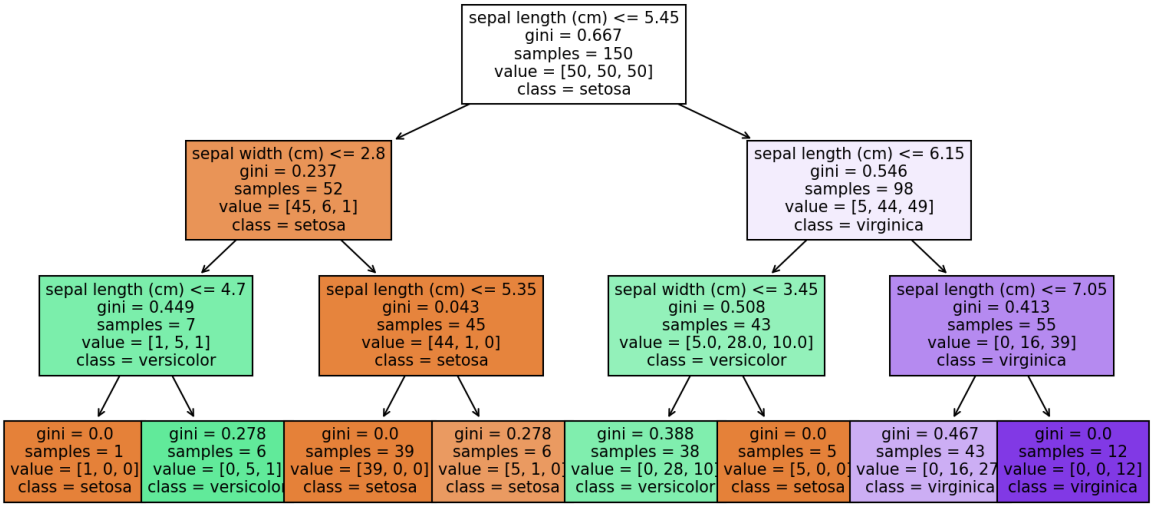
Drzewo Decyzyjne dla Zestawu Danych Iris z wykorzystaniem współczynnika Gini

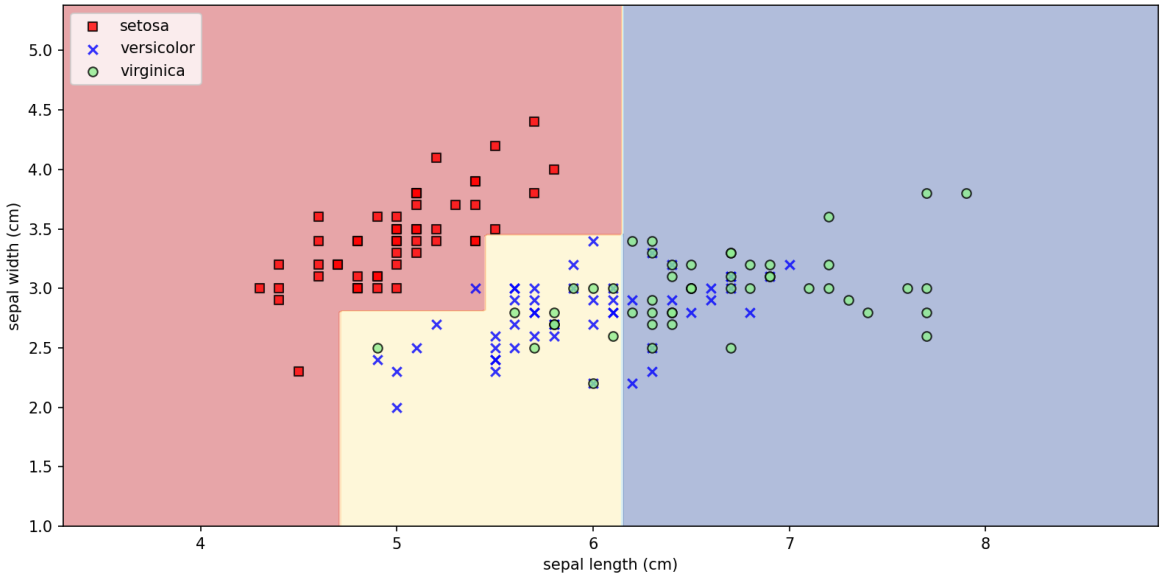
Wprowadzenie

W ramach zadania wykonano klasyfikację zbioru danych Iris przy użyciu modelu drzewa decyzyjnego. Do klasyfikacji użyto dwóch cech: szerokości i długości działki kielicha (sepal width i sepal length).



Wizualizacja Drzewa Decyzyjnego

Pierwsze drzewo decyzyjne stworzone dla zbioru danych Iris, wykorzystując współczynnik Gini, rozpoczyna się od korzenia z nieczystością Gini=0.667. Ta wartość wskazuje na równomierny rozkład klas na początku. W kolejnych węzłach obserwujemy spadek współczynnika Gini, co oznacza zwiększającą się czystość podziałów i lepszą separację klas. Struktura drzewa ukazuje, jak decyzje są podjęte na podstawie szerokości i długości działki kielicha (sepal width i sepal length), co prowadzi do efektywnej klasyfikacji klasy setosa, a mniej skutecznej dla versicolor i virginica.

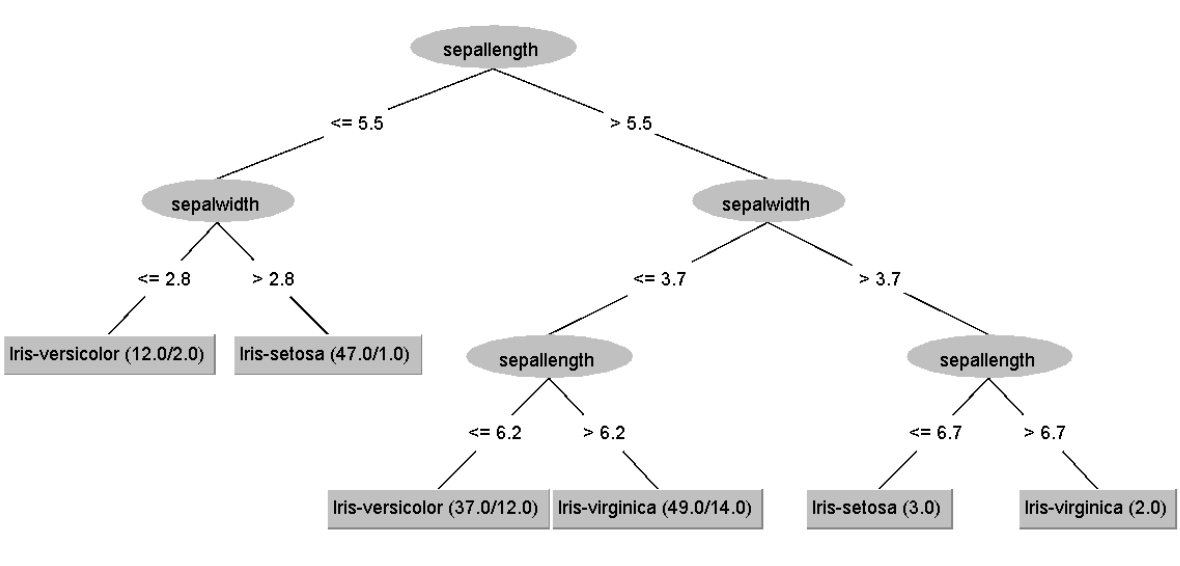


Granice Decyzyjne

Powyższy schemat ilustruje granice decyzyjne utworzone przez model. Kolory odpowiadają różnym klasom irysów. Punkty danych są oznaczone różnymi symbolami i kolorami w zależności od ich rzeczywistej klasy. Obszar niebieski odpowiada klasie setosa, żółty - versicolor, a czerwony - virginica.

Na schemacie widać, że klasa setosa jest dobrze odseparowana od pozostałych klas, co odzwierciedla węzły z Gini równym 0 w drzewie decyzyjnym. Granice dla klasy versicolor i virginica nakładają się na siebie, co wskazuje na pewne trudności w rozdzieleniu tych dwóch klas przez model. Widać to także na schemacie drzewa, gdzie Gini dla podziałów związanych z tymi klasami jest większy od zera.

WEKA



Wizualizacja drzewa

W programie WEKA użyto algorytmu J48, który jest implementacją C4.5, do klasyfikacji użytego wcześniej zbioru Iris. Drzewo decyzyjne ma maksymalną głębokość ustawioną na 3, co utrzymuje model prostym i wydajnym. Minimalna liczba obiektów w liściu (min num of obj) wynosi 2, co zapobiega zbyt szczegółowemu podziałowi danych. Proces walidacji, wykorzystujący 3 podziały (num of folds), służy do oceny jakości modelu, utrzymując generalizację podziału przy zachowaniu precyzji klasyfikacji różnych klas irysów.

Wnioski

Analiza modeli drzew decyzyjnych pokazała, że oba modele (python sklearn, WEKA) były skuteczne w identyfikacji klasy Iris setosa, co potwierdza niski współczynnik Gini (zbliżony do 0) oraz brak błędów klasyfikacji dla tej klasy w obu modelach (niepoprawnie zakwalifikowano pojedyncze irysy). Klasa iris setosa była łatwo oddzielana od pozostałych.

Jednakże, oba modele napotkały trudności w dokładnej klasyfikacji dwóch pozostałych klas: Iris versicolor i Iris virginica. Te klasy często były mylone ze sobą, co było widoczne w większych wartościach współczynnika Gini oraz liczbie błędów klasyfikacji w odpowiednich węzłach obu drzew. Drzewo z WEKA, mimo ograniczonej głębokości, również pokazało problem z błędami klasyfikacji dla tych dwóch klas (np. Iris-versicolor (48/2) oznacza, że 2 przypadki zostały niewłaściwie sklasyfikowane).