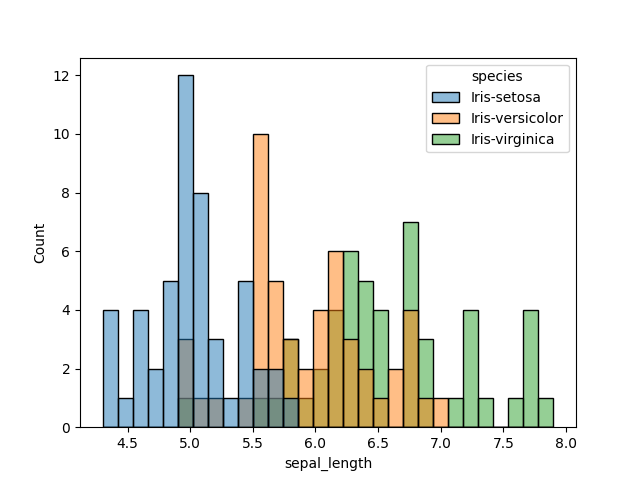
**Sprawozdanie Dane Iris i algorytm Kmeans**

**Wykonał: Jakub Marciniak**

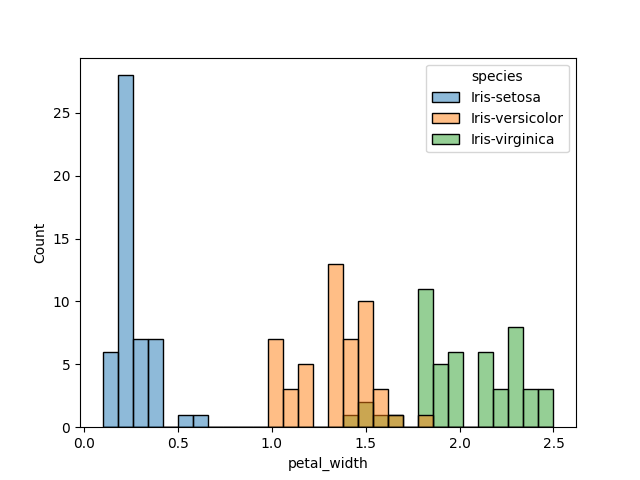
**Nr albumu: 389081**

**Histogramy**

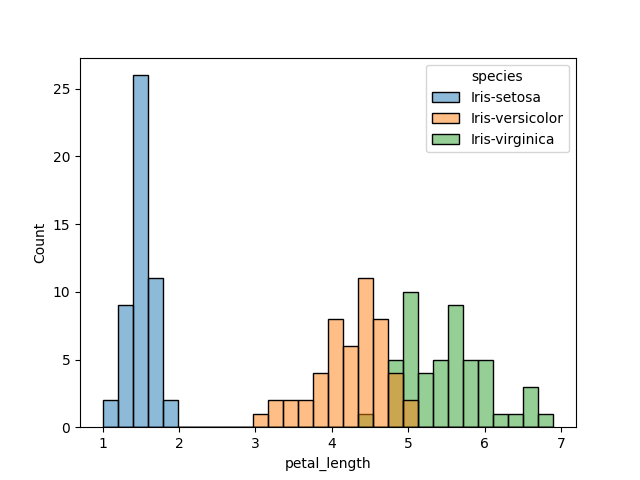
Na poniższym histogramie pokazany jest sepal\_length(długość kielicha), który jest pogrupowany po species. Możemy zauważyć, że różne dane nachodzą na siebie co świadczy o tym że jest ciężej jest rozróżnić dany gatunek za pomocą sepal\_length. Widać również, że najwięcej danych nachodzi na siebie z klas Iris-versicolor oraz Iris-Virginica, Iris-setosa ma duża ilość danych, które to nie nachodzą na inne klasy.



Na poniższym wykresie widnieje petal\_length, który został pogrupowany przy pomocy species.Widać tu jak Iris-setosa jest oddalona od pozostałych gatunków i w jakim przedziale jest jej najwięcej, jest to bardzo wąski zbiór, natomiast gatunki Iris-versicolor oraz Iris-virginica nachodzą się na siebie i może tu dochodzić do lekkich błędów w grupowaniu danych.



Poniższy histogram przedstawia petal\_length. Jest na nim też bardzo dobrze widoczne, że dane Iris-setosa jest oddzielona od reszty gatunków tak samo jak w przypadku z petal\_width, natomiast Iris-versicolor i Iris-virginica znowu na siebie zachodzą jednak nie tak bardzo jak w przypadku sepal\_length gdzie nakładanie na siebie wszytkich gatunków było bardzo duże.

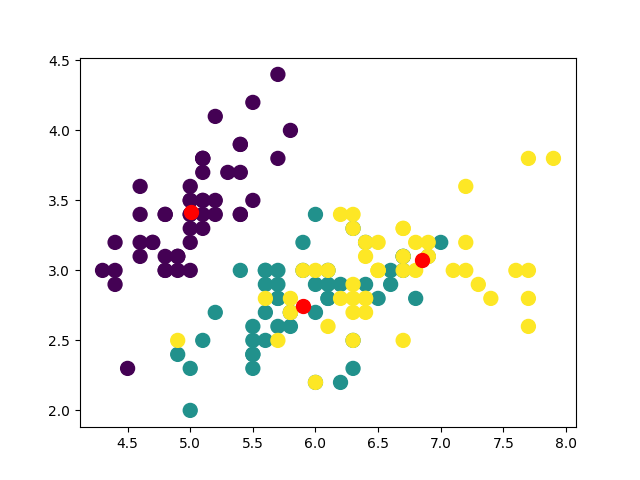


* Z powyższych histogramów wywnioskować można, że petal\_length oraz petal\_width są najlepsze żeby móc zidentyfikować dany gatunek kwiatu.
* Widać też dobrze, że gatunek Iris-setosa jest bardzo łatwy do oddzielenia, natomiast Iris-versicolor oraz Iris-virginica zachodzą na siebie co utrudnia poprawne rozgrupowanie ich

Wykresy 2d

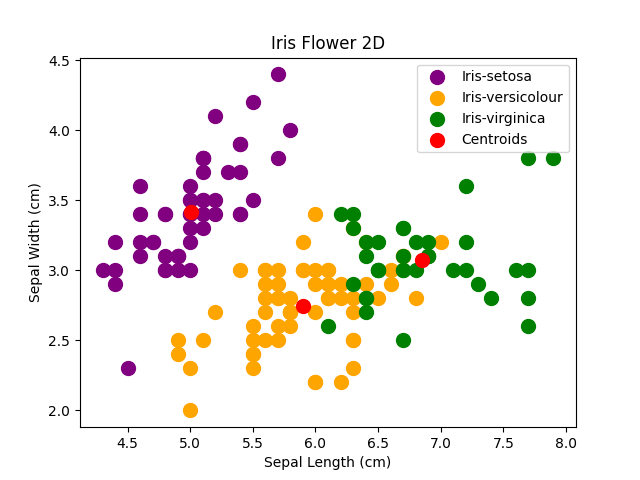
Mamy tu dwa wykresy 2d pokazujące jak wyglądają dane iris , jednakże na pierwszym obrazku widać że dane są pogrupowane w taki sposób jaki jest w bazie. Mamy bardzo dużo kropeczek żółtych, które świadczą że jest to Iris-virginica oraz turkusowy który odpowiedzialny jest za gatunek Iris-versicolor. Bardzo duża ilość nakłada się na siebie co utrudnia poprawną klasyfikacje kwiatków do odpowiedniego gatunku. Na obu obrazkach można zauważyć, że Iris-Setosa jest odległym zbiorem od pozostałych i klasteryzacja jej jest jak i w przypadku Kmeans jak i bez niego w 100% dobrze i nie wpływa na pozostałe utrudniając im klastrowanie.

Bez Kmeans



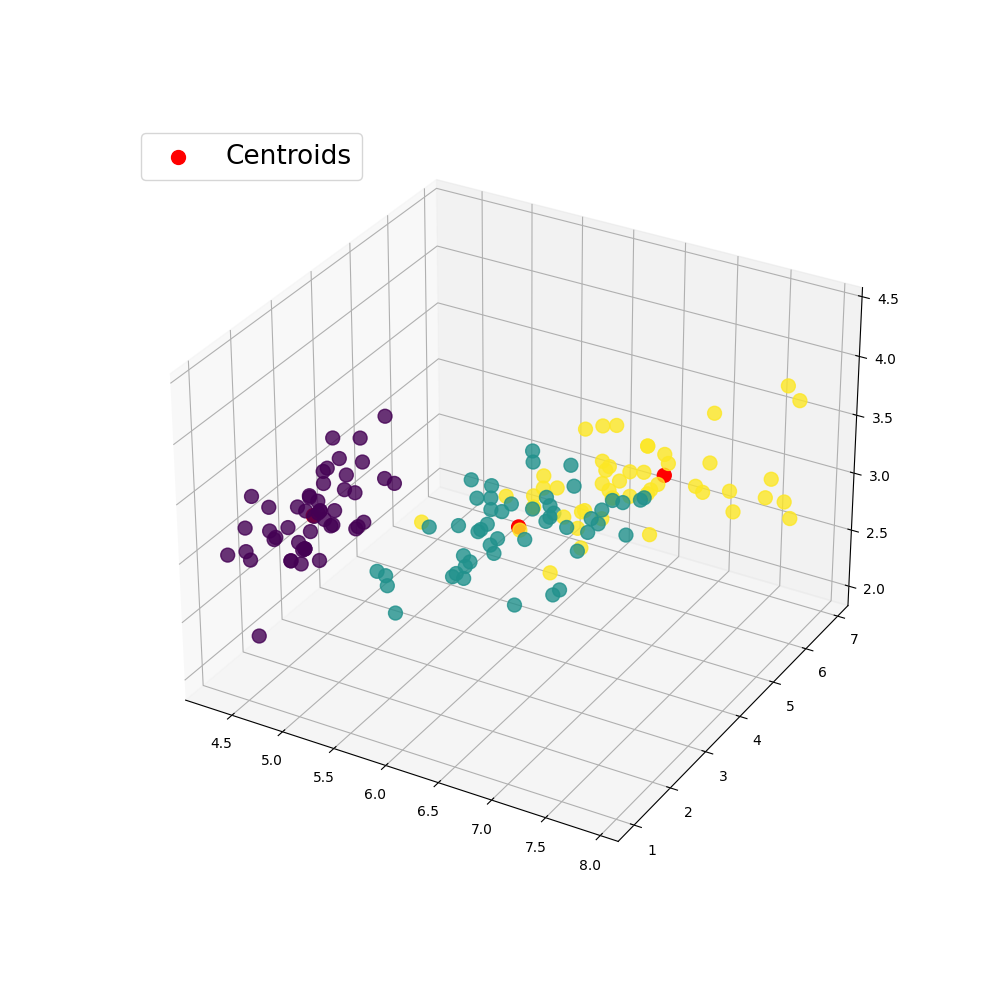
Drugi obrazek jest przedstawiony z Algorytmem Kmeans, na którym możemy zaobserwować w jaki sposób zostały przypisane wartości odpowiednich gatunków. Dzięki klasteryzacji gatunki są o wiele lepiej przypisane do odpowiednich grup. Przez co możemy łatwiej rozpoznać konkretny gatunek kwiatu. W porównaniu do wykresu powyżej możemy zaobserwować, że zaszło kilka różnic pomiędzy danymi które nie posiadały algorytmu Kmeans jak i tymi które ten algorytm posiadają. Dane stały się o wiele bardziej czytelne i łatwiejsze do klasyfikowania.

Z Kmeans



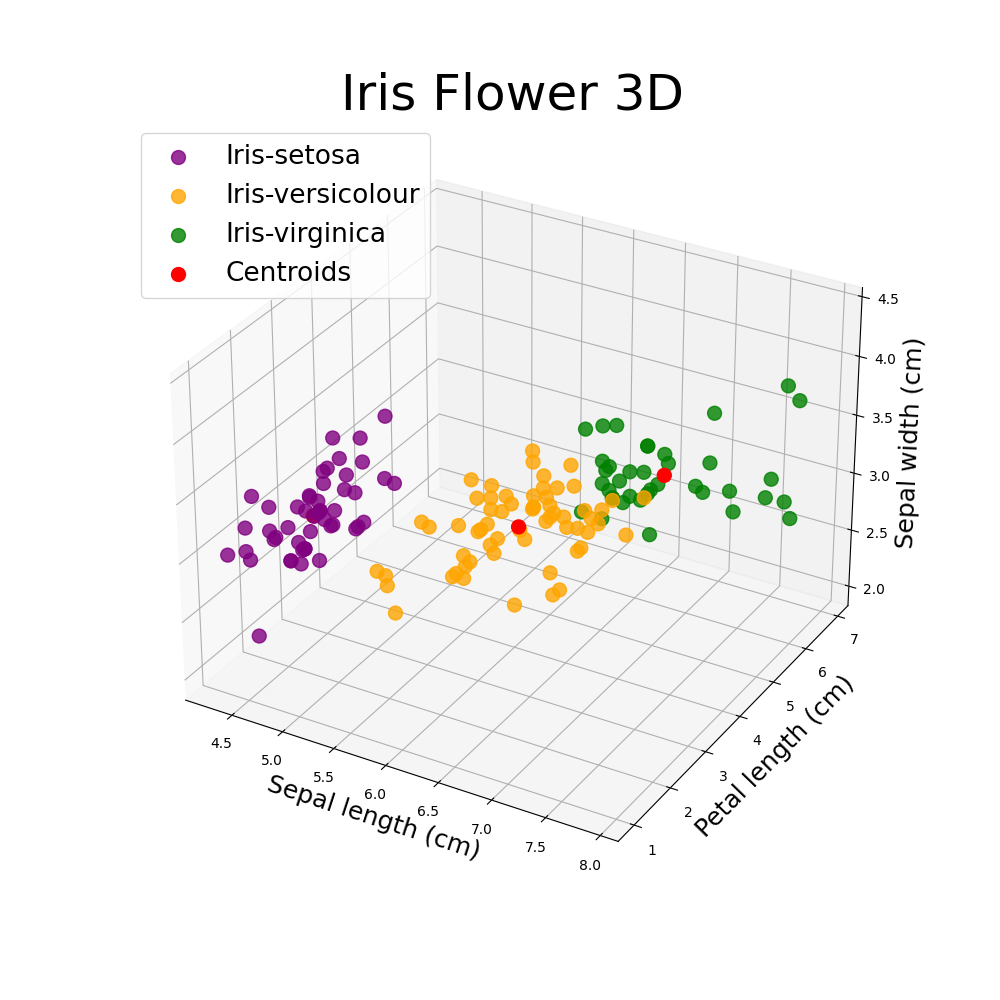
Wykresy 3D

Tym razem mamy wykres, który jest wykresem 3D. Można na nim zaobserwować dokładniejsze położenie każdego z gatunków. Widać dokładniej jak nakładają się różne dane Iris-virginica na Iris-versicolor. Widać też bardzo dokładnie, że zbiór gatunków Iris-setosa jest zbiorem odosobnionym od innych i jest on dobrze przydzielany jak widać to z tabel Conffusion matrix.  
**Bez KMeans**



Na wykresie mamy przedstawiony wykres 3D który został sklasteryzowany przez algorytm KMeans. Dzięki temu widać owiele lepsze grupowanie danych w stosunku do wykresu, który tego KMeans nie posiadał. Widać, że dane przynależą do odpowidnich każdej grupie środków(klastrów). Jest owiele lepsze dopasowanie gatunków Iris-virginica i Iris-versicolor do swoich środków natomiast tak jak zawsze Iris-setosa jest oddalona od pozostałych i w tym przypadku, zostaje ona sklasteryzowana bez problemów. Możemy zauważyć, że w miejscu gdzie jest czerwone kółeczko i tak jest ciężko zaklasyfikować dane do odpowiedniej grupy gdyż nakładają się na siebie w dużej ilości.

**Z Kmeans**



**Krótka legenda** classification\_report

**Precision – precyzja może być postrzegana jako miara dokładności klasyfikatora. Jest to stosunek prawdziwie pozytywnych wyników do sumy prawdziwie pozytywnych i fałszywych alarmów. Dla wszystkich przypadków sklasyfikowanych jako pozytywne, jaki procent był poprawny.**

**Recall** - zdolność klasyfikatora do prawidłowego znajdowania wszystkich pozytywnych wystąpień. Dla wszystkich przypadków, które były rzeczywiście pozytywne, jaki procent został prawidłowo sklasyfikowany.

**f1-score** – jest to ważona średnia harmoniczna precision oraz recall. Generalnie mówiąc f1-score jest mniejszy niż accuracy\_score, ponieważ zawiera on w swoich obliczeniach precision i recall. Do porównywania klasyfikatorów (w naszym przypadku species) należy używać średniej ważonej, a nie globalnej dokładności.

**Support** - to liczba rzeczywistych wystąpień klasy w określonym zbiorze danych.

**Confusion matrix dla 100 % zbioru – 150 kwiatki**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| Iris-setosa | 1.00 | 1.00 | 1.00 | 50 |
| Iris-versicolor | 0.77 | 0.96 | 0.86 | 50 |
| Iris-virginica | 0.95 | 0.72 | 0.82 | 50 |

całkowita dokładność zbioru wynosi – accuracy = 0.89%

Widać tu jak precision gatunku Iris-versicolor jest słaba w porównaniu do pozostałych możemy wywnioskować z tego, że bardzo duża ilość kwiatków została fałszywie sklasyfikowana, oraz recall wynosi 96%, stąd można zauważyć, że 96% z wszystkich pozytywnych wystąpień była zaklasyfikowana prawidłowo. Dla gatunku Iris-virginica można zaobserwować słaby recall oznacza to że dużo zostało zaklasyfikowane do precision aż 95% natomiast tylko 72% zostało dobrze zaklasyfikowanych. Jak to było widać z wykresów powyżej, Iris-setosa była oddaloną grupą której klasyfikacja nie stanowiła problemu, stąd właśnie precision, recall oraz f1-score na poziomie 100%. Ogólna dokładności dla zbioru to 89%

**Confusion matrix dla 90 % zbioru – 135 kwiatki**

confusion\_matrix

[[42 0 0]

[ 0 44 2]

[ 0 4 43]]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Predicted | | |  |
|  |  | Iris-setosa | Iris-versicolor | Iris-virginica | Total |
| Actual | Iris-setosa | 42 | 0 | 0 | 42 |
| Iris-versicolor | 0 | 44 | 2 | 46 |
| Iris-virginica | 0 | 4 | 43 | 47 |
|  | Predicted total | 42 | 48 | 45 | 135 |

Z podanej tabeli można wydedukować,

Z tabeli możemy wywnioskować, że ze 135 różnych kwiatków, 6 zostało źle sklasyfikowanych. Źle sklasyfikowane zostały Iris-versicolor na Iris-virginica w ilości 4 sztuk oraz Iris-virginica na Iris-versicolor w ilościach 2 sztuk. Natomiast Iris-setosa jest sklasyfikowana bardzo dobrze bez żadnych problemów i nieścisłości. Można to zaobserwować bardzo dobrze z wykresów 2D jak i 3D, które są wyżej.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| Iris-setosa | 1.00 | 1.00 | 1.00 | 42 |
| Iris-versicolor | 0.92 | 0.96 | 0.94 | 46 |
| Iris-virginica | 0.96 | 0.91 | 0.93 | 47 |

Z tabeli classification\_report dla 90% zbioru można zaobserwować, że Iris-setosa ma 100% dokładności, natomiast reszta gatunków już ma dokładność na gorszym poziomie. Widać tu że Iris-versicolor ma precision na poziomie 92% co znaczy, że mała część kwiatów została źle zaklasyfikowana natomiast recall na poziomie 96% świadczy, że bardzo dużą część została poprawnie zaklasyfikowana . Iris-virginica ma natomiast precision na poziomie 96% i widać dzięki recall na poziomie 91%, że to właśnie tylko tyle kwiatków było poprawnie zaklasyfikowanych.

całkowita dokładność zbioru wynosi - accuracy\_score = 0.9555555555555556

**Confusion matrix dla 75 % zbioru – 113 kwiatki**

confusion\_matrix

[[32 0 0]

[ 0 39 2]

[ 0 4 36]]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Predicted | | |  |
|  |  | Iris-setosa | Iris-versicolor | Iris-virginica | Total |
| Actual | Iris-setosa | 32 | 0 | 0 | 32 |
| Iris-versicolor | 0 | 39 | 2 | 41 |
| Iris-virginica | 0 | 4 | 36 | 40 |
|  | Predicted total | 32 | 43 | 38 | 113 |

Jak widać w tabeli dane w niej przedstawione składają się z 75% całego zbioru i mamy tu widoczne powtórzenie źle zaklasyfikowanych gatunków tak jak na wykresie powyżej.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| Iris-setosa | 1.00 | 1.00 | 1.00 | 32 |
| Iris-versicolor | 0.91 | 0.95 | 0.93 | 41 |
| Iris-virginica | 0.95 | 0.90 | 0.92 | 40 |

Podobna Sytuacja miała miejsce dla 90%. Wykres wyżej

całkowita dokładność zbioru wynosi - accuracy\_score = 0.9469026548672567

**Confusion matrix dla 50 % zbioru – 75 kwiatki**

confusion\_matrix

[[22 0 0]

[ 0 24 1]

[ 0 6 22]]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Predicted | | |  |
|  |  | Iris-setosa | Iris-versicolor | Iris-virginica | Total |
| Actual | Iris-setosa | 22 | 0 | 0 | 22 |
| Iris-versicolor | 0 | 24 | 1 | 25 |
| Iris-virginica | 0 | 6 | 22 | 28 |
|  | Predicted total | 22 | 30 | 23 | 75 |

Z tabeli widać że mamy tylko 24 Iris-versicolor natomiast przewidzianych zostało aż o 6 więcej, zostały one zaklasyfikowane błędnie jako Iris-versicolor, jednak powinna to być Iris-virginica. Za to Iris-virginica została tylko 1 raz błędnie zaklasyfikowana jako Iris-versicolor. O Iris-setosa nie ma co nawet mówić bo zawsze jest klasyfikowana na 100%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | Support |
| Iris-setosa | 1.00 | 1.00 | 1.00 | 22 |
| Iris-versicolor | 0.80 | 0.96 | 0.87 | 25 |
| Iris-virginica | 0.96 | 0.79 | 0.86 | 28 |

W Iris\_versicolor mamy precision na poziomie 80% ponieważ aż 6 versiolor zostało zaklasyfikowanych do virginica, natomiast recall jest na poziomie 96% więc aż 96% pozytywnych odpowiedzi została dobrze zaklasyfikowana. Dla Iris-virginica mamy 96% precision więc widać, że mamy bardzo dużo zgłoszeń czy to prawdziwych czy fałszywych. Teraz jednak spójrzmy na recall z którego wynika że mamy tylko 79% poprawnych zgłoszeń.

całkowita dokładność zbioru wynosi - accuracy\_score = 0.9066666666666666

**Confusion matrix dla 25 % zbioru – 38 kwiatki**

confusion\_matrix(y\_test, y\_predict)

[[10 0 0]

[ 0 12 1]

[ 0 1 14]]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Predicted | | |  |
|  |  | Iris-setosa | Iris-versicolor | Iris-virginica | Total |
| Actual | Iris-setosa | 10 | 0 | 0 | 10 |
| Iris-versicolor | 0 | 12 | 1 | 13 |
| Iris-virginica | 0 | 1 | 14 | 15 |
|  | Predicted total | 10 | 13 | 15 | 38 |

W tabeli z 25% całego zbioru mamy do czynienia ze złym dopasowaniem Iris-virginica na Iris-versicolor w ilości 1 sztuki oraz na odwrót Iris-versicolor na Iris-virginica również w ilości 1 sztuki. Przez co nie udało uzyskać się 100% dopasowania jednak dzięki wymianie 1 za 1 w tabeli poniżej precision, recall oraz f1-score dla Iris-versicolor wynosi 92% i dla Iris-virginica wynosi 93% dopasowania.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | Support |
| Iris-setosa | 1.00 | 1.00 | 1.00 | 10 |
| Iris-versicolor | 0.92 | 0.92 | 0.92 | 13 |
| Iris-virginica | 0.93 | 0.93 | 0.93 | 15 |

całkowita dokładność zbioru wynosi - accuracy\_score = 0.9473684210526315

**Confusion matrix dla 10 % zbioru – 15 kwiatków**

confusion\_matrix

[[7 0 0]

[0 5 0]

[0 0 3]]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Predicted | | |  |
|  |  | Iris-setosa | Iris-versicolor | Iris-virginica | Total |
| Actual | Iris-setosa | 7 | 0 | 0 | 7 |
| Iris-versicolor | 0 | 5 | 0 | 5 |
| Iris-virginica | 0 | 0 | 3 | 3 |
|  | Predicted total | 7 | 5 | 3 | 15 |

Z tabeli wynika, że wszystkie gatunki (species) zostały zaklasyfikowane bez żadnych pomyłek, żadna klasa nie została błędnie zaklasyfikowana do innej klasy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | Support |
| Iris-setosa | 1.00 | 1.00 | 1.00 | 7 |
| Iris-versicolor | 1.00 | 1.00 | 1.00 | 5 |
| Iris-virginica | 1.00 | 1.00 | 1.00 | 3 |

Wszystkie klasy zostały zaklasyfikowane bez żadnych błędów. Co na tak małej ilości danych nie jest aż tak trudne do uzyskania.

całkowita dokładność zbioru wynosi - accuracy\_score = 1.0