Skład zespołu:
Mateusz Winnicki

WSI ĆWICZENIE 5

Cel ćwiczenia

Tematem piątych ćwiczeń są modele bayesowskie. Zadaniem do wykonania będzie zaimplementowanie naiwnego klasyfikatora bayesowskiego i zastosowanie go do zbadania zbioru danych zawierającego informacje o gatunków kosaćców.

Przebieg ćwiczenia

Ćwiczenie rozpoczniemy od zbadania wykresów przynależności do określonej klasy w zależności od wybranych cech, np. szerokości i długości płatka kosaćca.

We właściwym badaniu samego algorytmu naiwnego klasyfikatora bayesowskiego, będziemy sprawdzać sprawność modelu w zależności od podziału zestawu danych na zestaw trenujące i testowy oraz wyznaczymy macierze błędów dla poszczególnych przebiegów, ze szczególnym uwzględnieniem precyzji oraz czułości w zależności od klasy. Sprawdzimy również wpływ posortowania zbioru uczącego na efektywność modelu. Badane statystki będą prowadzone na zbiorach zawierających rezultaty 10 uruchomień klasyfikatora.

Aby uruchomić program należy to zrobić z parametrami określającymi: lokalizację pliku z badanym zestawem danych, udział zestawu uczącego (np. dla wartości 0,80 otrzymamy podział na zestawy uczący i testowy 80-20)

Dodanie do linii *-str* włączy opcję sortowania danych w zestawie uczącym.

Przykład:
> *py .\classifier.py -f data/iris.data -ts 0.8 -srt*

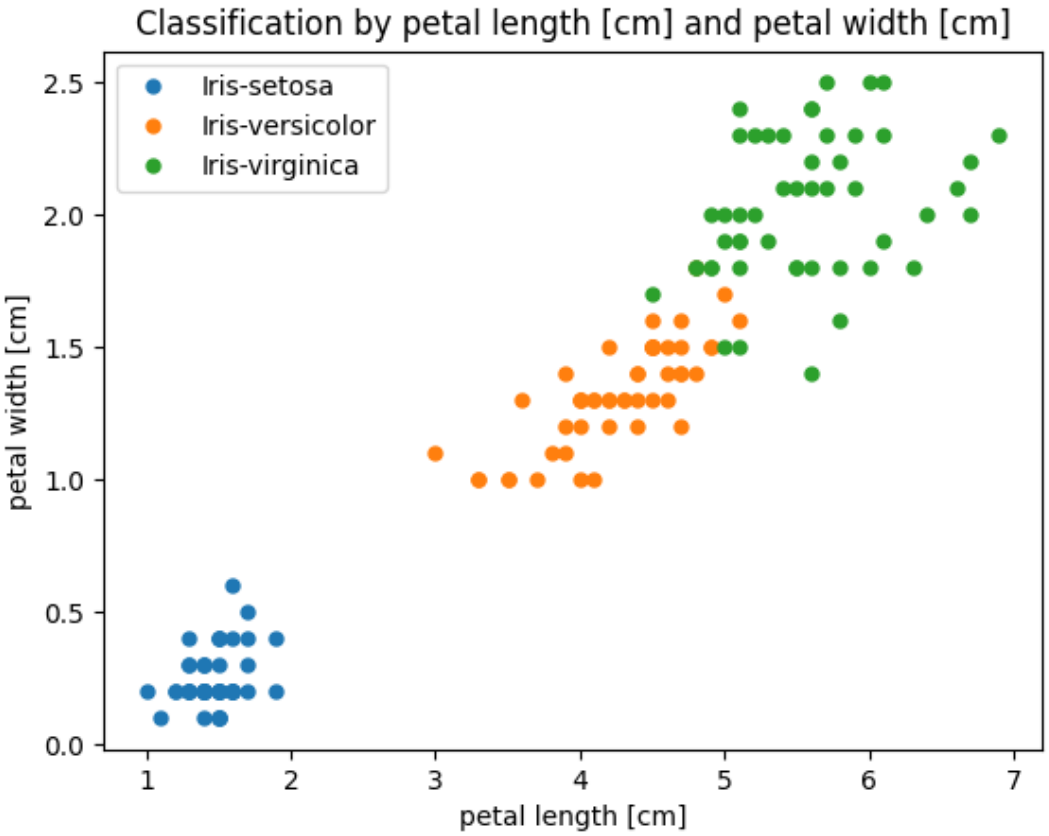
W pliku *classifier.py* w linii 200. znajduje się zakomentowane wywołanie funkcji tworzącej wykres klasyfikacji w zależności od podanych dwóch cech, można ją przetestować odkomentowując ją. Nowe wykresy pojawią się w folderze *plots*.

Moduły użyte w ćwiczeniu:

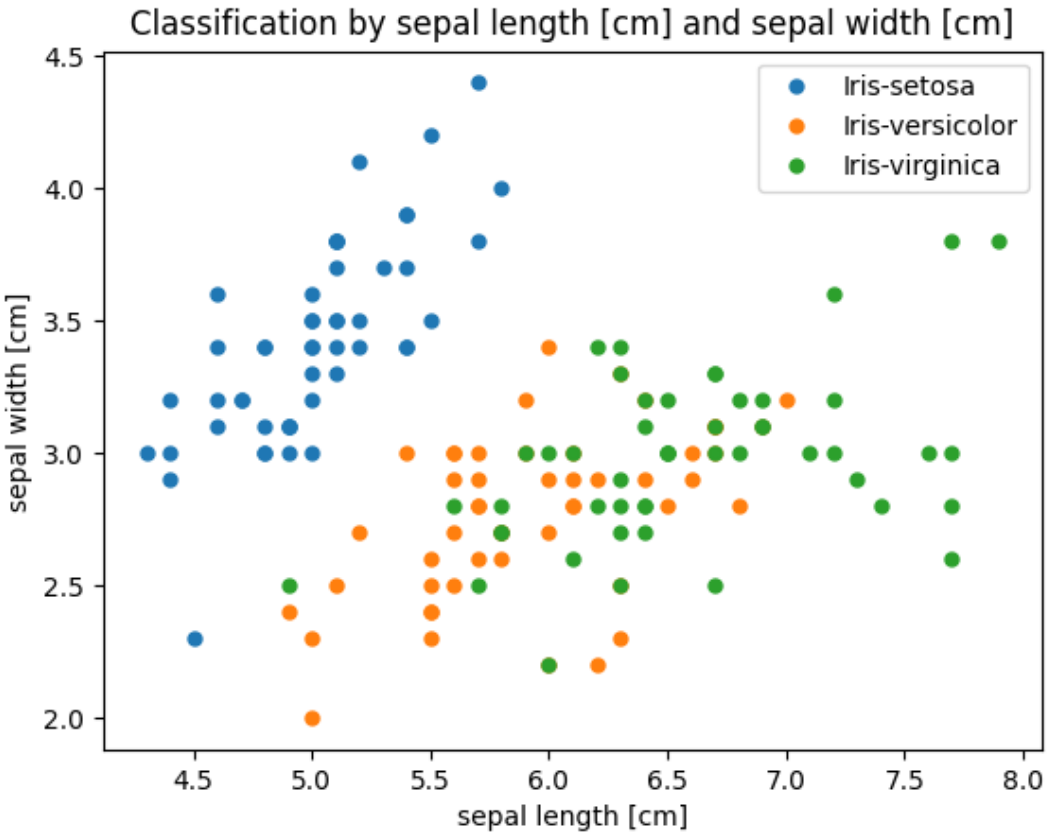
- *Numpy,*
- *Pandas,*
- *Matplotlib,*
- *Sklearn* (funkcja *train_test_split* do podziału zbioru na uczący i testowy),
- *Seaborn* (do ładnej prezentacji macierzy pomyłek w postaci heatmapy).

Wyniki ćwiczenia

Po wstępnej analizie badanego zbioru danych możemy odkryć kilka zależności, po których na oko jesteśmy w stanie określić wpływ danej cechy na klasyfikację gatunku kosaćca.



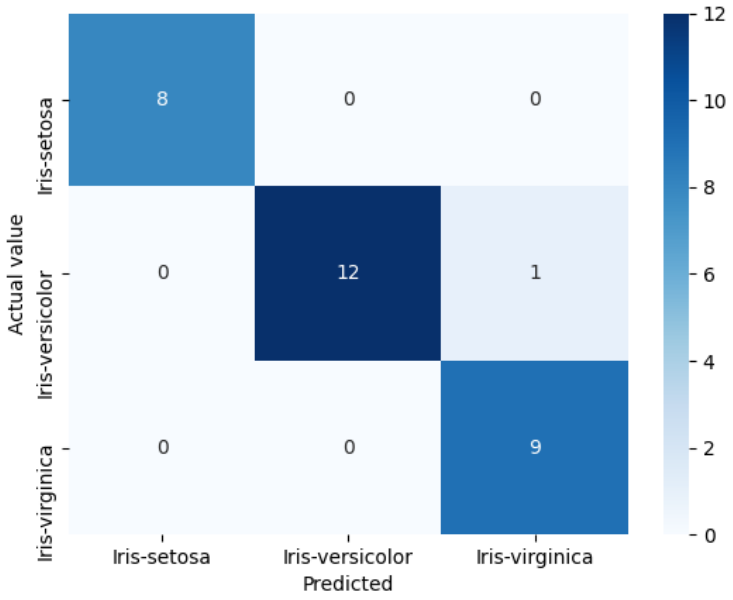
Na powyższym wykresie widać, że odmiana szczecinkowa charakteryzuje się zdecydowanie mniejszymi płatkami od odmiany różnobarwnej i wirginijskiej. Dla nowego kosaćca o wymiarach 1,5cm x 0,25cm z dużą dozą pewności możemy go zaklasyfikować jako odmianę szczecinkową, bez potrzeby wykorzystywania algorytmów uczenia maszynowego. Dwie pozostałe odmiany również mają swoje własne charakterystyczne rozmiary płatka, jednak ich klasyfikacja byłaby mniej pewna, ze względu na podobieństwo rozmiarów granicznych przypadków kosaćca różnobarwnego i wirginijskiego.



Również dla rozmiarów działki kielicha rozpoznanie odmiany szczecinkowej jest dosyć proste gołym okiem (choć niewątpliwie trudniejsze niż w przypadku rozmiarów płatka). W tym przypadku nie jesteśmy jednak w stanie rzetelnie określić, czy nowy okaz kosaćca o wymiarach długości działki kielicha z przedziału [5cm; 7cm] oraz szerokości z przedziału [2cm; 3,5cm] należy do odmiany różnobarwnej czy wirginijskiej.

Ze wstępnej analizy możemy wywnioskować, że informacje na temat rozmiarów płatka kosaćca mają większy wpływ w klasyfikacji nowego okazu od rozmiarów działki kielicha oraz że odmiana *iris-setosa* jest prostsza do zaklasyfikowania.

Pierwszy test działania zaimplementowanego klasyfikatora przeprowadzimy dla zbioru podzielonego na zbiór uczący i testowy w stosunku 80-20.

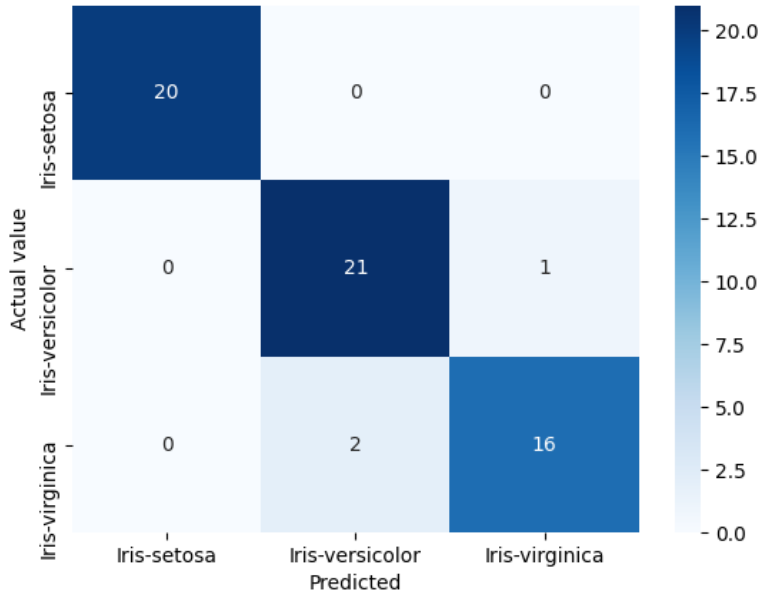


Rysunek 1 Tablica pomyłek dla podziału 80-20

	Dokładność klasyfikatora	**Precyzja**			**Czułość**		
		Satosa	***Versicolor***	***Virginica***	***Satosa***	***Versicolor***	***Virginica***
Min	0,96	1	0,88	0,88	1	0,86	0,86
Max	1	1	1	1	1	1	1
Średnia	0,99	1	0,96	0,98	1	0,97	0,96
Wariancja	0,01	0	0,06	0,04	0	0,05	0,05

Otrzymaliśmy bardzo wysoką średnią dokładność klasyfikatora. Precyzja i czułość w każdym z przebiegów były maksymalne w przypadku gatunku *satosa* oraz nieco niższe dla dwóch pozostałych.

Po zmniejszeniu rozmiaru zestawu uczącego do 60% początkowego zestawu danych, otrzymujemy następujące wyniki.



Rysunek 2 Tablica pomyłek dla podziału 60-40

	Dokładność klasyfikatora	**Precyzja**			**Czułość**		
		Satosa	***Versicolor***	***Virginica***	***Satosa***	***Versicolor***	***Virginica***
Min	0,92	1	0,88	0,66	1	0,74	0,88
Max	1	1	1	1	1	1	1
Średnia	0,97	1	0,95	0,91	1	0,92	0,94
Wariancja	0,02	0	0,04	0,09	0	0,07	0,05

Zmniejszenie zbioru trenującego poskutkowało nieznacznym spadkiem ogólnej dokładności klasyfikatora oraz precyzji i czułości klas *versicolor* i *virginica*. Metryki te nadal pozostały jednak maksymalne dla *satosy*.

Kolejne badanie przeprowadzimy dla zbioru trenującego równego 40% rozmiaru danego zbioru danych.

	Dokładność klasyfikatora	**Precyzja**			**Czułość**		
		Satosa	***Versicolor***	***Virginica***	***Satosa***	***Versicolor***	***Virginica***
Min	0,92	1	0,88	0,81	1	0,81	0,86
Max	0.98	1	96	1	1	1	0,96
Średnia	0,95	1	0,91	0,93	1	0,93	0,91
Wariancja	0,02	0	0,02	0,05	0	0,05	0,03

Jak widać niewielka przewaga rozmiaru zbioru testującego nad uczącym nie wpłynęła drastycznie na wyniki. Statystyki gatunku *satosa* nadal nieskazitelne.

Zmniejszmy rozmiar zbioru uczącego do zaledwie 10% zadanego zbioru danych.

	Dokładność klasyfikatora	**Precyzja**			**Czułość**		
		Satosa	***Versicolor***	***Virginica***	***Satosa***	***Versicolor***	***Virginica***
Min	0,75	1	0,77	0,47	0,08	0,61	0,69
Max	0,97	1	1	1	1	1	1
Średnia	0,92	1	0,91	0,81	0,89	0,84	0,91
Wariancja	0,06	0	0,07	0,15	0,27	0,13	0,08

Trenowanie nawet na tak mały zbiorze uczącym daje nam naprawdę zadowalające wyniki. Ponownie najwyższa precyzja przypadła klasie *satosa*, a najniższa klasie *virginica*. W kategorii czułości doszło do przetasowania, gatunek *satosa* pogorszył swoje wyniki. Jak widać z minimalnego wyniku i bardzo wysokiej wariancji, możemy się domyślić, że wynika to najprawdopodobniej z losowego przypadku, w którym być prawdopodobnie zbiór uczący (który zawierał jedynie 15 egzemplarzy) mógł cierpieć na niedobór informacji o kosaćcu szczecinkowym.

Na koniec zbadajmy jak posortowanie zbioru trenującego wpłynie na działanie klasyfikatora. Sprawdzimy to na podziale z najgorszymi wynikami, czyli 10-90.

	Dokładność klasyfikatora	**Precyzja**			**Czułość**		
		Satosa	***Versicolor***	***Virginica***	***Satosa***	***Versicolor***	***Virginica***
Min	0,85	1	0,59	0,68	0,87	0,58	0,54
Max	0,99	1	0,96	1	1	1	1
Średnia	0,94	1	0,84	0,91	0,98	0,89	0,94
Wariancja	0,03	0	0,11	0,09	0,04	0,12	0,16

Posortowanie zbioru uczącego spowodowało nieznaczny wzrost średniej dokładności klasyfikatora.

Na koniec sprawdźmy skrajną przewagę rozmiaru zestawu trenującego, w stosunku 90-10 ze zbiorem testowym.

	Dokładność klasyfikatora	**Precyzja**			**Czułość**		
		Satosa	***Versicolor***	***Virginica***	***Satosa***	***Versicolor***	***Virginica***
Min	0,82	1	0,75	0,25	1	0,5	0,5
Max	1	1	1	1	1	1	1
Średnia	0,95	1	0,93	0,87	1	0,91	0,89
Wariancja	0,04	0	0,09	0,22	0	0,15	0,16

Statystki są niższe niż w większości badanych przypadków, jednak ciężko tu mówić o przetrenowaniu modelu. Wyniki są zaniżane przez występowanie skrajnych wyników minimalnych dla gatunków *versicolor* oraz *virginica*.

Podsumowanie

Zgodnie z początkowymi przewidywaniami, kosaciec szczecinkowy był zdecydowanie prostszy do klasyfikacji od pozostałych dwóch gatunków. Zestaw danych na temat gatunków był na tyle zdywersyfikowany, że algorytm bardzo dobrze radził sobie ucząc się nawet na małych zbiorach. Oczywiście mniejszy zestaw treningowy skutkował spadkiem jakości klasyfikatora, jednak mimo to otrzymywaliśmy skuteczność powyżej 90%. Wpływu posortowania zestawu uczącego na efektywność modelu nie dostrzegłem.