P O L I T E C H N I K A Ś W I Ę T O K R Z Y S K A

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 SO2	Cecha wejściowa	Numeryczny	-	Nie
total_sulfur_dioxide	Całkowity SO2	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
Całkowity 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”

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 liczebności n genererowanych jest B zbiorów danych D_1, D_2, … , 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:

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

gdzie:

- $\hat{y}$ - przewidywany wynik końcowy modelu,
- $T_1(x), T_2(x), \ldots, T_B(x)$ - wyniki poszczególnych drzew decyzyjnych dla próbki x,
- B – liczba drzew w lesie,
- 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

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 b łędu 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

pandas

numpy

scipy

scikit-learn

matplotlib

seaborn

jupyter

os

sys

warnings

time

abc

3. Opis najważniejszych fragmentów kodu

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%)

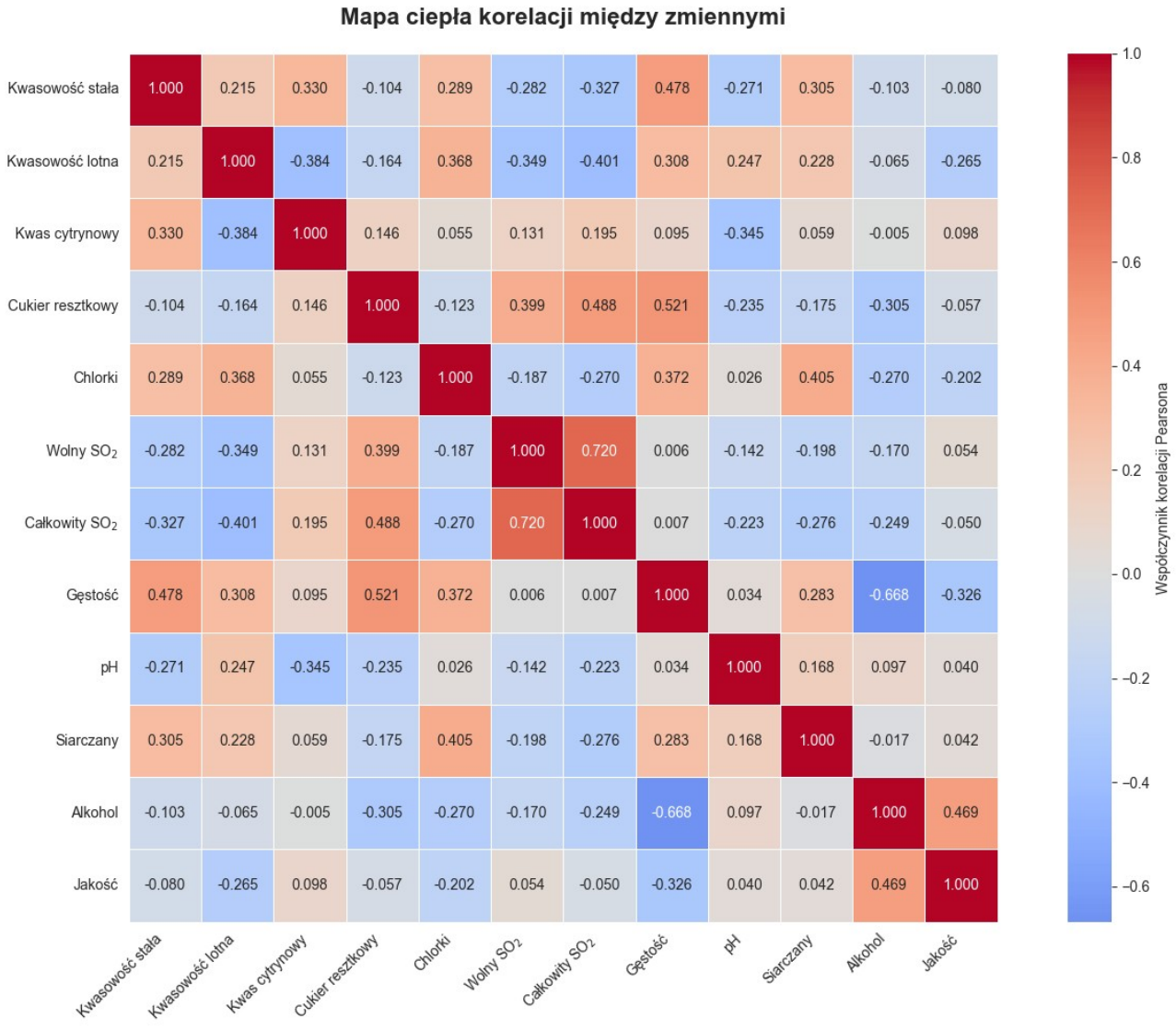
Wyniki zapisane w powyższej tabeli wykazują, iż najbardziej nieregularny rozkład wartości występuje w atrybucie kwas cytrynowy, gdzie wartości odchodzące od normy stanowią odsetek 7.83 procenta wszystkich wartości. Kolejno, w procesie analizy obliczono wartości odchylenia standardowego, średniej, mediany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

Tabela 4.2. Statystyki opisowe.

Atrybut	**Średnia**	**Odchylenie standardowe**	**Mediana**	**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 prawostronnie 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 SO_2 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.



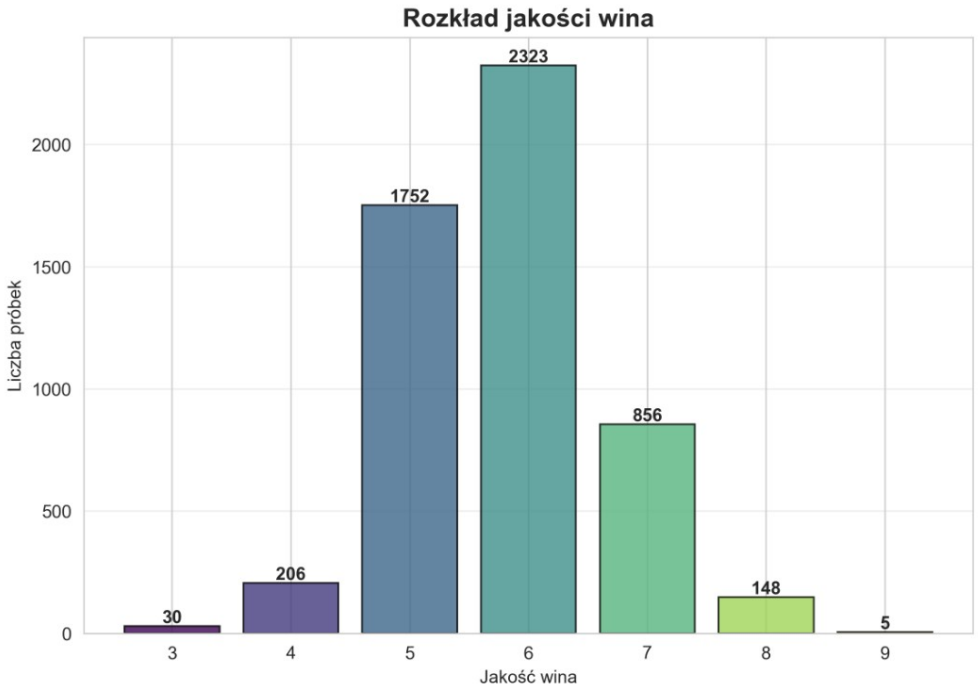
Rysunek 4.1. Wyniki uzyskane podczas analizy korelacji liniowej Pearsona.

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 * Alkohol + 2.12	22.0%	0.777	Dodatni
Gęstość	Jakość = - 96.84 * Gęstość + 102.11	10.7%	0.832	Ujemny
Kwasowość lotna	Jakość = - 1.39 * Kwasowość lotna + 6.27	7.0%	0.848	Ujemny
Model wielokrotny	Jakość = 0.39 * Alkohol + 30.31 * Gęstość – 1.37 * 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.

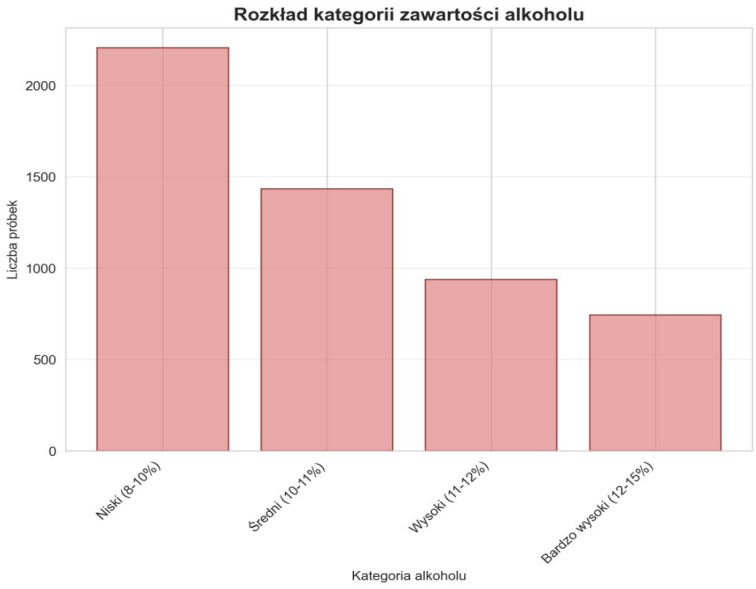


Rysunek 4.2. Rozkład jakości wina.

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.

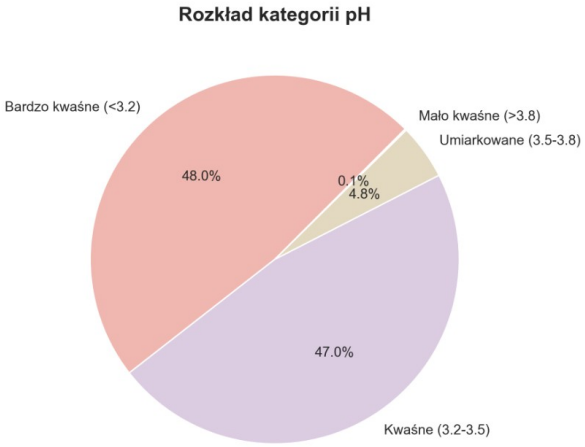
Rysunek 4.3. Wykres kołowy reprezentujący 10 najlepszych wartości siarczanów.

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.



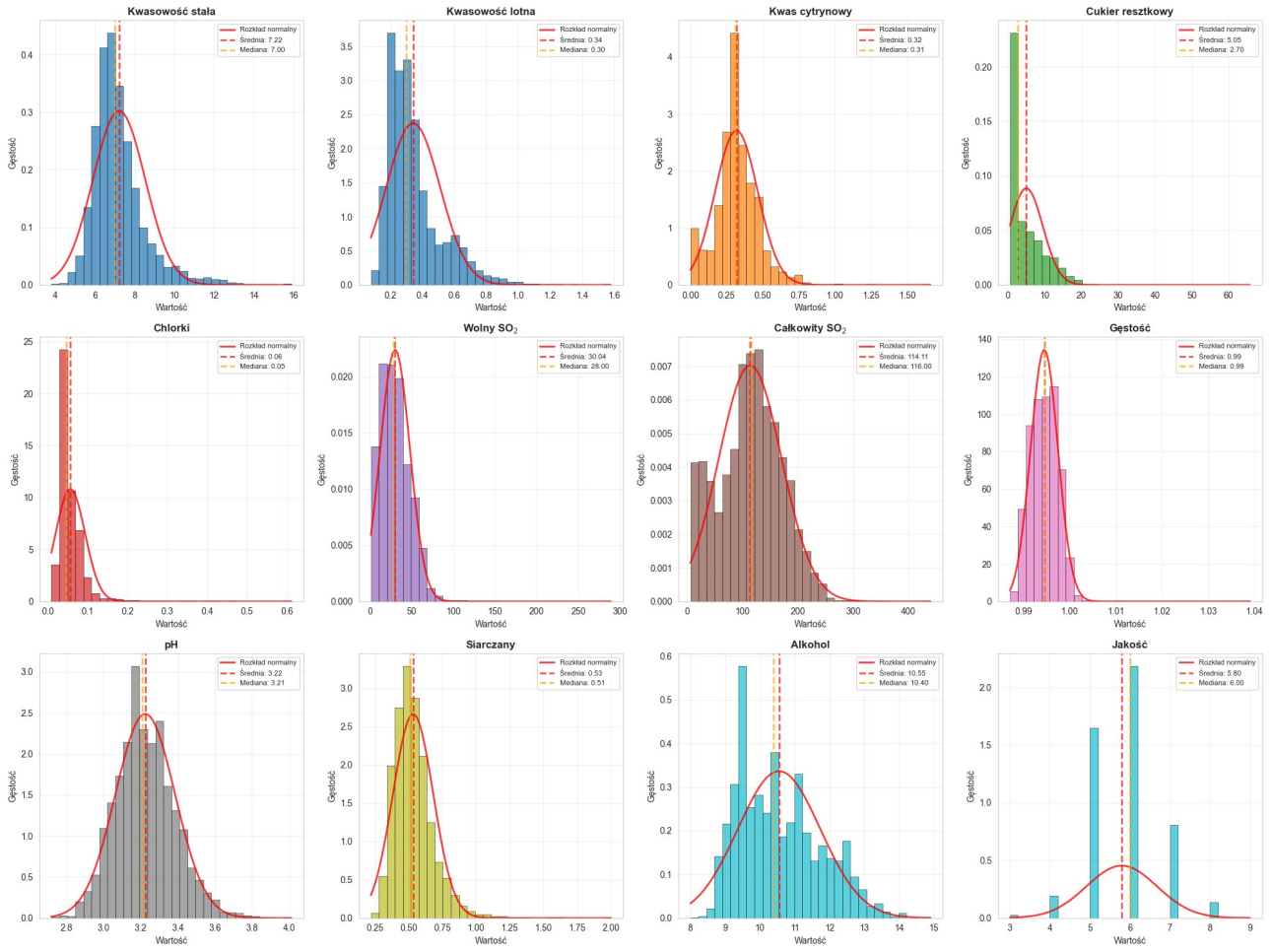
Rysunek 4.4. Rozkład kategorii zawartości alkoholu.

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.



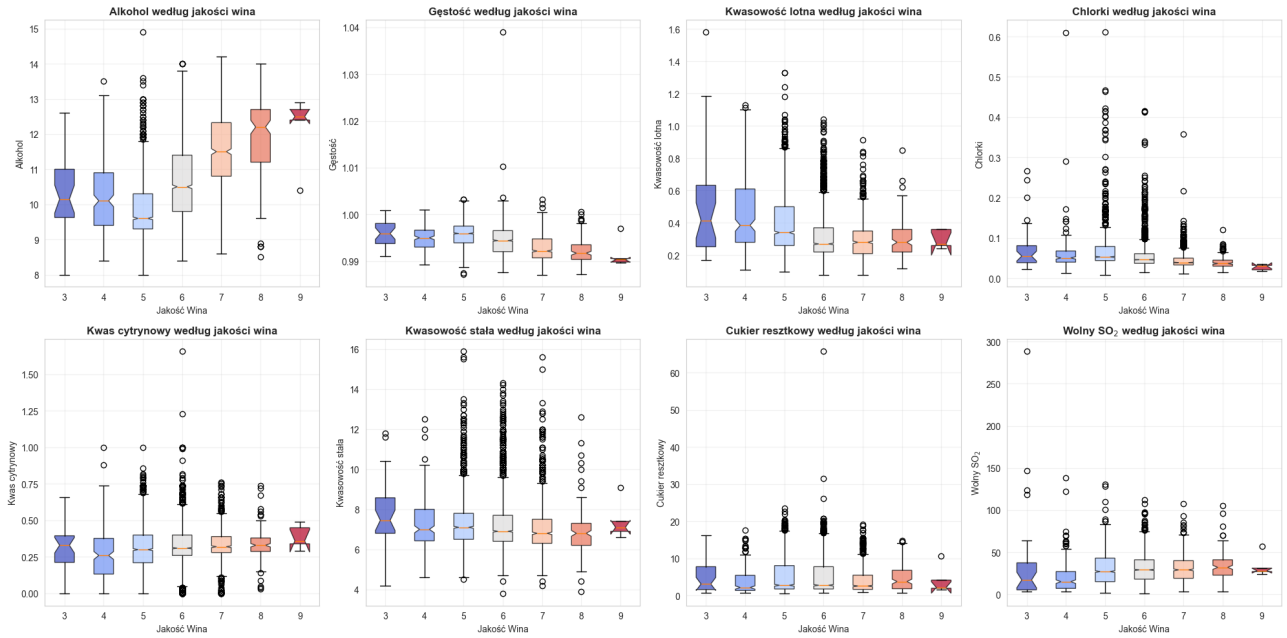
Rysunek 4.5. Rozkład kategorii pH.

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.



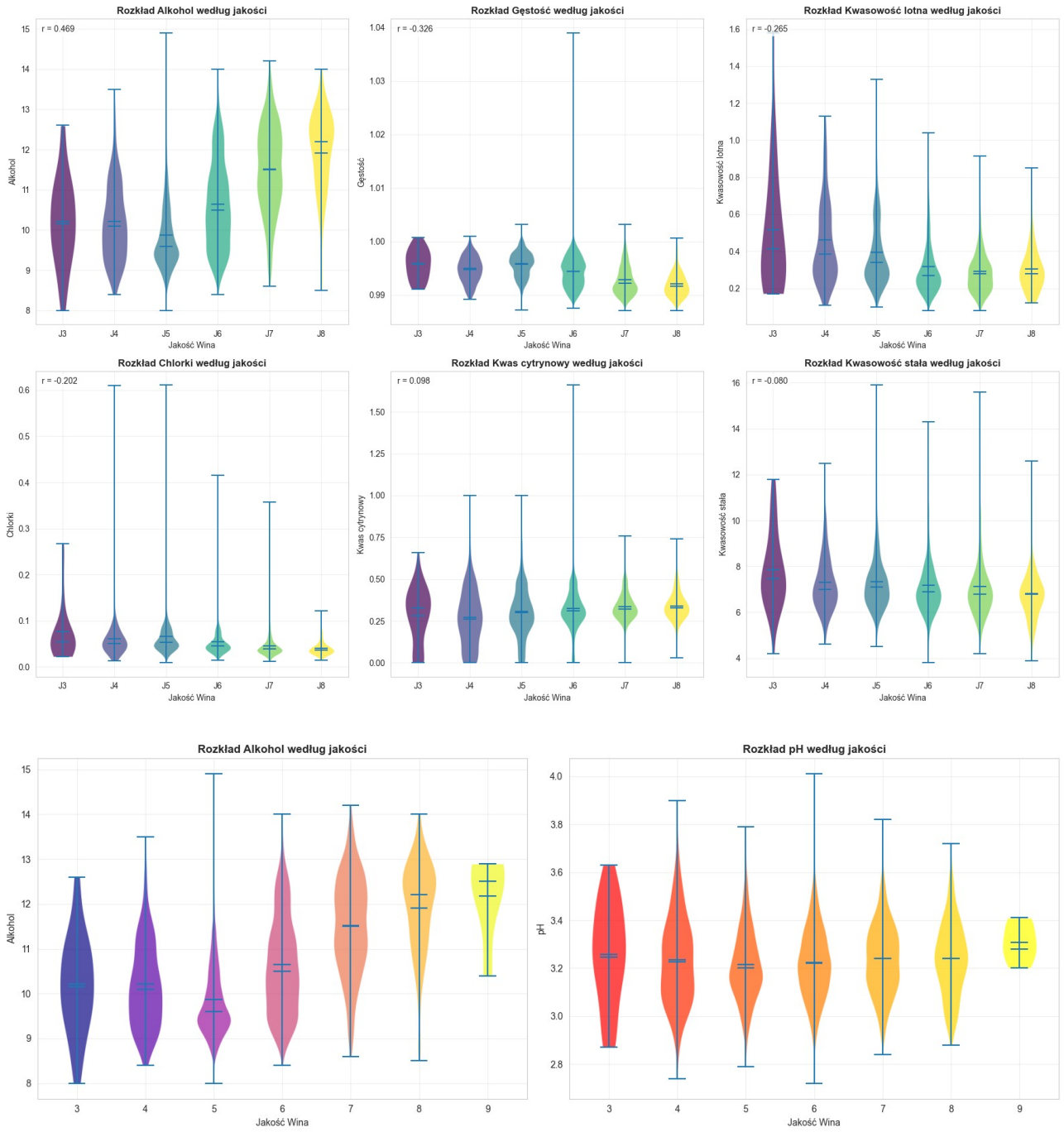
Rysunek 4.6. Histogramy pojedynczych atrybutów w zbiorze.

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 SO_2 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 pudełkowe zależności poszczególnych atrybutów od jakości wina.

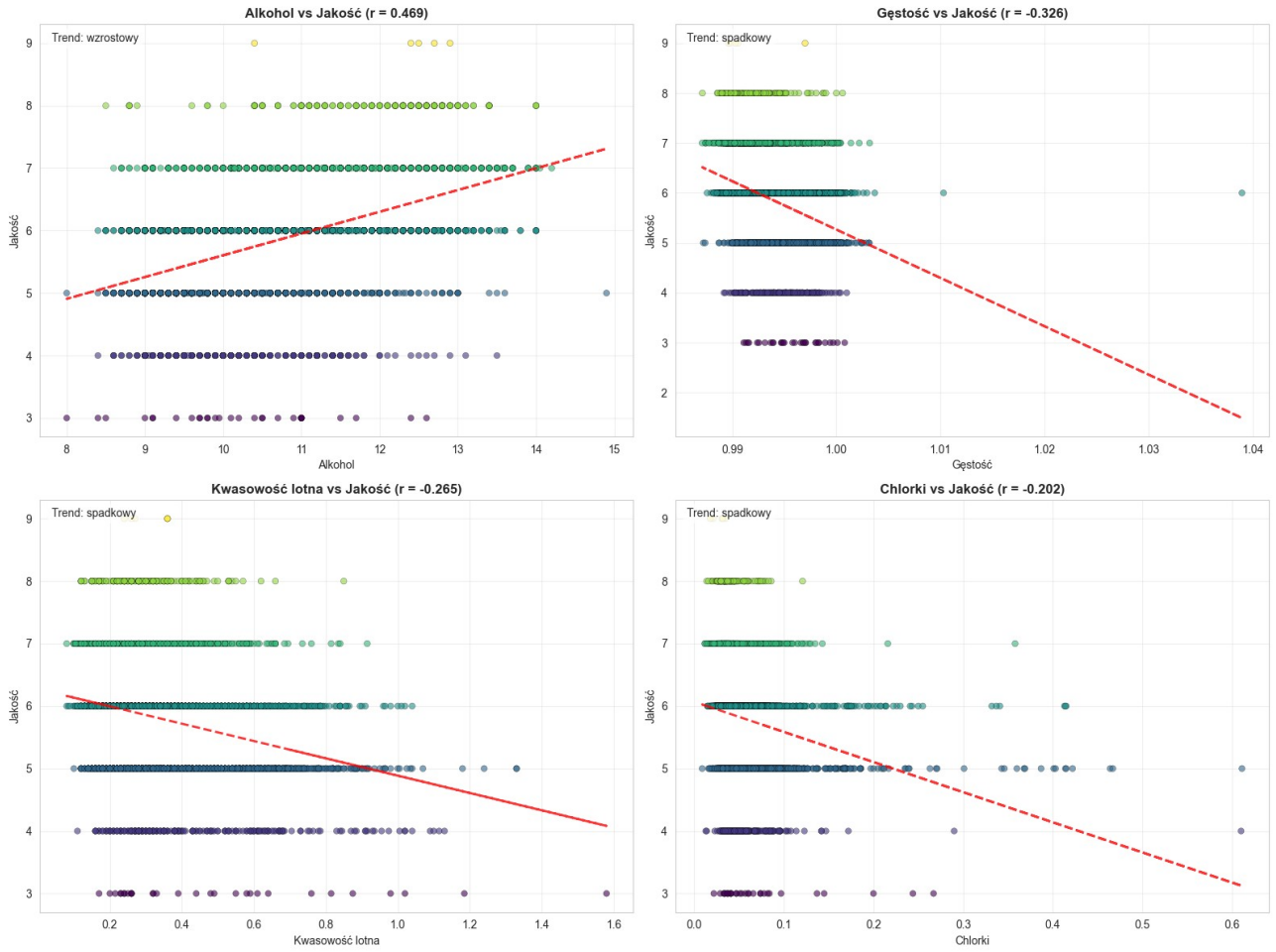
Prezentowane wykresy pudełkowe wizualizują rozkład parametrów fizykochemicznych wina w podziale na oceny jakości (od 3 do 9), pozwalając na ocenę tendencji centralnej oraz rozrzutu danych. Każde "pudełko" obejmuje środkowe 50% obserwacji (rozstęp międzykwartylowy), a pozioma linia wewnątrz wskazuje medianę, 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 (anomalie), które są najliczniejsze w klasach średnich (5 i 6), co świadczy o dużej niejednorodności win w tym segmencie cenowym. Wysokość pudełek informuje o stabilności procesu – spłaszczone pudełka 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 znikomą 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 górnych 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).

4.1. Zbiór „Heart Disease”

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 ≈ 68%), niska wartość F1-score (≈ 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 niezbalansowania 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),

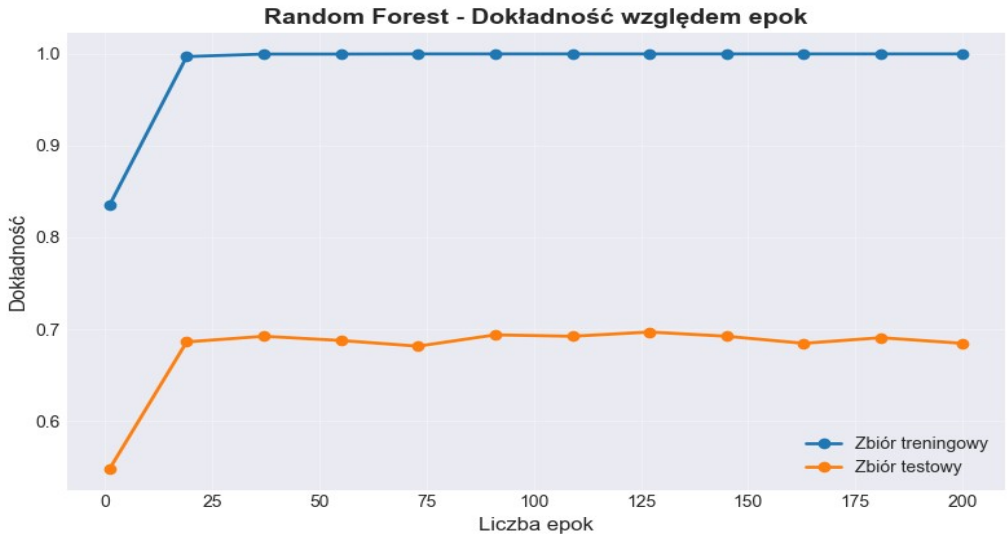
- minimalna liczba próbek w liściu – 1,
- minimalna liczba próbek wymagana do dokonania podziału – 2,
- liczba drzew w lesie – 200.

Poniższa tabela przedstawia uzyskane wartości metryk dla wskazanych parametrów.

Tabela 5.2. Wartości metryk uzyskane dla najlepszych parametrów modelu.

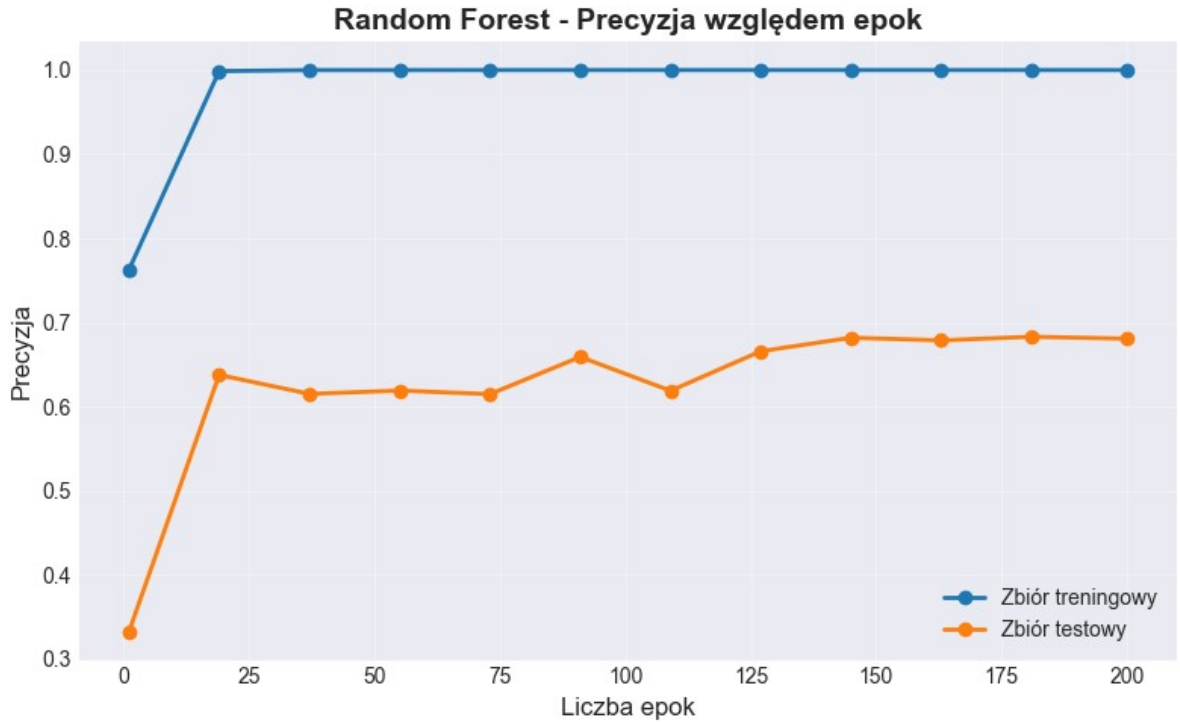
Metryka	Wartość
Średnia dokładność	0.6891
Średnia precyzja	0.5441
Średnia czułość	0.3608
Średni błąd kwadratowy	0.4293
Średni błąd bezwzględny	0.3482

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.

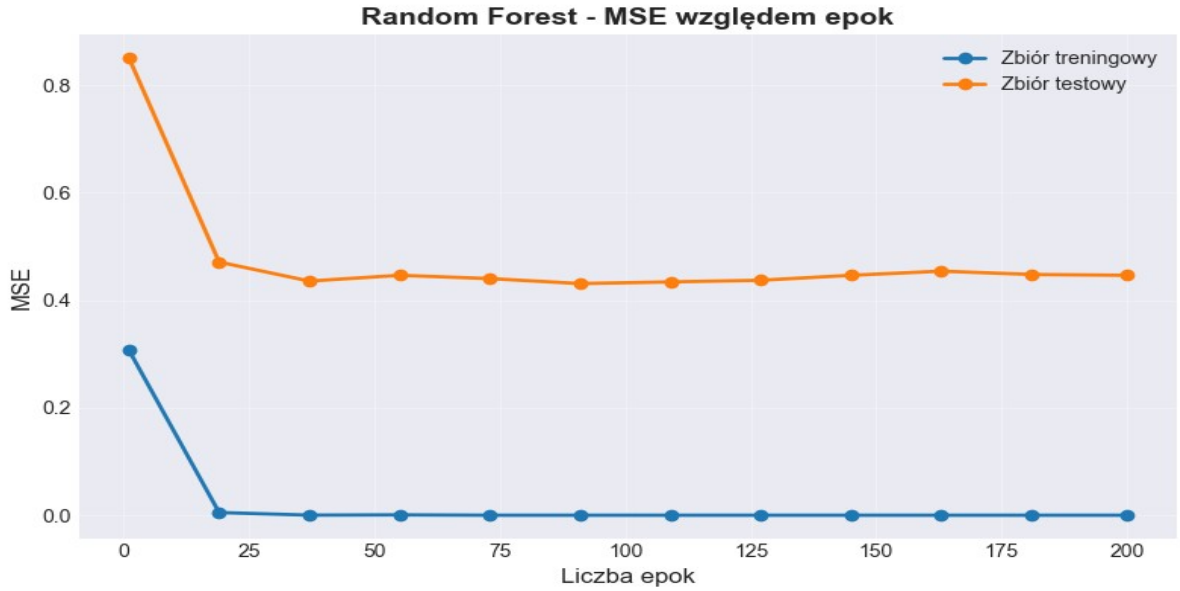


Rysunek 5.2. Wykres zależności precyzji względem liczby epok.

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.

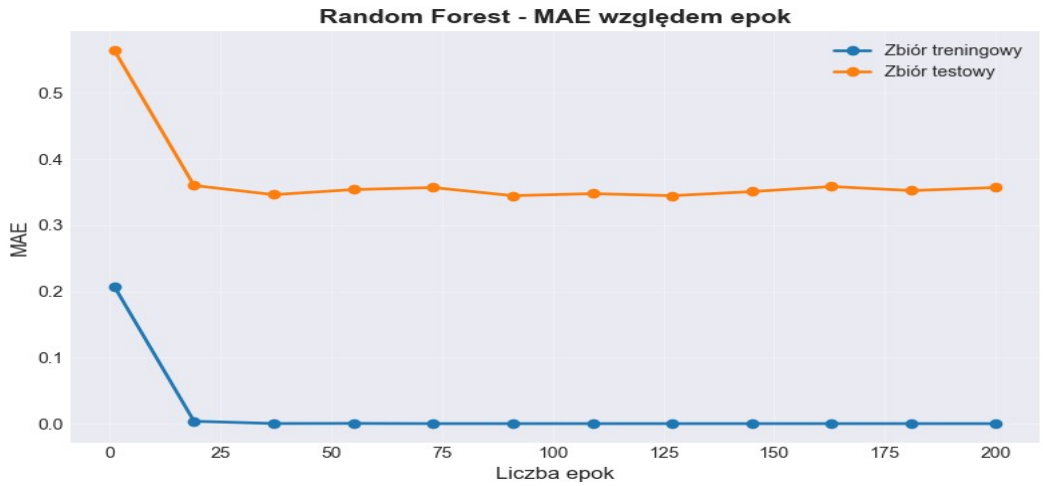
Przebieg krzywej czułości zawartej na powyższym rysunku charakteryzuje się szybkim nasyceniem. Po fazie wzrostowej w pierwszych 20 epokach, metryka osiąga poziom asymptotyczny w granicach 43% na zbiorze testowym. Dalszy proces treningowy nie przynosi przyrostu czułości, co sugeruje, że model wyczerpał 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.

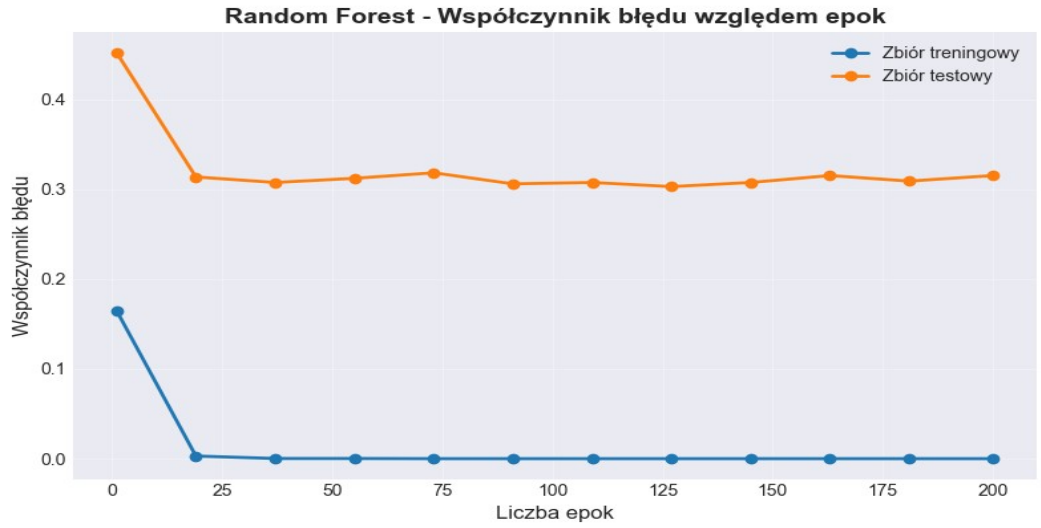
Na powyższym wykresie widoczne jest, iż po około 40 epokach wartość błędu średniokwadratowego na zbiorze testowym zatrzymuje się w przedziale 0.43-0.45 i pozostaje niewrażliwa na dalszy przebieg uczenia. Widoczna, duża i stała odległość między linią niebieską

(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.

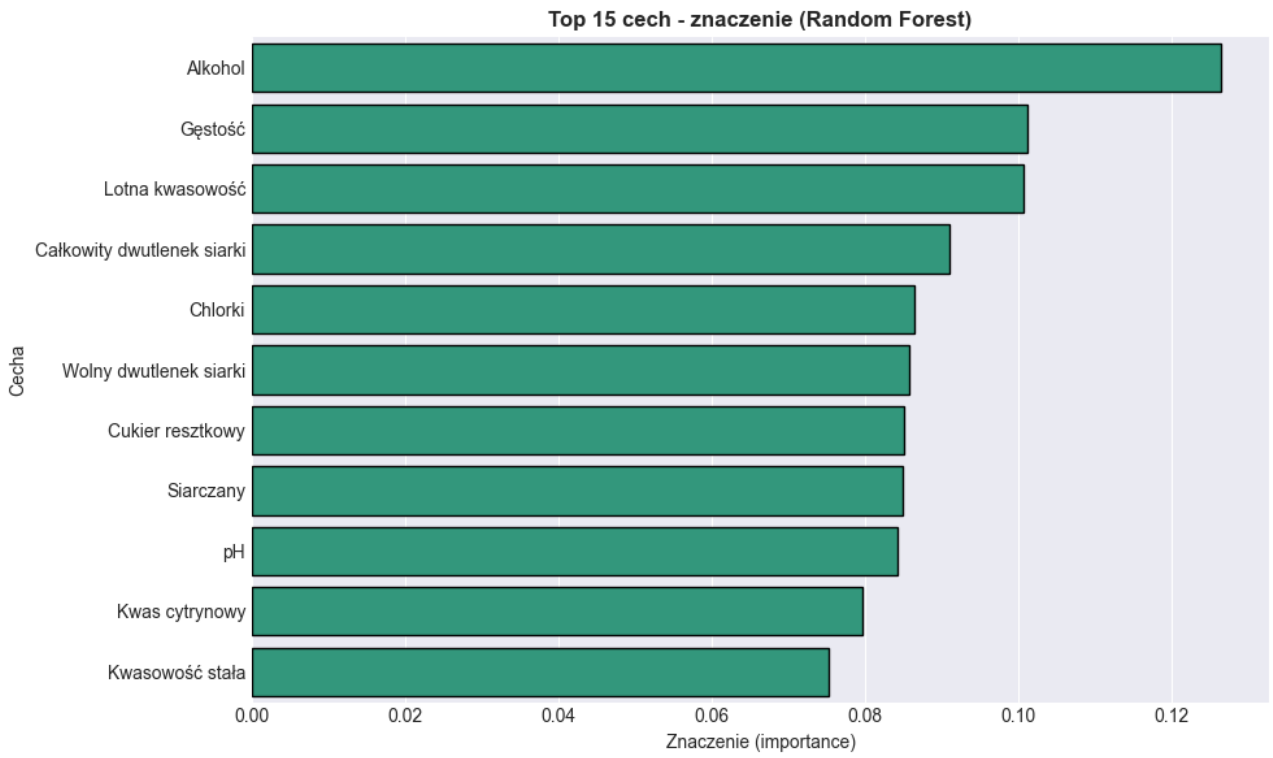
Z powyższego rysunku można odczytać, że Wstępna faza uczenia (pierwsze 25 epok) przynosi drastyczną redukcję średniego błędu na zbiorze testowym do poziomu ok. 0.35. Po tym punkcie krzywa staje się płaska. Oznacza to, że wydłużanie treningu o kolejne epoki nie sprawia, że model myli się mniej – błąd 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.



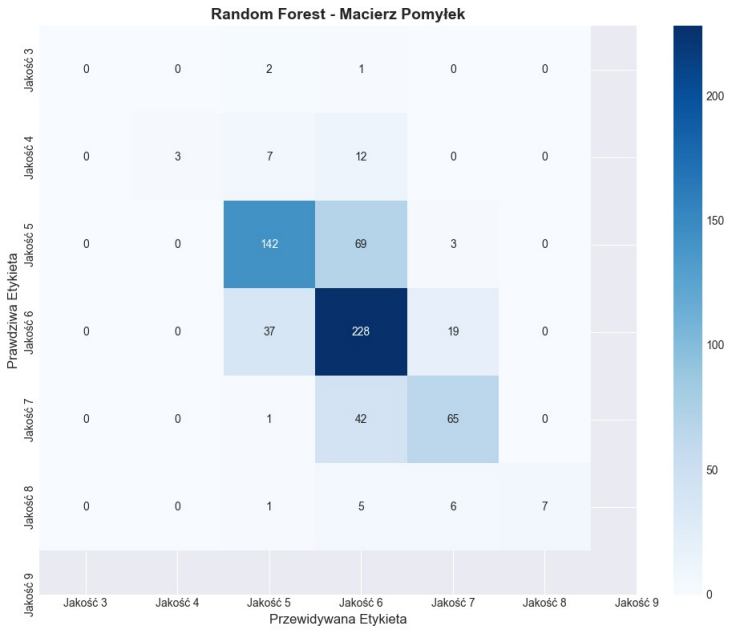
Rysunek 5.7. Piętnaście najważniejszych cech w zbiorze „Wine Quality" uzyskanych w procesie uczenia za pomocą algorytmu Random Forest.

Analiza ważności cech jednoznacznie wskazuje na alkohol jako dominujący predyktor, którego wpływ na decyzje modelu (waga > 0,12) znacząco przewyższa pozostałe parametry. Na kolejnych miejscach uplasowały 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.

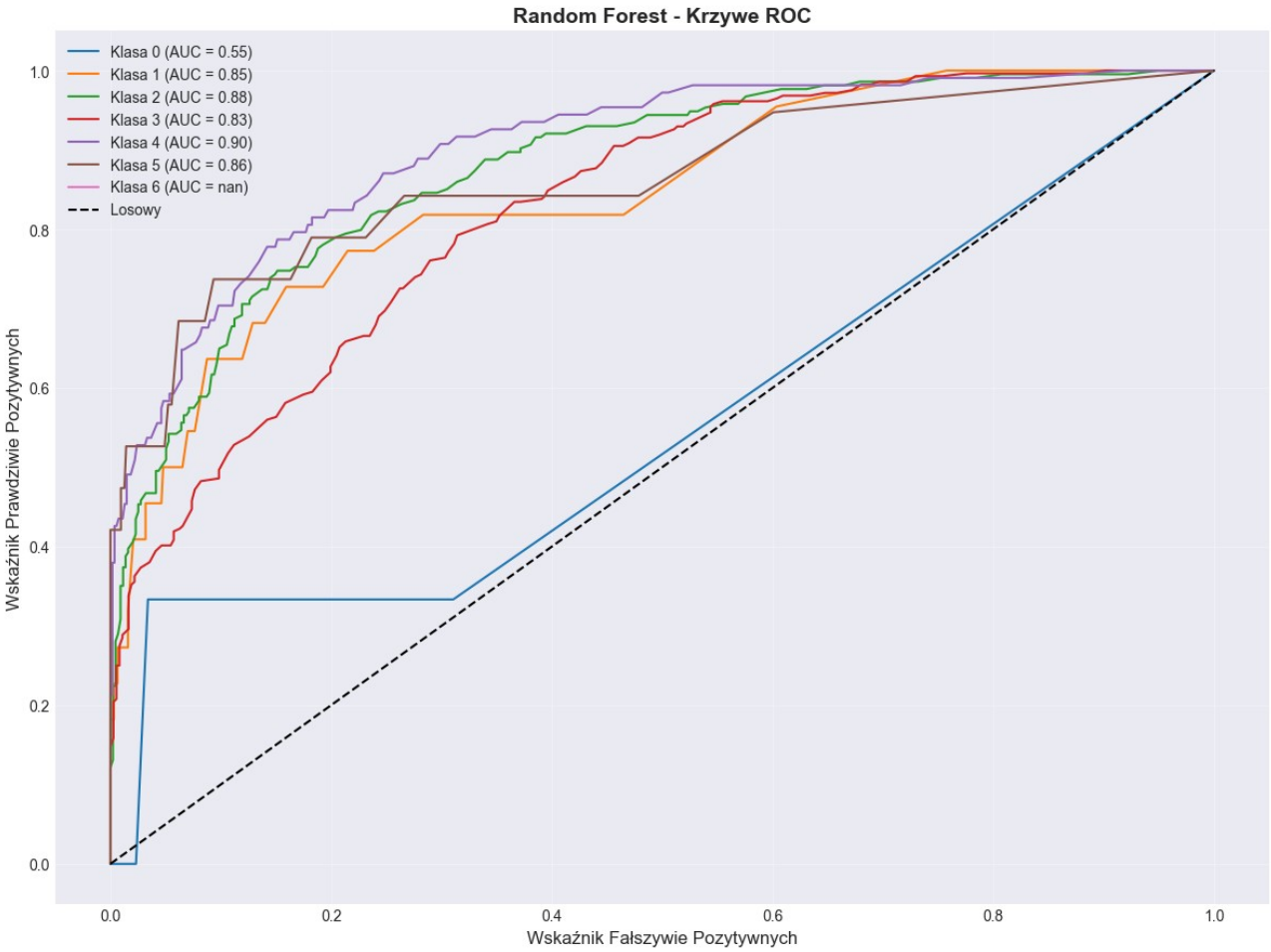
Metryka	Wartość
Dokładność	0.6846
Precyzja	0.6808
Czułość	0.4288
Średni błąd kwadratowy	0.4462
Średni błąd bezwzględny	0.3569
Pierwiastek ze średniego błędu kwadratowego	0.6679

Ostateczna weryfikacja na zbiorze testowym potwierdza zdolność generalizacji modelu, który osiągnął dokładność na poziomie 68,46%, przy zbliżonej wartości precyzji (68,08%). Istotna dysproporcja między wysoką precyzją 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 błędu średniokwadratowego (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 odcięcia. 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.

Etykieta	Precyzja	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 rzadszych (4 i 8) model przyjmuje strategię konserwatywną, co objawia się maksymalną precyzją (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) znacząco 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 ≈ 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,

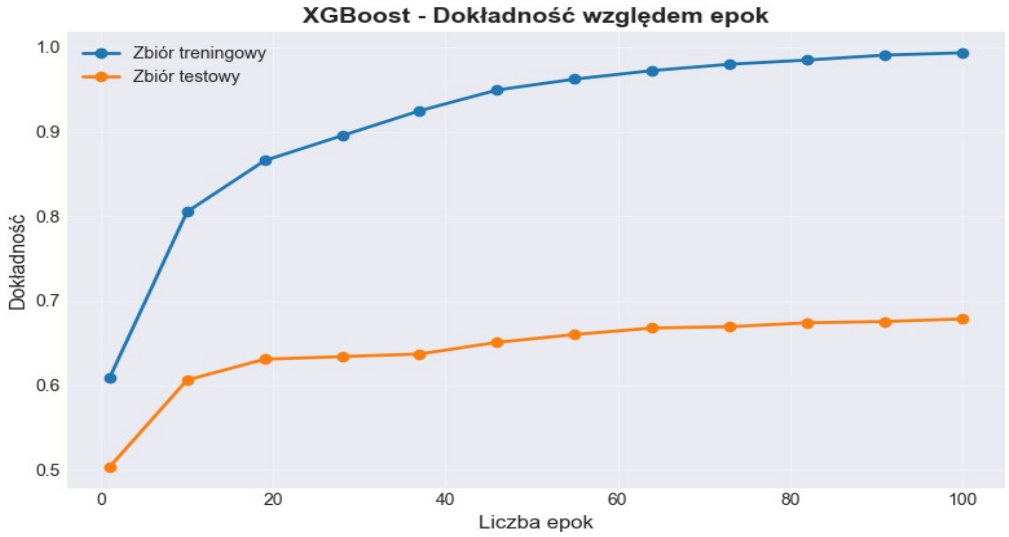
- liczba drzew w lesie – 100,
- ułamek próbek treningowych losowany do zbudowania każdego drzewa – 0.8.

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 precyzja	0.4984
Średnia czułość	0.3581
Średni błąd kwadratowy	0.4438
Średni błąd bezwzględny	0.3566

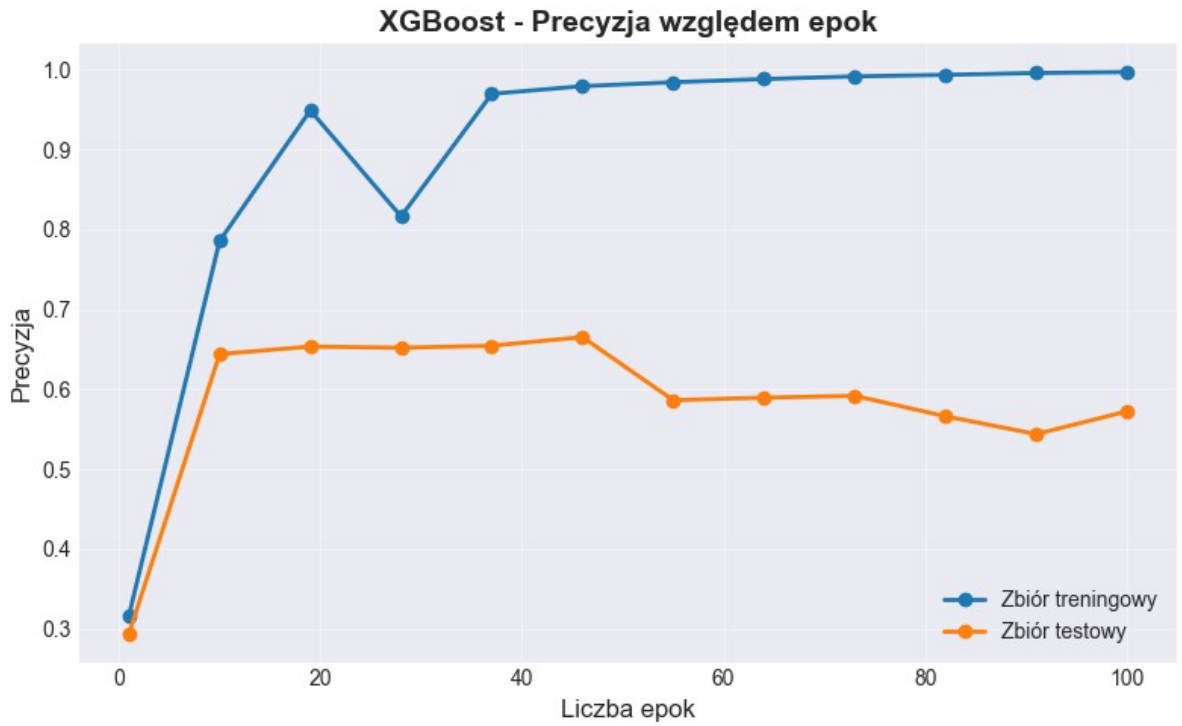
Proces optymalizacji hiperparametrów pozwolił na podniesienie skuteczności modelu. Po dostrojeniu (m.in. głębokość drzewa = 9, learning rate = 0.1), średnia dokładność wzrosła do 68,36%. Najważniejszym wskaźnikiem jest tutaj niski średni błąd bezwzględny (MAE = 0,3566). Oznacza on, że model po optymalizacji "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.

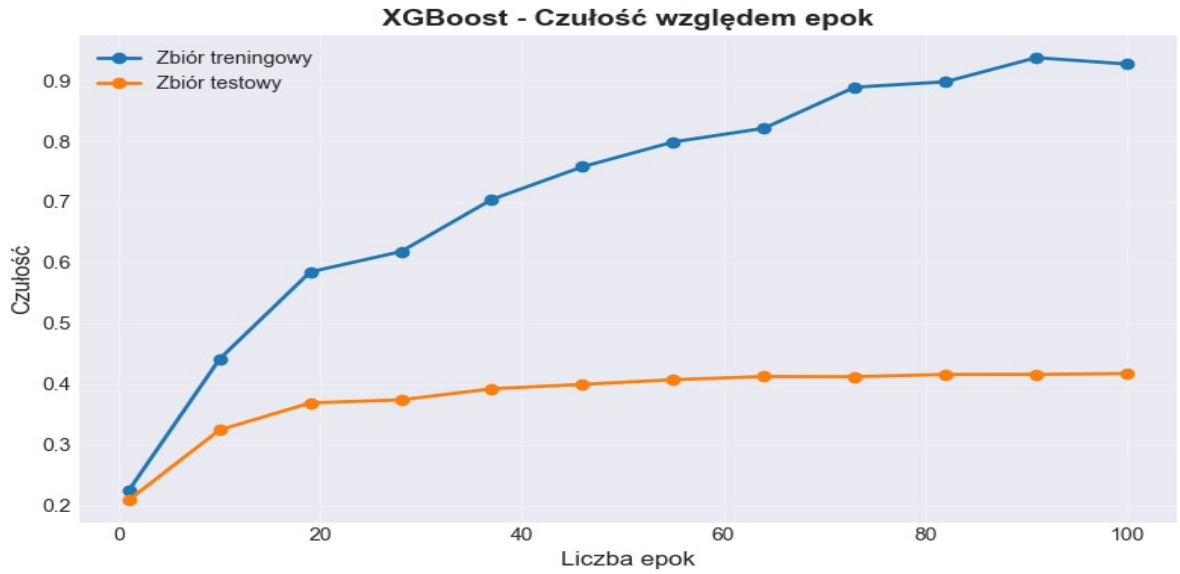
Analiza przebiegu uczenia wskazuje na bardziej stopniowy charakter przyrostu wiedzy w porównaniu do Lasów Losowych. Dokładność na zbiorze testowym rośnie sukcesywnie, osiągając stabilizację dopiero w okolicy 60–80 epoki na poziomie ok. 68%. Brak gwałtownego skoku w

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.



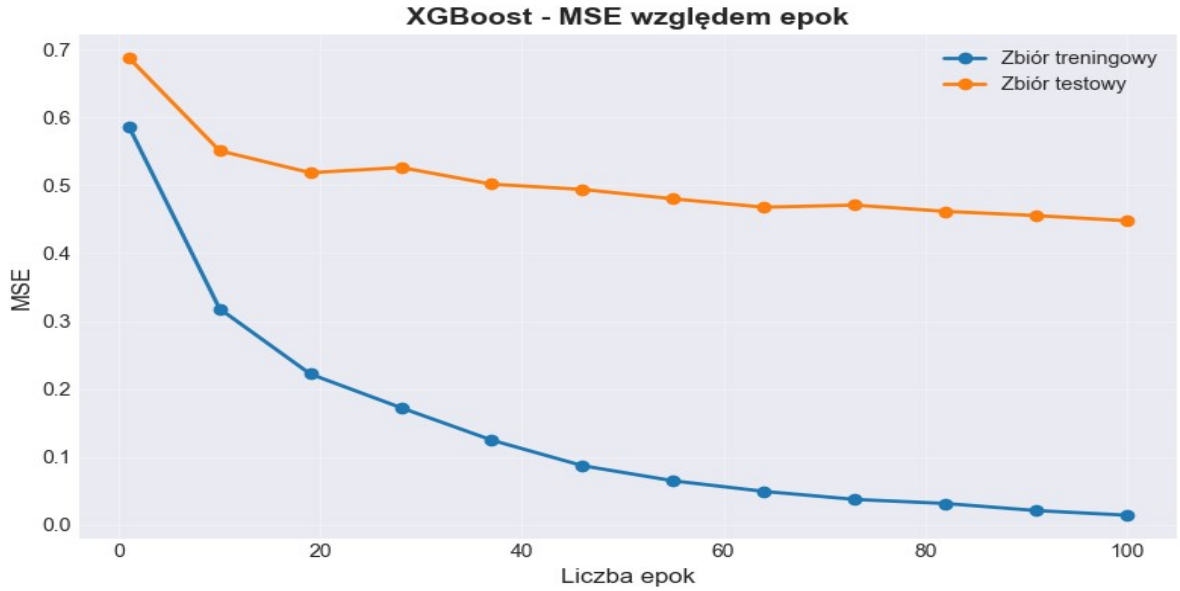
Rysunek 5.12. Wykres zależności precyzji względem liczby epok.

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.

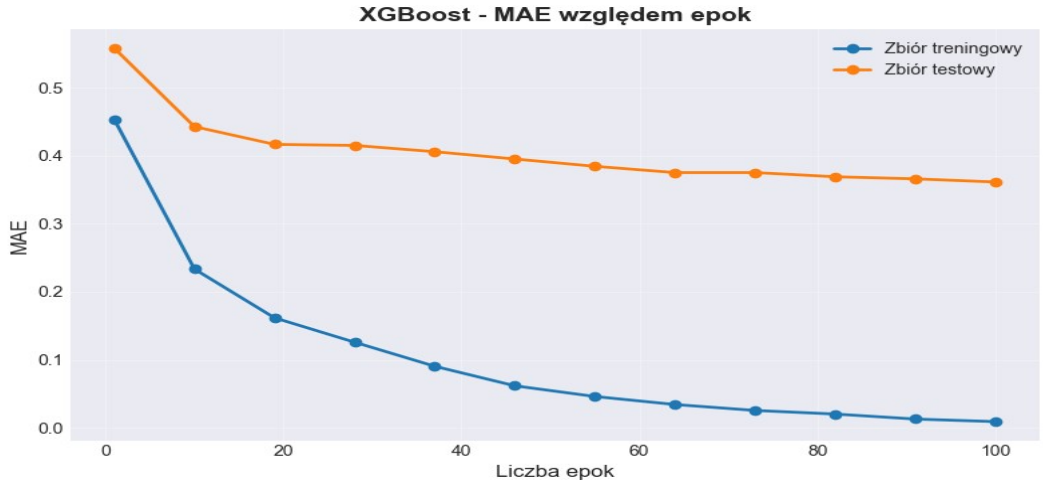
W przeciwieństwie do modelu Random Forest, krzywa czułości dla XGBoost wykazuje trend rosnący przez niemal cały okres treningu, nie osiągając tak szybkiego nasycenia. Czułość na zbiorze testowym wzrasta z poziomu 20% do ponad 41% przy setnej epoce. Oznacza to, że boosting (iteracyjne uczenie na błędach) 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.

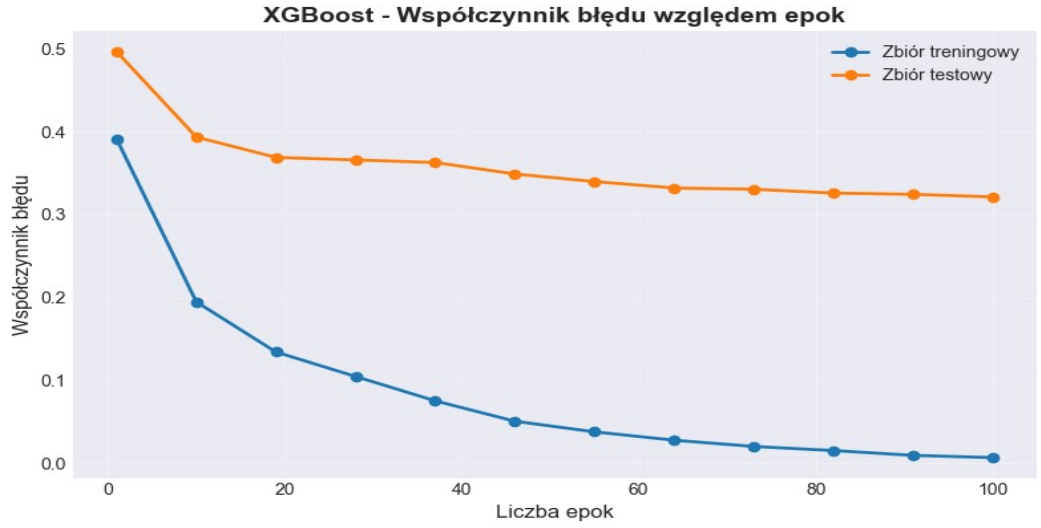
Wykres MSE potwierdza efektywność procesu minimalizacji funkcji straty. Błąd na zbiorze testowym spada monotonicznie do około 80. epoki, osiągając minimum na poziomie ~0.45. Warto zauważyć, że po przekroczeniu tego punktu krzywa testowa ulega delikatnemu wypłaszczeniu,

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.

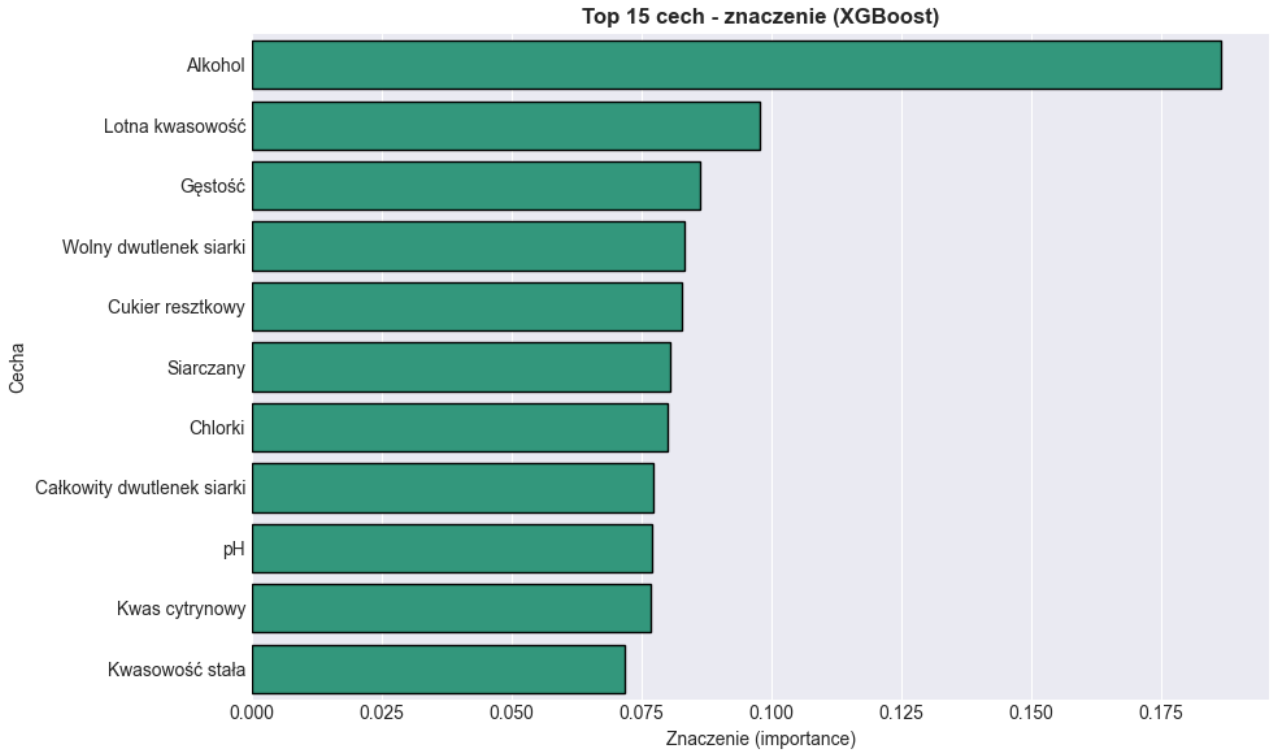
Przebieg MAE jest zbliżony do liniowego spadku w pierwszej połowie treningu. Ostateczna wartość błędu na poziomie ~0.36 oznacza, że model myli się średnio o około 1/3 klasy jakości. Stabilizacja tego parametru w późniejszych epokach (80-100) sugeruje, że model osiągnął limit możliwości predykcyjnych 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.



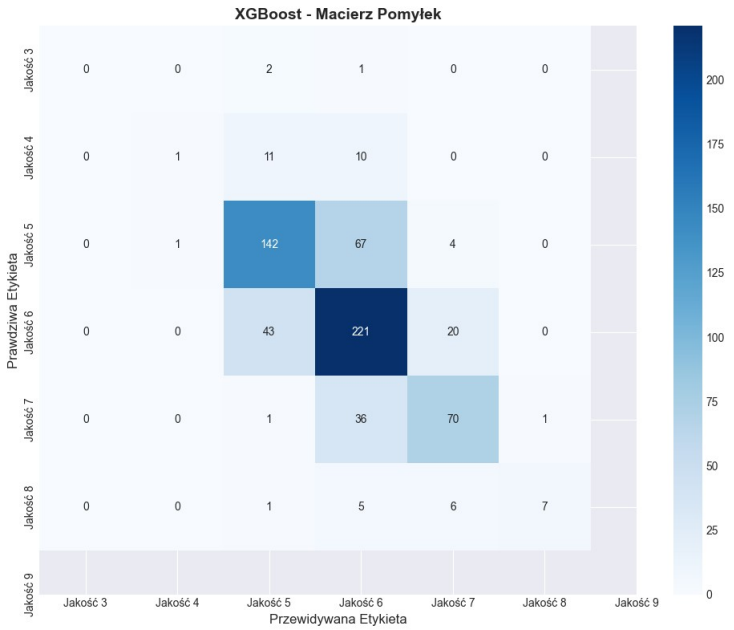
Rysunek 5.17. Piętnaście najważniejszych cech w zbiorze „Wine Quality" uzyskanych w procesie uczenia za pomocą algorytmu XGBoost.

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

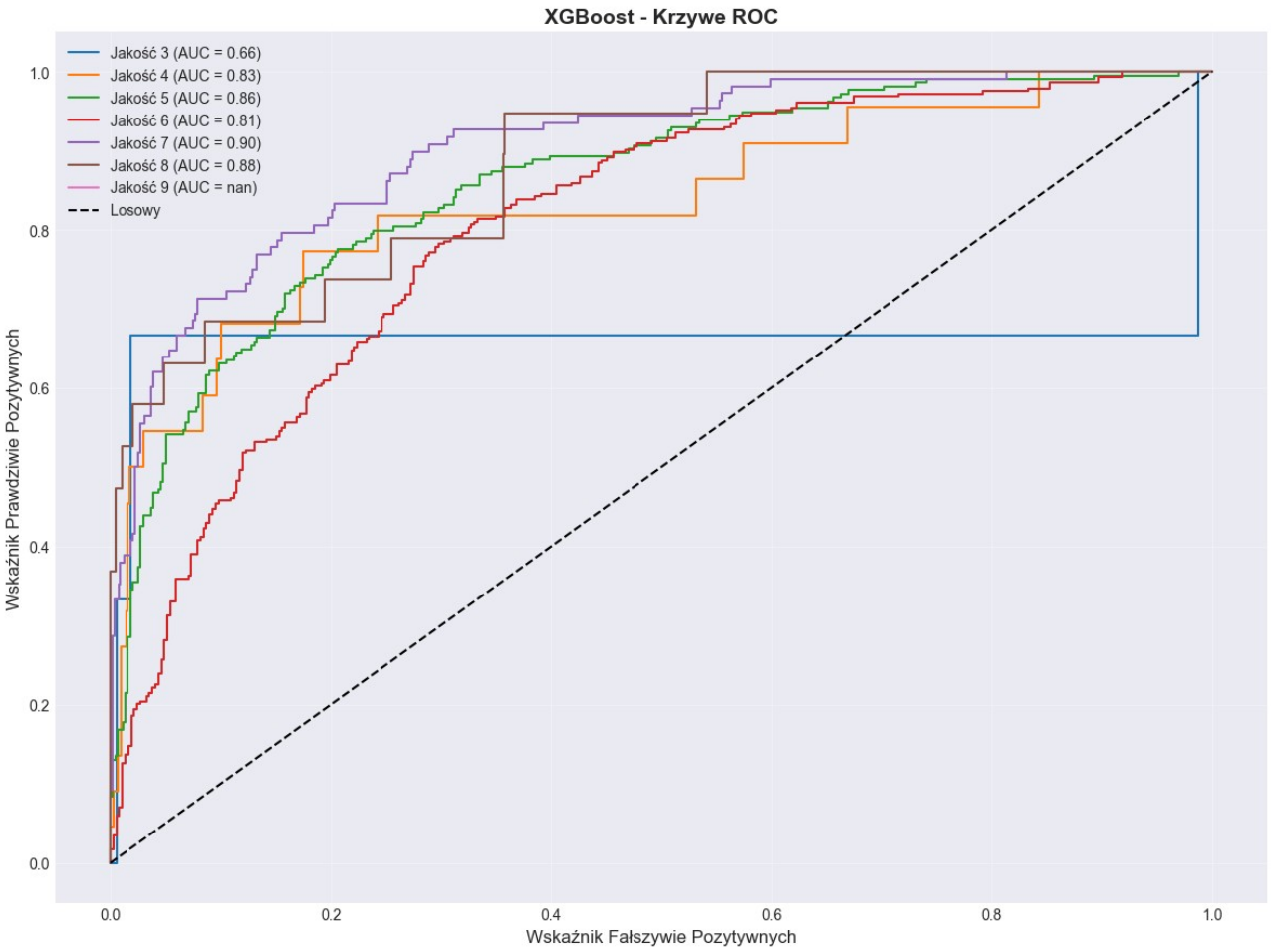
Wyniki na zbiorze testowym potwierdzają dobrą generalizację modelu XGBoost – dokładność (67,85%) jest zbliżona do tej uzyskanej na zbiorze treningowym, co wyklucza zjawisko przeuczenia (overfitting). W porównaniu do modelu Random Forest, XGBoost osiągnął nieco niższą precyzję (57,25%), ale utrzymał zbliżony 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 (Precyzja/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	Precyzja	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

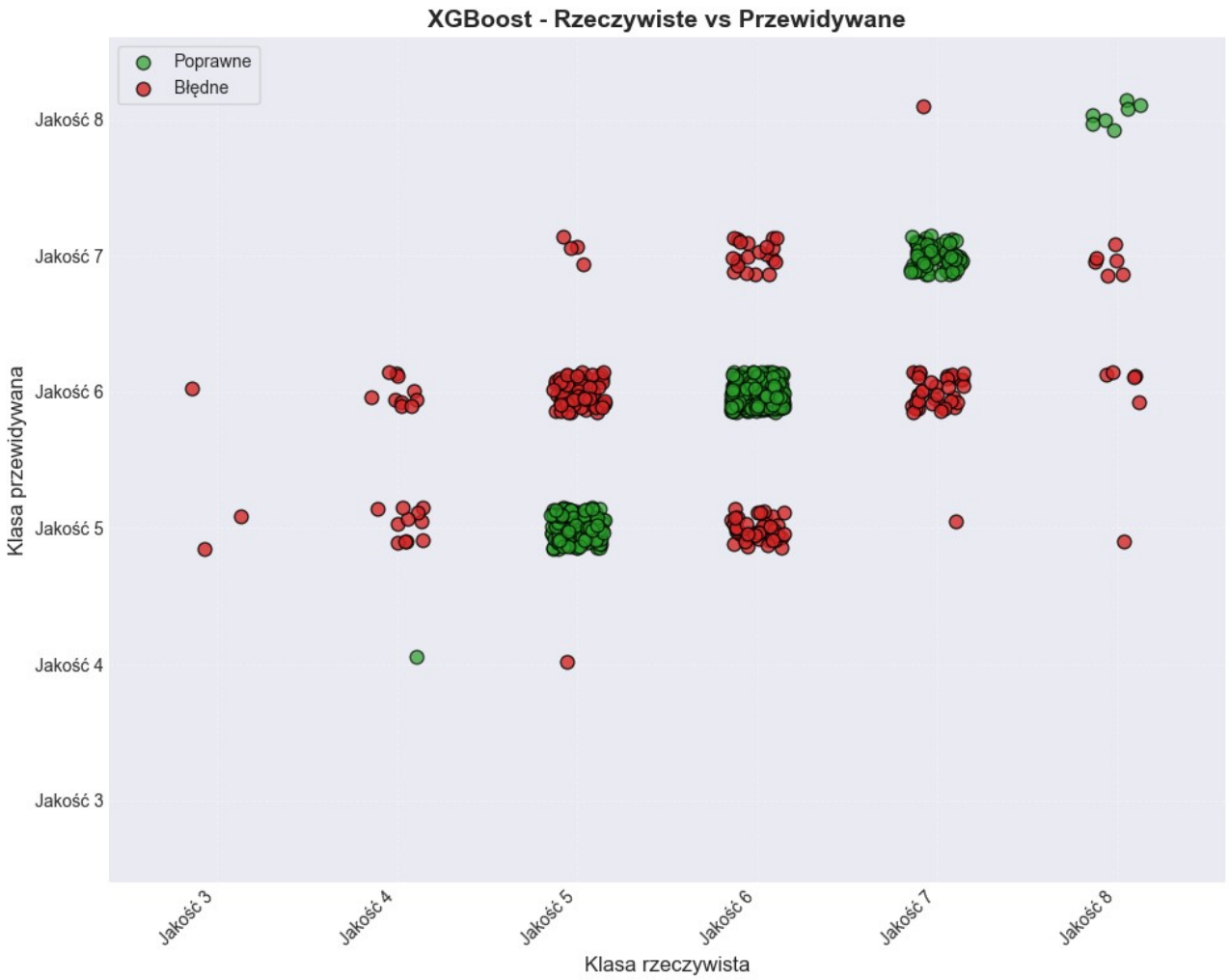
Szczegółowa dekompozycja skuteczności modelu w podziale na klasy ujawnia subtelną przewagę algorytmu XGBoost nad prostszymi metodami, szczególnie w obszarze klas trudnych. W odróżnieniu od wcześniejszych modeli, XGBoost wykazuje zdolność detekcji klasy 4, osiągając precyzję na poziomie 0,50, co mimo 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 precyzja (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 błędnym 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

5.2.2. Algorytm XGBoost

6. Podsumowanie

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.