Eksploracja danych, lista 2
Sprawozdanie

Mateusz Kubuszok, 179956

Wstęp

Analizowanym zbiorem danych jest baza informacji na temat win. Są one kategoryzowane wg kultywarów (od 1 do 3). Posiadają również cechy ilorazowe opisujące ich skład chemiczny – zawartość wybranych substancji mających wpływ na walory smakowe wina.

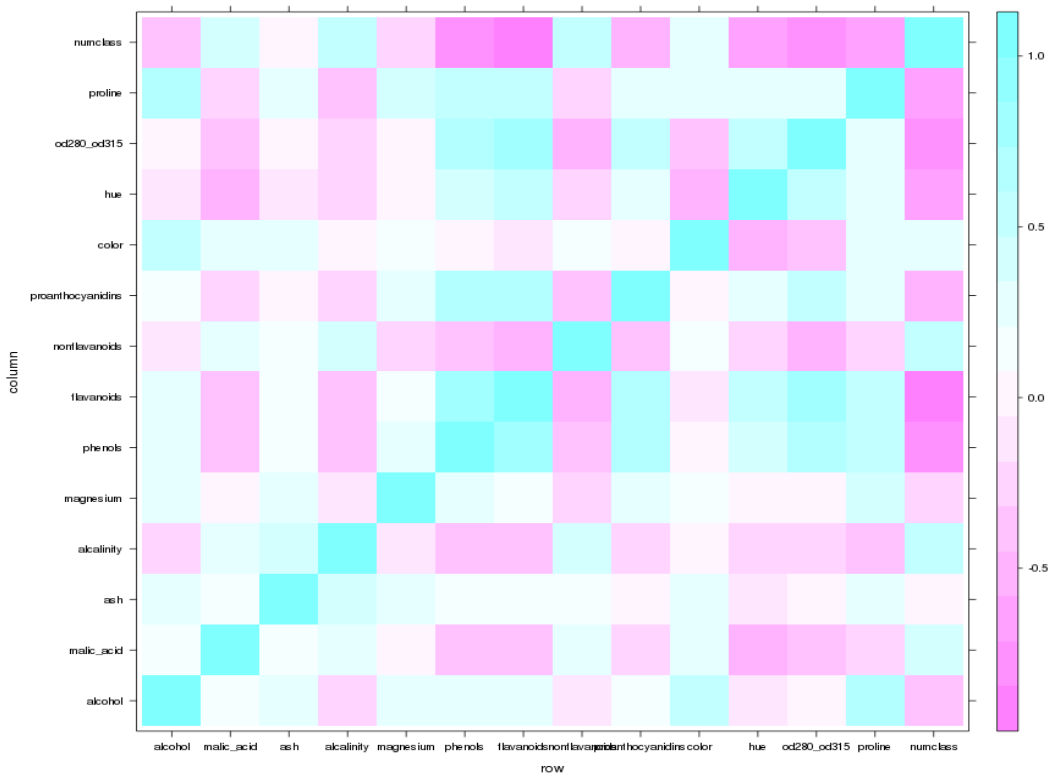
Poszczególne cechy to:

1) zawartość alkoholu,
2) zawartość kwasu jabłkowego,
3) zawartość popiołów,
4) kwaśność zawartych popiołów,
5) zawartość magnezu,
6) zawartość fenoli ogółem,
7) zawartość flawonoidów,
8) zawartość fenoli niebędących flawonoidami,
9) zawartość proantocyjanidyny,
10) intensywność barwy,
11) nasycenie barwy,
12) współczynnik OD280/OD315,
13) zawartość proliny.

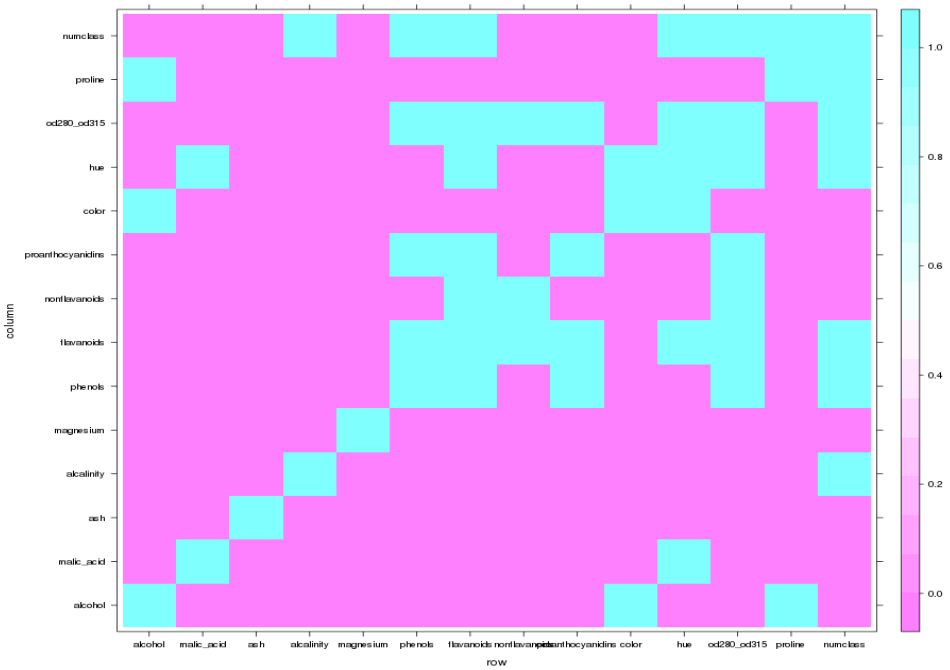
Rozpatrywanym zagadnieniem jest dyskryminacja cech wskazujących za przynależność win do danego kultywaru oraz stworzenie modelu umożliwiającego predykcję przynależności wina do odpowiedniej grupy.

Typowanie cech

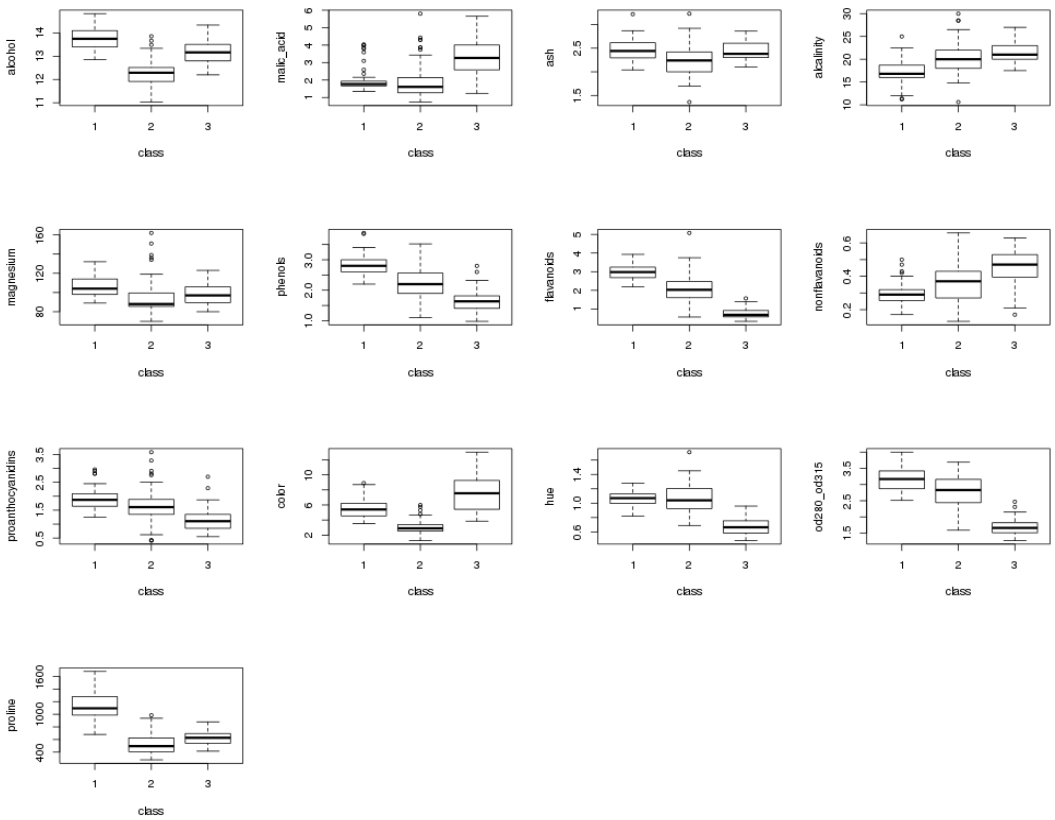
Aby ustalić związek cech z kultywarem możemy obliczyć kowariancję dla wszystkich cech. Otrzymamy wówczas wykres podobny do poniższego:



Patrząc na wartości kowariancji wybranej cechy z wartością ***numclass*** (numeryczna wartość klasy w odróżnieniu od nominalnej) typuje nam w najprostszy sposób kandydatów na cechy dyskryminujące (nie gwarantuje to jednak sukcesu). Dla zwiększenia przejrzystości możemy podkreślić wartości dla których zależności są szczególnie silne: większe od ***0.5*** lub mniejsze od ***-0.5***:



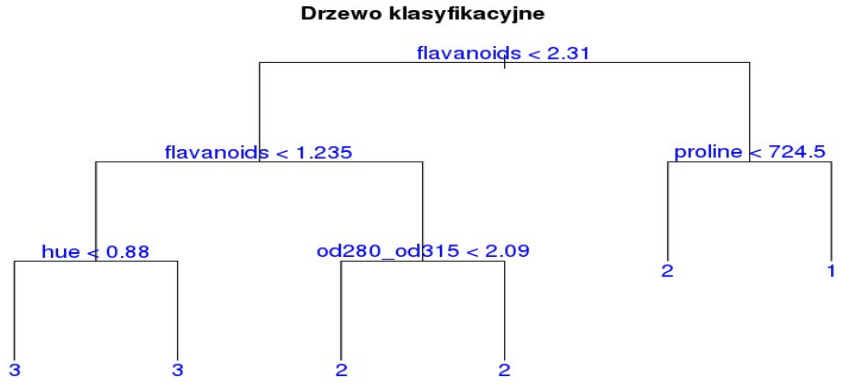
Wykres sugeruje nam, że dobrymi kandydatami będą wartości: zawartości ***proliny***, czynnika ***OD280/OD315***, nasycenia ***barwy***, zawartości ***flawonoidów***, zawartości ***fenoli*** oraz ***kwaśność*** zawartego popiołu. Wnioski te potwierdza wykres pudełkowy (sugerujący jednocześnie kilka innych kandydatów np. **barwę**):



Klasyfikacja

Wytypowane cechy układają się tam w sposób sugerujący zależność intensywności danej cechy od kultywaru, pomija jednak kilka cech które nie robią wrażenia zależności liniowej - klasa jest jednak zmienną normatywną i nie można jej na dłuższą metę traktować jako numeryczną. Możemy jednak sprawdzić jakość wytypowanych zmiennych budując modele predykcyjne przy pomocy algorytmów ***k najbliższych sąsiadów*** oraz ***drzew klasyfikacyjnych***.

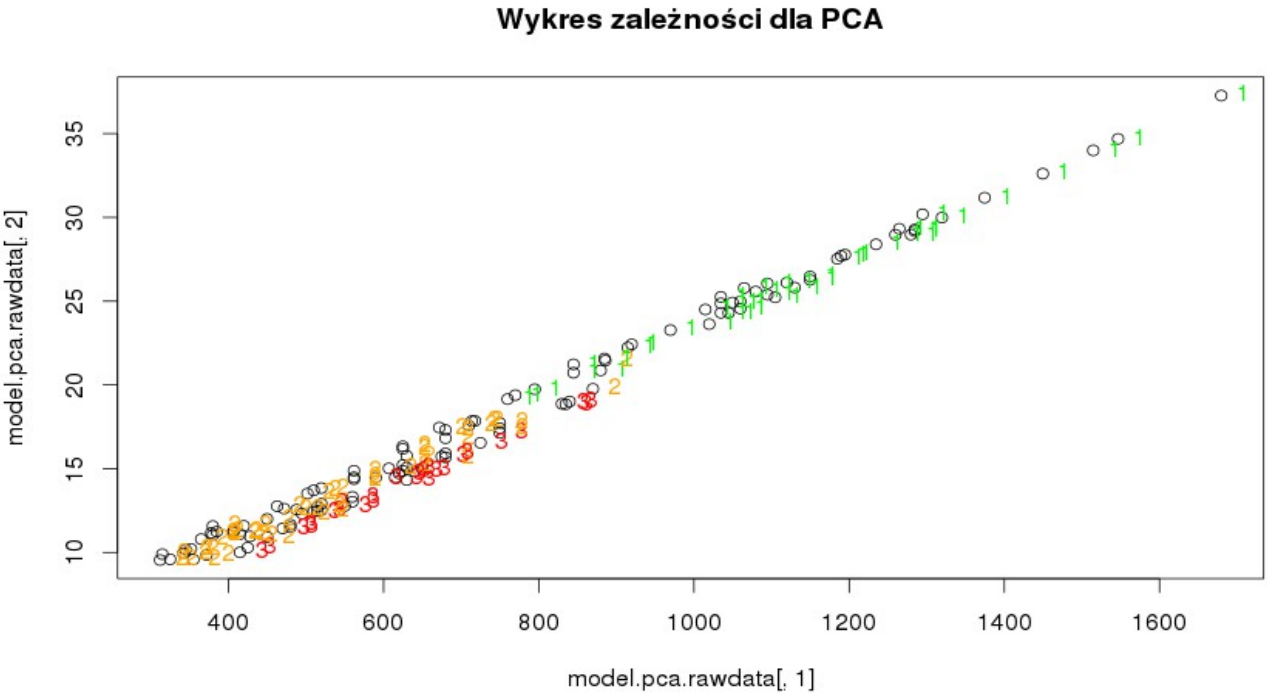
Dane podzielmy losowo na 2 zbiory: uczący oraz testowy. Podziału dokonamy tak, aby zbiór uczący zawierał 2/3 zaś testowy 1/3 pierwotnego zbioru danych. Po zastosowaniu algorytmu ***tree*** z pakietu ***tree*** otrzymamy drzewo podobne do poniższego:



Po użyciu go na zbiorze testowym oraz porównaniu przewidywanych wyników z rzeczywistymi otrzymano błąd klasyfikacji równy ***17%***. Zastosowawszy metodę k najbliższych sąsiadów dla k równego ***5*** otrzymano błąd równy ***38%***. Zmiany wartości k na inne nie przyniosły wyraźnej poprawy skuteczności predykcji.

Redukcja wymiarów

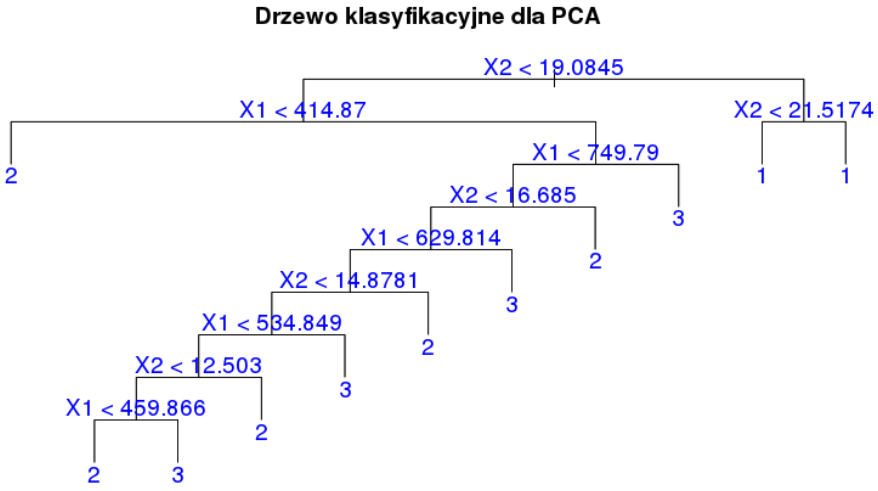
Ponieważ wszystkie porównywane cechy opisują wartości numeryczne możemy podjąć próbę redukcji wymiarów algorytmem **PCA**. Po redukcji do 2 wymiarów możemy otrzymany model w całości przedstawić na wykresie:



Dla tak przedstawionych danych widać, że wina pierwszej klasy mają wyraźnie inne cechy niż wina 2 i 3 klasy. Odróżnienie 2 ostatnich jest już jednak nieco trudniejsze, gdyż ich zbiory częściowo zachodzą na siebie. Przydatne wydaje

się sprawdzenie zachowania algorytmów klasyfikujących dla tak przygotowanych danych.

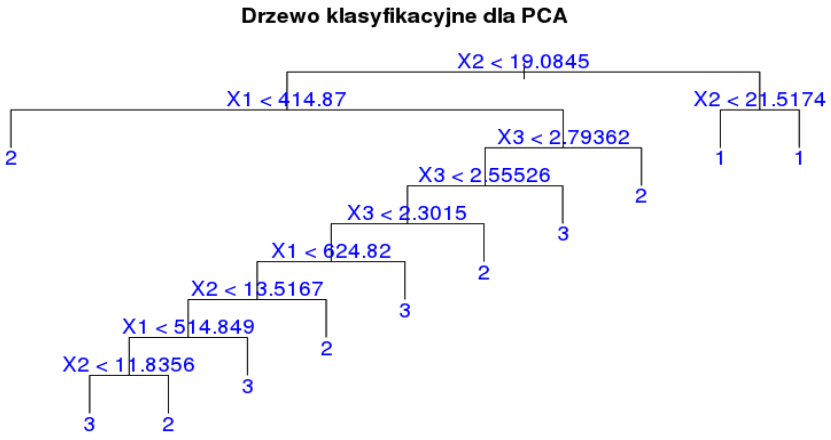
Drzewo klasyfikacyjne dla danych o zredukowanych wymiarach:



pogorszyło jednak wyniki do błędu równego ***23%***. Dla k najbliższych sąsiadów otrzymano z kolei błąd równy ***35%***, a więc lepszy.

Po zwiększeniu liczby wymiarów do 3 otrzymany wykres:

który nie dodaje znacząco wiele informacji. Wygenerowane dla niego drzewo klasyfikacyjne:
nie powoduje poprawy predykcji (błąd ***38%***), podobnie dla kNN (***36%***).



Podsumowanie

Najlepsze wyniki predykcji klasy wina osiągnięto dla algorytmu drzew klasyfikacyjnych bez redukcji wymiarów. Pozostałe przypadki (drzewa klasyfikacyjne z redukcją wymiarów, oba przypadki dla k najbliższych sąsiadów) osiągnęły porównywalne ale wyraźnie gorsze wyniki.

Dalsze próby uruchamiania całej procedury pokazały, że jakoś predykcji w dużej mierze zależy od dobrego zbioru uczącego – kilkakrotne próby pokazywały lepsze jak i gorsze osiągnięcia algorytmów, choć ich względne osiągi były podobne – drzewa klasyfikacyjne bez redukcji wymiarów miały zawsze najlepsze wyniki, drzewa działające na zredukowanych wymiarach nieco gorsze osiągi, najgorsze zaś algorytm k-najbliższych sąsiadów.

Prawdopodobnie dodanie kilku cech nie wytypowanych w pierwszym etapie mogłoby poprawić zdolność odgadywania kultywaru na podstawie składu chemicznego powstałego z nich wina. Jednak już teraz możliwe jest odgadnięcie ich z dużym prawdopodobieństwem.