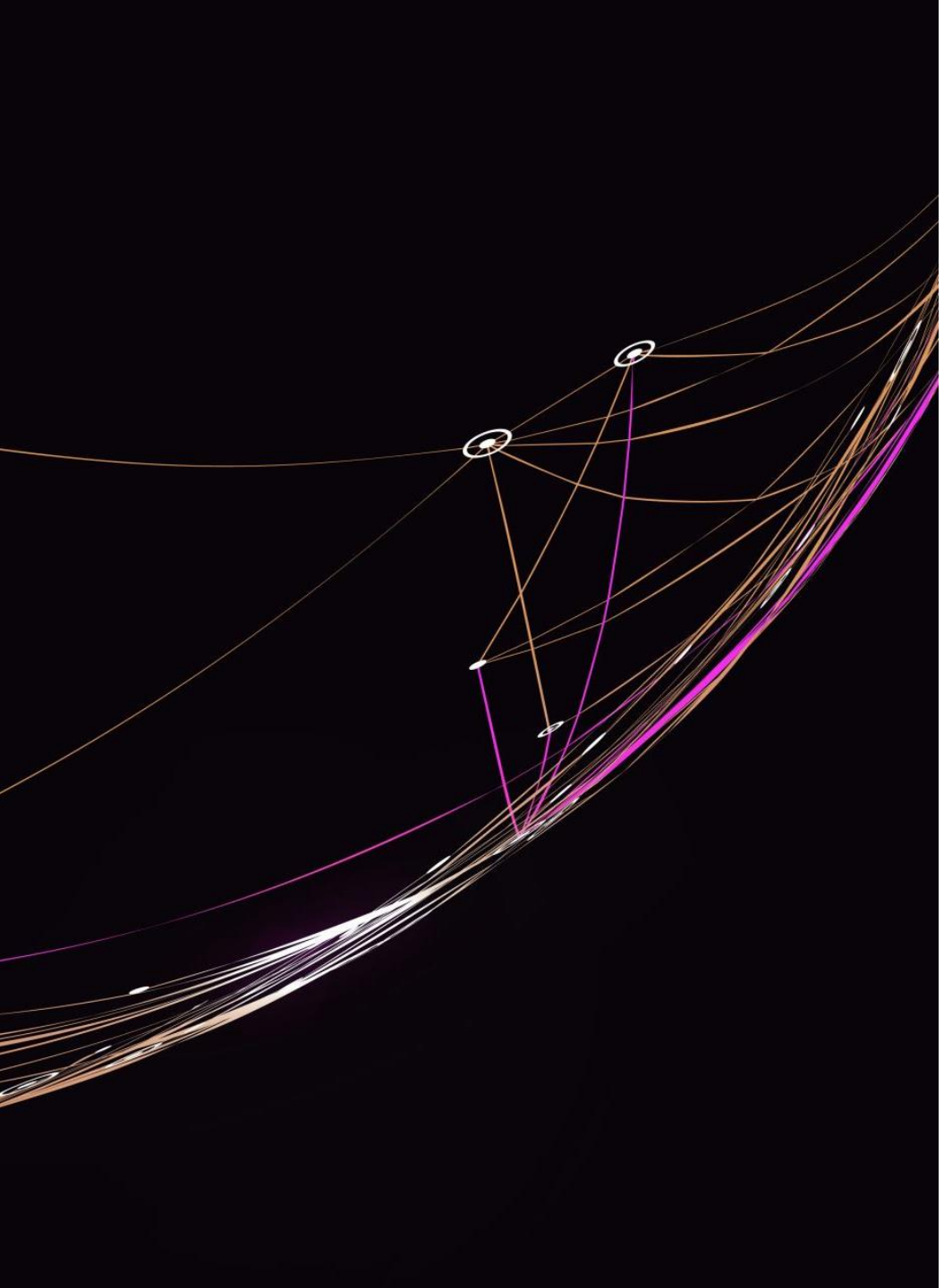


ADMINISTRAÇÃO

IBM0112 DATA MINING

Árvores de decisão

Cassius Figueiredo

Características

- Uma árvore de decisão chega a uma predição por meio de uma série de perguntas relacionadas a pertencer ou não à determinados grupos.
- Cada questão deve ter apenas duas respostas (sim ou não), o que leva à geração de uma árvore binária.

Características

- Iniciamos pela pergunta conhecida como nó raiz (*root node*) e vamos percorrendo a árvore por seus ramos de acordo com os fatores que levem a decidir por determinado grupo ou não até chegar em uma folha (*leaf node*).
- A proporção alcançada na folha indicará a probabilidade procurada, no caso de uma aplicação em problemas supervisionados de classificação.

Motivação

- Em áreas de negócios, por exemplo, as árvores de decisão podem ser utilizadas para definir perfis de consumidores ou até previsões de quem poderá pedir demissão.
- Na área financeira são utilizadas para precificar ativos.
- Em gestão de projetos, possui aplicações em análise de riscos.

Como funciona?

© Can Stock Photo - csp39294735

Árvores de regressão vs. classificação

- As árvores de regressão são utilizadas quando a variável dependente é contínua. Já as árvores de classificação são utilizadas para variáveis dependentes categóricas.
- No caso das árvores de regressão, o valor obtido nas folhas é o valor médio das observações que estão naquela região. Logo, se um novo dado a ser avaliado cair na mesma região, a predição será feita por meio deste valor médio.

Árvores de regressão vs. classificação

- Ambas as árvores dividem o conjunto de variáveis independentes (também conhecido por espaço de predição) em regiões distintas, sem interseção.
- Para simplificar, pode-se considerar estas regiões independentes como nossa caixa divisória com parafusos e pregos.

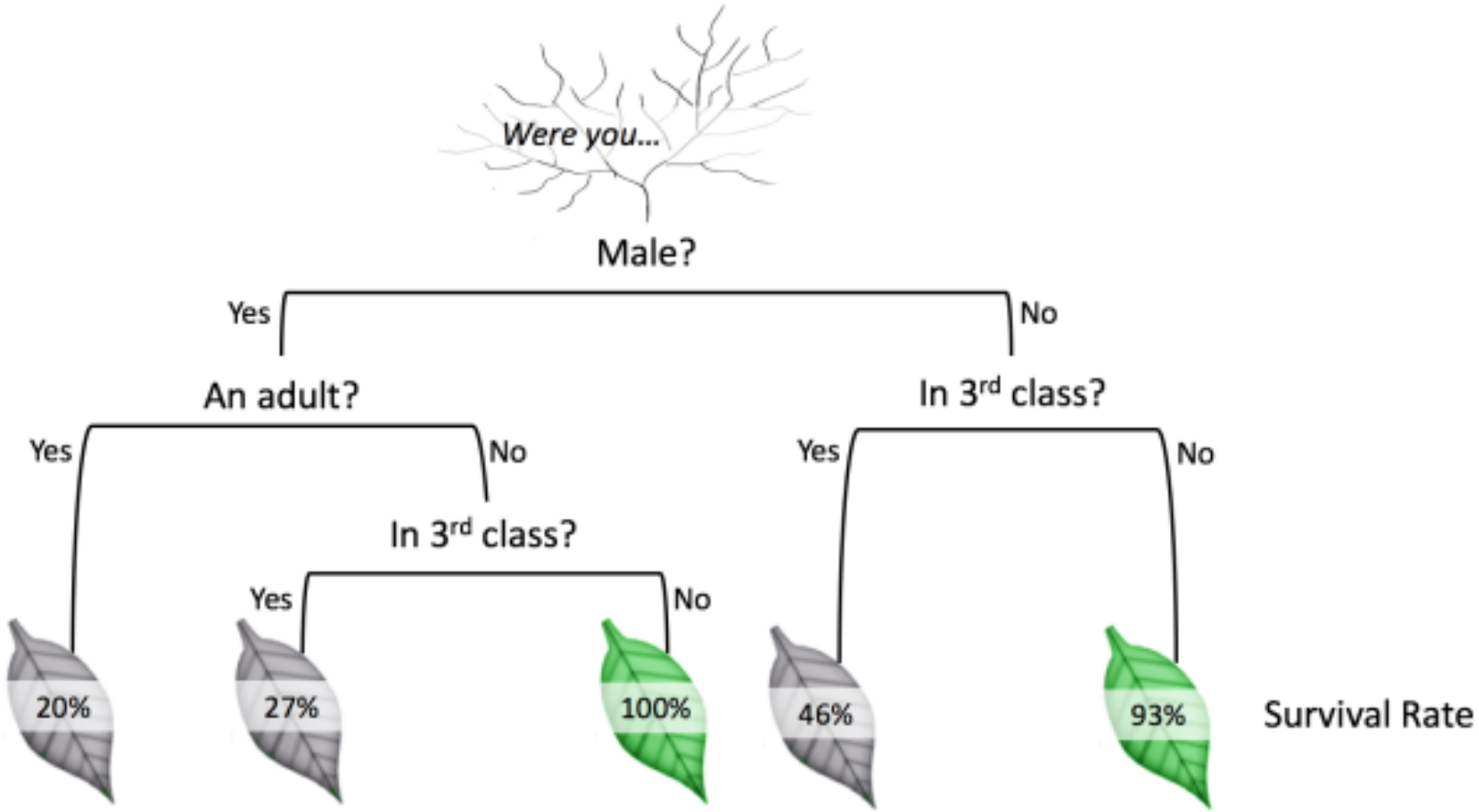
Árvores de regressão vs. classificação

- Ambas as árvores seguem uma abordagem top-down gulosa (greedy, em inglês) conhecida por divisão recursiva binária (recursive binary splitting).
- É “top-down” pois inicia pelo “topo” da árvore, onde todas as observações pertencem à mesma região e vai dividindo o conjunto de forma binária em quebras sucessivas.
- É chamado “gulosa” pois o algoritmo avalia uma quebra de cada vez, sem se preocupar com futuras quebras.

Exemplo

- Vamos ver um exemplo relacionado ao acidente ocorrido com o navio Titanic.
- O objetivo é saber quais grupos de passageiros teriam maior probabilidade de sobrevivência.
- O conjunto de dados foi compilado originalmente pelo British Board of Trade, para investigar o acidente.
- A imagem a seguir apresenta o resultado do algoritmo, após sua execução.

Exemplo


Were you...
Male?
Yes
No
An adult?
In 3rd class?
Yes
No
Yes
No
In 3rd class?
Yes
No
20%
27%
100%
46%
93%
Survival Rate

Exemplo

- Avaliando a árvore gerada, podemos ver que a probabilidade de sobrevivência seria maior se alguém pertencesse ao grupo de mulheres das cabines de primeira e segunda classes ou ainda ao grupo de crianças do sexo masculino, também das cabines de primeira e segunda classes.

Criando uma árvore de decisão

- Já percebemos o quão fácil é interpretar uma árvore de decisão.
- Vamos ver a seguir como elas são geradas.

Algoritmo

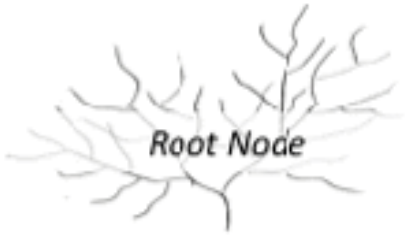
- A árvore de decisão começa pela divisão do grupo inicial em dois subgrupos, ambos com dados similares.
- A seguir, repete-se o procedimento de divisão binária em cada subgrupo.
- Desta forma, a cada divisão teremos uma menor quantidade de dados, porém estes serão mais homogêneos.

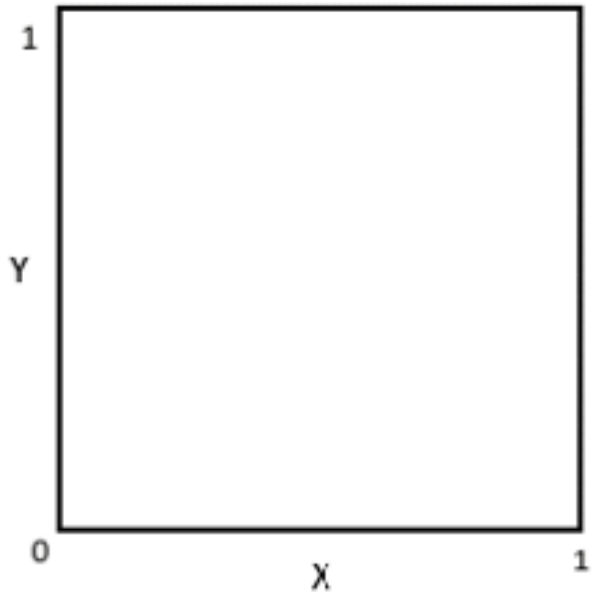
- O princípio das árvores de decisão é baseado no fato que se isolarmos os diferentes grupos em ramos diferentes da árvore, todos que pertençam a estes grupos terão uma previsão similar.

Algoritmo

- O processo de particionar dados visando a obtenção de grupos homogêneos é chamado partição recursiva (recursive partitioning) e envolve apenas dois passos:
  - Descobrir o fator binário que divida o conjunto de dados em dois grupos, da forma mais homogênea possível.
  - Repetir o passo 1 em cada um dos subgrupos até que determinada condição de parada seja alcançada.

Algoritmo





Critérios de parada

- Critérios de parada podem ser definidos de várias formas, exemplos:
  - Parar quando os dados de uma folha pertencerem a uma determinada categoria/valor;
  - Parar quando uma folha tiver menos de cinco dados;
  - Parar quando novas divisões não melhorarem a homogeneidade do subgrupo.

Critério para divisão de grupos

- Um importante conceito para que algoritmos como árvores de decisão funcionem, corretamente é o critério de divisão a ser utilizado.
- A pureza pode ser compreendida com o quanto um grupo é homogêneo, em sua constituição. Porém ainda assim, a homogeneidade está fundamentada no processo matemático que utilizamos para avalia-la:
  - Gini Index
  - Information Entropy
- E qual a relevância? A métrica de pureza escolhida pode impactar o resultado de seu modelo e deve ser avaliada.

Gini vs. Entropy

- Tecnicamente, devem ser testadas no cenário em questão para avaliação.
- O Gini Index varia entre 0 e 0,5 enquanto a entropia varia entre 0 e 1.
- A entropia é mais cara computacionalmente, por conta da inclusão de logaritmos em sua fórmula.
- Dois excelentes artigos, com exemplos, seguem, para quem quiser entender um pouco mais como funcionam.
  - Gini Index vs Information Entropy
  - Decision Trees: Gini vs Entropy

Flexibilidade

- A partição recursiva faz uso apenas das melhores perguntas binárias para formar a árvore de decisão.
- Desta forma, a presença de variáveis irrelevantes não afeta o resultado final.
- Além disso, as questões binárias impõe uma divisão central dos dados, logo, as árvores de decisão são bem robustas quanto aos valores extremos (i.e. outliers).
- Aceita mistura de variáveis contínuas e discretas.
- Não exige nenhum processo de normalização de variáveis para que todas estejam na mesma escala.

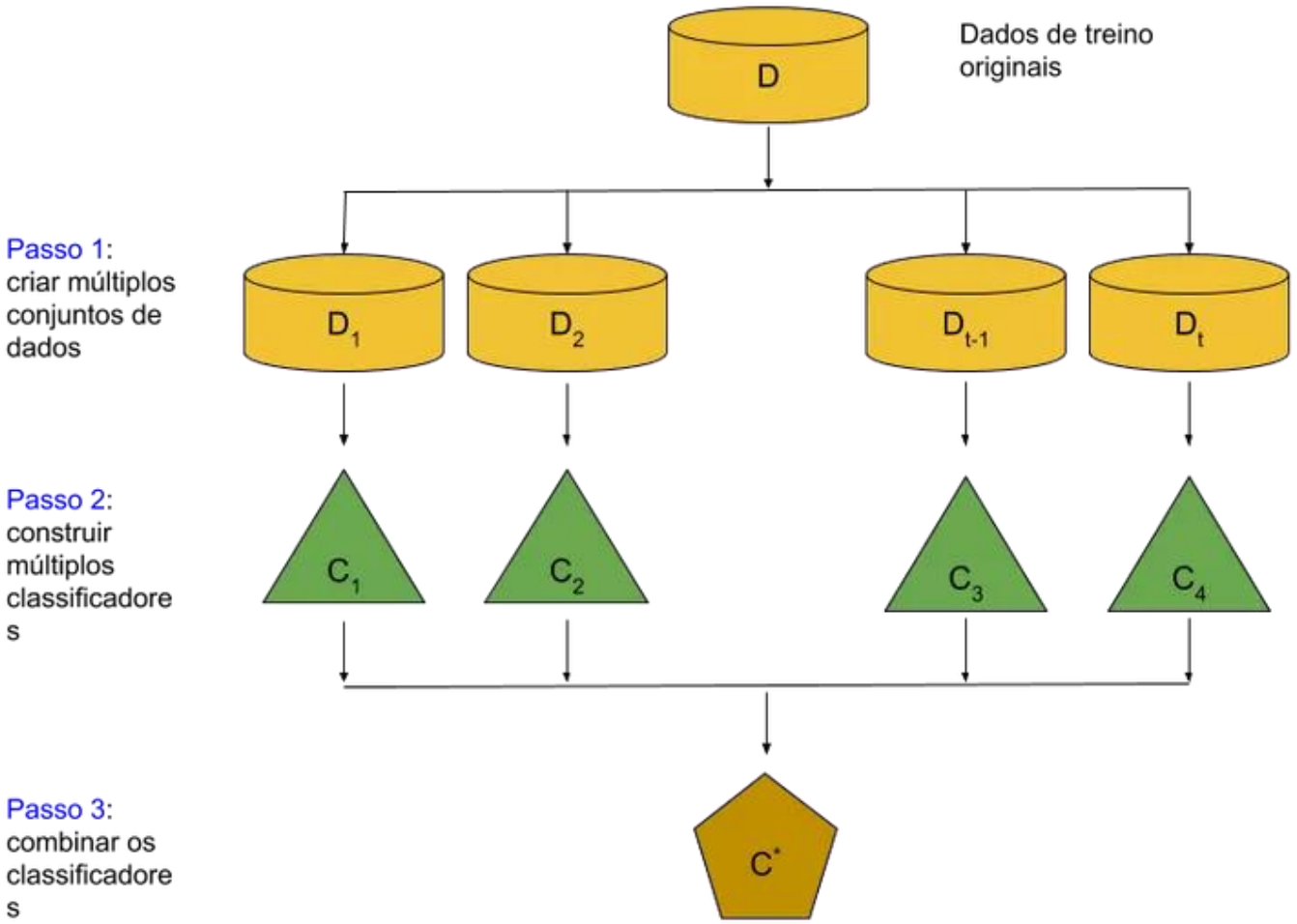
Limitações

- Fazer uso das melhores questões binárias para divisão dos dados pode não levar às predições mais precisas.
- Utilizar divisões assimétricas no início do processo pode levar a predições melhores ao final.
- Ao usarmos uma única árvore como modelo para nosso problema podemos ser severamente afetados por problemas de variância.

Ensembles - Bagging

- É uma técnica usada para reduzir a variância das previsões.
- Combina o resultado de vários classificadores, modelados em diferentes sub-amostras do mesmo conjunto de dados.
- Os classificadores são combinados usando média, ou mediana, de acordo com a necessidade.
- Os valores combinados são geralmente mais robustos do que um único modelo.

Bagging


D
Dados de treino originais
Passo 1: criar múltiplos conjuntos de dados
D_1
D_2
D_{t-1}
D_t
Passo 2: construir múltiplos classificadores
C_1
C_2
C_3
C_4
Passo 3: combinar os classificadores
C^*

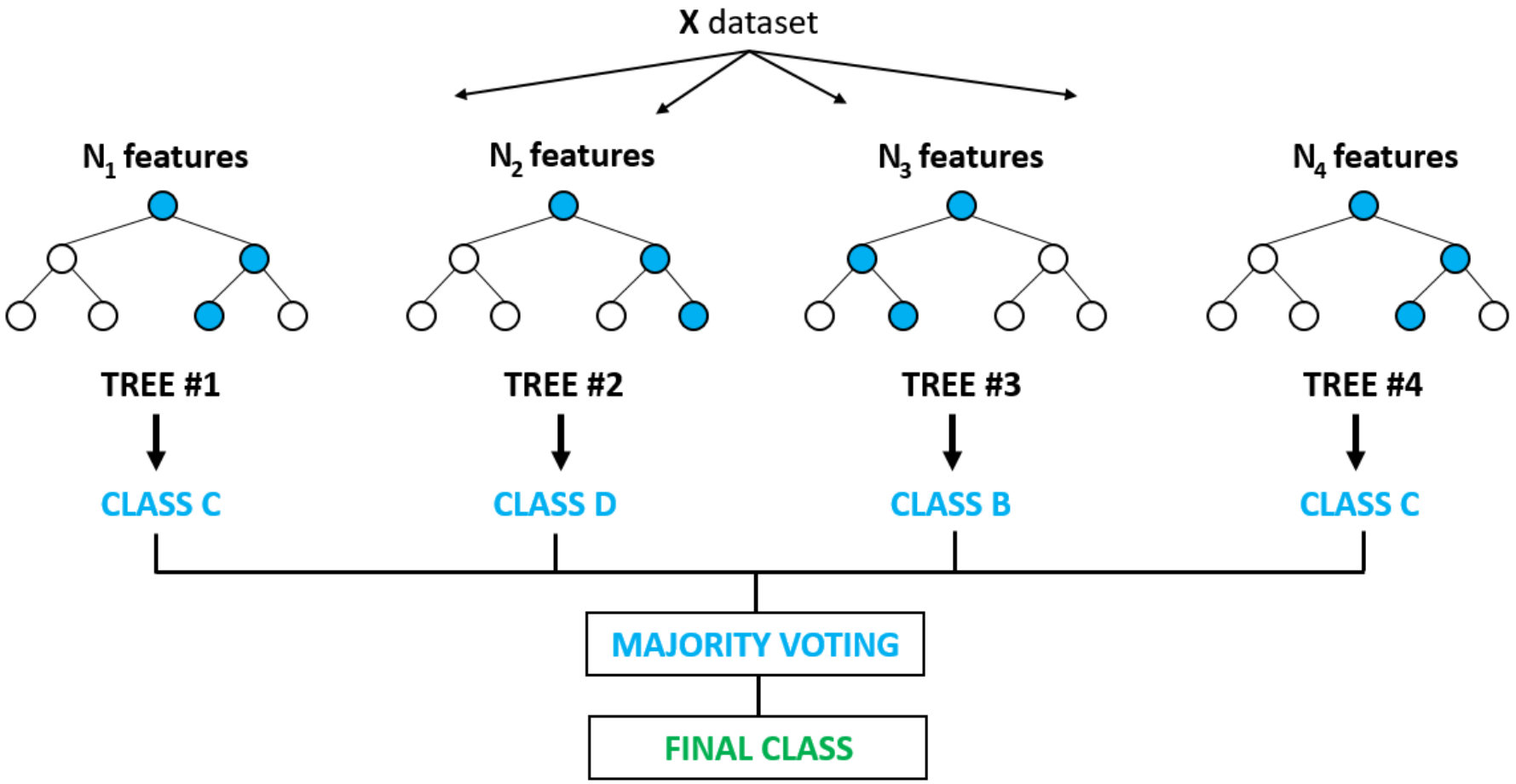
Random Forest

- É uma evolução das árvores de decisão, com o objetivo de reduzir a variância das previsões.
- É considerada o "canivete suíço" em ciência de dados. Ou seja, quando você não consegue pensar num algoritmo (seja qual for a situação), vale a tentativa de usar uma Random Forest!

Random Forest

- Random Forest é um método de aprendizagem de máquina versátil e capaz de executar tarefas de regressão e de classificação.
- É um tipo de método de aprendizado chamado ensembles, onde modelos são combinados para formar um modelo mais forte.

Random Forest



Fonte: Abilash R.

Random Forest - Vantagens

- Aplicável a problemas de classificação e regressão, porém melhor performance em problemas de classificação.
- Trata grandes volumes de dados e com muitas dimensões.
- Implementa métodos para equilibrar erros em conjuntos de dados onde as classes são desbalanceadas.

Random Forest - Desvantagens

- Apesar de resolvê-los, não é tão boa para problemas de regressão, uma vez que não fornece previsões precisas para variáveis contínuas.
- Pode ser considerada como uma caixa preta para quem faz modelagem estatística – temos muito pouco controle sobre o que o modelo faz. Na melhor das hipóteses, conseguimos apenas experimentar diferentes parâmetros.
- Perde-se a capacidade de geração de regras da árvore de decisão simples. Pode-se gerar, no máximo, um indicador de importância da variáveis.

Ensembles - Boosting

- Outra opção é selecionar estrategicamente as árvores (no lugar de uma seleção aleatória) de forma que a predição de cada uma das árvores geradas melhore gradativamente.
- A ideia é transformar modelos de aprendizagem fracos em fortes.
- Esta técnica é chamada boosting.
- Gradient Boosting (GBM), LightGBM e XGboost, são exemplos de aplicação desta técnica.