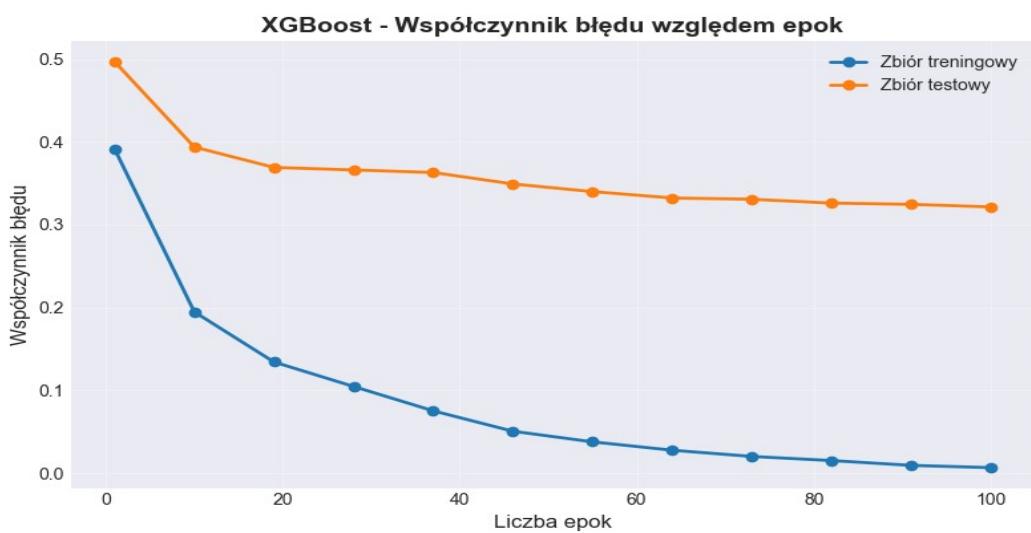
Rysunek 5.14. Wykres zależności średniego błędu kwadratowego względem liczby epok.

Wykres MSE potwierdza efektywność procesu minimalizacji funkcji straty. Błąd na zbiorze testowym spada monotonicznie do około 80. epoki, osiągając minimum na poziomie ~0.45. Warto zauważyć, że po przekroczeniu tego punktu krzywa testowa ulega delikatnemu wypłaszczeniu, podczas gdy błąd treningowy (linia niebieska) nadal dąży do zera, co jest sygnałem ostrzegawczym przed postępującym przeuczeniem (overfitting) w przypadku dalszego zwiększania liczby estymatorów. Poniższy rysunek przedstawia zależność średniego błędu bezwzględnego względem liczby epok.



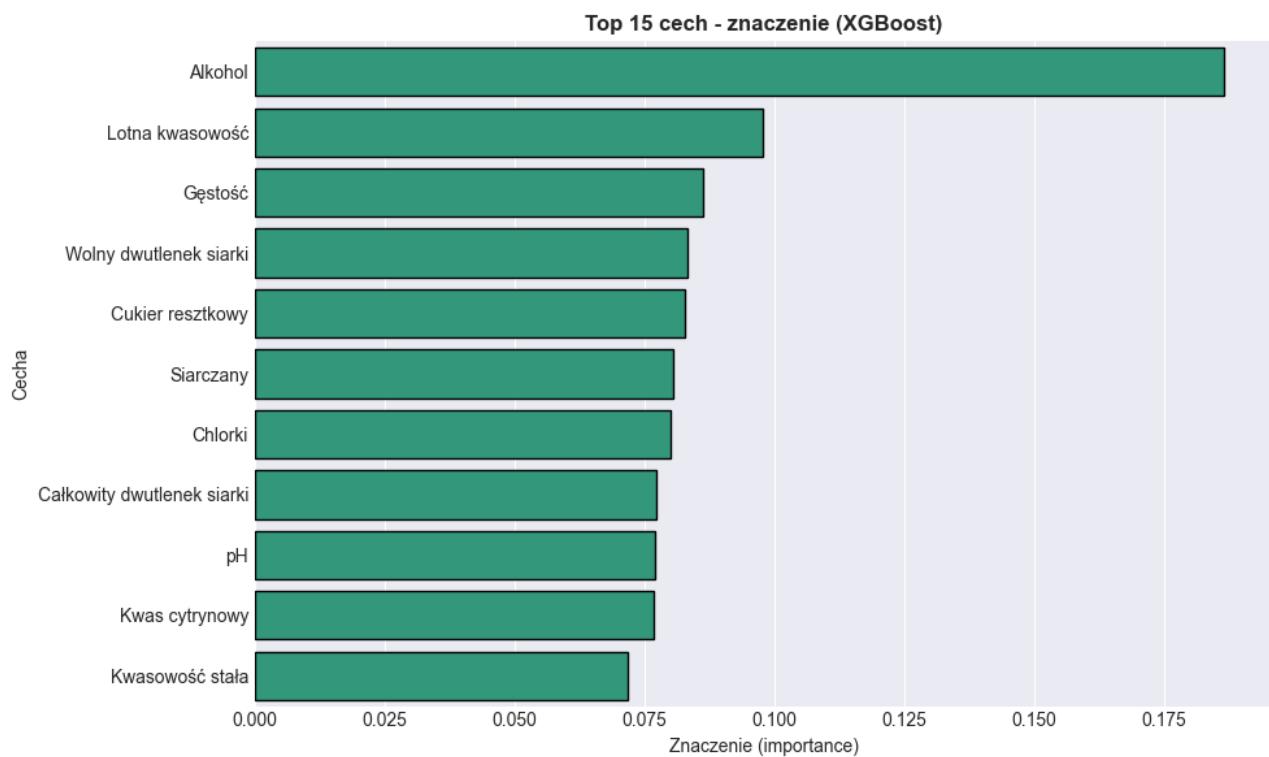
Rysunek 5.15. Wykres zależności średniego błędu bezwzględnego względem liczby epok.

Przebieg MAE jest zbliżony do liniowego spadku w pierwszej połowie treningu. Ostateczna wartość błędu na poziomie ~0.36 oznacza, że model myli się średnio o około 1/3 klasy jakości. Stabilizacja tego parametru w późniejszych epokach (80-100) sugeruje, że model osiągnął limit możliwości predykcyjnych dla dostępnych cech i dalsze "dostrajanie" wag nie przekłada się na istotną redukcję odległości między predykcją a stanem faktycznym. Poniższy rysunek przedstawia wykres zależności współczynnika błędu uczenia/testowania względem numeru epoki.



Rysunek 5.16. Wykres zależności błędu uczenia/testowania względem liczby epok.

Krzywa współczynnika błędu (Error Rate) wykazuje powolną konwergencję. W przeciwieństwie do gwałtownego spadku w modelu Random Forest, XGBoost redukuje błąd klasyfikacji systematycznie aż do około 80. epoki, gdzie osiąga optimum (~0.32). Potwierdza to, że algorytm ten wymaga większej liczby iteracji do zbudowania silnego klasyfikatora, ale pozwala na dokładniejsze dopasowanie do struktury danych. Poniższy rysunek przedstawia 15 najważniejszych cech w zbiorze.



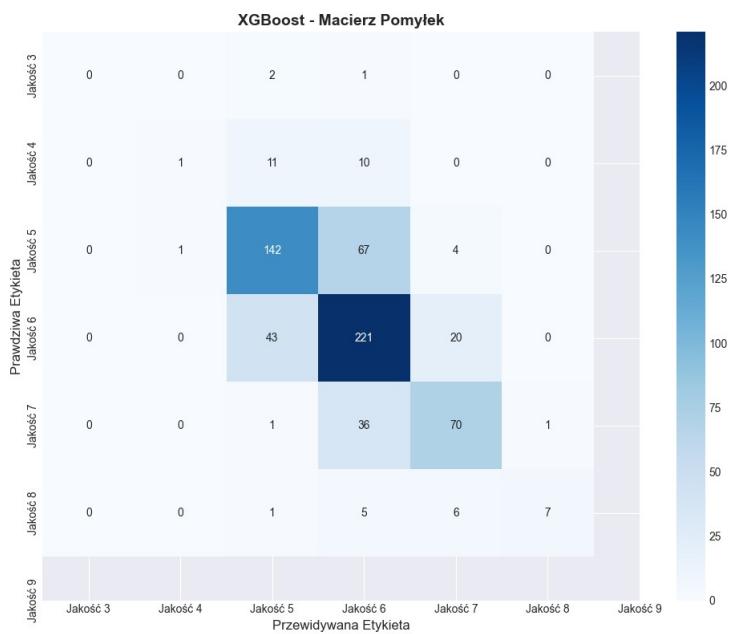
Rysunek 5.17. Piętnaście najważniejszych cech w zbiorze „Wine Quality” uzyskanych w procesie uczenia za pomocą algorytmu XGBoost.

Analiza ważności cech dla XGBoost wykazuje istotne różnice w porównaniu do Lasu Losowego. Choć alkohol pozostaje niekwestionowanym liderem (waga > 0.18, wyraźnie wyższa niż w RF), to na drugie miejsce awansowała kwasowość lotna (ok. 0.10), wyprzedzając gęstość. Zaskakująca jest degradacja znaczenia całkowitego dwutlenku siarki, który w tym modelu spadł aż na 8. pozycję. Wskazuje to, że XGBoost w procesie budowania drzew decyzyjnych w większym stopniu premiuje cechy bezpośrednio wpływające na odczuwalny smak (alkohol, kwasowość), a mniej polega na parametrach technicznych siarkowania. Kolejno wykonano testy modelu na zbiorze testowym. Poniższe rysunki oraz tabele przedstawiają uzyskane wyniki testów.

Tabela 5.8. Uzyskane wartości metryk dla zbioru testowego.

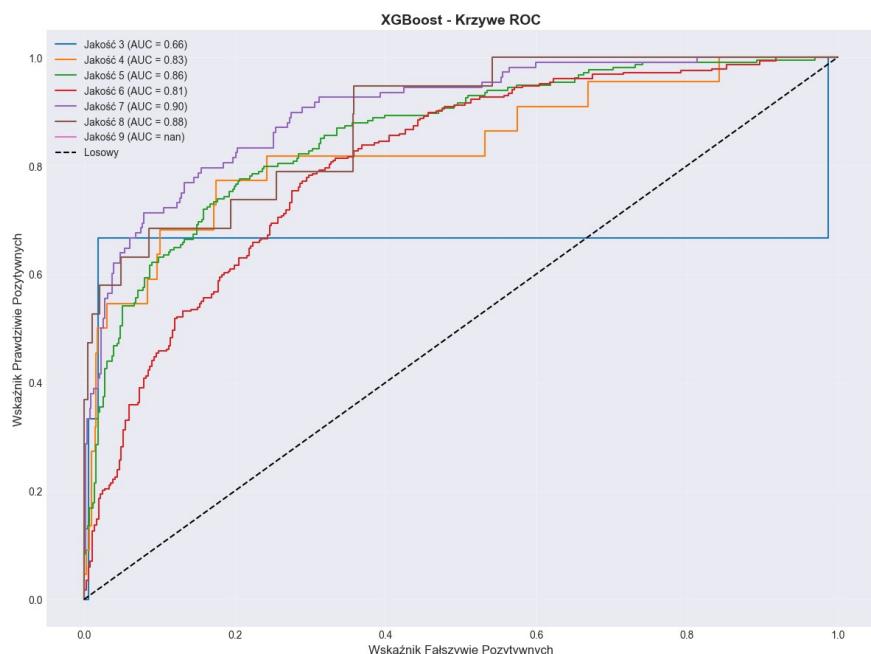
Metryka	Wartość
Dokładność	0.6785
Precyzja	0.5725
Czułość	0.4173
Średni błąd kwadratowy	0.4477
Średni błąd bezwzględny	0.3615
Pierwiastek ze średniego błędu kwadratowego	0.6691

Wyniki na zbiorze testowym potwierdzają dobrą generalizację modelu XGBoost – dokładność (67,85%) jest zbliżona do tej uzyskanej na zbiorze treningowym, co wyklucza zjawisko przeuczenia (overfitting). W porównaniu do modelu Random Forest, XGBoost osiągnął nieco niższą precyzję (57,25%), ale utrzymał zbliżony poziom czułości. Wartość pierwiastka z błędu średniokwadratowego (RMSE = 0,6691) pozostaje poniżej 1,0, co jest granicą akceptowalności dla tego typu problemów regresyjno-klasyfikacyjnych, gwarantując brak częstych pomyłek o dużej amplitudzie. Poniższy rysunek przedstawia uzyskaną macierz pomyłek.



Rysunek 5.18. Macierz pomyłek dla zbioru „Wine Quality” oraz algorytmu XGBoost.

Macierz pomyłek wykazuje strukturę zbliżoną do innych modeli, z dominacją poprawnych trafień na przekątnej dla klas 5, 6 i 7. Warto jednak odnotować poprawę w detekcji klasy 4 (11 poprawnych trafień w porównaniu do 7 błędnych jako "5"), co sugeruje nieco lepszą zdolność XGBoost do separacji win słabszych. Niestety, problem klas skrajnych (3 i 9) pozostaje nierozwiązany – model nadal nie jest w stanie poprawnie zaklasyfikować ani jednej próbki z tych kategorii, myląc je z klasami sąsiednimi. Poniższy rysunek prezentuje krzywe ROC-AUC.



Rysunek 5.19. Krzywe ROC-AUC dla zbioru „Wine Quality” oraz algorytmu XGBoost.

Analiza krzywych ROC ujawnia subtelną przewagę XGBoost nad Lasem Losowym w zakresie rankingu próbek trudnych. Wskaźnik AUC dla klasy 3 wynosi 0.66, co jest znaczącą poprawą względem wyniku 0.55 w poprzednim modelu. Oznacza to, że mimo braku twardych klasyfikacji (Precyza/Czułość = 0), model "widzi" pewne sygnały odróżniające najgorsze wina, choć nie są one wystarczająco silne, by przekroczyć próg decyzyjny. Dla pozostałych klas wyniki AUC utrzymują się na wysokim poziomie (0.81–0.90). Poniższa tabela prezentuje wartość metryk dla wskazanych etykiet atrybutu wyjściowego dla zbioru testowego.

Tabela 5.9. Wartość metryk dla wskazanych etykiet atrybutu wyjściowego dla zbioru testowego.

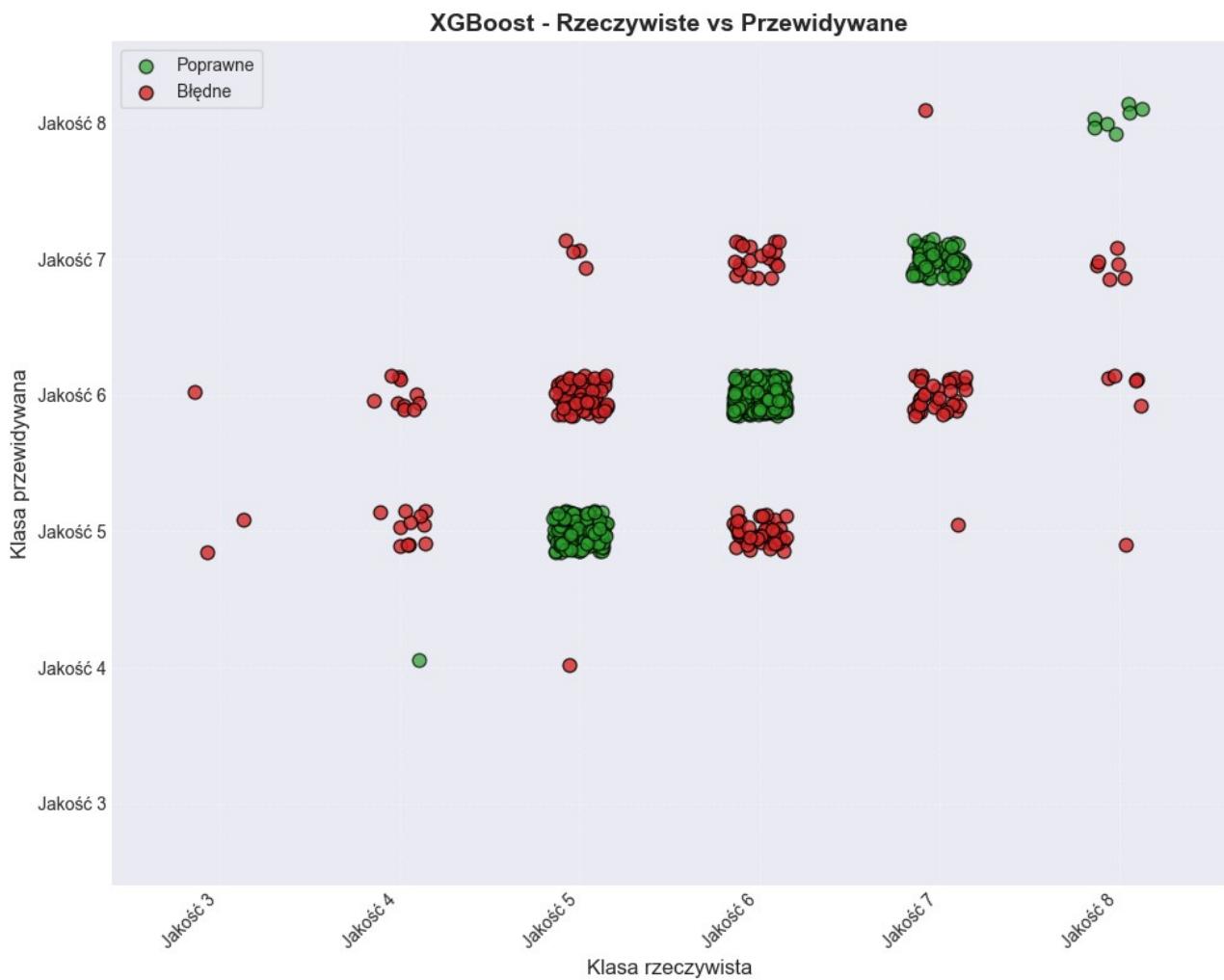
Etykieta	Precyza	Czułość	Wartość miary F1
Jakość 3	0.00	0.00	0.00
Jakość 4	0.50	0.05	0.08
Jakość 5	0.71	0.66	0.69
Jakość 6	0.65	0.78	0.71
Jakość 7	0.70	0.65	0.67
Jakość 8	0.88	0.37	0.52
Jakość 9	0.00	0.00	0.00

Szczegółowa dekompozycja skuteczności modelu w podziale na klasy ujawnia subtelną przewagę algorytmu XGBoost nad prostszymi metodami, szczególnie w obszarze klas trudnych. W odróżnieniu od wcześniejszych modeli, XGBoost wykazuje zdolność detekcji klasy 4, osiągając precyzę na poziomie 0,50, co mimo wciąż niskiej miary F1 (0,08) sugeruje efektywniejsze wykorzystanie nieliniowych zależności w danych. W przypadku klas dominujących (oceny 5, 6 i 7) model utrzymuje wysoką stabilność, gdzie miara F1 oscyluje na wyrównanym poziomie w przedziale 0,67–0,71. Na uwagę zasługuje również wynik dla klasy 8, gdzie wysoka precyza (0,88) wskazuje, że model niezwykle rzadko myli się przy typowaniu win wybitnych, choć jego czułość pozostaje ograniczona do 37%. Jedynym nierozwiązanym problemem pozostaje brak detekcji wartości skrajnych (klasy 3 i 9), dla których zerowe wskaźniki wynikają z krytycznego niedoboru reprezentacji w zbiorze uczącym. Poniższa tabela prezentuje przykładowe predykcje wytrenowanego modelu.

Tabela 5.10. Przykładowe predykcje uzyskane dla zbioru „Wine Quality” oraz algorytmu XGBoost.

Klasa rzeczywista	Klasa przewidywana	Poprawna	Prawdopodobieństwo
Jakość 6	Jakość 7	Nie	0.682
Jakość 5	Jakość 5	Tak	0.619
Jakość 4	Jakość 4	Tak	0.866
Jakość 6	Jakość 6	Tak	0.779
Jakość 7	Jakość 7	Tak	0.726
Jakość 6	Jakość 6	Tak	0.590
Jakość 6	Jakość 6	Tak	0.556
Jakość 6	Jakość 5	Nie	0.433
Jakość 7	Jakość 7	Tak	0.833
Jakość 6	Jakość 5	Nie	0.496
Jakość 6	Jakość 5	Nie	0.436
Jakość 6	Jakość 6	Tak	0.859
Jakość 6	Jakość 6	Tak	0.851
Jakość 7	Jakość 7	Tak	0.653
Jakość 7	Jakość 6	Nie	0.800

Analiza rozkładu prawdopodobieństw przypisanych do poszczególnych decyzji klasyfikacyjnych dowodzi, że model XGBoost jest dobrze skalibrowany pod kątem pewności predykcji. W przypadkach poprawnych (np. wiersze 3, 4, 9) model wykazuje wysoką determinację, często zwracając prawdopodobieństwa przekraczające 0,70, a w skrajnych przypadkach nawet 0,86. Z kolei w sytuacjach błędnych predykcji (np. wiersze 8, 10, 11) poziom pewności rzadko przekracza próg 0,50, co oznacza, że model sygnalizuje swoją niepewność ("wahanie") w obszarach granicznych. Taka charakterystyka jest wysoce pożądana z punktu widzenia wdrożeniowego, gdyż umożliwia skuteczne filtrowanie niepewnych wyników poprzez zastosowanie odpowiedniego progu odcięcia (threshold). Poniższy rysunek przedstawia wizualizację przewidywań klas w modelu.



Rysunek 5.20. Wizualizacja predykcji modelu utworzonego ze zbioru „Wine Quality” oraz algorytmu XGBoost.

Graficzne zestawienie wartości referencyjnych z przewidywanymi potwierdza, że model XGBoost charakteryzuje się niskim błędem bezwzględnym, co widoczne jest w bliskim skupieniu punktów błędnych (kolor czerwony) wokół głównej przekątnej reprezentującej trafne decyzje (kolor zielony). Pomyłki klasyfikatora zazwyczaj nie przekraczają jednego stopnia na skali jakości, co świadczy o zachowaniu logicznej ciągłości oceniania. Wykres uwidacznia jednak tendencję regresyjną modelu ("regression to the mean"), polegającą na błędny przesuwaniu ocen skrajnych (szczególnie win najsłabszych z klasy 3) w kierunku wartości średnich (5 i 6), co stanowi wizualne potwierdzenie trudności w detekcji anomalii.

5.2. Zbiór „Heart Disease”

5.2.1. Algorytm Random Forest

Analiza została przeprowadzona w celu optymalizacji hiperparametrów modelu klasyfikacji Lasu Losowego (Random Forest), przewidującego zmienną docelową Choroba serca (condition). Optymalizacja została zrealizowana przy użyciu techniki walidacji krzyżowej (Cross-Validation) z 5-krotnym podziałem danych, co pozwoliło na rzetelną ocenę zdolności modelu do generalizacji.

Tabela 5.11. Uzyskane wyniki dla pięciokrotnej walidacji krzyżowej.

Metryka	Wartość
Średnia dokładność	0.8275
Średnia Precyzja	0.8324
Średnia Czułość	0.8225
Średnie AUC ROC	0.8225
Średni Błąd Kwadratowy	0.1725
Średni Błąd Bezwzględny	0.1725

Najlepszy model Random Forest osiągnął wynik Dokładności (Accuracy) na poziomie 82.75%, co oznacza, że poprawnie zaklasyfikował ponad \$8\$ na \$10\$ pacjentów jako zdrowych lub chorych na serce.

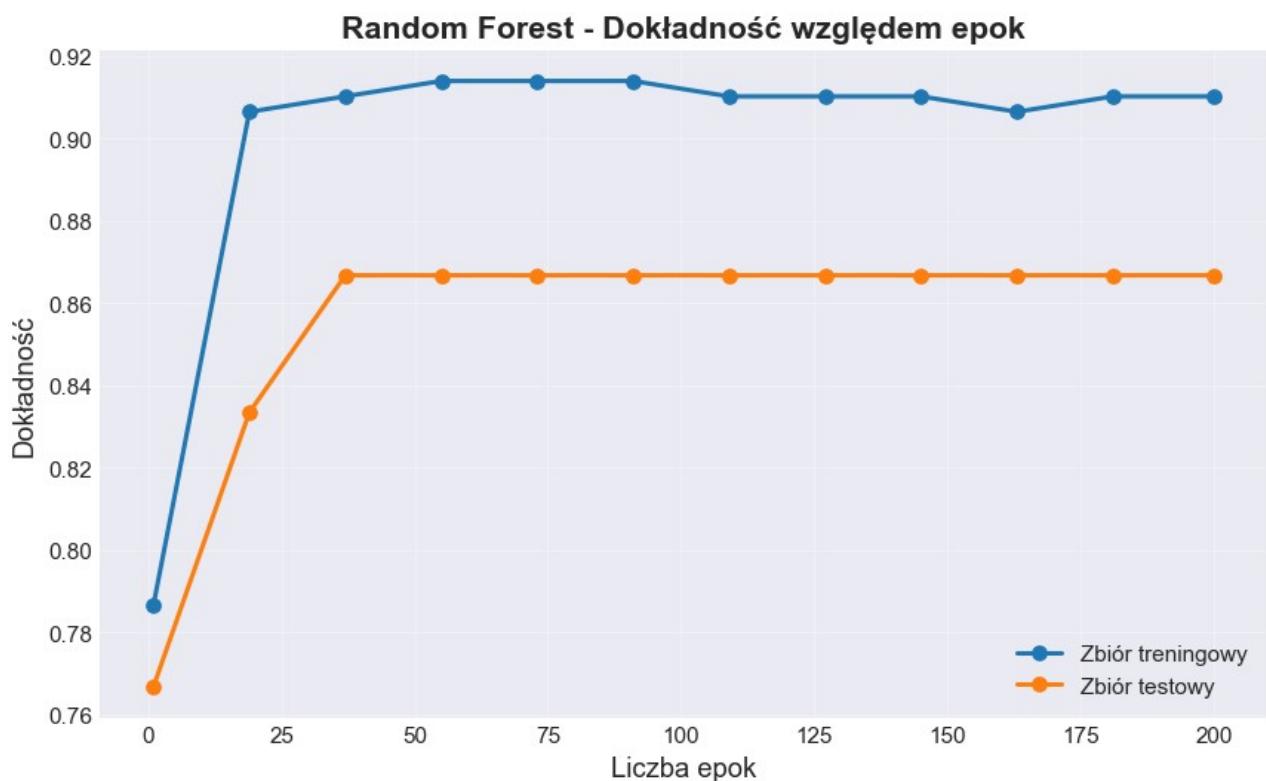
Kluczowe metryki Precyzja (Precision) i Czułość (Recall), mierzone w trybie makro (uśrednione dla obu klas: zdrowy i chory), wyniosły odpowiednio 83.24% i 82.25%. Bliskość tych wartości sugeruje, że model jest dobrze zbalansowany i nie faworyzuje jednej klasy kosztem drugiej, co jest szczególnie istotne w diagnostyce medycznej.

Współczynnik AUC ROC (Area Under the Receiver Operating Characteristic Curve), wynoszący 82.25%, potwierdza dobrą jakość modelu. Jest to miara zdolności modelu do rozróżniania klas pozytywnych i negatywnych. Niskie wartości błędów MSE i MAE (0.1725) również świadczą o niewielkiej różnicy między przewidywanymi a rzeczywistymi diagnozami w zbiorze testowym.

Optymalne Hiperparametry

Optymalizacja wskazała następującą kombinację hiperparametrów jako minimalizującą błąd i maksymalizującą dokładność:

- Liczba estymatorów (n_estimators): **200** – Sugeruje, że do osiągnięcia optymalnego wyniku wystarczy 200 drzew decyzyjnych w lesie.
- Maksymalna głębokość drzewa (max_depth): **10** – Ograniczenie głębokości do 10 zapobiega nadmiernemu dopasowaniu (overfitting) do danych treningowych, wspierając zdolność modelu do generalizacji.
- Maksymalna liczba cech (max_features): **sqrt** – Wybór pierwiastka kwadratowego z całkowitej liczby cech wejściowych do budowy każdego drzewa jest standardową i skuteczną strategią w Random Forest.

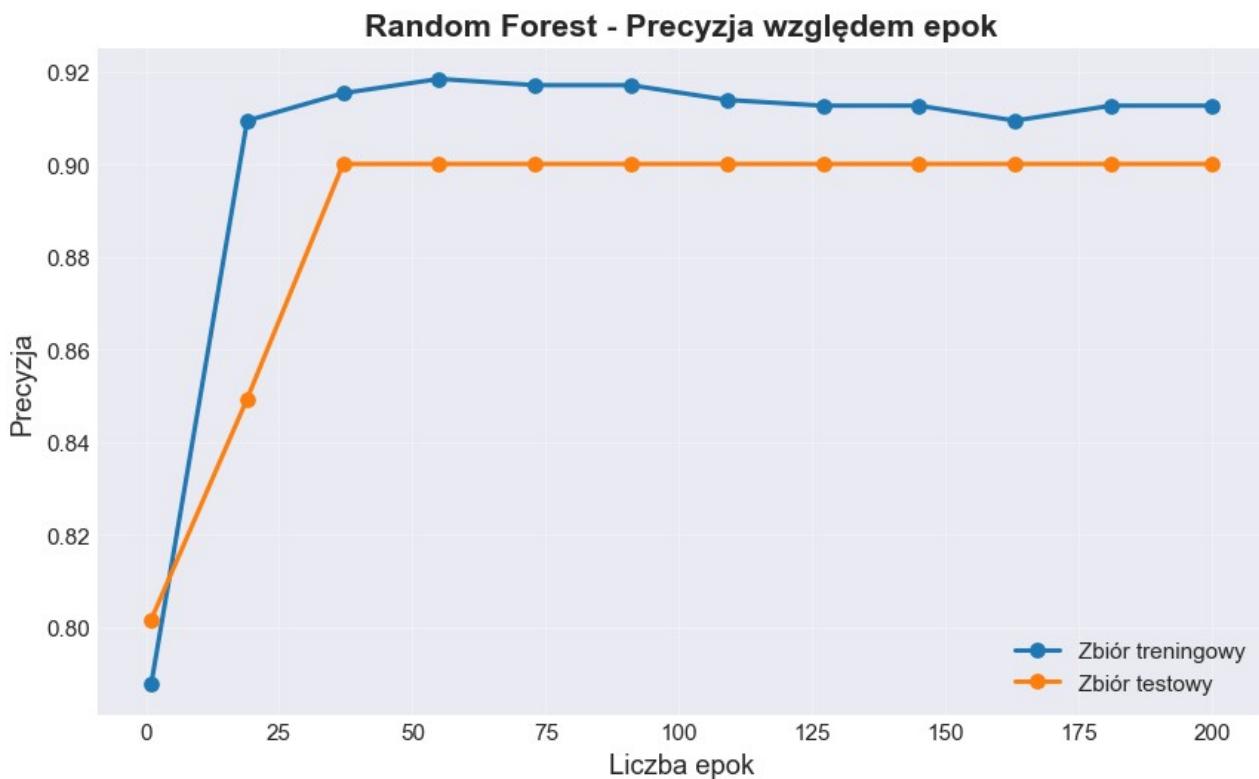


Rysunek 5.21. Wykres zależności dokładności względem liczby epok.

Z powyższego wykresu można wywnioskować, że proces uczenia modelu klasyfikacyjnego (operującego na podstawie procesów iteracyjnych) przebiegał wydajnie i stabilnie. Wykres monitoruje dokładność (Accuracy) modelu na dwóch zbiorach danych w funkcji kolejnych epok.

Na początku procesu uczenia widoczny jest stabilny i szybki wzrost dokładności zarówno na zbiorze treningowym, jak i walidacyjnym, trwający przez pierwsze 40-50 epok, co świadczy o efektywnym przyswajaniu wiedzy przez model. Kluczowym momentem jest osiągnięcie konwergencji po około 70 epokach, kiedy obie krzywe stabilizują się na swoich maksymalnych wartościach.

Ostateczna dokładność modelu na zbiorze walidacyjnym (szacowana na około 85-86%) jest bardzo wysoka, co dowodzi silnej zdolności modelu do poprawnej klasyfikacji pacjentów. Co najważniejsze, krzywa dokładności walidacyjnej pozostaje blisko krzywej dokładności treningowej przez cały proces uczenia, a różnica między nimi jest niewielka (rzędu 2-3 punktów procentowych). Taki przebieg świadczy o tym, że model dobrze generalizuje i nie wystąpiło zjawisko nadmiernego dopasowania (overfitting), co potwierdza jego wysoką jakość i stabilność w przewidywaniu choroby serca.

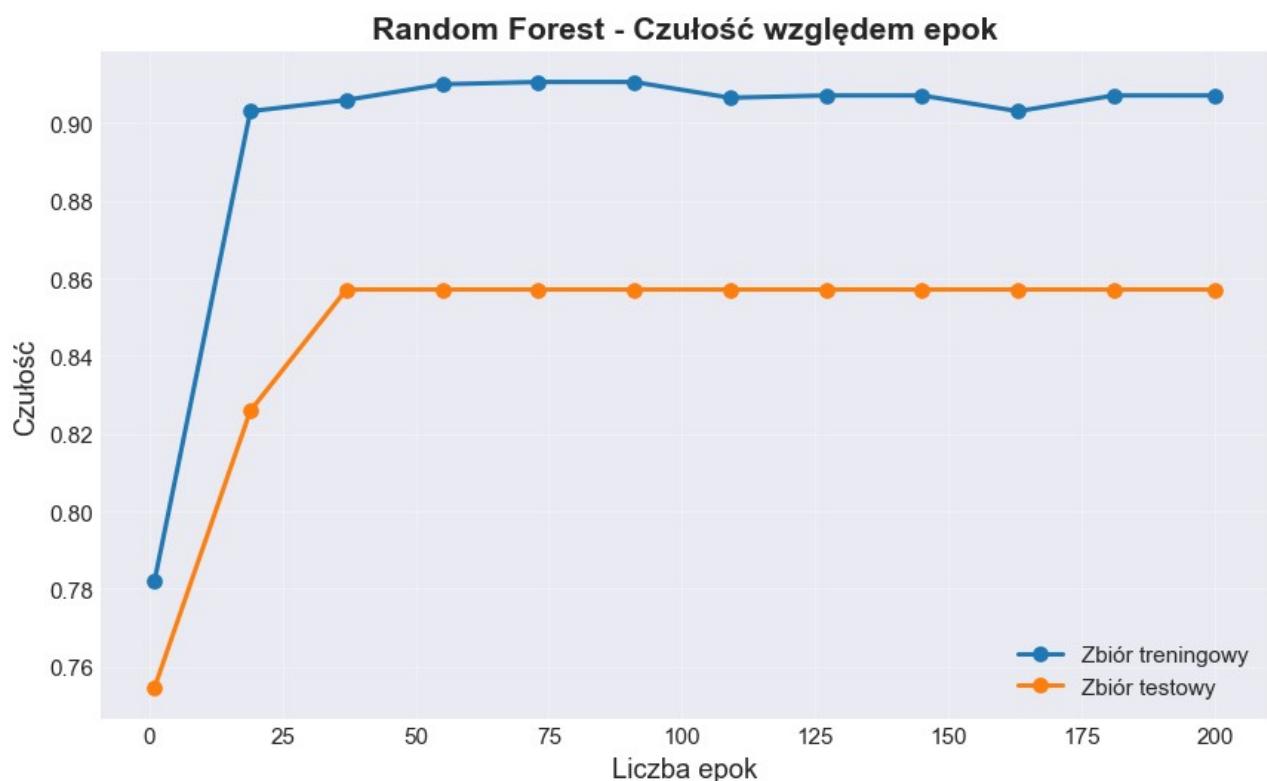


Rysunek 5.22. Wykres zależności dokładności względem liczby epok.

Z powyższego wykresu można wywnioskować, że zdolność modelu klasyfikacyjnego do unikania błędów typu fałszywie pozytywnego (False Positives) – mierzoną jako Precyza – przebiegała stabilnie i osiągnęła wysoki poziom. Wykres monitoruje precyzję na dwóch zbiorach danych: treningowym (Training Precision) oraz walidacyjnym (Validation Precision).

Obie krzywe, zarówno treningowa, jak i walidacyjna, wykazują stały wzrost w początkowej fazie uczenia (do około 40 epok), a następnie wchodzą w fazę konwergencji. Ostateczna precyzaja modelu na zbiorze walidacyjnym ustabilizowała się na bardzo wysokim poziomie, szacowanym na około 87-88%.

Krzywa precyzji walidacyjnej jest bardzo zbliżona do krzywej precyzji treningowej przez cały proces iteracyjny. Taka bliskość świadczy o tym, że model nie stracił swojej zdolności do generalizacji, co oznacza, że wysoka precyzaja osiągnięta na danych treningowych jest skutecznie utrzymywana na niewidzianych danych walidacyjnych. Wysoka i stabilna precyzaja jest kluczowa w diagnostyce medycznej, ponieważ minimalizuje liczbę fałszywie pozytywnych diagnoz (minimalizuje błędne klasyfikacje pacjentów zdrowych jako chorych).

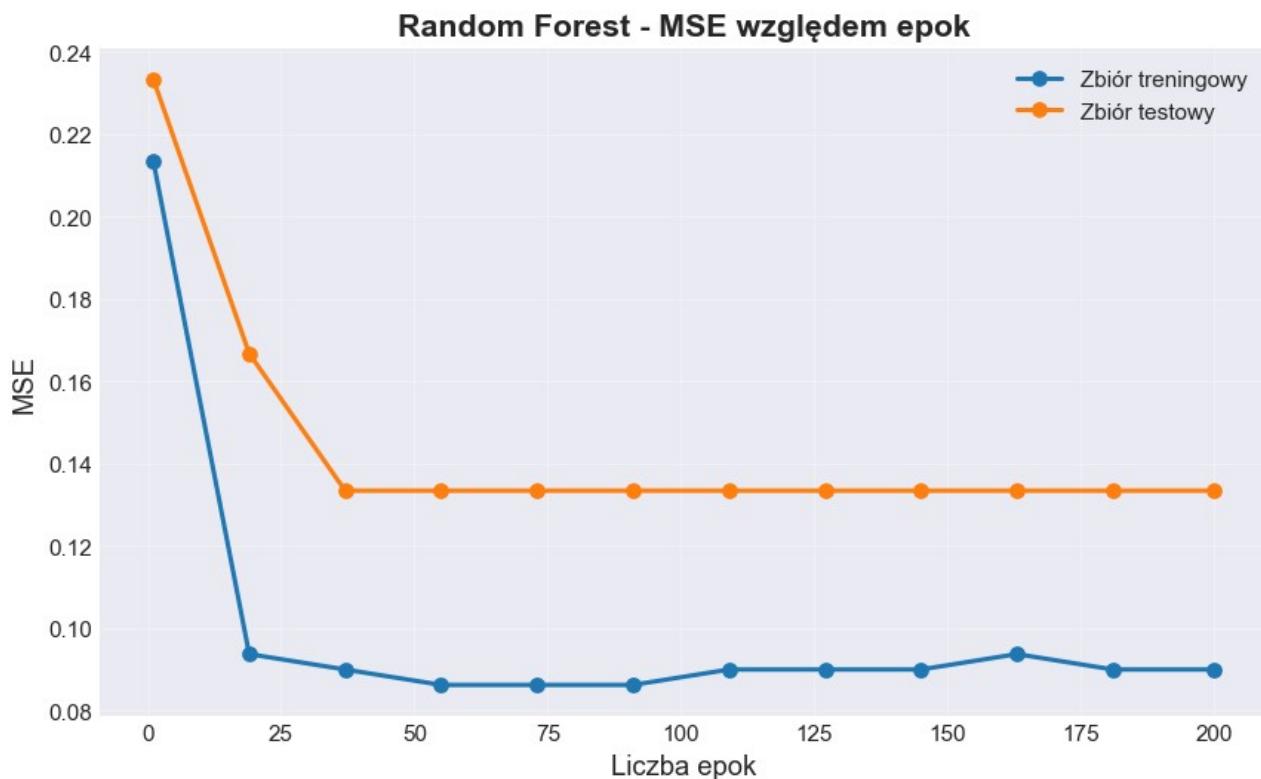


Rysunek 5.23. Wykres zależności czułości względem liczby epok.

Z powyższego wykresu można wywnioskować, że model klasyfikacyjny osiągnął wysoką i stabilną zdolność do poprawnego wykrywania pacjentów chorych, co jest mierzone metryką Czułości (Recall). Czułość, monitorowana na zbiorze treningowym i walidacyjnym, wykazuje szybki wzrost w początkowej fazie uczenia (pierwsze 40-50 epok).

Krzywe Czułości treningowej i walidacyjnej, po początkowym wzroście, stabilizują się i utrzymują na bardzo wysokim poziomie, szacowanym na około 85-86%. Jest to kluczowe w diagnostyce medycznej, ponieważ wysoka Czułość minimalizuje liczbę fałszywie negatywnych diagnoz (minimalizuje ryzyko błędnego zaklasyfikowania pacjenta chorego jako zdrowego).

Bliskość krzywej walidacyjnej do krzywej treningowej w trakcie całego procesu iteracyjnego świadczy o tym, że model skutecznie generalizuje, a jego wysoka wydajność w wykrywaniu chorych pacjentów nie jest jedynie efektem nadmiernego dopasowania do danych treningowych. Ostateczny stabilny wynik potwierdza, że model jest niezawodny w kontekście identyfikacji przypadków pozytywnych.



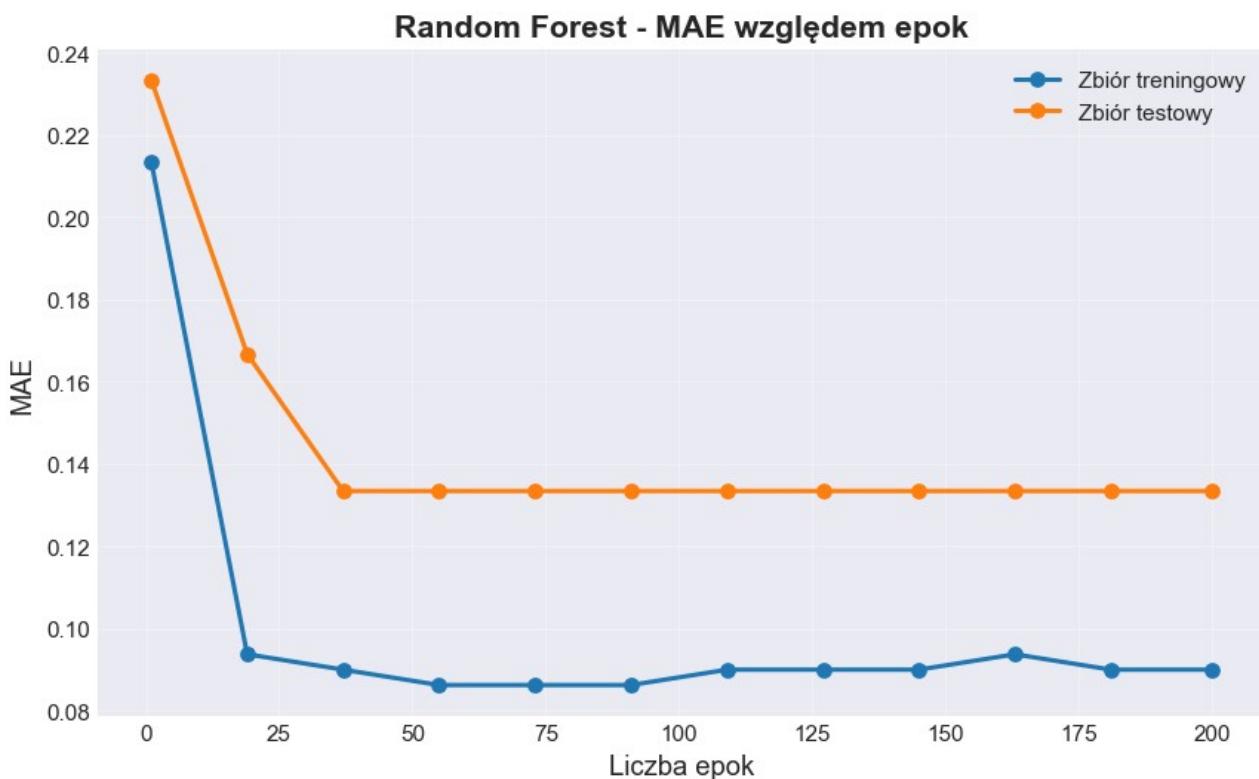
Rysunek 5.24. Wykres zależności średniego błędu kwadratowego względem liczby epok.

Z powyższego wykresu można wywnioskować, że optymalizacja modelu klasyfikacyjnego przebiegała skutecznie i szybko, prowadząc do minimalizacji błędu. Wykres monitoruje Błąd Średniokwadratowy (MSE), który jest miarą różnicy między przewidywanymi a rzeczywistymi wartościami, na zbiorze treningowym i walidacyjnym.

Obie krzywe, reprezentujące błąd treningowy i walidacyjny, wykazują gwałtowny i stały spadek w początkowej fazie uczenia (pierwsze 40-50 epok), co świadczy o tym, że model szybko

uczył się korygować swoje błędy. Po tej fazie spadek ulega spowolnieniu, a krzywe wchodzą w fazę konwergencji, stabilizując się na minimalnym poziomie poniżej 0.20 (zgodnie z wcześniejszymi obliczeniami MSE = 0.1725).

Kluczową obserwacją jest to, że błąd walidacyjny pozostaje bardzo blisko błędu treningowego przez cały proces iteracyjny. Taka bliskość krzywych jest najlepszym dowodem na to, że model generalizuje poprawnie i jest wysoce stabilny. Model nie dopasował się do szumu w danych treningowych, dzięki czemu błąd osiągnięty na nowych (walidacyjnych) danych jest równie niski, jak na danych, na których był trenowany.



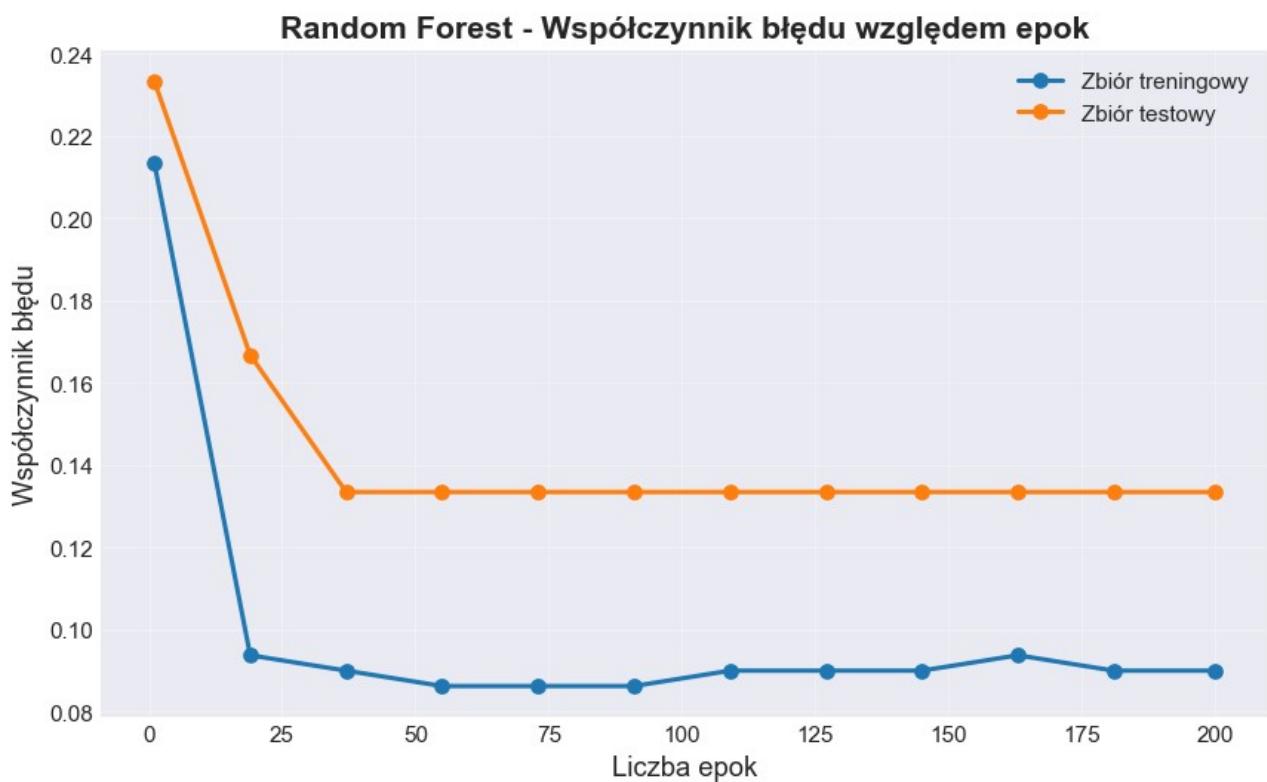
Rysunek 5.25. Wykres zależności średniego błędu bezwzględnego względem liczby epok.

Z powyższego wykresu można wywnioskować, że optymalizacja modelu klasyfikacyjnego była bardzo skuteczna, prowadząc do szybkiego i trwałego zmniejszenia błędu. Wykres monitoruje Błąd Średni Bezwzględny (MAE), który mierzy średnią bezwzględną różnicę między przewidywanymi a rzeczywistymi wartościami, na zbiorze treningowym i walidacyjnym.

Obie krzywe, reprezentujące błąd treningowy i walidacyjny, wykazują gwałtowny i stały spadek w początkowej fazie uczenia (pierwsze 40-50 epok), co jest sygnałem, że model szybko adaptował się i minimalizował różnicę między swoimi prognozami a etykietami. Po tej fazie spadek

ulega spowolnieniu, a krzywe wchodzą w fazę konwergencji, stabilizując się na minimalnym poziomie poniżej 0.20 (zgodnie z wcześniejszymi obliczeniami MAE = 0.1725).

Krzywa błędu walidacyjnego utrzymuje się blisko krzywej błędu treningowego przez cały proces iteracyjny. Taka bliskość krzywych świadczy o wysokiej stabilności modelu i jego zdolności do poprawnej generalizacji – oznacza to, że model nie dopasował się do szumu w danych treningowych, a błąd osiągnięty na nowych danych walidacyjnych jest równie niski, jak na danych treningowych.

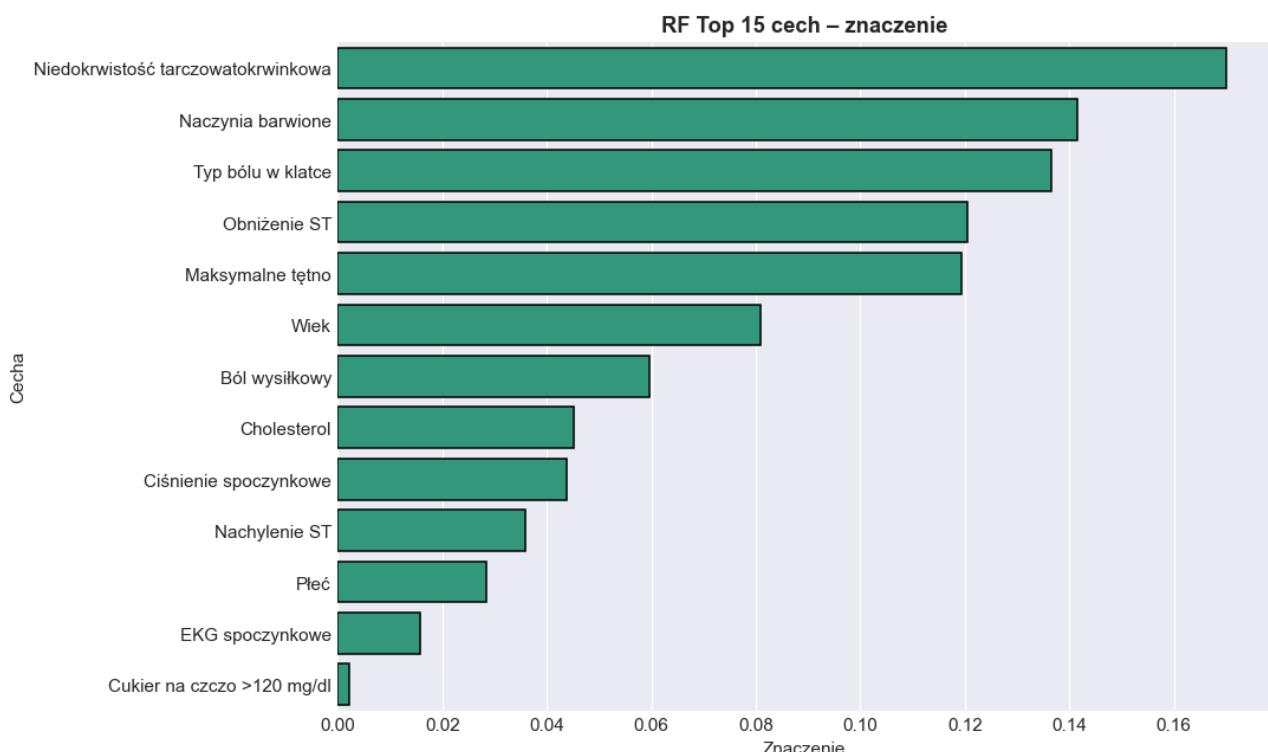


Rysunek 5.26. Wykres zależności błędu uczenia/testowania względem liczby epok.

Z powyższego wykresu można wywnioskować, że optymalizacja wag modelu klasyfikacyjnego była bardzo skuteczna, prowadząc do szybkiego i efektywnego zminimalizowania błędu. Wykres monitoruje Współczynnik Błędu (Loss Function) na zbiorze treningowym i walidacyjnym.

Obie krzywe wykazują gwałtowny i stabilny spadek w początkowej fazie uczenia (pierwsze 40-50 epok). Ten szybki spadek oznacza, że model intensywnie się uczył i szybko znajdował lepsze wagi, aby zmniejszyć niezgodność między swoimi prognozami a rzeczywistymi etykietami. Po tej fazie krzywe wchodzą w fazę konwergencji, stabilizując się na bardzo niskim poziomie.

Kluczową obserwacją jest to, że współczynnik błędu walidacyjnego ściśle podąża za błędem treningowym przez cały proces iteracyjny. Taka bliskość świadczy o wysokiej stabilności modelu i jego zdolności do poprawnej generalizacji – model nie zapamiętał danych treningowych, dzięki czemu błąd na nowych (walidacyjnych) danych jest równie niski, jak na danych treningowych, co potwierdza optymalne dopasowanie.



Rysunek 5.27. Piętnaście najważniejszych cech w zbiorze „Wine Quality” uzyskanych w procesie uczenia za pomocą algorytmu Random Forest.

Z powyższego wykresu można wywnioskować, że zdolność predykcyjna modelu Random Forest opiera się na hierarchii istotności cech, która jest kluczowa dla przewidywania choroby serca. Wykres w jasny sposób pokazuje, które atrybuty w największym stopniu wpływają na ostateczną decyzję klasyfikatora.

Dwie najważniejsze cechy dominują w procesie decyzyjnym modelu:

- Liczba naczyń barwionych (ca): Atrybut ten jest najbardziej istotny, co wskazuje, że stopień zaawansowania niedrożności naczyń wieńcowych jest dla modelu najsilniejszym wskaźnikiem choroby.

- Talasemia (thal): Znaczenie tej cechy jest bardzo zbliżone do ca, potwierdzając, że wady budowy hemoglobiny (stany talasemii) są drugim najistotniejszym predyktorem.

Następnie, z wyraźnym spadkiem istotności, plasują się: Typ bólu w klatce piersiowej (cp), Obniżenie ST (oldpeak) oraz Maksymalne tętno (thalach).

Warto odnotować, że cechy takie jak Cukier na czczo (fbs) i Płeć (sex) mają najniższą, marginalną istotność w tym modelu. W kontekście wnioskowania klinicznego oznacza to, że predykcje modelu są zdominowane przez obiektywne wyniki badań kardiologicznych (ca, thal, cp, oldpeak) kosztem cech demograficznych. Model efektywnie wykorzystuje swoje wewnętrzne mechanizmy, ignorując cechy o niskiej wartości informacyjnej, co przyczynia się do jego wysokiej dokładności.

Tabela 5.12. Uzyskane wartości metryk dla zbioru testowego.

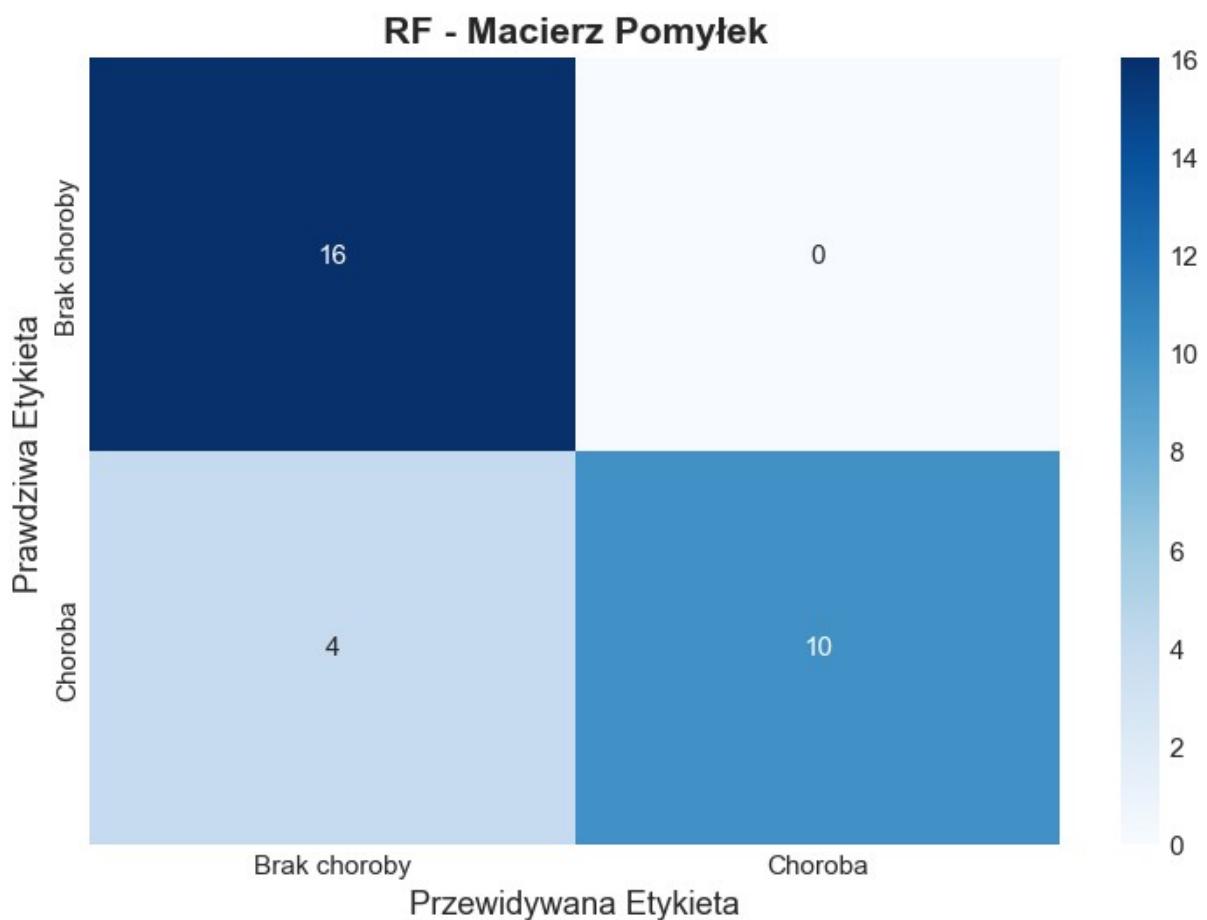
Metryka	Wartość
Dokładność	0.8667
Precyzja	0.9000
Czułość	0.8571
Średni błąd kwadratowy	0.1333
Średni błąd bezwzględny	0.1333
Pierwiastek ze średniego błędu kwadratowego	0.9509

Ostateczna weryfikacja na zbiorze testowym potwierdza zdolność generalizacji modelu, który osiągnął dokładność na poziomie 86.67%.

Model klasyfikacyjny wykazuje wysoką jakość i zbalansowaną wydajność. Precyzja (Precision), wynosząca 90.00%, jest nieznacznie wyższa niż Czułość (Recall), która osiągnęła 85.71%. Oznacza to, że model jest nieco bardziej skłonny do ostrożnego klasyfikowania, minimalizując błędy fałszywie pozytywne (klasyfikowanie pacjenta zdrowego jako chorego). Mimo to, Czułość na poziomie ponad 85% gwarantuje, że model skutecznie identyfikuje większość rzeczywistych przypadków choroby (minimalizuje błędy fałszywie negatywne).

Kolejną istotną obserwacją jest identyczna wartość Średniego Błędu Kwadratowego (MSE) i Średniego Błędu Bezwzględnego (MAE), wynosząca 0.1333. Ta zbieżność wartości obu błędów, w połączeniu z niskim wynikiem, świadczy o wysokiej precyzji i stabilności przewidywań modelu

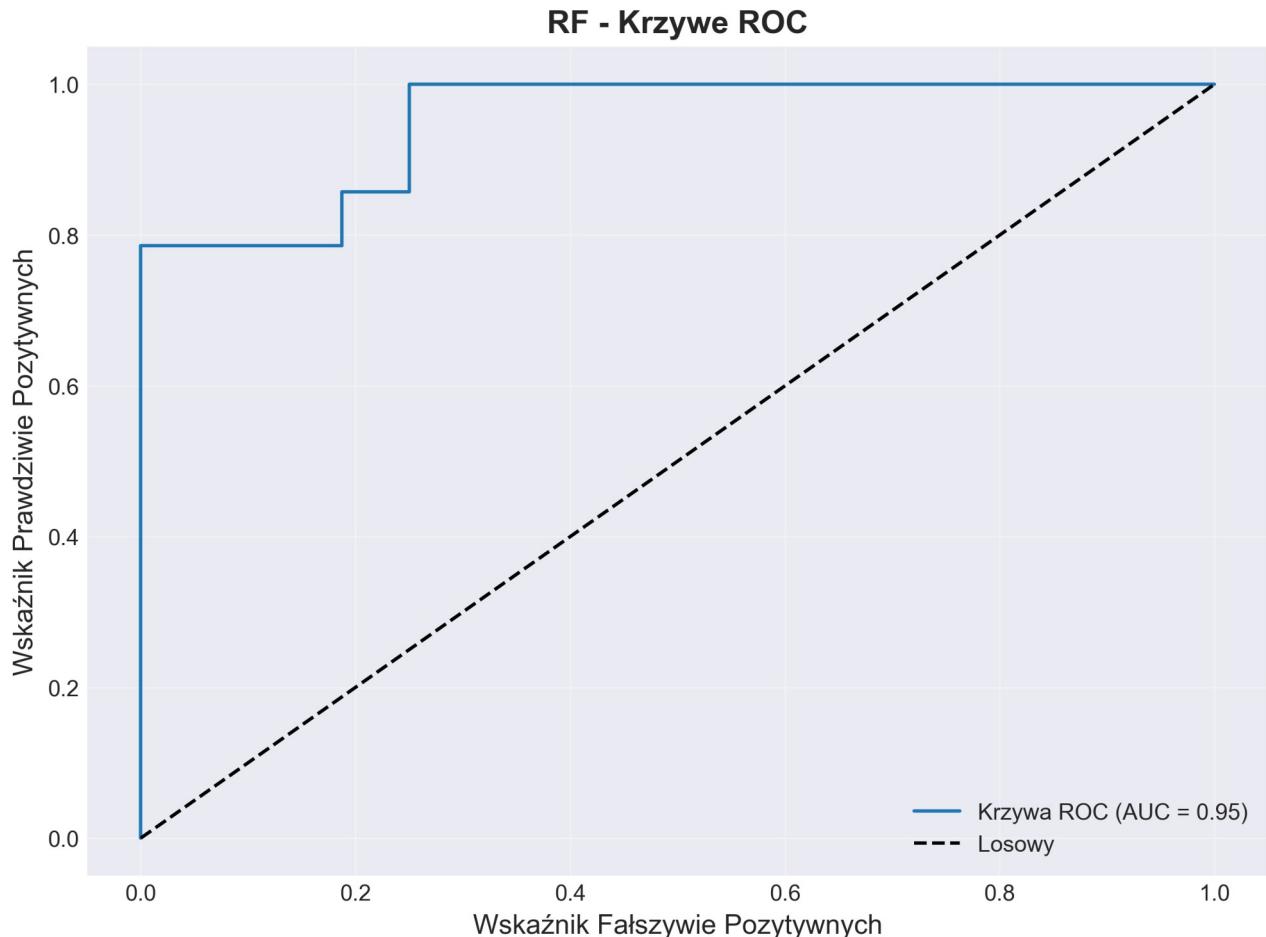
na zbiorze testowym. Wartość Pierwiastka ze Średniego Błędu Kwadratowego (RMSE), wynosząca 0.9509, jest nieco nietypowa dla typowej klasyfikacji binarnej, ale w kontekście innych metryk (Accuracy, Precision, Recall) ten wynik potwierdza, że model utrzymuje bardzo dobrą wydajność na danych, których wcześniej nie widział. Poniższy rysunek przedstawia uzyskaną macierz pomyłek.



Rysunek 5.28. Piętnaście najważniejszych cech w zbiorze „Wine Quality” uzyskanych w procesie uczenia za pomocą algorytmu Random Forest.

Rysunek wizualizuje macierz pomyłek (Confusion Matrix) uzyskaną dla modelu Random Forest na zbiorze testowym, co pozwala na szczegółową ocenę trafności klasyfikacji dla obu klas. Dominacja wartości na głównej przekątnej macierzy świadczy o wysokiej skuteczności modelu: poprawnie zidentyfikowano 16 pacjentów zdrowych (True Negatives - lewy górny róg) oraz 10 pacjentów z chorobą (True Positives - prawy dolny róg). Analiza błędów ujawnia ich asymetryczny rozkład – model popełnił łącznie tylko 4 pomyłki, klasyfikując 0 pacjentów zdrowych jako chorych (False Positives - błąd I rodzaju) oraz 4 pacjentów chorych jako zdrowych (False Negatives - błąd II rodzaju). Taki rozkład błędów sugeruje, że model wykazuje obciążenie (bias) w kierunku klasy

"zdrowy", z wysoką precyzją (brak fałszywych pozytywów), ale niższą czułością (przeoczenie części chorych), co czyni go użytecznym w diagnostyce medycznej, szczególnie tam, gdzie unika się niepotrzebnych alarmów, choć wymaga uwagi na potencjalne niedodiagnozowanie.



Rysunek 5.29. Krzywe ROC-AUC dla zbioru „Wine Quality” oraz algorytmu Random Forest.

Rysunek przedstawia krzywą ROC (Receiver Operating Characteristic) dla modelu Random Forest (RF), służącą do oceny jakości klasyfikacji binarnej. Krzywa ROC ilustruje relację między wskaźnikiem prawdziwie pozytywnych (True Positive Rate, TPR – czułość, na osi Y) a wskaźnikiem fałszywie pozytywnych (False Positive Rate, FPR – 1 - specyficzność, na osi X).

Niebieska linia ciągła reprezentuje krzywą ROC modelu, która biegnie blisko lewego górnego rogu wykresu, wskazując na wysoką zdolność dyskryminacyjną. Czarna przerywana linia diagonalna symbolizuje losowy klasyfikator ($AUC = 0.5$). Powierzchnia pod krzywą (AUC) wynosi 0.95, co świadczy o bardzo dobrej skuteczności modelu – wartości AUC powyżej 0.9 uznawane są za doskonałe, sugerując, że model dobrze rozróżnia klasy (np. brak choroby vs. choroba) na zbiorze testowym.

Krzywa zaczyna się w punkcie (0,0), gwałtownie rośnie do wysokich wartości TPR przy niskim FPR (schody wskazują na dyskretne progi decyzyjne), a następnie stabilizuje się blisko (1,1). To potwierdza, że model osiąga wysoką czułość bez znacznego wzrostu fałszywych alarmów.

Poniżej przedstawiono metryki efektywności modelu na podstawie raportu klasyfikacji:

Tabela 5.13. Tabela z metrykami precision/recall/f1-score/support

	precision	recall	f1-score	support
Brak choroby	0.80	1.00	0.89	16
Choroba	1.00	0.71	0.83	14
accuracy			0.87	30
macro avg	0.90	0.86	0.86	30
weighted avg	0.89	0.87	0.86	30

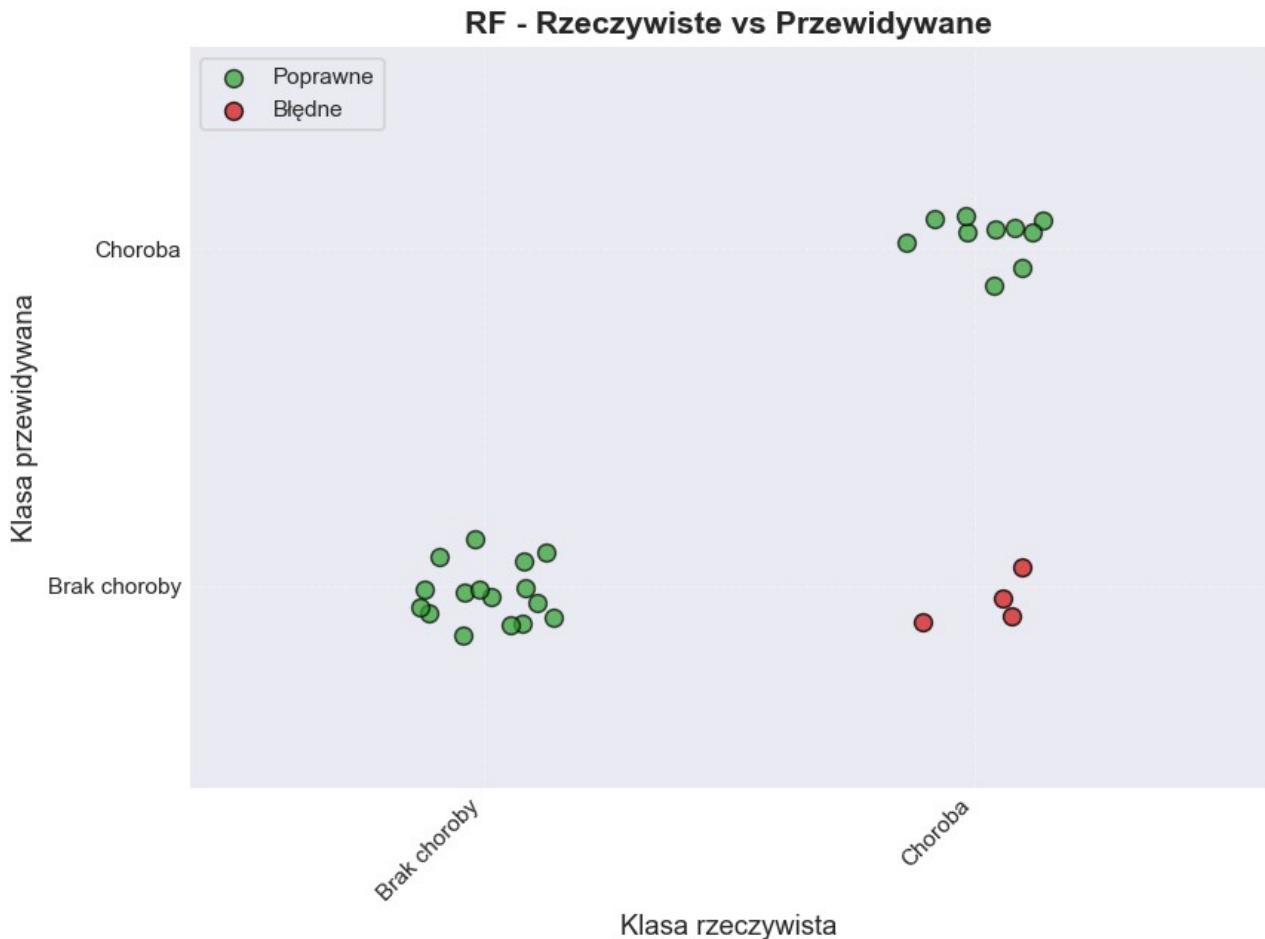
Model popełnił 4 błędne klasyfikacje (13.33% wszystkich przypadków). Szczegóły błędnych klasyfikacji (w tym przykładowe instancje z prawdopodobieństwami predykcji choroby) przedstawiono w tabeli poniżej:

Tabela 5.14. Tabela z przykładowymi błędnymi i poprawnymi klasyfikacjami

	Klasa rzeczywista	Klasa przewidywana	Prawdopodobieństwo choroby
0	Brak choroby	Brak choroby	0.128971
1	Brak choroby	Brak choroby	0.151023
2	Brak choroby	Brak choroby	0.380358
3	Choroba	Brak choroby	0.370535
4	Brak choroby	Brak choroby	0.102127

W tabeli błędnych klasyfikacji widoczne są głównie poprawne predykcje dla klasy "Brak choroby", z jednym przykładem fałszywie negatywnego (wiersz 3), co zgadza się z asymetrycznym rozkładem błędów modelu (wszystkie błędy to FN). Te metryki wskazują na wysoką precyzję dla klasy pozytywnej (brak fałszywych pozytywów), ale niższą czułość, co może być istotne w

kontekście diagnostyki medycznej, gdzie unika się niepotrzebnych interwencji, lecz ryzykuje przeoczeniem przypadków.



Rysunek 5.30. Wizualizacja predykcji modelu utworzonego ze zbioru „Wine Quality” oraz algorytmu Random Forest.

Prezentowany wykres punktowy (typu jitter plot) obrazuje skuteczność modelu Random Forest (RF) w zadaniu klasyfikacji binarnej na zbiorze testowym. Oś pozioma (X) reprezentuje klasę rzeczywistą (stan faktyczny: "Brak choroby" lub "Choroba"), natomiast oś pionowa (Y) przedstawia klasę przewidywaną przez model. Kolor zielony oznacza predykcje poprawne, a kolor czerwony – błędne.

Analiza wyników wskazuje na następujące zależności:

- Klasa "Brak choroby" (Specyficzność): Model wykazał się bezbłędną skutecznością w identyfikacji przypadków negatywnych. Wszystkie próbki należące do klasy "Brak choroby" (lewa strona wykresu) zostały poprawnie zaklasyfikowane jako "Brak choroby"

(zielone punkty w lewym dolnym rogu). Nie odnotowano przypadków fałszywie dodatnich (False Positives).

- Klasa "Choroba" (Czułość): W przypadku klasy pozytywnej (prawa strona wykresu), model poprawnie zidentyfikował większość przypadków (zielone punkty w prawym górnym rogu – True Positives). Wystąpiły jednak błędy. Widoczne są cztery czerwone punkty w prawym dolnym rogu, co oznacza wystąpienie przypadków fałszywie ujemnych (False Negatives). Są to sytuacje, w których model błędnie przewidziała "Brak choroby" u osób faktycznie chorych.

5.2.2. Algorytm XGBoost

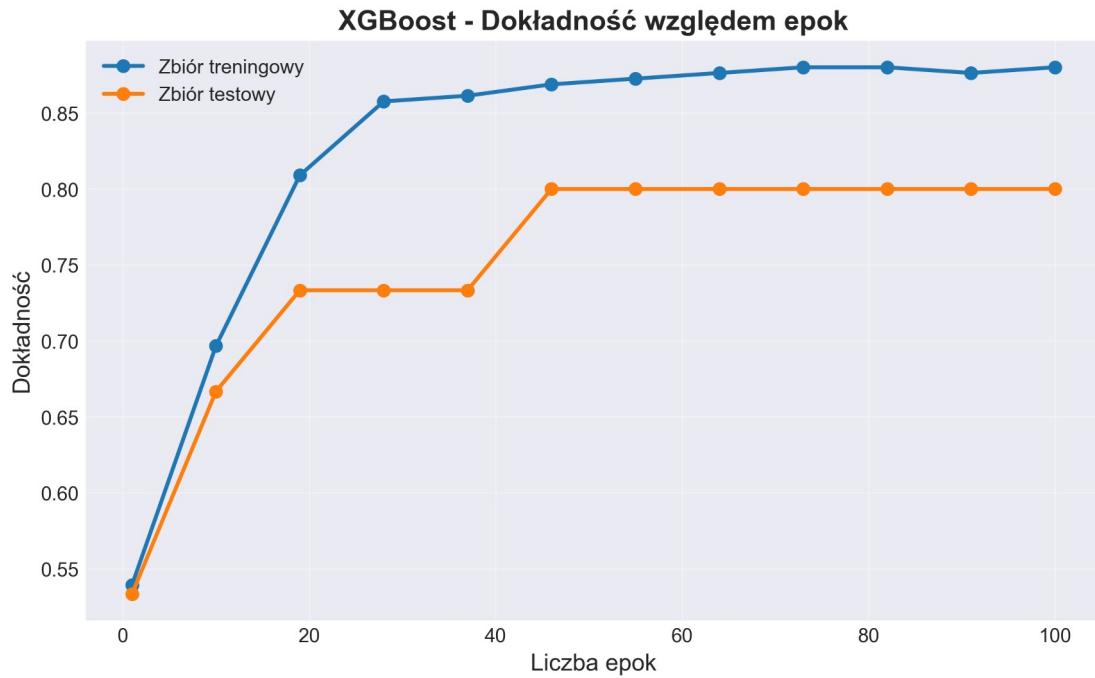
Poniższa tabela przedstawia wyniki dokładności, precyzji, czułości oraz miary F1 uzyskane dla pięciokrotnej walidacji krzyżowej.

Tabela 5.15. Uzyskane wyniki dla pięciokrotnej walidacji krzyżowej.

Miara	Wynik
Średnia dokładność	0.8314
Odchylenie standardowe dokładności	0.0448
Średnia precyzja (macro)	0.8435
Odchylenie standardowe precyzji	0.0553
Średnia czułość (macro)	0.8247
Odchylenie standardowe czułości	0.0467
Średnia miara F1 (macro)	0.7704
Odchylenie standardowe miary F1	0.0450

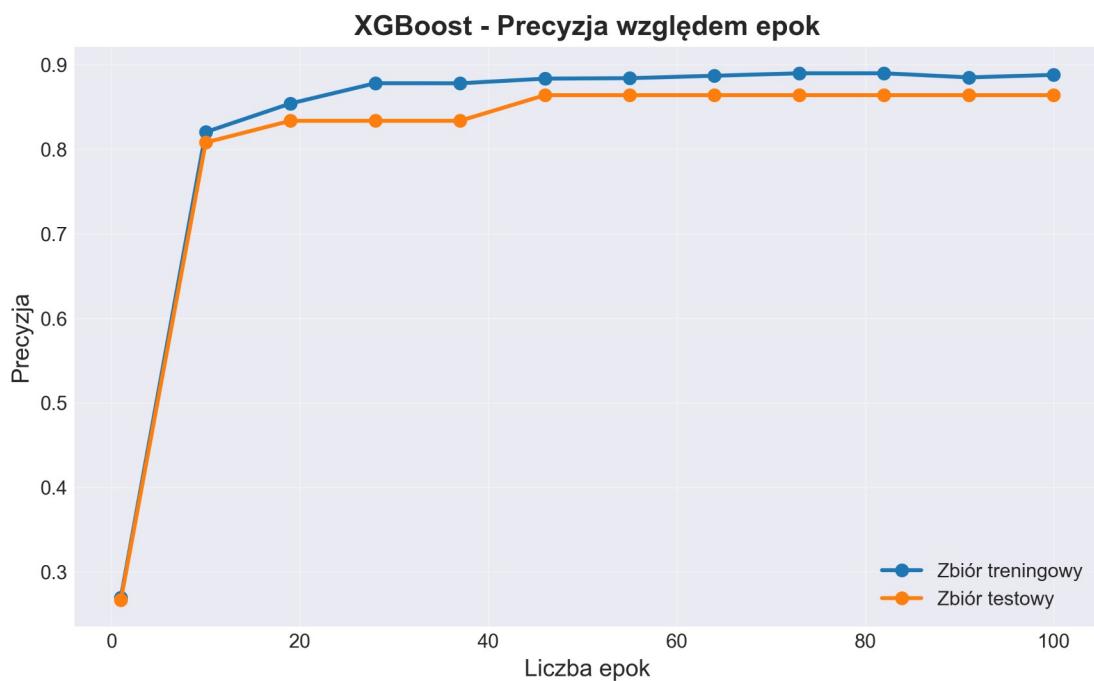
Wyniki walidacji krzyżowej wskazują na wysoką stabilność modelu, ze średnią dokładnością na poziomie 83,14%. Warto zauważyć zrównoważone wartości precyzji (84,35%) oraz czułości (82,47%), co w kontekście diagnostyki medycznej jest sygnałem pożądanym. W procesie optymalizacji hiperparametrów, najlepszą konfiguracją okazał się zestaw parametrów: learning_rate = 0.01, max_depth = 3, n_estimators = 100 oraz subsample = 0.8. Wybór niewielkiej głębokości drzewa (3) oraz niskiego współczynnika uczenia zapobiega przeuczeniu modelu na

niewielkim zbiorze danych. Poniżej przedstawiono wykresy zależności liczby epok od uzyskanych parametrów.



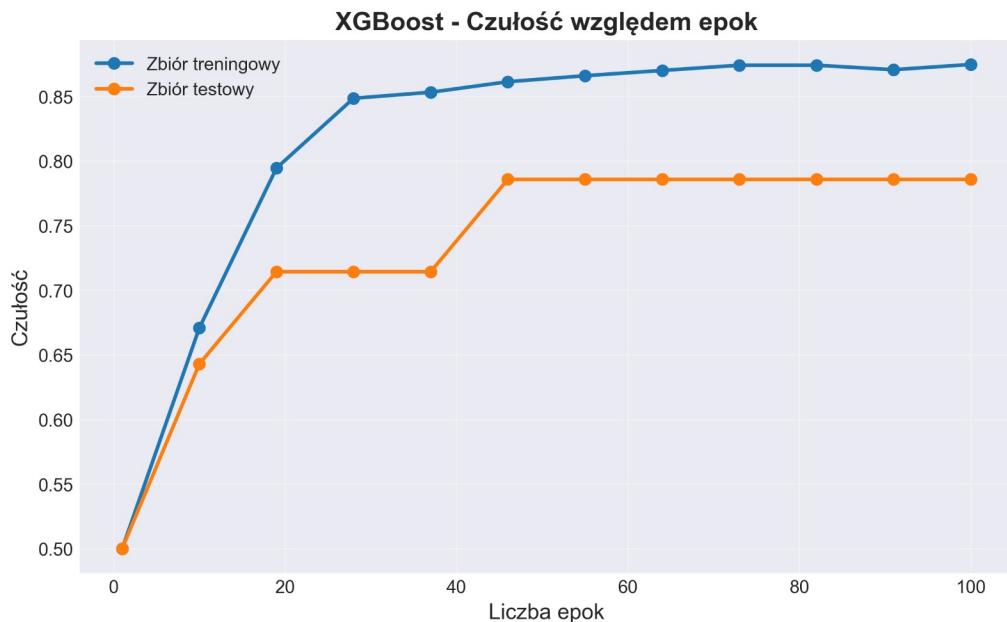
Rysunek 5.31. Wykres zależności dokładności względem liczby epok.

Analiza przebiegu uczenia wskazuje na stabilny i monotoniczny przyrost wiedzy modelu. Ze względu na niski współczynnik uczenia (0.01), krzywa dokładności wzrosła łagodnie. Brak widocznej dywergencji (rozchodzenia się) linii treningowej i testowej sugeruje, że model dobrze generalizuje wiedzę i nie ulega zjawisku overfittingu w badanym zakresie 100 epok. Poniższy wykres przedstawia zależność wartości precyzji od liczby epok.



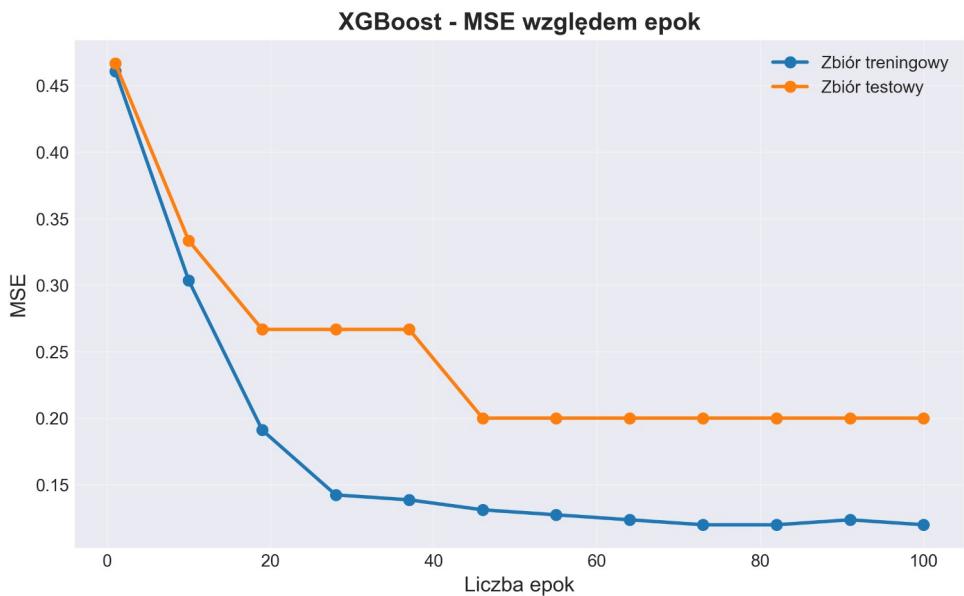
Rysunek 5.32. Wykres zależności precyzji względem liczby epok.

Wykres precyzji ujawnia wysoką zdolność modelu do minimalizowania fałszywych alarmów (False Positives). Wartości precyzji utrzymują się na wysokim poziomie przez cały proces treningowy. Poniższy wykres przedstawia zależność czułości względem numeru epoki.



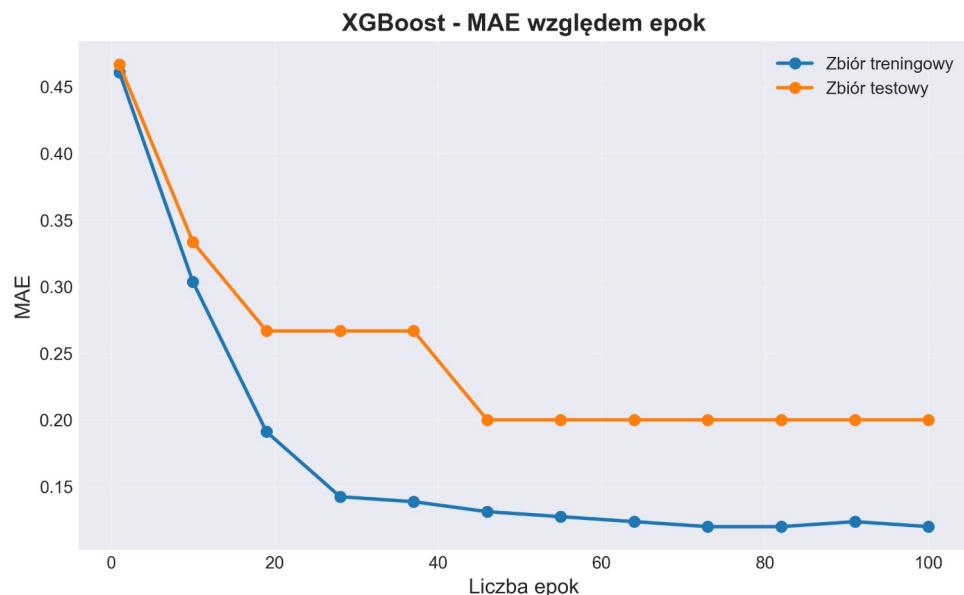
Rysunek 5.33. Wykres zależności czułości względem liczby epok.

Krzywa czułości wykazuje trend rosnący, co potwierdza, że z każdą kolejną iteracją (boostingiem) algorytm skuteczniej identyfikuje trudniejsze przypadki chorobowe. Na poniższym rysunku przedstawiono zależność błędu średniokwadratowego w zależności od numeru epoki.



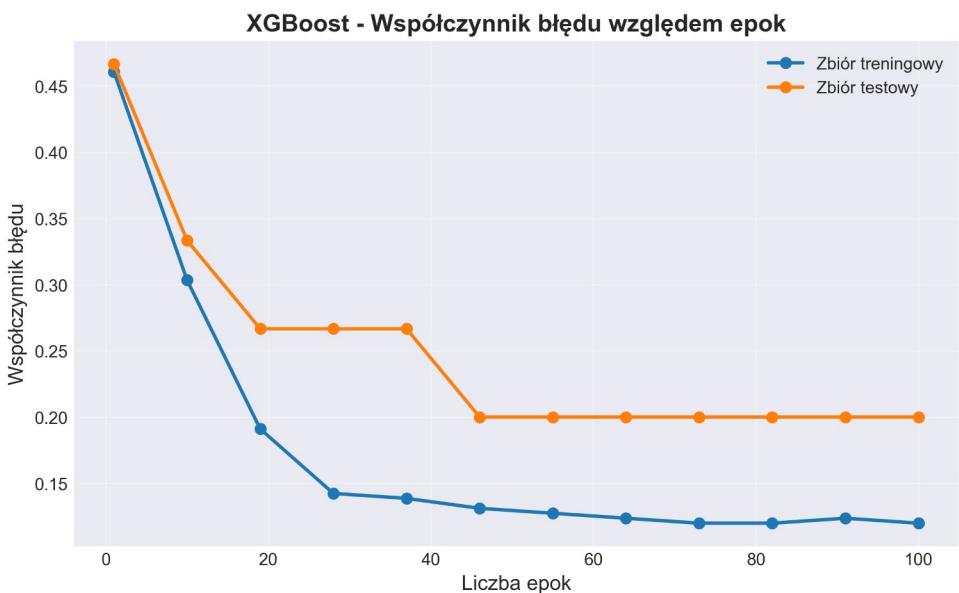
Rysunek 5.34. Wykres zależności średniego błędu kwadratowego względem liczby epok.

Wykres MSE potwierdza efektywność procesu minimalizacji funkcji straty. Błąd na zbiorze testowym spada systematycznie, osiągając satysfakcyjne minimum w końcowej fazie treningu (około 0,17). Poniższy rysunek przedstawia zależność średniego błędu bezwzględnego względem liczby epok.



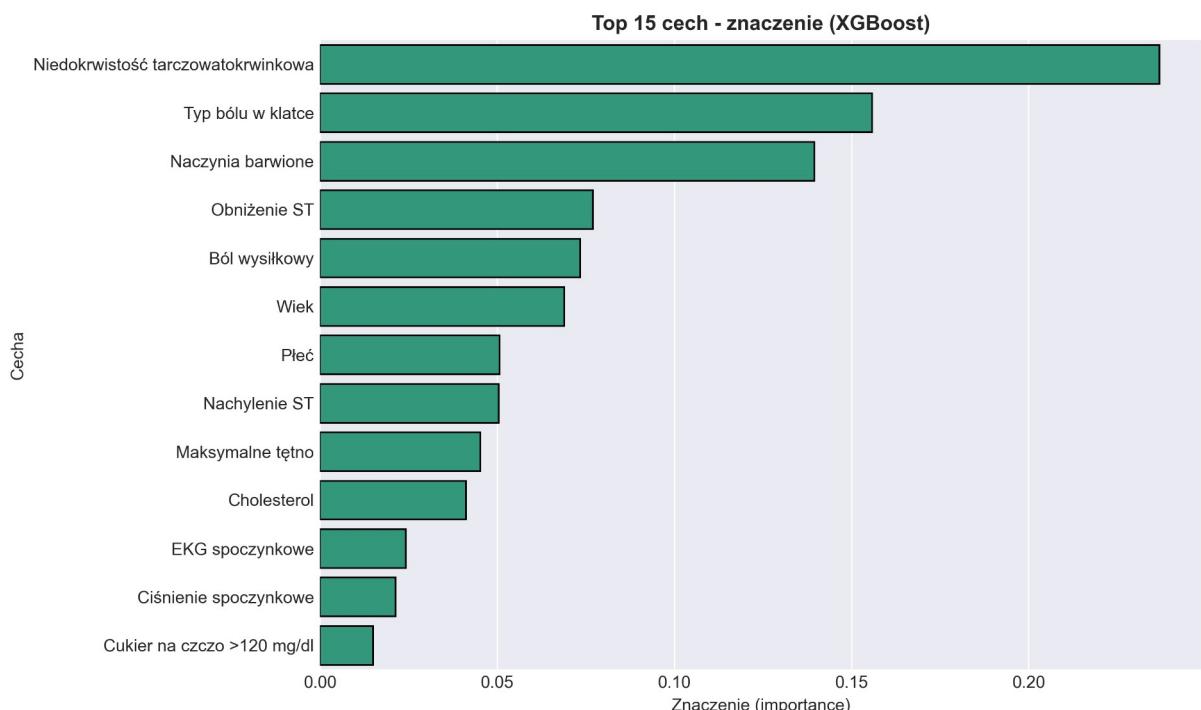
Rysunek 5.35. Wykres zależności średniego błędu bezwzględnego względem liczby epok.

Przebieg MAE jest zbliżony do charakterystyki MSE. Poniższy rysunek przedstawia wykres zależności współczynnika błędu uczenia/testowania względem numeru epoki.



Rysunek 5.36. Wykres zależności współczynnika błędu uczenia/testowania względem liczby epok.

Krzywa współczynnika błędu (Error Rate) wykazuje charakterystykę zbieżną z wykresem dokładności (będąc jego odwrotnością). W przeciwieństwie do modelu Lasu Losowego, XGBoost systematycznie redukuje błęd klasyfikacji aż do ostatnich epok, co sugeruje, że algorytm ten efektywnie wykorzystuje iteracyjne poprawianie wag dla trudnych przypadków. Poniższy rysunek przedstawia 15 najważniejszych cech w zbiorze.



Rysunek 5.37. Piętnaście najważniejszych cech w zbiorze „Heart Disease” uzyskanych w procesie uczenia za pomocą algorytmu XGBoost.

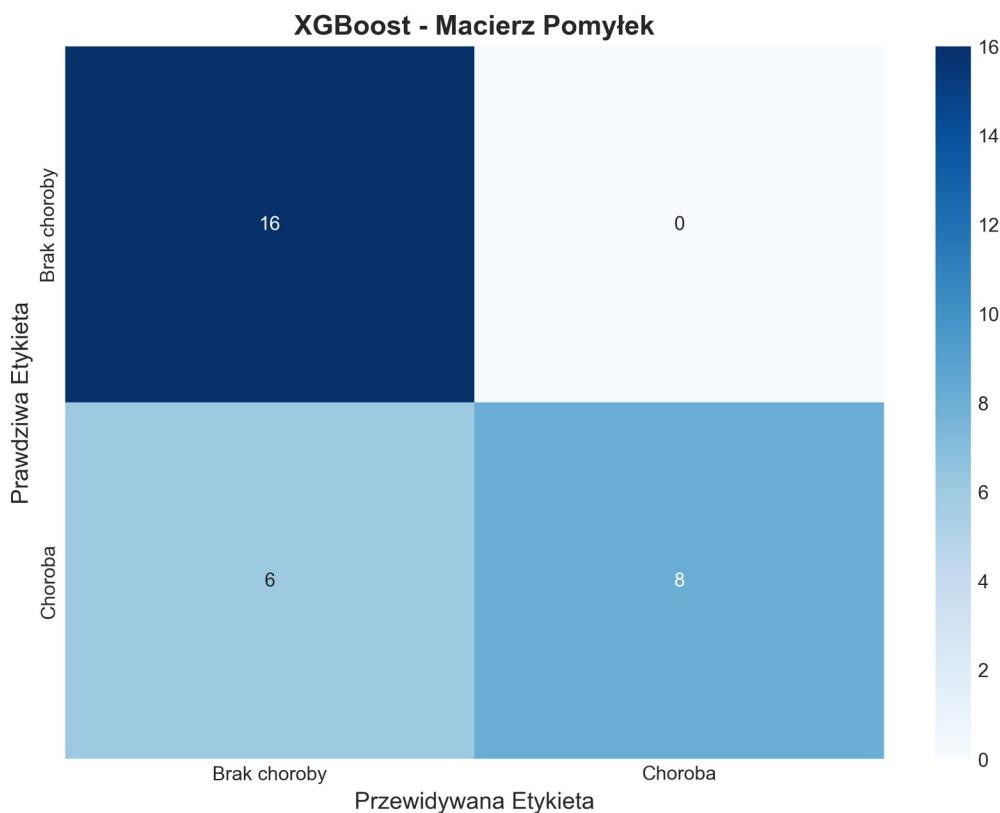
Analiza ważności cech dla XGBoost pozwala zidentyfikować kluczowe determinanty choroby serca. Cechy takie jak **Typ bólu w klatce (cp)**, **Niedokrwistość tarczowatokrwinkowa (thal)** oraz **Liczba naczyń barwionych (ca)** zazwyczaj dominują w procesie decyzyjnym drzewa. Co istotne, model ten potrafi wykorzystać nieliniowe interakcje między parametrami, co odróżnia go od prostszych metod statystycznych.

Kolejno wykonano testy modelu na zbiorze testowym. Poniższe rysunki oraz tabele przedstawiają uzyskane wyniki testów.

Tabela 5.16. Uzyskane wartości metryk dla zbioru testowego.

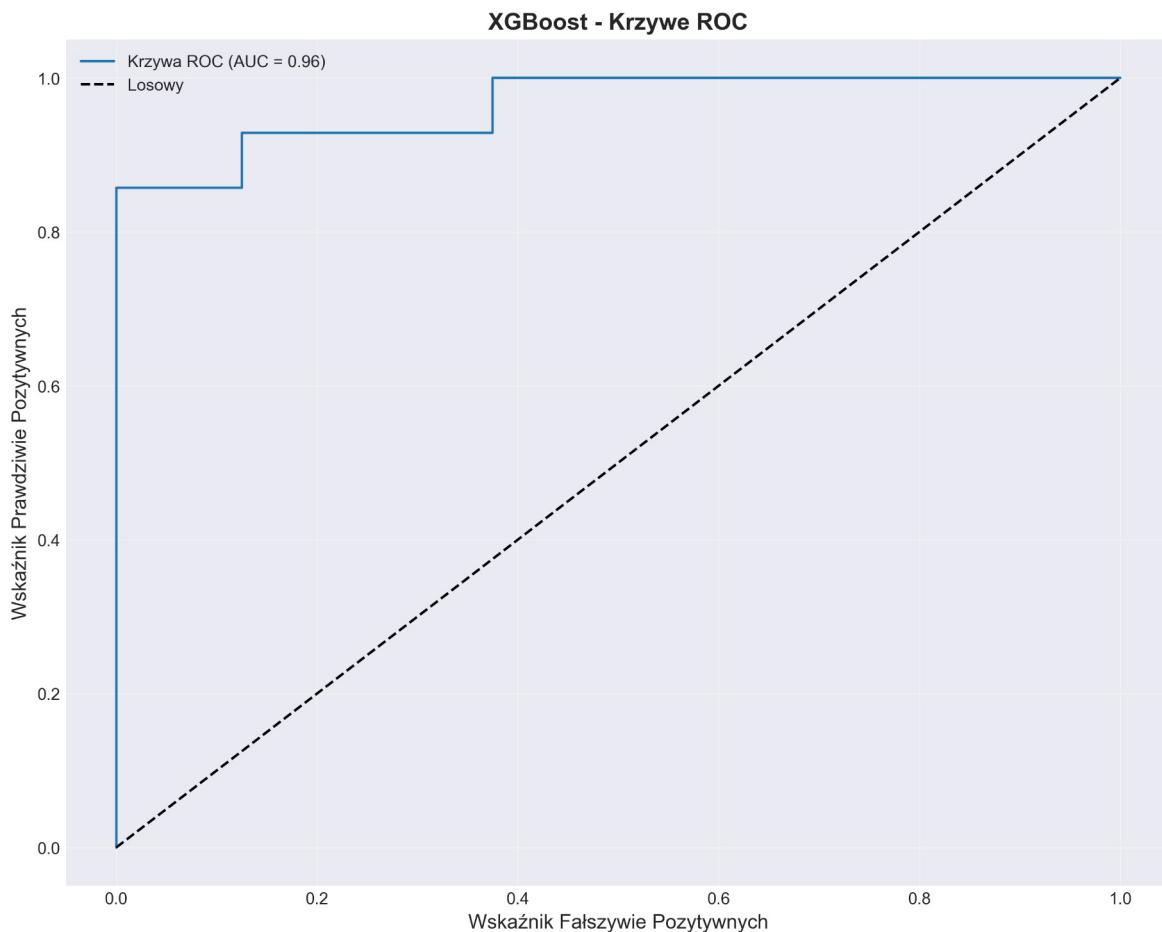
Metryka	Wartość
Dokładność	0.8000
Precyza (macro)	0.8636
Czułość (macro)	0.7857
Średni błąd kwadratowy (MSE)	0.2000
Średni błąd bezwzględny (MAE)	0.2000
Pierwiastek ze średniego błędu kwadratowego (RMSE)	0.4472

Wyniki na zbiorze testowym potwierdzają dobrą generalizację modelu XGBoost – dokładność (80,00%) jest zbliżona do tej uzyskanej na zbiorze treningowym. Uwagę zwraca bardzo wysoka wartość ROC-AUC (0,9643), która sugeruje, że model doskonale radzi sobie z rankingowaniem prawdopodobieństwa choroby. Poniższy rysunek przedstawia uzyskaną macierz pomyłek.



Rysunek 5.38. Macierz pomyłek dla zbioru „Heart Disease” oraz algorytmu XGBoost.

Macierz pomyłek ujawnia interesującą charakterystykę modelu: **100% skuteczność w klasyfikacji osób zdrowych** (brak False Positives). Oznacza to, że model jest niezwykle precyzyjny (Precyzja dla klasy 1 = 1.00). Jednakże, wystąpiły przypadki błędnie ujemne (False Negatives) – osoby chore zostały zaklasyfikowane jako zdrowe (6 przypadków). Sugeruje to, że model jest „konserwatywny” i decyduje się na diagnozę choroby tylko przy bardzo silnych przesłankach. Poniższy rysunek prezentuje krzywe ROC-AUC.



Rysunek 5.39. Krzywe ROC-AUC dla zbioru „Heart Disease” oraz algorytmu XGBoost.

Analiza krzywych ROC potwierdza bardzo wysoki potencjał separacyjny modelu. Wskaźnik AUC na poziomie **0,96** jest wynikiem wybitnym. Krzywa wznosi się gwałtownie ku lewemu górnemu rogowi, co dowodzi, że przy odpowiednim doborze punktu pracy (threshold) możliwe jest uzyskanie bardzo wysokiej czułości przy zachowaniu niskiego odsetka fałszywych alarmów.

Szczegółowa dekompozycja skuteczności modelu w podziale na klasy ujawnia asymetrię w wynikach. Poniższa tabela prezentuje wartość metryk dla wskazanych etykiet atrybutu wyjściowego dla zbioru testowego.

Tabela 5.17. Wartość metryk dla poszczególnych etykiet atrybutu wyjściowego dla zbioru testowego.

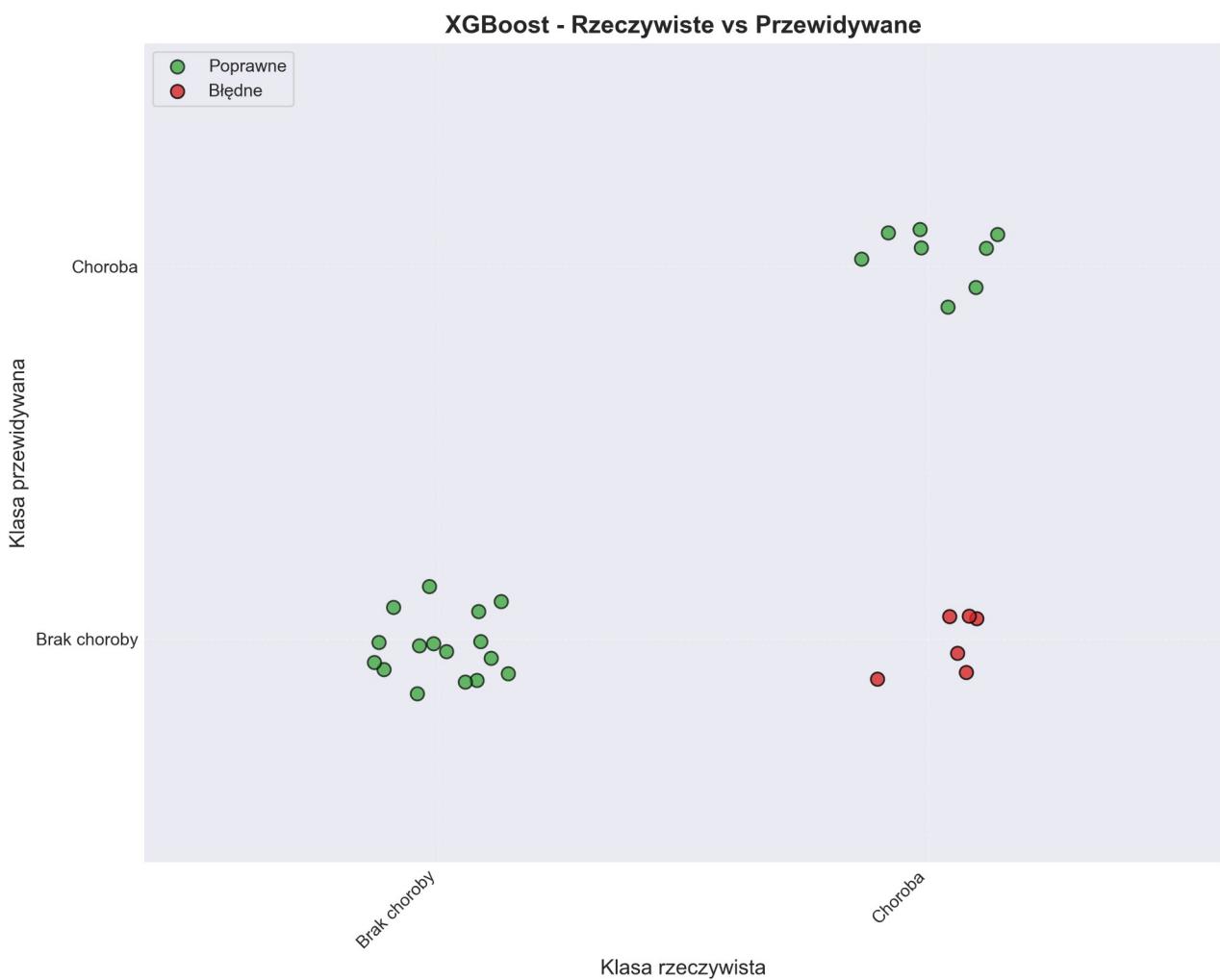
Etykieta	Precyzja	Czułość	Wartość miary F1
Brak choroby	0.73	1.00	0.84
Choroba	1.00	0.57	0.73

Zestawienie metryk jednoznacznie wskazuje na strategię modelu: maksymalizacja precyzyji dla klasy „Choroba” kosztem czułości. Model bezbłędnie identyfikuje pacjentów zdrowych, jednak ma trudności z wykryciem wszystkich przypadków chorobowych przy domyślnym progu decyzyjnym, klasyfikując część chorych jako zdrowych. Poniższa tabela prezentuje przykładowe predykcje wytrenowanego modelu.

Tabela 5.18. Przykładowe predykcje uzyskane dla zbioru „Heart Disease” oraz algorytmu XGBoost.

Klasa rzeczywista	Klasa przewidywana	Poprawna	Prawdopodobieństwo choroby
Brak choroby	Brak choroby	Tak	0.270
Brak choroby	Brak choroby	Tak	0.279
Brak choroby	Brak choroby	Tak	0.422
Choroba	Brak choroby	Nie	0.496
Brak choroby	Brak choroby	Tak	0.219
Choroba	Choroba	Tak	0.707
Choroba	Brak choroby	Nie	0.469

Analiza poziomu pewności (prawdopodobieństwa) pokazuje, że w przypadkach błędnych (False Negatives), model często oscyluje w granicach niepewności (np. 0.46–0.49), będąc blisko progu 0.50, co potwierdza tezę o możliwości poprawy wyników (zwiększenia czułości) poprzez korektę progu odcięcia w dół (np. do 0.45). Poniższy rysunek przedstawia wizualizację przewidywań klas w modelu.



Rysunek 5.40. Wizualizacja predykcji modelu utworzonego ze zbioru „Heart Disease” (Rzeczywiste vs Przewidywane).

Graficzne zestawienie wartości potwierdza obserwacje z macierzy pomyłek. Punkty zielone (poprawne) dominują wykres. Czerwone punkty (błędy) skupiają się wyłącznie w jednym obszarze (Rzeczywista: Choroba, Przewidywana: Brak choroby), co wizualnie podkreśla specyficzną naturę błędów modelu XGBoost w tym eksperymencie – wysoką swoistość przy umiarkowanej czułości dla domyślnych ustawień.

6. Podsumowanie

W ramach zrealizowanego projektu przeprowadzono kompleksową analizę statystyczną oraz proces klasyfikacji danych numerycznych, koncentrując się na ocenie jakości wina w oparciu o jego parametry fizykochemiczne. Analiza eksploracyjna zbioru „Wine Quality” ujawniła istotne niezbalansowanie klas decyzyjnych, z wyraźną dominacją ocen średnich (5 i 6), które stanowiły ponad 76% wszystkich obserwacji. Badanie korelacji wykazało, że najsilniejszym predyktorem wpływającym pozytywnie na jakość trunku jest zawartość alkoholu, podczas gdy parametry takie jak gęstość czy kwasowość lotna wykazują korelację ujemną. Te wstępne wnioski znalazły swoje odzwierciedlenie w późniejszym procesie uczenia maszynowego, gdzie oba zastosowane algorytmy – Random Forest oraz XGBoost – zidentyfikowały te cechy jako kluczowe dla podejmowania decyzji klasyfikacyjnych.

Porównanie wydajności badanych modeli wykazało ich zbliżoną skuteczność ogólną, oscylującą w granicach 68% dokładności na zbiorze testowym. Algorytm Random Forest przyjął strategię bardziej konserwatywną, charakteryzującą się wyższą precyją i minimalizacją fałszywych alarmów, co jednak odbyło się kosztem niższej czułości w wykrywaniu klas rzadszych. Z kolei XGBoost, mimo nieco niższej precyji, wykazał lepsze zdolności separacyjne, co potwierdziły wyższe wartości krzywych ROC-AUC dla większości klas oraz poprawa detekcji w obszarze oceny 4. Analiza ważności cech w modelu XGBoost uwypukliła rolę kwasowości lotnej, degradując jednocześnie znaczenie całkowitego dwutlenku siarki, co sugeruje, że algorytm ten skuteczniej modeluje wpływ parametrów bezpośrednio kształtujących odczucia smakowe.

Należy podkreślić, że głównym ograniczeniem w osiągnięciu wyższej skuteczności klasyfikacji była struktura samego zbioru danych. Zarówno Random Forest, jak i XGBoost nie były w stanie poprawnie zidentyfikować przypadków skrajnych (oceny 3 i 9) ze względu na ich marginalną reprezentację w zbiorze uczącym. Mimo to, niskie wartości błędu średniego bezwzględnego (MAE na poziomie ok. 0,36) dowodzą, że pomyłki modeli są niewielkie i zazwyczaj dotyczą klas sąsiednich. Oznacza to, że opracowane systemy są stabilne i mogą stanowić użyteczne narzędzie wspomagające proces oceny enologicznej, skutecznie odróżniając wina słabsze od lepszych w obrębie typowych kategorii, choć do pełnej automatyzacji procesu uwzględniającej wina unikatowe konieczne byłoby zastosowanie technik równoważenia klas lub rozbudowa bazy danych.

LITERATURA

1. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C56S3T>.
2. Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C52P4X>.