Siódme laboratorium z przedmiotu Wprowadzenie do sztucznej inteligencji

Autor
Łukasz Jaremek
310710

Data
styczeń 2022

Praca wykonana samodzielnie.

Użyte technologie

Język:
- Python 3.9.9

Biblioteki ponad standardowe:
- matplotlib 3.5.1
- pandas 1.3.5

Model Bayesowski

Problem:

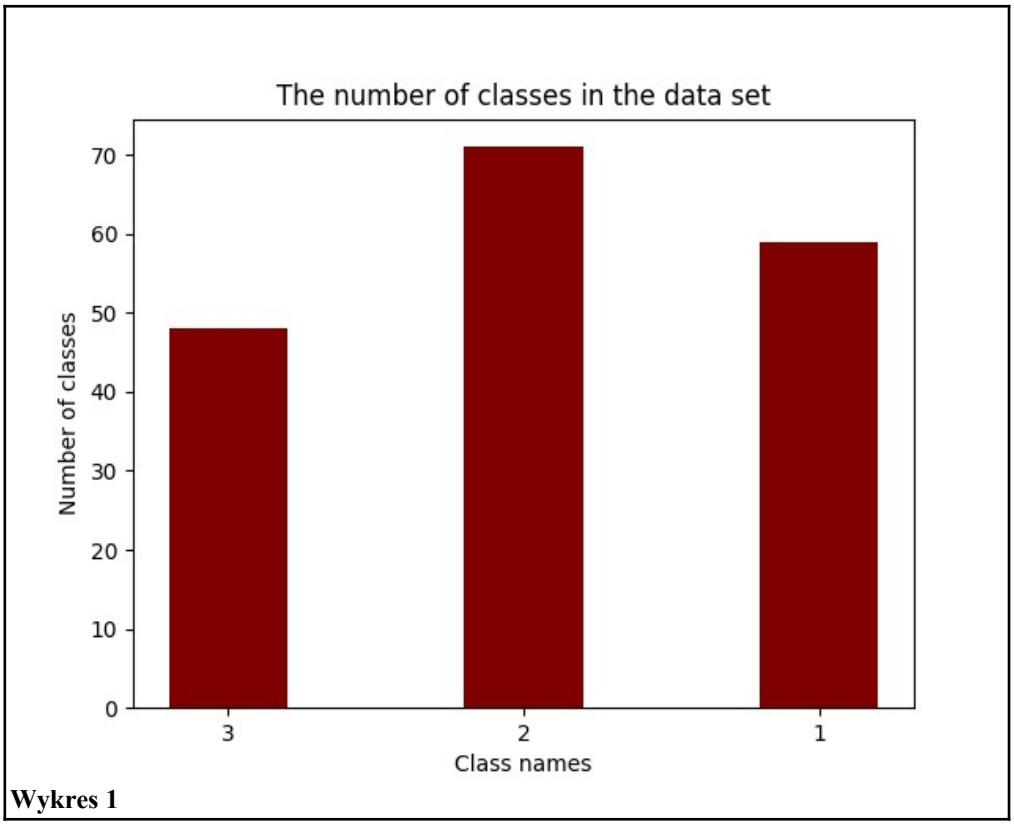
Zaimplementowanie naiwnego klasyfikatora Bayesowskiego do klasyfikacji win na podstawie zbioru danych zawierającego chemiczną analizę trzech włoskich win.

Analiza zbioru danych

Zbiór zawiera 178 rekordów, które zawierają 3 klasy.
Każda z nich posiada 13 atrybutów w postaci liczb zmiennoprzecinkowych:

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

Rozkład liczebności klas przedstawiony jest na wykresie 1.



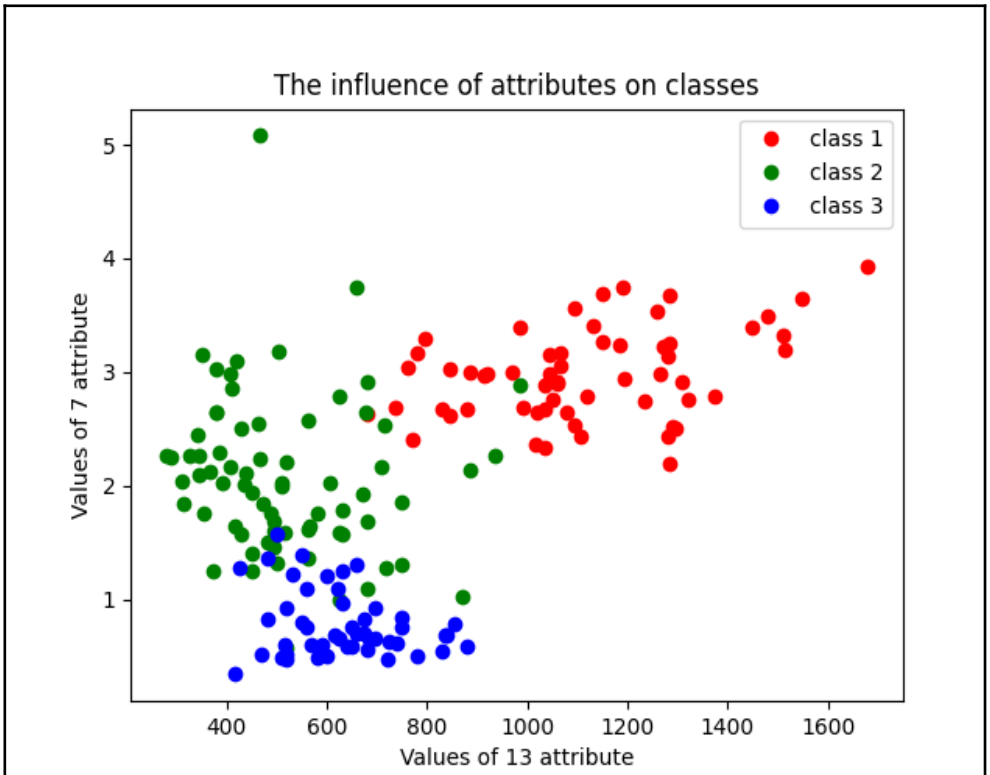
Wykres 1

Jak widać, liczby wystąpień klas są w zbliżonych ilościach.

Separowalność atrybutów

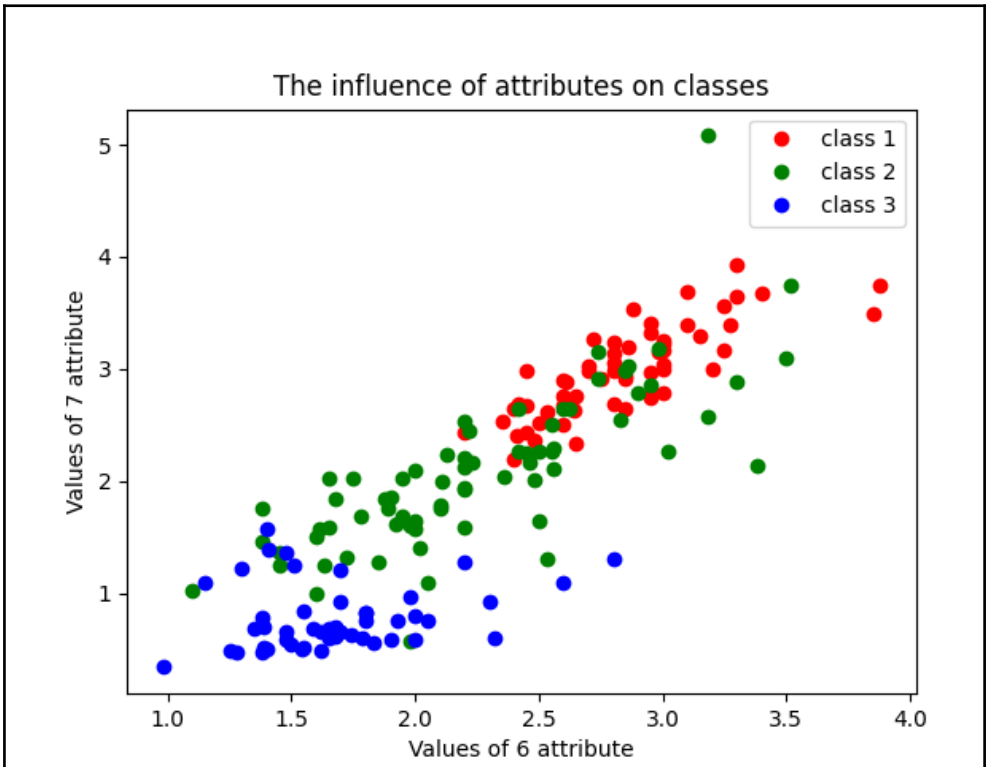
Sprawdźmy czy istnieją atrybuty, które da się odseparować aby wyraźnie zobaczyć, że każda klasa charakteryzuje się własnymi wartościami.

Na wykresie 2 można bez większych problemów odseparować od siebie wszystkie 3 klasy na podstawie atrybutów 7 (Flavanoids) i 13 (Proline).



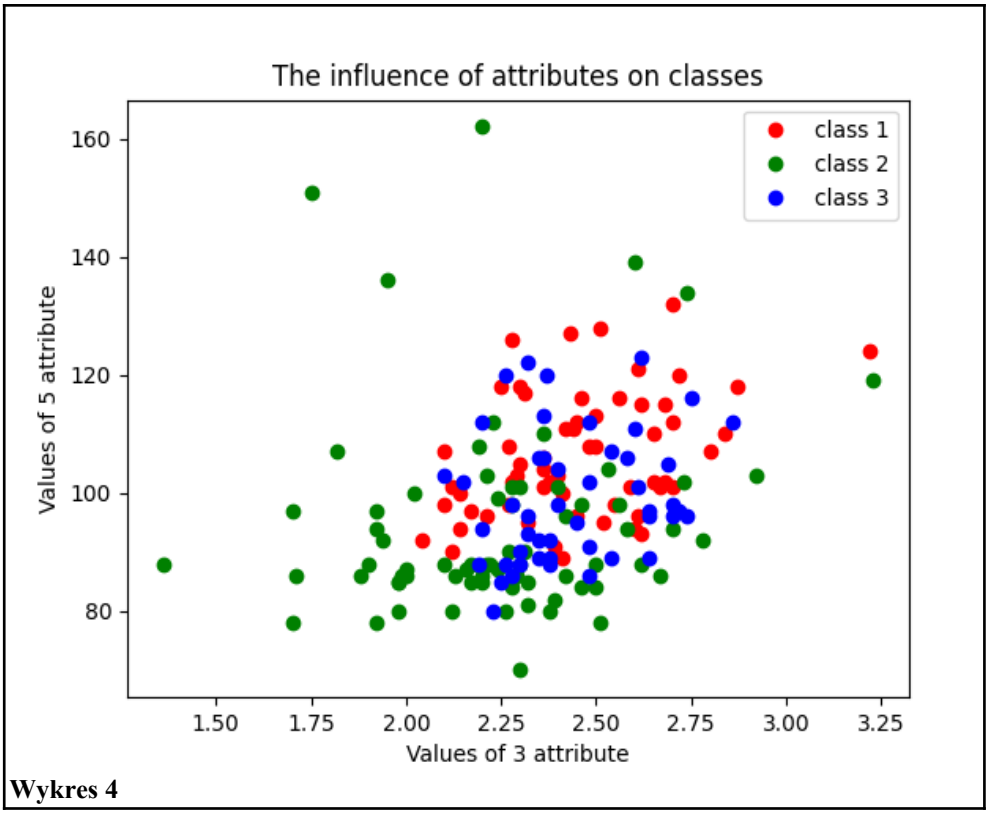
Wykres 2

Wykres 3 przedstawie częściowo separowalne klasy. O ile klasę trzecią można łatwo odróżnić, to klasy pierwsza i druga są już cięższe do jednoznacznego określenia.



Wykres 3

Natomiast wykres 4 pokazuje sytuację, która jest najczęściej spotykana – nie da się odseparować klas na podstawie tylko dwóch parametrów. Aby to zrobić, potrzebujemy kombinacji różnych parametrów oraz pracy na wielu wymiarach, co dla człowieka nie jest przystępną opcją.



Wykres 4

Skuteczność modelu

Sprawdźmy skuteczność modelu. Podzielę zbiór danych na treningowy oraz testowy w różnych proporcjach a następnie porównam wyniki.

1. Podział danych 0.5/0.5 na treningowe/testowe:

	1	2	3
1	28	0	0
2	3	34	2
3	0	0	22

Macierz pomyłek klasy 1:

	Wartości rzeczywiste	
Wartości oczekiwane	28	3
	0	56

Macierz pomyłek klasy 2:

	Wartości rzeczywiste	
Wartości oczekiwane	34	0
	5	50

Macierz pomyłek klasy 3:

	Wartości rzeczywiste	
Wartości oczekiwane	22	2
	0	62

TPR: 1.0
FPR: 0.03
PPV: 0.91
ACC: 1.0

2. Podział danych 0.65/0.35 na treningowe/testowe:

	1	2	3
1	12	2	0
2	0	28	0
3	0	0	20

Macierz pomyłek klasy 1:

	Wartości rzeczywiste	
Wartości oczekiwane	12	0
	2	48

Macierz pomyłek klasy 2:

	Wartości rzeczywiste	
Wartości oczekiwane	28	2
	0	32

Macierz pomyłek klasy 3:

	Wartości rzeczywiste	
Wartości oczekiwane	20	0
	0	40

TPR: 1.0
FPR: 0.0
PPV: 1.0
ACC: 1.0

3. Podział danych 0.8/0.2 na treningowe/testowe:

	1	2	3
1	11	0	0
2	0	13	0
3	0	0	11

Macierz pomyłek klasy 1:

	Wartości rzeczywiste	
Wartości oczekiwane	11	0
	0	24

Macierz pomyłek klasy 2:

	Wartości rzeczywiste	
Wartości oczekiwane	13	0
	0	22

Macierz pomyłek klasy 3:

	Wartości rzeczywiste	
Wartości oczekiwane	11	0
	0	24

TPR: 1.0
FPR: 0.0
PPV: 1.0
ACC: 1.0

Można zauważyć niezależnie od podziału danych bardzo dobre wyniki modelu. Dane zostały początkowo pseudo-losowo posortowane. Nie przeprowadziłem analizy bez posortowania danych ponieważ dane domyślnie są posortowane zgłędem klas, co oznacza, że aby model działał poprawnie należało by dokonać takiego podziału danych, aby zbiór treningowy łapał pierwsze dwie klasy oraz część trzeciej. Wtedy zbiór testowy zawierałby tylko fragment trzeciej klasy, co by wykazywało 100% skuteczność modelu – co oczywiście nie byłoby miarodajne.

Podsumowanie laboratorium

Implementacja modelu była zaskakująco prosta w porównaniu do materiału przedstawionego na wykładach. Model pomimo małych ilości danych wykazuje wysoką skuteczność, lecz ze względu na postać dostarczonych danych nie można było przeprowadzić sensownej próby bez sortowania danych.