Decision Tree

Jest to supervised learning algorytmem, który może być wykorzystany zarówno do regresji jak i klasyfikacji. Może wykorzystywać zarówno dane ciągłe jak i atrybutowe.

Podstawowa terminologia:

Root Node – reprezentuje całą populację. Jest ona następnie dzielona na dwa lub więcej homogenicznych podzbiorów.

Splitting – proces rozdziału *node* na dwa lub więcej *sub-nodes.*

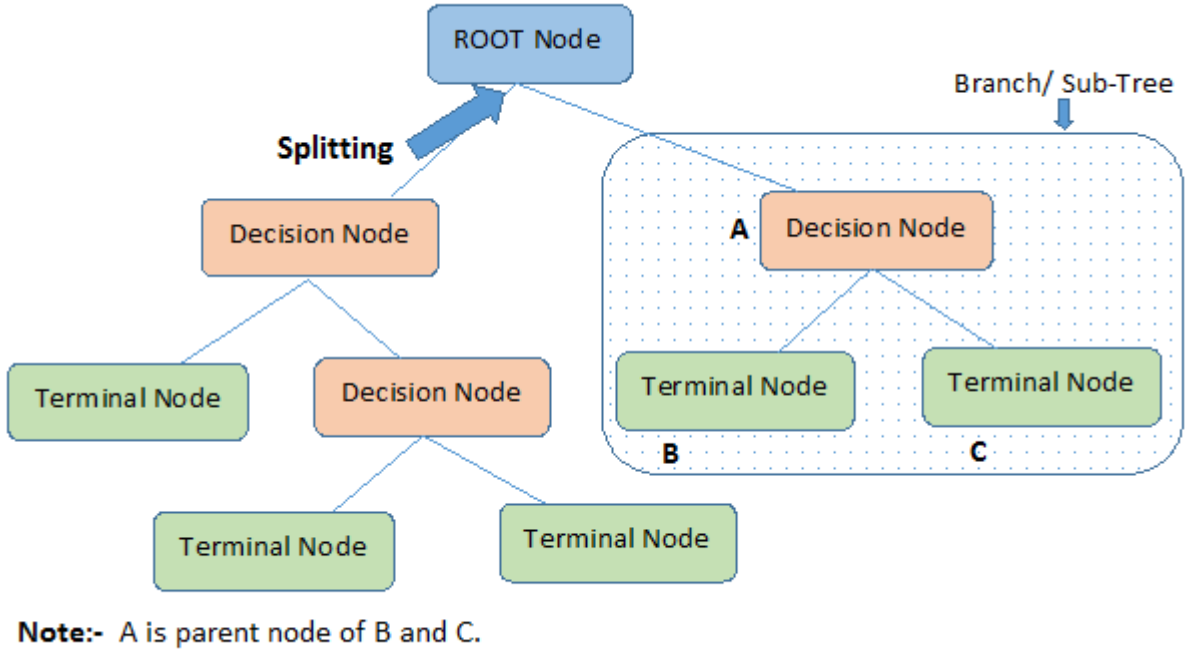
Kiedy *sub-node* rozdziela się na kolejne *sub-nodes* jest nazywany **Decision Node**

Nodes które się nie rozdzielają to **Terminal Node** lub **Leaf**

Usuwanie *sub-nodes* nosi nazwę **Pruning** (przeciwne działanie *Splitting*)

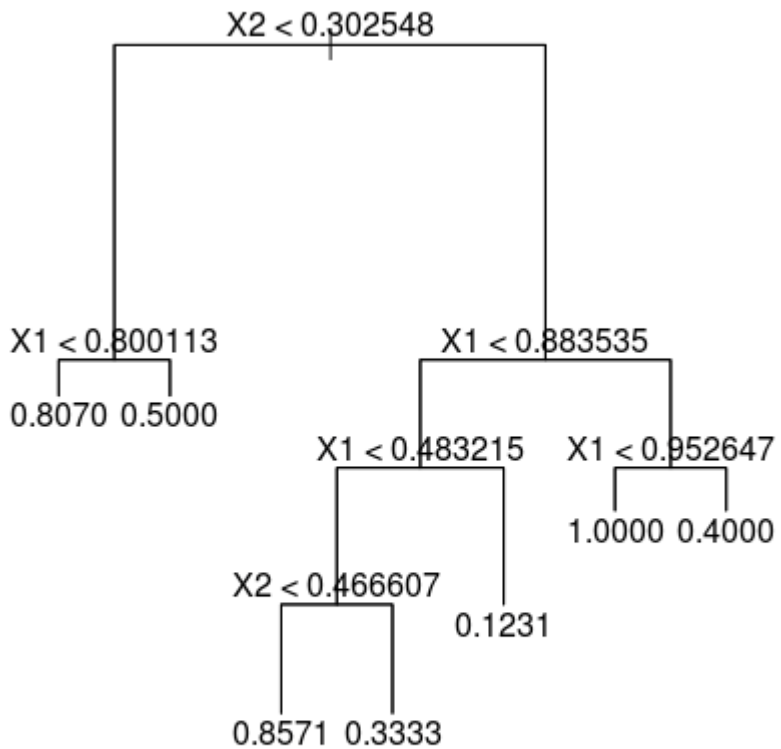
Wydzielona część całego drzewa – **Branch**

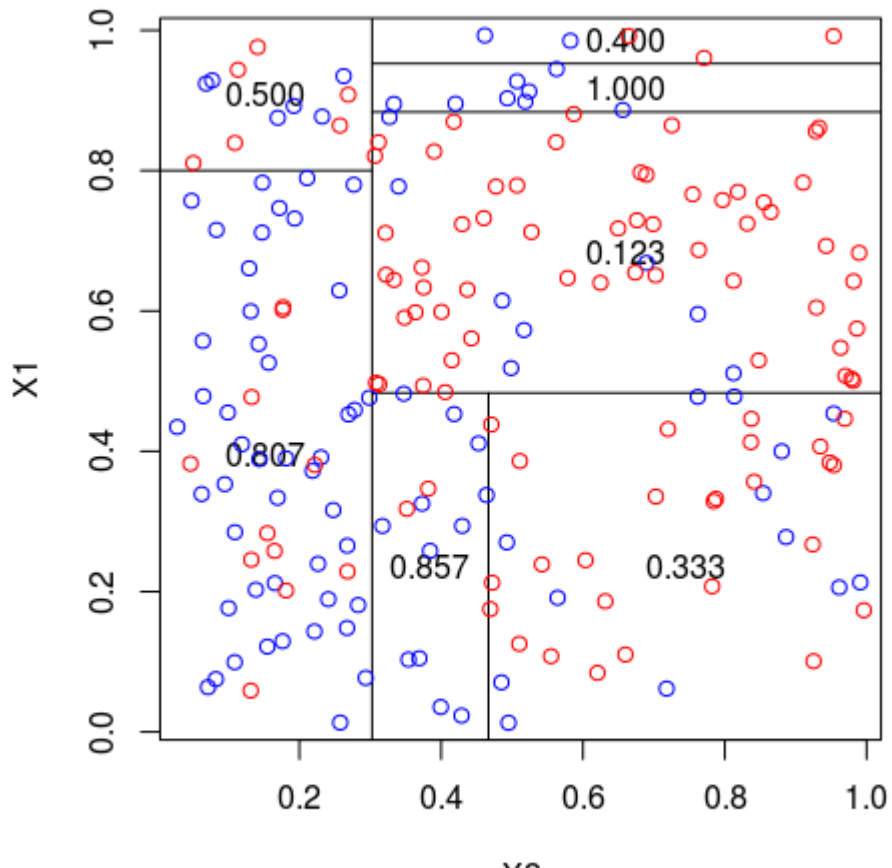
Node, który jest dzielony na *sub-nodes* to **parent node** a *subnode* to **childs** of *parent node*.



Note:- A is parent node of B and C.

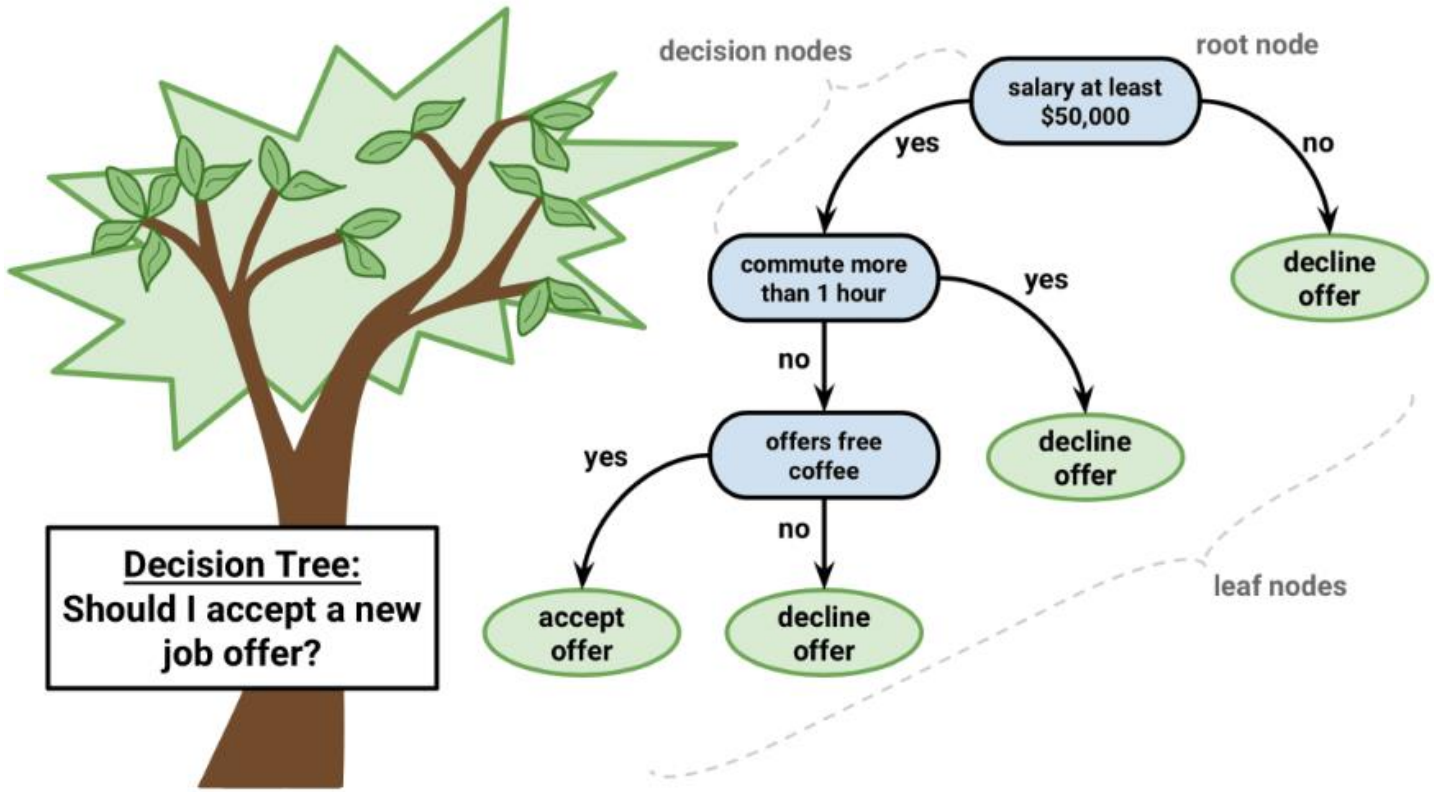
Regression Tree





Dopasowanie drzewa do losowych wartości – splitting jest przypisany do jednej osi. W rezultacie w przestrzeni wielowymiarowej będą tworzone p-dimensional „hyperblock”

Classification Tree



Zalety

Intuicyjna metoda, łatwa do wytłumaczenia. Odzwierciedla dobrze sposób podejmowania decyzji przez człowieka. Może być przedstawiona graficznie.

Wady

Brak dokładności predykcyjnej (małe zmiany w danych mogą prowadzić do dużych zmian w końcowym drzewie, podzial na dwie części w różnych proporcjach da różne wwyniki).

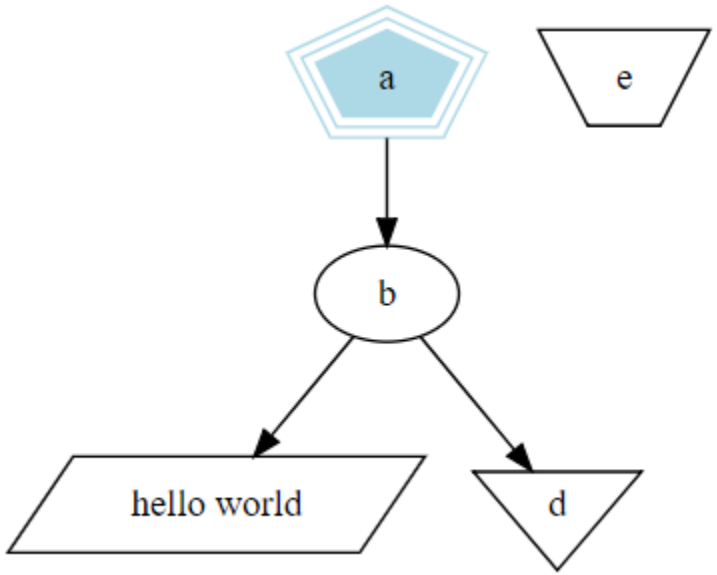
Rozwiązaniem jest agregacja wielu decision trees przy użyciu takich metod jak**: bagging, random forest, boosting** dokładność predykcyjna może być znacznie poprawiona)

WIZUALIZACJA

www.webgraphviz.com

dotguide.pdf

```
digraph G {
 a -> b -> c;
 b -> d;
 a [shape=polygon,sides=5,peripheries=3,color=lightblue,style=filled];
 c [shape=polygon,sides=4,skew=.4,label="hello world"]
 d [shape=invtriangle];
 e [shape=polygon,sides=4,distortion=.7];
 }
```



WebGraphviz is Graphviz in the Browser

Enter your graphviz data into the Text Area:

(Your Graphviz data is private and never harvested)

Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5

```
digraph g{
  rankdir=LR;
  "webgraphviz" -> "@" -> "gmail" -> "." -> "com"
}
```

Generate Graph!

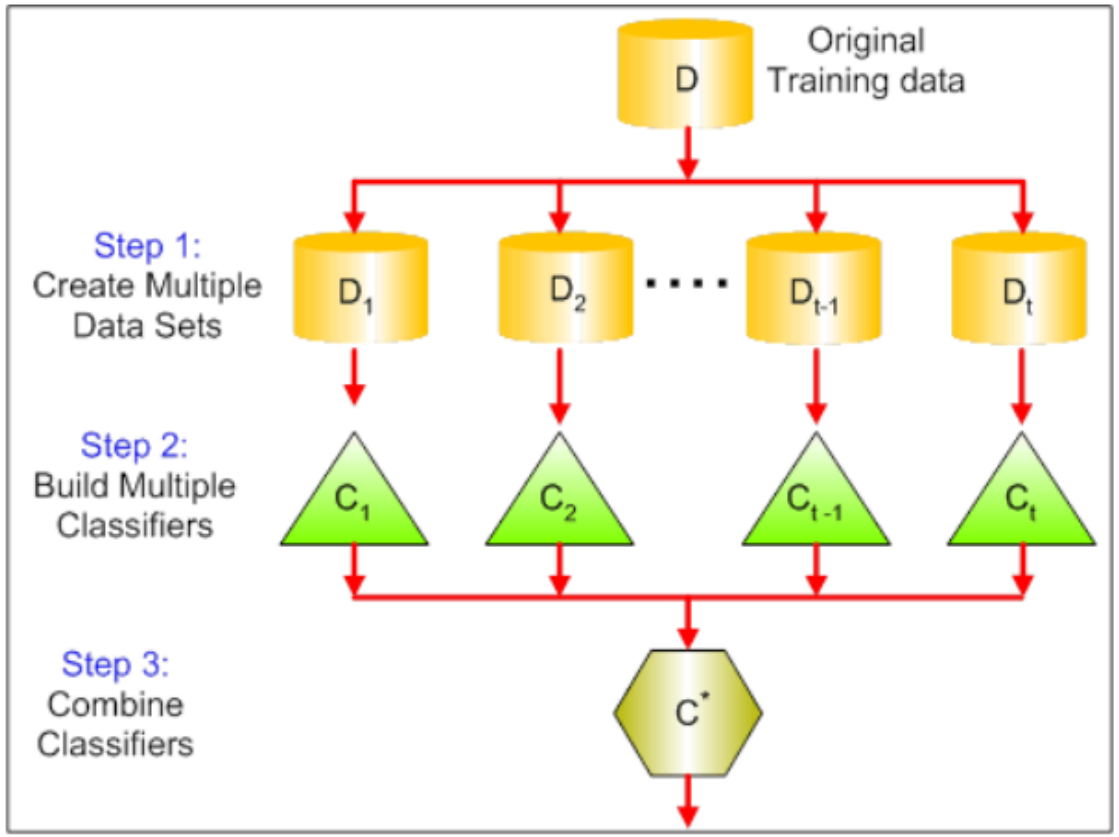
Bagging lub bootstrap aggregation

Jest techniką mającą na celu redukcję wariancji predykcji. Osiąga się to przez łączeniewielu klasyfikacji na podzbiorach tego samego zbioru.

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x)$$

B – różnych podzbiorów treningowych

$\hat{f}_{bag}(x)$ - uśredniona wartość predykcji



1. Rozbijamy zbiór danych na szereg podzbiorów (nawet setki lub tysiące drzew)
2. Budujemy klasyfikatory (implementacja algorytmu) dla każdego podzbioru (models i predictions)
3. Wyznaczamy średnią wartość ze zbioru predykcji.

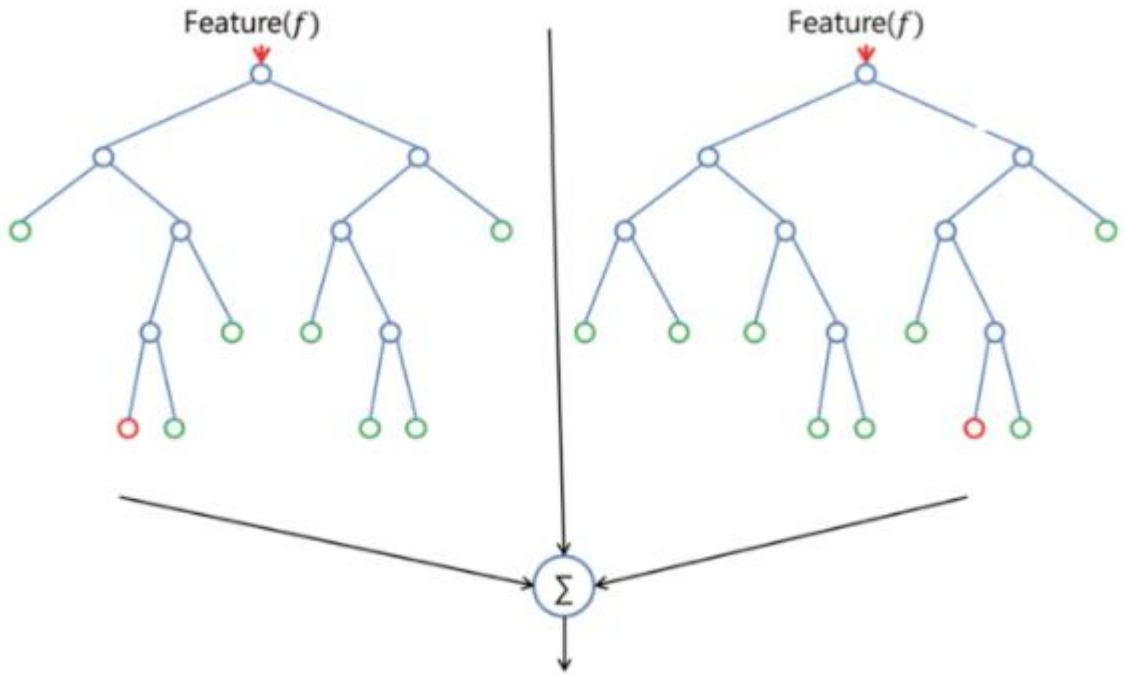
Random forest

Wszechstronny, łatwy do użycia algorytm ML który tworzy, nawet bez optymalizacji hiper-parametrów bardzo dobre modele. Jest jednym z najczęściej używanych algorytmów , ze względu na jego prostotę oraz dlatego, że może być używany zarówno do klasyfikacji i regresji.

Random forest jest supervised algorytmem – tworzy „forest” składający się z wielu Decision Trees i łączy je razem aby otrzymać lepsze i stabilniejsze wyniki predykcji.

Random Forest wprowadza (oprócz redukcji *dimensions*, uwzględniania *missing values*, *outliers*) istotne udoskanelenie które decorraletes trees.

Podobnie jak w bagging budowanych jest wiele drzew decyzyjnych, ale kiedy są one budowane wprowadza się w ten proces randomizację na 2 etapach (losowego wybierania punktów i wybierania losowo części features).



Random forest z dwoma drzewami

Hiperparametry:

n_estimators

Liczba drzew budowanych prze algorytm, zanim nastąpi uśrednianie. Generalnie większa liczba drzew zwiększa stabilność i wydajność modelu.

max_features

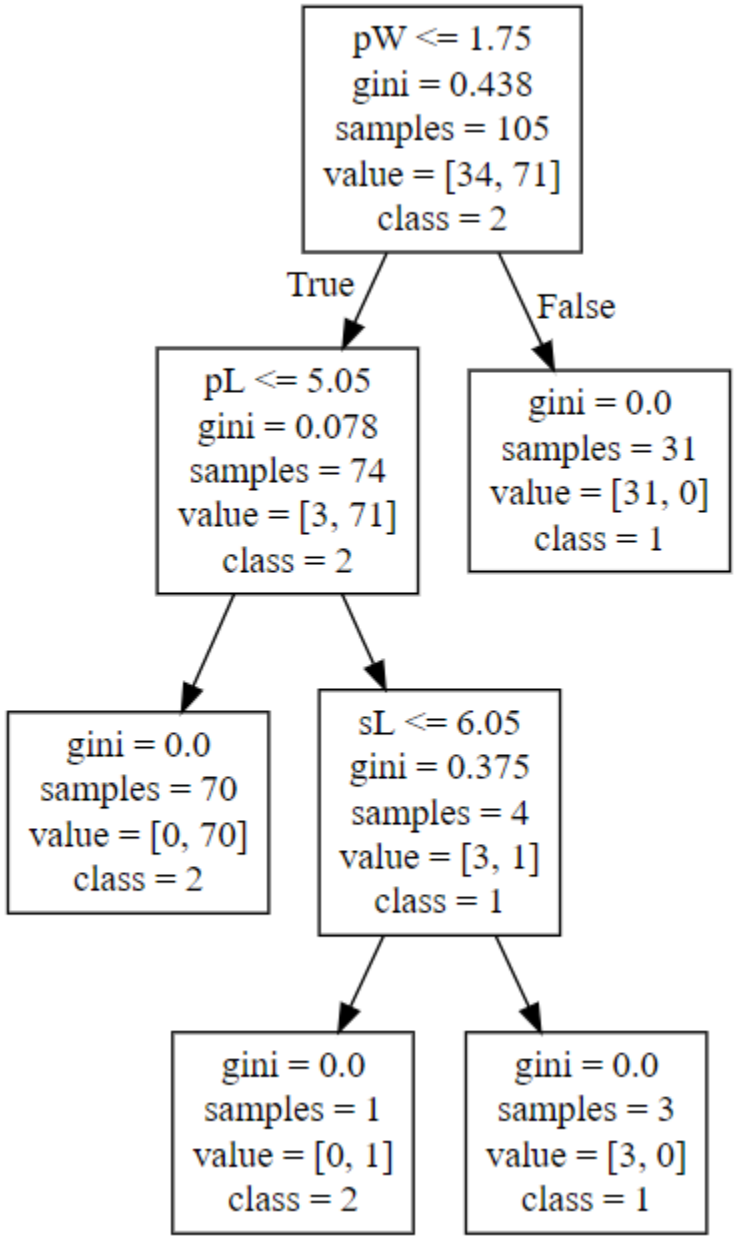
Maksymalna liczba features po której następuje rozdzielenie

min_sample_leaf

Minimalną liczbę of leafs, wymaganą do rozdzielenia.

random_state

Zapewnia powtarzalność modelu



samples - liczba elementów w node

value – liczebność klas

gini - gini impurity of the node (jak bardzo klasy są zmieszane)