Wizualizacja i eksploracja danych biznesowych

Programowanie w językach skryptowych

Wiktoria Kozieł 79864

Spis treści

Wprowadzenie do projektu 2
Eksploracyjna analiza danych 3
Podgląd danych 3
Statystyki opisowe 4
Wykrycie braków danych 5
Histogramy 6
Boxploty 7
Heatmapa korelacji 8
Przygotowanie danych do analizy 9
Czyszczenie danych 9
Przekształcenie zmiennych wejściowych (normalizacja i skalowanie) 9
Podział na zbiór treningowy i testowy 9
Modele klasyfikujące jakość 10
Model 1: Random Forest (RF) 10
Model 2: K-Nearest Neighbors (KNN) 11
Ocena jakości 12
Modele regresyjne szacujące jakość 13
Model 1: Regresja liniowa (Linear Regression) 13
Model 2: Gradient Boosting Regressor 13
Zaokrąglenie wyników i ocena 13
Wykresy błędów, rzeczywiste vs przewidywane 14
Porównanie modeli klasyfikujących i regresyjnych 16
Plusy zastosowanych modeli 16
Minusy zastosowanych modeli 16
Wnioski 17
Grupowanie (clustering) 18
KMeans 18
Wizualizacja i profilowanie grup 19
Podsumowanie i wnioski 20

Wprowadzenie do projektu

Celem projektu jest przeprowadzenie kompleksowej analizy zbioru danych zawierającego informacje fizykochemiczne dotyczące białych win portugalskich. Analiza obejmuje zarówno eksplorację danych, jak i budowę modeli predykcyjnych oraz próbę grupowania obserwacji w celu lepszego zrozumienia struktury danych.

Projekt realizowany jest w ramach dwóch przedmiotów: *Programowanie w językach skryptowych* oraz *Wizualizacja i eksploracja danych biznesowych*. W ramach pracy wykorzystywane są narzędzia analityczne języka Python oraz biblioteki umożliwiające analizę statystyczną, wizualizację danych i budowę modeli uczenia maszynowego.

Główne etapy projektu to:

- Eksploracyjna analiza i przygotowanie danych
- Budowa modeli klasyfikujących jakość wina
- Budowa modeli regresyjnych szacujących jakość
- Porównanie skuteczności modeli
- Grupowanie obserwacji na podstawie cech fizykochemicznych

Wyniki analizy mogą znaleźć zastosowanie m.in. w przemyśle winiarskim w celu usprawnienia procesu oceny jakości produktów na podstawie ich właściwości chemicznych, bez konieczności subiektywnej oceny sensorycznej.

Eksploracyjna analiza danych

Podgląd danych

Dane zawierają 4898 obserwacji i 12 zmiennych, z czego 11 opisuje cechy fizykochemiczne wina, a ostatnia (quality) to ocena jakości w skali od 0 do 10. Poniżej zaprezentowano przykładowe dane wejściowe:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	7.0	0.27	0.36	20.7	0.045	
1	6.3	0.30	0.34	1.6	0.049	
2	8.1	0.28	0.40	6.9	0.050	
3	7.2	0.23	0.32	8.5	0.058	
4	7.2	0.23	0.32	8.5	0.058	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	45.0	170.0	1.0010	3.00	0.45	
1	14.0	132.0	0.9940	3.30	0.49	
2	30.0	97.0	0.9951	3.26	0.44	
3	47.0	186.0	0.9956	3.19	0.40	
4	47.0	186.0	0.9956	3.19	0.40	

	alcohol	quality
0	8.8	6.0
1	9.5	6.0
2	10.1	6.0
3	9.9	6.0
4	9.9	6.0

Statystyki opisowe

Poniżej przedstawiono podstawowe statystyki opisowe dla wszystkich zmiennych ilościowych. Widzimy zróżnicowanie m.in. w zawartości alkoholu oraz poziomie siarki w winie, co może mieć wpływ na końcową jakość produktu.

	fixed acidity	volatile acidity	citric acid	residual sugar	\
count	4898.000000	4897.000000	4897.000000	4898.000000	
mean	6.854788	0.278224	0.334170	6.395090	
std	0.843868	0.100798	0.121023	5.075612	
min	3.800000	0.080000	0.000000	0.600000	
25%	6.300000	0.210000	0.270000	1.700000	
50%	6.800000	0.260000	0.320000	5.200000	
75%	7.300000	0.320000	0.390000	9.900000	
max	14.200000	1.100000	1.660000	65.800000	

	chlorides	free sulfur dioxide	total sulfur dioxide	density	\
count	4897.000000	4897.000000	4896.000000	4895.00000	
mean	0.045768	36.696140	138.470282	0.99403	
std	0.021848	99.558576	43.068772	0.00299	
min	0.009000	2.000000	9.000000	0.98711	
25%	0.036000	23.000000	108.000000	0.99173	
50%	0.043000	34.000000	134.000000	0.99375	
75%	0.050000	46.000000	167.000000	0.99610	
max	0.346000	6900.000000	630.000000	1.03898	

	pH	sulphates	alcohol	quality
count	4898.000000	4898.000000	4898.000000	4897.000000
mean	3.188267	0.489847	10.514267	5.877884
std	0.151001	0.114126	1.230621	0.885727
min	2.720000	0.220000	8.000000	3.000000
25%	3.090000	0.410000	9.500000	5.000000
50%	3.180000	0.470000	10.400000	6.000000
75%	3.280000	0.550000	11.400000	6.000000
max	3.820000	1.080000	14.200000	9.000000

Wykrycie braków danych

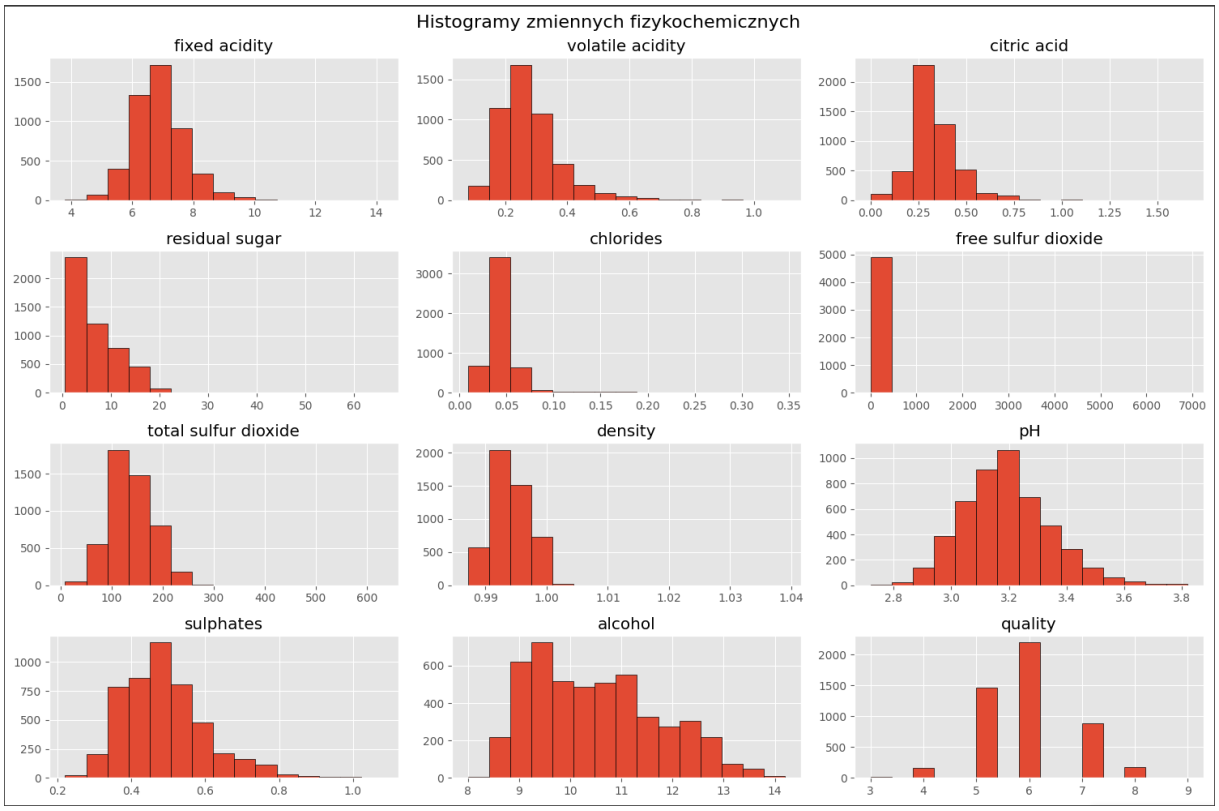
W danych zauważalne są pojedyncze braki – głównie w kolumnach volatile acidity, citric acid, chlorides, free sulfur dioxide, total sulfur dioxide, density, quality. Skala braków jest niewielka (1–3 wartości), dlatego na dalszym etapie zostaną one usunięte.

	0
fixed acidity	0
volatile acidity	1
citric acid	1
residual sugar	0
chlorides	1
free sulfur dioxide	1
total sulfur dioxide	2
density	3
pH	0
sulphates	0
alcohol	0
quality	1

dtype: int64

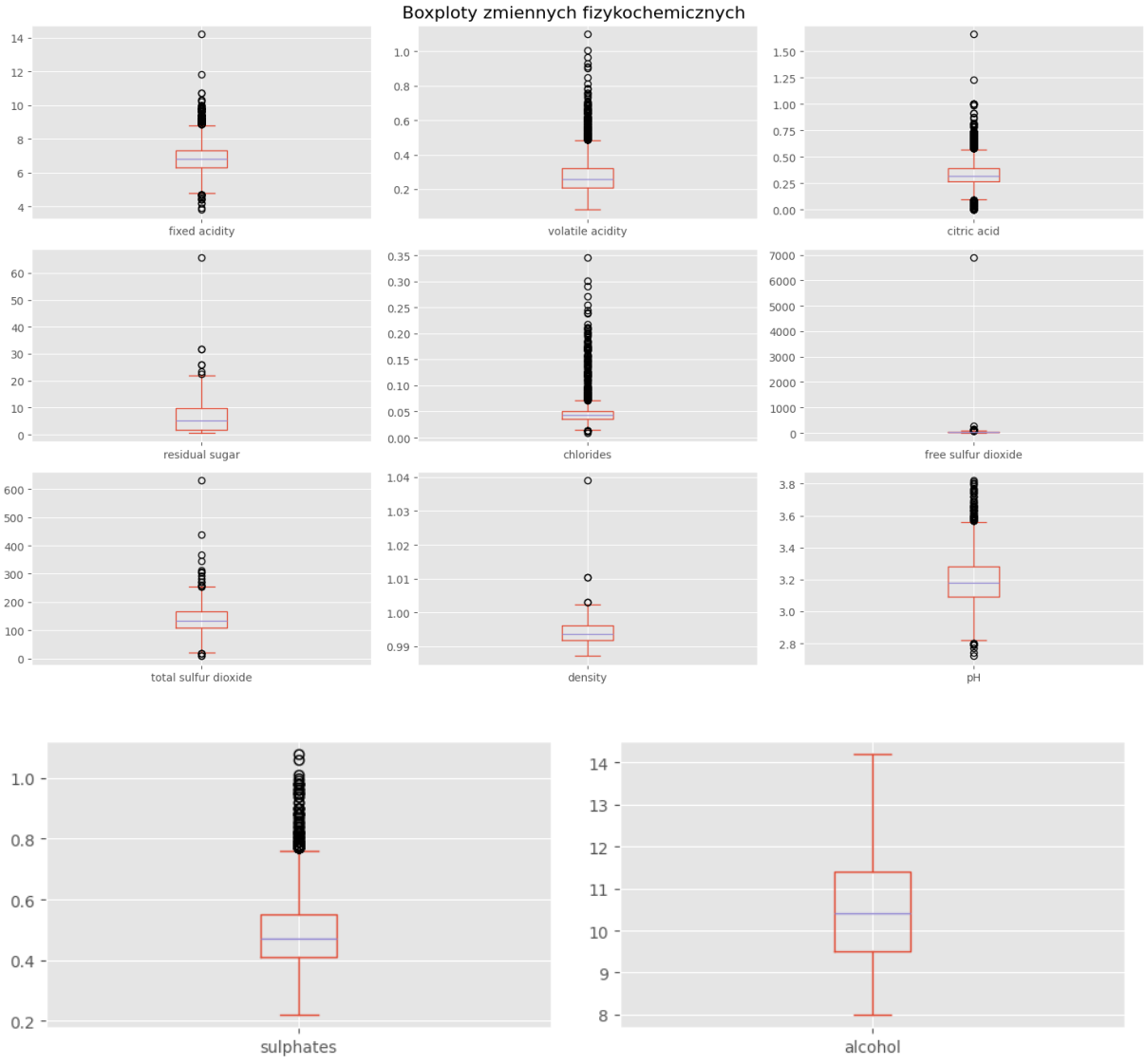
Histogramy

Poniżej przedstawiono histogramy wszystkich zmiennych fizykochemicznych zawartych w zbiorze danych. Widzimy, że wiele cech (np. residual sugar, free sulfur dioxide) ma rozkład skośny, a niektóre zmienne, takie jak alcohol, wykazują rozkład zbliżony do normalnego. Histogramy pomagają w ocenie rozrzutu i identyfikacji potencjalnych wartości odstających.



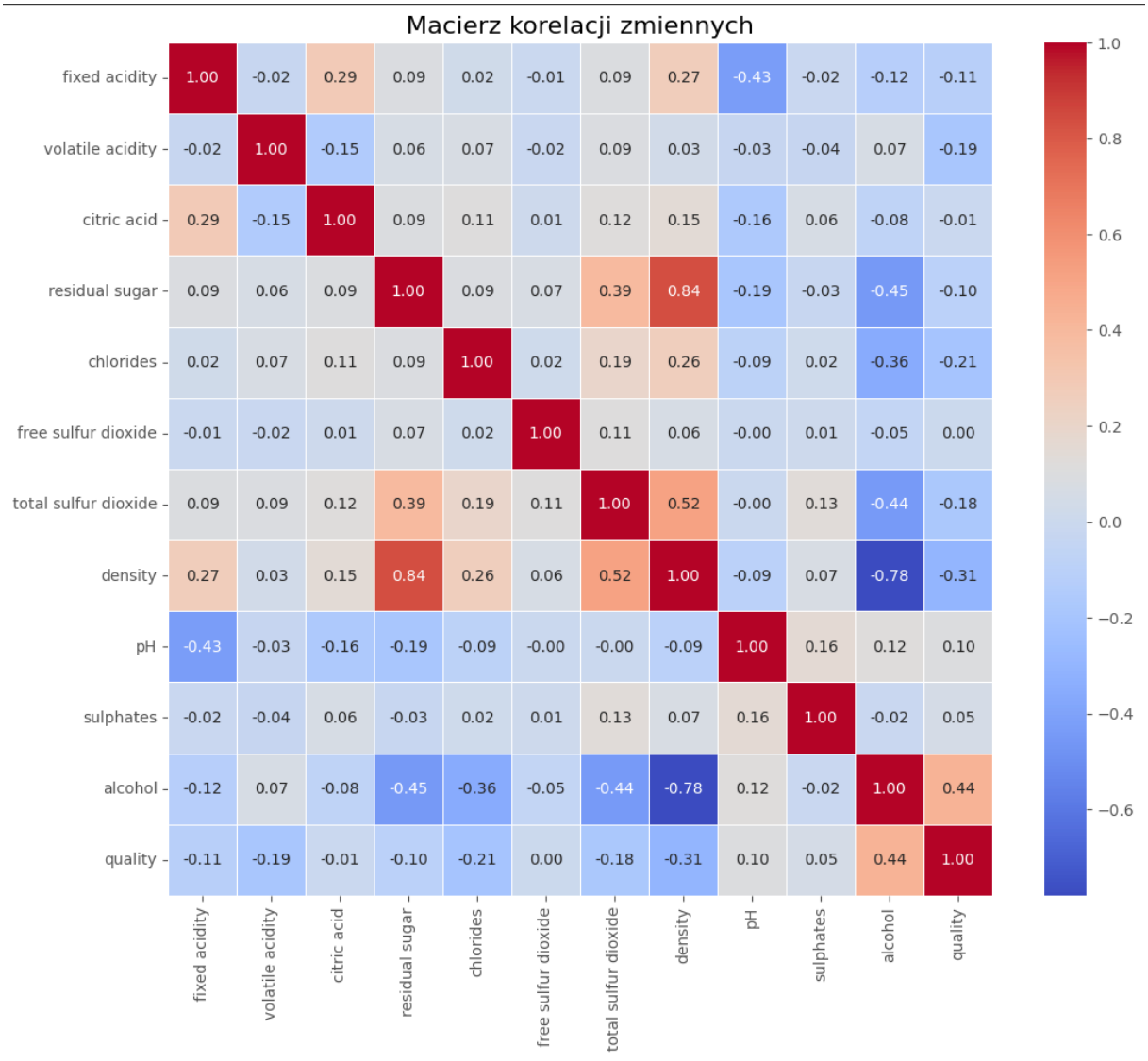
Boxploty

Na poniższych wykresach pudełkowych przedstawiono rozrzut oraz wartości odstające dla zmiennych fizykochemicznych (z wyłączeniem zmiennej celu quality). Widzimy, że cechy takie jak residual sugar, free sulfur dioxide oraz total sulfur dioxide zawierają wiele wartości odstających, co może mieć wpływ na budowę modeli predykcyjnych. Boxploty pomagają zidentyfikować zmienne o dużym rozrzucie oraz ewentualne potrzeby transformacji danych.



Heatmapa korelacji

Poniższa macierz korelacji przedstawia siłę i kierunek powiązań pomiędzy zmiennymi. Zauważalna jest dodatnia korelacja zmiennej alcohol z jakością (quality) oraz ujemna korelacja m.in. density i volatile acidity z jakością. Wysoka korelacja występuje też pomiędzy free sulfur dioxide a total sulfur dioxide, co sugeruje możliwą redundancję tych zmiennych.



Przygotowanie danych do analizy

Czyszczenie danych

Ze względu na niewielką liczbę braków danych, zdecydowano się usunąć wiersze zawierające wartości puste. Po oczyszczeniu danych, końcowy zbiór zawiera 4888 obserwacji.

Przekształcenie zmiennych wejściowych (normalizacja i skalowanie)

Zmiennym wejściowym nadano skalę standardową (średnia 0, odchylenie standardowe 1), aby wyeliminować wpływ różnic w jednostkach miar na działanie wybranych algorytmów uczenia maszynowego.

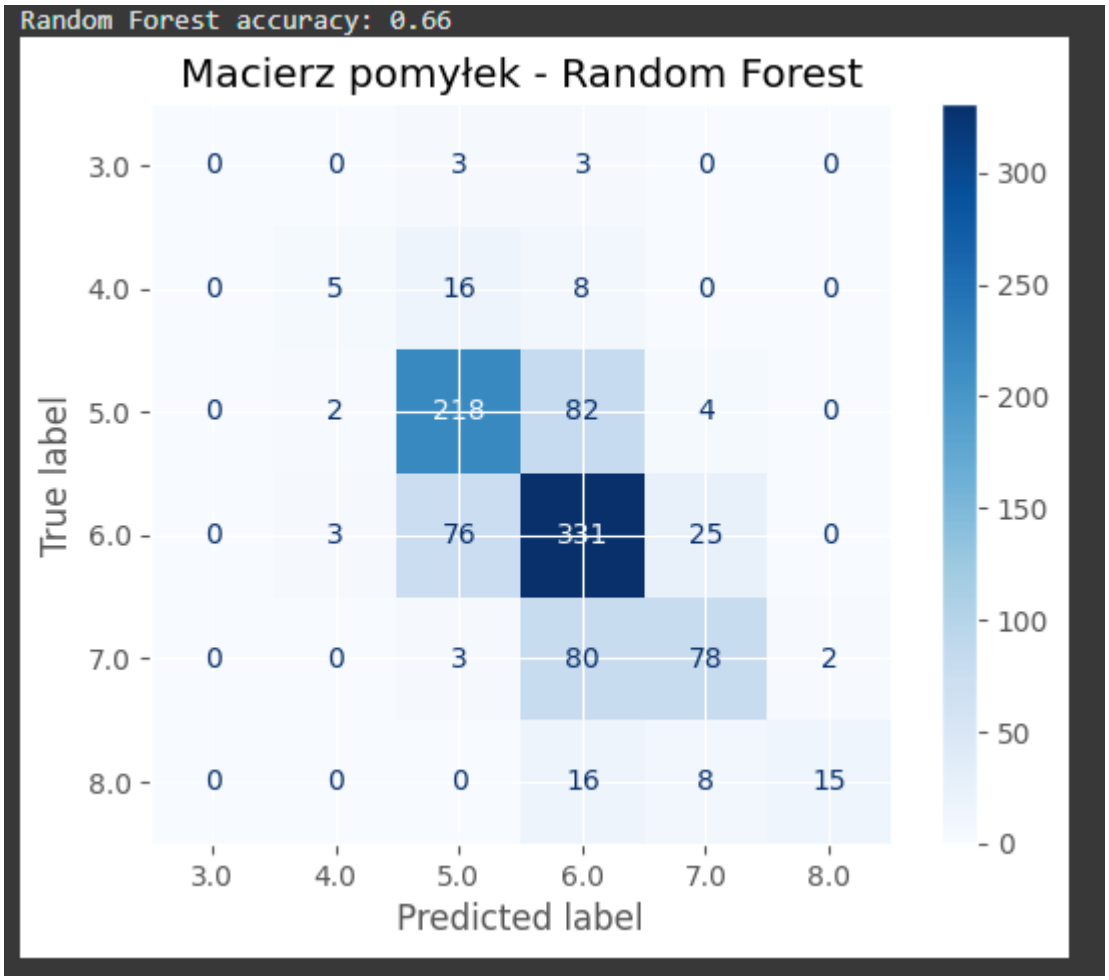
Podział na zbiór treningowy i testowy

Dane zostały podzielone na zbiór treningowy (80%) oraz testowy (20%) w celu oceny skuteczności modeli predykcyjnych na danych nieuczestniczących w procesie uczenia.

Modele klasyfikujące jakość

Model 1: Random Forest (RF)

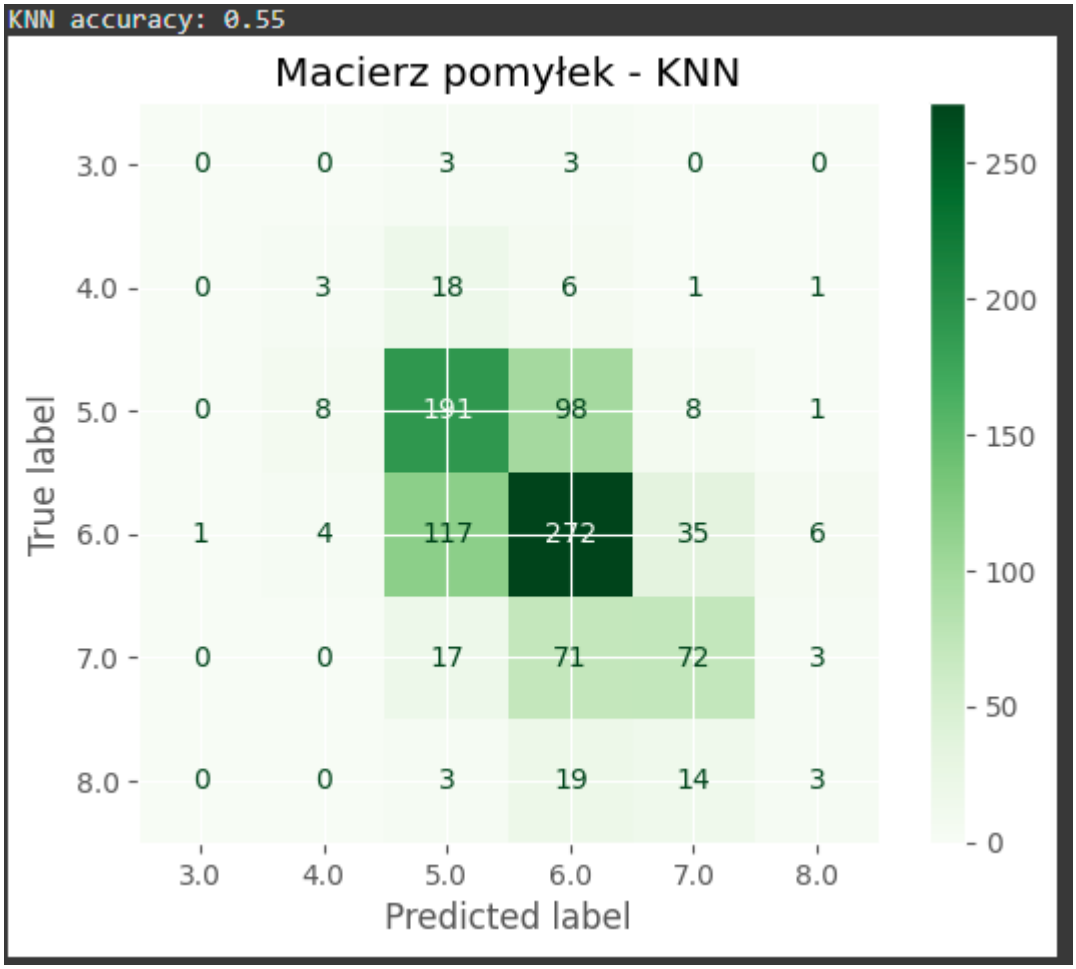
Pierwszym zbudowanym modelem klasyfikującym był Random Forest, czyli las losowy. Jest to algorytm oparty na wielu drzewach decyzyjnych, który sprawdza się dobrze w przypadku danych o różnej skali i nieliniowych zależnościach. Model osiągnął dokładność na poziomie 66%. Na poniższym wykresie przedstawiono macierz pomyłek obrazującą jakość predykcji.



Model 2: K-Nearest Neighbors (KNN)

Drugim wykorzystanym modelem klasyfikującym był K-Nearest Neighbors (KNN), który opiera się na odległości pomiędzy obserwacjami w przestrzeni cech. Dla liczby sąsiadów k=5, model osiągnął dokładność na poziomie 55%

Model dobrze działa z przeskalowanymi danymi, ale może być mniej odporny na dane odstające. Na poniższym wykresie przedstawiono macierz pomyłek.



Ocena jakości

W celu oceny skuteczności działania modeli klasyfikujących jakość wina, obliczono dokładność (accuracy) predykcji oraz przedstawiono macierze pomyłek dla każdego z algorytmów.

Random Forest osiągnął dokładność na poziomie 66%, co świadczy o dobrej ogólnej jakości predykcji. Model ten wykazuje odporność na wartości odstające i dobrze radzi sobie z nieliniowymi zależnościami.

K-Nearest Neighbors osiągnął dokładność na poziomie 55%. Wynik jest nieco niższy, co może być spowodowane większą wrażliwością na rozkład danych oraz obecnością klas o małej liczbie próbek.

Na podstawie macierzy pomyłek zauważono, że oba modele mają tendencję do mylenia klas sąsiadujących (np. jakość 5 i 6), co jest naturalne przy zadaniu wieloklasowej klasyfikacji na zbiorze o ciągłej naturze zmiennej celu.

Modele regresyjne szacujące jakość

Model 1: Regresja liniowa (Linear Regression)

Pierwszym modelem regresyjnym była regresja liniowa. Celem było przewidzenie wartości zmiennej quality traktowanej jako ciągła.

Po zaokrągleniu wyników do najbliższej liczby całkowitej, obliczono dokładność klasyfikacji, która wyniosła 51%.

Model osiągnął również wartość błędu średniokwadratowego (MSE) na poziomie 1.18, średniego błędu bezwzględnego (MAE) na poziomie 0.61, a współczynnik determinacji R^2 wyniósł -0.49.

Model 2: Gradient Boosting Regressor

Drugim modelem regresyjnym był Gradient Boosting Regressor — algorytm bazujący na zespołowym uczeniu słabych modeli (tzw. boosting).

Po zaokrągleniu przewidywanych wartości do najbliższej liczby całkowitej, model osiągnął dokładność 56%.

Wyniki oceny błędów: MSE = 0.49, MAE = 0.55, R^2 = 0.38. Model ten poradził sobie zauważalnie lepiej niż regresja liniowa, co sugeruje, że zależności między zmiennymi są nieliniowe.

Zaokrąglenie wyników i ocena

Ponieważ zmienna quality w oryginalnym zbiorze danych ma postać liczby całkowitej (0–10), przewidywane przez modele regresyjne wartości zostały zaokrąglone do najbliższej liczby całkowitej. Umożliwiło to porównanie ich skuteczności z modelami klasyfikującymi.

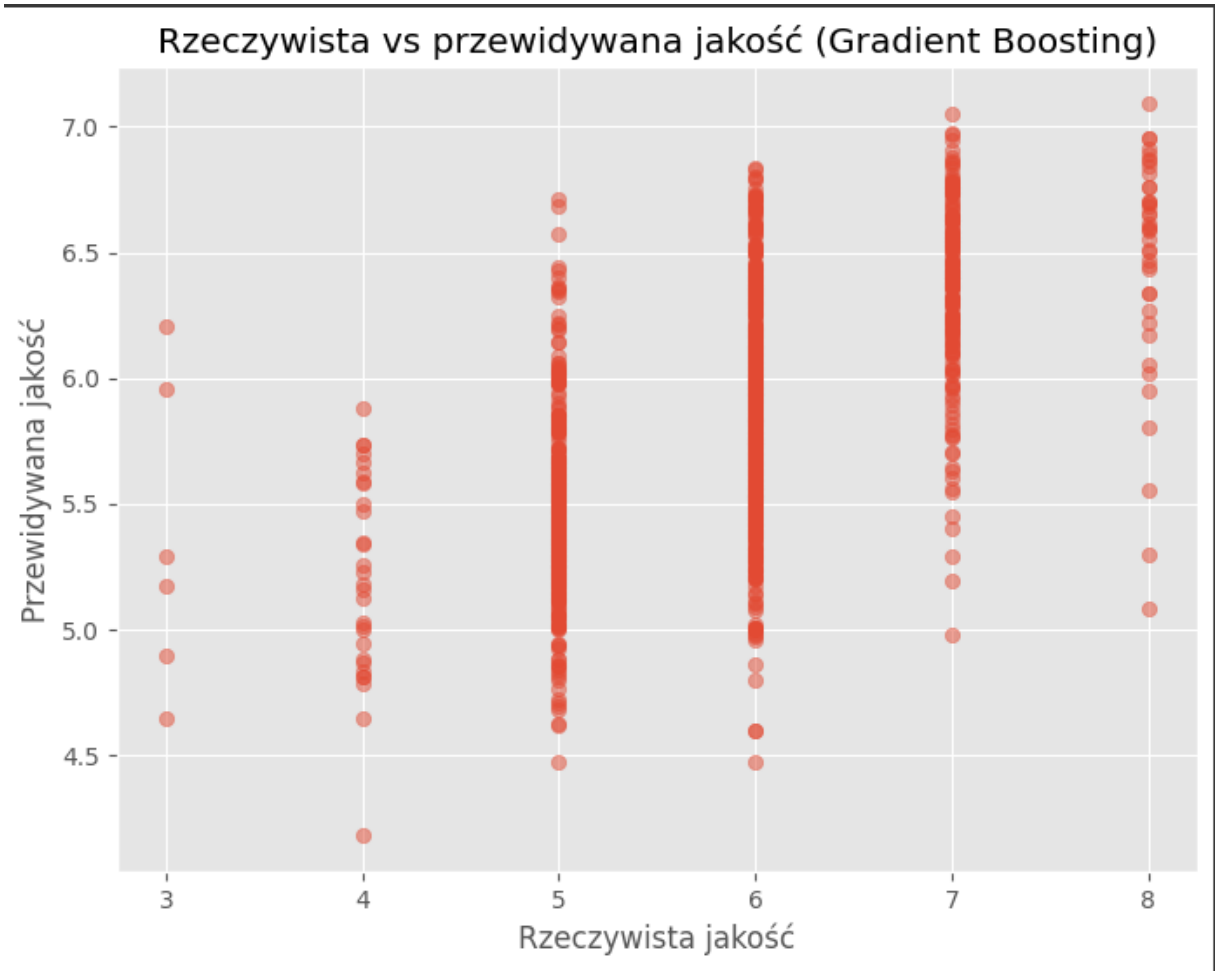
Model	Accuracy	MSE	MAE	R^2
Linear Regression	51%	1.18	0.61	-0.49
Gradient Boosting Regressor	56%	0.49	0.55	0.38

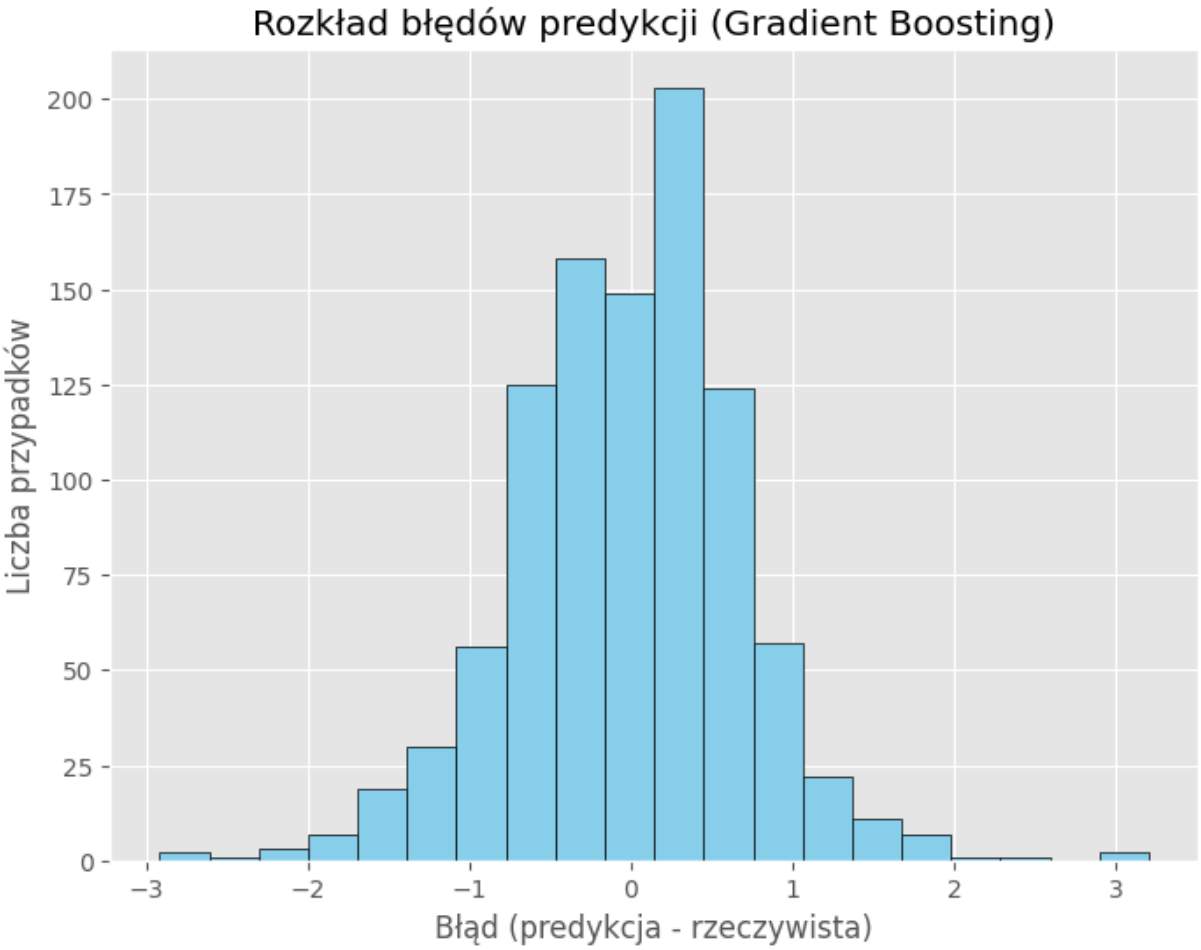
Gradient Boosting poradził sobie zdecydowanie lepiej niż regresja liniowa, co potwierdza obecność nieliniowych zależności w danych.

Wykresy błędów, rzeczywiste vs przewidywane

Poniższy wykres przedstawia porównanie wartości przewidywanych przez model Gradient Boosting z rzeczywistymi ocenami jakości wina. Idealny model generowałby punkty blisko przekątnej wykresu.

W kolejnym wykresie zaprezentowano histogram błędów predykcji. Widzimy, że większość błędów oscyluje wokół zera, co oznacza poprawność działania modelu, jednak występują też pojedyncze odchylenia o wartości ±2.




Rozkład błędów predykcji (Gradient Boosting)
Liczba przypadków
Błąd (predykcja - rzeczywista)
0
25
50
75
100
125
150
175
200
−3
−2
−1
0
1
2
3

Porównanie modeli klasyfikujących i regresyjnych

Plusy zastosowanych modeli

Random Forest – bardzo dobra dokładność, odporny na szum, nie wymaga skalowania danych, dobrze działa przy wielu zmiennych.

Gradient Boosting – skuteczny w regresji, dobrze radzi sobie z nieliniowościami, daje dokładne przewidywania po odpowiednim zaokrągleniu.

KNN – prosty, intuicyjny algorytm, dobre wyniki przy dobrze przeskalowanych danych.

Regresja liniowa – łatwa do interpretacji, szybka, przydatna jako punkt odniesienia.

Minusy zastosowanych modeli

Random Forest – mniej interpretowalny (czarna skrzynka), może przeuczyć się przy małych zbiorach.

Gradient Boosting – wolniejszy trening, wymaga strojenia hiperparametrów dla najlepszych wyników.

KNN – bardzo wrażliwy na skalę i wartości odstające, wolniejszy przy większych zbiorach.

Regresja liniowa – niska skuteczność, zakłada liniowe zależności, które w tym zbiorze nie występują.

Wnioski

W niniejszym rozdziale zestawiono wyniki uzyskane przez modele klasyfikujące i regresyjne w przewidywaniu jakości wina.

Modele klasyfikujące (Random Forest, KNN) od razu operowały na klasach (liczbach całkowitych), podczas gdy modele regresyjne (Linear Regression, Gradient Boosting) wymagały zaokrąglenia wyników do postaci klasy.

Model	Accuracy(%)
Random Forest (klasyfikacja)	66%
KNN (klasyfikacja)	55%
Linear Regression (regresja)	51%
Gradient Boosting (regresja)	56%

Na podstawie wyników można zauważyć, że Random Forest był najskuteczniejszym modelem pod względem dokładności klasyfikacji.

Wśród modeli regresyjnych najlepszy rezultat uzyskał Gradient Boosting, co potwierdza jego zdolność do uchwycenia złożonych, nieliniowych relacji pomiędzy zmiennymi.

Analiza przeprowadzona na zbiorze danych dotyczącym jakości białych win portugalskich pozwoliła zidentyfikować cechy mające największy wpływ na ocenę końcową produktu. W szczególności zaobserwowano, że zawartość alkoholu, poziom kwasowości oraz gęstość wina korelują z jego jakością.

Eksploracyjna analiza danych wykazała obecność wartości odstających oraz nienormalnych rozkładów niektórych zmiennych, co mogło wpłynąć na skuteczność niektórych modeli predykcyjnych.

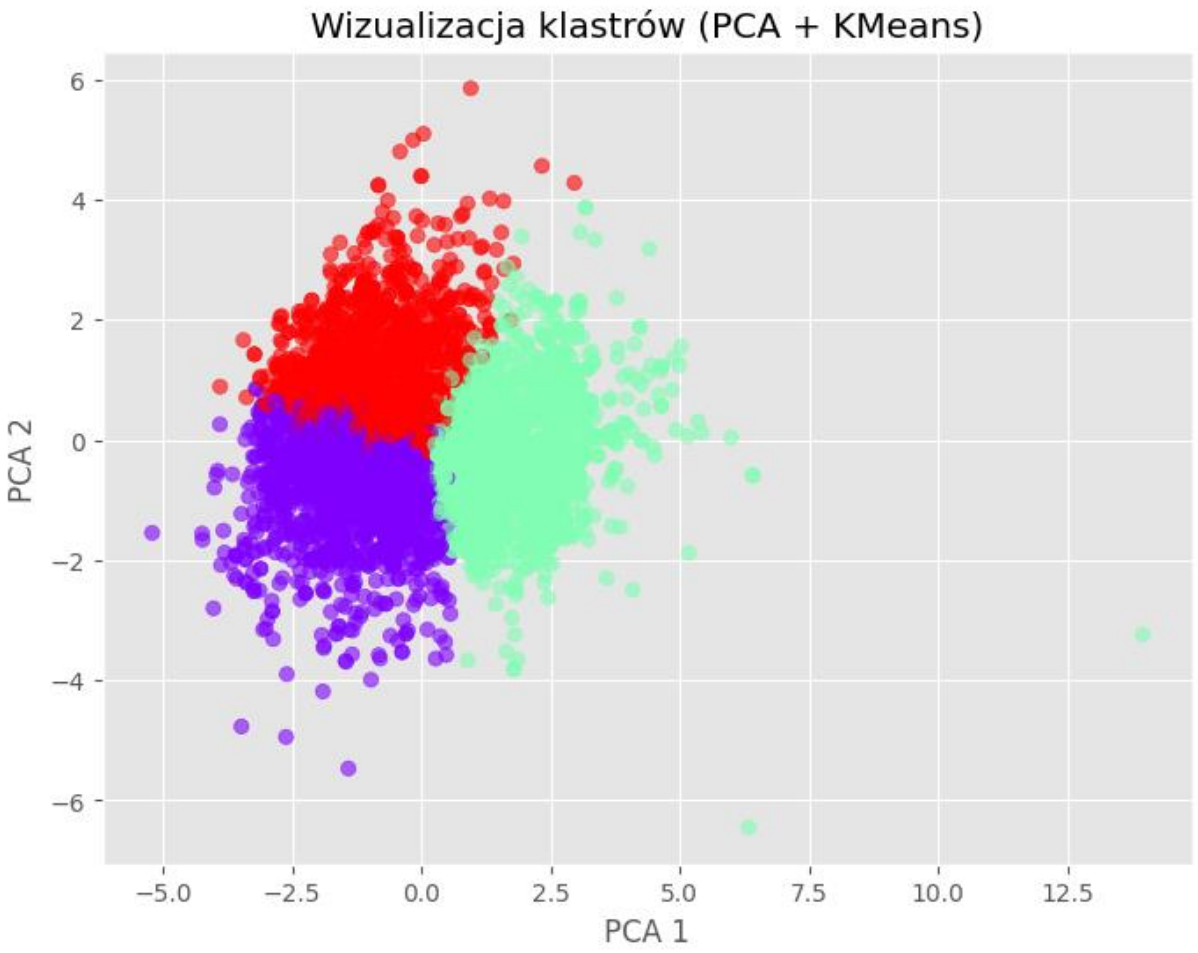
W ramach części predykcyjnej projektu zaobserwowano, że modele klasyfikujące jakość wina wykazały wyższą skuteczność niż modele regresyjne. Najlepsze rezultaty osiągnął algorytm Random Forest, który uzyskał dokładność klasyfikacji na poziomie 66%. Z kolei spośród modeli regresyjnych najwyższą skuteczność osiągnął Gradient Boosting, uzyskując dokładność 56% po zaokrągleniu wyników. Mimo nieco niższych wyników regresji, modele te pozwalają lepiej zrozumieć zmienność i przewidywaną jakość próbek, co może być istotne w zastosowaniach biznesowych.

Grupowanie (clustering)

KMeans

Grupowanie danych przeprowadzono za pomocą algorytmu KMeans z liczbą klastrów ustawioną na 3. Zastosowano wcześniej przekształcone dane fizykochemiczne (bez zmiennej quality). Wizualizację wykonano przy użyciu metody PCA do redukcji wymiarowości.

Na poniższym wykresie przedstawiono dane pogrupowane według przypisanych klastrów. Kolory reprezentują różne grupy, które mogą wskazywać na różne typy win.



Wizualizacja i profilowanie grup

Na podstawie centroidów klastra obliczono średnie wartości cech fizykochemicznych w każdej z trzech grup. Na tej podstawie można wnioskować, że jedna z grup charakteryzuje się wyższą zawartością alkoholu i siarczynów, podczas gdy inna wykazuje większą kwasowość.

Wizualizacja pozwala zauważyć naturalne skupiska danych, które mogą odpowiadać różnym profilom win, np. bardziej wytrawnych lub słodszych.

Cluster	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
0	6.227153	0.276390	0.286659	3.395510	0.040280
1	6.970909	0.285590	0.363091	11.093967	0.055613
2	7.428273	0.270585	0.351957	3.874617	0.039474

free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
31.646915	122.845449	0.992182	3.306542	0.522291	11.067202	6.114844
48.223691	170.646556	0.997034	3.153862	0.495129	9.476253	5.582920
27.857939	115.590529	0.992334	3.096650	0.446497	11.194896	5.977716

Podsumowanie i wnioski

Celem projektu było przeprowadzenie eksploracyjnej analizy danych dotyczących właściwości fizykochemicznych białych win portugalskich oraz przewidywanie ich jakości z wykorzystaniem różnych metod uczenia maszynowego. Przeanalizowano zbiór zawierający 4898 obserwacji, który po oczyszczeniu z brakujących danych zawierał 4888 rekordów.

Na podstawie przeprowadzonej analizy stwierdzono, że zawartość alkoholu, poziom kwasowości oraz gęstość wina mają największy wpływ na ocenę jakości. Dane cechowały się rozrzutem i obecnością wartości odstających, co uzasadniało potrzebę zastosowania bardziej zaawansowanych modeli predykcyjnych.

W części klasyfikacyjnej najlepszy rezultat osiągnął model Random Forest, uzyskując dokładność 66%, natomiast w regresji najlepszy był Gradient Boosting z dokładnością 56% po zaokrągleniu wyników. Porównanie modeli wykazało, że klasyfikatory lepiej radzą sobie z przewidywaniem klas jakości, podczas gdy modele regresyjne umożliwiają bardziej szczegółowe szacowanie wartości.

W końcowym etapie projektu przeprowadzono grupowanie danych przy użyciu algorytmu KMeans, co pozwoliło wyróżnić trzy wyraźne klastry o różnych profilach cech chemicznych. Profilowanie tych grup może mieć praktyczne zastosowanie np. w segmentacji win pod kątem preferencji konsumentów lub optymalizacji procesów produkcyjnych.

Uzyskane wyniki potwierdzają zasadność wykorzystania analizy danych i uczenia maszynowego w ocenie jakości produktów spożywczych. Projekt dostarczył także cennych doświadczeń w zakresie pracy z danymi rzeczywistymi, modelowania predykcyjnego oraz wizualizacji i interpretacji wyników.