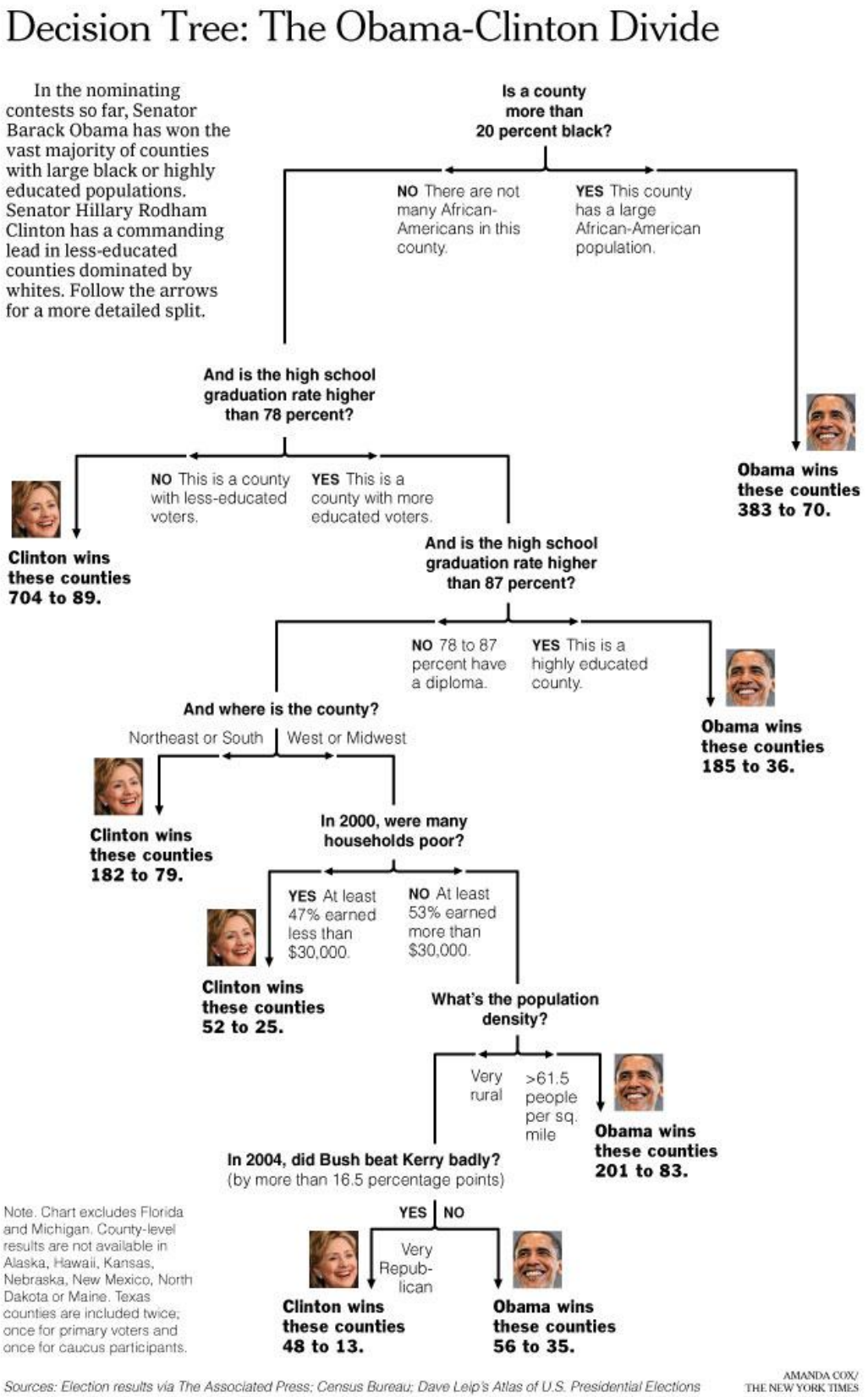
Drzewa decyzyjne

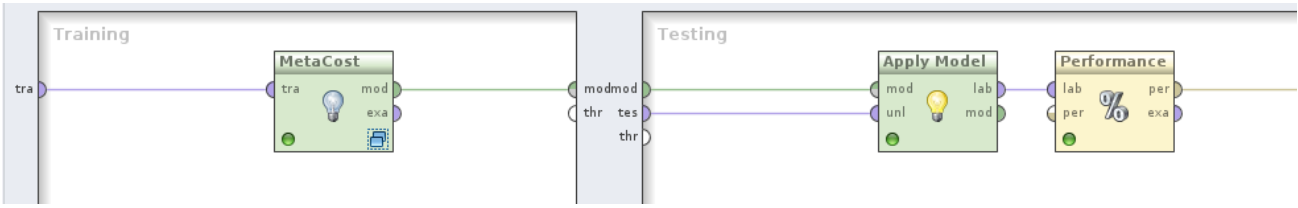
Celem laboratorium jest zapoznanie studentów z podstawowymi metodami indukcji drzew decyzyjnych.



Rapid Miner

- Uruchom narzędzie RapidMiner
- Utwórz prosty przepływ polegający na wczytaniu zbioru danych `bank.csv` i uruchomieniu operatora `Decision Stump`. Obejrzyj uzyskany model. Sprawdź, w jaki sposób zmiana kryterium podziału zbioru wpływa na kształt modelu. Zamień operator `Decision Stump` na operator `Decision Tree` i ponownie zbuduj oraz przeanalizuj model. Następnie wyłącz pre- i post-processing i sprawdź, jaki wpływ miało to na kształt modelu
- Dodaj do przepływu dyskretyzację atrybutów (operator `Discretize`, podział na 3 przedziały), oraz zmień operator `Decision Tree` na kolejno: `CHAID`, `ID3`, oraz `Decision Tree (weight-based)`. W ostatnim przypadku jako operator wewnętrzny do ważenia atrybutów wykorzystaj operator `Weight by Correlation`.
- Utwórz nowy przepływ zawierający operatory `Read ARFF` (wczytaj plik `mushroom.arff`), `Set Role` (wskaż atrybut class jako typu `label`), `Replace Missing Values` (pozostaw domyślne parametry), oraz uruchom proces walidacji krzyżowej wykorzystując operator `Cross-Validate`. Jako operatory wewnętrzne walidacji zastosuj najpierw `Decision Tree`, a potem `Random Forest`. Zaobserwuj zmiany w generowanych modelach, zwróć uwagę, jaki wpływ na model losowy ma zwiększenie puli dostępnych atrybutów.
- Uruchom narzędzie RapidMiner i załaduj zbiór `fertility.csv`. Zapoznaj się z opisem zbioru danych.
- Utwórz przepływ w którym za pomocą walidacji krzyżowej ocenisz jakość modelu drzewa decyzyjnego. Jako miarę oceny podziału przyjmij miarę *gini_index*. Obejrzyj uzyskany model i macierz pomyłek. Czy model dobrze radzi sobie z rozpoznawaniem nietypowych próbek?
- Przyjmij, że klasą pozytywną są normalne próbki spermy. Umieść operator `Decision Tree` wewnątrz operatora `Meta Cost` i zbuduj macierz kosztów, w którym błąd *false positive* będzie dwukrotnie droższy niż błąd *false negative*. Obejrzyj uzyskaną macierz pomyłek.
- Postaraj się uzyskać rozwiązanie, w którym czułość klasy “O” (ang. *recall*) przekroczy 50%. Co dzieje się z ogólną dokładnością klasyfikatora?
- Zamień operator `Decision Tree` na `Random Tree` i skonfiguruj operator w taki sposób, aby w każdej iteracji dysponował połową atrybutów do wyboru. Ponownie postaraj się tak skonfigurować macierz kosztów, aby uzyskać czułość klasy “O” powyżej 50% przy jak najwyższej ogólnej dokładności modelu.

- Zamień operator `Meta Cost` na operator `Tree to Rules` pozostawiając wewnątrz indukcję drzewa decyzyjnego przy użyciu miary *gini_index*. Obejrzyj uzyskany model regułowy.



Orange Data Mining

- Uruchom narzędzie Orange Data Mining. Pobierz zbiór danych `flags.tab` i zapoznaj się z jego opisem
- Użyj operatora `Select Attributes` do wskazania, który atrybut jest zmienną celu.
- Użyj operatora `Edit Domain` aby wartościom zmiennej celu nadać czytelne nazwy
- Postaraj się zbudować jak najdokładniejszy klasyfikator który umożliwia przewidywanie dominującej religii w państwie na podstawie cech charakterystycznych flagi tego państwa.
- Prześlij dane do operatora `Test Learners` i skonfiguruj operator w taki sposób, aby ocena modeli odbywała się na podstawie walidacji krzyżowej.
- Wykorzystaj jednocześnie: głosowanie większościowe, naiwny klasyfikator Bayesa, drzewo decyzyjne i klasyfikator regułowy. Obejrzyj tablicę z sumarycznymi wynikami. Czym różnią się od siebie poszczególne klasyfikatory?
- Wyślij wszystkie klasyfikatory do operatora `ROC` i obejrzyj uzyskane krzywe.