Projekt PUMA

Wiktor Hosumbek
Szymon Joszko

Plan prezentacji

- Cel projektu
- Zbiór danych
- Metody
- Wyniki
- Podsumowanie

Cel projektu

Celem projektu była klasyfikacja jakości win na podstawie ich składu. Do wykonania klasyfikacji użyliśmy pięciu metod uczenia maszynowego i je porównaliśmy

Zbiór danych

Zbiór danych użytych w projekcie pochodzi ze strony UCI dokładny link: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

“P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.”

Zbiór danych

Zbiór ten dotyczy jakości portugalskiego wina “Vinho Verde” i jest podzielony na dwa podzbiory, jeden dla czerwonego wina a drugi dla białego. Dane wejściowe zawierają w sobie tylko dane psychochemiczne takie jak ph, zawartość alkoholu, A nie zawierają informacji na temat marki, ceny, typu winogron itd. jest to spowodowane ochroną prywatności wytwórni tych win. Dane są prawdziwe i nie posiadają brakujących argumentów.

Daną wyjściową jest ocena w skali rosnącej (0-10) im wyższa, tym lepsza jakość wina. Dane te nie są zbalansowane co oznacza że istnieje w bazie znaczna większość win średnich niż win bardzo słabych czy wybitnie dobrych.

Baza danych posiada 4898 rekordów wina białego i 1599 rekordów wina czerwonego co daje łącznie 6497 rekordów

Zbiór danych

Dane wejściowe:

1 - kwasowość stała

2 - kwasowość lotna

3 - kwas cytrynowy

4 - cukier resztkowy

5 - chlorki

6 - wolny dwutlenek siarki

7 - całkowity dwutlenek siarki

8 - gęstość

9 - pH

10 - siarczany

11 - alkohol

Dane wyjściowe:

12 - Jakość

Metody

Do wykonania klasyfikacji użyliśmy i porównaliśmy 5 metod:

1. Metoda SVM
2. Drzewo decyzyjne
3. Naiwny Klasyfikator Bayesowski
4. Regresja logistyczna
5. Lasy losowe

Metoda SVM

Maszyna Wektorów wspierających (Support Vector Machine) ma na celu znalezienie takiej prostej(hiperpłaszczyzny separującej), która oddziela przykłady ze zbioru treningowego z maksymalnym marginesem.

Metody znajdowania najlepszych parametrów:

GridSearchCV() - przeszukuje podane parametry jeden po drugim

RandomizedSearchCV() - przeszukuje podane parametry losowo

obie metody służą do znajdowania najlepszych parametrów pracy algorytmu.

Metoda SVM

Zakresy parametrów:

Do uczenia klasyfikowania do wszystkich klas użyliśmy parametrów:

```
parameters = {'kernel': ('linear', 'rbf'),
              'C': [2 ** -2, 2 ** 2],
              'gamma': [2 ** -2, 2 ** 2],
              'degree': [1, 2, 3, 4]}
```

Metoda SVM

Zakresy parametrów:

Do uczenia klasyfikowania do trzech klas użyliśmy parametrów:

```
parameters = {'kernel': ('linear', 'rbf'),

              'C': [2 ** -6, 2 ** 6],

              'gamma': [2 ** -6, 2 ** 6],

              'degree': [1, 2, 3, 4, 5, 6, 7, 8]}
```

Dodatkowo dla zbioru wina czerwonego użyliśmy poszerzonego zakresu C i gamma do [2**-4,2**4] ale nie przyniosło to lepszych wyników

Drzewo decyzyjne

Drzewo decyzyjne to nieparametryczna metoda uczenia maszynowego nadzorowanego stosowana do klasyfikacji i regresji. Jej celem stworzenie jest modelu przewidującego wartość docelową poprzez utworzenie reguł decyzyjnych na podstawie cech danej próbki.

Opis:

- Wiedza jest reprezentowana w postaci drzewa.
- Węzły drzewa określają sposób podziału przestrzeni cech na obszary/klasy.
- Liście drzewa określają klasę, do której należy klasyfikowany obiekt.
- Proces klasyfikacji polega na przejściu od korzenia drzewa do liści.

Naiwny Klasyfikator Bayesowski

Naiwny klasyfikator bayesowski - jest prostym probabilistycznym klasyfikatorem zakłada on wzajemną niezależność zmiennych niezależnych (naiwność).Nazywany też jako „model cech niezależnych”. Model prawdopodobieństwa można wyprowadzić korzystając z twierdzenia Bayesa.

Opis:

- Twierdzenie Bayesa określa prawdopodobieństwa warunkowe dwóch zdarzeń warunkujących się wzajemnie.
- Wyliczane prawdopodobieństwo to prawdopodobieństwo a posteriori.
- Wnioskowanie bayesowskie polega na sekwencyjnym wykorzystaniu reguły Bayesa.
- Wnioskowanie bayesowskie pozwala na aktualizację prawdopodobieństw, które mogą służyć do aktualizacji prawdopodobieństw zajścia zdarzeń z nimi współzależnych

Regresja logistyczna

Metoda do szacowania prawdopodobieństwa przynależności przykładu do określonej klasy. Jest to klasyfikator binarny czyli jeśli prawdopodobieństwo przekracza 50% to próbka należy do klasy pozytywnej i w odwrotnym przypadku do negatywnej

Lasy losowe

Metoda zespołowego uczenia maszynowego, która polega na konstruowaniu wielu drzew decyzyjnych w czasie uczenia i generowaniu klasy, która jest dominantą Losowe lasy decyzyjne poprawiają tendencję drzew decyzyjnych do nadmiernego dopasowywania się do zestawu treningowego.

Wyniki

- wyniki dokładności klasyfikacji dla 10 klas wina czerwonego

	Decision Tree	gaussianNB	SVM	Random Forest	Logistic Regression
zb. treningowy	1.0	0.5460232350312779	0.5915996425379804	0.683646	0.589812
zb. testowy	0.5770833333333333	0.5479166666666667	0.6041666666666666	0.652083	0.620833

Wyniki

- wyniki dokładności klasyfikacji dla 10 klas wina białego

	Decision Tree	gaussianNB	SVM	Random Forest	Logistic Regression
zb. treningowy	1.0	0.458284714119 01983	0.99679113185 53092	0.580222	0.542299
zb. testowy	0.57278911564 62585	0.448979591836 7347	0.58503401360 54422	0.527211	0.521769

Wyniki

- wyniki dokładności klasyfikacji dla 3 klas wina czerwonego

	Decision Tree	gaussianNB	SVM	Random Forest	Logistic Regression
zb. treningowy	1.0	1.0	1.0	0.999106	0.993744
zb. testowy	0.9958333333333333	1.0	1.0	0.989583	0.991667

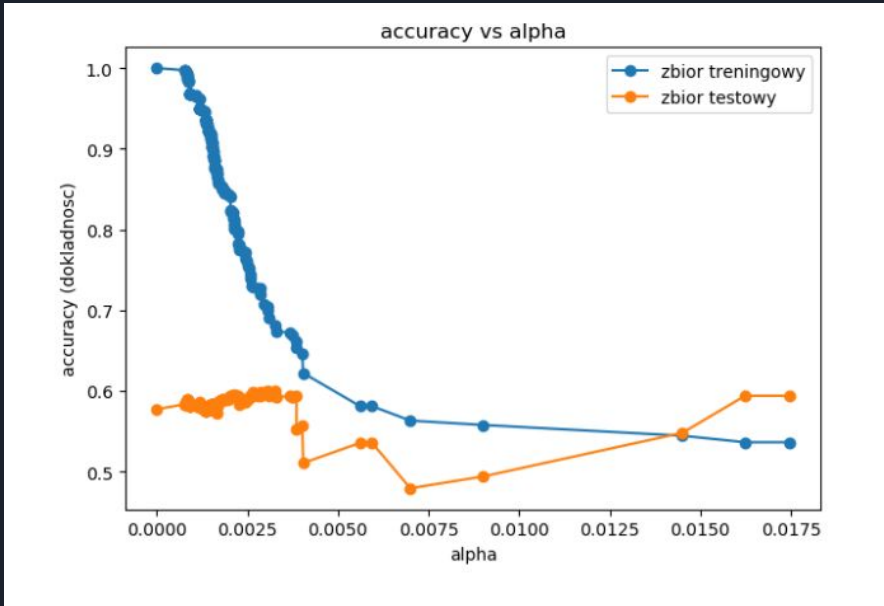
Wyniki

- wyniki dokładności klasyfikacji dla 3 klas wina białego

	Decision Tree	gaussianNB	SVM	Random Forest	Logistic Regression
zb. treningowy	1.0	0.999708284714119	1.0	0.989790	1.000000
zb. testowy	1.0	0.999319727891564	1.0	0.987075	1.000000

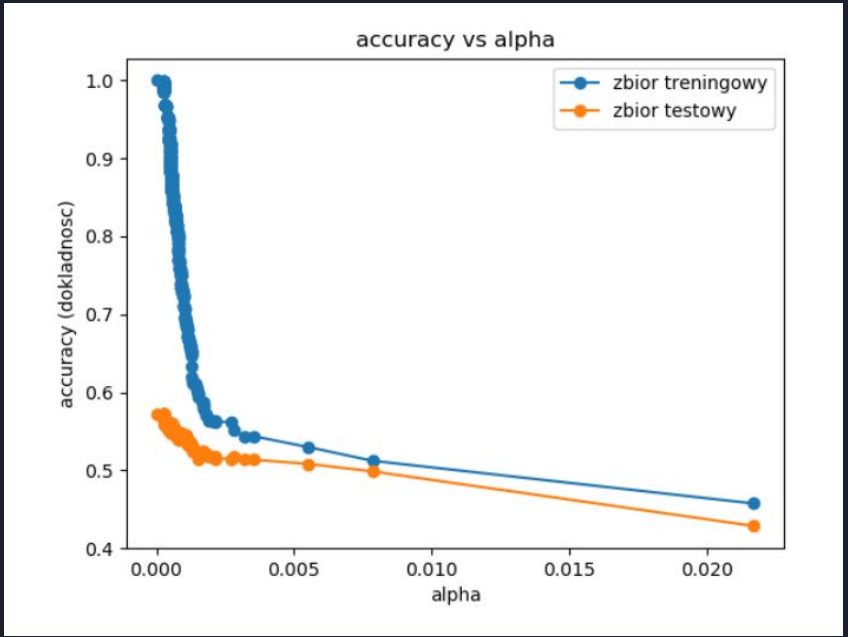
Wyniki

- Wyniki accuracy vs alpha dla wina czerwonego i 10 klas



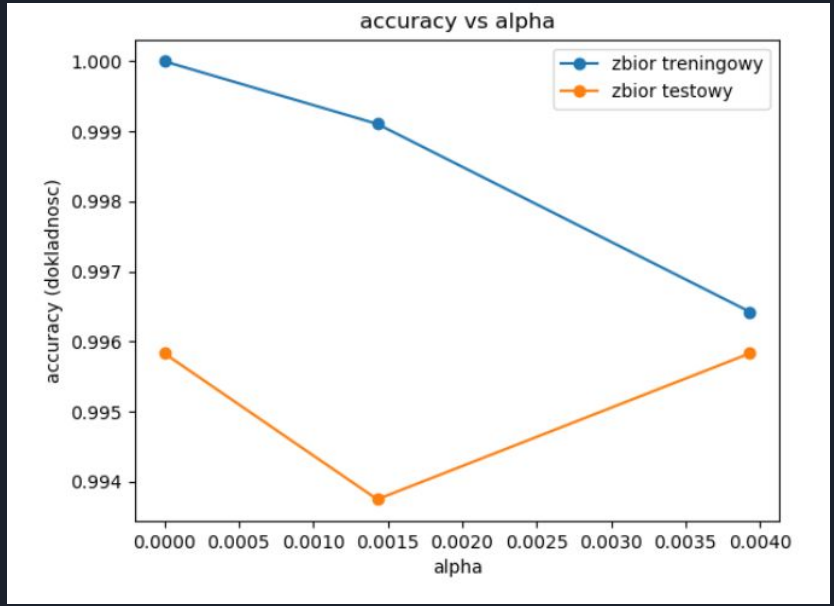
Wyniki

- Wyniki accuracy vs alpha dla wina białego 10 klas



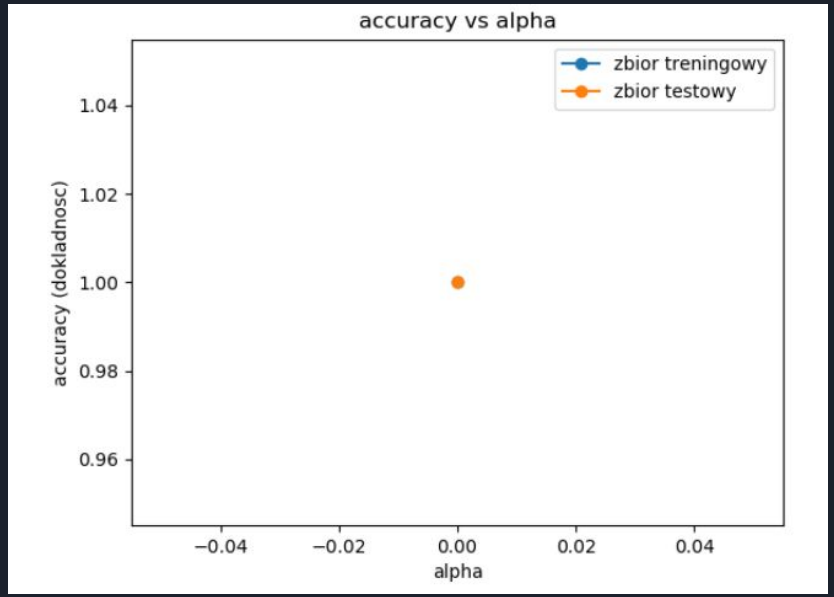
Wyniki

- Wyniki accuracy vs alpha dla wina czerwonego i 3 klas



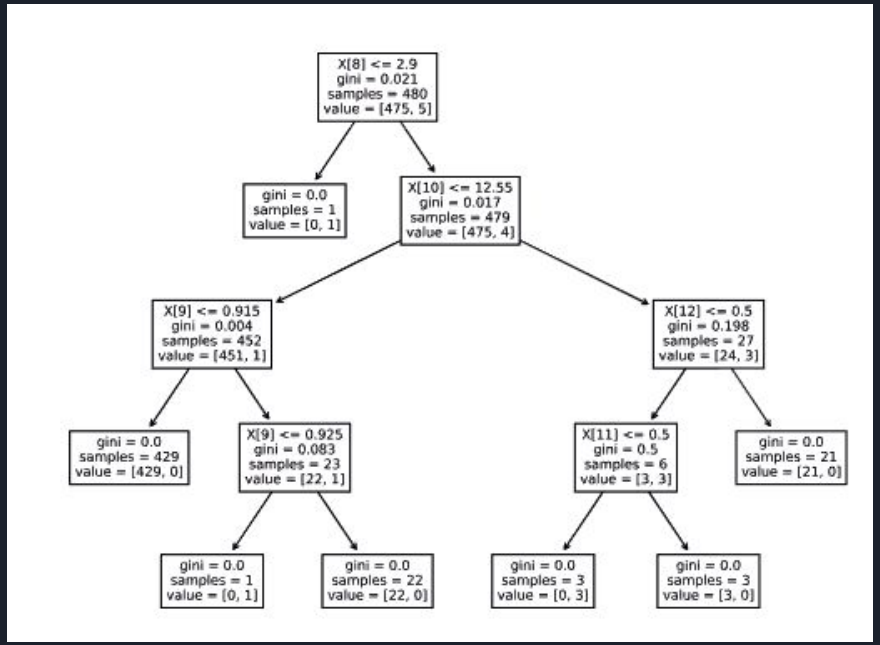
Wyniki

- Wyniki accuracy vs alpha dla wina białego i 3 klas



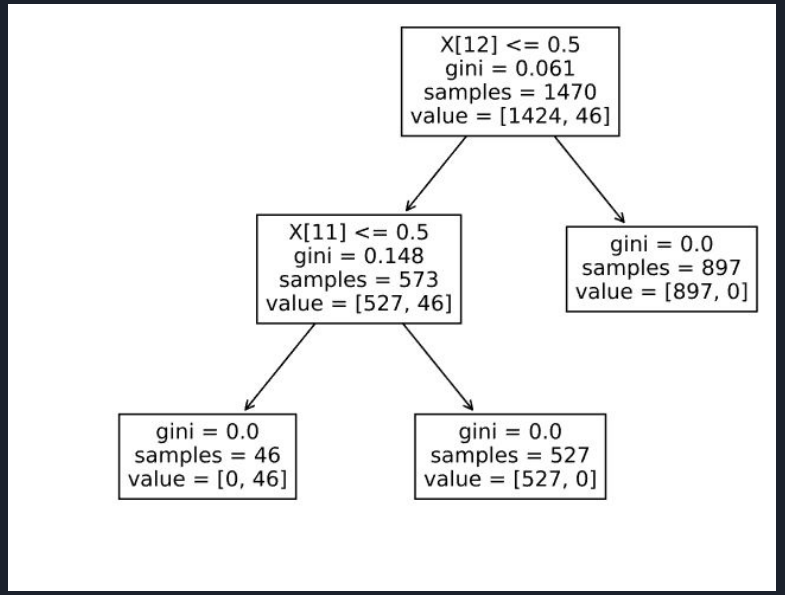
Wyniki

- wygląd drzewa dla wina czerwonego i 3 klas



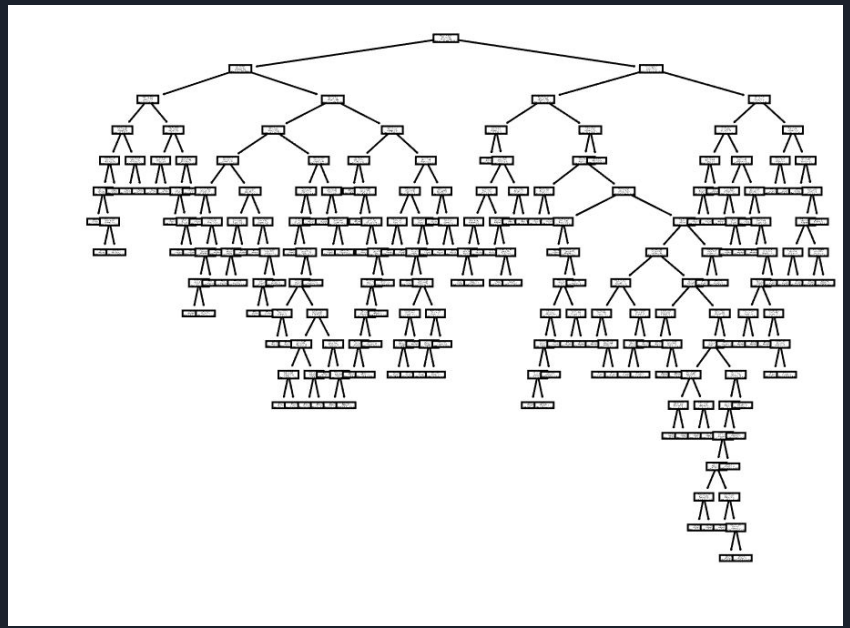
Wyniki

- wygląd drzewa dla wina białego i 3 klas



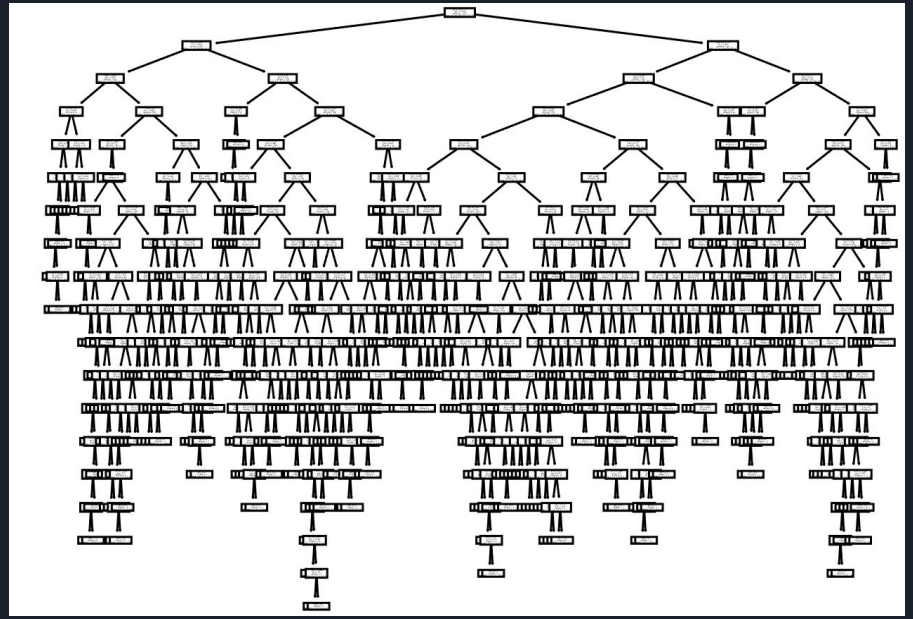
Wyniki

- wygląd drzewa dla wina czerwonego i 10 klas



Wyniki

- wygląd drzewa dla wina białego i 10 klas



Podsumowanie

Do sklasyfikowania jakości wina na podstawie jego składu użyliśmy pięciu metod z różnymi parametrami. Ze względu na nie równomierne rozproszenie danych (większość danych klasyfikowała się do środkowego zakresu) żadna z metod nie osiągnęła pożądanego przez nas efektu. Do klasyfikacji na 10 różnych jakości najlepsza okazała się metoda lasu losowego - dla wina czerwonego i metoda svm dla wina białego. Nie przekroczyły one ale dokładności 70% dla zbioru testowego. Wobec tego postanowiliśmy klasyfikować wina do 3 grup. Niskiej jakości, średniej i wysokiej jakości. Dla tak złagodzonych kryteriów, każda z pięciu testowanych metod osiągnęła dokładność powyżej 98% dla zbioru testowego. Wszystkie metody dawały porównywalne rezultaty na bardzo zadowalającym poziomie.