Sprawozdanie z laboratorium:
Uczenie Maszynowe

Case study

27 czerwca 2019

Prowadzący: dr hab. inż. Maciej Komosiński

Autorzy: **Jakub Tomczak** inf127083 ISWD jakub.pa.tomczak@student.put.poznan.pl

Zajęcia środowe, 15:10.

Oświadczam, że niniejsze sprawozdanie zostało przygotowane wyłącznie przez powyższego autora, a wszystkie elementy pochodzące z innych źródeł zostały odpowiednio zaznaczone i są cytowane w bibliografii.

1 Zbiór danych

Do case study został wybrany zbiór *Wine Quality* [1]. Zawiera on 4898 próbek wina białego i 1599 czerwonego typu *Vinho Verde* pochodzącego z północnej Portugali. Każda próbka jest badaniem fizykochemicznym, które zawiera 12 własności:

- fixed acidity
- voltaile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

Każdy z atrybutów jest atrybutem liczbowym, co więcej, są to wartości ciągłe. Jedyną operacją jaka została przeprowadzona na zbiorze danych jest standaryzacja potrzebna w przypadku niektórych klasyfikatorów. Zmienną wyjściową jest ocena wina, podana jako liczba naturalna w zakresie 0-10. Do case study zostały wykorzystane próbki wina białego.

2 Eksploracyjna analiza danych

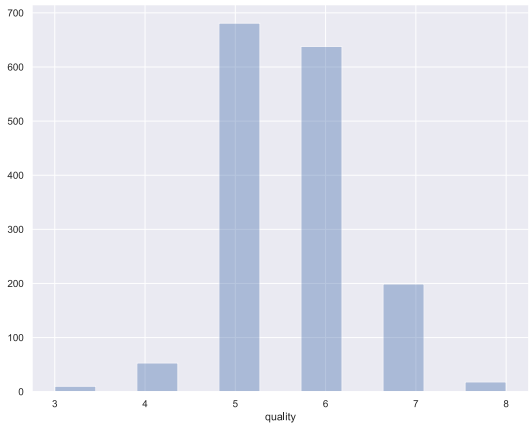
Do analizy został wybrany zbiór z winami białymi ze względu na większą liczbę próbek w zbiorze. Jeżeli chodzi o atrybut decyzyjny *quality* to w obu przypadkach jest on mocno niezbalansowany co widać na obrazku 1. W obu zbiorach nie występują oceny 0, 1, 2 oraz 10; dodatkowo w zbiorze win czerwonych nie ma żadnego przykładu dla oceny 9.

2.1 Macierz rozproszenia

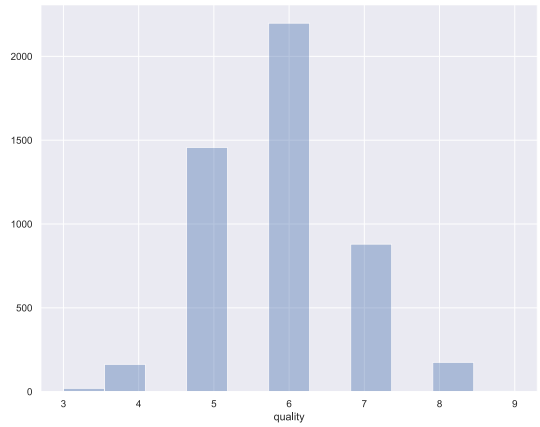
W celu wykrycia zależności między atrybutami wykonana została macierz rozproszenia. Wykresy na rysunku 2 pokazują pewne zależności, np. między atrybutem *fixed acidity* a *pH* - ujemna korelacja, czy między *density*, a *fixed acidity* - dodatnia korelacja, co potwierdzają wyniki w macierzy korelacji na rysunku 3. Zauważalna jest także dodatnia korelacja między *residual sugar*, a *density*.

2.2 Principal component analysis

Po rozkładzie na główne składowe wzięto wektor 5 głownych składowych o największym wkładzie, które wyjaśniały ponad 70% wariancji (73.02%) na które składały się atrybuty *fixed acidity*, *citric acid*, *total sulfur dioxide*, *alcohol*, *sulphates* i za ich pomocą przetransformowano zbiór treningowy oraz testowy (po standaryzacji). Wykres wariancji wyjaśnionych oraz łącznej wariancji wyjaśnionej za pomocą tych składowych jest przedstawiony na obrazku 4. Zbiór danych po transformacji za pomocą *PCA* został wykorzystany do trenowania klasyfikatorów jednakże wyniki były bardzo zbliżone do tych bez transformacji do nowej przestrzeni, tak więc nie są one ujęte w sprawozdaniu.



(a) Histogram dla win czerwonych.



(b) Histogram dla win czerwonych.

Rysunek 1: Porównanie histogramów atrybutu decyzyjnego *quality*

3 Podział danych

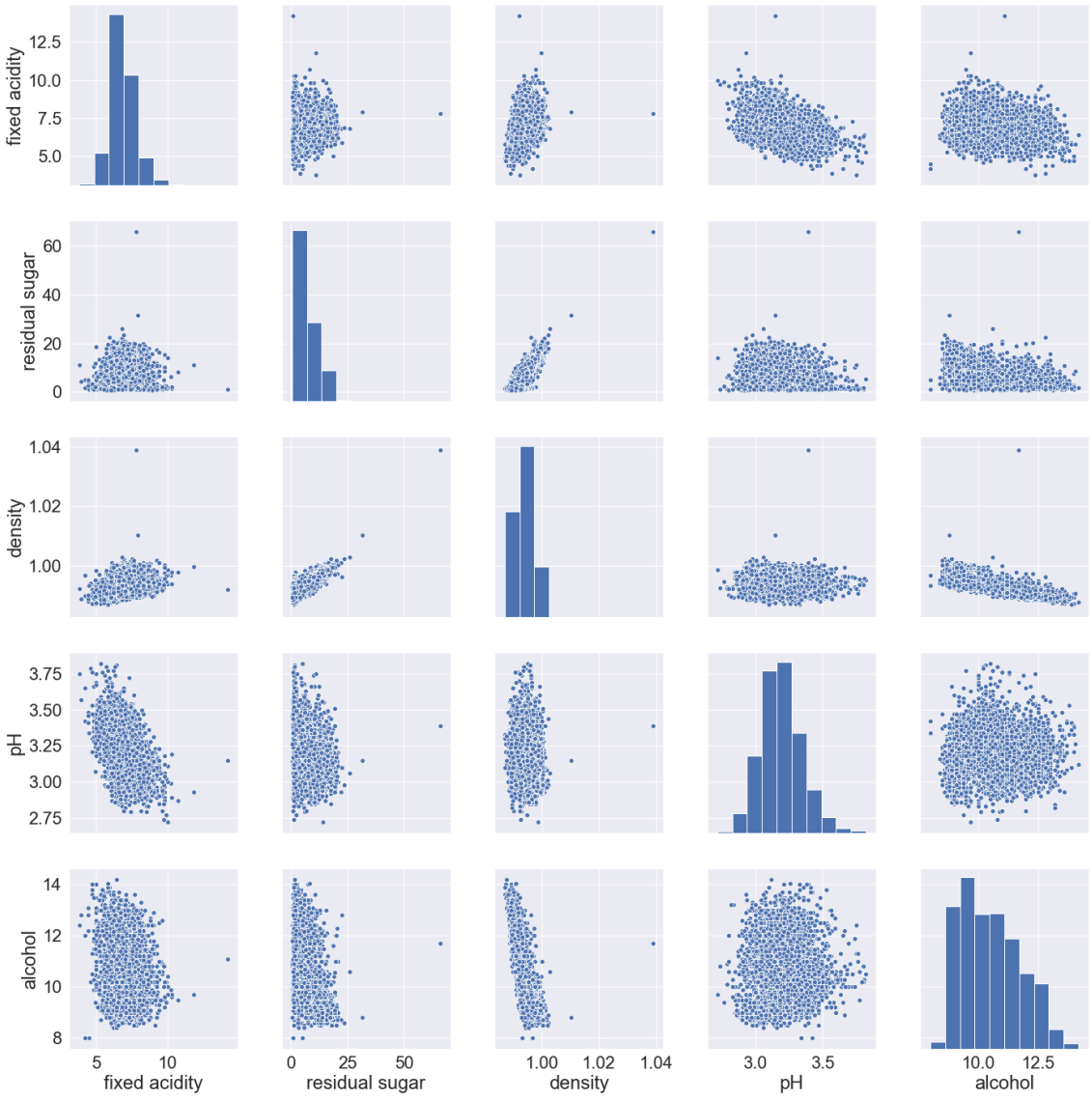
Dane przed podziałem na zbiór treningowy oraz testowy zostały przeskalowane przy użyciu klasy `StandardScaler`, z drugiej strony podczas korzystania z drzew decyzyjnych, czy losowego lasu wykorzystano dane nieustandaryzowane. Jak już zostało wcześniej wspomniane, dane są niezbalansowane, stąd pomysł na podzielenie danych na 3 nowe etykiety - *low*, *medium* oraz *high*, które odpowiadają jakości wina. Podział na sztuczne klasy został przeprowadzony dwukrtonie. Jak widać, na rysunku 1 zbiór win białych zawiera jedynie klasy *3, 4, 5, 6, 7, 8, 9*, stąd pierwszy podział taki, że klasy *3, 4* należą do klasy *low*, klasy *5, 6, 7* są mapowane na *medium*, natomiast klasy *8,9* są mapowane na klasę *high*. Taki sposób mapowania wydaje się dosyć naturalny, chociaż jak widać na rysunku 5a podział ten jest dalej niezbalansowany. Z tego powodu został wykonany drugi podział wg. schematu: *3, 4, 5* do *low*, klasa *6* do *medium*, natomiast klasy *7,8,9* do *high*. Jak widać na rysunku 5b taki podział jest zbalansowany.Owe dodatkowe podziały zostały wykonane w celu sprawdzenia różnic w działaniu klasyfikatorów w przypadku różnego zbalansowania danych.

	low	medium	high
Podział 1	3.74%	92.59%	3.67%
Podział 2	33.48%	44.88%	21.64%

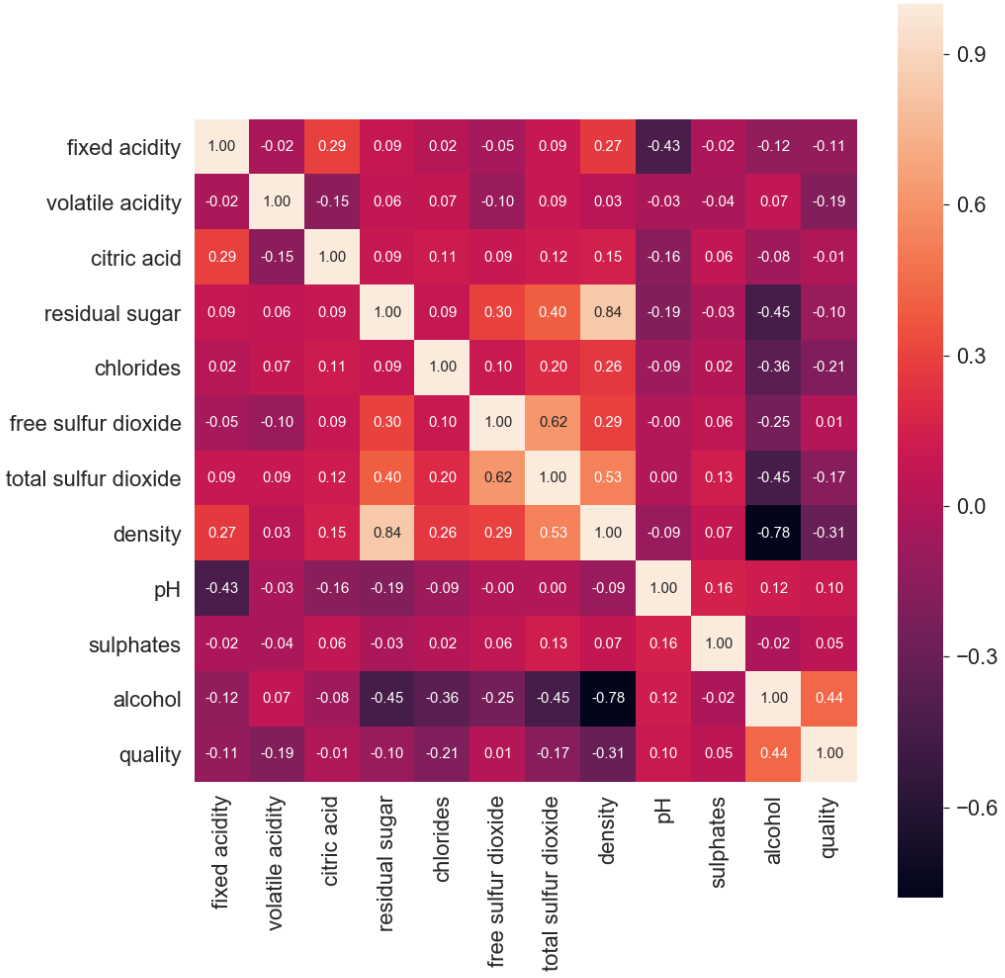
4 Klasyfikacja

4.1 Drzewa decyzyjne

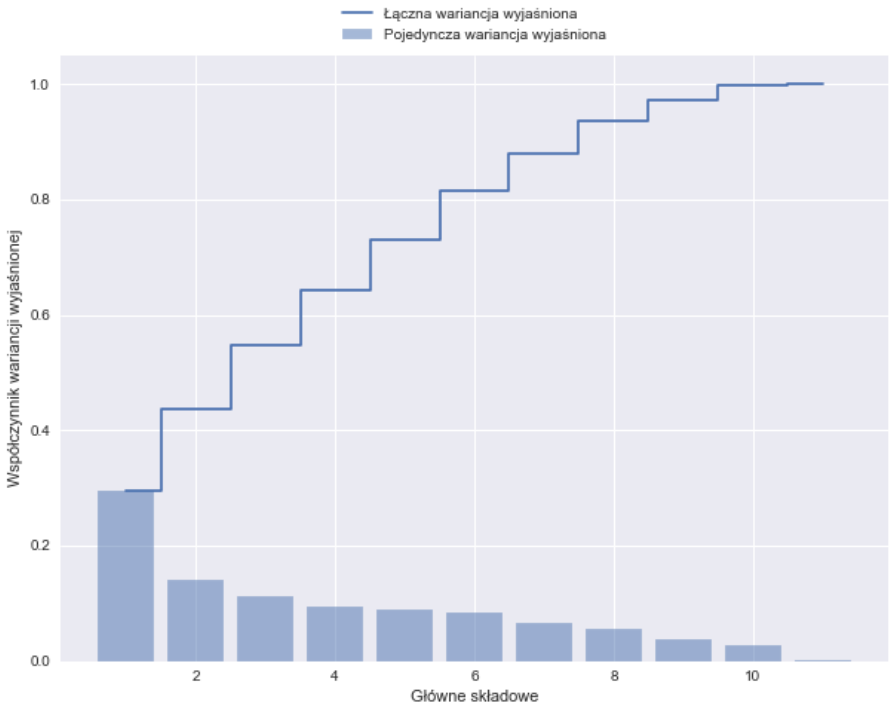
Pierwszy klasyfikatorem jest drzewo decyzyjne dla którego jako kryterium podziału wybrano entropię. Została użyta domyślna implementacja z biblioteki *Sklearn - DecisionTreeClassifier*. Dla pierwotnego podziału na klasy drzewo wykazało trafność na poziomie 45%, najlepszym podziałem okazuje się podział niezbalansowany na 3 klasy, w przypadku którego uzyskano



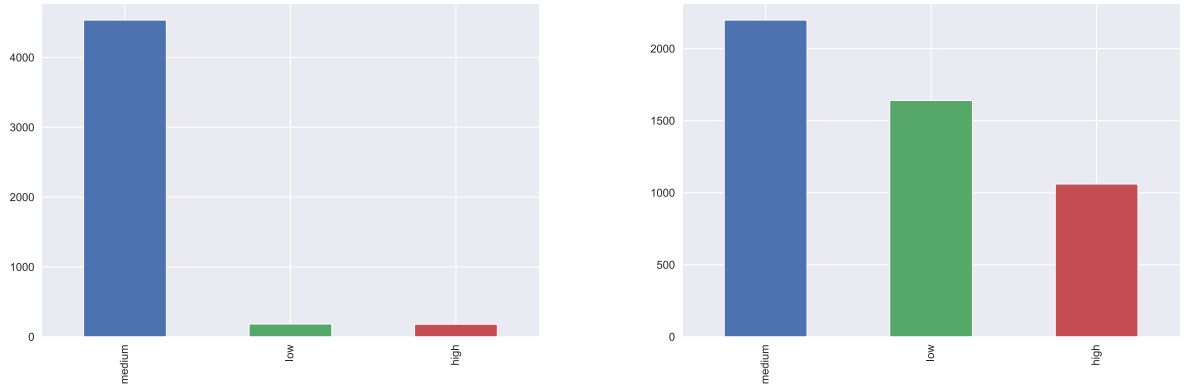
Rysunek 2: Macierze rozproszenia ukazujące związki między poszczególnymi cechami w zbiorze danych.



Rysunek 3: Macierz korelacji między cechami.



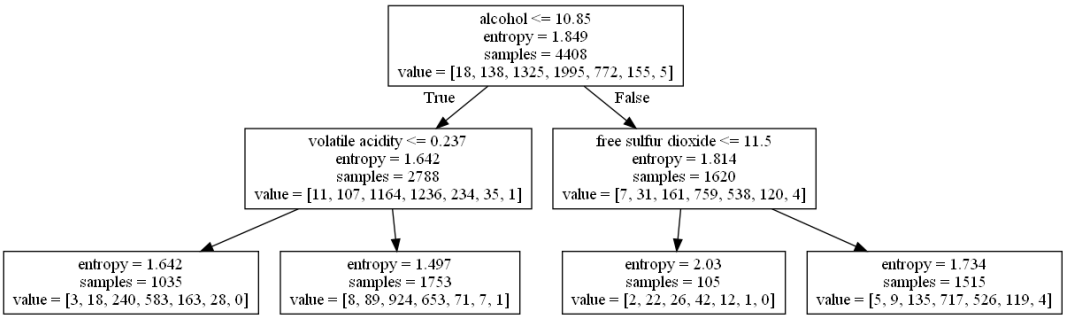
Rysunek 4: Wykres wariancji wyjaśnionej.



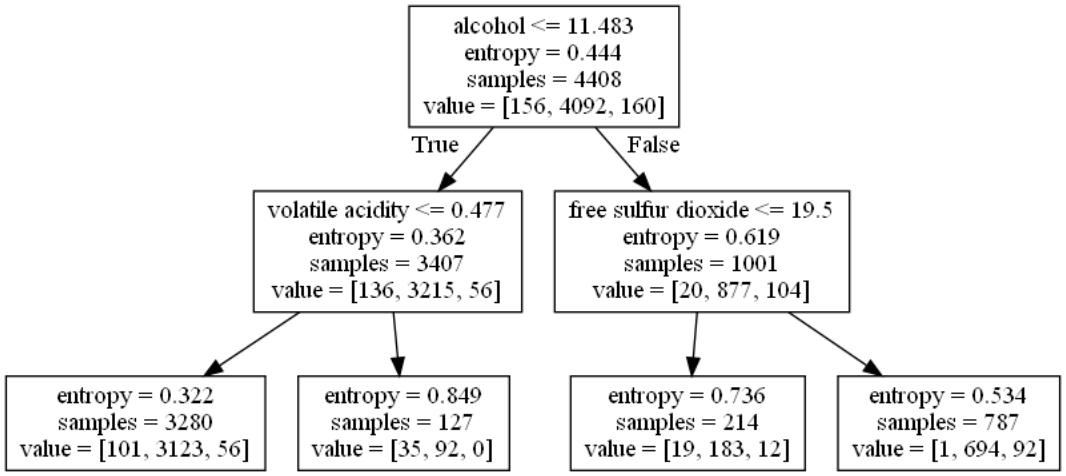
(a) Podział niezbalansowany.

(b) Podział zbalansowany.

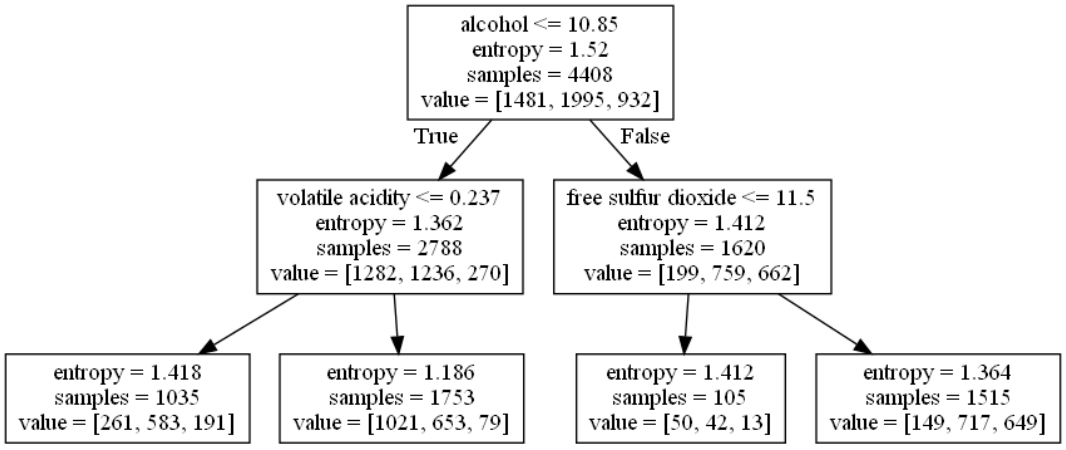
Rysunek 5: Porównanie histogramów atrybutu decyzyjnego *quality* po zastosowaniu nowych podziałów.



(a) Podział oryginalny (11 klas).

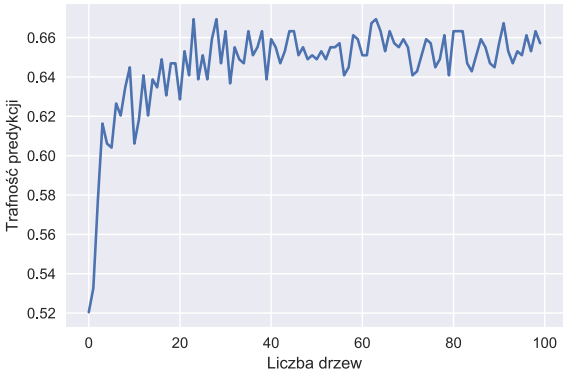


(b) Podział niezbalansowany.

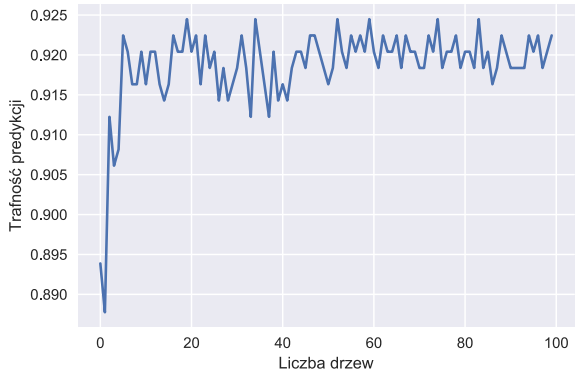


(c) Podział zbalansowany.

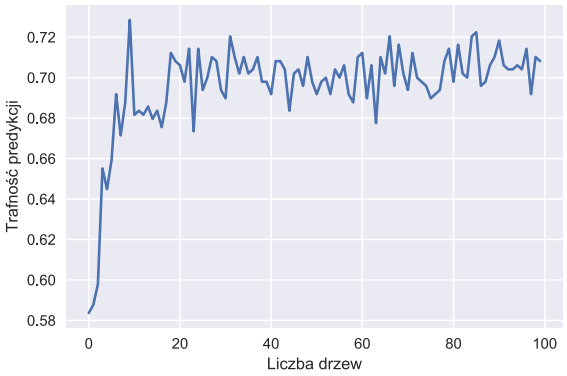
Rysunek 6: Porównanie drzew decyzyjnych wygenerowanych dla różnych podziałów.



(a) Zależność trafności predykcji od liczby drzew przy podziale oryginalnym.



(b) Zależność trafności predykcji od liczby drzew przy podziale niezbalansowanym dla 3 klas.



(c) Zależność trafności predykcji od liczby drzew przy podziale zbalansowanym dla 3 klas.

Rysunek 7: Porównanie drzew decyzyjnych wygenerowanych dla różnych podziałów.

trafność na poziomie 90% dla zbioru testowego. W przypadku trzeciego podziału - na 3 zbalansowane klasy otrzymano 49%, czyli niewiele więcej niż w przypadku oryginalnego podziału. Te wyniki są prezentowane dla drzew o maksymalnej głębokości 2, które są zaprezentowane na obrazkach 7. Szczegółowe wyniki dla drzew o głębokościach od 2 do 5 dla poszczególnych podziałów są zaprezentowane w tabeli 1.

Za pomocą lasu losowego wyznaczono ważności poszczególnych atrybutów, zostały one podsumowane w tabeli 2. Najważniejszym czynnikiem wpływającym na ocenę okazała się zawartość alkoholu w winie. Widać to na przykładowych drzewach na rysunku 7 (została ustawiona maksymalna głębokość 2 dla czytelności).

Na rysunku zaprezentowano trafności predykcji w zależności od liczby drzew w danym lesie. Na rysunku 7c, iż dla podziału zbalansowanego na 3 klasy dochodzimy do ponad 70% trafności, dla oryginalnego podziału nie przekraczamy 67% jednak i tak są to o wiele lepsze wyniki niż dla pojedynczego drzewa jak zaprezentowano w tabeli 1.

	2	3	4	5
Podział oryginalny	45.51%	46.12%	44.88%	49.18%
Podział na 3 niezbalansowane klasy	90.4%	90.4	90.81%	90.4%
Podział na 3 zbalansowane klasy	49.38%	48.57%	51.22%	52.04%

Tabela 1: Trafności predykcji w zależności od głębokości drzewa oraz rodzaju podziału.

Atrybut	Podział oryginalny	Podział niezbalansowany	Podział zbalansowany
alcohol	13.2%	9.8%	13.2%
density	9.83%	8.26%	9.8%
volatile acidity	9.66%	10.9%	9.9%
total sulfur dioxide	9.14%	8.7%	8.8%
free sulfur dioxide	8.94%	11.31%	8.6%
residual sugar	8.63%	8.7%	8.2
citric acid	8.52%	9.07%	8.3%
pH	8.11%	8.2%	8.36%
chlorides	8.06%	7.8%	8.9%
sulphates	7.9%	8.12%	7.9%
fixed acidity	7.79%	8.7%	7.8%

Tabela 2: Ważność poszczególnych atrybutów wyznaczone za pomocą losowego lasu w zależności od podziału.

Podział oryginalny	`GaussianNB`	`BernoulliNB`
Podział oryginalny	42.65%	41.83%
Podział na 3 niezbalansowane klasy	85.51%	90.4%
Podział na 3 zbalansowane klasy	49.38%	45.91%

Tabela 3: Trafności predykcji w zależności od klasyfikatora oraz rodzaju podziału.

4.2 Naive Bayes

W przypadku tego klasyfikatora wykorzystano dwie implementacje `GaussianNB` oraz `BernoulliNB`. W tabeli zostały zagregowane wyniki dla różnych podziałów 3. Jak widzimy wyniki są zbliżone do wyników drzewa decyzyjnego, najlepszą trafność otrzymujemy dla podziału niezbalansowanego na 3 klasy.

4.3 SVM

Ostatnim klasyfikatorem wykorzystanym do predykcji jakości wina jest SVM, wykorzystano tutaj klasę `SVC` - czyli *Support vector classifier*. Oprócz domyślnych parametrów dodano także wagi - parametr *class_weight*. W pierwszym przypadku zbadano trafność dla domyślnego parametru - bez balansowania klas. W drugim przypadku ustawiono parametr *class_weight* na ‚balanced', dla którego wagi są dobierane wg wzoru 1, tak więc są one odwrotnie proporcjonalne do częstotliwości występowania danej klasy w danych wejściowych. W trzecim przypadku jako wagi wzięto sumę przykładów podzieloną przez liczność danej klasy. Parametr C został ustawiony na wartość 1, natomiast parametr *kernel* został ustawiony na wartość ‚rbf'. W przypadku *SVM* również najlepszą trafność otrzymano dla podziału niezbalansowanego dla 3 klas, wynik 90% jest praktycznie taki sam jak dla drzew decyzyjnych, czy klasyfikatora

	class_weight=None	class_weight='balanced'	class_weight=calculated
Podział oryginalny	53.26%	43.06%	43.06%
Podział na 3 niezbalansowane klasy	90.61%	64.08%	70.0%
Podział na 3 zbalansowane klasy	59.18%	57.95%	59.7%

Tabela 4: Trafności predykcji w zależności od rodzaju zbalansowania (parametr `class_weight`) oraz rodzaju podziału.

bayesowskiego. W tabeli 4 zostały podsumowane wyniki dla klasyfikatora *SVM*.

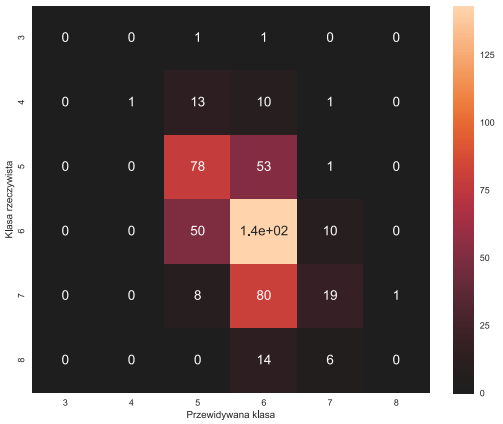
$$\frac{n_{samples}}{n_{classes} * np.bincount(y)} \tag{1}$$

5 Podsumowanie

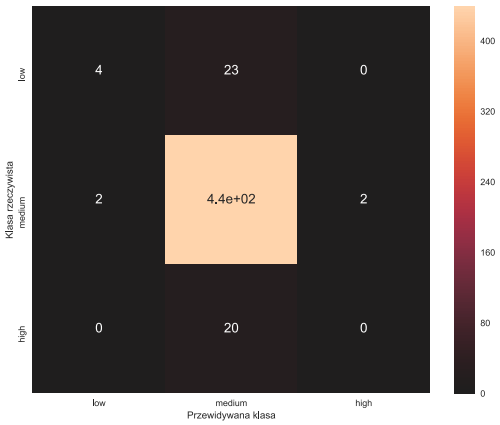
W badanym zbiorze danych okazało się, iż niezbalansowanie danych mocno wpływa na wyniki każdego z klasyfikatorów. Okazuje się, iż niezbalansowane dane są problemem dla każdego z nich i należy mocno zadbać o to, by zbierane dane były zbalansowane. Jeżeli jednak nie da się tego osiągnąć to należy pamiętać, iż na przykład miara *accuracy*, nie jest w tym przypadku najlepszą miarą. W wynikach pojawia się tylko miara trafności klasyfikacji, jednakże analizując macierze pomyłek dla każdego z klasyfikatorów doszedłem do wniosku iż miara *F1-score* jest lepsza w tym przypadku niż *accuracy*, która bierze tylko dobrze zaklasyfikowane przykłady. Dla przykładu, w przypadku użycia klasyfikatora typu drzewo decyzyjne otrzymaliśmy następujące macierze pomyłek widoczne na obrazku 8. Jak widać, w przypadku gdzie trafność była najlepsza - zbiór niezbalansowany dla 3 klas, i wynosiła ona 90.4% dla drzewa o głębokości 5, otrzymaliśmy następujący wektor ocen F-1 `[0.24, 0.94, 0.0]`. Natomiast, dla drzewa o takiej samej maksymalnej głębokości, uczonego na zbiorze zbalansowanym otrzymaliśmy trafność 52%, jednakże w tym przypadku wektor ocen F-1 wygląda trochę lepiej `[0.58, 0.49, 0.48]` choć i tak jest daleki od ideału. Dopiero dla sieci neuronowej o warstwach ukrytych `(121, 60, 30)` udało się dojść do lepszego niż poprzednie wyniku trafności dla zbioru oryginalnego oraz zbalansowanego i dosyć dobrego wyniku F-1 `[0.71, 0.64, 0.70]`.

Literatura

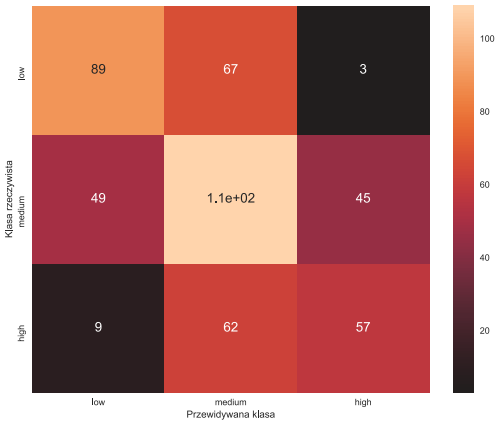
[1] Wine quality dataset. Machine learning repository, 2009. https://archive.ics.uci.edu/ml/datasets/Wine+Quality.



(a) Macierz pomyłek dla podziału oryginalnego.



(b) Macierz pomyłek dla podziału niezbalansowanego, 3 klasy.



(c) Macierz pomyłek dla podziału zbalansowanego, 3 klasy.

Rysunek 8: Macierze pomyłek dla drzewa decyzyjnego o maksymalnej głębokości 5.