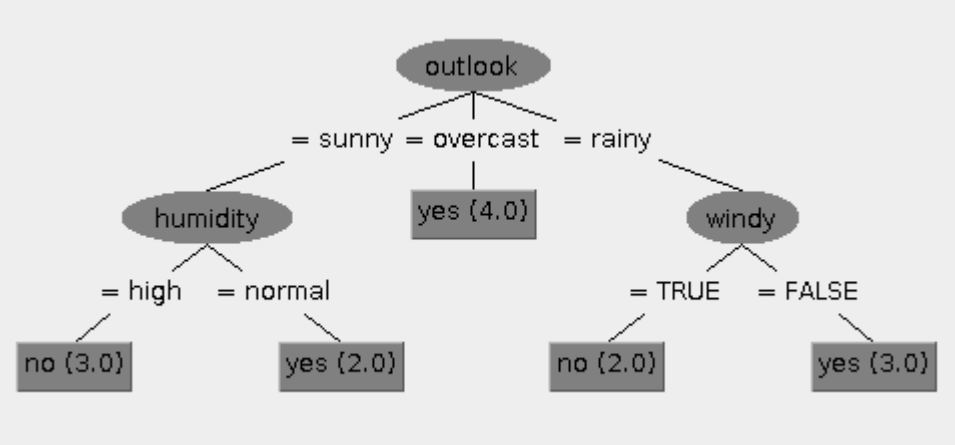
Drzewa decyzyjne



Drzewo decyzyjne to graficzna metoda wspomagania procesu decyzyjnego, stosowana w teorii decyzji. Algorytm drzew decyzyjnych jest również stosowany w uczeniu maszynowym do pozyskiwania wiedzy na podstawie przykładów.

Zad.1 Jak Twoim zdaniem wyglądałoby drzewo decyzyjne dla zestawu danych poniżej (podejmujemy decyzję Enjoy = yes/no na podstawie pozostałych parametrów)? Narysuj je na kartce.

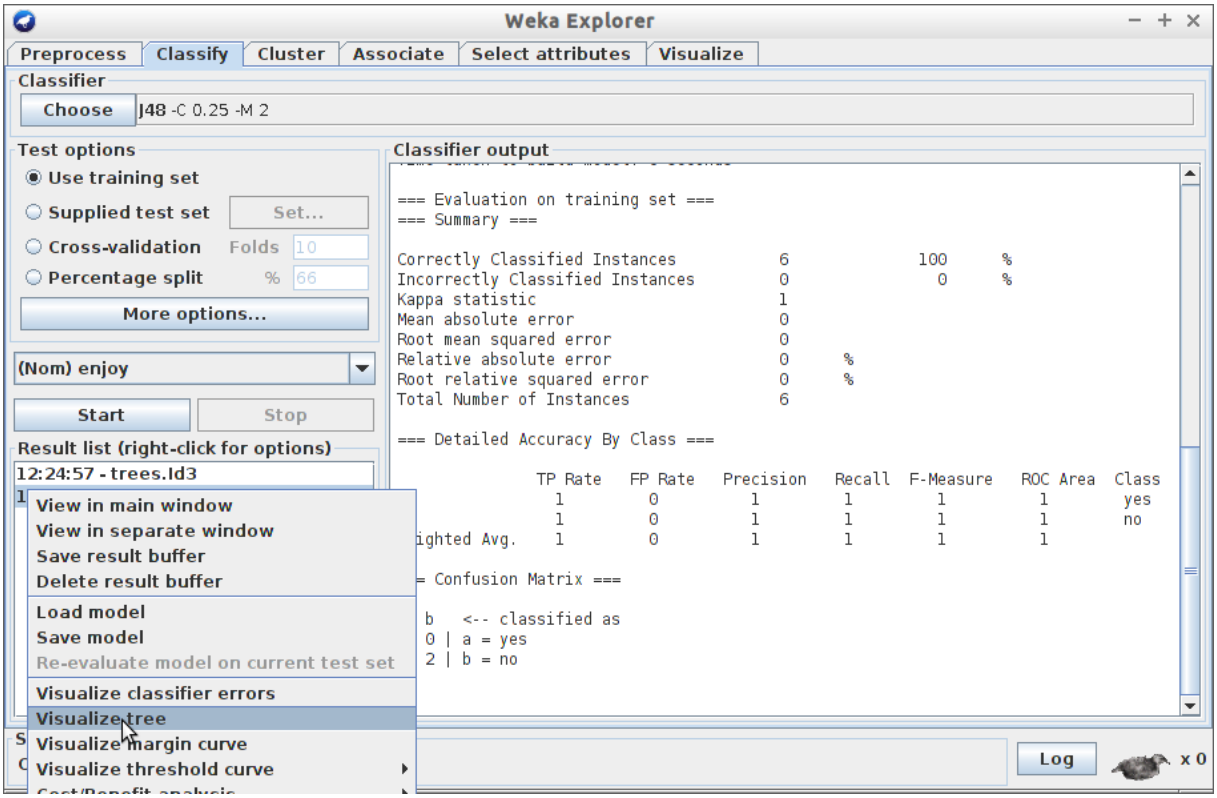
Sky	AirTemp	Humidity	Wind	Water	Forecast	Enjoy
sunny	warm	normal	strong	warm	same	yes
sunny	warm	high	strong	warm	same	yes
rainy	cold	high	strong	warm	change	no
sunny	warm	high	strong	cool	change	yes
cloudy	warm	normal	weak	warm	same	yes
cloudy	cold	high	weak	cool	same	no

1. Jaki jest rozmiar zbioru uczącego?
2. Ile atrybutów występuje w zbiorze uczącym?
3. Ile jest instancji jest pozytywnych (*Enjoy=yes*) a ile negatywnych?
4. Który z atrybutów najlepiej rozdziela dane? ;)

Zad2.

1. Otwórz pakiet WEKA.
2. Wczytaj plik `swimming.arff` ze zbioru danych.
3. Przeanalizuj dane za pomocą WEKI.
4. Kliknij w zakładkę **Clasify**.
5. Za pomocą przycisku **Choose** wybierz klasyfikator J48 (C4.5)

6. Upewnij się, że w oknie *Test options* zaznaczona jest opcja *Use training set*. Uwaga! **Nie** powinniśmy korzystać z tej formy testowania - dlaczego? Tutaj jesteśmy zmuszeni, z uwagi na niewielki zbiór uczący…
7. Kliknij w przycisk **Start** i zapoznaj się z uzyskanymi wynikami.
8. Następnie zwizualizuj drzewo tak jak to pokazano poniżej:



Porównaj wyniki ze swoimi obliczeniami.

Zad3. Poprawność klasyfikacji

- Załaduj plik `credit-g.arff` do Weki. Zawiera on dane uczące dla systemu, który na podstawie atrybutów zawartych w pliku powinien określać czy dany zestaw wartości atrybutów wskazuje na wiarygodnego klienta banku, czy też nie - czy można przyznać mu kredyt, czy jest to ryzykowne.
- Przejdź do zakładki **Classify** i wybierz algorytm J48.
- W obszarze *Test options* wybierz opcje *Percentage split* z wartością 66% Oznacza to, ze 66% danych posłuży do uczenia, a 34% do walidacji. Jakie to ma znaczenie?
- Uruchom algorytm. Ile procent przypadków zostało poprawnie zaklasyfikowanych? Czy to dobry wynik?
- Przejdź do zakładki **Preprocess** i zobacz jak wygląda rozkład atrybutu określającego czy danych zestaw jest *dobry* czy *zły*. Jaka byłaby skuteczność algorytmu który niezależnie od wartości atrybutów „strzelałby” że użytkownik jest wiarygodny?
- Dlaczego przed przystąpieniem do klasyfikacji, warto wcześniej przyjrzeć się danym?

oryginalne klasy	pozytywne (relewantne)	negatywne (nierelewantne)
pozytywne (wyszukane)	TP	FN
negatywne (niewyszukane)	FP	TN

TN (*true negative*) - liczba prawidłowych klasyfikacji stanu normalnego

FP (*false positive*) - liczba nieprawidłowych klasyfikacji stanu normalnego jako patologicznego

FN (*false negative*) - liczba nieprawidłowych klasyfikacji stanu patologicznego jako normalnego

TP (*true positive*) - liczba prawidłowych klasyfikacji stanu patologicznego

Wówczas wrażliwość odpowiada mierze: $$sensitivity = \frac{TP}{(TP + FN)}$$

zaś czułość definiowana jest jako: $$specificity = \frac{TN}{(FP + TN)}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$