Raport z klasyfikacji irysów

Maria Galanty - kandydatka na stanowisko Junior Data Scientist

Luty 2020

1 Wstep

Zadanie polegało na zbudowaniu modelu, który sklasyfikuje gatunki kwiatów. W przedstawionym rozwiazaniu wykorzystałam Pythona.

2 Eksploracja danych

Dane zawierały 150 obserwacji oraz pieć zmiennych: długość, szerokość płatków i listków oraz gatunek irysów. Każda klasa liczyła równo 50 okazów, jest to bardzo dobrze zbalansowany zbiór danych. Wystapiła jedna brakujaca zmienna, która została uzupełniona przy pomocy średniej wartości z najbardziej zbliżonych obserwacji. W jednym z rekordów wystapiła wartość '2,2', która najprawdopodobniej była literówka i została poprawiona na '2.2'.

3 Klasyfikator

Przy budowie modelu wykorzystałam drzewo decyzyjne. Ze wzgledu na bardzo dobrze zbalansowany zbiór danych, do ewaluacji modelu użyłam cross-walidacji i metryki *accuracy* - suma wszystkich dobrze sklasyfikowanych obserwacji podzielona przez wszystkie obserwacje w zbiorze treningowym/testowym. Przy cross-walidacji (umożliwiajacej sprawdzenie, czy model nie jest przeuczony) podzieliłam zbiór na 5 cześci, co pozwoliło mi trenować model na 4 z nich i testować na pozostałej. Uzyskałam średnia testowa precyzje modelu wynoszaca 0.96.